Predictive Analytics for Wind Power Forecasting

PhD Thesis

Rosemary Tawn

Supervisors: Jethro Browell and David McMillan

Institute for Energy and Environment

Department of Electronic and Electrical Engineering

University of Strathclyde, Glasgow

May 8, 2022

Abstract

At the most basic level, forecasts are needed when decisions must be made in the present but are based on future conditions. There are many uses for forecasts within the energy industry, and in particular in relation to renewable generation types where future generation is uncertain and depends on weather conditions, be that wind, solar irradiation, cloud cover or sea conditions. Thus, forecasts are an essential part of the running of the electricity system, from day-to-day scheduling and trading decisions to long-term system planning. This thesis selects and aims to address three current problems in the practical implementation of wind power forecasting methods: firstly, the affect of several different occurrences of missing data in the forecasting process and how this can be mitigated; secondly, the difficulty in producing a forecast that accurately predicts ramps; and thirdly, the need for skilful site- and task-specific forecasts weeks ahead to inform maintenance decisions. Missing data is the result of incomplete datasets, data latencies and new sites lacking historic power data. Missing historic values affect the ability of statistical models to learn site characteristics and fit an accurate model while data latencies mean forecast inputs are not always available when new forecasts are issued. Several different mitigation methods for each occurrence of missing data are explored via case studies for a Vector Autoregressive model implementation. Complex relationships between wind speed and power, sudden ramps in power and imperfect models can reduce the skill of individual forecast models. Improvements through combination of several different forecasting models is explored and a forecast combination method that explicitly incorporates forecasts of ramp rate proposed to improve power forecasts around times of ramps. A lack of suitably tailored forecasts for a given decision also reduces the likelihood of uptake of a new model or data source for certain applications. A novel job-specific index of ‘number of useful hours’ is forecast for subseasonal-to-seasonal timescales and its calibration and usefulness for the case of crane hire for maintenance decisions is assessed. This work has also produced guidance and recommendations for the implementation of a very short-term statistical forecasting system for Natural Power Consultants.

Contents

List of Figures

List of Tables

List of Acronyms

ANFIS	Adaptive Neuro-Fuzzy Inference System
AR	Autoregressive
AR(I)MA	Autoregressive (Integrated) Moving Average
BM	Balancing Mechanism
CDF	Cumulative Distribution Function
CEEMDAN	Complete Ensemble Empirical Mode Decomposition with Adaptive Noise
CRPS	Continuous Ranked Probability Score
EA(WR)	East Atlantic (Western Russia)
ECMWF	European Centre for Medium-range Weather Forecasts
ELM	Extreme Learning Machine
EMD	Empirical Mode Decomposition
EMOS	Ensemble Model Output Statistics
ESO	Electricity System Operator
GAM(LSS)	Generalised Additive Model (of Location, Scale and Shape)
GBM	Gradient Boosted Machine
IMF	Intrinsic Mode Function
LASSO	Least Absolute Shrinkage and Selection Operator
LSTM	Long Short-term Memory

MAPE	Mean Absolute Percentage Error
MAR	Missing at Random
MCAR	Missing Completely at Random
MC(MC)	Markov Chain (Monte Carlo)
MJO	Madden-Julian Oscillation
MIDAS	Met Office Integrated Data Archive System
ML	Machine Learning
MNAR	Missing Not at Random
NAO	North Atlantic Oscillation
(N)MAE	(Normalised) Mean Absolute Error
NN	Neural Network
NWP	Numerical Weather Prediction
OLP	Optimal Linear Pool
PC(A)	Principal Component (Analysis)
PDF	Probability Density Function
PPV	Positive Predictive Value
RF	Random Forest
RMSE	Root Mean Squared Error
SCA	Scandinavian Pattern
SCADA	Supervisory Control and Data Acquisition
SSA	Singular Spectrum Analysis
SVM	Support Vector Machine
SVR	Support Vector Regression
S2S	Subseasonal-to-seasonal
TPR	True Positive Rate

TSO	Transmission System Operator
VAR	Vector Autoregressive
VMD	Variational Mode Decomposition

Statement of Authorship

Output	Contribution	Chapter relates to
Paper 'A review of very short-term wind and solar power forecasting' [3]	Literature review; case study implementation; writing of paper except first half of introduction and forecast evaluation section of case study, which the first draft was written by Jethro Browell.	2
Paper 'Missing data in wind farm time series: Properties and effect on forecasts' [4]	Literature review; case study implementation of methods with guidance from supervisor; 95% of paper writing; presentation at PSCC.	3
Paper 'Quantile combination for the EEM20 wind power forecasting competition' [5]	25% code development; report sections on normalisation and quantile combination; 50% of conference presentation.	4
Presentation 'Forecast combination and adaptation for improved ramp forecasts', *Wind Energy Science Conference*, online, May 2021	Literature review; case study implementation of methods with guidance from supervisor; conference presentation.	4
Paper 'Subseasonal-to-seasonal forecasting for wind turbine maintenance scheduling', accepted for publication in *Wind* (May 2022)	Forecast generation and evaluation (supervision from Jethro Browell), first draft writing. Editing joint with Jethro Browell.	5

Acknowledgements

I would like to acknowledge Natural Power Consultants Ltd and The DataLab for their sponsorship of this PhD project and support along the way, including advice and support as well as sharing of data. Many thanks to my supervisors, Iain Dinwoodie, David McMillan and particularly Jethro Browell, for his regular advice and always insightful suggestions. Thanks also to my family, Tom, Edwyn and my cohort of fellow students for their support.

Chapter 1

Introduction

Arrhenius first discovered that carbon dioxide released from burning of fossil fuels exacerbates the atmospheric greenhouse effect in 1896 [6]. Between 1750 and 2021, humans have emitted over 1.65 trillion tonnes of CO_2 [7], resulting in an average temperature rise of 1.19°C above pre-industrial levels [8]. In 2018 the Intergovernmental Panel on Climate Change published a special report on the impacts of global warming exceeding 1.5°C [9], detailing climate projections for both 1.5 and 2°C of warming above pre-industrial levels. They find limiting warming to 1.5°C is necessary to keep sea level rise, ocean acidification, species loss (both in the oceans and on land) and security of food and water supply within safe limits. Modelled pathways staying within or very close to 1.5°C require reaching net zero emissions by 2050 with a 45% reduction in anthropogenic CO_2 emissions from 2010 levels by 2030. This scale of reduction requires large scale international change across all sectors. The UK has the 8th largest cumulative carbon emissions by country [7] and as such has an even stronger responsibility to rapidly reduce emissions than countries in the global south.

Reducing dependence on fossil fuels and moving to renewable energy generation is a significant part of the transition to net zero. Most methods of renewable energy generation are methods of electricity production. While the overall share of energy provided by electricity in the UK was 17% in 2019 [10], this is forecast to increase as technologies such as electric vehicles and the electrification of heat are more widely adopted. Table 1.1 shows the lifetime carbon intensity of different electricity generation

Table 1.1: Total lifetime carbon intensity estimates for electricity by generation type, including direct emissions, infrastructure, and methane. From the IPCC AR5 report [1]; this was published in 2014 so technological improvements and decarbonisation of the grid are likely to have decreased emissions further from these figures.

Technology	Carbon footprint (gCO2eq/kWh)
Coal	740–910
Gas	410–650
Nuclear	5.5–26
Solar PV	18–180
Onshore wind	7–56
Offshore wind	8–35
Marine	5.6–28
Geothermal	6–79
Hydropower	1–2200

methods, and it is clear that coal and gas produce significantly more emissions per kWh generated than other technologies. Wind, marine and nuclear show the lowest carbon footprints and efficiency improvements in technologies since this data was published in 2014 may well lower these further. Nuclear projects, however, are significantly more expensive than new wind power capacity (the 2021 strike price agreed for Hinckley point C, the newest nuclear plant being built in the UK, is £106/MWh [11] whereas all the strike prices in the latest wind Contracts for Difference round are below £50/MWh [12]) and marine energy technologies are not yet mature enough to be built at scale.

One of the primary benefits of renewable energy is its role in displacing energy generation by fossil fuels and the associated effects not only on the climate but also on health and wellbeing. Air pollution from fossil fuel-related emissions was responsible for 3.61 million excess deaths worldwide in 2015 [13]. Reducing consumption of fossil fuels also reduces demand on shipping: an estimated 40% of all goods transported by the shipping industry are just fossil fuels for burning elsewhere in the world [14]. The shipping industry is itself responsible for 2.8% of global emissions [15], so simply reducing transport of fossil fuels would have a significant impact on global emissions even before a reduction in emissions from the end fossil fuel use is taken into account. Fossil fuels also have wide ranging negative effects on ecosystems, from both routine pollution for example groundwater contaminated by coal mines [16] and from one-off events like the Deepwater Horizon oil spill which caused serious harm to all species in the area, from ‘unprecedented’ deaths in marine mammals to record sea turtle strandings and developmental abnormalities in fish, with both deep and shallow water coral reefs and more than 2100km of coastal habitat also affected [17]. Aerosol pollutants from burning of fossil fuels also disrupt hydrological cycles. The complete phase out of fossil fuels is estimated to increase rainfall by up to 70% in regions of India, up to 30% in China and 40% over the Sahel, giving additional drought resilience and agricultural benefits for large portions of the global population [13].

In recent years there has been growing public awareness and momentum behind environmental campaigns, not least from the ‘Fridays for Future’ protests started by Greta Thunberg in 2018. Around 100,000 people attended the climate march in Glasgow during COP26 [18] and subsequent campaigning led to oil giant Shell dropping out of plans to drill in the Cambo oil field, with the project now put on hold [19].

While more and more renewable generation is being commissioned and built, this alters the power system: the variability of renewable-based power production impacts operators’ ability to manage the power system effectively [20], where many operational decisions rely on accurate forecasts [21]. As the proportion of energy generated from renewable sources increases, forecasting has become a ‘central tool’ [22] for the operation of the power system and trading of energy [23], with applications ranging from unit commitment and economic dispatch to management of grid constraints to allocation of reserves [22]. Wind energy is beginning to offer other ancillary services beyond power production, for example frequency response [24] and even black start capability [25]. However, there are still challenges and changes needed to allow operation of a grid with near 100% renewables. Four main conditions were outlined by Holttinen et al. [24]: system stability, system adequacy, revision of reserves and the balancing system (including ‘continual improvemfis aents’ to forecasting methods for renewables), and finally development of grid infrastructure. Along these lines, forecast accuracy measured through Mean Absolute Error (MAE) has been improving for all horizons over the last decade [24].

One system where forecasts are required in the UK is the Balancing Mechanism

(BM), the system operator's system for ensuring supply and demand balance and that grid constraints are respected. Wind farms over a certain capacity are required to participate in the BM and must submit a 'Physical Notification' (PN) of their power output to the Electricity System Operator (ESO) for every half-hour settlement period for the next 24 hours. The forecast value for each settlement period may be updated up to one hour before it begins. Thus the production of very short-term forecasts up to 3 hours ahead (when including time to generate and accept new values) are an obligation for these sites. Currently, generators are "required to use reasonable endeavours to ensure that the data held by the ESO ... is accurate at all times" [26]. However, there is no defined maximum error stipulation, although it is possible that poor forecasts could incur fines from the regulator OFGEM and preclude sites from providing other ancillary services in the future. Forecast performance from BM data shows a bias to overforecasting at many sites (Figure 1.1), possibly due to financial compensation incentives during curtailments: curtailed sites are paid for the difference in energy between their PN value for power and curtailed power. The ESO also produces its own in-house forecasts with incentive from OFGEM to minimse the error of these.

1.1 Research motivation, aims and objectives

There are key qualities a forecast must have in order to be adopted as a useful source of information about the future. Three main properties are:

- functional: able to produce forecasts under real-life operating conditions
- accurate: forecast skill is high enough to warrant its use, including skill at the model's worst-performing times
- relevant: the quantity that is being forecast gives useful information for the given application - generally that means information that allows subsequent decisions to be made.

This thesis identifies times or applications where forecasts within the wind industry might not meet some of these criteria, and aims to address current issues by proposing

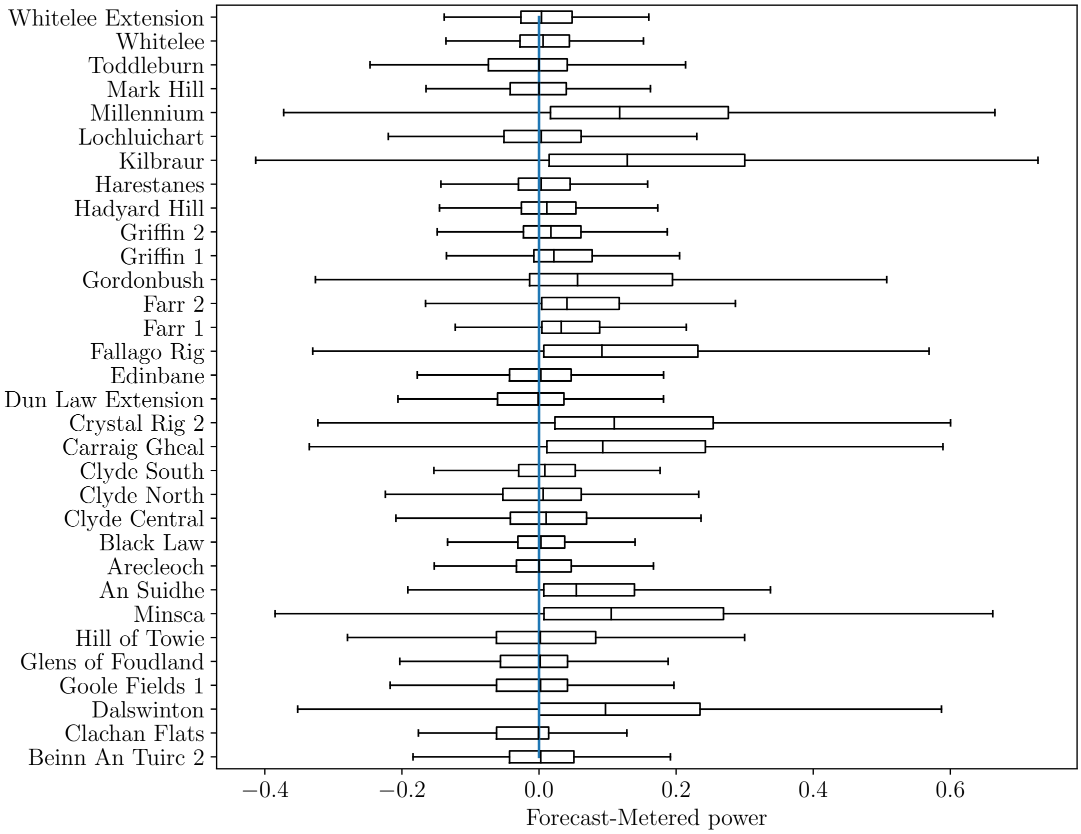


Figure 1.1: Difference between forecast power (PN) and metered power for a selection of sites participating in the UK BM. Data from Elexon (`bmreports.com`), covering the time period January 2014–July 2016. Overforecasting is much more common than underforecasting.

novel methods to improve the functionality, accuracy and relevance of forecasts. Examples of models that are not functional would include one that requires data that is not available at the point a forecast is generated, or that takes longer to run than the forecast horizon. A less extreme example would be a forecast where some of the data is missing which may result in no value, or a nonsensical value, being returned by the forecast model if not anticipated and addressed appropriately. This is very common in live forecasting, but not often studied in detail in the academic literature. There are many unavoidable data quality issues that may mean a forecast is not produced or gives an unrealistic value, so these potential sources of error in the forecast must be identified and methods put in place to avoid negative impact on the final forecast accuracy where possible. While continuous development of forecasting models has seen improvement in forecast accuracy over the last decade [24], there can still be times when forecast accuracy is lowered. This may be due to data quality issues, or at times that are particularly difficult to forecast. Times of ramps, rapid changes in power, are one such time that is difficult to forecast accurately. Furthermore, accuracy criteria are often more stringent at times of larger financial risk (or reward). For example, curtailment payments to wind farms through the UK's Balancing Mechanism are proportional to the difference between forecast power and the requested curtailment level. Therefore, underforecasting at these times would reduce the payment the wind farm receives. Large underforecasts, especially at times of high power when curtailments are more likely, are therefore much less desirable than large overforecasts. It was determined from work in Chapter 3 that these types of forecast errors often occur at times of upwards ramps, where the forecast ramp is slower or later than observed. Therefore, methods to improve the accuracy of power forecasts around times of ramps is addressed in Chapter 4. The final key forecast functionality is relevance. This is particularly applicable when generic forecasts are used, rather than a forecast tailored to the site and specific purpose. For wind energy, standard publicly available weather forecasts display ground-level wind speeds but turbines can experience quite different wind speeds at hub height, at the top of the tower nearly 100m high. Discussions with industry identified on-site maintenance work as an area where a generic publicly available weather forecast is often used. While this is an accessible forecast, it is not tailored to the exact wind farm location or the purpose. Maintenance activities often set a safety limit on wind speed and so it is the amount of time below the safety limit, rather than the exact wind speed, that is crucial for these decisions. Thus, a gap in the currently available forecasting products was identified and Chapter 5 addresses this.

1.2 Outline of approach

The first identified motivation to address is the treatment of missing data in the forecast generation process. To understand this problem, real-world data was analysed for amounts and type of missing data, including missing data patterns and the relationship between missing points and other values in the dataset. Three different areas where missing data occurs were identified and missing data patterns were replicated in a complete dataset to enable mitigation approaches to be tested and compared against forecasts made with 'perfect' data.

To address forecast accuracy at critical times of ramps in power, a combination approach was developed that takes not only individual forecasts from different models but also explicit information about ramps through a forecast ramp rate feature. This method was tested for its ability to forecast ramp events as well as its overall skill as a power forecast.

Finally, wind turbine maintenance scheduling was identified as an application currently lacking forecasts of relevant quantities. A new index related to useful hours within weather windows was proposed and forecasts of this quantity produced to allow equipment hiring decisions on these timescales. This is demonstrated through a case study for crane hire.

1.3 Overview of operational forecasting system at Natural Power

Informed by my PhD work, Natural Power have updated their operational forecasting system with the aim of providing increased accuracy in the very short-term forecasts that form part of the PN submission. As a forecast provider for a group of sites, they were able to implement a spatio-temporal model to take advantage of information from multiple locations. My work on missing data informed the input features used, strength of regularisation and methods to accommodate for missing data built into this model, taking into account the logistical constraints (server space available and other data feeds besides power). I also suggested inclusion of a lognormal transform to reduce the forecast bias at high and low powers seen in the initial model.

1.4 Thesis structure

Chapter 2 presents a literature review of very-short-term methods for wind and solar power forecasting, drawing comparisons between the two fields and ending with a case study benchmarking contrasting methods and setting out good practice in forecast evaluation. Chapter 3 analyses the properties of missing data relevant for wind power forecasting and proposes methods to mitigate this. Chapter 4 sets out a method for forecast combination tailored to improve power forecasts around times of ramps. Chapter 5 details a site- and task-specific metric for use in maintenance scheduling when operations are subject to wind speed safety limits. Finally, Chapter 6 summarises the conclusions from this work.

Chapter 2

Review of wind and solar forecasting literature

The contents of this chapter are reproduced from the published literature review, 'A review of very short-term wind and solar power forecasting' in Renewable and Sustainable Energy Reviews [3].

Abstract

Installed capacities of wind and solar power have grown rapidly over recent years, and the pool of literature on very short-term (minutes- to hours-ahead) wind and solar forecasting has grown in line with this. This chapter reviews established and emerging approaches to provide an up-to-date view of the field. Knowledge transfer between wind and solar forecasting has benefited the field and is discussed, and new opportunities are identified, particularly regarding use of remote sensing technology. Forecasting methodologies and study design are compared and recommendations for high quality, reproducible results are presented. In particular, the choice of suitable benchmarks and use of sufficiently long datasets is highlighted. A case study of three distinct approaches to probabilistic wind power forecasting is presented using an open dataset. The case study provides an example of exemplary forecast evaluation, and open source code allows for its reproduction and use in future work.

2.1 Introduction

The increasing penetration of wind and solar energy in power systems around the world necessitate new ways of operating energy systems and markets. The variability and limited predictability of the wind and solar resource introduces uncertainty for planners and operators on all time scales, from seconds and minutes ahead, to decadal variability [27, 28] and climate change [29].

Forecasting plays a central role in minimising this uncertainty on operational time scales from real-time to a few days ahead [22]. Quantifying uncertainty is also necessary for 'optimal' decision-making and risk management. Forecast uncertainty is quantified in probabilistic forecasts which most commonly take the form of prediction intervals, predictive probability density functions (univariate or multivariate), or trajectories/scenarios, though other formats exist.

It is important to distinguish between *short-term* forecasting, with lead-times of hours to days ahead, and *very short-term* forecasting, with lead-times of minutes to hours ahead. The World Meteorological Organisation defines the *very short-term* range as up to 12 hours ahead [30], but in energy forecasting the distinction is generally methodological rather than at fixed lead time although neither convention is consistently applied. The term *nowcasting* is also used to refer to very short-term forecasting in the meteorology community, but here we will use *very short-term* throughout for consistency. The main source of predictability on short-term time scales comes from Numerical Weather Prediction (NWP), whereas the main sources of predictability on very short-term time scales are recent observations. NWP is not well suited to very short-term forecasting because of the time required for data assimilation and computation, and additional uncertainty introduced by weather-to-power conversion which is greater than natural variability on very short-term time scales. For the purposes of this review, which focuses on very short-term forecasting, we are concerned with forecasting methods based on recent observations and timescales where NWP adds limited or no value.

Wind and solar forecasts of the minutes and hours ahead are required by power system operators to manage the balance of supply and demand, and electricity market participants to trade energy. For instance in Denmark, the country with the highest penetration of wind energy in the world, the Transmission System Operator's forecasts "are in five minute resolution and are updated every few minutes using all the latest information available" [31]. While countries such as Denmark are employing best practices such as leveraging real-time power production data, novel methods for producing increasingly accurate forecasts are continually being proposed, new sources of observational data are becoming available, and new ways of sharing data between parties are emerging. Current large collaborative forecasting projects include the European Smart4Res project [32] and several studies commissioned by the US Department of Energy's Solar Forecasting 2 program, such as the open-source solar forecast arbiter for forecast evaluation and benchmarking [33]. This article reviews these advances beyond current state-of-the-art operational forecasting systems and discusses their relative merits and potential evolution.

The expansion of wind and solar energy and research necessitates regular reviews and synthesis of advances, yet despite sharing many common features, wind and solar forecasting are often reviewed in isolation, perhaps a result of the relatively later development of solar power forecasting compared to wind [34]. Both wind speed and solar irradiance exhibit spatio-temporal correlation as a result of their dependence on large-scale meteorological phenomena. As such, some methods are effective for both wind and solar applications, such as time-series methods supplemented with exogenous inputs or multi-variate extensions which capture spatial correlations between multiple sites. In the recent history of very short-term wind and solar power forecasting one field has learned from the other. In this paper we identify potential opportunities for further advances in the same vein.

Both solar power forecasting [35–40] and wind power forecasting [41–45] have been reviewed recently individually. However, very short-term horizons have received little attention in these reviews and neither have advances in very short-term wind and solar forecasting been compared. There are only two exceptions we are aware of: Sweeney et al. [46] who consider wind, solar and hydro power together and discuss very short-term lead-times briefly, but do not systematically review the field and instead provide a vision for renewable energy forecasting in the future; and Barbieri et al. [47] whose primary focus is very short-term solar but who also briefly mention transferable approaches from wind literature.

Previous reviews provide detailed analysis of various modelling approaches; for solar forecasting, Antonanzas et al. [35] examine several approaches to persistence models for solar forecast benchmarks, Inman et al. [37] cover clear sky models in depth, and Ahmed et al. [39] include particular detail on deep learning models and sky imaging. There have been several reviews of combined, or hybrid, models [41,44,48] while Giebel et al. [42] provide an overview of the history of very short-term forecasting as well as models using NWP inputs. Jensen et al. [49] detail a wide range of solar evaluation metrics, including for event-based forecasts, e.g. forecast performance for ramps. Foley et al. [43] give average values for error metrics for different forecast horizons. Current state of the art and future directions suggested include greater prevalence of probabilistic forecasting [35, 50], increased focus on the economic impact of forecasts on decision making [35, 49], weather classification or regime-based approaches [39, 46], and use of high resolution — including turbine level — data and data marketplaces [46]. Yang et al. [40] use a text mining approach to map forecasting and model terminology, before also highlighting six key recent works. Inman et al. [37] identify the forecasting of ramp events as a particular challenge for renewable energy integration in general. Key recommendations from these reviews include the need for a general database of geographically dispersed sites to test models on [51], consistent benchmarking approaches across research papers [36] and common evaluation metrics [35]. Lauret et al. [52] recommend the Continuous Ranked Probability Score (CRPS) score for probabilistic forecast evaluation.

This review proceeds with a description of the systematic literature search that has been performed and a high-level bibliometric analysis (Section 2.2), after which very short-term solar (Section 2.3) and wind (Section 2.4) power forecasting are reviewed before a summary of common research methods and comparisons between the wind and solar literature are drawn in Section 2.5. While this review is by no means exhaustive, it is intended to give an overview of the variety of approaches that have been proposed in recent years. A case study based on an open dataset is presented in Section 2.6 in order to reproduce and compare three distinct classes of statistical model commonly employed for wind and solar forecasting but that are seldom compared to one another. The findings of this review and advances in very short-term wind and solar power forecasting are discussed in Section 2.7 which also speculates as to the direction of future research in this area.

2.2 Summary of papers reviewed

We used Web of Science to conduct a literature search[1], up to and including the end of 2020, for publications on short-term and very-short-term wind and solar forecasting. The number of works in this area has clearly been increasing substantially throughout the last decade, in line with the increases seen in both wind and solar generation globally (Fig 2.1). This suggests the importance of forecasting these variable generation technologies increases as their penetration on the grid increases [42].

For short-term methods where lagged on-site measurements are the predominant data input, models often fall into two broad types: traditional time series regression, and Machine Learning approaches. Of the papers examined in this review, we found 24% included some type of regression or time series model, and 62% included a Machine Learning (ML) model. A list of all papers included is given in Table A.1 of Appendix A. Figure 2.2 shows a general summary of forecasting approaches across the literature.

The subsequent sections 2.3 and 2.4 cover the top 50 most cited results stratified by the number of publications in each year and selected by the Web of Science search[1]. This selection has been limited to publications in 2014 or later, as the aim of this work is to focus on recent trends and developments in wind and solar forecasting. The literature from this search has also been supplemented with other references and works already known to the authors.

[1] The search query used was ((TS=(((”wind speed” OR ”wind power” OR ”solar” OR ”renewable generation”) NEAR/5 (”forecast*” OR ”predict*”)) AND (”short term” OR ”short-term” OR ”very-short-term”) NOT(”hydro” OR ”thermal”)))) AND LANGUAGE: (English)

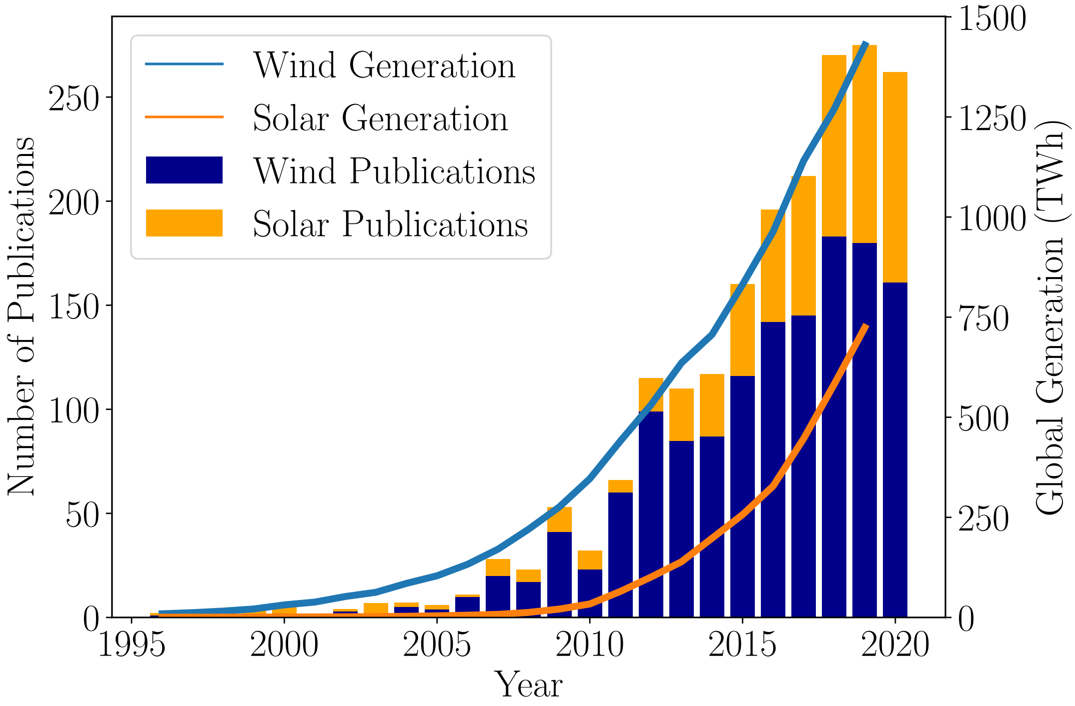


Figure 2.1: Forecasting publications broken down by wind and solar as a stacked bar chart, also plotted with global energy generation through time. Generation data provided under CC BY 4.0, Hannah Ritchie & Max Roser, `ourworldindata.org/renewable-energy`.

2.3 Solar power forecasting

Solar projects tend to have smaller installed capacities relative to wind projects: in the UK as of May 2020, the average solar installation has a capacity of 1.29 kW, with only 1.8% of these exceeding 4kW [53]. Of the larger UK projects requiring planning applications, the average installed capacity across 1171 projects was 7.2 MW, compared to an average of 29.6 MW across 778 wind projects [54]. As such, solar generation tends to consist of a greater number of smaller projects than wind. Sweeney et al. [46] note that decentralised small scale energy sources often contribute to localised grid congestion problems, increasing the importance of accurate forecasts for grid management. Very small solar systems such as household installations are often ‘behind-the-meter’, with no power production data available to forecasters and as such are often instead incorporated in ‘net demand’ (rather than power production) forecasts [35].

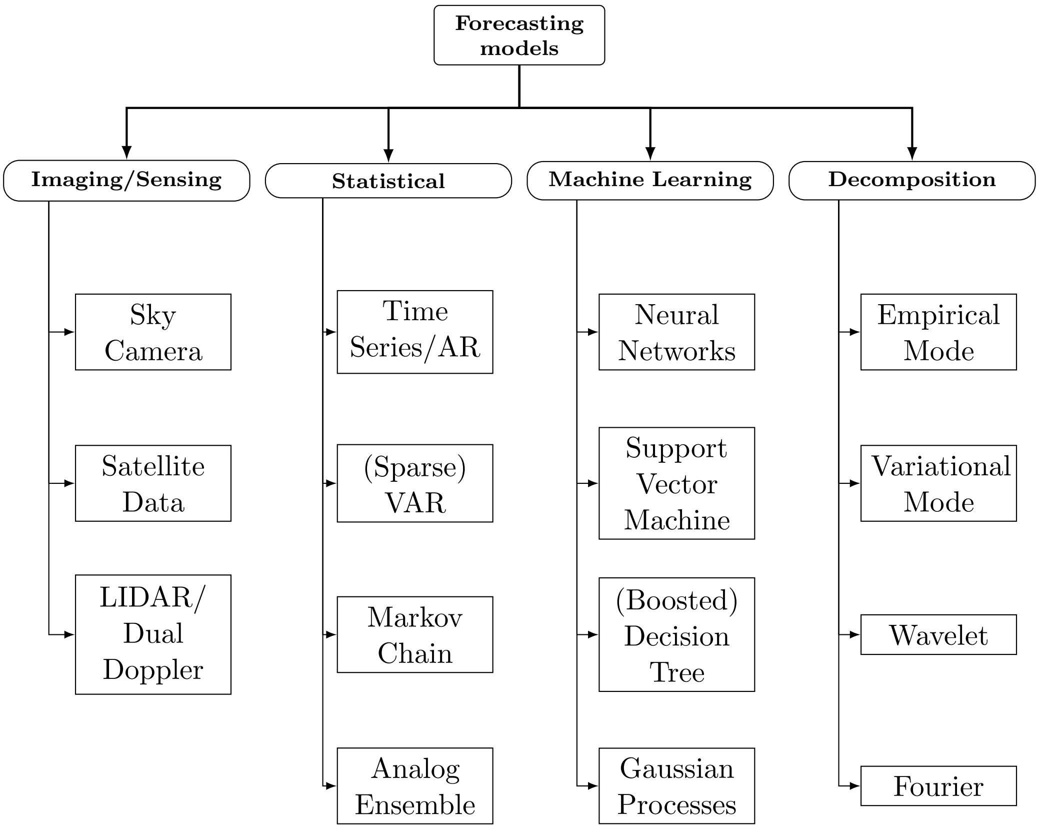


Figure 2.2: Diagram of forecasting model techniques. Neural Networks include ELM, RNN, CNN, LSTM etc; Decision tree methods include Random Forest and Gradient Boosted Trees. Methods may also be implemented in an adaptive or online framework, or include regime switching. They may also be used for probabilistic as well as deterministic forecasts.

Solar power production follows strong seasonal and diurnal patterns due to the changing path of the sun, which defines the maximal possible irradiation for a given location, time and date. This is known as 'clear sky' irradiation, which can be well defined by various models [37]. In addition, the passage of clouds create shadows that introduce stochastic variability in the power time series and is much more challenging to predict [35]. Atmospheric aerosols may also reduce surface irradiation and therefore power output. This may be caused by natural phenomenon such as salt from sea spray, dust storms and soot from wildfires, or man-made pollution. A case study in West Africa found a reduction in power in the range 13-37% due to dust aerosols [55]. The physical condition of the panels can also affect production. For example, accumulation of dirt and dust have been shown to reduce energy production by 2-6%, and snow cover can also reduce power output completely if thick enough [56].

2.3.1 Image-based methods

Imaging techniques may be applied to either ground-based systems or satellite images to determine and predict future cloud cover, used in turn to forecast solar irradiance or solar power directly. Ground based sky imaging has mainly been used for high temporal resolution forecasts up to 30 minutes ahead. For a cloud at an altitude of 2km and a speed of 10 ms^{-1}, this represents a field of view of 154°. The focus of this method on very short time horizons is two fold: field of view and cloud formation and dissipation limit the skill of this method out to longer horizons [57], while it also fills a gap that several other data sources don't currently have the spatial or temporal resolution to match [58] (satellite images generally have a 15 minute or slower update time for example). Methods using propagation of current observed cloud conditions are common such as cloud motion displacement [59] or determination and propagation of shadow position using cloud base height measurements in conjunction with images [60]. In this work clouds were also classified by type, although persistence still outperformed this method at a horizon of 25 minutes. Pitfalls of sky imaging systems may include errors due to perspective, image saturation in pixels close to the sun and soiling of the cameras [40]. There is also additional expense associated with maintaining a camera system on site.

Lago et al. [61] train an irradiance model using satellite and weather forecast data as inputs and ground measurements of solar irradiance at a group of sites in the Netherlands as the target variable. The learned model may then be used more generally at other sites without the need for ground measurements, and in fact this generalised model also outperformed models trained with local ground data. This approach is perhaps more suited to forecasting a group of sites rather than a single location, as a small subset of sites that do have ground-based measurements is also needed for model training. It would be interesting to test the generalisation of this approach to other climate regimes and more geographically dispersed sites. Harty et al. [62] also use both satellite and NWP data. However, they take a slightly different approach, producing cloud motion vector fields from both information sources and combing these via ensemble Kalman filter. Their method improves upon using a single information source for intra-hourly forecasts for a city region. Bellinguer et al. [63] modelled spatio-temporal dependencies, with different models fitted conditional on NWP geopotential height. A combination of satellite data, where the 10 most informative pixels are chosen via mutual information, and on-site power measurements are used as inputs. Carriere et al. [64] note that different information sources tend to be most beneficial at different forecast lead times, so proposed an approach including several information sources. Irradiance time series from satellite data, NWP forecasts and lagged on-site power and temperature are supplied to the model, leading to good performance across a range of horizons up to 36 hours ahead. Non-parametric probabilistic forecasts were produced through an analog ensemble, using sets of similar past observations.

2.3.2 Probabilistic methods

Probabilistic methods allow quantification of uncertainty in the forecast and can facilitate proper risk analysis in applications. However, only a portion of the solar forecasting literature considers probabilistic forecasts and within this there is still sometimes a focus on general prediction intervals rather than full predictive densities.

Prediction interval approaches include a method using the variability of a time series about its mean [65]. Alternatively an ‘uncertainty metric’ may be determined from ensemble forecasts for points in a reference dataset, which is then used to look up the expected error (then used as a prediction interval) using a nearest neighbours approach [66].

Full density forecasts may be parametric, where the predictive distribution is specified by a small number of parameters (e.g. the mean and variance of a Gaussian distribution), or non-parametric with no assumed distributional shape. Golestaneh et al. [67] find that solar forecast error distributions are not easily fitted by any common parametric distribution, so propose non-parametric quantile forecasts using lagged power alongside meteorological measurements in an improved Extreme Learning Machine (ELM) model. This was demonstrated on a high resolution (1 minute) dataset which may not always be available. Gaussian process regression has also been proposed with an extension to give less weight to the observations that were more likely to be outliers [68].

2.3.3 Machine learning

Various machine learning techniques have been proposed for solar forecasting as they can allow for nonlinear relationships [69, 70] and learn from data without the need to make assumptions about the relationships between variables. For the very short term (up to one hour ahead), Rana et al. [71] showed that on-site power measurements can provide skilful forecasts and NWP inputs (solar irradiance, temperature, humidity and wind speed) don't further improve forecast skill. They used an ensemble of Neural Networks, which outperformed a Support Vector Regression (SVR) model. In other work, Sivaneasan et al. [72] found that feature engineering of a 'cloud cover index' from humidity and rainfall measurements and use of previous forecast errors as Neural Network (NN) inputs showed improved performance compared to a NN trained without these. Long Short-term Memory (LSTM) networks are a common choice for time series problems; Lee et al. [66] demonstrate their use with the dropout technique to produce ensemble forecasts. Alternative techniques to generate an uncertainty interval were also compared in this work. ELM models may overcome problems of overfitting and local minima associated with NN approaches. To reduce computational complexity, Majumder et al. [69] used a low rank kernel ELM along with variational mode decomposition to address the nonstationarity of solar time series. This model was tested across a range of horizons (15 minutes to 1 day ahead). In other work using a cost function based on generalised correntropy for the ELM improved performance, possibly due to increased robustness to outliers [73]. Tang et al. [74] also used an ELM to forecast solar power, in combination with pre-processing of inputs using an entropy method. The probabilistic approach of Golestaneh et al. [67] is also based on ELM and performs favourably in comparison to both persistence and climatology as well as other ELM variants.

Abuella et al. [75] used an ensemble of SVR models to generate day-ahead forecasts from NWP data; the 24 different forecasts are then combined to give the final probabilistic forecast via a Random Forest (RF). This approach shows improvement over individual models and could be appropriate for combining shorter-term forecasts. Eseye et al. [76] also used NWP variables as inputs to an Support Vector Machine (SVM) model, additionally applying wavelet decomposition. However, the number of decomposed series were chosen based on previous literature rather than optimised on the given data.

For models with multiple data processing steps as well as model fitting, it may be advantageous to optimise all hyper-parameters for all parts of the model process simultaneously: Li et al. [77] found a 53% improvement just by using simultaneous optimisation.

Spatio-temporal relationships have been considered by including irradiance measurements from nearby sites as forecast inputs [70]. Not only is the proposed model shown to outperform Autoregressive (AR) methods, but boosted regression trees outperform both NN and SVR models. These models were developed only on times where clear sky irradiance exceeded a threshold, limiting their applicability to forecasts for dawn and dusk times.

2.3.4 Other methods

There has also been focus on utilising spatio-temporal dependencies between sites for solar forecasting. Agoua et al. [78] propose a Vector Autoregressive (VAR) model normalising the input power time series by simulated power to make the time series stationary. They find Least Absolute Shrinkage and Selection Operator (LASSO) is the most effective variable selection procedure, and that conditioning on surface wind speed also adds skill to the forecasts.

The Sun4Cast system developed in the USA utilises several data sources and diverse models before producing a final forecast through a weighted combination [79]. The very short-term models include a sky imaging system, regression tree on pyranometer data, satellite imaging with advection and an NWP model tailored to solar forecasting with a high refresh rate [80]. They found benefits from each model for different lead times or climate scenarios, giving an effective combined model.

Several of the studies mentioned in previous sections may also be classified as hybrid methods. The term ‘hybrid forecast’ is often used to refer to methods where more than one forecasting method is combined into a final forecast: this may be simply through combining forecasts from different models [66, 75], applying some form of decomposition to the original time series and fitting different models to each resulting series [76], or where multiple different input data sources are processed separately before being combined [62, 80], for example satellite data and irradiation measurements. Hybrid methods often outperform a single model method, especially where a diverse set of individual models are combined. A full recent review of hybrid models for solar forecasting is given by Guermoui et al. [48].

2.4 Wind power forecasting

The very nature of the wind presents forecasting challenges: the state of the atmosphere can never be fully known, meaning wind speed is treated as a stochastic process affected by many factors, from large scale weather systems down to local terrain. Of course the variable of interest in forecasting is often not wind speed but power. The relationship between wind speed and power is dynamic and nonlinear [20] which adds complexity and makes forecast power particularly sensitive to wind speed in between cut-in and rated wind speed. Wind power forecast errors are typically heteroscedastic and auto-correlated. Furthermore, production is bounded between zero and the rated capacity of a turbine or farm. These properties violate common assumptions in statistical modeling, such as independent and identically normally distributed errors, and should receive careful treatment in sophisticated forecasting methods.

Wake effects can influence the power output of turbines in the ‘shadow’ of others and this is highly related to wind direction. A power drop of around 30% of capacity was seen between the first and second row at Horns Rev when the wind direction is such that a turbine is directly behind another [81]. In cold climates, icing can reduce power output by as much as 40% [82]. Losses in operating efficiency over time could also affect forecast accuracy: turbine aging is estimated to cause a typical decrease in output of 0.2% per year in the first 5 years [83], although this also includes losses due to increased downtime. Data feed quality also affects the performance of models where site measurements are used as model inputs [4].

2.4.1 Regression-based methods

Past on-site measurements are widely collected by wind farm owners and are often valuable inputs when forecasting a few hours ahead. Simple time series methods based on Autoregressive moving average (ARMA) models are well established [42] and still form the basis of ongoing research. Zhou et al. [84] showed that a dynamic combination of an Autoregressive integrated moving average (ARIMA) model with recent measurements as inputs and an AR model with inputs from NWP models is an improvement over either individual model. Other approaches using Autoregressive models in conjunction with other models are detailed in Section 2.4.4 on hybrid methods.

VAR models have been proposed to capitalise on spatial dependencies between geographically dispersed sites; since the number of model coefficients grows with the square of the number of sites, sparse models have been employed to reduce computational time and model complexity while improving forecast performance. For the case study presented by Cavalcante et al. [85], a standard VAR model with no regularisation is shown to give improvement of around 5.9% over an AR model for a 2-hour ahead forecast, while introducing sparsity through LASSO regularisation gives a further 1% improvement. Grouping the LASSO penalty by whether an input is a lag of the predictor or not (i.e. diagonal vs off-diagonal elements) seemed to give the best results. An adaptive LASSO estimation algorithm is proposed by Messner et al. [86] to track potential changes in the VAR coefficients in an online fashion, yielding improvements relative to the equivalent static model for 15-minute resolution data and lead-times greater than 30-minutes.

Dowell et al. [87] developed probabilistic forecasts based on the logit-normal distribution in a VAR framework for 5-minute ahead wind power forecasting; training on a window of most recent data allowed for changes in the sparsity through time. In a deterministic setting without the logit-normal transformation, and based on hourly mean powers, this method was outperformed by the LASSO-VAR approach. Correlation between farms has also been used to determine the sparsity of a VAR model [88], where the overall sparsity and number of non-zero coefficients for each farm can also be controlled. This was shown to outperform a standard LASSO-VAR model, but not compared against the sparsity structured LASSO in [85].

Capturing changes in VAR coefficients over time has been considered in adaptive frameworks where changes are tracked in an online setting [85, 87]. These adaptive methods improve over static equivalents, but inherently track changes with some lag and smoothing. Explicitly conditioning VAR coefficients on large-scale weather patterns was found to improve wind speed predictions from 1–6 hours ahead [89] but has not been applied to wind power.

For sites that wish to benefit from the improvements of spatio-temporal forecasting without revealing potentially commercially sensitive information, privacy preserving approaches have been developed. These may be grouped into three broad categories, each with their own disadvantages [90]: data transformation that may lead to a trade-off between privacy and model accuracy; multi-party computation [91] which may require a central coordinator and where similarity between model inputs and targets may lead to a breach in data confidentiality, or where using encryption techniques significantly increases computation time; and decomposition into parallel sub-problems which require iterative solutions - and each iteration progressively reveals more information to the participating data owners.

2.4.2 Machine learning

As with solar forecasting, various machine learning techniques have been applied to very short-term wind power forecasting. A comparison of SVR, decision trees and Random Forest models found Random Forest to give the lowest mean absolute percentage error [92] although no feature engineering was explored, which has been shown to play a significant role in good model performance in other works [93]. Correction of the output of an SVM using a Markov Chain showed improvement over a basic SVM approach [94]. The probabilistic output of a Markov Chain appears to be discarded in favour of a point forecast with 'fluctuation intervals'. A combination of two kernels (wavelet and polynomial) in an SVM model improved wind speed forecasts relative to the use of just a wavelet kernel [95]. The recent trend in wind speed (increasing, decreasing or stable) was also used to train separate models for these regimes, giving a slight improvement over a single model for all conditions.

Wang et al. [96] applied a multi-objective approach to NNs, having separate objective functions for bias and variance. Similar multi-objective approaches have also been used on decomposed time series and are detailed in the section on decomposition methods [97,98]. Khodayar et al. [99] used autoencoders for unsupervised feature learning and 'rough' neurons to better process noisy data, showing superior performance to other NN models. To its credit, forecast evaluation is based on a full year of out-of-sample data using the open source Western Wind dataset [100]. Neural Networks were also used by Rodríguez et al. [101] for 10-minute-ahead microgrid control.

Graph Neural Networks were used along with an LSTM for feature extraction to identify and utilise spatio-temporal relationships between sites by Khodayar et al. [102], giving improvement over both persistence and other ML benchmarks. Inclusion of other metrics such as maximum observed error and correlation matrix of forecasts as well as usual average error metrics enhanced the analysis in this work, and the use of an open dataset is also a good step towards replication and comparison of research methods.

Hossain et al. [103] also used convolutional NNs and Gated Recurrent Unit (GRU) layers for feature selection and processing of multiple input data sources respectively. They found improvement over other ML approaches at two case study sites.

De-noising of wind speed time series using Singular Spectrum Analysis (SSA) along with a fuzzy Neural Network model outperformed ARIMA and other NN implementations for a group of sites in China [104]. A novel neighbourhood LSTM network was proposed by Zhang et al. [105] and claims to take causality, rather than just correlation, between variables into account, outperforming other ML methods in the study. Chen et al. [106] compared artificial intelligence methods with Autoregressive models, finding that both an artifical Neural Network and an Adaptive Neuro-Fuzzy Inference System (ANFIS) marginally outperform an ARMA model for 10 minutes ahead forecasts, but that the ARMA model has superior performance for hour-ahead forecasts.

Many of the hybrid and decomposition approaches detailed in the following sections also make use of machine learning models.

2.4.3 Decomposition methods

Decomposition methods are based on the premise that the wind speed or power time series contain different frequency signals with different characteristics, and that modelling each of the decomposed series separately can lead to overall improvement in forecast skill [107–109]. Empirical Mode Decomposition (EMD) is based purely on the data and splits the original time series into several Intrinsic Mode Functions (IMFs), which can each have time varying frequency. As such, this method is applicable to nonlinear and non-stationary data [110].

Ensemble EMD, adding a noise term to the original signal before the decomposition, may be used to minimise mode mixing between the IMFs. Using ensemble EMD, Zhang et al. [109] applied an ANFIS model to those IMFs classed as 'nonlinear' and a seasonal ARIMA model to those classed as 'periodic'. However, the judgement of which model to apply seems to have been made manually which may not be appropriate for real-world applications. Similarly, IMFs may be classed as high or low frequency signals, with different models applied to each; Liu et al. [107], used an LSTM network for low frequency signals to capture longer-term trends, with an Elman NN for higher frequency IMFs. Similarly, a combination of ARIMA and NNs has been demonstrated to fit probabilistic forecasts to decomposed series [111]. An alternative approach fitted multiple different NNs to each IMF, with the final forecast for each IMF being a weighted combination of these [112].

To reduce the number of models estimated, Lu et al. [108] used permutation entropy to group similar IMFs. An SVM was then used to forecast each series, outperforming both methods with no decomposition and those with decomposition but not using the permutation entropy approach. Decomposition has been combined with multi-objective optimisation for both accuracy and stability. This has been implemented with both Elman [97] and wavelet Neural Networks [98]. In both works the proposed methods outperformed single objective models.

Wavelet decomposition also results in the decomposition of a time series into multiple signals with different typical frequencies; it was found that further decomposing the highest frequency of these series improved forecasts [113]. Variational Mode Decomposition (VMD) is another decomposition technique, where each mode has a limited bandwidth. Zhang et al. [114] found this outperforms EMD for the sites analysed.

2.4.4 Hybrid (combination) models

Hybrid models are based on the premise that a combination of several forecasts from different models, or where models use different information sets as inputs, commonly outperform a single model [115]. This does rest on the assumption that no model is the true representation of the underlying data generating process, as this single model, if known, would outperform any combination of 'misspecified' forecasts [116]. However, in many 'real-life' applications, either the true process is not known or no individual forecaster or model has access to the complete information needed to generate the 'perfect' model. This is certainly true of wind power forecasts, where the final value of power output is the result of complex physical interactions to produce the wind speed seen by the turbine, as well as the performance of the individual turbine and any imposed control actions.

The simplest method of forecast combination is a linear weighting approach where forecasts are combined as a simple weighted sum, often with the restriction of non-negative weights that sum to one. This approach was used for the combination of an SVM and radial basis function NN, where weights were found via forecast correlation with the actual time series for four different wind speed regimes for each month [117]. While specifying the model weights according to a correlation measure eliminates the need for estimation of the weights as free parameters, it may not guarantee the optimal combination.

Xiao et al. [118] used linear weights to combine five different models, with the weights optimised both by minimising forecast errors ('traditional' approach) and using a particle swarm optimisation. Including all five individual models in the final combination consistently gave best results as opposed to dropping some model(s) completely, with the particle swarm optimised weights outperforming the traditional approach for this case. Zhou et al. [84] found improvement using a small sliding window of previous forecast errors to adaptively combine forecasts, although only linear ARIMA type models were considered.

Nonlinear combination of an ensemble of neural network forecasts was achieved by a genetic programming algorithm [119]. Both lagged power measurements and NWP variables were used as inputs for one hour ahead forecasts, with feature selection to find the subset of 'informative' inputs although the results of this were not reported. Ouyang et al. [120] takes a slightly different approach, determining significant input variables by Granger causality and building a separate univariate model for each of these. A multilayer perceptron was found to be best for combining the univariate predictions in the second stage of the model, and outperformed multivariate models. Lin et al. [121] proposed a probabilistic forecast combination method, also using a weight coefficient for each model and combining both parametric and nonparametric forecast distributions. It is based on open data from GEFcom2014 [122]. Deterministic forecasts from a range of ML models have also been used as inputs for probabilistic combined forecasts [123]. While none of the individual models showed improvement over persistence for 1 hour ahead forecasts, the final combined model gave a significant (30%) improvement and beat persistence at all sites tested.

2.4.5 Probabilistic methods

A quantile loss function with an LSTM network was used to generate interval forecasts [124]. Attention mechanisms for automatic weighting of input features and extracting trends through time appear to improve the sharpness of the forecasts.

Jiang et al. [125] used separate objective functions to maximise the interval coverage and minimise the interval width of a forecast power interval independently. This allows the user to choose from a set of pareto-optimal solutions according to their preferred trade-off between coverage and interval width. A deep learning approach using a convolutional Neural Network was found to outperform persistence and other shallow networks across seasons, quantile levels and for a wide range of forecast horizons [126].

A Markov Chain (MC) approach where transition probabilities between discrete power levels are modelled gives probabilistic forecasts without assuming a distributional shape [127]. A large number of power levels may lead to transition probabilities of zero in this method simply because they are not observed in the training data; a Bayesian approach where prior transition probabilities can be specified would mitigate this.

The Weibull distribution is commonly used to model wind speed distributions; Bracale et al. [128] propose a mixture of two Weibull distributions to allow for bimodal distributions, fitting the mean with an ARIMA model and the remaining parameters through Bayesian inference. This approach outperformed both persistence and single distribution models for hour ahead forecasts.

The point forecast accuracy of an LSTM and the good probabilistic reliability of a Gaussian Process regression model were combined and found to outperform other time series methods both on point forecast accuracy and probabilistic performance [129].

2.4.6 Turbine-level data and remote sensing

Wind farms comprise multiple, sometimes hundreds, of individual wind turbines, forming a hierarchy which may be exploited to improve forecast performance [130]. Furthermore, if spread over a sufficiently large area, up-wind turbines may detect changes in wind speed early enough to inform very short-term forecasts for the farm as a whole. Similarly, measuring the wind speed up-wind of the wind farm using remote sensing may provide valuable information for very short-term forecasts.

Jiang et al. [131] proposed use of time series from a neighbouring turbine and selection of forecast inputs via grey correlation analysis to improve individual turbine's forecasts. Along with an SVM model and cuckoo search for parameter optimisation, this does appear to improve forecasts relative to persistence, ARIMA and other SVM models. This model doesn't take account of the changing relationships between turbines as wind direction changes, for which a dynamic model may be more suitable.

A spatio-temporal Gaussian Process has also been proposed to predict turbine- and farm-level power production for 1- to 12-hours ahead [132], improving over non-spatial approaches to a comparable degree as spatio-temporal models on multiple wind farms. To its credit, this study is based on an open dataset [133]. One month of training data is used to train models on on a 6-hour rolling basis, which may impact some methods more than others.

Turbine-level forecasting using inputs from similar turbines (found through clustering algorithms) and an LSTM network showed promise over other ML benchmarks for 90-minutes ahead forecasts [134], although there was no discussion of how this translates to farm-level forecasts or consideration of hierarchical approaches for this.

Both lidar and radar technologies have been deployed at wind farms to measure the wind resource, though forecasting has not been the primary motivation. Wurth et al. [135] review minute-scale forecasting, with scanning lidar and radar identified as promising technologies; while use cases exist they are underdeveloped. Valledcabres et al. [136] use dual doppler radar observations of up-wind wind speed to improve 5-minute ahead predictions of 1-minute mean power. Scanning lidar have also shown potential to improve forecasts for minutes-ahead horizons [137] but suffer from data reliability issues in fog or rainy conditions.

2.5 Research methods

While very short-term power forecasting is an evolving area, certain methods are applied more commonly for different lead times. This is partly due to physical restrictions (for example cloud formation and dissipation and changes in wind direction limit the predictability of image based methods to long horizons) but also due to practical limitations, such as the latency in data assimilation, low temporal resolution, and low refresh rate typical of NWP models. However, higher spatial and temporal resolution NWP products are becoming available with hourly re-fresh rates, as provided by NOAA's High Resolution Rapid Refresh [138] and the UK Met Office's UKV [139] and MOGREPS systems [140], for example. Higher resolution and refresh rates are offered by emerging technologies such as Whiffle's so called 'finecasting' approach and NOAA's experimental 'Warn-on-Forecast' product [141]. These advances bring the ability of NWP to model and predict physical processes to ever shorter lead-times where they have not traditionally out-performed statistical methods based on local observations. The two approaches are complimentary and state-of-the-art, site-specific forecasting systems combine both NWP and statistical processing of live site data.

Imaging techniques for minutes-ahead applications have had greater attention in the solar literature, whilst models including spatio-temporal relationships have focused more on wind power forecasting. Methods producing a forecast as a probability distribution are becoming more widespread, although there is more focus on probabilistic forecasting in the wind community. Solar forecasts sometimes only give a single confidence interval which might not have a formal definition in terms of probability coverage [65, 66]. Probabilistic forecasts are not always evaluated using probabilistic metrics, or only for one interval rather than the whole distribution [68, 72, 124].

Confidence in the significance of results may be undermined by use of limited case study datasets with a length of days to weeks rather than a year or more [142]. In particular, results of model evaluation carried out entirely on data from one season at one site may not generalise to other seasons, weather conditions or other locations. The shorter the dataset, the smaller the probability the data contains a wide range of weather (cloud or wind) conditions; this increases the risk of poor performance when forecasting for conditions not included in the training set. A long dataset covering multiple sites would be expected to allow more robust conclusions on model performance to be drawn. Use of small datasets is seen in both solar [72, 74, 77] and wind [94, 95, 104, 107, 109, 111, 118, 129, 131] studies.

Papers on novel methods do not always include appropriate benchmarks such as naive models or established best-in-class methods; we found both solar and wind papers which only compare models to their own variations [65,66,69,74,84,94–96,101,125,127]: this is in line with a survey by Doubleday et al. [36], who find that 8 of 42 solar forecasting papers surveyed did not include a benchmark other than variants of the same model. They recommend comparison to two benchmarks, one highly reliable but more naive approach and one closer to state-of-the-art. Testing against benchmarks that are significantly different from the proposed model would allow for a clear comparison with other methods. Consistent benchmarks across papers and publishing code alongside for reproducibility would not only strengthen confidence in reported results, but allow easy comparison of state-of-the-art approaches.

We have found it is sometimes unclear how, or if, data has been partitioned to perform out-of-sample evaluation [92, 97, 98, 105, 112, 113, 117]. A brief clear description of the training and testing sets or cross validation approach used would be beneficial in these cases.

For wind power forecasting, a proportion of work is based on wind speed, rather than power, forecasts [89, 99, 111]: while wind speed forecasts may well be more appropriate for some applications, it is worth noting that grid or trading decisions require power forecasts. The conversion from wind speed to wind power is complex and nonlinear in itself and so models reporting skill in forecasting wind speeds are not guaranteed to provide the same level of skill if used to forecast power instead. Likewise, papers based on wind speed datasets where a power variable has been simulated by passing values through a power curve will not be representative of the noisy power data seen operationally [101, 106]. Open source wind power datasets [34, 122] now allow testing of models on power (rather than wind speed) data when that best fits with the aim of the study. For solar power, solar irradiation forecasts are analogous to wind speed forecasts in that they forecast a proxy for power, but not power itself. Although the relationship between irradiance and power output for solar is less complex than the wind speed-power relationship, results of studies based on irradiance forecasts are not guaranteed to generalise to power forecasts.

One of the strengths of Machine Learning based research papers seems to be the general prevalence of data preprocessing including data cleaning and variable transformations such as principal component analysis [102, 104]. Preprocessing is also seen in ML approaches to solar forecasting [72, 74].

2.6 Case study

The following case study is provided, along with underlying data and code, to serve as an example of good practice, and to highlight features of different approaches to very short-term forecasting. Three distinct methods found in the preceding literature review are implemented and evaluated using data from the wind track of GEFcom2014 [122].

The Vector Autoregressive (VAR) method was chosen to demonstrate modelling of spatio-temporal dependencies for a set of sites, while the Markov Chain (MC) method is a computationally simple nonparametric approach where no distributional shape is assumed. Finally, the decomposition (Empirical Mode Decomposition (EMD)) approach was chosen as a contrasting method and is often only benchmarked against other decomposition approaches. The GEFcom2014 data is publicly available and comprises two years (2012 and 2013) of hourly resolution data, including both wind power measurements and Numerical Weather Predictions (NWPs) of wind speed at 10 and 100m for ten wind farms in Australia. However, here we only use the wind power and lagged values thereof for very short-term methods. The data is separated into the same training and testing sets for each method to allow fair comparison of the forecast errors. The first third of the data (up to 2012-08-31) is used for model training and hyperparameter optimisation while the remaining 16 months are used for forecast evaluation. Forecasts are generated for 1 to 6 hours ahead.

All data and code for this case study is available at `doi.org/10.5281/zenodo.5070758` for reference or to use as benchmarks for other research. All methods implemented in this case study produce probabilistic forecasts to allow full description of the forecast distributions. For the Vector Autoregressive (VAR) and Empirical Mode Decomposition (EMD) approaches, parametric probabilistic forecasts were produced by log-transforming the data and fitting a Gaussian distribution to the transformed values. The log transform is defined as

$$y = \ln\left(\frac{x}{1-x}\right) \quad , \quad 0 < x < 1. \tag{2.1}$$

Power values x in the range $[0, \epsilon]$ are rounded to $x = \epsilon$ (and likewise for high powers

$[1-\epsilon, 1]$ rounded to $x = 1-\epsilon$) to prevent infinities in the transformed data. This is effectively the same as putting all the probability mass from these boundary regions at ϵ and $1-\epsilon$. The mean of the distribution of the transformed values is modelled with VAR or EMD, and the variance of the distribution is also modelled (in transformed space) to fully specify the Gaussian predictive distribution. Both a constant value for the variance, and a simple exponential smoothing model, were tested and the approach that minimised the pinball score chosen at each forecast horizon. Once the mean and variance are specified, quantile forecasts may be generated from the forecast Gaussian distribution, before applying the inverse transform

$$x = (1 + e^{-y})^{-1} \tag{2.2}$$

to produce quantile forecasts for the wind power.

2.6.1 Vector Auto-Regression

A simple vector autoregressive model with Least Absolute Shrinkage and Selection Operator (LASSO) has been implemented, using the log normal transformation in the same way as set out by Dowell et al. [87]; after transformation, the data is modelled as normally distributed and forecasts are defined by the mean and variance of this distribution. In the transformed space, the mean of the distribution is modelled by a vector autoregressive process whereby the previous m lags from all sites are included as inputs and the regularisation parameter λ allows control over the strength of regularisation (and therefore the number of non-zero input coefficients) in the model. λ and m were optimised through cross validation on the training set, before the final model was fitted on the training data and used to forecast for all of the test set. The value of ϵ was set to 0.01. The variance of the forecast distribution was modelled as constant, and found using the variance of the residuals for the training folds (found via cross validation). An exponential smoothing model for the variance was also tested but did not provide any improvement over the constant variance model.

2.6.2 Empirical Mode Decomposition

Empirical Mode Decomposition allows a signal to be separated into different sub-series (called Intrinsic Mode Functions (IMFs)), each with different characteristic frequencies. Because it is an empirical approach based on the data, it allows for time-varying frequencies in the decomposed series and, by definition, the individual IMFs (plus the final residual) sum to the original signal for all time points. This allows different models to be fit on the different series, with the aim of improving the overall model fit. For example, one model may be better suited to forecasting the higher frequency IMFs, and a different model for the low frequency ones.

Following Zhang et al. [109] we chose Autoregressive (AR) and Adaptive Neuro-Fuzzy Inference System (ANFIS) models as candidate models for each IMF. We fit a separate model to each IMF with no grouping of IMFs. We implement Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) to minimise mode mixing and to improve the spectral separation of modes [143].

A sliding window approach was implemented, with sliding windows over the training data used to find the optimal number of lags to include as inputs in the ANFIS model, the best performing model for each IMF (AR or ANFIS), the optimal number of IMFs and the optimal window length. Finding the 'best' model for each IMF relies on the assumption that better IMF forecasts will also sum up to a better overall forecast for the original series. CEEMDAN was applied to each window separately as the values of individual IMFs can change, particularly at the boundaries, for different windows. Although this is computationally more expensive than decomposing the entire series at once and all decomposed series are still guaranteed to sum to the original time series under this simpler approach, decomposing the series separately for each window guarantees no 'information leakage' from future values occurs in the decomposition. Once the types of model (AR or ANFIS), number of IMFs to use and sliding window length has been found, forecasts can be generated for all sliding windows in the testing set.

We found that the ANFIS model performed worse than an AR model for all IMFs. It is possible that increasing the number of inputs to the ANFIS model would give improvement, but this also results in a significantly higher computational cost. Five or six IMFs were found to give the best results.

It should be noted that due to the computational time associated with fitting several different models to all IMFs for many windows, choice of the optimal model for each IMF and optimisation of parameters was only done for one step ahead forecasts, rather than optimising separately for every forecast horizon.

2.6.3 Markov chain method

Similar to AR methods where the forecast is dependent on previous values, Markov Chains assume the Markov property: the state at time $t+1$ depends only on the state at time t. Power values are discretised into a finite number of states and the transition probabilities between states result in a forecast probability for each power state, given the previous power observation. These probabilities may then be converted into quantiles to produce a nonparametric probabilistic forecast distribution. A frequentist approach to building a MC model would involve the calculation of a 'transition matrix' of probabilities of transitioning from each (discrete) state at time t, to each state at time $t+1$. The maximum likelihood estimates for transition matrix entries are found using counts of transitions between states from the training data [127]. This produces a transition matrix entirely dependent on the observed training data and may lead to transition probabilities equal to zero between certain states, simply because that transition was not observed over the training period. The uncertainty on the transition matrix entries is not accounted for. A Bayesian approach introduces priors to help account for this; the end forecast probabilities for each state are then effectively an integral over all possible values for the transition matrix entries, taking prior estimates and observed transitions into account. We follow Chen et al. [144] and use a Dirichlet prior which allows for a neat analytic solution. In the Bayesian formulation, the forecast distribution is

$$p(y|x) \propto \int p(y|\Theta)p(\Theta)p(x|\Theta)d\Theta \tag{2.3}$$

where x is the training data and Θ are the transition matrix values: θ_{ij} represents the probability of transitioning from state i to state j. $p(\Theta)$ are the prior probabilities, given by the Dirichlet distribution; for a MC with K discrete states the l^{th} row is given by

$$p(\Theta) = \frac{1}{B(\alpha)} \prod_{j=1}^{K} \theta_{lj}^{\alpha_j - 1}. \tag{2.4}$$

$p(x|\Theta)$ is the likelihood function:

$$p(x|\Theta) = \prod_{i=1}^{K} \prod_{j=1}^{K} \theta_{ij}^{n_{ij}} \tag{2.5}$$

The value of α_{lj} has to be specified for each lj element in the transition matrix, i.e. there are K^2 prior values for a MC with K discrete states. It is reasonable to assume that a transition to a more similar (closer) state is more likely than a large jump in power between time steps, so we constrained the prior values to adhere to this by defining

$$\alpha_{lj} = K - |l - j|. \tag{2.6}$$

To be able to optimise the importance of the observed data relative to the priors, a 'scaling factor' c was also introduced. For a forecast input state l, the final forecast probabilities are then

$$p(y|x) \propto \begin{pmatrix} cN_{l1} + \alpha_{l1} - 1 \\ cN_{l2} + \alpha_{l2} - 1 \\ cN_{l3} + \alpha_{l4} - 1 \\ \cdots \\ cN_{lK} + \alpha_{lK} - 1 \end{pmatrix} \tag{2.7}$$

Finally, this vector is normalised so that the elements sum to 1 (i.e. the total probability across all states is one). The full details of this derivation are given in Appendix B.

Transition probabilities may change over time, so a sliding window using only the most recent data points was employed. The length of this sliding window was optimised over the training set, as well as the number of discrete states K and the scaling factor c giving the relative importance placed on the counts versus the priors. Forecasts could then be produced for the testing period.

2.6.4 Persistence

Persistence models are a common benchmark for time series forecasting methods [42], as they are very simple but often hard to beat on short time scales. Forecasts based on a Gaussian distribution are used, where the mean is equal to the most recent observation and the standard deviation is found from the standard deviation of residuals in the training set:

$$\hat{y}_{t+k|t} \sim \mathcal{N}(y_t, \sigma_t) \quad , \quad \text{where} \quad \sigma_t = \sqrt{\frac{1}{T}\sum_{t=k+1}^{T} (y_t - y_{t-k})^2} \quad . \tag{2.8}$$

2.6.5 Forecast evaluation

When developing new forecasting methods and tools it is necessary to establish some criteria by which success and improvement upon existing practice are defined. Error metrics such as Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) are regularly used in academic literature and practice, with the forecast with the best score declared *the best*. In this case study, as well as exploring three distinct forecasting methods from the literature, we also offer an exemplary comparative evaluation of their performance. We focus on quantitative evaluation, following Messner et al. [142] in particular, but also comment on other important qualitative issues, such as computational time and interpretability [145]. In what follows we briefly introduce the metrics and scores we will employ, and direct readers to [49, 142, 145] and other sources referenced therein for more detailed discussion.

Evaluating deterministic forecasts

We employ two established metrics to evaluate deterministic forecast performance: MAE and RMSE. In both cases, metrics are defined for specific lead-times k steps ahead. For T forecasts $\hat{y}_{t+k|t},\ t = 1, ..., T$ of y_t made k steps ahead at time t, MAE and RMSE are given by

$$\mathrm{MAE}_k = \frac{1}{T}\sum_{t=1}^{T} |y_{t+k} - \hat{y}_{t+k|t}| \quad , \quad \text{and} \tag{2.9}$$

$$\mathrm{RMSE}_k = \sqrt{\frac{1}{T}\sum_{t=1}^{T} (y_{t+k} - \hat{y}_{t+k|t})^2} \quad . \tag{2.10}$$

MAE is favoured by many practitioners due to its ease of interpretation, whereas those in the modelling community may prefer RMSE as it is analogous to the loss function used in model estimation when the objective is to produce conditional expectations as forecasts. Similarly, MAE corresponds to the conditional median. The preference for MAE or RMSE may depend on the individual application and whether the related cost function of the decision scales linearly or quadratically with forecast error. In any case, the ranking of forecasts by performance would not change if one metric were used instead of the other. Normalised forms of MAE and RMSE, where the error is expressed as a proportion of total power capacity, are also frequently used.

Evaluating probabilistic forecasts

Probabilistic forecasts are evaluated according to the principle of minimising sharpness subject to reliability. Reliability is verified using reliability diagrams, and then reliable forecasts may be discriminated using the Pinball Loss (also known as the 'quantile score'). Pinball loss for probability level α and lead-time k is given by

$$\mathrm{Pinball}_{\alpha,k} = \frac{1}{T}\sum_{t=1}^{T} \left(\hat{q}^{(\alpha)}_{t+k|t} - y_{t+k}\right)\left(\mathbb{1}(y_{t+k} \leq \hat{q}^{(\alpha)}_{t+k|t}) - \alpha\right) \tag{2.11}$$

where $\hat{q}^{(\alpha)}_{t+k|t}$ is the predictive quantile of y_{t+k} with probability level α made at time t. Pinball loss is typically averaged across probability levels α to give a single score for the forecasting method being evaluated. It is a proper score, i.e. the minimum score coincides with the optimal estimate of that quantile. As it is estimated per quantile, it is possible to investigate the relative performance of the forecasts across the distribution as well as overall performance across a set of quantiles.

Comparing forecasts

Skill scores are a useful way of comparing forecasting methods because they are unit-free. For a given metric, the *skill* of the candidate method with performance metric M relative to a reference method with performance M_{ref} is given by

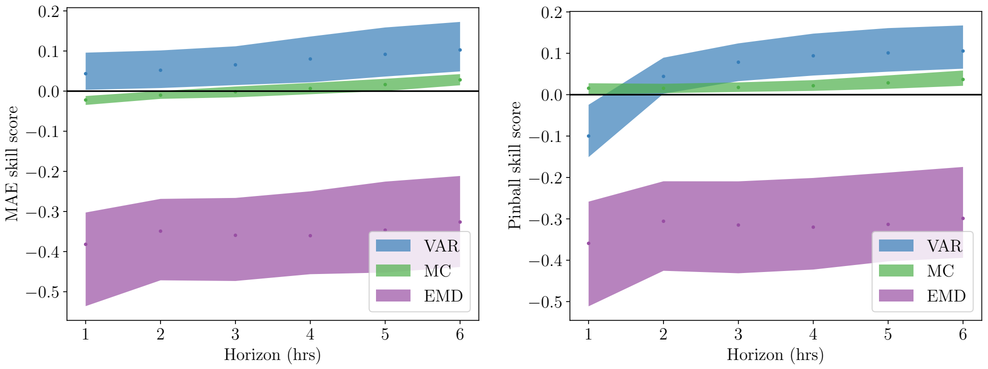
$$\frac{M_{\text{ref}} - M}{M_{\text{ref}} - M_{\text{pref}}} \tag{2.12}$$

where M_{pref} is the 'perfect' score for the given metric, which is zero in many cases. A skill score of zero indicates no improvement relative to the reference method and positive skill score indicates superior performance. A common reference method in wind and solar power forecasting is persistence (or *smart* persistence in the case of solar) and provides a robust benchmark for very short-term forecasts. On longer lead-times climatology is a more common reference method, e.g. the seasonal average for a given time of year.

A practical limitation on forecast evaluation is the finite number of forecast-observation pairs available when calculating metrics and skill scores. As a result, it can be difficult to establish whether any observed difference in performance will generalise or whether it is the result of sampling variation. Bootstrapping is a popular non-parametric method for quantifying the impact of sampling variation [146] involving re-sampling forecast errors and calculating error metrics multiple times in order to estimate sampling variation. To get bootstrapped confidence intervals for a given forecast metric for a set of forecast errors, bootstrap resampling will produce n groups of forecast errors where each group is the same size as the original set of forecast errors by random sampling with replacement. The desired forecast metric is calculated for each of these groups separately to produce a list of n values of the metric. The desired confidence intervals are quantiles of this set of sampled metrics, so the 95% confidence interval would be defined by the 2.5% and 97.5% quantiles. If variation in metrics/skill scores overlap then any difference in performance is unlikely to generalise and may be a result of sample variation. Additionally, bootstrapped skill scores provide greater discrimination than independently re-sampled metrics [142]. However, care must be taken where serial correlation is present, which is often the case in forecasting tasks. Failure to account for serial correlation, e.g. by employing a block bootstrap, would result in over-confident results. Alternative tests for the significance are also available, such as the Diebold-Mariano test [147].

Results

Skill scores were calculated for all zones, using timestamps with forecast-observation pairs available for all zones and averaging the skill score from each zone. Figure 2.3 shows the variation in skill score with forecast horizon for all three models in this case study, relative to probabilistic persistence. For deterministic measures (MAE and RMSE) – which are based only on the q50 value – the VAR model outperforms both persistence and the other models tested, for all horizons. However, the MC model has the best pinball score for one step ahead forecasts, perhaps due to its nonparametric nature and therefore lack of distributional assumptions. The decomposition (EMD) approach is significantly worse than persistence for all horizons.

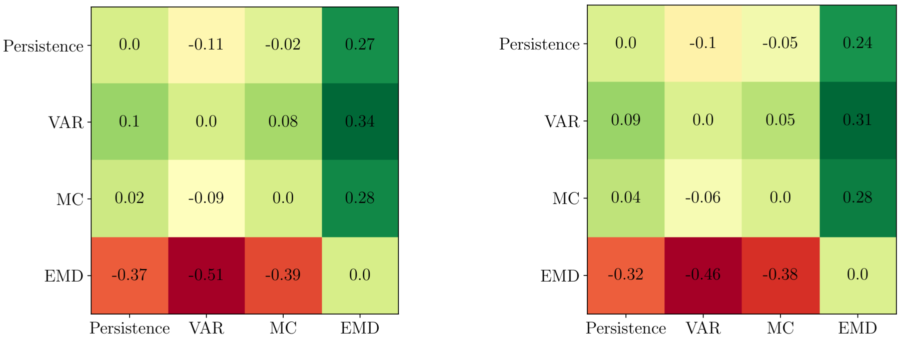


(a) MAE skill score against forecast horizon

(b) Pinball skill score against forecast horizon

Figure 2.3: Skill scores of case study models, relative to probabilistic persistence model. The 95% interval of bootstrap samples is shown. Positive values indicate improvement over persistence. VAR=Vector Autoregressive, MC=Markov Chain, EMD=Empirical Mode Decomposition.

It is also beneficial to compare each model to each other model; matrices of the mean skill score between models are presented in Figure 2.4 for a 2 hour ahead horizon. A positive value indicates the model on the y-axis outperforms that of the model on the x-axis. This clearly shows the EMD approach has the most extreme skill scores, whereas the relative performance of the other models are closer. The VAR model is the only one to outperform all other models. RMSE shows almost identical skill scores to MAE, and pinball skill scores are also similar.



(a) Matrix of skill scores of MAE

(b) Matrix of Pinball skill scores

Figure 2.4: Matrices of MAE (left) and Pinball (right) skill scores for the 2h-ahead forecast produced by all combinations of models implemented. A positive value indicates the model on the y-axis outperforms that of the model on the x-axis. The VAR model outperforms all others in terms of both MAE and Pinball metrics at this horizon. VAR=Vector Autoregressive, MC=Markov Chain, EMD=Empirical Mode Decomposition.

For probabilistic forecasts, the best forecast should be sharp, subject to reliability. This cannot be judged from a single score value such as pinball loss, and so reliability diagrams also play an important role in probabilistic forecast evaluation. Relative Empirical frequency has been plotted, so that a perfect forecast would have a value of zero. For example, it would be expected that in a perfect forecast distribution, the observed power would be less than the q20 quantile forecast 20% of the time and the difference between expected and observed frequencies (the relative empirical value) would be zero. Figure 2.5 shows the reliability across the q5-q95 quantiles for the case study models. Both persistence and to a lesser extent the VAR forecasts display the s-shaped curve associated with too broad a forecast distribution, while the MC and EMD forecasts show bias (under and over-forecasting respectively). The confidence intervals derived from bootstrap resampling show the deviations from 'perfect' reliability are significant for all models.

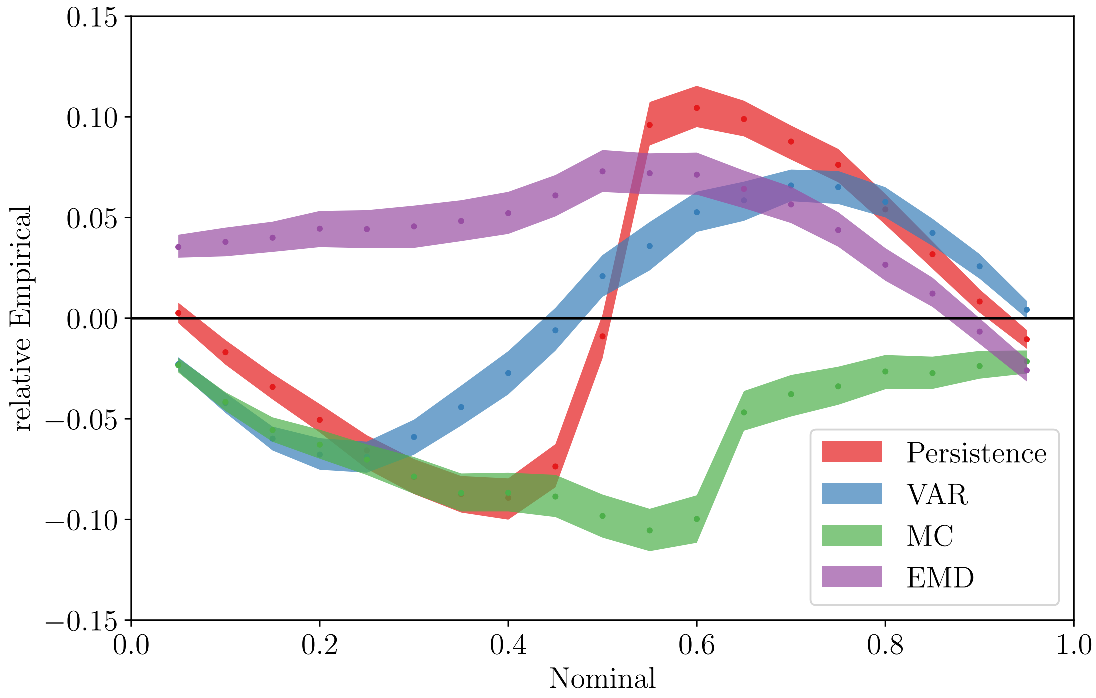


Figure 2.5: Relative reliability of two hour ahead forecasts for the case study models at zone 4. A relative empirical frequency of zero represents ideal reliability. VAR=Vector Autoregressive, MC=Markov Chain, EMD=Empirical Mode Decomposition.

2.6.6 Summary

This case study shows examples of contrasting methods for very short-term wind power forecasting and their relative performances, with the VAR approach proving the most skilful. This is perhaps not surprising given it is the only model to use inter-site dependencies in the forecasts. The Markov Chain model produces a nonparametric forecast, meaning no prior knowledge or assumption of the forecast distribution is needed. It shows superior performance for one step (one hour) ahead forecasts, but its skill is lesser for longer horizons, likely due to the fact it only uses one forecast input (the lag one power value). Decomposition models seem unlikely to provide competitive performance to other methods unless very different models are optimal for the different IMFs, and it can be unclear how best to choose which model to fit to each series. Grouping the IMFs before model fitting was not explored in this work; while this may improve forecast performance, it requires an additional step of forecast setup tuning which would significantly increase the time and effort needed to optimise the forecast setup.

The MC model is the most computationally fast of the models, effectively only requiring to discretise the power values and count the number of transitions between each level. In this study we have fixed the structure of the priors and only tuned their strength relative to the observations, as tuning each prior individually would be much more complex, but other structures could also be explored. While the VAR model takes slightly longer to fit (around a second to train once, tested with 25 different regularisation strengths for 6 different numbers of lags, for each forecast horizon), it fits one model for all locations simultaneously and predicting from a fitted model is still fast. However, the EMD model is significantly more complex both to train and to predict from, due to the additional decomposition step and then models fit separately for all of the decomposed series.

None of these models are perfect, but are intended to serve as open source examples for benchmarking future research and a demonstration of good practice in forecast evaluation. While this case study is only demonstrated for wind data, the code has been made available and could also be applied to solar data, although care must be taken when preparing the data to account for the diurnal and annual cycles. Models that augment inputs (e.g. sky images and other weather data) show improvements [46] but such data were not available here.

2.7 Discussion and future work

Demand for ever more accurate very short-term wind and solar power forecasts has motivated a growing volume of research over the past decade, a trend which shows no signs of slowing. The vast majority of published research focused on wind power in the first half of the decade, but solar has been catching up and in 2019 there was one solar publication for every two in wind. In both cases there has been a shift to probabilistic forecasting, with authors citing benefits for users that become more acute as penetration of wind and solar increases.

The parallel development of very short-term wind and solar forecasting has benefited both fields. Approaches initially developed for wind power, such as exploiting spatio-temporal dependency, have been successfully adapted for solar. Similarly, the relatively well established use of remote sensing data in solar is showing potential for wind. Satellite images capture cloud motions on relevant time scales offering significant benefit for solar forecasting, but no equivalent for surface wind speeds has been demonstrated. A comparison may also be drawn between sky cameras and LIDAR; both are dedicated hardware for measuring the approaching solar or wind resource, respectively. Sky cameras are established tools for very short-term solar forecasting whereas only a few examples of scanning LIDAR for wind exist, likely due in part to significant differences in hardware and maintenance costs. If a sufficient economic incentive (or regulatory necessity) emerges for more accurate very short-term wind power forecasts, remote sensing may represent a suitable opportunity for forecast improvement.

There has also been an increase in the number of proposed methods involving some form of forecast combination from multiple individual models or contrasting data sources, particularly for wind but also for solar power forecasting as the field becomes more mature. The potential for improvement through forecast combination is explored in Chapter 4 of this thesis. In addition to theoretical advances, practical considerations have been the subject of recent research, including handling quality issues and data sharing. Data quality may be compromised by communications failures or operator actions, such as curtailment or integration with co-located storage. When a wind or solar farm is curtailed or metered alongside a co-located battery, its power output is no longer representative of local weather conditions with negative consequences for training forecast models and operational forecasts based on live power data. Where there is a broad literature on this topic in general, application to very short-term wind forecasting has only been considered in work that forms the basis of Chapter 3 of this thesis [4]. A related challenge which has received almost no attention in the literature is the prediction and/or utilisation of *power available signals* from curtailed wind and solar farms on very short-term forecast horizons. Plant controllers can typically produce accurate predictions of present power available but not forecasts of future values. Data sharing between wind and solar farms is necessary in order to capitalise on spatio-temporal information for very short-term and regional forecasting. Some data owners prefer not to share data they consider to be commercially sensitive or private, but it may be possible for them to do so in such a way that improves forecast performance while preserving privacy. In the absence of open data or a central forecast provider, privacy preserving sharing for spatio-temporal very short-term wind power forecasting was first proposed by Pinson et al. [148], later developed by Zhang et al. [91], and recently reviewed by Gonçalves et al. [90]. Furthermore, data markets have been proposed to provide a financial incentive to share data in this way for renewables forecasts [149]. Further development is required to refine such algorithms, which can be demanding in terms of both computation and communication requirements, to develop compelling business models for data markets, and to ensure that they are cyber-secure.

Time-series based methods for both wind and solar forecasting have benefited from contributions from a range of disciplines including statistics, signal processing and machine learning among others. The application and adaptation of optimisation techniques capable of scaling to high-dimensional time series prediction is a good example of this. The significance of proposed methods risks being undermined when case studies are evaluated on small private datasets (hours to days, rather than months to years) and only compared to variations on the same approach. Often methods are evaluated for wind speed forecasting and their suitability for application wind power forecasting is not discussed or verified. Guidance and recommendations for forecast evaluation, including dataset size and properties, benchmarks and significance testing may be found in [36, 142, 145].

The reproducibility of energy forecasting research has improved in general over the past decade with use of open datasets and publication of code becoming more common. The Western Wind dataset [100] covers a large number of US locations, but power data is simulated (using wind speeds and a manufacturer power curve) rather than direct measurement; for this reason it might not be the most appropriate dataset to validate forecasting models on. A number of forecasting competitions have also released datasets, notably the GEFcom series which also publish descriptions of top performing methods and their performance, which may serve as benchmarks for future innovations. However, none of these competitions have featured very short-term forecasting to date, instead focusing on day-ahead time scales where the main task is post-processing numerical weather predictions. Competition formats based on providing training data comprising input-output pairs and a test set of only inputs (with corresponding output held by the organisers for evaluation) does not translate well to time series forecasts where lagged values are a necessary input. A competition focused on very short-term forecasting would be more challenging to run (e.g. running truly live, or requiring participants to submit software) but could make a valuable contribution to the field.

Finally, it is worth noting that methods discussed in this chapter have primarily been developed for use forecasting on very short-term time scales; other methods and input data become more skilful at longer time scales. Chapter 5 discusses the production of forecasts on subseasonal-to-seasonal timescales (2 weeks to a month ahead) where large scale atmospheric patterns become the dominant source of predictability and both different methods and different data sources are used.

Chapter 3

Missing data in wind power time series: Properties and effect on forecasts

This chapter is based on the work presented in the paper 'Missing data in wind farm time series: properties and effect on forecasts' published in Electric Power Systems Research [4]. The text from this article has been edited and extended here.

Abstract

Missing or corrupt data is common in real-world datasets; this affects the estimation and operation of analytical models where completeness is assumed or required. Statistical wind power forecasts utilise recent turbine data as model inputs, and must therefore be robust to missing data. We find that wind power data is 'Missing Not at Random', with missing patterns also related to the forecast output. Approaches for dealing with this missing data in training and operation are proposed and evaluated through a case study, leading to a suggested forecasting methodology in the presence of missing data. In the training set, missing data was found to have significant negative impact on performance if simply omitted but this can be almost completely mitigated using multiple imputation. Greater increase in forecast errors is seen when input data are

missing operationally, and re-training forecast models using the remaining inputs is found to be preferable to imputation.

3.1 Introduction

In a real world setting data is not always tidy and complete; this can cause problems for models that are designed only to work with complete datasets. There are three main instances of missing data that occur in the normal course of the production of a wind power forecast, all with different best approaches for being overcome and consequences if not dealt with properly. Firstly, a large dataset of input/output cases are needed to train the model. Any missing historic values will affect the completeness of this training dataset; statistical models like autoregression require a large complete training set for model fitting and if this training set contains missing values, model fitting may return unexpected or NaN values if no method of treating them is implemented. For example, if ordinary least squares is used to find the minimum of the error function through matrix multiplication and NaNs are retained in the X matrix, they will propagate into the final matrix of model coefficients. This type of 'operational' missing data is expected at the majority of sites and encompasses all missing data that may be routinely expected, including both times when values are actually missing e.g. due to a data recording failure, but also times when the recorded power output is not reflective of unconstrained generation. These times include when a turbine is down for maintenance or during curtailments. While operational missing data is common, the level is expected to be low enough that reasonable mitigation is possible with suitable methods. The second instance of missing data is in the inputs to the forecast generation process once a model has been trained. The intention is to regularly re-issue forecasts using the most up-to-date site information: however, data arrival latencies can vary by site and through time. In general, more remote sites or older sites with a slower data connection are likely to have longer average latencies but also data arrival times can vary from a few minutes to over 24 hours for the same site. This means there is no guarantee of what information will be available to use as a forecast input at the time of issue. Again, simply adding a 'NaN' for any missing values will result in a 'NaN' value being issued as a forecast, which is clearly not appropriate. Thirdly, newer wind farms will have very limited historic on-site data which is needed for model training and so methods to reproduce these measurements for a long single time period are needed. This is particularly important when a setting up a forecast model for multiple sites where it is guaranteed the historic data will be of different lengths at different sites. In these cases other data sources may be useful to fill long gaps. The optimal methods are likely to depend on the length and number of missing periods across all variables. It is clear that missing data in each of these three areas must be addressed in order to set up an operational forecasting model. As there are several possible approaches for missing data treatment in each case, it is important to identify the most appropriate option. This work focusses on the generation of forecasts for multiple sites within one model, allowing development of missing data methods that take advantage of inter-site dependencies which also improve overall forecast accuracy.

While it is possible to make intuitive predictions about what missing data methods are likely to result in the best forecasts, an empirical case study is needed to test out the techniques on real data. It is expected that methods that result in information loss (e.g. dropping rows with missing values and thus reducing the size of the training dataset) will perform more poorly than those that use all the available data. The availability of datasets from multiple sites has advantages in the process of missing data as independent locations are less likely to be missing concurrently, meaning techniques that infer missing values based on the remaining data from the same time point are applicable. In theory, access to both wind speed and power measurements allows inference of one from the other through a power curve, though the mechanisms for missing data mean both variables will often be missing simultaneously. The disadvantage of missing data from multiple site data streams is the increased likelihood of a given time point experiencing missing data at one or more locations, meaning a much higher proportion of time points will experience some missing data compared to a single site time series.

After a review of existing missing data methods across subject areas (Section 3.1.1), the properties of missing data across a set of example sites are found in Section 3.2. Case studies for all of the missing data instances described above are presented in Sections 3.4

and 3.5, exploring forecast performance under various missing data scenarios and with selected mitigation techniques.

3.1.1 Previous literature

There has been little work to describe the general missing properties seen in wind farm operational data, and while some works have considered missing data in wind power time series for other applications, its impact on forecasts has not been assessed. The impact of missing data on monthly and annual average measurements was discussed for wind energy resource assessments [150] along with the corresponding impact on revenue [151]. Other applications include power curve estimation [152], wind farm control [153] and fatigue assessment [154], sensor fault diagnostics [155] and site-level data for wind integration studies [156].

In very short-term wind power forecasting studies, subsets of data with missing values are often simply omitted, which may bias model estimates and is not an option when producing operational forecasts. Recent works have focused on high dimensional modelling [86], dynamic models [89] and data sharing via privacy preserving algorithms [91], for example, but with the implicit assumption of data completeness.

Other research has presented methods for filling missing data in a wind time series; however, the simulated missing values are selected randomly throughout the time series [157] which does not reflect real patterns of missing data. Lotfi et al. [158] uses imputation by simple autoregressive or moving average models which are not suited to filling extended periods of missing data. The purpose of filling in a time series is generally to allow further analysis, for example to calculate energy yields from power or to detect sensor failures. By only reporting the accuracy of the imputation process itself, the financial or decision-making consequences of the proposed imputation methods are not addressed.

Fields that often utilise longitudinal studies, such as medical trials and political behaviour studies, have traditionally encountered significant levels of missing data [159] and as such have developed methods to quantify and account for its effects on study outcomes. Central to these methods is the classification of missing data into one of three types [160]: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). For a set of observed values of an input variable X^{obs}, and missing observations X^{miss}, the probability that any single observation x_t is missing is $p(x_t \in X^{miss})$ and may, in general, depend on the values of any variables in the dataset (both observed and unobserved, input and predictor variables). Allowing the probability of missingness to depend on variable values means the probability is now conditional on these other factors:

$$p(x_t \in X^{miss}) = p(x_t \in X^{miss}|x_t, y_t) \tag{3.1}$$

where y_t represents other variables measured at the same time point. The three types of missing data may be expressed in terms of the dependences (or lack of) on x_t and y_t.

Data may be classified as MCAR when the probability of a data point being missing is completely independent of any variables in the dataset, that is to say, there is a constant probability of missingness for all values and Equation 3.1 becomes

$$p(x_t \in X^{miss}) = C. \tag{3.2}$$

In this case, the remaining complete data has the same distribution as the original population with missing data and so there will be no bias introduced to the model estimation. However, this type of missing data is unusual in reality as the mechanisms that produce missing values are very rarely completely random.

MAR data occurs when ‘missingness’ in one variable is independent of its own value but does depend on the value of another variable. An example of this would be prices on a stock market that is closed at weekends: these missing points are conditional on another variable such as day of the week but don’t depend on what the actual stock values would have been if the stock market was open on that day. A MAR condition is expressed as

$$p(x_t \in X^{miss}) = p(x_t \in X^{miss}|y_t) \tag{3.3}$$

Many commonly used methods described in later sections assume the MAR condition; however, it is often impossible to positively verify a MAR pattern in data [161].

Finally, if the probability of a point being missing is dependent on the value it would have taken, then the missing pattern is classified as MNAR. In this case, the distribution of values in the remaining observed cases is not the same as that in the missing data points and so any model analysis ignoring this discrepancy will produce biased results. In reality, many cases of missing data do in fact depend on - or display correlation with - the missing value itself; for example in a health study, individuals who are more susceptible to disease are more likely to miss an interview due to illness. In this case there is a dependency between the missing value and the probability of missingness, and thus any method designed for data analysis or model fitting must take this into account. This is usually done through some knowledge of the mechanism for missingness for that given scenario, or information about the prior distribution [162,163]. An incorrect assumption of MAR or MCAR patterns will introduce bias to final results, as the properties of the observed dataset are not representative of the whole population. As such, this situation is often described as 'non-ignorable' due to the need for direct modelling of the missing mechanism and Equation 3.1 cannot be simplified.

Since the possible subset of methods used to deal with missing data depends on the type of missingness, it is important that researchers are aware of the differences and begin any data analysis with an attempt at characterising their missing data. Often knowledge of the circumstances of data collection and any underlying processes will give a good idea of the type of missingness likely present; however, given the importance of correct classification for all further data analysis, a formal categorisation is desirable. The likelihood ratio test first proposed by Little [164] distinguishes between MAR and MCAR data; however, a conclusive categorisation of data as non-ignorable missing is much more complex and often not practically possible. Beyond contextual justification, sensitivity analysis may be used where different missing data models are applied and the categorisation inferred from the accuracy of the final model results [162,165]; however, this can quickly become labour-intensive when considering several complex models. For regression models, the missing indicator method provides a suggestion of the significance of a difference in distribution between observed and missing values [166]. For this, an extra binary variable is created to indicate whether the corresponding data point in the original variable is missing (if all variables are to be tested, this will result in a matrix the same size as the original training matrix, but with a 1 in place of every present value and a 0 in place of every missing value). The indicator variable(s) are then included in the regression analysis in addition to all the original variables: if the regression coefficient for the indicator variable(s) is significant, the missing data follows a different distribution from the observed data and it is therefore classed as non-ignorable missing.

Shmueli [167] distinguishes between approaches for causal explanation versus prediction at all stages in the modelling process, and notes that when dealing with missing data in a prediction setting, the relationship between the missing points and the response variable is the dominant factor in determining the best method to use. Previous work has compared the performance of missing data methods on classification trees for binary response data [168] with a short extension to logistic regression. However, there is very little work showing the performance of missing data methods on forecasts of a continuous response variable. Further, the ability to utilise dependencies between variables (such as modelling multiple wind farms in the same VAR model) may affect the relative performance of missing data methods.

The extent of missing data and the variable structure within a study or application may also determine the effectiveness of various techniques; for example, a large proportion of missing data or missing points spread across many variables will result in a small subset of the data being complete. Large information losses may affect some forecasting methods more than others. The application of statistical models to short term forecasting uses a historic dataset to train a model, before the fitted model is applied to the most recent data to obtain forecasts. As such, there are two data sets used in the generation of a forecast: a historical set of all available past measurements and a set of forecast input measurements to generate the new forecast. Missing data in a forecast input cannot just be deleted as the input is necessary to generate a forecast. Because the number of forecast inputs is generally orders of magnitude smaller than the number of data points in the historic training dataset, a missing point in the forecast inputs represents a larger proportional loss of information and may be expected to have a greater detrimental effect on the forecast error than a single missing point in the training data. Of course, a larger number of missing points are also expected in the larger training dataset. The large size of the training dataset allows the modelling of overall trends and dependencies so the model is able to capture the general, typical relationships in the data with minimal disturbance from noise or abnormal values in individual time points. The main methods found in the literature are outlined below and deal with missing data in a static dataset such as the training data for a statistical model. 'forecast input imputation' and 'forecast re-training' methods describe methods for dealing with missing forecast inputs.

Complete and available case analysis

The two simplest methods commonly used are complete and available case analyses. Complete case analysis involves ignoring any time points or study members with only partial information [165] (Figure 3.1). As discussed earlier, this method will produce biased results if data is not MCAR because the sub-sample of observed data is no longer a random sample from the whole population. Available case analysis aims to reduce the amount of information lost through the systematic deletion used in complete case analysis, and uses all observed values of any single variable to compute its properties. However, it is not clear when this will produce unbiased statistics, and values in covariance matrices lose physical meaning through the treatment of all variables individually [166]. Weightings based on a model for the probability of a point being missing may be used to counteract the bias induced by these simple missing data treatments [169] (Figure 3.2), but where missing values are spread over many variables the reduction in size of the training data set may still contribute to degradation in forecast performance.

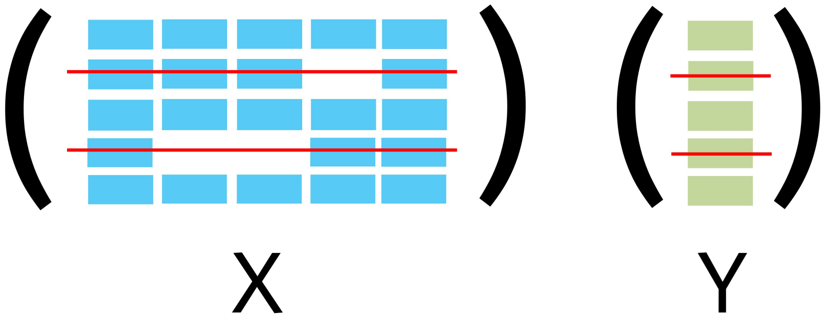


Figure 3.1: Rows with missing data are deleted in complete case analysis

Imputation

Another popular ad-hoc method is mean imputation, where any missing values are simply filled with the mean value for that variable [170]. This allows for preservation of the sample means, but sample variance is reduced as the real variance of the unknown values will be greater than zero. For an application in forecasting where quantification of the uncertainty in a value is perhaps even more important than the value itself, this is clearly inappropriate. Mean imputation constitutes one case of single imputation, the procedure of filling any missing value with an informed guess. This class of models also includes hot-deck imputation, used for time series data and where a missing value is filled with the last observed value. While this may be an acceptable method for individual missing points, this is not suited to data with long spans of missing data as in the case of a turbine down for maintenance.

Because filling values results in an apparently complete data set, uncertainty estimates that do not take into account additional error from loss of information at missing values will produce consistently over-confident results. However, imputation does retain all the non-missing data points while creating a complete rectangular data set for model training. Multiple imputation reconciles the need for complete data with an accurate estimation of final uncertainty in any results derived from the data by filling

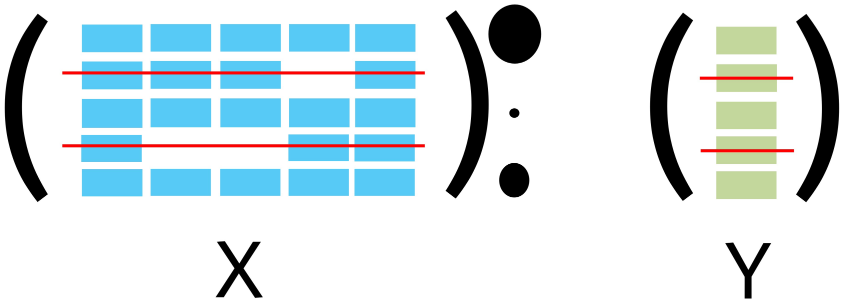


Figure 3.2: Rows with missing data are deleted and the remaining rows are weighted by the inverse of their modelled probability of containing missing data

the data set several times according to a probability distribution for the missing values [162,166,171,172]. This produces multiple completed data sets, each of which may then be used for any further analysis with the ultimate result becoming a combination of those from each imputed data set. This process carries the additional uncertainty from missing values throughout all analysis, giving accurate estimates for desired quantities and their confidence levels. The distribution of missing values must be known to allow multiple imputation, or more involved statistical methods such as a Markov chain Monte Carlo algorithm [173, 174] should be utilised. Multiple imputation is versatile and independent of subsequent models used for data analysis; however, care must be taken to correctly identify the distribution of missing values [175]. It also assumes variables follow a multivariate normal distribution; while this is generally not strictly true, it has been shown that estimates found assuming a multivariate normal distribution are generally as good as those from more complex and rigorous alternatives [171].

Maximum likelihood

The alternative to filling in missing values is to develop a forecast model that incorporates the possibility for missing values. As the name suggests, maximum likelihood methods involve maximising the probability of the real data (including the points

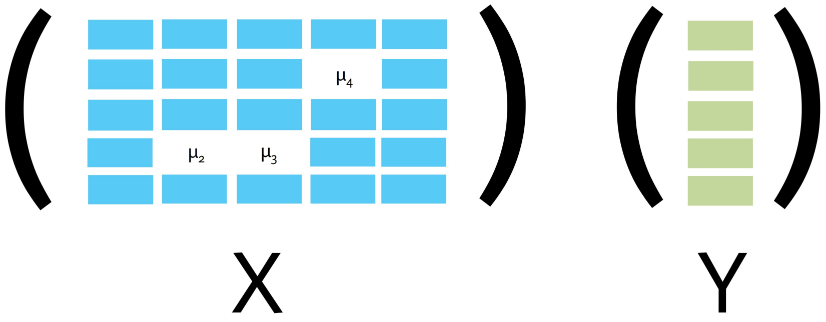


Figure 3.3: Points with missing data are filled with the column mean.

classed as missing) arising from the model fitted to it. Maximum likelihood techniques are applied in statistics outside of missing data scenarios by formulating an expression for the probability of observing the real data, given the underlying parameters. The maximum of this expression with respect to the model parameters then gives the maximum likelihood estimate of the model. Missing data adds an extra element to this maximisation problem. The EM (Expectation Maximisation) algorithm was developed by Dempster and allows calculation of maximum likelihood in the presence of missing values [163]. It involves iterating over two steps: for the first step, the expected values of the population distribution parameters are assumed, allowing computation of the complete data likelihood conditional on these parameter guesses. This likelihood is then used in the maximisation step to recalculate the population distribution parameters. This process of alternate steps is repeated until all estimates converge to stable values. More recent work suggests contemporary algorithms and machine learning techniques may produce faster convergence than the EM algorithm [176]. In a forecasting context, a PEM (prediction error minimisation) approach may be preferred as it directly optimises the end forecast errors [177].

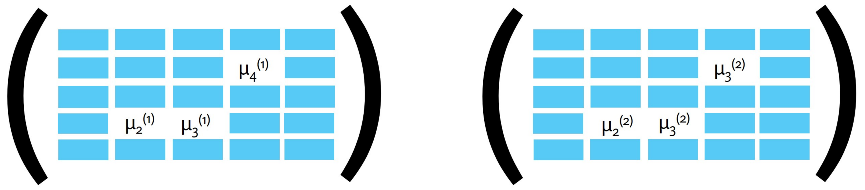


Figure 3.4: Points with missing data are filled with an estimate of the value, sampled from a probability distribution. This is repeated n times to get n complete training datasets that together retain uncertainty about the values at the missing points.

Other approaches

Signal processing applications have used dimensionality reduction of the large matrix of data using Hankel structures [178, 179]. This has been implemented on data from phasor sensors monitoring electricity grid frequency but is not commonly used in a forecasting context. This method also relies on collinearity between variables which may not be strong between certain wind farm sites, and may also vary through time depending on weather regimes and prevailing wind directions. Methods based on machine learning have also been put forward [180–183] but may be less interpretable and more computationally complex to implement and run.

3.2 Missing data properties

It is important to first establish the amount and type of missing data seen in real wind turbine Supervisory Control and Data Acquisition (SCADA) datasets, in order to select appropriate mitigation techniques to test in the following case studies and to ensure the missing data cases studied are realistic scenarios.

For wind turbines, any measurements taken during non-routine operation may be considered invalid or missing as they are not representative of the unconstrained behaviour that data analysis is generally aiming to capture, i.e. power production may not match what the wind farm or turbine is normally capable of in those wind conditions. As such, three main sources of missing data were identified: data missing in the raw time series due to sensor measurement, recording or communications failures; missing periods due to site-wide maintenance works; and curtailments (when controller action is taken to limit power output). The shutdown of individual turbines may be compensated for by renormalising site power production and so is not considered as missing data. Curtailment here means times when the site has accepted a BOA (Bid Offer Acceptance) instruction from the Electricity System Operator (ESO), requiring the entire site output to be limited to a set power value below what it would be producing under unconstrained operation. This is generally due to grid constraints or lower demand than supply. The proportion of time points in the series affected by each of these missing sources was found separately in addition to the combined effect. It was assumed that wind speed measurements were still valid when the turbine was not operational for maintenance or during curtailments, although it should be noted that the anemometry system on the turbines has been designed for accurate measurement during normal operation through the use of a nacelle transfer function and thus the accuracy of these measurements will be reduced when the turbine rotor is not turning [184]. The distributions of lengths of missing data are also found, giving insight into the likely missing data mechanism.

3.2.1 Testing for MNAR patterns

Planned maintenance activities are often scheduled for times with lower wind speeds and any work on a turbine will have an associated maximum safe wind speed over which activities will be cancelled; this suggests a correlation between times of missing data due to planned maintenance and the value of missing variables, making the data MNAR. Wind farm sites may be more likely to be curtailed close to rated power from grid constraints limiting power flows; again this would cause an MNAR data pattern from curtailments. In addition to the physical justifications given above, an MNAR pattern was tested for using the missing indicator method [166] for each site wind speed and power variable. Neither contextual justification nor the missing indicator method give a definitive classification of the data as MNAR [161], but they do give a suggestion of whether it is reasonable to assume MNAR as the most likely missing data type. The missing indicator method involves adding a binary indicator variable into a regression model to encode whether the original variable is present or missing for every time step. A missing indicator is included for every input and a linear regression using one lag at each site is performed. The model is then formulated as

$$y = X\beta_X + Z\beta_Z \quad , \tag{3.4}$$

where Z is the matrix of indicators, with one for each element in X. An example input matrix for a regression, and its corresponding indicator variables, is:

$$X = \begin{pmatrix} 6 & 4 & \mathrm{NaN} & 6 \\ 9 & 8 & 0 & 3 \\ 3 & \mathrm{NaN} & 6 & 9 \\ 8 & \mathrm{NaN} & 1 & 2 \end{pmatrix} ; \; Z = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \tag{3.5}$$

Missing elements in X are then filled with zeros to allow model computation. Because all variables are also normalised to zero mean, this is equivalent to mean imputation. The main effect of this is the decreased relative significance of that particular variable's coefficient (zero contribution from that variable to the end forecast for that input/output pair implies greater contribution from the other variables). This only serves to slightly decrease the significance of the variable coefficient in β_X which is then used to rule out site forecasts where the variable's coefficient is not significant, making the end results slightly more stringent if anything. The desired output from this test is the significance of the coefficients rather than an accurate forecast, when the use of mean imputation in this way may be of more concern. Coefficients in β_Z that are significantly different from zero imply that the mean of the missing data points is not equal to that of the non-missing data. Therefore, the observed and unobserved data follow different distributions and the data for this variable can be said to be MNAR at the chosen significance level.

It is not necessary to include more than one lag of each variable, as the pattern of missing data will be the same in each lag (each is a shifted copy of the original time series). P-values from a t-test are used to determine significance of the regression coefficients at the 5% confidence level. As a Vector Autoregressive (VAR) model framework is used, each variable has a significance associated with its use in the forecast for each site separately. The proportion of site forecasts for which the indicator variable is significant then gives a suggestion of the likelihood that variable displays MNAR missing data.

3.3 Missing data mitigation methods

Data may be missing from the training dataset in both the forecast input (X) and target variable (Y) matrices, compromising parameter estimation, or input data required to generate a forecast may be missing, compromising the production of operational forecasts. In addition to the ideal case with no missing data, the following missing data scenarios are simulated to evaluate the effect of different types of missing data on forecast performance:

- Missing training data: five levels of missing training data in X are synthesised, mimicking patterns seen in real life datasets
- Missing spatial forecast inputs at a single time instance, i.e. the columns in X for multiple sites with the same time lag are missing when generating a forecast (after a model is trained)
- Missing temporal forecast inputs from a single location: the columns in X for multiple lags from a given target site are missing when generating a forecast (after a model is trained)

From this, five cases are tested. Cases 1-4 represent missing live forecast inputs needed to generate a forecast, whereas case 5 examines the effect of missing training data with three treatment methods tested. A forecast is output for all sites and the percentage difference (‘worsening’) from the full case with no missing data is calculated at each site before being averaged over all sites.

Case 1: varying the number of sites included as model inputs

Case 2: varying the length of time a single site has been unavailable for (lags are missing cumulatively, from the most recent one)

Case 3: varying the single lag that is missing, i.e. comparing the effect of missing information at one site from 30 minutes versus 60 or 90 minutes ago.

Case 4: varying the number of sites with the most recent piece of information missing. This is important for deciding whether to issue a new forecast that includes very recent information at one or two sites, or whether to stick with an older forecast where information is known for all sites.

Case 5: Varying the amount of missing data in the training dataset, trialled with three different methods for mitigating it.

The structure of the forecasting process, including the missing data methods tested, is set out in Figure 3.5. The methods listed are described in detail in the next two sections. Table 3.1 shows the various case studies and the reasons for choosing these particular scenarios.

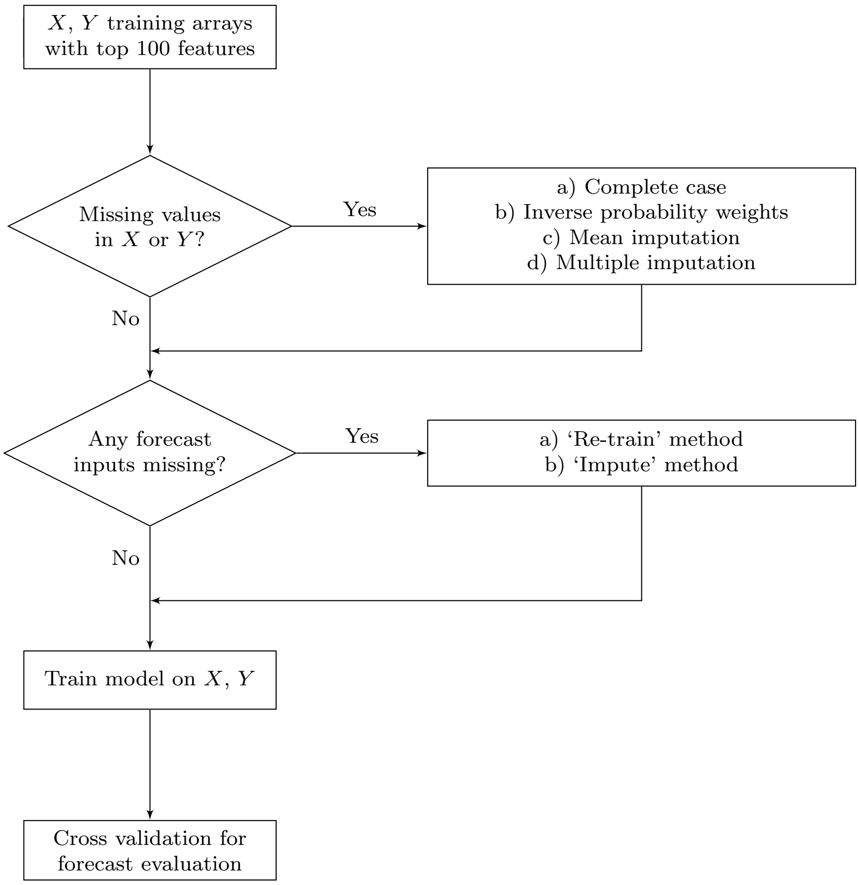

X, Y training arrays with top 100 features
Missing values in X or Y?
Yes
No
a) Complete case
b) Inverse probability weights
c) Mean imputation
d) Multiple imputation
Any forecast inputs missing?
Yes
No
a) 'Re-train' method
b) 'Impute' method
Train model on X, Y
Cross validation for forecast evaluation

Figure 3.5: Flowchart of the process for identifying and dealing with missing data within forecasts. Multiple methods a), b), c), d) were applied to find the optimal approach for each stage.

Case study	Where missing?	Description	Why chosen	Shows
1	live inputs	Whole sites are completely excluded	Study the effect of number of sites in the model.	More sites = more complex model but greater probability of similar sites, improving skill. Effect of distance between sites was not explicitly tested.
2	live inputs	Different lengths of missing data tested	How much worse is a long period offline than short?	Effect of a longer missing period is cumulative but greatest loss in skill from missing the most recent information
3	live inputs	Different single missing lags tested	How important is the most recent information?	The most recent lag is the largest contributor to forecast skill.
4	live inputs	How many sites missing most recent lag	Can other sites compensate affected when many sites drop out simultaneously?	Forecasts do worsen when more sites are missing but the effect is quite gradual
5	all through training data	Missing data in historic set from data quality issues and curtailments	This occurrence of missing data is almost inevitable; case study investigates what methods mitigate it best	Imputation methods perform best and multiple imputation beats mean imputation, though it is more computationally complex
6	large chunks of training data	Replicates sites with limited historic data, eg new sites	To find out which alternative information sources may be valuable for filling long histories	No one best method but both reanalysis and Balancing Mechanism (BM) data can help.

Table 3.1: Summary of case study cases, their purpose and main results. The results of cases 1-5 are described in Section 3.4.2 and case 6 is described in Section 3.5.

3.3.1 Missing values in training data

Due to the large size of the training dataset needed for model fitting, the likelihood of any given variable (column in X) containing any missing data is high. Because variables originate at different sites, it is unlikely that missing values will be simultaneous across sites. This means the proportion of individual time points (rows in X) that contain any missing values is higher than the overall proportion of missing data. Starting with a complete dataset so that a comparison with the 'ideal' case can be made, the levels and patterns of missingness seen in real datasets must be reproduced. The dataset displaying real patterns of missing values is labelled as dataset A, and the complete dataset where missing patterns are replicated is labelled dataset B. The missing data patterns observed in dataset A were replicated in dataset B in the case study to allow comparison to the complete data case through the use of the same dataset but with missing values introduced in a realistic pattern. Although the creation of MNAR missing data patterns has been studied [51], the methods focus on datasets with a small number of variables or where the 'rules' for missingness can be simply simulated. The availability of a dataset with 'real' missing data allowed for a nearest neighbours approach to reproduce missing patterns. The two pairs of most correlated sites between datasets A and B were found using the R^2 correlation coefficient and then used to calculate the Euclidean distance between power values in Y in datasets A and B. For each row in Y in dataset B, the most similar row (nearest neighbour) for the two most correlated sites in Y in dataset A was found. The missing data pattern from the corresponding row in X in dataset A was then reproduced in that input/output pair of dataset B to give the 'closest' reproduction of missing data, labelled 'medium'. Datasets with deliberately higher and lower levels of missing data were created following the same procedure but using a different number of nearest neighbours and picking the highest or lowest missing data pattern within this subset as the one to replicate (Table 3.2, 'knn' column). An approach using the probability of missing data in a certain variable given the output power was also tested but resulted in all input/output pairs containing missing data, leaving no training data for the complete case analysis. This is perhaps due to the lack of dependency between missing data across variables in this method.

	knn	missing data %	% rows with missing data
low	3	1.36%	42%
low-medium	2	2.48%	56%
medium	1	6.15%	76%
medium-high	2	9.57%	94%
high	3	11.65%	99%

Table 3.2: Missing data created in the complete dataset. 'knn' gives the number of nearest neighbours selected from and '% rows with missing data' indicates the reduction in size of the available dataset when using complete case analysis.

Forecasts are now generated using dataset B, both using the complete unaltered dataset to get the 'ideal' forecast performance, and with missing data applied and the four different missing data techniques set out in Figure 3.5 are tested. Using complete case analysis [165] will result in a large amount of information loss and a greatly reduced number of time points for model training - loss of up to 99% of rows with high levels of missing data (Table 3.2). As the missing data is likely MNAR, complete case analysis will result in a biased training dataset as the fitted model is not representative of all behaviour seen over the training period. However, complete case analysis is widely used in practice as it is a very simple method to implement and thus it is included in the case study to understand the potential detrimental effect of this common method. Since complete case analysis is often used and well understood, the extended method of correction by applying inverse probability weights to counteract bias is also tested [169]. In this method, the probability of a row being complete is first found through logistic regression (the model is trained on the same inputs as the final forecast, but with a binary output representing whether that set of inputs is complete or contains missing values). Rows that have a low probability of being complete (high probability of being missing) are likely to be under-represented in the remaining complete case set, so a high weight (calculated as the reciprocal of the probability of being complete) is used to correct the skewed representation of the population by the complete cases. Due to the importance of variable selection for this model, principal component analysis was used to ensure no linear dependencies between inputs. Care must be taken that a probability of zero or less is not returned for any rows that are actually complete as this would result in an infinite or negative weight as well as being nonsensical (the probability of being complete must be greater than zero for every row that is observed to be complete). It is expected that this method will outperform complete case analysis due to the additional steps taken to correct sample bias, but may still not be ideal as information in partially missing training rows is still completely discarded in the final model fit. Mean imputation, where all missing values assume the mean value for that variable, is another commonly used and easily understandable technique. While it does preserve all the available information for use in the model fitting process, all missing points on a given variable are filled with a single value (the mean), resulting in a lower overall variance than the true values would have had. In the case of MNAR missing data, the mean of the missing values is not equal to the mean of the remaining values, so mean imputation will introduce a population bias into the training dataset.

Multiple imputation is also implemented; this method requires assumption of the distribution of the missing values to pick replacement values from. By repeating the whole imputation process n times, n complete datasets are created with the variation in the imputed values between them. By conducting all subsequent analysis using each imputed dataset separately, n final forecast errors are produced with the variation between them giving the additional uncertainty introduced by the presence of missing values. However, in practice the distribution of missing values is rarely known, as is the case for missing SCADA values. When the shape of the missing values' distributions are not known, assumption of a multivariate normal distribution is generally sufficient even if not strictly correct [171]. In a multivariate model like this case study, other related variables are available to form a model to predict the missing values within a Markov Chain Monte Carlo (MCMC) framework based on that described by Schunk [174]. Assuming missing values are spread across all variables (columns in X), missing points must first be filled in with an initial value to allow the iterative process of imputing values to take place. Commonly the mean is taken as the starting value. If only one variable in X has missing values then no initialisation is needed and the imputation step need only be carried out once.

For the iterative process, each of the missing values of each variable are imputed in turn (starting with the variable with the lowest amount of missing data), by modelling the probability distribution for each missing value and picking a value from this as the new imputed value. The whole process of imputing all missing points across all variables is repeated until some convergence criteria is met. For this case study a regression model is used to predict the mean of the missing values, $\hat{\mu} = X\beta$ where β is found by ordinary least squares and X includes all the other model variables. The imputed value y is then picked from the normal distribution

$$N(\hat{\mu}, \hat{\sigma}^2) \text{ , where} \tag{3.6}$$

$$\hat{\sigma}^2 = \frac{1}{n-k}(Y^TY - Y^TX(X^TX)^{-1}X^TY) \text{ ;} \tag{3.7}$$

n is the number of observations (number of rows in X) and k is the number of variables (number of columns in X). Y are the values for the variable we are trying to impute, for all the complete rows, with X containing the other available variables for the same rows. Care must be taken to ensure $\hat{\sigma}^2$ is positive; negative values can result from an ill conditioned X^TX matrix, meaning its inverse is not precisely calculable: $(X^TX)(X^TX)^{-1} \neq I$. Reducing or removing collinear variables or using principal component analysis can solve this issue.

Convergence may be judged on mean change in imputed values for a chosen variable between each iteration, or other criterion metrics taking position and dispersion into account [174]. Figure 3.6 shows convergence of the MCMC process, measured as the average difference in imputed values between iteration steps for the variable with the highest level of missing data.

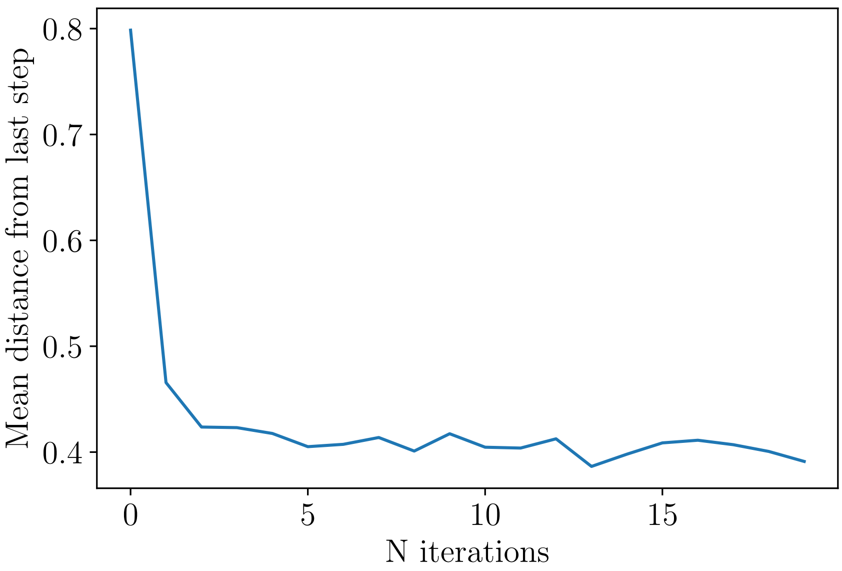


Figure 3.6: Convergence of the MCMC imputation algorithm.

3.3.2 Missing inputs for forecast generation

Missing input data (when incoming data feeds with the most recent information are down) mean a new forecast cannot be generated without model adjustments or additional steps. Data arrival latencies vary both between sites and through time, meaning the particular variables with most recent measurements available will vary. Some simplified cases are tested to examine the effect of different scenarios. For each different case 1-4 of missing data as laid out in Section 3.3, two approaches to deal with missing inputs are considered. In the first approach, alternative models are fit which do not require the missing value(s), and in the second, missing data are filled with estimates. In the first case, named the *'re-train'* method, the linear model is re-configured and re-trained without the missing forecast input (columns are dropped from X and corresponding elements from β). Forecast training is not computationally expensive as long as the number of forecast inputs (number of columns in X) is not too large, hence the selection of the most informative 100 inputs as a precursor step in the forecasting process to allow fast re-estimation. In the second case (the *'impute'* method), a regression model is fitted to predict the missing values(s) using the available forecast inputs.

The original forecasting model with all forecast inputs is then used. This model also requires training and evaluation of a new regression model to impute values for every different combination of variables missing, but again this requires no more than a few seconds of extra computational time. If multiple forecast horizons at multiple sites are being trained in one model as is the case with the VAR model used in the case studies in this chapter, the re-train approach involves re-estimating ((ninputs-1) × nsites × nhorizons) coefficients when one forecast input is dropped, while the impute method uses the previously trained main forecasting model and the secondary model used to find the value of the missing input only requires estimation of (ninputs-1) coefficients.

3.4 Case study: SCADA datasets with missing values

The dataset used to study the missing data properties in real SCADA is labelled as 'dataset A', and comprises wind speed and power time series with missing data from 30 European wind farms (though with the majority in the UK). The set has a mean site capacity of 41.8 MW and a range between the different sites of 129.8 MW. Two years of 10-minute resolution data was re-sampled to 30-minute resolution in line with the time resolution used in the case study. The dataset of complete time series at half-hour resolution for 10 UK wind power sites is labelled 'dataset B'.

First the properties of missing values in dataset A are investigated: Figure 3.7 shows the levels of missing data seen at the sites analysed. Although the majority of sites displayed low levels of missing data with medians of 2.70% and 1.57% for power and wind speed respectively, there are several sites with much larger proportions of missing data, even over 30% at one site. Some sites may have a much larger proportion of missing data from curtailments if their bid price is favourable so they are asked to curtail more frequently. Older sites may experience higher rates of sensor failures and have less capable data transfer infrastructure. More remote sites may also see longer data latencies and possible more instances of missing data. The proportion of data missing is broken down by cause (no data recorded, maintenance or curtailments) in Table 3.3. This shows raw data missing is the main cause, followed by curtailments and then maintenance activities.

	Power	Wind Speed
raw data	5.32%	4.86%
maintenance	0.57%	0%
curtailments	2.89%	0%
Overall	5.71%	4.86%

Table 3.3: Mean missing data proportions by type, averaged across all sites. Due to overlap between different missing data causes, the total proportion of missing data is less than the sum of each individual type.

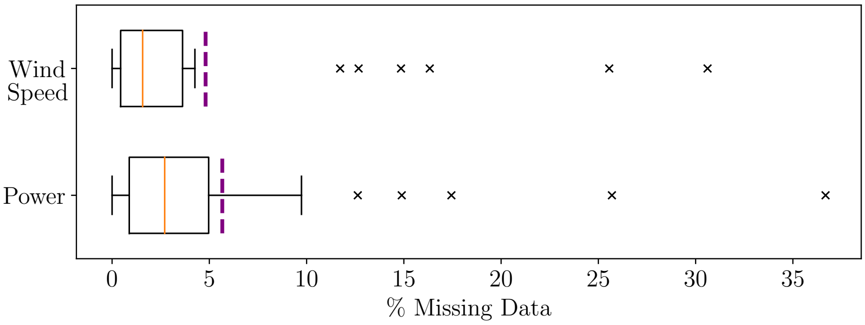


Figure 3.7: Percentages of missing data seen overall in power and wind speed variables for a group of 30 wind farms. The dashed purple line represents the mean value.

Figure 3.8 shows the distributions of lengths of missing values as well as the distributions of times between missing values. It is important to note the long tail of the distributions is omitted from the figure for clarity; however, the maximum observed missing lengths extended to 29 days for both wind speed and power. The mean length of a period of missing values is 3 hours, with a mean length of non missing periods (i.e. mean time between instances of missing data) of 48 hours. The presence of long sections of missing data strengthen the hypothesis that the mechanism causing missing data is not completely random. The distributions of missing data for power and wind speed are also similar, as the dominant cause of missing data is raw points missing which is likely to affect both measured variables.

The percentage of site forecasts where the corresponding missing indicator variable was classed as significant was found (Figure 3.9). On average across all variables, the missing indicator was significant in the forecast 60.8% of the time. This suggests it is likely the MNAR data pattern applies to both wind speed and power measurements across the majority of sites studied. Any model that ignores MNAR missing data will

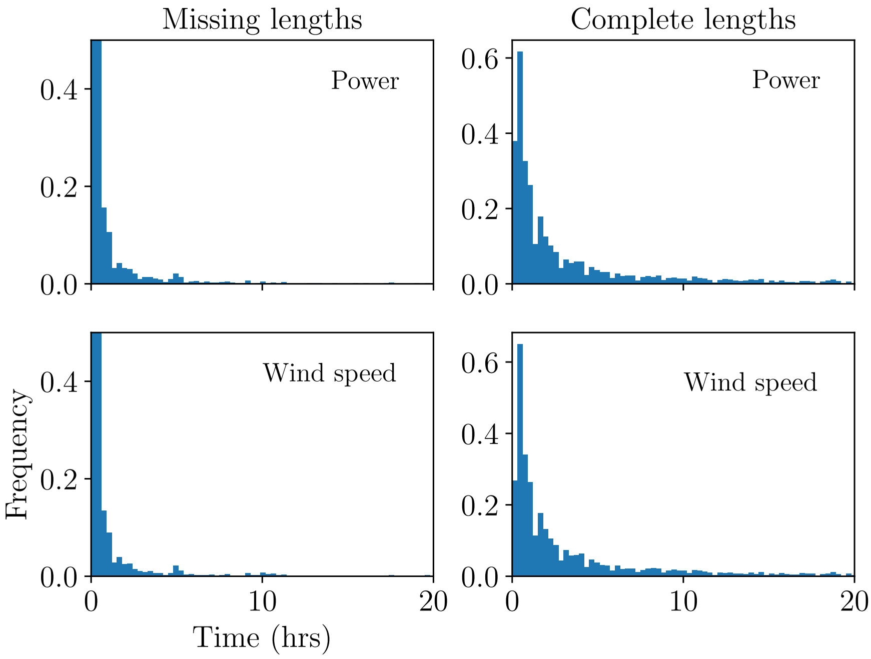


Figure 3.8: Distributions of the lengths of gaps in the data (missing lengths) and lengths between any missing entries (complete lengths), for power (ap) and wind speed (ws).

under-represent behaviour seen under missing data scenarios in the training dataset and therefore might be expected to perform worse under these conditions.

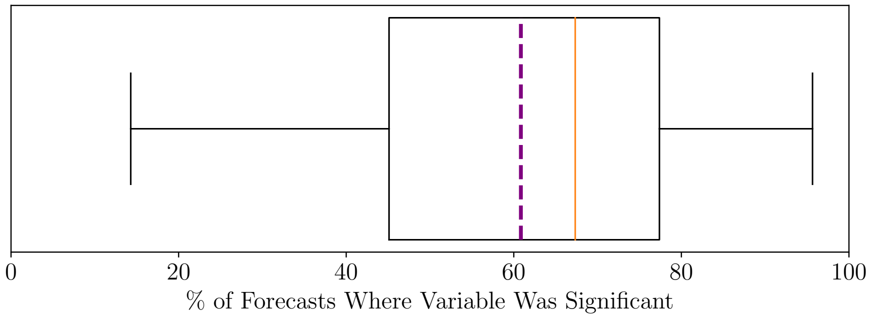


Figure 3.9: Percentage of forecasts where the original variable is classed as MNAR by the missing indicator test. Dashed purple line represents the mean.

In order to test the effect of missing data at different points relative to the ideal case of complete data, a case study covering a range of missing data scenarios is conducted. A VAR forecasting model was used to test the effect of different types of missing data, with errors evaluated through 5-fold cross-validation. Input variables comprised 60 lags (representing 30 hours) of both power and wind speed, two-hourly mean and standard deviation measures, and monthly and diurnal dummy variables to account for seasonality and day/night variations. LASSO regularisation was used to perform feature selection as a precursor step to the model fitting [185], with the top 100 features kept as inputs for the final model. The optimum regularisation parameter value for each fold combination is found through nested cross validation prior to forecast training and testing, using the top 100 most significant features. Forecasts are evaluated using Normalised Mean Absolute Error (NMAE), where MAE is divided by site capacity in order to compare sites of different sizes. For all the missing data cases, a 2.5-hour ahead horizon is used as statistical forecasts tend to outperform both persistence and numerical weather model based methods on this time scale, and thus this is a typical horizon for use of this type of model. An example application is the UK BM where gate closure (after which forecasts may not be modified) is 1 hour before the given half-hour settlement period; forecast horizons used are generally 1.5 to 3 hours ahead to include time to send and receive the forecasts.

As a benchmark to compare worsening in performance due to missing data, the forecast model with no missing inputs was evaluated and compared to a simple persistence model. The VAR model outperforms persistence at every site for all forecast horizons beyond 30 minutes ahead, with improved relative performance at longer forecast horizons as displayed in Figure 3.10).

3.4.1 Forecasting process assumptions

A preliminary step to select out the most informative 100 features is employed, which allows faster calculation of the final step of model training and testing with a set value of regularisation parameter. However, re-running only the final step of model training and testing when different inputs are missing requires two assumptions:

1. Missing a small number of the top 100 features makes negligible difference to the final forecast error

2. The optimum regularisation parameter is the same.

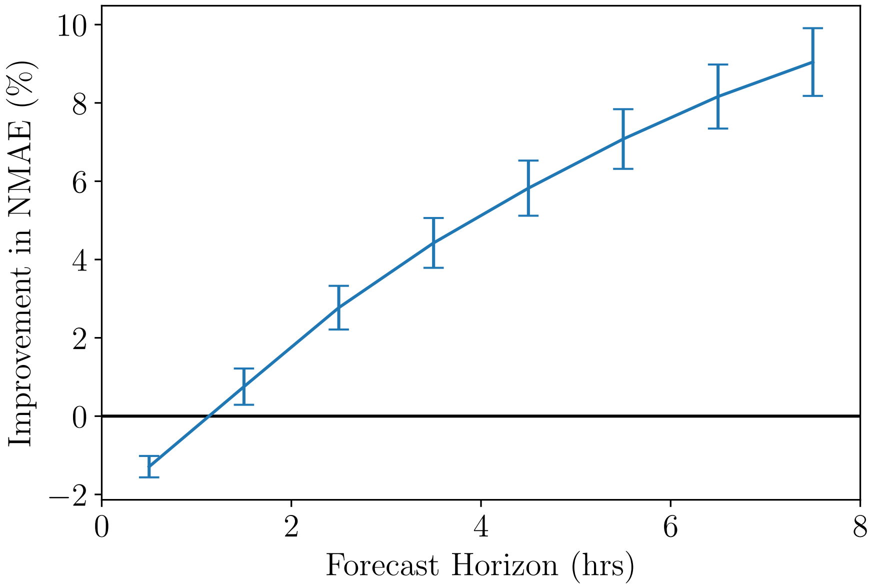


Figure 3.10: Model error relative to the persistence model, at varying forecast horizons, when no data is missing. Positive Δ NMAE represents model error lower than persistence error. Error bars represent the variation over different sites.

Both these assumptions must be tested in the presence of missing data, and also for producing forecasts at different horizons. Given the high computation time (on the order of hours) needed to re-run all steps of feature selection for the final model, a small increase in error for a drastically faster model would still be preferred. When the complete forecasting process was run with missing data present, an improvement in error of 0.02% at the missing site was seen compared to re-running only the final step using forecast inputs determined by the complete case. An even smaller difference of 0.002% was seen at the other sites with no direct missing inputs. For the case with no missing data, the standard deviation of NMAE of the forecast errors found through bootstrap resampling is 0.026%. Given the error differences between re-running the whole forecasting process and just the last step are smaller than the standard deviation of the forecast errors, we can conclude there is no significant difference in forecast performance between these two approaches. As such, retraining only the final model without repeating the feature selection process is justified in the presence of missing data. It is also assumed the optimal set of forecast inputs is independent of forecast horizon. It was found that forecasts with a longer horizon display very little difference in error (0.001%) between a model trained specifically for that horizon and one using the features and regularisation parameter optimised for a 1-step-ahead forecast. Both scenarios result in a change in the optimum regularisation parameter of more than 20% but in both cases this clearly has a negligible impact on final forecast error. The coefficients affected most by regularisation are those with small magnitude, which by definition will also be the ones with the smallest impact on the final forecast.

3.4.2 Case study results

Missing inputs for forecast generation

In the first case tested, all forecast inputs for one or several sites are missing, seen for example during a long communications failure with a site. Forecast performance at a missing site clearly benefits from the inclusion of other sites with complete inputs in the forecasting model as seen in Figure 3.11. Increasing the number of sites in the model may increase the probability of a complete site similar to the missing site, therefore improving forecast errors. Explicit testing of the impact of distance between sites and the impact this has on forecast skill was not investigated, although this would be an interesting extension to this work. In particular, showing the added value gained for forecast accuracy by including a nearby site as a forecast input might encourage participation in data sharing and data market mechanisms. Filling missing forecast inputs using the complete forecast inputs (the ‘impute’ method) shows slightly worse performance than retraining a model without them, as the extra step of predicting the missing values adds to the uncertainty in the final forecast.

Case 2 examines the effect of the length of a missing period at a single site and is plotted in Figure 3.12; as may be expected, removing the most recent lags makes the greatest difference to forecast error as these lags tend to carry the highest weight in the regression model (indicating they are the best predictors). Forecasts continue to worsen with increasing length of missing period, but the largest proportion of the loss of forecast skill comes from missing the most recent information.

Case 3 evaluates the impact of a single piece of information loss. Figure 3.13 shows a 2% increase in forecast error at the missing site when the most recent piece of

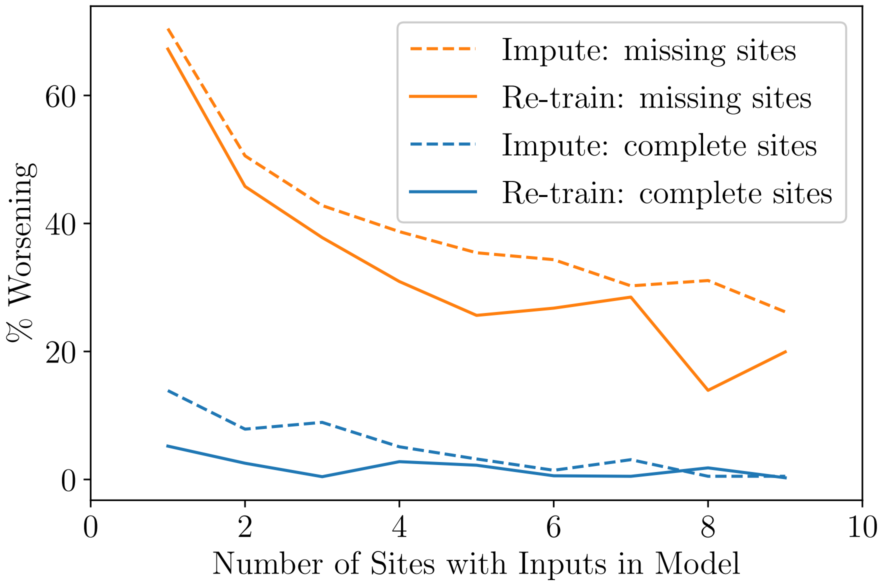


Figure 3.11: Case 1: Impact of number of sites included in VAR model on forecast error. Forecast error improves when more sites are included in the model, particularly for sites that are missing forecast input data.

information is missing, but other lags missing have negligible impact. The increase in forecast error of approximately 3% that is seen at the other sites with no data missing may be due to time delays between weather conditions at different sites. Imputing missing forecast inputs made no significant difference to the forecast skill.

In Case 4, the effect of data missing simultaneously across sites is shown by evaluating forecast performance with lag 1 missing at a number of sites. As expected, an increased number of sites with the most recent information missing results in a worsening of forecast performance across all sites, but notably more so at missing sites, shown in Figure 3.14.

Missing values in training data

The effect of data missing in the training set used to fit the forecasting model is shown in Figure 3.15. Forecast performance is evaluated for complete case analysis, inverse probability weightings, mean imputation and multiple imputation. A greater proportion of missing data significantly reduces the number of complete rows remaining in

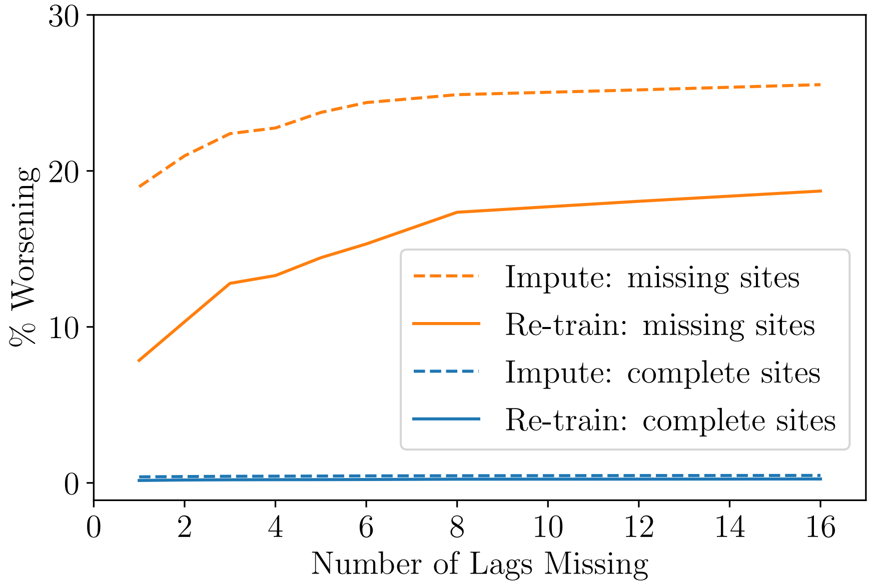


Figure 3.12: Case 2: Impact of cumulative number of lags missing from one site on forecast error (normalised to site capacity). Forecasts at missing sites are worse for longer missing periods.

the training dataset (Table 3.2), decreasing the accuracy of the model fit. This disproportionately affects the methods where incomplete rows are not used for model fitting, leading to much worse performance of the complete case and inverse probability weights methods at higher missing data levels. At 11.65% missing data, forecasts using a complete case approach are 19% worse than when the training dataset is complete. Correcting the bias of the complete case approach through inverse probability weightings improved forecasts at missing data levels of 9% or more, although care must be taken to choose a suitable number of components in the principal components analysis used for this. Mean imputation benefits from use of all available information, leading to a worsening in foreecast NMAE of only 1.27% when 11.65% of data is missing.

Multiple imputation, where a complete dataset is recreated before model estimation, shows the best performance with worsening of only 0.72% at a missing data proportion of 11.65%. The benefit of multiple imputation is particularly pronounced at high levels of missing data, although even more modest improvements in forecast skill for lower missing data levels can still be worthwhile. Multiple imputation also results in more

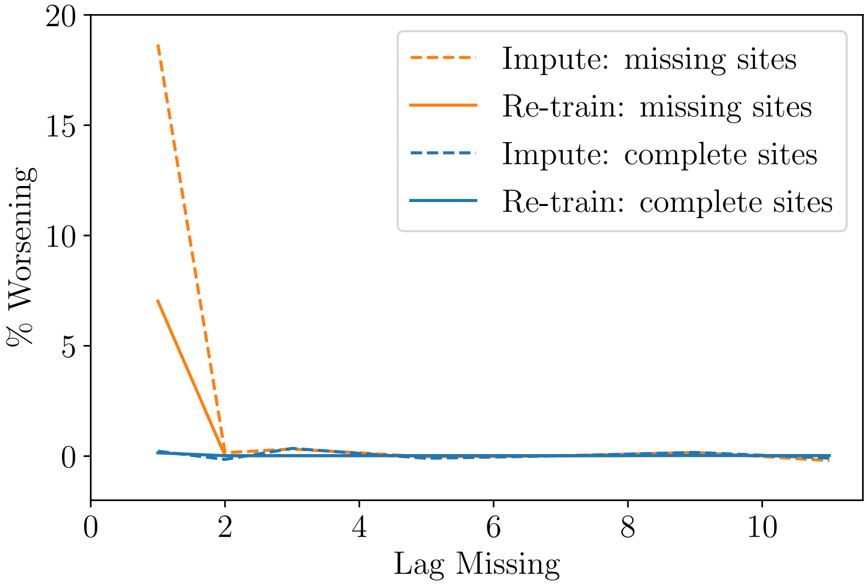


Figure 3.13: Case 3: Impact of an individual lag missing from one site on forecast error (normalised to site capacity).

consistent forecast errors across the set of sites analysed; 80% of sites displayed the lowest forecast errors when multiple imputation was used, and the worsening for multiple imputation at the remaining two was less than 0.3% greater than the worsening for mean imputation.

3.5 Case study: creating artificial historic datasets

While most sites are expected to provide on-site measurements on a live basis in order for a forecast to be generated, in some cases this information is not available. If a site is newly constructed or has recently changed ownership, the site owner may not have access to historic SCADA measurements. Smaller sites may also lack the capability to, or choose not to, share live data feeds with an external forecast vendor. In these cases, other alternative information sources are available. Numerical Weather Prediction (NWP)-based reanalysis models estimate and then propagate forward the atmospheric state through a gridded physics-based atmospheric model. For sites that are not newly built and that are large enough to participate in the UK BM, another data

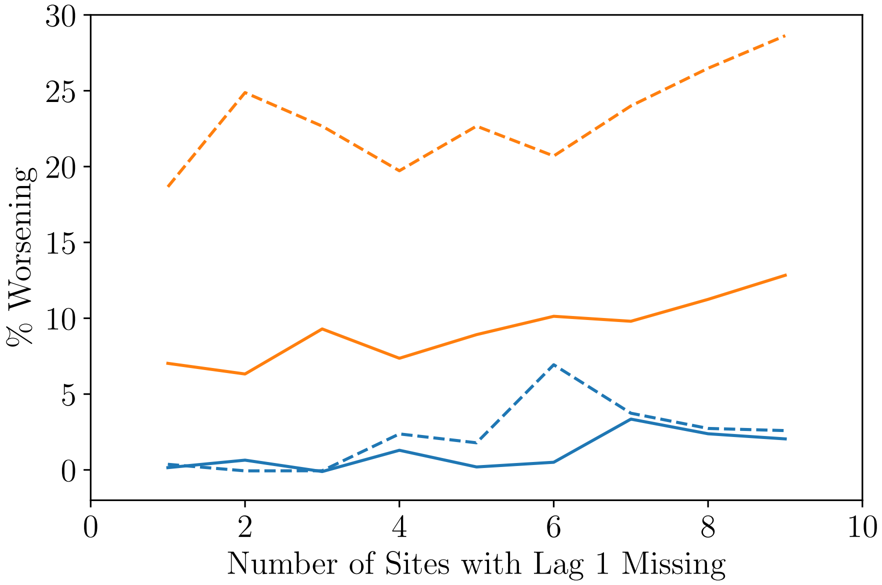


Figure 3.14: Case 4: Impact on forecast error of number of sites with lag 1 missing. The more sites missing data simultaneously, the worse the error, and the sites with data missing are most affected. Legend is the same as for Figures 1.11–1.13.

source is half-hourly resolution metered power data that is publicly available through Elexon's BM reporting site (bmreports.com/bmrs/). This is expected to be more accurate than power values generated from renalysis data as it is a direct measurement from the site, rather than a post-processed output from a coarse simulation model. However, this only provides power and not wind speed values. A case study is conducted on nine UK sites that participate in the BM, to determine the practicality of either reanalysis or BM data in supplementing historic site data for model training (case 6 in Table 3.1). Forecasts were evaluated for 1, 3 and 5 sites having no, 3 or 6 months' worth of SCADA. Where available, it was assumed the SCADA was for the time period directly before the forecast evaluation period. For the time points where 'no SCADA' was simulated, all values in the forecast training and testing inputs as well as the training output for that site were replaced with the 'artificial' (reanalysis based or BM based) values. Only the SCADA values were preserved in Y_{test} so that all the data supplied to the forecast model uses the artificial data, but the forecasts themselves are evaluated against the 'real' values.

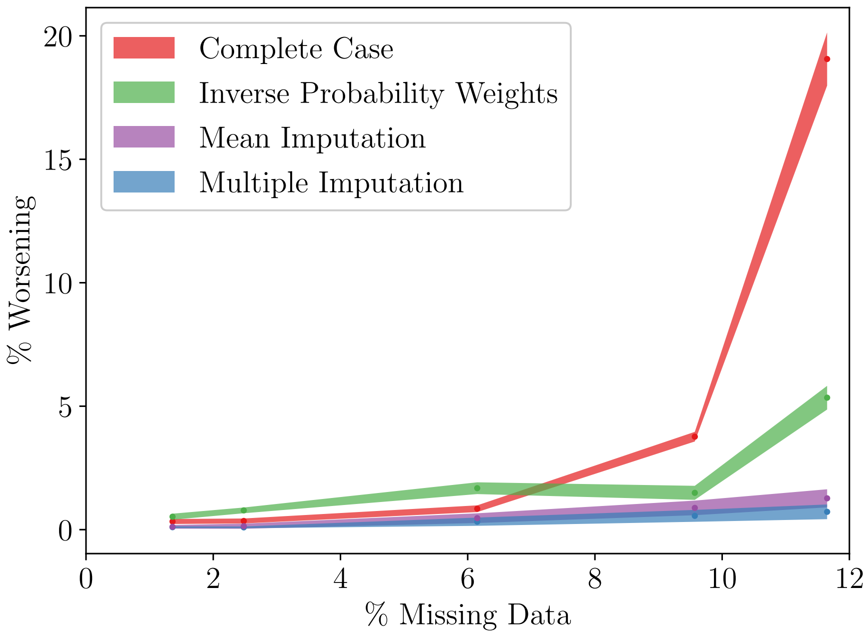


Figure 3.15: Case 5: Missing data in the training dataset.

Examples of global, publicly available reanalysis datasets include MERRA-2 from NASA and ERA5 from the European Centre for Medium-range Weather Forecasts (ECMWF). Analysis by Olauson [186] compares the accuracy of these datasets for both country-wide and individual turbine's wind power generation. They found the ERA5 data displayed higher correlations with real measurements, with 20% lower errors than MERRA-2. ERA5 provides hourly wind speed and direction on a 0.25° × 0.25° grid (roughly 28 × 15 km in Scotland), which is twice the resolution provided by MERRA-5. The ERA5 reanalysis dataset is available to download from the Copernicus climate data store [1]. Reanalysis datasets require several steps of post-processing to get from wind speeds at fixed heights and grid points, to wind speed and power at the site location and at hub height. First, bi-linear interpolation was used to obtain a wind speed value at the site co-ordinate from the four closest grid points. Wind speeds are available at both 10 and 100m, allowing fitting of a logarithmic wind profile to scale the wind speeds to hub height. The log law relates mean wind speed $\bar{u}(z)$ and height z [187]:

[1] cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land

$$\bar{u}(z) = \frac{u^*}{\kappa} \left[\ln\left(\frac{z}{z_0}\right) + \Psi \right] \tag{3.8}$$

where the component Ψ depends on atmospheric stability, and can be neglected under an assumption of neutral wind conditions. z_0 is the roughness length, u^* is the friction velocity and κ is the von Karman constant. In a plot of $\ln(z)$ against $\bar{u}(z)$, u^*/κ is given by the gradient and the intercept denotes $-u^*/\kappa \ \ln(z_0)$. Once u^*/κ and z_0 are known, Equation 3.8 may be used to find the wind speed at hub height. A bias correction may be applied to wind speeds by a fit of reanalysis wind speed versus SCADA wind speed values [188]. While this is designed to remove any biases in the reanalysis model and go some way to accounting for site-specific properties like terrain, this is only possible in practice if SCADA is available to make this comparison. For this case study, no bias correction is included due to the assumption that this process would only be used for sites with very little or no SCADA data available. Finally, wind speeds are passed through a power curve to produce power as well as wind speed values for forecast inputs. Manufacturer power curves are smoothed to better represent whole-farm output through application of a Gaussian kernel with standard deviation of 1.5ms^{-1}, as suggested by Olauson [186].

The dataset used for this case study is labelled 'dataset C' and comprises 9 sites with a 19-month history available for training in a 5-fold cross-validation framework. These sites were chosen as they all have SCADA datasets spanning the same 19-month time period and also participate in the BM. Table 3.4 shows the forecast errors for all simulations, comparing the substitution of missing history with ERA5 and BM data, as well as an option labelled 'drop'. This alternative approach doesn't include any forecast input from the site(s) with limited historic data, instead training a model with only inputs from other sites with the output forecast(s) for the missing site(s) trained on BM data. The general relationship between length of training dataset and forecast errors is displayed in Figure 3.17.

For sites where missing historic data is simulated, the forecast error is compared with the case with complete historic data and worsening is calculated at each site before the averaging across all missing sites. For simulations where more than one site is

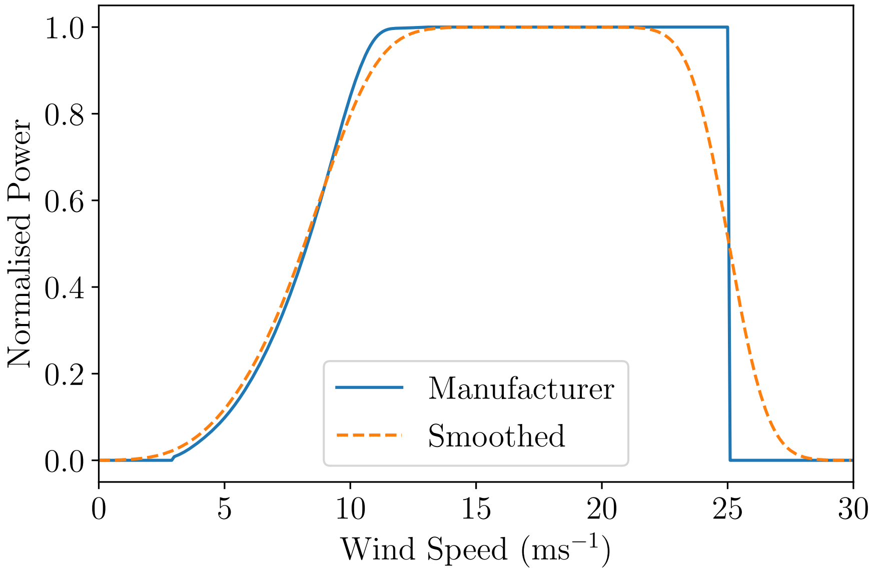


Figure 3.16: Manufacturer power curve and curve smoothed with a Gaussian kernel. Standard deviation of the Gaussian kernel was taken to be 1.5 ms^{-1}.

missing historic data, a random subset of sites is chosen 10 times and the results of each simulation are also averaged. The standard deviation, found through bootstrapping, of forecast errors in the case where all sites have complete history is 0.33%. As such, any difference in forecast error of less than 0.33 % is deemed not significant, and therefore recorded in brackets in Table 3.4.

In general, it can be seen that filling the training dataset with ERA5 produces the lowest errors in most forecast scenarios; however, for a low number of sites (1 or 2) with no historic data, filling with BM values is the best method. The advantage of BM data is that it comprises measurements of actual power at the wind farm; the several additional preprocessing steps needed to get power values from raw reanalysis outputs are likely to add to any inherent model error. However, ERA5 data has the advantage of providing an estimate of wind speed that is not available from BM data. When a large proportion (5 out of 9) of sites have limited historic datasets, the ‘drop’ method gives the best end forecasts. While some general patterns for the optimal way to extend training datasets by large portions are given in this case study, it is based on a small set of simplified scenarios and the optimal approach may vary with geographic location,

Table 3.4: Forecast errors when the training dataset at some site(s) has been artificially generated. The figures shown are the percentage worsening between the error in this case and the error when all sites have historic SCADA available. N_s is the number of sites with limited or no historic data, while m is the number of months of historic data available for those site(s). Values in brackets are smaller than the standard deviation of ideal forecasts.

N_s	m	ERA5	BM	drop
1	0	11.6	**10.32**	24.19
1	3	$(\simeq \mathbf{10^{-3}})$	3.46	3.79
1	6	$(\simeq \mathbf{10^{-3}})$	1.99	1.60
2	0	17.18	**14.76**	26.4
2	3	**2.37**	6.22	3.79
2	6	**1.40**	4.03	1.60
5	0	**13.67**	16.61	36.15
5	3	5.63	8.50	**3.79**
5	6	3.47	5.91	**1.60**

dataset resolution and overall available length at all sites.

3.6 Conclusions

The properties of missing data in real SCADA time series have been found, before the effect of various missing data scenarios on forecast skill were simulated through case studies. Real wind power data is shown to have typical median levels of missing data of 2.70% for the power variable and 1.57% for wind speed. However, some sites may display levels up to 36%, greatly reducing forecast skill. Data is Missing Not at Random, meaning care must be taken to use an appropriate missing data technique. The impact of missing data on wind power forecasts in an autoregressive framework has been demonstrated, with the most appropriate mitigation methods identified. The key results are summarised:

- Missing training data can have a significant impact on results if not dealt with appropriately; multiple imputation is found to be the best of the methods considered here to compensate for this
- If inputs to an operational forecast model are missing, retraining the model with-

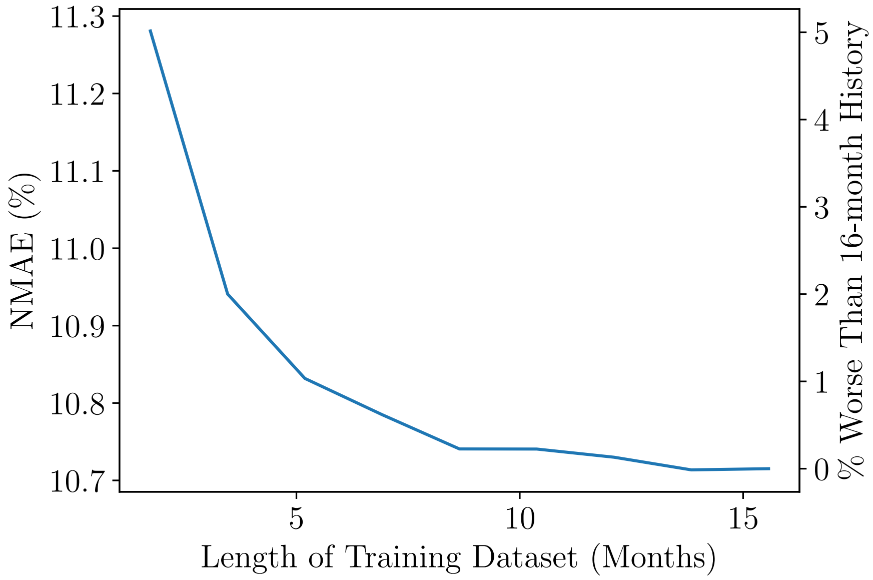


Figure 3.17: Variation in forecast error with the length of available training dataset

out these inputs results in better performance than filling the missing values using a regression model based on available inputs

- Forecast error improves across all sites when more sites are included in the model, with particular improvement at sites that are missing forecast input data; therefore, spatio-temporal models including a greater number of sites are generally more robust to missing data
- Forecasts continue to worsen with increasing length of missing period, but the largest proportion of the loss of forecast skill comes from missing the most recent information
- When a subset of sites have a short historic dataset available, ERA5 reanalysis data scaled to site location and hub height and passed through a power curve provides a good substitute. Where applicable and only one or two sites have short datasets, Balancing Mechanism data may be used.

While these results are from case studies using a VAR forecasting model, future work could extend this to other models. It is expected the results would be similar, as the change in forecast skill is likely related more to the loss of information from the missing variable(s) than the modelling framework itself. In summary, awareness of the properties of missing data, its potential impact on model performance and use of suitable mitigation techniques is essential to realise that model's full potential.

Chapter 4

Forecast combination with adaptation for improved ramp forecasts

Abstract

Ramps in power are some of the most important times for decision making, where severe ramps may require actions such as curtailment of turbines to limit a fast ramp up, or utilisation of energy storage or even load curtailment to balance a severe downward ramp [189]. Meanwhile, forecast skill is often lower around power ramps due to timing and amplitude errors, reducing ability to manage the power system optimally. Previous research [190,191] has developed specialised ramp forecasts that predict the probability of ramp events and their characteristics, but this is often separate from a power forecast. It has also been suggested that combined forecasts may serve to smooth out power fluctuations such as ramps [189].

This chapter presents a forecast combination approach with the aim of improved performance around times of ramps. Ramps are explicitly modelled through forecasts of ramp rate, which are passed to the forecast combination step along with the individual model power forecasts. Day-ahead forecasts based on the outputs of physical weather models are able to predict the presence of a ramp ahead of its arrival at a site but may

include magnitude or timing errors. In contrast, very short-term (up to 6 hours ahead) models based on site power measurements are sensitive to recent changes in observed conditions but do not have ‘awareness’ of incoming weather fronts until they arrive. The resulting forecasts are evaluated not only on overall performance but also ability to correctly forecast ramps.

4.1 Introduction

Forecast combination can be important in a number of scenarios: commonly, different models perform best at different forecast horizons so it makes sense to blend them, particularly at the crossover horizon. A statistical model which uses recent on-site measurements is likely to perform best for very short-term horizons (up to 6 hours ahead), and a Numerical Weather Prediction (NWP) model generally performs better for longer horizons, with skill up to 10 days ahead [192]. In the very short-term, conditions are most likely to equal present measurements plus some variation. NWP takes a large number of global atmospheric measurements and uses a sophisticated physics-based atmospheric model to propagate these conditions forward in time. The relatively longer computation time of NWP models means the time difference between input information and output forecasts is necessarily longer than for statistical models, reducing their skill at very short-term horizons. However, they have greater skill for longer horizons, although there are uncertainties present in both the initial conditions and the model parameters. On even longer scales, subseasonal-to-seasonal (S2S) forecasts make use of atmosphere-sea-ice-land interactions to capture longer term variations in atmospheric conditions; forecasts on these horizons are not included in combination work in this chapter but are explored in Chapter 5.. It has been found that a combination of several forecasts from different models, or where models use different information sets as inputs (as statistical and NWP models do), often outperforms a single model [115]. This does rest on the assumption that no model is the true representation of the underlying data generating process, as this single model, if known, would outperform any combination of ‘misspecified’ forecasts [116]. However, in many ‘real-life’ applications either the true process is not known or no individual forecaster or model has access to the complete information needed to generate the 'perfect' model. This is certainly true of wind power forecasts, where the final value of power output is the result of complex physical interactions to produce the wind speed seen by the turbine, as well as the performance of the individual turbine and any imposed control actions. In addition to improved forecasts in terms of a lower error metric (or sharper forecast densities while maintaining reliability), it has been suggested the error distribution of combined forecasts may be closer to Gaussian than that of the individual forecasts [193].

The simplest of forecast combination methodologies is the equally weighted linear opinion pool, ELP. This is equivalent to taking a simple average of all forecasts, so that the value from each of the forecasts contributes equally to the combined forecast value. Alternatively, the optimally weighted linear opinion pool (OLP) method uses a weighted average of the model forecasts where the combining weights w_i may be optimised for the best final forecast. The set of weights $w_i \ \forall \ i = 1, 2, ...N$ may be constrained such that all $w_i \geq 0$ and $\sum_i^N w_i = 1$ to guarantee that the resultant combined forecast falls in the same range as the member forecasts - but this may place unnecessary restrictions on the parameters. Thus, two benchmarks based on the Optimal Linear Pool (OLP) are tested: constrained and unconstrained OLP. In the literature, an intercept to correct for overall bias has also been proposed [194] and Bordley shows this is equivalent to the inclusion of a Gaussian prior in a Bayesian framework [195]. Estimation of optimal weights may be achieved through minimisation of the variance of the forecast errors [196] which takes into account forecast covariances, maximisation of a likelihood function [197] (or minimisation of a score function), or by simply weighting forecasts in a ratio given by the inverse of the individual forecast's error [198].

In practice it has been shown that ELP often outperforms an OLP forecast combination; this has been dubbed the 'forecast combination puzzle' [199,200]. This may be due to increased variance from the estimation of the weights for OLP, particularly with small sample sizes. Weights are optimised over a training sample, while these values may not be optimal over out-of-sample forecasts. Further work has aimed to classify cases where OLP should be used over ELP, finding that equal weights provide improved performance when sample size is small, the model forecasts have similar variances or there is correlation between the errors of different forecast models [201]. Equal weights are also preferred when there has been a structural change within the time series, for example a location shift, as this means the optimal weights are further from the true optimal combination. Recursively estimating the weights based on forecast performance over a window of the most recent time points addresses this problem and has been shown to outperform ELP on inflation forecasts [202].

Other pooling variations have also been proposed, including trimming (where the outlying forecasts are discarded) [203], explicit inclusion of a persistence forecast (or other lagged values) even though this information is contained in other forecasts [204], automatic selection of best combination strategy [205], or a logarithmic opinion pool [206]. The generation of different forecasts from the same model using different windows of training data has also been investigated [207], showing robustness against structural breaks.

Most of the aforementioned literature deals with point forecasts, whereas probabilistic forecast performance must be evaluated over the whole distribution. Quantile forecasts may be produced by combining point forecasts [208]; this may be useful where only point forecasts are available from individual models, but it is not clear how this method performs relative to combining full probability forecasts. Hall et al. [206] suggest a method for combining the forecast moments (mean and variance), applying these to an assumed distributional shape for the final forecast. Unless there is good reason for specifying the shape of the forecast distribution, this seems like a potentially restrictive approach. The same paper also suggests a direct Bayesian combination of the densities, allowing more flexible final distributions, although this method involves estimation of covariance matrices for both the mean and variance moments.

For quantile forecasts, the quantiles may be combined independently of each other (assuming the same probability levels have been forecast for all individual models), or as a whole distribution. For a given time point t, probability level α and individual quantile forecasts for that level $y_{t,i}^{(\alpha)}$, the OLP combination of N individual forecasts is given by

$$\hat{y}^{(\alpha)}_{t,\mathrm{comb}} = \sum_{i}^{N} w_i^{(\alpha)} \hat{y}^{(\alpha)}_{t,i} \quad . \tag{4.1}$$

This is effectively finding a weighted average of points on the Cumulative Distribution Function (CDF) with the same y-value (averaging by quantile) as seen in Figure 4.1. From here on, this will be referred to as 'OLP by quantile'.

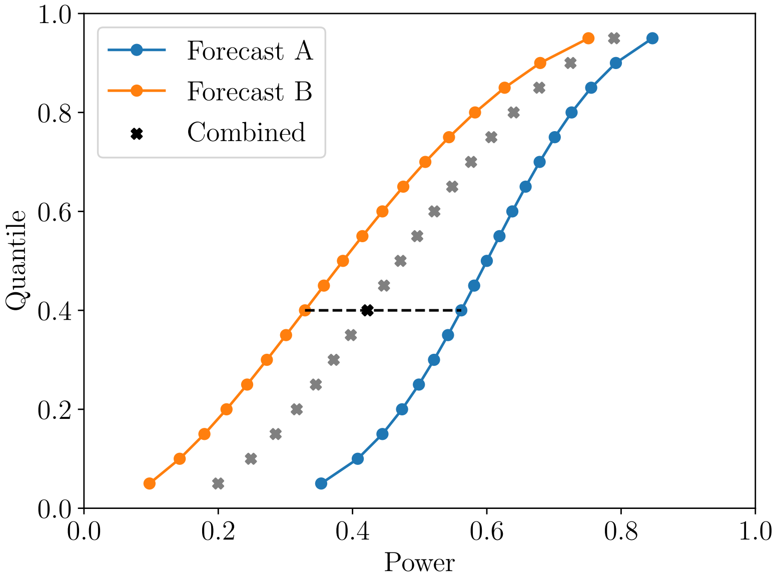


Figure 4.1: Forecast combination by probability level; the final α-quantile of the combined distribution is a combination of the α-quantiles of the individual forecasts.

Combining forecasts by CDF value has also been proposed [197]. This may be done analytically in some simple cases such as combination of Gaussian distributions, but where this is not possible or where the distribution is nonparametric and defined by quantiles it may be done using a set of points on the CDF. In this case, the combination is a weighted sum of CDF values for a set of common y values — or in the context of wind power forecasts, power levels. This is shown graphically in Figure 4.2 and is expressed mathematically as

$$F(\hat{y}_t)_{\mathrm{comb}} = \sum_{i}^{N} w_i F_i(\hat{y}_t) \tag{4.2}$$

where $F(y)$ is the Cumulative Distribution Function of the forecast distribution for a power value y.

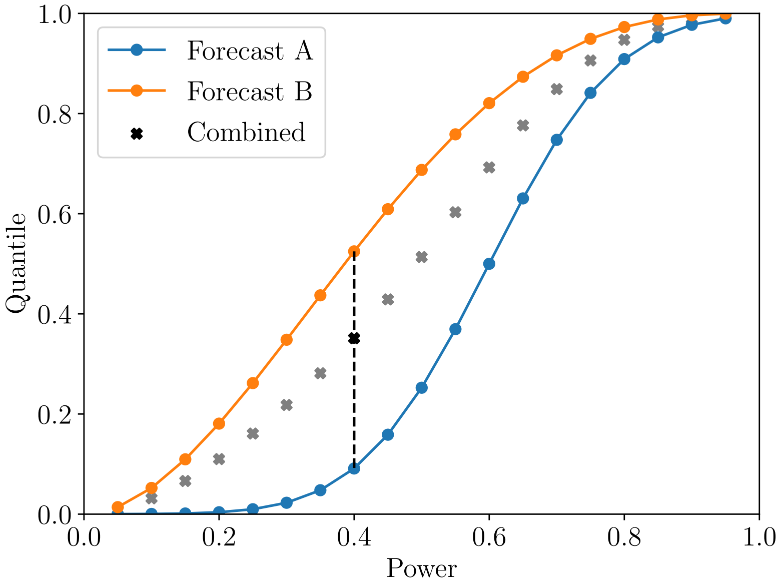


Figure 4.2: Forecast combination by power level; quantile forecasts must be interpolated to set y-values (power in this case) before combining

Ranjan's theoretical results show that any linear opinion pool will lack both sharpness and calibration [197], even when the individual model forecasts were calibrated. A beta-transformed linear opinion pool is proposed instead, allowing for a nonlinear relationship between the individual and combined forecasts and incorporating recalibration into the forecast combination process. This method has been shown to successfully combine calibrated and uncalibrated forecasts in a range of applications including wind power forecasting [197,209]. The beta transformed OLP may also be applied by quantile or by power. For OLP by power, Equation 4.2 is updated to

$$F(\hat{y}_t)_{\text{comb}} = B\left(\sum_{i}^{N} w_i F_i(\hat{y}_t)|\alpha, \beta\right) \tag{4.3}$$

where $B(x|\alpha, \beta)$ is the CDF of the beta distribution with parameters α and β.

4.2 Very short-term forecasts

In order to test forecast combination methods, a set of forecasts from individual models must first be created. The forecasts presented in the case study in Chapter 2 are used as the base forecasts in this chapter. All methods implemented in this case study produce probabilistic forecasts, allowing testing of forecast combination methods specifically for combining probability distributions.

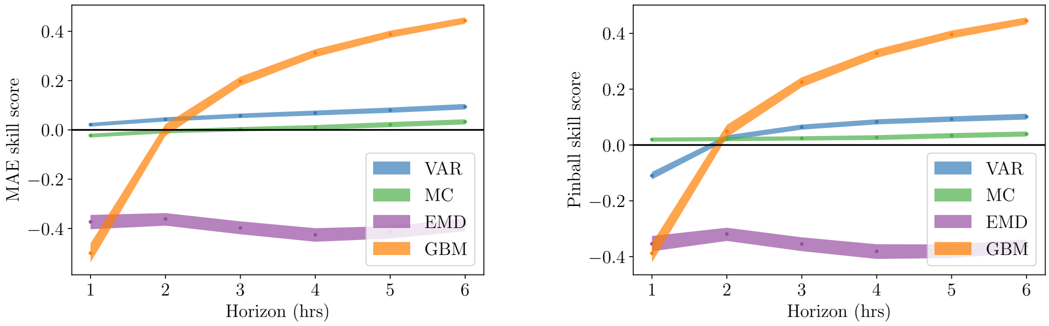
4.3 Day-ahead forecasts

Methods utilising output of an NWP weather model generally outperform very short-term methods (based only on the recent site power measurements) for horizons beyond 3 to 6 hours [42]. NWP models are able to predict major future events such as a weather front passing over a site that would not be picked up in site power data in advance. As such, they contain additional information about future weather conditions not encapsulated in very short-term forecasts and may provide benefit in a combination scheme. Day-ahead forecasts are also produced for the same zones and times as very short-term forecasts, based on the 100m wind speed and direction forecasts provided in the GEFcom2014 data [122]. Directly modelling power from wind speed requires a complex nonlinear relationship to be fitted; a Gradient Boosted Machine (GBM) is a suitable choice for this as it is able to incorporate conditional dependencies between variables as well as feature selection through the boosting procedure. Nonparametric probabilistic forecasts may be produced through the use of a quantile loss function. This approach was used very successfully by Landry [93] who won the wind track of GEFcom2014 with a GBM approach and extensive feature engineering. The day ahead forecasts produced here are produced for only one forecast horizon (24 hours ahead) and follow a very similar two stage methodology, where an initial GBM model is trained using engineered features to produce single site $\hat{y}_t^{(0.5)}$ forecasts (stage 1). Then in stage 2, inter-site dependencies are included in the final GBM forecasts by using the $\hat{y}_t^{(0.5)}$ forecasts from all other sites as additional inputs. Sparse regression trees were used to filter out the most uninformative features before both stages. Of all the features engineered, only those with a feature importance greater than 1 at key probability levels ($\alpha = 0.1, 0.5, 0.9$) were kept. If one of the sinusoidal inputs (sin/cos of either the hour of day or month of year) or one of the U or V wind components was selected, the other variable in the pair was also retained as a GBM input. Since NWP models are typically run for horizons out to at least 10–15 days ahead and a new set of forecasts is commonly generated every 6 hours, it is assumed that NWP for the leading 4 hours (28 hours ahead for a day ahead forecast) will also always be available for use as features. Table 4.1 details the inputs tested and their average feature importances across all zones. Features labelled 'energy' are produced by passing wind speed through a basic power curve.

4.4 Evaluation of individual forecasts

Skill scores were calculated for all zones, using timestamps with forecast-observation pairs available for all zones and averaging the skill score from each zone. Figure 4.3 shows the variation in skill score with forecast horizon for all individual forecast models, relative to probabilistic persistence. The decomposition (EMD) approach is significantly worse than persistence for all horizons. The day-ahead model using a Gradient Boosted Machine shows inferior performance for the very shortest horizons (less than three hours ahead) but its performance relative to the very short-term models rapidly improves with forecast horizon. The two hour ahead forecast horizon in particular has no one model that is clearly the best. While the results in Figure 4.3 are for zone 1, the same general patterns are seen across all ten zones. It is also worth noting that knowledge of curtailments at a site is essential, to avoid evaluating the forecasts on times when you 'see' a ramp in the on-site data (the beginning or end of a curtailment) that isn't due to weather conditions and therefore couldn't (and arguably shouldn't) have been forecast.

Probabilistic forecasts should be sharp, subject to reliability. This cannot be judged from a single score value such as Pinball loss, and so reliability diagrams also play an important role in probabilistic forecast evaluation. Relative Empirical frequency has been plotted, so that a perfect forecast would have a value of zero. For example, it



(a) MAE skill score against forecast horizon

(b) Pinball skill score against forecast horizon

Figure 4.3: Skill scores of case study models for zone 1, relative to probabilistic persistence model. The 95% interval of bootstrap samples is shown. Positive values indicate improvement over persistence. While only a single 24-hour ahead forecast is produced for the GBM model, the skill score varies with horizon as this is calculated relative to persistence at each forecast horizon.

would be expected that in a perfect forecast distribution, the observed power would be less than the q20 quantile forecast 20% of the time and the difference between expected and observed frequencies (the relative empirical value) would be zero. Figure 4.4 shows the reliability across the q5–q95 quantiles for the case study models. At the shorter forecast horizons, Vector Autoregressive (VAR) forecasts and at 1 hour ahead Empirical Mode Decomposition (EMD) display the s-shaped curve associated with too broad a forecast distribution, while the Markov chain and EMD forecasts show bias (under and over-forecasting respectively). While the GBM forecasts are closest to being calibrated, the confidence intervals derived from bootstrap resampling show the deviations from ‘perfect’ reliability are significant for all models.

4.5 Combination benchmarks

Four variations of common models discussed in the introduction are used as combination benchmarks: constrained OLP, unconstrained OLP, beta-transformed OLP by quantile and beta-transformed OLP by power.

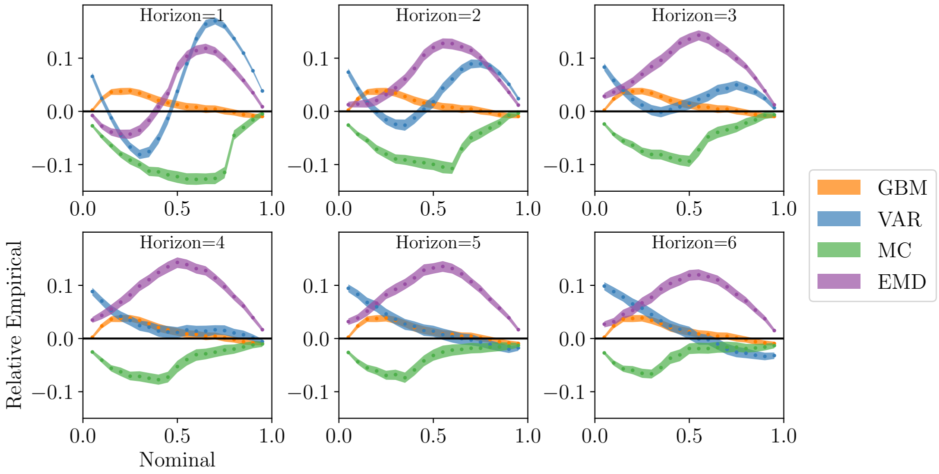


Figure 4.4: Relative reliability for the individual model forecasts at zone 1. A relative empirical frequency of zero represents ideal reliability. Intervals show the 95% bootstrapped interval.

Parameter estimation

Each of these methods based on a linear pool involves a set of parameters (the weights w_i and, in the case of a beta transformation, the shape parameters α and β) that must be optimised. Ranjan [197] gives the likelihood function over which to maximise when combining by power (as in Equation 4.2) — but both the CDF and Probability Density Function (PDF) must be known for this, and thus is suited better to parametric forecasts. To follow this method for quantile forecasts, an estimate of the PDF would have to be found from the quantile points of the CDF, so instead parameters are estimated by directly minimising the Pinball score. For combination by power level, the function `contCDF` from the R package `ProbCast` [210] was used to interpolate CDF values, with exponential tails fitted beyond the lowest and highest quantiles [211].

4.6 Novel combination method for improvement of forecasts around ramps

A combination approach that allows for inclusion of information about upcoming ramps, as well as combination of single-model forecasts is proposed. In this chapter, a set of distinct very short-term models is supplemented with a day-ahead model to gain the benefits of diverse forecast inputs. This method extends previous research in the following ways:

- Using probabilistic, rather than deterministic, forecasts as inputs to the combination and ramp adjustment stage, with feature selection to determine the most informative quantiles
- Working in a quantile regression framework to produce calibrated quantile forecasts without the restriction of a parametric distribution: this is done using a GBM with quantile loss function, also allowing interaction between inputs
- Including forecasts of ramp rate that are conditional on the power level and also capture inter-site dependencies
- Comprehensive evaluation of both general model skill and performance around ramp events, distinguishing also between upwards and downwards ramps as these events lead to different general actions within the power system
- Benchmarking against other forecast combination approaches where forecasts are not adapted for ramp events

4.6.1 Ramp rate forecasts

Ramp rates were forecasted and provided as an additional feature to the novel forecast combination method. The ramp rate r, which is the gradient of the power time series p_t, was calculated for every time point as $r_t = p_t - p_{t-1}$. A rolling average was then applied to produce a smooth time series of ramp rate where sudden spikes in gradient are averaged out while changes in gradient that are consistent across neighbouring times (i.e. a ramp event) are retained. Both 5-hour and 10-hour rolling averages were calculated and then forecast to be used as forecast combination inputs.

Ramp forecasts were produced via a Generalised Additive Model (GAM) [212], allowing for nonlinear smooth relationships between inputs and the target variable. In general GAMs can include linear functions and both univariate and bivariate smooth functions of inputs. Here all inputs were bivariate smooths, allowing each ramp rate feature to be conditional on the power level at that time. For J input ramp rate features $x_1^r, x_2^r, ...x_J^r$ and their corresponding power values $x_1^p, x_2^p, ...x_J^p$,

$$y = \sum_{j}^{J} f_j(x_j^r, x_j^p) \tag{4.4}$$

where all f_j are smooth functions, in this case cubic splines. Splines are made up of a number of basis functions which, when summed with different weights on each basis function, are able to produce a wide range of shapes of smooth functions. For a bivariate spline composed of K basis functions b_k,

$$f(x^r, x^p) = \sum_{k=1}^{K} \beta_k b_k(x^r, x^p) \ . \tag{4.5}$$

Each input was a lag of ramp rate coupled with the corresponding power value for that lag, resulting in a ramp rate forecast conditional on power level. Lags up to 12 hours for the forecast site plus lag 1 from all nine other sites (coupled with the power level of lag 1 at that site) were included, making a total of 21 splines each made up of several basis functions. The package `glmnet` was used to fit a linear model with Least Absolute Shrinkage and Selection Operator (LASSO) to the basis functions, allowing for regularisation over the supplied inputs. Optimal regularisation penalty was found via cross validation on the training set, before ramp rate forecasts were generated for the test set used in the forecast combination stage.

4.6.2 Forecast combination

A GBM was chosen as the base method for combining quantile forecasts and ramp rate forecasts from the individual models. As the inputs themselves are now quantile forecasts, the input features that will be most important for each final quantile forecast are expected to be quite different across the probability levels. As such, preliminary feature selection and model parameter optimisation is conducted independently for each probability level. The `lightgbm` [213] implementation of GBM is used for efficiency in fitting all the separate models needed. The out of sample test set quantile forecasts

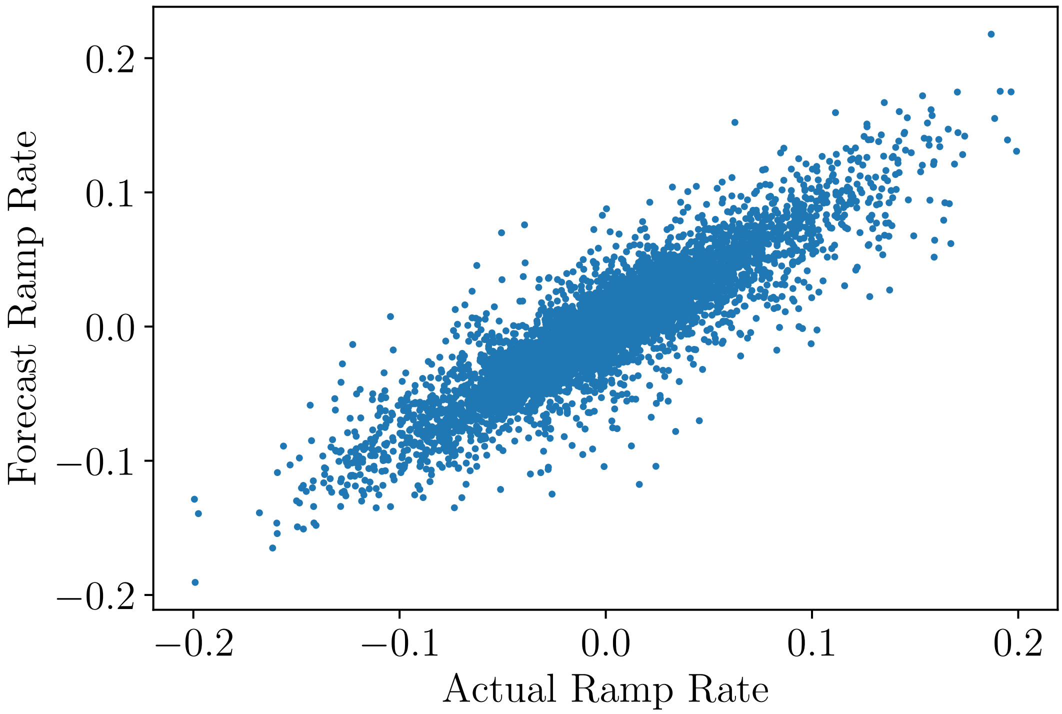


Figure 4.5: Forecast vs actual ramp rates for 1 hour ahead forecasts at zone 1.

from each individual model were supplied as features and GBM parameters (maximum tree depth, maximum number of leaves) were optimised using a grid search and nested cross validation over these time points to maximise the number of out of sample final combined forecasts available for evaluation. The number of iterations was also checked to make sure it was in line with the minimum point of training error, avoiding both under and overfitting. Sparse decision trees were used to determine the most important features in the same way as for the day ahead forecasts; here, the 30 most important features were retained and passed as inputs to the combination model. GBM combinations both with and without the ramp features were produced to evaluate the benefit from the inclusion of ramp rate forecasts. The most recent observed ramp rate was also included in the ramp rate combination model as well as forecasts for the 5- and 10-hour rolling mean ramp rates.

4.7 Results and discussion

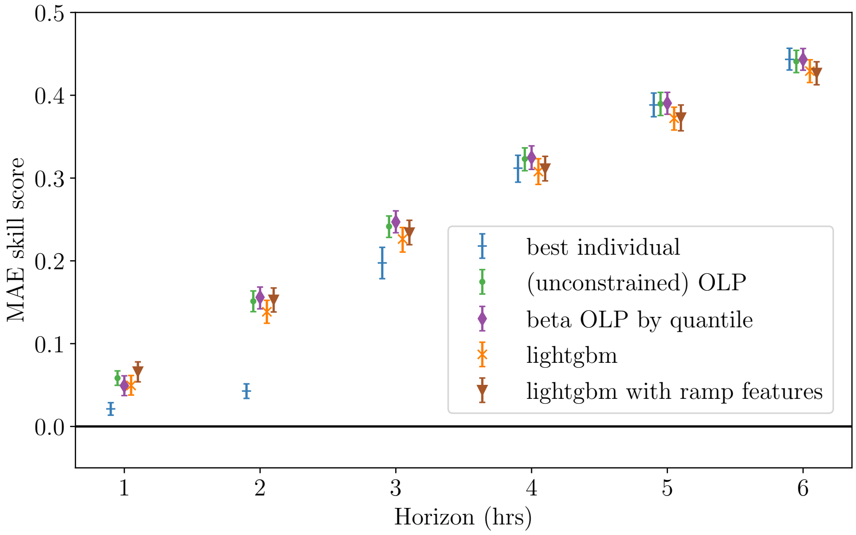
The final combined forecasts were evaluated for all the points in the testing set and again Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Pinball loss are evaluated as well as reliability on its own. The best individual model at each site and each horizon is also included for comparison. The plots shown in this section are all for zone 1, but plots for the other zones are displayed in Appendix C.

Firstly, overall forecast performance across all time points is evaluated. Figure 4.6 shows model skill scores relative to persistence for each forecast horizon. Only MAE and Pinball skill scores are shown for brevity, but the results for RMSE are almost identical to those for MAE. Compared to the best individual model, all the combined models give greatest improvement at the two hour ahead horizon, which is the horizon at which no single model was significantly more skilful than any other. It is difficult to tell from this plot alone whether any of the combination models are significantly better than others. To evaluate this more clearly, a matrix of skill score where each combination model is directly compared to each other one is plotted in Figure 4.7. A Diebold-Mariano test [116] for each pair of models is also conducted and the significance level of the result indicated on the same plot.

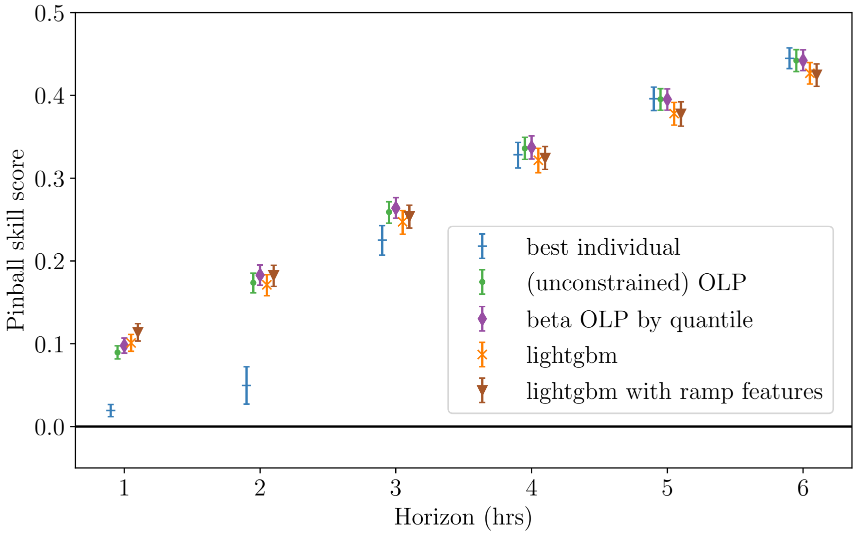
For the shortest forecast horizons (up to 3 hours ahead), the lightgbm with ramps model is generally the best performing model. This is particularly true at the one hour ahead horizon, where it is significantly better than all other models at the majority of sites. However, the difference in performance between the lightgbm-ramps model and the other combination schemes lessens and becomes less significant with increasing forecast horizon. All the forecast combination methods significantly outperform the best individual model for the shortest forecast horizons, showing benefit from the pooling of different forecasting methods. For the longer horizons (4 to 6 hours ahead), the relative performance of each combination scheme is more variable across the different locations and is much less likely to be statistically significant than for shorter horizons. The gbm-based combination methods (both with and without ramps) are significantly worse at 6 of the 10 zones for the longer horizons and they are not significantly better than any other methods at the remaining four locations. Given the variability of wind, the most recent ramp rate is much less valuable for longer horizons and the skill of ramp rate forecasts also decreases with increasing forecast horizon. It may also be that the strength of complex relationships between features and output power is so much weaker at longer horizons that a simpler method is more favourable. There are no real significant differences between the three OLP based combination methods for longer horizons either. None of the combination methods are consistently better than the best individual model, but neither are the OLP methods worse than the best individual model.

Choosing the forecast with the best Pinball score is not necessarily the best option if that forecast is not reliable; it is important to check reliability by itself as well. The general patterns seen across all horizons at each zone tend to be similar for all combination models. However, the beta transformed OLP methods tend to give the best reliability across the whole forecast distribution, especially for longer forecast horizons (3 hours ahead and longer). In fact, reliability seems to improve with forecast horizon. This may be related to the fact that the day-ahead GBM model displayed the best reliability of all the individual models, and is the best performing model at longer horizons (therefore recieving greater weight than other models at these horizons). While the GBM based combination methods (lightgbm and lightgbm-ramps) tend not to be as well calibrated overall, they show better reliability relative to the other models at the 1 hour ahead forecast horizon. This matches with the skill score results. All combination methods show better reliability than the best individual model at 6 of the 10 zones, while there is almost always at least one combination model that outperforms the best individual model at the remaining zones too. Confidence intervals are omitted for clarity.

So far forecast evaluation has focused solely on overall average performance across all timestamps; while it is important that any forecast model has good general skill, there are also times where having forecast skill may be especially valuable. One example of this is times of ramps in power, where having foresight allows better electricity system planning for grid operators. Forecast of ramps may also have implications for electricity

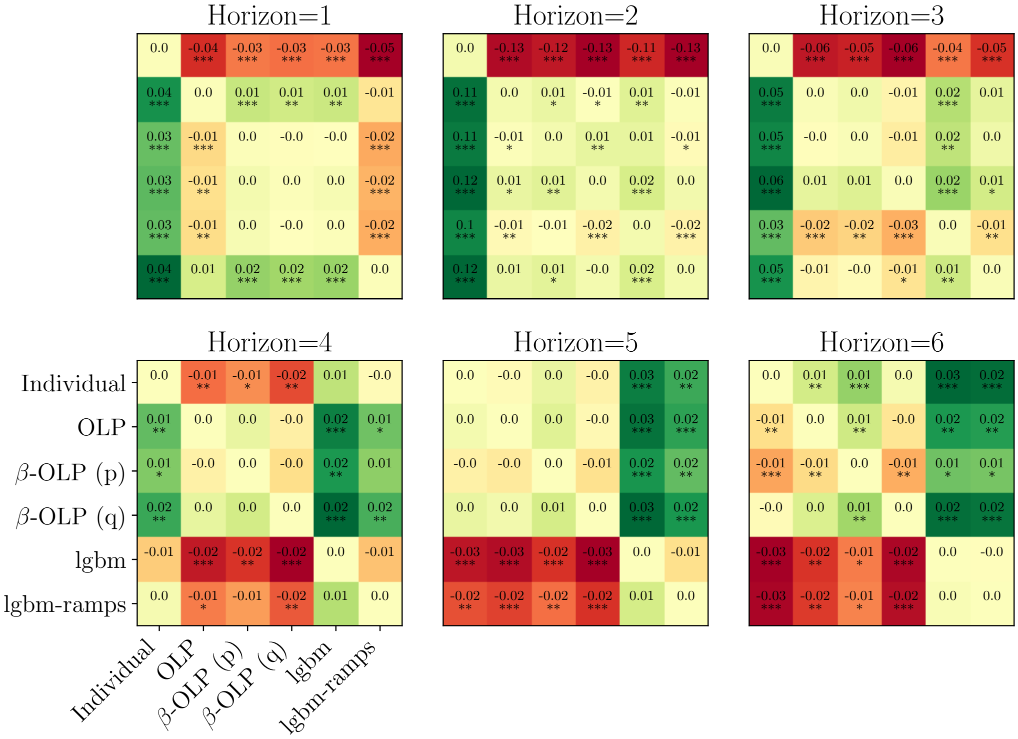


(a) MAE skill score against forecast horizon

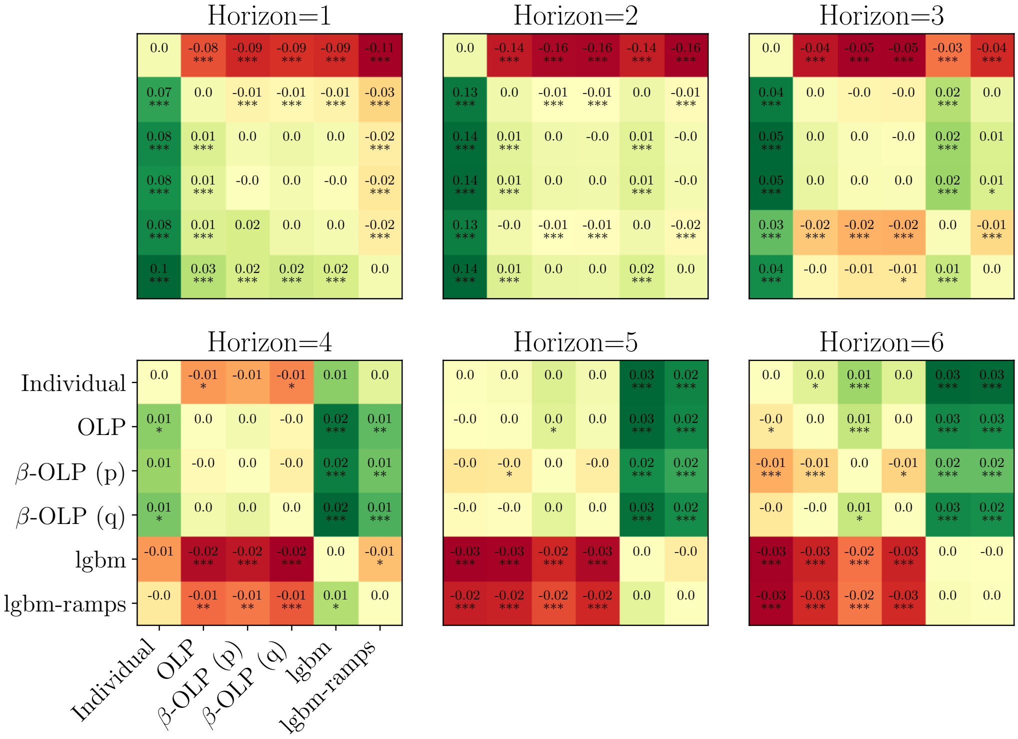


(b) Pinball skill score against forecast horizon

Figure 4.6: Skill scores of combination models at zone 1, relative to probabilistic persistence model. The 95% interval of bootstrap samples is shown. Positive values indicate improvement over persistence. Constrained OLP is omitted as it is outperformed by unconstrained OLP for every zone and horizon.



(a) Matrix of MAE skill scores. *(Figure continues onto next page)*



(b) Matrix of Pinball skill scores

Figure 4.7: Matrices of skill scores for each pair of forecast combination methods. A positive value indices the model on the y-axis outperforms that on the x-axis. β-OLP (p) denotes the beta-transformed OLP by power, and (q) by quantile. Asterisks denote significance level of Diebold-Mariano test values: *=0.05; **=0.01; ***=0.001.

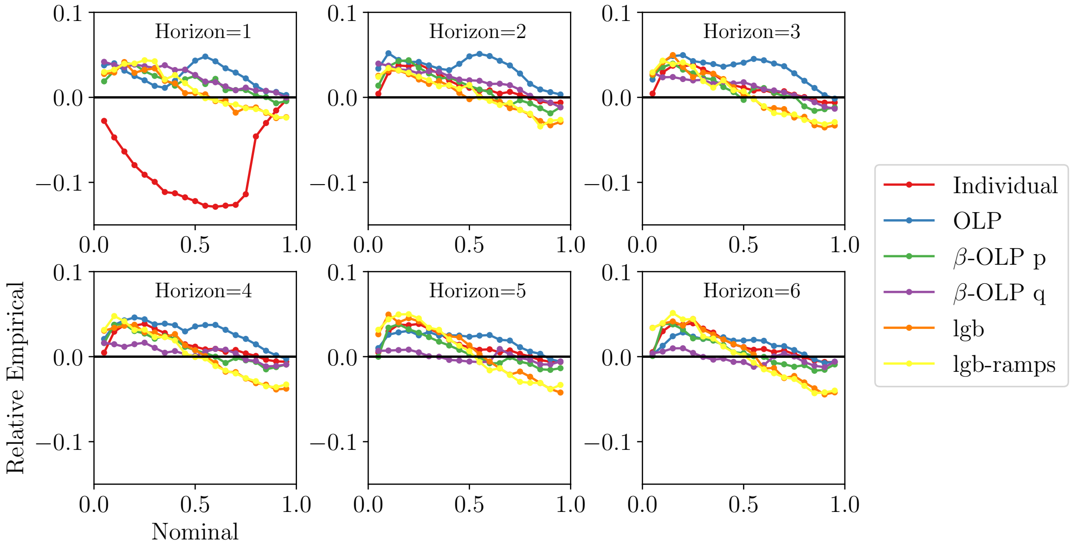


Figure 4.8: Relative reliability for the combined model forecasts at zone 1. A relative empirical frequency of zero represents ideal reliability.

traders and turbine operators. Therefore, the performance of each forecast combination method is also evaluated at times of ramps. A ramp may be defined as any time period with a change greater than a certain percentage of capacity, within a certain period of time. Figure 4.9 shows the distribution of ramps occurring within a 4-hour window. It can be seen that the distribution of upwards and downwards ramps is generally symmetrical, and that larger ramps are less common as might be expected.

Initial results discussed below are presented for a ramp with a change in power greater than 50% of site capacity, within a period of four hours. Upwards and downwards ramps are also distinguished, as these can result in quite different actions by the system operator [214]. Using this ramp definition, all time points in the forecast evaluation set can be classified as either an upwards ramp, downwards ramp or no ramp. The same classification is also applied to the actual power. We can then distinguish the times the forecast correctly predicted each type of ramp event via a confusion matrix. This may be done separately for each zone, combination model and forecast horizon. For example, the confusion matrix for the 1 hour ahead OLP forecast at zone 1 is shown in Table 4.2. Several metrics can be calculated from each confusion matrix; four relevant ones have been selected for analysis. They are based on True Positive (TP),

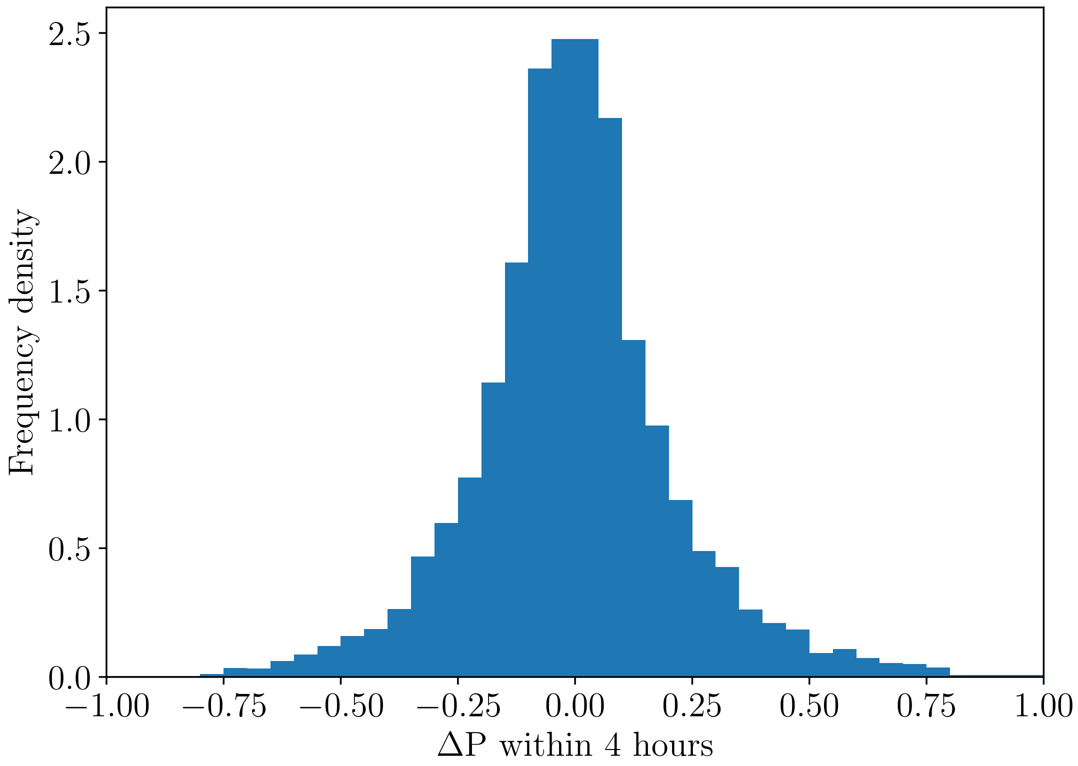


Figure 4.9: Distribution of ramps at zone 1.Each ramp is the maximum observed change in power within a 4-hour time window.

False Negative (FN) and False Positive (FP) counts. These may be calculated separately for upwards and downwards ramps. For example, $\text{TP}^{(u)}$ is the number of true positives for upwards ramps, i.e. an upwards ramp was forecast and an upwards ramp occurred. Similarly $\text{FP}^{(d)}$ would be the number of false positives for downwards ramps (the number of instances where a downwards ramp was forecast but didn't occur): this is a sum of times where a downwards ramp was forecast, and either no ramp or a positive ramp occurred. The number of false negative upwards ramps, $\text{FN}^{(u)}$ would be the total number of times an upwards ramp was not forecast (so either no ramp or a downwards ramp was forecast) but an upwards ramp did occur. For the example confusion matrix in Table 4.2, $\text{TP}^{(u)} = 262$; $\text{FP}^{(d)}$=6+29=35; $\text{FN}^{(u)}$=173+6=179. Then we can define a set of metrics which assess the skill of the forecasts in capturing upwards and downwards ramps separately. Three of these metrics are calculated separately for up and down ramps: the True Positive Rate (TPR), Positive Predictive Value (PPV) and

F1-score are calculated separately for upwards and downwards ramps. Their general definitions are

$$\begin{aligned}\mathrm{TPR} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\\ \mathrm{PPV} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\\ \text{F1-score} &= 2\frac{\mathrm{PPV}\times\mathrm{TPR}}{\mathrm{PPV}+\mathrm{TPR}}\quad .\end{aligned}\tag{4.6}$$

		Actual		
		Up Ramp	**Not Ramp**	**Down Ramp**
Forecast	**Up Ramp**	262	27	17
	Not Ramp	173	7526	188
	Down Ramp	6	29	230

Table 4.2: Confusion matrix for 1 hour ahead OLP forecast for zone 1

Each of these metrics is given for each forecast combination model and forecast horizon in Figure 4.10 for zone 1. The perfect score for all metrics is one, when all events are either true positives or true negatives. It should be noted that values of zero are included in brackets as these represent times where no true positives were observed in the evaluation set. While techniques such as cross validation were used to maximise the size of both the forecast model training dataset and the evaluation set, there are still some sites where few ramps were seen in the available data and therefore some of the metrics are based on few counts. In the case of no events at all, a ‘nan’ value is displayed.

The TPR, also known as the recall or sensitivity of the forecast, gives the proportion of true events (up or down ramps) that were correctly forecasted. The changes in TPR with model and forecast horizon are similar for both upwards and downwards ramps, with slightly lower values for negative (downwards) ramps. For the 1 hour ahead horizon, the best individual model has the best TPR for all zones; this model is the Markov Chain model which is built on transition probabilities so might be expected

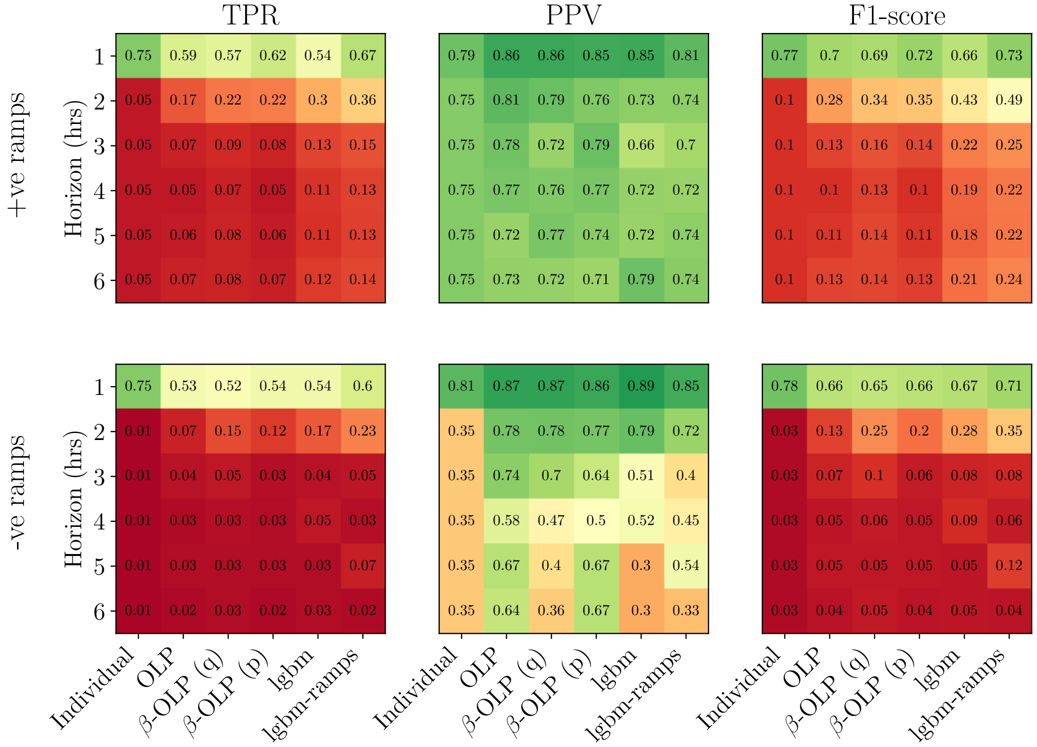


Figure 4.10: Confusion matrix scores for all combination models and forecast horizons for zone 1 separated by upwards (top row) and downwards (bottom row) ramp events

to capture ramps well. The only exception to this is the lightgbm-ramps model which slightly outperforms at zone 10 for positive ramps. For horizons of 2 or more hours ahead, the TPR rate is much lower (around 0.06 or 6% for positive ramps and 3% for negative ramps at zone 1). The best individual and lightgbm-ramps models still tend to outperform the OLP combination methods at these longer horizons. The lightgbm-ramps model shows the best TPR rate for all longer horizons at 6 of the 10 sites, and is best for at least some of the horizons for the remaining 4 sites. The TPR score for the lightgbm-ramps model tends to decrease more slowly with forecast horizon than the other models, suggesting slightly more true events are being picked up in the forecasts through the inclusion of ramp rate forecasts in the combination method. Finally, while the relative performance of the models is similar across locations, the absolute TPR value varies with location. This may suggest ramps at some locations are easier to detect than at others, although the number of ramps observed will also influence these results.

Positive predictive value (PPV), or precision, gives the proportion of forecasted ramps that were forecasted correctly. This is one of the most relevant metrics, as it gives a measure of confidence in the forecast when a ramp is predicted. Again, PPV is higher for the shorter forecast horizons. Scores are generally higher than than those for TPR, suggesting while the ramps that are forecast tend to be more accurate, the forecast still misses a significant portion of real ramps. There is no consistent difference in PPV between upwards and downwards ramps, across all the sites. The OLP-based methods, particularly OLP and β-OLP by power, tended to perform slightly better although this is not clear cut at all zones and all models performed quite similarly. The lightgbm-ramps model performs worse in particular at zones 7 and 8, suggesting it over-predicts ramps here.

The F1-score is defined as the harmonic mean of TPR and PPV, combining two desirable attributes: high levels of detecting true ramps, and high levels of positive forecast accuracy (when a ramp is forecast, it occurs). As such, the patterns seen in the F1-score are also a combination of those seen in the TPR and PPV. TPR, PPV and F1-score are all concerned with the quality of the forecast for positive events (upwards or downwards ramps) and while these are the events that have the most serious system consequences if they are not prepared for, the forecast performance at times of no ramps is also important. Other metrics that cover this, such as True Negative Rate (TNR) may also be evaluated but have been left out of this analysis for brevity.

Finally, the Accuracy score (Figure 4.11) gives the proportion of all events that were forecasted correctly (the sum of the diagonal elements in the confusion matrix, divided by the sum of all elements). This may be the least useful metric as it is mostly, but not entirely, dominated by instances of non-ramp events. Accuracy is around 0.9 or above for all zones and horizons except zone 10. No particular forecast combination model is considerably more accurate than any other.

So far ramp scores have been presented for one definition of a ramp, whereas different real-life applications may depend on different definitions of a ramp. Both the time interval and the magnitude of the power change may affect the power system actions

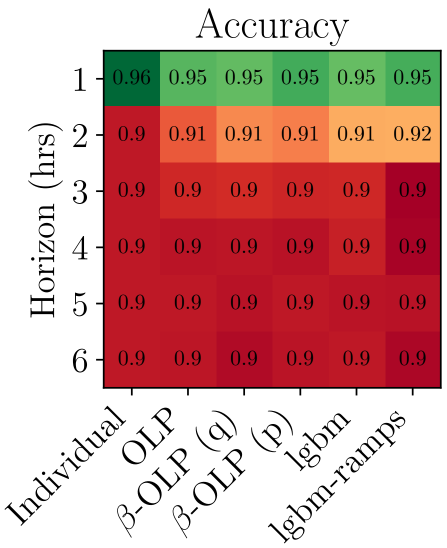


Figure 4.11: Accuracy scores for all combination models and forecast horizons for zone 1

taken to prepare for the ramp. Figure 4.12 shows the F1-score for a range of ramp magnitudes and durations. As might be expected, the most severe ramps (bottom left corner) show the lowest scores and the smallest changes in power over the longest time period have the highest score. In general the more severe the ramp, the less frequently they will occur. A very small number of extreme ramps, and a large number of time points with no or small ramps, in the training set is likely to lead to a model with poor performance for the extreme events. If an application requires high quality forecasts of extreme events, other approaches, for example using extreme value theory, may be more appropriate.

4.8 Conclusions

Four individual model forecasts were created, three based on recent measurements and one on NWP. None of these models showed good calibration before forecast combination. Four benchmark combination methods were employed: both constrained and unconstrained OLP, and a beta-transformed OLP combined by both quantile and power. A new nonlinear forecast combination method based on a GBM is proposed, including a variant that explicitly takes forecasts of ramp rate as well as individual power forecasts into the combination model. All forecast combination models showed higher skill scores (on both MAE and Pinball score) than the best individual model for horizons 1-3 hours ahead, with the greatest improvement for 2 hours ahead where the skill of individual models was most similar. The benefit of forecast combination was less clear 4-6 hours ahead where the individual GBM model performed much better than the other individual models.

The new combination approach based on a lightgbm model with ramp features gave significantly better skill than all other combination models 1 hour ahead (and up to 3 hours ahead at some locations), but the simpler linear combinations outperformed it at longer (4-6 hours) horizons. While forecast combination clearly did result in improvements in forecast skill over individual model forecasts, there are still limitations to this approach. It was observed that horizons with one much more skilful individual model saw little benefit from forecast combination. Inclusion of a very poor individual forecast may worsen the final combination so care should still be taken to ensure the individual forecasts are sufficiently skilful that they add value to the final forecast. The diversity of forecast inputs and computation methods is also likely to affect the size of benefit gained from combination. The proposed forecast combination model does not account for structural changes in the skill of the individual models; an adaptive forecast combination approach would be needed for this. This may be more important for other applications like demand forecasting where sudden changes in behaviour are more likely.

Times of ramps were also identified and the ability of the forecast to predict these assessed. It was found that the lightgbm-ramps model correctly forecasts the highest proportion of true ramps 2 or more hours ahead at 6 of the 10 locations tested, but also has a tendency to over-predict the frequency of ramps at non-ramp times. Overall, there is no one model that is consistently better at forecasting ramps across all locations and horizons.

The choice of combination model will depend on the individual site characteristics and the specifics of the wider problem (for example, if there is a greater penalty for missing a ramp than forecasting a false positive). A larger dataset would be needed with significance tests employed to fully understand differences between models. Integrating these forecasts with the financial consequences of system actions would better define their usefulness and limitations.

Table 4.1: Table of GBM input features and their relative feature importances averaged across the q10, q50 and q90 models and also averaged across all zones.
* Calculated using the Yamartino method [2]

Input	Average feature importance
U10	1.21
V10	1.03
U100	3.64
V100	2.68
10m energy	0
100m energy lag 4	1.06
100m energy lag 3	0.36
100m energy lag 2	1.10
100m energy lag 1	18.59
100m energy	0
100m energy lead 1	2.48
100m energy lead 2	0.59
100m energy lead 3	1.28
100m energy lead 4	1.83
100m energy / 10m energy	0.39
average energy over lags 1:4	27.44
average energy over leads 1:4	12.74
10 m windspeed	1.63
100m windspeed lag 4	0.07
100m windspeed lag 3	0.03
100m windspeed lag 2	0.14
100m windspeed lag 1	0.06
100m windspeed	14.40
100m windspeed lead 1	0.08
100m windspeed lead 2	0.02
100m windspeed lead 3	0.02
100m windspeed lead 4	0.05
percentage change in 100m windspeed since lag 1	0.14
wind shear	2.63
st dev of wind direction*	0.39
sin(hour of day)	1.03
cos(hour of day)	0.06
sin(month of year)	0.32
cos(month of year)	2.48

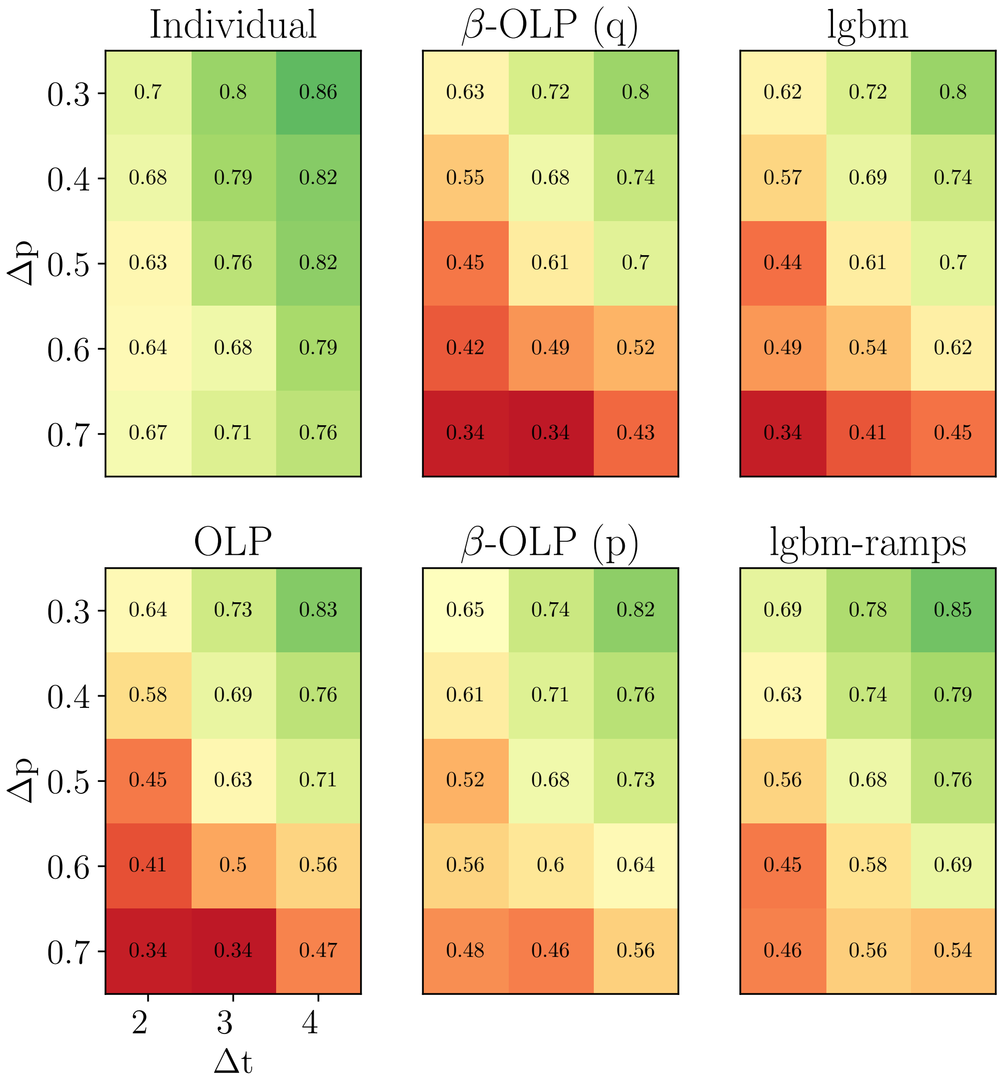


Figure 4.12: F1 scores for varying ramp definitions, for all combination models at zone 1. Values are for a 1 hour ahead forecast.

Chapter 5

Subseasonal to seasonal forecasting for wind turbine maintenance scheduling

This chapter is based on the work presented in the paper 'Subseasonal-to-seasonal forecasting for wind turbine maintenance scheduling' which has been accepted for publication in Wind. The text from this article has been edited and extended here.

Abstract

Certain wind turbine maintenance tasks require specialist equipment, such as a large crane for heavy lift operations. Equipment hire often has a lead time of several weeks but equipment use is restricted by future weather conditions through wind speed safety limits, necessitating an assessment of future weather conditions. This chapter sets out a methodology for producing subseasonal-to-seasonal (up to 6 weeks ahead) forecasts that are site- and task-specific. Forecasts are shown to improve on climatology at all sites, with fair skill out to six weeks for forecasts of both variability and weather windows. For the case of crane hire, a cost-loss model identifies the range of electricity prices where the hiring decision is sensitive to the forecasts. While there was little difference in the hiring decision made by the proposed forecasts and the climatology

benchmark at most electricity prices, the repair cost per turbine is reduced at lower electricity prices.

5.1 Introduction

While the skill of traditional Numerical Weather Prediction (NWP)-based forecasts generally decreases with increasing forecast horizon, there are some applications where decisions must be made at longer horizons and therefore longer-range forecasts may be beneficial. In the case of wind energy, turbines must be maintained, through both annual servicing campaigns and ad-hoc maintenance of worn or failed parts to ensure maximum operating time and efficiency, and therefore energy output, from the turbine. Most of the mechanical equipment for the turbine is housed in the nacelle at the top of the tower, and so a large proportion of maintenance tasks require a technician to physically climb the turbine and enter the nacelle. Once a wind farm has been built, Operations and Maintenance (O&M) is the largest ongoing cost for the wind farm owner and as such it is in their interests to minimise these costs while maximising turbine availability. . Activities such as work in the nacelle or crane use have strict safety limits on the wind speed so work may only be carried out below these limits. Maintenance tasks also often require the hire of cranes, contractors and other equipment that must be booked in advance often with a wait time of several weeks. Knowledge of the likely weather conditions can allow for improved scheduling, by allowing an initial decision to be made further ahead or by giving more information than just relying on the average conditions for the time of year. For example, knowing if a given week is expected to be more or less windy than average would allow planning of the number of jobs to schedule for that week. Extended spells of unusually low wind, sometimes also coupled with high demand due to cold weather, can make power system management more difficult (and expensive); Subseasonal-to-Seasonal (S2S) forecasts can give an indication of these unusual conditions further in advance, allowing preparations and corrective actions to be taken [215].

Subseasonal forecasts are generally defined as 10 days to one month ahead, with seasonal forecasts covering one to seven months ahead [215]. They cover the ‘gap’ between conventional weather models and longer-term seasonal predictions. NWP-based forecasts typically have low accuracy beyond the two week horizon due to imperfect knowledge of atmospheric conditions, imperfect physical model and chaotic nature of the atmosphere. However, atmospheric interactions with other systems such as the ocean, ice and soil moisture produce 'forcing' effects, and therefore atmospheric changes, on longer timescales [216] which S2S forecasts aim to capture through atmosphere-ocean-ice-land coupling. The output S2S forecast from a weather forecast provider such as the European Centre for Medium-range Weather Forecasts (ECMWF) is a set of ensembles. These are produced through perturbations to both the model initial conditions and to the model physics. ECMWF provides 50 perturbed ensemble members plus one un-perturbed one for example. Taken together, this set of ensemble members includes the uncertainty in the forecast and can be used for further probabilistic analysis of the future conditions. While the skill of S2S forecasts is not high enough to produce a calibrated probabilistic forecast for a specific hour or day, there is some skill in forecasts averaged over longer periods (e.g. weekly mean conditions) and determining if conditions are likely to be more or less windy than average for that time of year. While by no means a complete description of future conditions, partial information can still aid decision making and planning on these timescales [192, 217].

The potential for S2S forecasts spans a wide range of sectors with varying levels of familiarity in using forecast information [192, 217]. The agricultural sector is quite familiar with the use of forecasts to make decisions on when to irrigate, apply fertilizer and pesticides, and when to harvest, but could still benefit from information on the S2S timescale. Meanwhile public health decision makers are generally less familiar with the use of forecasts but could benefit from advance warning of a likely heatwave or cold snap, or disease outbreaks such as malaria where there is a strong weather dependence. Other sectors that could benefit from forecasts on the S2S timescale are preparation for extreme events such as floods and droughts, water management including onset of rainy seasons, and other energy applications such as hydropower and system operator planning. White et al. [217] detail a variety of case studies spanning all these applications.

Methods for forecasting persistent and extreme cold events include trajectory analysis of the two main Empirical Orthogonal Functions (PCs) of 500 hPa geopotential height to represent mid-tropospheric flow [218]. The impact of the Madden-Julian Oscillation (MJO) for forecasts across Europe was found to be asymmetric, with negative phases of the MJO showing positive skill but little skill associated with positive MJO phases. They demonstrate the use of trajectories in EOF phase space to show transitions between atmospheric states as the forecast evolves. The MJO was also found to be a source of predictability for heavy tropical precipitation; Specq [219] found that 'MJO phases are typically precursors of rainfall east of the positive convective anomaly'. They found an improved hit rate for predicting onset of heavy rains across the Western Pacific and Africa for week 2 of the forecasts but not elsewhere, further, an increased hit rate was also accompanied by an increase in false alarms. A statistical model using feature engineered inputs from S2S weather models was developed to forecast North Atlantic tropical cyclones, giving improvement over purely NWP model based forecasts [220]. Other work on forecasting of tropical cyclones has focussed on the predictability of storm surges [221]. Using 10m wind speed and sea level pressure as forcing variables coupled with a hydrodynamic model, they found some skill for past extreme events (hurricanes Katrina and Isabel) but this was limited to 4-10 days ahead. They found skill is sensitive to model initialisation conditions, the number of ensemble members and the horizontal resolution of the model.

Research in S2S forecasts for energy applications is a relatively new and growing field; the S2S prediction project has been working to 'improve forecast skill and understanding on the subseasonal-to-seasonal scale' since 2013 [222] and the S2S4E (Subseasonal-to-seasonal forecasting for energy) project [223] has produced a body of work specific to energy applications in the last 4 years. One of the project's main outputs was a decision support tool tailored to wind, solar, hydropower and demand forecasts and displaying forecast skill on a coarse grid across the globe. While this will hopefully encourage potential users in the energy industry to consider the benefits of S2S forecasts, the economic value and range of applications and therefore the potential of S2S forecasts in the energy industry has not been fully investigated [224]. The same work also notes some of the key challenges for S2S forecasting such as: the terminology gap between forecasters and decision makers; the best use of probabilistic information in decision making; the long-term benefits delivered by S2S forecasts versus the immediate benefits demanded by business; and current meteorolgical modelling limits.

Meteorological organisations produce forecasts on subseasonal-to-seasonal timescales from sophisticated models put together by experts. This involves not only understanding and modelling of physical atmospheric processes but also, for S2S forecast models, modelling of slower-varying interactions between the atmosphere and other physical systems such including land (e.g. soil moisture), ice, and the ocean. Moreover, weather models require assimilation of a huge number of measurements representing the initial state of the atmosphere and also simulation across a global grid with a large number of points. Understanding of physical processes that drive certain weather patterns and teleconnections can improve NWP models and aid in developing good forecasts.

Beerli et al [225] show that a particularly extreme (weak or strong) polar vortex is related to strong coupling between the troposphere and stratosphere, leading to more persistent phases of the North Atlantic Oscillation (NAO), which is in turn associated with certain patterns of wind speed across Europe. They focus only on winter months where predictability is higher, however. The Quasi-biennial Oscillation (QBO) consists of alternating westerly and easterly winds in the tropical stratosphere and this has been found to improve prediction of the Northern hemisphere stratospheric polar vortex a month ahead [226]. Jung [227] note that understanding of polar atmospheric processes is limited compared to processes at lower latitudes; an experiment where models are 'relaxed' towards reanalysis data in the polar region shows that improved polar forecasts would lead to improvements in subseasonal midlatitude forecasts, but only for some regions. Benefit was shown for northern areas of North America, Eastern Europe and Northern Asia in particular.

The work on seasonal forecasting for energy focuses on transforming forecasts of weather variables into more relevant energy system specific quantities, applying various postprocessing correction methods and evaluating the skill of these forecasts for specific energy applications. A common approach involves implementing Principal Component

Analysis (PCA) on the forecasts from all grid points over a large geographical region (for example the North Atlantic and Europe). This identifies key regimes, or patterns, that can then be linked to variables of interest at the chosen location. This may simply involve finding the variable's distribution conditional on one of the principal components [228], or making some kind of 'weather index', be that an index representing the current phase or amplitude of a certain atmospheric mode like the NAO or MJO [229], or a value calculated for local conditions from the principal components [230].

One of the first works to investigate subseasonal-to-seasonal predictability for wind speeds was Lynch et al [231], who found skill in country-wide weekly mean wind speed out to 14-20 days. A recent review [232] found that while the body of academic research on subseasonal forecasting for wind has expanded in the last few years, there is still work to be done to transfer this research into useable forecasting products. They summarised that the NAO, East Atlantic (EA) and Scandinavian Pattern (SCA) are the main patterns that explain weather conditions, but that different patterns are most important at different times of year and for different weather variables. In agreement with [225] they identify the polar vortex as another important source of predictability for European winter forecasts.

Lledò et al. [228] show that forecasts of the MJO can lead to skilful probabilistic daily mean wind speed forecasts up to 36 days ahead. However, they find that the current numerical weather model used to produce S2S forecasts at the ECMWF does not reproduce the teleconnections from the MJO phases to European weather and that currently, strong MJO events are not skilfully predicted more than 10 days ahead. Although the North Atlantic Oscillation accounts for approximately a third of European circulation variability, Lledò et al. [229] examined a wider range of teleconnection indices and found forecast skill from all four tested (the NAO, EA, East Atlantic Western Russia (EAWR) and SCA). They also note that combining forecasts from different providers improves ensemble mean correlation, and that results can be sensitive to hindcast period length and number of ensemble members. While seasonal forecasting linked to wind speeds often focuses on the most windy winter months [233], Lledò et al. [228, 229] examine skill for all seasons separately.

Alonzo et al. [230] used polynomial regression on PCA components to produce a site-specific index with which to produce conditional forecast distributions. Recalibration of the forecast ensembles improved the results and led to calibrated forecasts that were 30% sharper than climatology for their French case study. Cortesi et al. [234] uses clustering to identify regimes and maps locations and times of year where each regime had particularly strong skill in reproducing wind speeds. Wind energy production has also been forecast on a subseasonal-to-seasonal timescale using capacity factors for three main turbine classes [216]. While this technique allows forecasts to be made for any location without on-site data, this may result in lower skill compared to site-tailored forecasts.

Gonzalez et al [235] apply an adaptive linear combination approach to produce final forecasts that use multiple forecast models as inputs along with features like climatology and persistence derived from reanalysis datasets. They find skill relative to climatology, the best individual NWP model and other non-adaptive combination benchmarks for forecasts 2-5 weeks ahead with particular improvement further out, when forecasting national electricity demand from 2m temperatures. The adaptive nature ensures the model is able to adjust to relative changes in skill of the inputs, such as when a model update is released. Hwang [236] presents an alternative approach to S2S forecasting, focussing on statistical methods. They present two statistical models using customised feature selection and a nearest neighbours-based approach that are averaged to produce final forecasts. This methodology proved successful in NOAA's Subseasonal Climate Forecast Rodeo, with 40-169% improvements over the competition benchmarks for temperature and precipitation forecasts.

Work has also been done on larger geographical scales, which is needed for electricity system planning or studying whole country renewable penetration. National demand, wind and solar forecasts were produced for 28 European countries [237] with consistent skill at the 5–11 days horizon but variable skill for longer horizons, depending on location and specific distribution of renewables. On national aggregation levels, the complex interactions between the electricity system and the weather require a considered approach: Bloomfield et al. [233] used 'targeted circulation types' to identify weather drivers for the electricity system in Europe. They note that the strength of response between a weather pattern and the electricity system is strongly dependent on the capacity and location of renewable generation installed. Van der Wiel et al. [238] also defined regimes by their impact on energy variables, rather than the traditional approach that maximises circulation variance explained. They reconstruct 2000 years of data over Europe to examine the skill of S2S forecasts at identifying extreme energy shortfall events (where there is high demand but low renewable generation).

Existing research on maintenance scheduling focuses heavily on decisions specific to offshore wind farms such as vessel choice [239–242] or route planning between turbines [240, 241, 243]. Optimisation of the installation phase which involves hire of specialist equipment [244], optimisation of downtime relative to whole power system concerns like reserves [245], and prioritisation of maintenance by component type [246] have also been studied. However, perhaps due to the relatively lower cost of maintenance for onshore farms compared to offshore, there has been much less of a focus on maintenance planning for onshore operations. Methods based on maximising usage of weather windows that have been applied offshore also apply to onshore operations where wind speed safety limits apply, such as crane usage or work in the turbine nacelle.

Barlow et al. [247] note a distinction between ‘performance’ and ‘condition’ of an asset — for example, an old component that is still operating well may not have resulted in any loss of generation efficiency but its risk of failure has increased. They propose an approach that takes both performance and condition into consideration when deciding when to repair or replace assets. Pelajo et al. [248] combine wind and electricity price forecasts up to 7 days ahead to make a decision on when to take turbines offline for annual maintenance using a cost benefit approach. Their results show a 30% reduction in costs compared to going offline at a random point in the maintenance season. A cost-loss model has also been used for day-ahead go/no go decisions by Browell et al. [249]. They showed probabilistic forecasts of access windows, as opposed to deterministic, increased the number of days worked and decreased lost revenue. They also note that other works assume perfect forecasts [241, 243].

Aside from strategic decisions in the design stage, none of these works consider decisions that must be made weeks ahead such as crane or additional crew hire. Thus, there is no overlap between the maintenance scheduling literature and the work done on S2S forecasting for renewables.

Given the identified lack of tailored forecasts for wind turbine maintenance decisions weeks ahead, this chapter aims to demonstrate the production of forecasts of relevant quantities and assess their skill, both as raw forecasts compared to a benchmark, and also whether use of these forecasts can result in improved decision-making for problems such as equipment hire. A methodology for this is presented through a case study, where forecasts are developed and evaluated for eight wind farms in Scotland.

5.2 Methodology

The case study presented here aims to evaluate the presence (or lack) of skill in post-processed S2S wind speed forecasts at hub height for use by the wind energy industry (the hub is the section where the blades attach to the turbine body). Several possible forecast metrics are investigated along with a demonstration of possible forecast use in decision making. Essentially, the two main questions this case study aims to answer are: is there skill in S2S wind speed forecasts downscaled to a wind farm?, and does this skill translate into improved decision making for maintenance scheduling tasks? The overall modelling process is laid out in Figure 5.1 and the individual steps and data partitioning is further explained in the subsequent individual sections. Raw S2S forecasts are available for a grid of locations covering the whole globe and multiple vertical levels in the atmosphere. The methodology presented in this work produces forecasts for task-relevant metrics at the site of interest from this large scale data with forecasts of specific weather variables.

5.2.1 Data

The case study for this section focusses on onshore wind farms in Scotland as this is a geographical area with a relatively high concentration of onshore wind. Onshore wind in Scotland is a relatively mature sector, so there are several sites with fairly

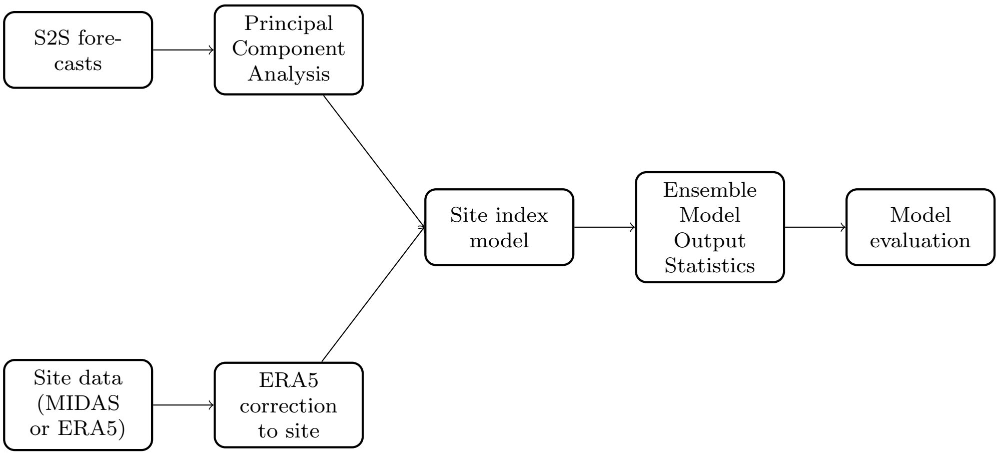

S2S forecasts
Principal Component Analysis
Site index model
Ensemble Model Output Statistics
Model evaluation
Site data (MIDAS or ERA5)
ERA5 correction to site

Figure 5.1: Flowchart of data and modelling steps to produce the final site index forecasts. The Principal Component Analysis step identifies large-scale weather patterns in the grid of forecasts over a whole European and North Atlantic area; the ERA5 correction to site fits a Generalised Additive Model (GAM) to correct ERA5 windspeeds to more closely match the short history of observed site wind speeds. The site index model takes the S2S principal components as inputs and is trained to forecast a site specific index given by the site data for the training set. The Ensemble Model Output Statistics step corrects any bias or dispersion problems with the ensemble of site index forecasts before forecasts are evaulated.

long historical datasets available. Eight such sites are used in this study. ECMWF forecasts of key variables on a grid covering the North Atlantic and Europe region (80°N/-90°W/20°S/30°E) were downloaded via the S2S database[1] [250]. New forecast runs are produced twice a week and all forecasts available since the 2019 model update were used, from June 2019 to May 2021. Forecast values were downloaded for horizons 0–42 days ahead in 24 hour increments. In addition, the same horizons were downloaded for 20 years of hindcasts for each forecast date, i.e. for a forecast on the 1st January 2020, the hindcast dates of 1st January 2019, 2018, etc are also downloaded. This gives a set of hindcasts over the 'hindcast times' 1999–2019 and then a two-year set of forecasts over the 'forecast times' June 2019–May 2021. When forecasts are later averaged into weekly values, 1 week ahead includes forecasts with horizon 1–7 days ahead, 2 weeks ahead includes the 8–14 day ahead forecasts and so on.

Geopotential height or mean sea level pressure are the most common meteorological

[1] `https://apps.ecmwf.int/datasets/data/\gls{s2s}`

variables used to study S2S weather patterns as they describe large scale atmospheric processes on these timescales [228–230, 234, 238]. In addition to this, we included other variables related to hub height wind speed. As the S2S forecasts do not currently have a specific 100 m wind speed variable, several variables related to this (both 1000 and 925 hPa wind speeds and mean sea level pressure) were included as well as 10 m wind speed.

5.2.2 Site wind speed data

In an ideal world, there would exist measured wind speed data for the site of interest at hub height, possibly at all turbine locations, and for the same time period (1999–2021) as for the ECMWF hindcasts and forecasts. However, very few wind farms have been operating this long and so measured wind farm datasets tend to be much shorter. Therefore, two approaches have been tested: first we have used wind speed time series from the Met Office Integrated Data Archive System (MIDAS) dataset [251] for weather stations where a complete history over the time period is available. This allows assessment of the skill of the forecasts when trying to model measured wind speeds without extra simulated data. However, the majority of these weather stations measure wind speed at 10 m as opposed to the much taller hub heights of turbines that are of interest. To also quantify the 'real-life' skill and usefulness of S2S forecasts we also used ERA5 reanalysis data from the nearest grid point to selected wind farms and corrected this long time series with the shorter measured time series available from the site itself. A comparison of various reanalysis products for surface and near surface wind speeds found the ERA5 data provides the closest agreement with observations [252]. Both of the MIDAS and ERA5 datasets provide mean wind speeds at hourly resolution, allowing for not only calculation and forecasting of weekly mean wind speed but also of other quantities such as standard deviation of the hourly wind speed values across the week and metrics relating to specific wind speed thresholds, for example proportion of the time wind speed is below the safe threshold for work in the nacelle.

All the eight wind farms assessed are in Scotland, with two single farms (WF 1 and 3) and two groups of three geographically close farms (2a, 2b, 2c and 4a, 4b, 4c).

The four MIDAS stations selected are those geographically closest to these 4 areas that also recorded wind speed measurements for the required time period and are labelled numerically to match the wind farms (so MIDAS 1 is close to WF 1, etc). Table 5.1 describes the general location and terrain of the wind farms used.

Wind Farm	Description
1	South-West Scotland, moorland with some forestry nearby
2a, b, c	Mid Scotland, close to East coast, rolling hills
3	Aberdeenshire, forestry land
4	Moray, open moorland

Table 5.1: Summary of case study wind farms, and their general locations.

5.2.3 Initial transformations and corrections

The ERA5 wind speed time series v_{ERA5} required correction to the site wind speeds v_s through a GAM, with a cubic spline of the ERA5 wind speed along with a cyclical cubic spline ($f_c(\cdot)$) of the day of the year (doy) to account for seasonal variability:

$$v_s = f(v_{ERA5}) + f_c(\text{doy}) \quad . \tag{5.1}$$

This was implemented with the `gam` function from the `mgcv` package in `R`, using its default numerical optimisation method. The basis dimension of the splines, k, was checked with `gam.check` to ensure it was sufficiently large. All available time points were used for training the GAM before predicting the corrected wind speed for all times in the hindcast and forecast sets.

The first step in processing the raw S2S forecasts was to identify the large scale patterns and atmospheric 'modes' present. Principal Component Analysis allows reduction of multi-dimensional data to fewer dimensions, where the new components are all orthogonal and components account for a sequentially decreasing proportion of variance in the original data. If $\mathbf{D}$ is a $t \times v$ matrix containing all the forecast values for a given variable with each spatial grid point in a separate column (normalised to zero mean) and each time point for a given forecast horizon in a separate row, the transformation

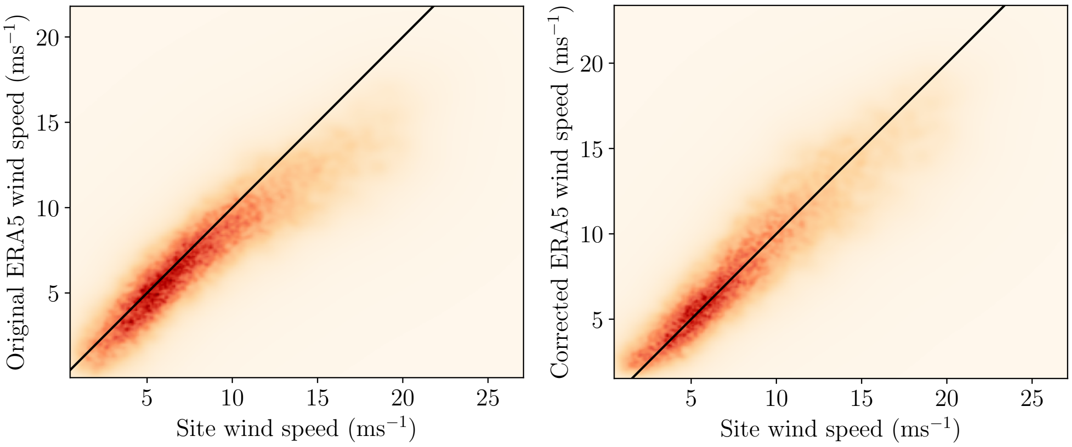


Figure 5.2: ERA5 wind speeds compared to site wind speeds before and after correction. Shown for WF1.

matrix $\mathbf{F}$ to produce the principal components will have dimensions $v \times v$ (or $v \times n$ if only the first n components are kept) and is made up of the column eigenvectors of the covariance matrix of $\mathbf{D}$, such that the eigenvector corresponding to the largest absolute eigenvalue is in the first column and so on. Then the transformed principal components $\mathbf{T}$ are given by

$$\mathbf{T} = \mathbf{DF}. \tag{5.2}$$

The first n columns of $\mathbf{T}$ give the first n principal components; n was selected so the retained principal components explain 90% or more of the variance in the original data — this is typically 20–40 components for the weather variables used here. Since forecasts are not available for 100m wind speed directly and several related weather variables are used instead, we have performed PCA on each of these separately as well as a PCA with them all together, to determine the optimal approach. The zero-hour ahead forecasts for the hindcast times were used to fit the PCA transformation before it was applied to the 1–42 days ahead forecasts for all the available times (both hindcast and forecast times).

We can also use the ‘reverse’ transformation to project single principal components back onto 'map' space, using $\mathbf{TF}^{-1} = \mathbf{D}$ with $\mathbf{T}$ being the i^{th} eigenvector to get a map of the i^{th} principal component. This allows us to observe the spatial patterns associated with the principal components as seen in Figure 5.3.

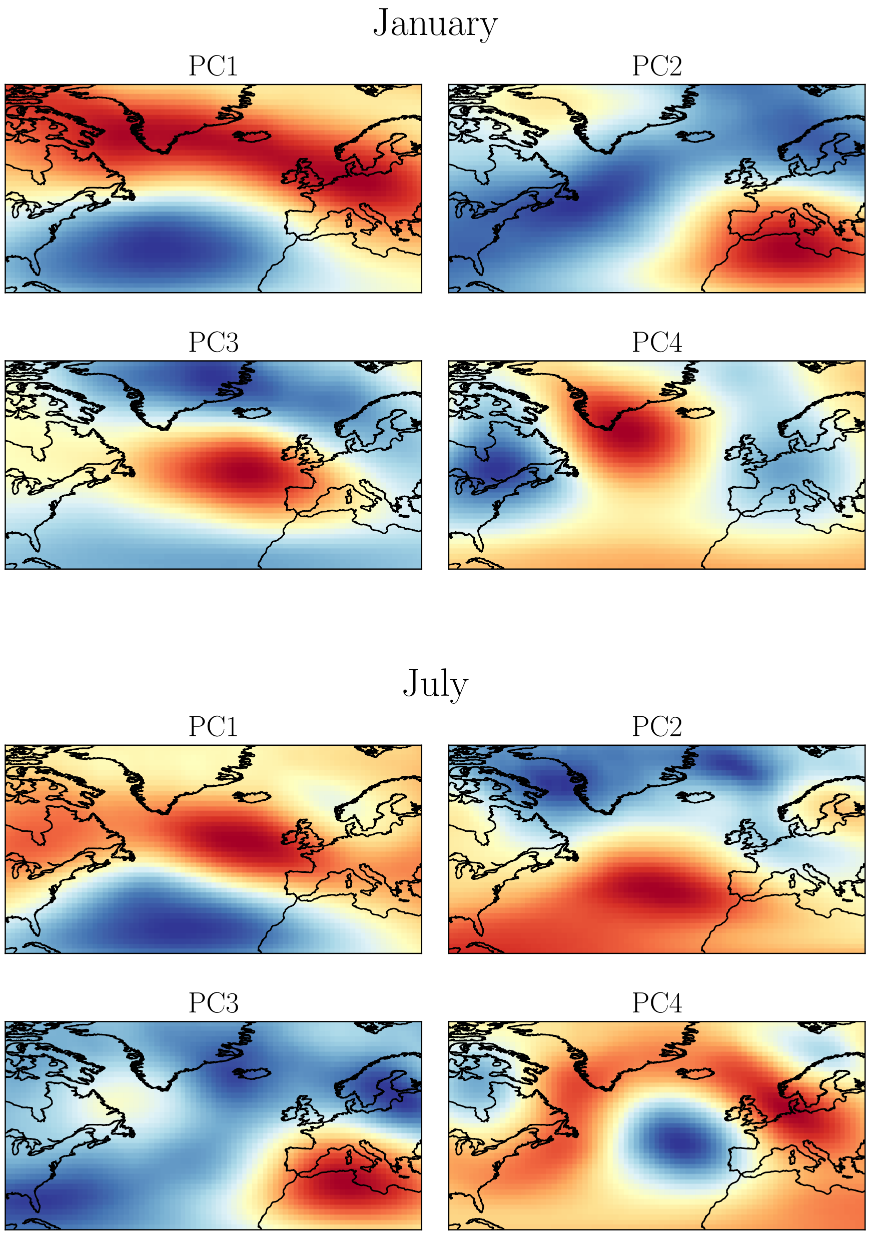


Figure 5.3: Principal components of 500 hPa geopotential height mapped back to geographical space, for January and July.

5.2.4 Site-specific index models

While the transformed Principal Component (PC) data now represents the main atmospheric patterns, this is still general to the whole North Atlantic-European region. Therefore, a model is needed to relate the whole-region regimes (represented by the PCs) to the local wind conditions at the site of interest. A similar approach was used by Alonzo et al. [230] with a polynomial regression model to generate site indices. Index models are trained for three different metrics, weekly mean wind speed, weekly standard deviation of wind speed and a 'useful hours' index tied to the safety limit for a given task.

Index forecasts for the hindcast time points are required for recalibration of the ensemble index forecasts (the Ensemble Model Output Statistics step), so we must produce out of sample index forecasts for the hindcast times as well as the forecast times. As such, cross validation is used where the 20-year hindcast dataset is broken into folds of 2 years each. Index forecasts for each fold of the hindcast times are generated by training on all the remaining hindcast data, however, the year immediately after the test fold is left out of the training folds as there can exist some long-term patterns or correlations in the atmospheric conditions. For example, to generate index forecasts for 2000–2001, all the remaining hindcast times apart from 2002 are used as the training set. To generate index forecasts for the forecast (as opposed to hindcast) times, the entire set of hindcasts is used as the training set. Models are fitted separately for each weekly forecast horizon and forecasts are made for each ensemble member.

Mean wind speed index

Weekly mean wind speed is derived from the hourly wind speeds within the week $\{w_T, w_{T+1}, ..., w_{T+168}\}$ and is defined as

$$W_T = \frac{1}{168} \sum_{t=T}^{T+168} w_t \quad . \tag{5.3}$$

for the week beginning $t = T$. The forecast index of weekly mean wind speed is produced by fitting a linear model where the input features are the principal components of the weather variables. If the weekly mean of the i^{th} principal component of the j^{th} weather variable is denoted as $\overline{\text{PC}}(i, j)$,

$$\hat{W}_T = \beta_0 + \sum_{j=1}^{3} \sum_{i=1}^{n} \beta_{i,j} \overline{\text{PC}}_T(i, j) \quad . \tag{5.4}$$

A linear model is used to generate indices; as there are many input features (tens of PCs for 3–6 different weather variables), a Least Absolute Shrinkage and Selection Operator (LASSO) approach using regularisation for feature selection is used. This means instead of finding the optimal β coefficients by purely minimising the squared errors, an additional penalty term proportional to the sum of the magnitudes of the β coefficients is added to the loss function:

$$\min_\beta L_1 = \left(W - \beta_0 - \sum_{j=1}^{3} \sum_{i=1}^{n} \beta_{i,j} \overline{\text{PC}}(i, j) \right)^2 + \lambda \left(|\beta_0| + \sum_{j=1}^{3} \sum_{i=1}^{n} |\beta_{i,j}| \right) \quad . \tag{5.5}$$

This extra penalty serves to shrink the absolute size of coefficients, which introduces sparsity, and the value of λ determines the ‘severity’ of this penalty. This is optimised through nested cross validation on the training set. Different configurations of inputs for the index model are tested: extra weekly features (weekly standard deviation, min and max of wind speed) are engineered and included as inputs; and the three weather variables related to 100 m wind speed are transformed into PCs separately to produce three sets of PC, or all together in one transformation. The optimal configuration in terms of whether to group weather variables for the PC transformation, and whether to use extra features, was determined separately for each final index metric using the MIDAS dataset as the ‘real’ wind speeds and is given in Table 5.2. It is assumed that the relative performance of these different model configurations will be very similar when an index model is trained on the corrected ERA5 data instead of the MIDAS wind speed data, so the same optimal model configurations were used for the remainder of the work.

Variability index

On the day of a maintenance activity, the measured hub (or crane) height wind speed is the deciding factor as to whether it is safe for a job to go ahead or not. However when these activities are planned weeks ahead and in whole-week chunks, there is significant uncertainty in the mean wind speed forecast and information about the conditions within the week is lost. Two weeks could have the same mean wind speed while one contains very stable weather for the whole week and one varies between periods of very low and very high wind speeds within the week. Thus, a forecast of the variability of wind conditions within each week is also useful, giving another point of information to base decisions on. This may be particularly relevant when the mean wind speed forecast is close to the safety threshold for the given activity. Standard deviation is used as the measure of inter-week variability and is calculated from the hourly data for that week:

$$S_T = \sqrt{\frac{\sum_{t=T}^{T+168} |w_t - W_T|}{168}} \quad . \tag{5.6}$$

Extra weekly features are derived from the hourly principal components, including weekly standard deviation $\mathrm{PC}^{(\sigma)}$, weekly minimum $\mathrm{PC}^{(\min)}$ and weekly maximum $\mathrm{PC}^{(\max)}$. These are also included in the linear model as linear features with coefficients γ, δ and ϵ respectively:

$$\begin{aligned} \hat{S}_T = \beta_0 + \sum_{j=1}^{3}\sum_{i=1}^{n} \beta_{i,j}\overline{\mathrm{PC}}_T(i,j) + \sum_{j=1}^{3}\sum_{i=1}^{n} \gamma_{i,j}\mathrm{PC}_T^{(\sigma)}(i,j) \\ + \sum_{j=1}^{3}\sum_{i=1}^{n} \delta_{i,j}\mathrm{PC}_T^{(\min)}(i,j) + \sum_{j=1}^{3}\sum_{i=1}^{n} \epsilon_{i,j}\mathrm{PC}_T^{(\max)}(i,j) \quad . \end{aligned} \tag{5.7}$$

Regularisation in the form of a LASSO penalty shrinks coefficients, leading to a more sparse model. The lasso penalty is applied to all $\beta_{i,j}, \gamma_{i,j}, \delta_{i,j}$ and $\epsilon_{i,j}$ coefficients in the same way.

Weather window indices

The third site index forecast produced links the weather conditions to a specific maintenance task of interest by forecasting the number of hours in the week that are safe to do that task. Each maintenance activity has a safe wind speed limit associated with it and takes a certain amount of time to perform. Therefore, we need a window of at least x hours where the wind speed is always below y ms^{-1} to perform a given activity. Two obvious metrics to measure this on a weekly basis would be the number of hours in the week below the wind speed threshold, or the number of weather windows in the week. However, there are pitfalls to both of these. Counting the total number of 'safe' hours in the week doesn't tell us anything about the length of weather windows available, so that 10 hours below the wind speed threshold could mean a continuous block of 10 hours available, or 10 separate blocks of 1 hour windows which is much less useful for a task that takes several hours. If we instead report the number of weather windows of a certain minimum time length, a value of 1 window in a week could mean a single window of the minimum time duration, or that the wind speed is below the threshold for the entire week; the second of these scenarios would allow a lot more work to be done in practice than the first. The final metric used is therefore a combination of these: the number of hours in the week contained in a weather window of a certain minimum length. This way, every hour that is counted is valuable for work as it is guaranteed to be in a usefully long window and there is still a distinction between a single long and a single short weather window. This metric doesn't differentiate between one long and several shorter windows, but the minimum duration constraint ensures each window counted is long enough for maintenance activities to be performed in it. Thus, the two parameters defining the weather windows (duration and wind speed threshold) are both determined by the maintenance activity of interest. All weather windows are counted, regardless of whether they fall within the usual working week (9–5 Mon-Fri) or not, as the decision to work at nights or weekends will depend on other factors including the urgency of the job and availability of staff to work other hours. The number of useful hours metric $H_T^{(p,q)}$ is the number of hours in the week where wind speed is below p ms^{-1} for q hours or more at a time and is modelled as

$$\hat{H}_T^{(p,q)} = \beta_0 + \sum_{j=1}^{3}\sum_{i=1}^{n} \beta_{i,j}\overline{\mathrm{PC}}_T(i,j) \quad . \tag{5.8}$$

As for the mean wind speed and variability indices, LASSO regularisation is employed for feature selection.

Table 5.2: Table of index model input features. gph=geopotential height; ws=wind speed. *925 and 1000hPa wind speed and mean sea level pressure are all included in one PC transform together.

Variable(s)	PCs kept	Features	Index models		
			Mean ws	Variability	Weather window
500 hPa gph	20	weekly mean	✓	✓	✓
		weekly sd		✓	
		weekly min,max		✓	
10 m ws	40	weekly mean	✓	✓	✓
		weekly sd		✓	
		weekly min,max		✓	
100m variables*	20	weekly mean	✓	✓	✓
		weekly sd		✓	
		weekly min,max		✓	

5.2.5 Ensemble Model Output Statistics

The ECMWF S2S forecasts provide 51 ensemble members, generated through perturbations to the intial conditions for the model run and the model physics. This encapsulates uncertainty information about the forecast as well as just the ‘best’ single forecast. However, there may still be bias, or over or under dispersion of the ensemble members, or both. Ensemble Model Output Statistics (EMOS) aims to correct this to provide a final unbiased, calibrated forecast. From Figure 5.4 we can see the index forecasts have virtually no bias, while a time series plot (Figure 5.5) shows the ensemble members are significantly under dispersed — this is also shown by the U-shaped verification rank histogram (Figure 5.6).

In this work, the EMOS method used is similar to that in Schuhen et al. [253] and is based on fitting of a parametric distribution to features of the ensemble mem-

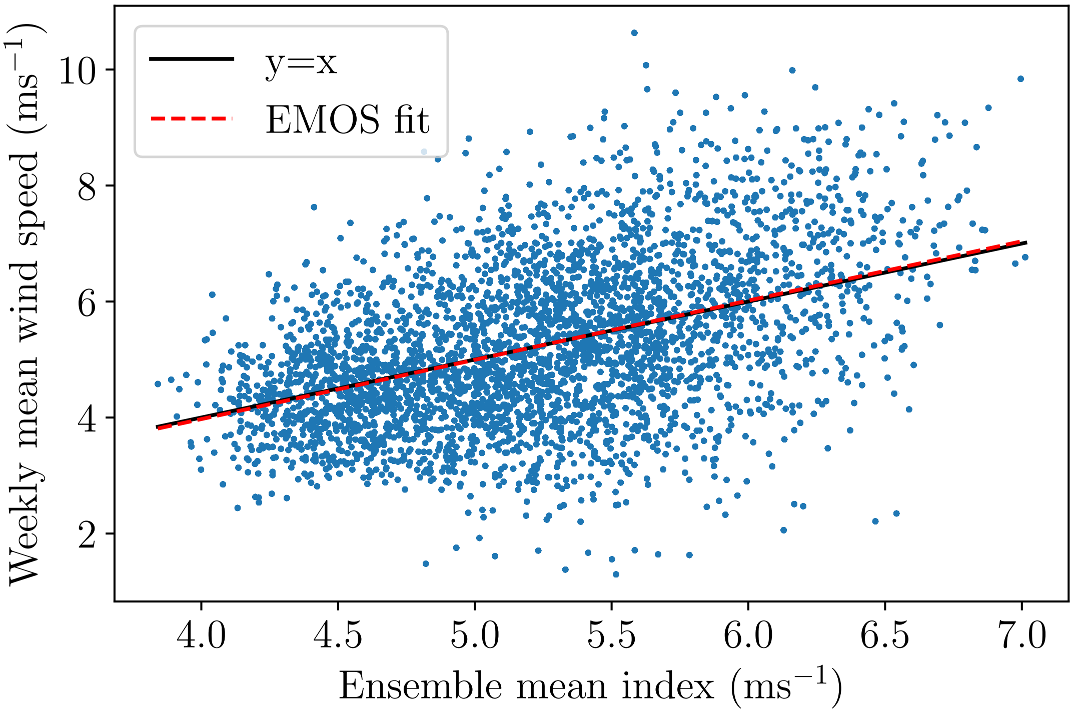


Figure 5.4: Weekly mean wind speed index forecast versus observed weekly mean forecast, for 2 weeks ahead forecasts at the MIDAS 4 station.

bers in the framework of a Generalised Additive Model of Location, Scale and Shape (GAMLSS) [254] — so all distributional parameters, not just the mean, can be modelled explicitly. The choice of parametric distribution to fit was aided by the `R` function `fitDist` from the `gamlss` package, which allows fitting of a group of possible distributions with the optimal distribution having the lowest AIC score. In `fitDist`, all distributions are fit with constant values for all parameters; this is then used to select a distribution that more closely matches the overall distribution of the target variable. For the mean wind speed and variability indices, the set of 'realplus' distributions were considered, as wind speeds (and their standard deviations) may only take real positive values. The gamma distribution was found to most closely fit the real distribution of weekly mean wind speeds and variabilities so this was used as the base parametric distribution to fit the EMOS correction model (a GAMLSS). The gamma distribution has two parameters, location μ and scale σ. First we define two metrics of the K ensemble index forecasts x_k, their mean value $\overline{x}$ and mean difference $\overline{\Delta x}$ [255]. This

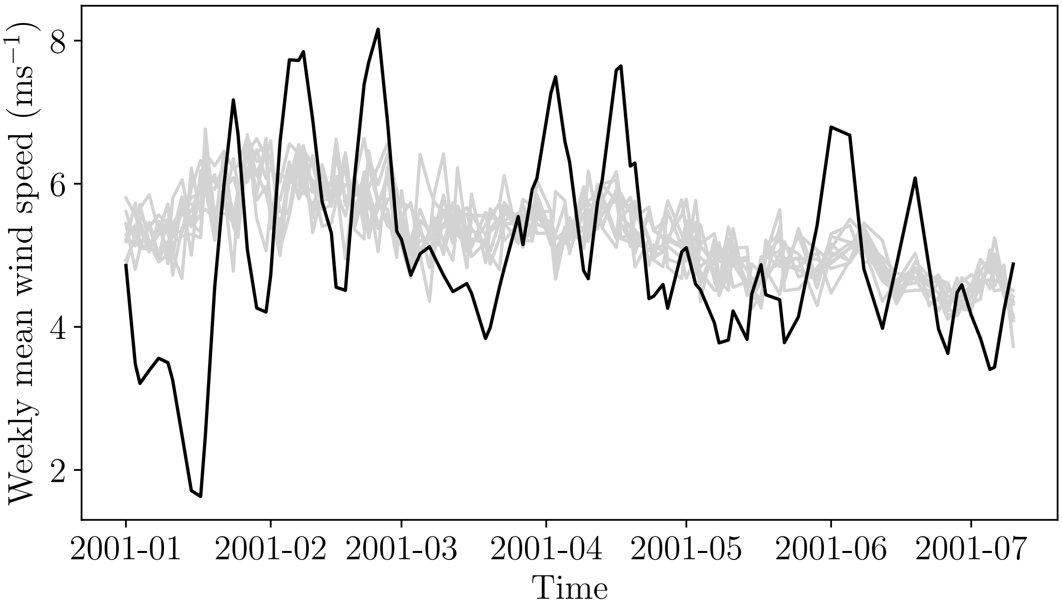


Figure 5.5: Time series of all individual ensemble member 2 weeks-ahead index forecasts (light grey), overlaid with the actual weekly mean wind speeds (black). There is clear underdispersion in the set of ensemble members.

value of mean difference takes into account the distance between each pair of ensemble members, and is less sensitive to outliers compared to measures like standard deviation.

$$\overline{x}_t = \frac{1}{K}\sum_{k=1}^{K} x_{k,t} \tag{5.9}$$

$$\overline{\Delta x}_t = \frac{1}{K^2}\sum_{k=1}^{K}\sum_{k'=1}^{K}(x_{k,t} - x_{k',t}) \tag{5.10}$$

The two gamma distribution parameters μ_t and σ_t can then be modelled as functions of $\overline{x}$ and $\overline{\Delta x}$

$$\log(\mu_t) = \beta_{\mu,0} + \beta_{\mu,1}\overline{x}_t \tag{5.11}$$

$$\log(\sigma_t) = \beta_{\sigma,0} + \beta_{\sigma,1}\overline{x}_t + \beta_{\sigma,2}\overline{\Delta x}_t \quad , \tag{5.12}$$

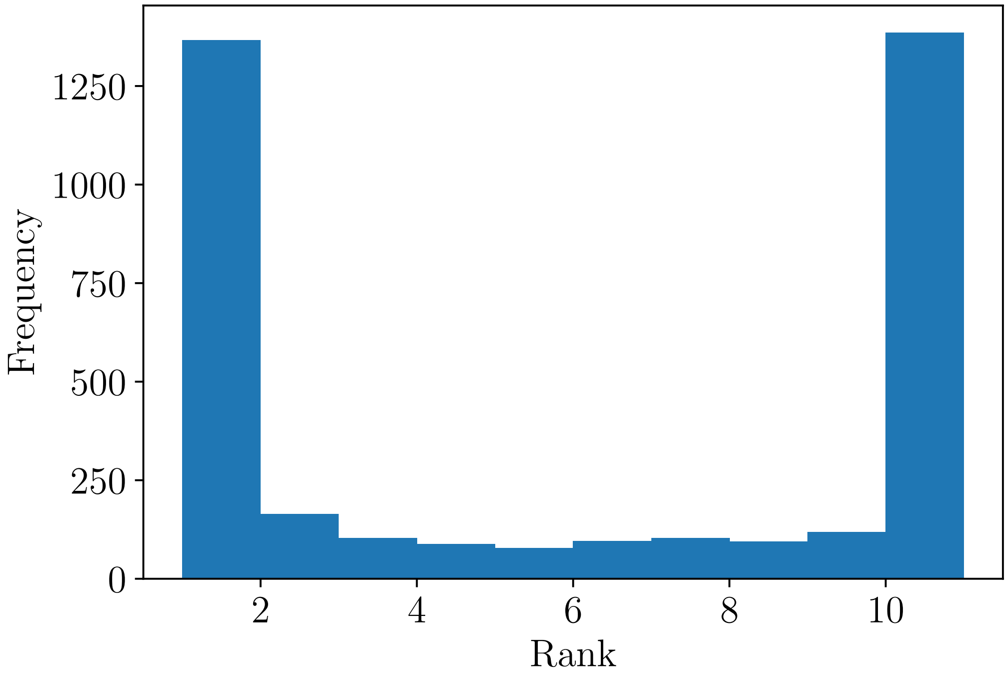


Figure 5.6: Verification rank histogram of all 2 weeks ahead ensemble member index forecasts, showing the ‘U’ shape characteristic of underdispersion.

giving the final forecast distribution

$$\hat{y}_t \sim \text{Gamma}(\hat{\mu}_t, \hat{\sigma}_t) \quad . \tag{5.13}$$

The β coefficients are estimated by maximum likelihood using the `gamlss` function . Log link functions in equations (5.11) and (5.12) ensure positive values for μ_t and σ_t.

A different distribution was needed for the weather window index forecasts, as the number of available hours in a week is bounded at zero and 168. This can be normalised to span [0,1] and will have probability masses on the bounds, as any week with a consistently high wind speed will have zero available hours and there are also weeks where the whole week is in a weather window. Therefore the zero- and one-inflated Beta distribution (BEINF) is chosen as the parametric distribution to fit in the EMOS stage for the weather window index forecasts. The BEINF distribution has 4 parameters, which in `R` are the location μ, scale σ and two parameters related to the amount of probability assigned at zero and one, ν and τ. These are modelled as functions of the same ensemble mean $\overline{x}_t$ and ensemble mean difference $\Delta\overline{x}_t$ as in Equations (5.9) and (5.10):

$$\text{logit}(\mu_t) = \beta_{\mu,0} + \beta_{\mu,1}\overline{x}_t \tag{5.14}$$

$$\text{logit}(\sigma_t) = \beta_{\sigma,0} + \beta_{\sigma,1}\overline{x}_t + \beta_{\sigma,2}\overline{\Delta x}_t \tag{5.15}$$

$$\log(\nu) = \beta_{\nu,0} \tag{5.16}$$

$$\log(\tau) = \beta_{\tau,0} \quad . \tag{5.17}$$

If the relations

$$\begin{aligned} \boldsymbol{\alpha}_t &= \mu_t(1/\sigma_t - 1) \\ \boldsymbol{\beta}_t &= (\mu_t - 1)(\sigma_t - 1)/\sigma_t \\ p_0 &= \nu/(\nu + \tau + 1) \\ p_1 &= \tau/(\nu + \tau + 1) \end{aligned} \tag{5.18}$$

are used to convert from location and scale parameters used in `R` to those more commonly used to describe the beta distribution and the boundary probabilities, the final forecast distribution is given by

$$\text{BEINF}(x|\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, p_0, p_1) = p_0\delta(x) + (1 - p_0 - p_1)f_B(x|\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) + p_1\delta(x - 1) \tag{5.19}$$

where $0 \leq x \leq 1$ and $p_0 + p_1 \leq 1$.

5.2.6 Benchmark climatology

It is important to benchmark any proposed method to be able to quantify its performance relative to other methods. For seasonal forecasts, the most common benchmark is climatology — where the forecast is an average of historic values for that particular time of year, in this case month of the year. We use the time points equivalent to the range of S2S hindcasts (i.e. 1999–2019) to produce the distribution of target values (e.g. weekly mean wind speed) for each month of the year — this is then the forecast climatology for the forecast times matching those we produce model forecasts for from the S2S forecasts. Climatology forecasts are produced for both the MIDAS data and the corrected ERA5 site data so S2S forecasts trained on each may be compared to climatology produced with the same data. Anecdotally, industry users who work on-site use their personal experience of past site conditions at different times of year to inform their work and decisions — this could be described as an informal type of climatology informed decision.

5.3 Results and analysis

5.3.1 Mean wind speed forecasts trained on MIDAS data

Firstly, two main variations to the model configuration are tested: grouping the weather variables related to 100m wind speed in one Principal Component transform, and calculating separate sets of PCs for each of these variables. Secondly, we have to choose the features derived from these PCs that are used as inputs to the index model: while PCs and MIDAS data/site measurements are of hourly resolution, index forecasts are produced for weekly intervals. All model configurations take weekly mean values of the PCs as inputs, but it is also possible to provide other features such as weekly standard deviation and minima/maxima for each PC. This may be more relevant when modelling certain metrics such as measures of variability rather than mean values. All these modelling configurations were tested for all the MIDAS sites, with very similar results displayed at all sites and shown for one location in Figure 5.7. This and all following skill scores are calculated relative to climatology. Confidence intervals are calculated through bootstrap resampling as detailed in Section 2.6.5. This shows no significant difference in skill between model configurations, a result which holds across all the MIDAS sites. This initial evaluation of skill also indicates skill above zero up to and including three weeks ahead, which is promising. The metric of Pinball loss used to calculate these skill scores is a function of both the reliability and sharpness of the forecasts. However, a calibrated forecast with a slightly lower skill score is still preferable over forecast with a better skill score but that is not calibrated. The Pinball score is chosen as it is a proper score, i.e. the minimum score coincides with the optimal estimate of that quantile. As it is estimated per quantile, it is possible to investigate the relative performance of the forecasts across the distribution as well as overall performance across a set of quantiles. For example, figure 5.12 shows skill relative to climatology for the more extreme quantiles estimated as well as skill in the central part of the forecast distribution.Calibration (or reliability) is displayed in Figure 5.8. This shows that all variations of linear model tested display very similar reliability to each other. Since using the 100m PCs together and not including extra features results in a faster runtime, this model is adopted for the weekly mean wind speed index. However, the extra features are always included for a variability index forecast as the extra features are directly related to inter-week variability which is the target for this index. The exact features and numbers of principal components used are given in Table 5.2.

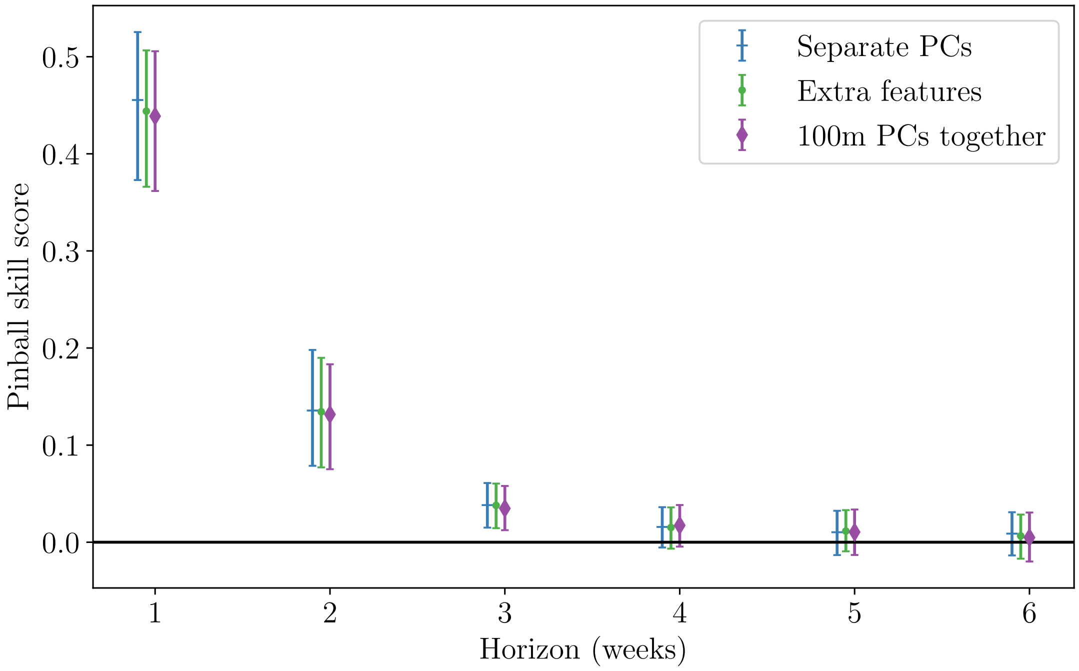


Figure 5.7: Pinball skill score of weekly mean wind speed index, at MIDAS 1 and relative to climatology. The three index model configurations are with separate Principal Components, with extra weekly features (standard deviation, min and max) of the PCs inputs, and where all 100m weather variables are transformed into PCs together. Error bars show 95% bootstrapped confidence intervals.

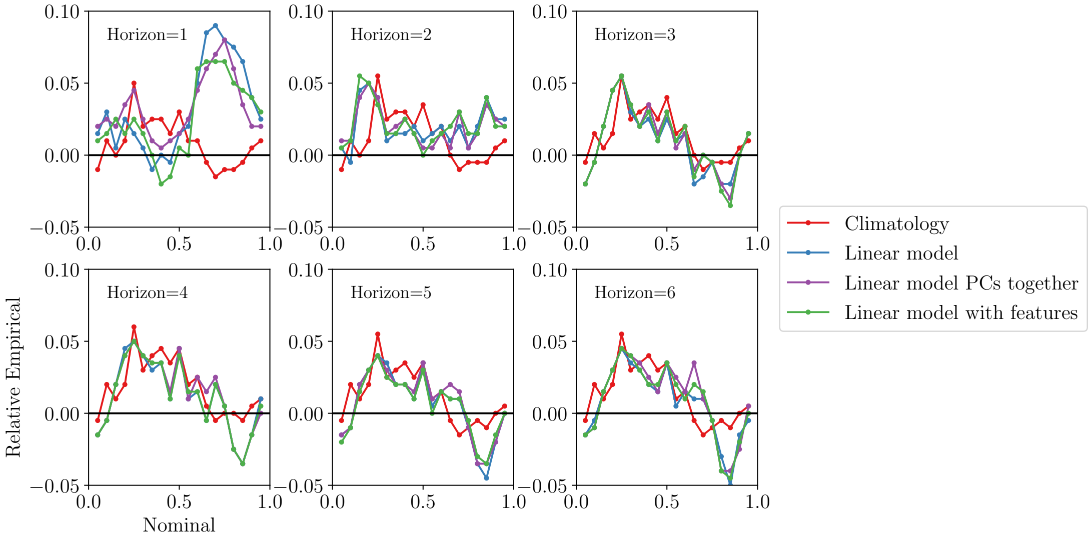


Figure 5.8: Normalised reliability diagram for mean wind speed index and the equivalent climatology forecast, at MIDAS 1. The forecast labelled 'linear model' is calculated with separate PCs and no extra engineered features. Horizons are in number of weeks.

5.3.2 Mean wind speed forecasts trained on ERA5 data

It is assumed that the optimal model configurations for forecasts trained on MIDAS data (grouping 100m windspeed-related variables and only including inter-week standard deviation as a feature for variability forecasts) also holds for forecasts trained on ERA5 data. The skill of the forecasts trained on corrected ERA5 data (for the actual wind farm sites) is presented grouped by location, with area 1 shown in Figure 5.9 and the remaining three areas in Appendix D. Forecasts with skill $0 < x < 0.15$ are described as 'fair', $0.15 < x < 0.3$ are 'good' and > 0.3 are 'very good'. This follows the descriptions for skill levels given by the S2S4e project's decision support tool [217][2]; however, these are purely indicative based on expert judgement and the skilfulness of a forecast depends on its usefulness in decision making. All the wind farm site forecasts show very good skill one week ahead, fair skill between 0.1 and 0.15 2 weeks ahead and continuing fair skill 3 weeks ahead. The ERA5 forecasts show similar skill to the MIDAS forecasts, with slight increases in skill over the MIDAS forecasts for areas 2 and 4 for the shorter horizons. Reliability diagrams for the MIDAS 1 site and Area 1

[2] \gls{s2s}4e-dst.bsc.es/#/

(based on ERA5 data at WF1) are shown in Figures 5.10 and 5.11 respectively. On these plots, a perfectly reliable forecast would lie on the line $y = 0$. Both climatology and linear model forecasts are reliable across the forecast distribution, the only exception being some of the higher quantiles in the one week ahead linear model forecast. This horizon is not the main focus for S2S decision making and typically, alternative data sources and forecasting methods (based on a standard NWP forecast) would be used for this shorter horizon so this slight lack of calibration is not hugely important.

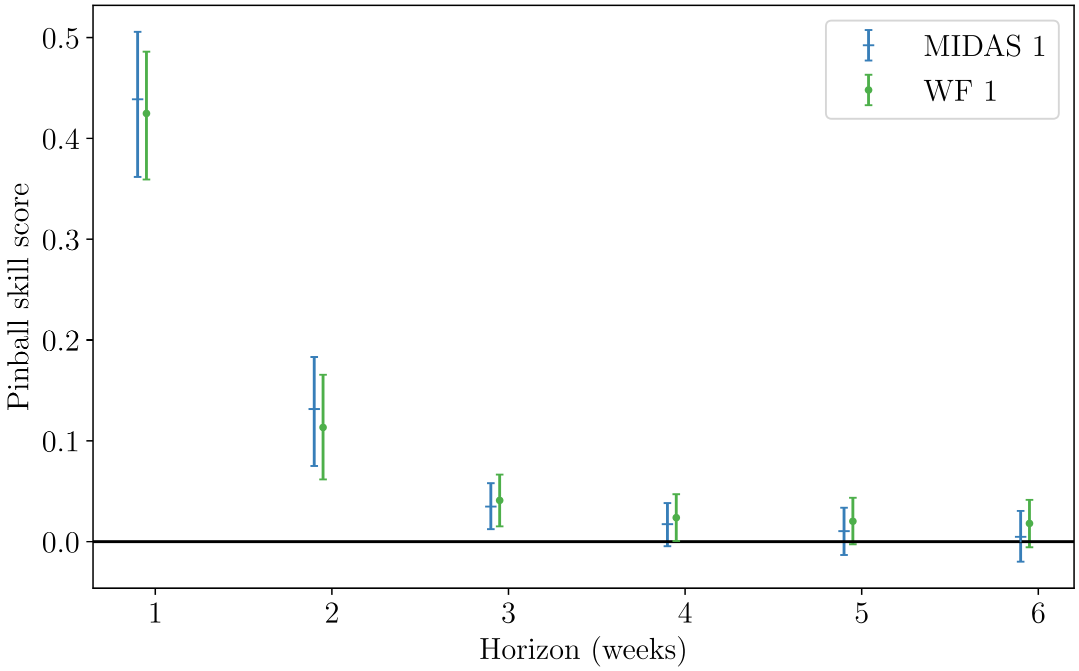


Figure 5.9: Pinball skill score of weekly mean wind speed index relative to climatology. The forecasts labelled 'WF 1' are based on corrected ERA5 data for that wind farm. Error bars show 95% bootstrapped confidence intervals. MIDAS skill score is calculated relative to climatology of the MIDAS wind speed data and WF1 skill score is calculated relative to climatology of the corrected ERA5 data.

The Pinball skill scores shown in Figures 5.7 and 5.9 are an average of the Pinball score across the forecast distribution. However, there may be more skill relative to climatology in different parts of the distribution, for example the centre vs the tails. Climatology is calculated for both MIDAS and corrected ERA5 data so each index model is compared to climatology that has been produced from the same data. Figure 5.12

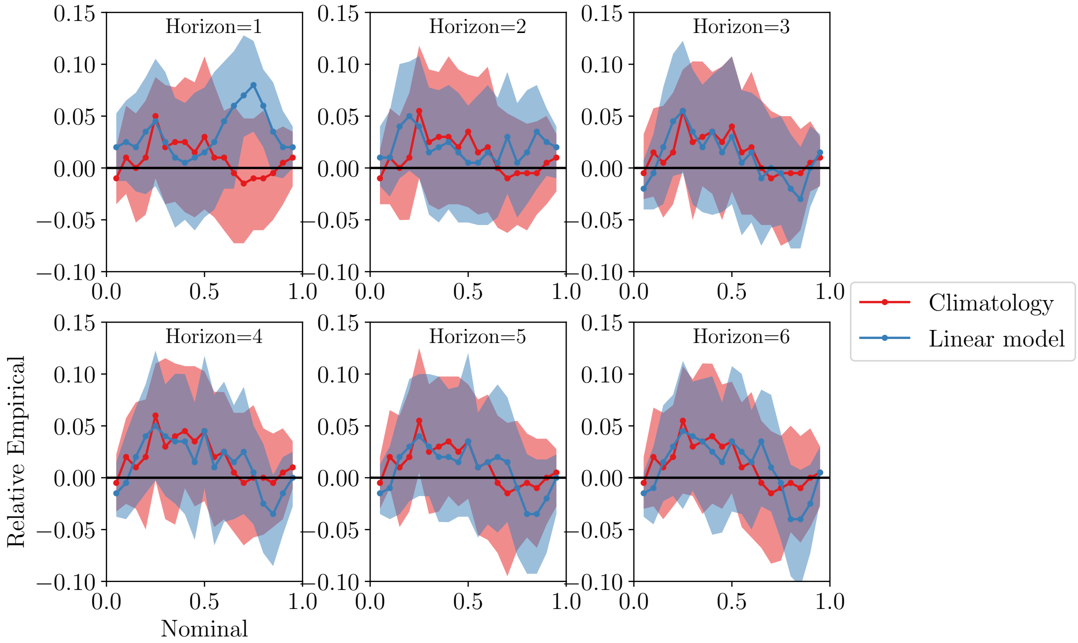


Figure 5.10: Reliability of weekly mean wind speed forecast at MIDAS 1. Intervals show 95% bootstrapped confidence bands.

shows the skill for a 2 week ahead forecast for each quantile level of the distribution. This is still a skill score relative to climatology rather than absolute Pinball values, and shows there is skill across the whole distribution with only slightly decreased skill at the most extreme (q5 and q95) quantiles assessed.

For the geographically close groups of sites at areas 2 and 4, it is possible to share resources such as cranes, parts and people easily between sites. As such, it is useful to understand any correlations in forecast performance between these sites. For example, when forecasts are underpredicting at site A, what does that likely mean for forecast performance at site B? To examine this, the realisation value is passed through the inverse cumulative distribution function to get the forecast probability of that value occurring. These 'realisation probabilities' should be uniformly distributed if the forecasts are calibrated. Because they indicate where in the distribution the real value fell, they can show times of under or overforecasting. A scatter plot of these values shows the correlation in the forecast bias between neighbouring sites. Figure 5.13 shows strong positive correlations between sites, indicating similar simultaneous forecast bi-

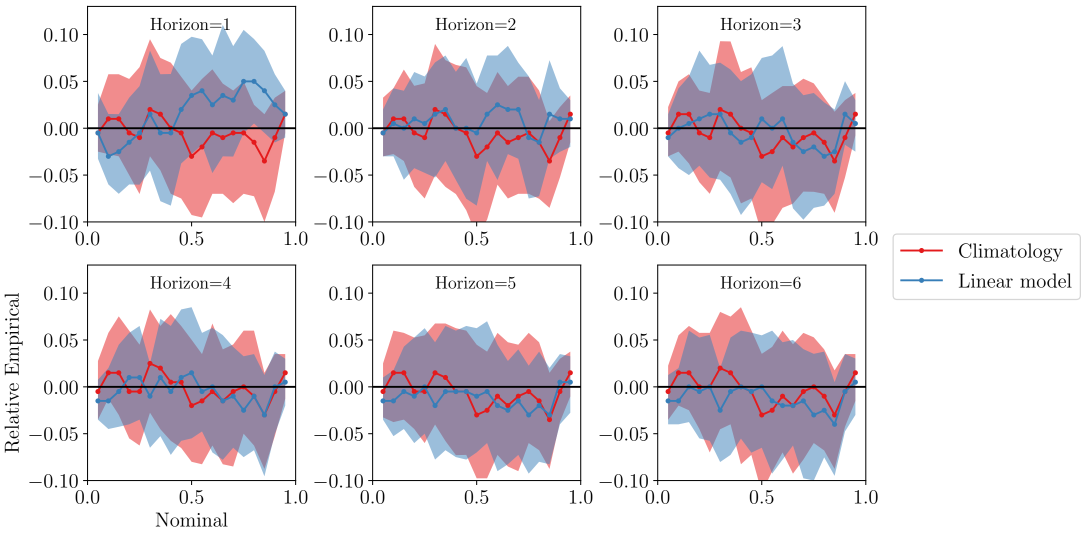


Figure 5.11: Reliability of weekly mean wind speed forecast at WF1. Intervals show 95% bootstrapped confidence bands.

ases. The least strong correlations (site 4a with 4b and 4c) also correspond to the largest geographical distance between wind farms.

5.3.3 Variability indices

The variability index is trained on the standard deviation of measured hourly wind speeds within the week. Figure 5.14 shows the skill of variability index forecasts for area 1. As for mean wind speed, the MIDAS-based and ERA5-based forecasts give very similar skill, except for area 2 at shorter horizons. In general, the skill of the variability forecasts is lower than that of mean wind speed forecasts for 1-2 week ahead forecasts; however, some skill persists to the longer (4-6 weeks ahead) horizons at all sites. Perhaps this is driven by a relationship between large scale weather patterns and variability — for example the NAO phase (the sea level pressure difference anomaly between Iceland and the Azores) is related to the number and strength of winter storms in Europe.

The reliability diagrams for variability forecasts again for both MIDAS (Figure 5.15) and ERA5 at the wind farm (Figure 5.16) show generally calibrated forecasts that don't display any consistent forecast bias or over or under-dispersion. The one week ahead

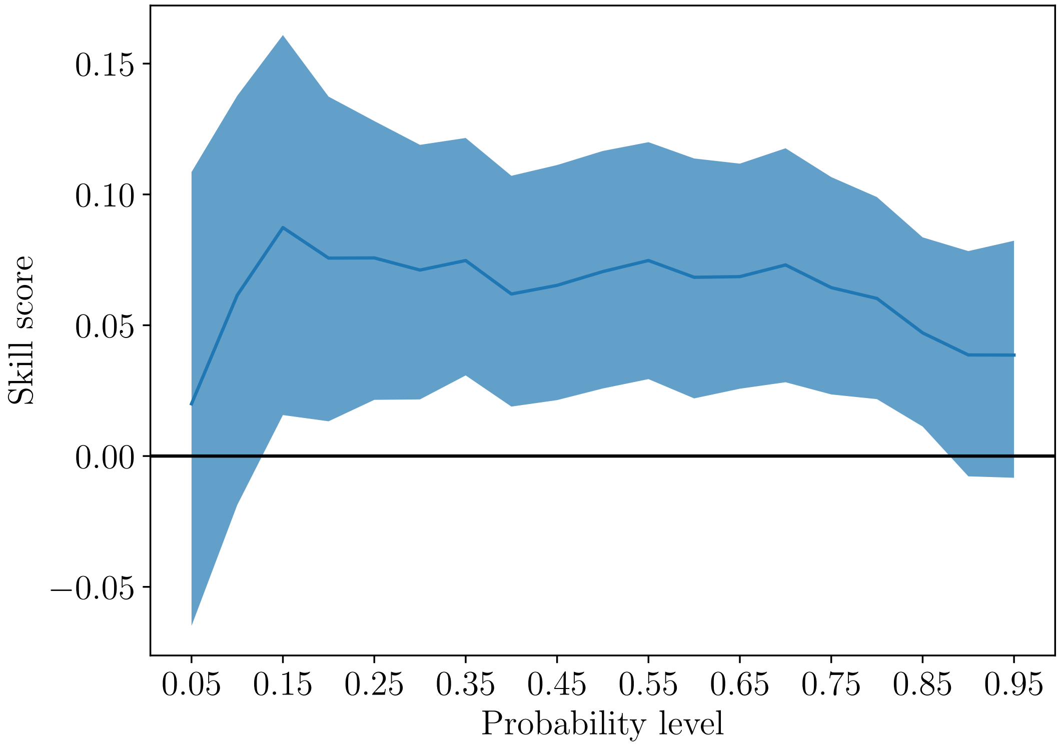


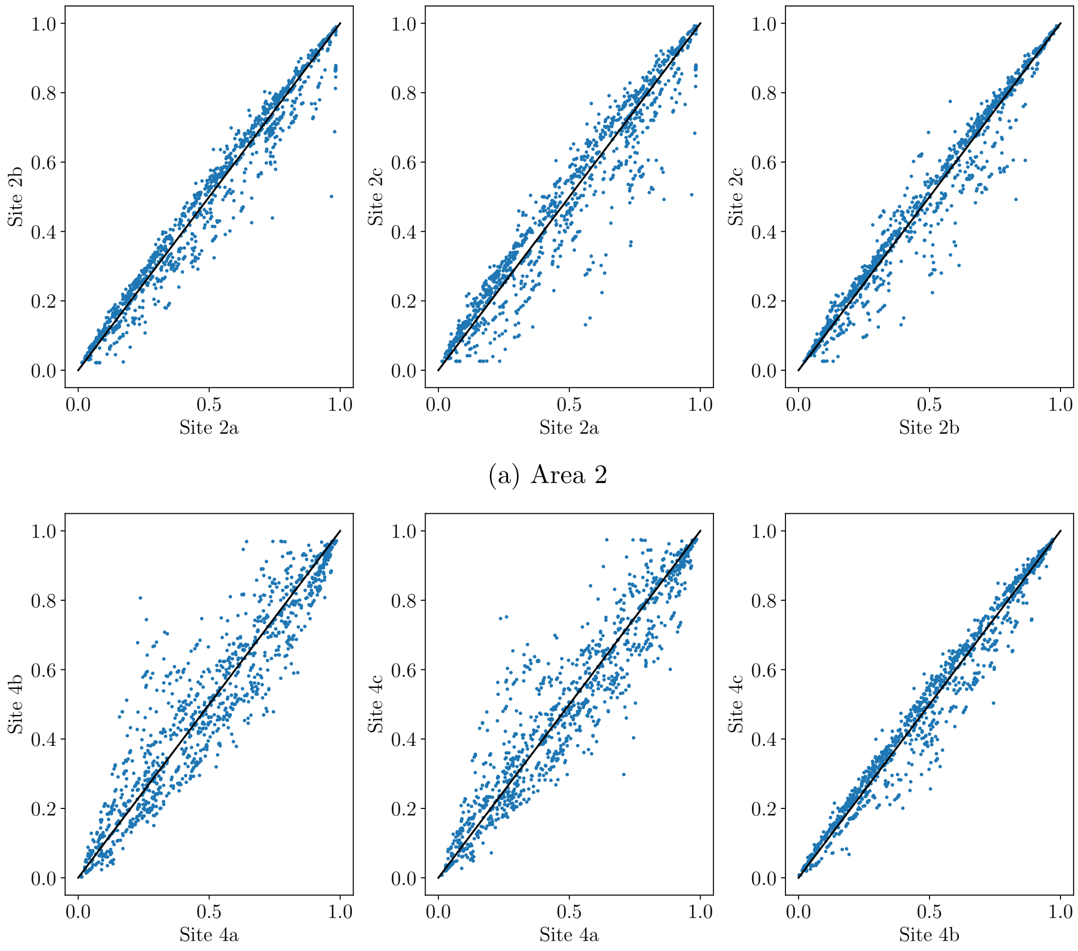
Figure 5.12: Pinball skill score of weekly mean wind speed index relative to climatology at WF 1, across the forecast distribution. The two week ahead forecast is shown here.

horizon again tends to show the greatest departure from perfect reliability. These observations apply to all locations, with the plots for other areas presented in Appendix C.

5.3.4 Weather window indices and economic value of forecasts

The weather window indices and subsequent cost-loss analysis is based on an example where a crane is needed for a maintenance task. The safe wind speed threshold for crane use depends on the crane model and object being lifted, but is assumed to be 7 ms^{-1} in this work[3]. It is also assumed that a minimum of an 8 hour window is needed to complete a maintenance task. As an indicative figure, in the 20 years of hindcast data at one site, 50.8% of time points were below the 7 ms^{-1} threshold. Of the 5003 unique weather windows, 2582 of them lasted 8 hours or longer.

[3] Exact limits are tied to the specific job and equipment and are set out in the 'authorised work procedure'. Interviews with site operations teams indicated typical crane lift limits are between 5 and 8 ms^{-1}.



(a) Area 2

(b) Area 4

Figure 5.13: Scatter plots of inverse probability of realisations. The solid black line shows y=x.

A fan plot of the weather window index (Figure 5.17) shows the forecasts follow seasonal trends well and do also have some sensitivity to shorter term spikes although the forecast distribution is generally quite broad. The weather window index forecasts have very good skill at all sites for 1 week ahead, fair skill for 2-3 weeks ahead and some skill at most sites beyond that (Figure 5.18). All forecasts are calibrated across all horizons with the exception of the 1 week ahead forecasts for the area 4 sites.

To perform a cost-loss calculation, the form of the losses incurred for the crane maintenance problem must be determined. It is assumed that any turbines that need

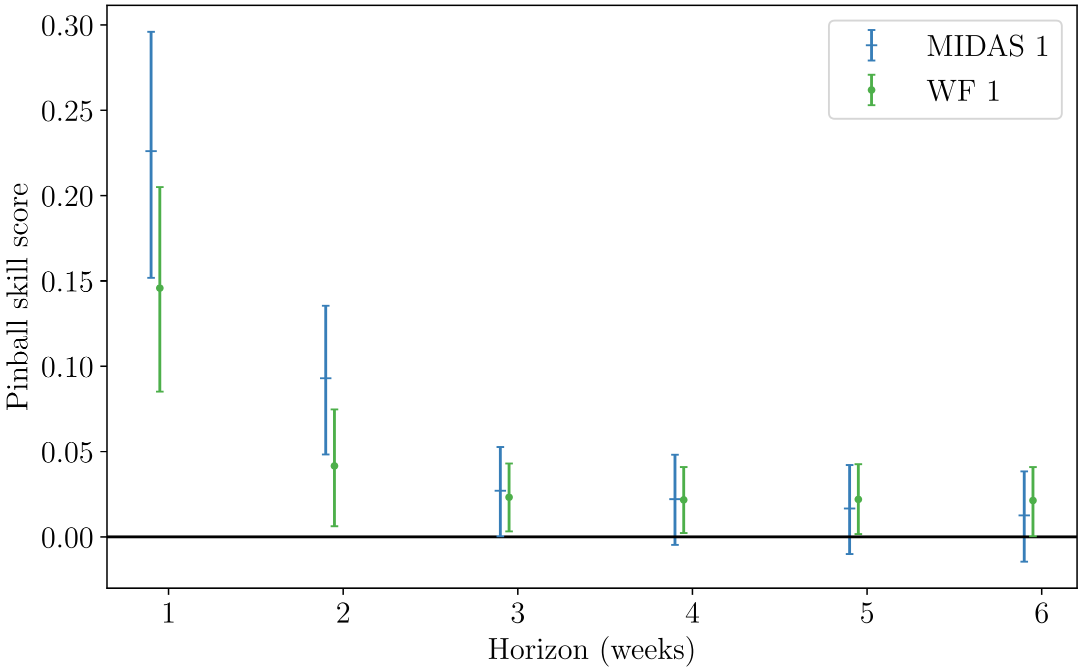


Figure 5.14: Pinball skill score of weekly variability (weekly standard deviation of hourly wind speeds) relative to climatology. The forecasts labelled ‘WF 1’ are based on corrected ERA5 data for that wind farm. Error bars show 95% bootstrapped confidence intervals.

to be fixed are currently not operational, so that losses consist of the lost energy whilst the turbine is down. Then the expectation value of lost energy for the week can be calculated. An alternative assumption would be that the turbines can continue to operate before the repair, in which case lost energy is only incurred from the downtime while the repair takes place. The expected lost energy is given by

$$\int_0^1 L(x)p(x)\ dx \tag{5.20}$$

where x denotes the number of useful window-hours in the week, $L(x)$ is the loss function and $p(x)$ is the forecast distribution. $L(x)$ represents the losses incurred when a turbine is not fixed; in this case, this is assumed to be the value of lost energy. To calculate this, the hourly resolution site wind speeds are passed through a turbine power curve before being summed into a value for weekly lost energy per turbine. Figure 5.20 shows the relationship between number of useful hours x and the lost energy for that

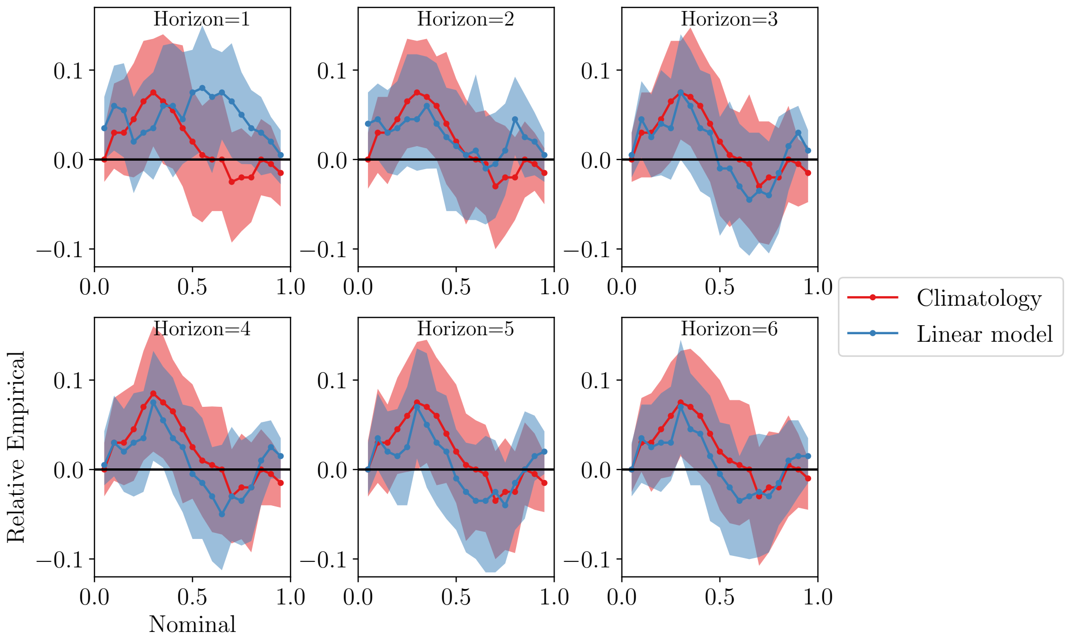


Figure 5.15: Reliability of variability forecast at MIDAS 1. Intervals show 95% bootstrapped confidence bands.

week $L(x)$. This relationship is modelled as quadratic, so the loss function becomes

$$L(x) = ax^2 + bx + c \tag{5.21}$$

and the expected lost energy (Equation 5.20) becomes

$$\int_0^1 \left(ax^2 + bx + c\right) \left[p_0\delta(x) + (1 - p_0 - p_1) f_B(x|\boldsymbol{\alpha}, \boldsymbol{\beta}) + p_1\delta(x - 1)\right] dx \tag{5.22}$$

where f_B is the beta distribution with parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\delta(x)$ is the dirac delta function centred at 0. Using the properties $\int_{x_1}^{x_2} f(x)\delta(x-a) = 0$ when $a \leq x_1$ or $a \geq x_2$ and the recursion relation for the beta function$B(\boldsymbol{\alpha}+1, \boldsymbol{\beta}) = \boldsymbol{\alpha} B(\boldsymbol{\alpha}, \boldsymbol{\beta})/(\boldsymbol{\alpha} + \boldsymbol{\beta})$, Equation 5.22 evaluates to

$$(1 - p_0 - p_1)\left[\frac{a\ \boldsymbol{\alpha}_t(\boldsymbol{\alpha}_t + 1)}{(\boldsymbol{\alpha}_t + \boldsymbol{\beta}_t)(\boldsymbol{\alpha}_t + \boldsymbol{\beta}_t + 1)} + \frac{b\ \boldsymbol{\alpha}_t}{\boldsymbol{\alpha}_t + \boldsymbol{\beta}_t} + c\right] \quad . \tag{5.23}$$

Here a, b and c are estimated from a quadratic fit to the relationship shown in Fig-

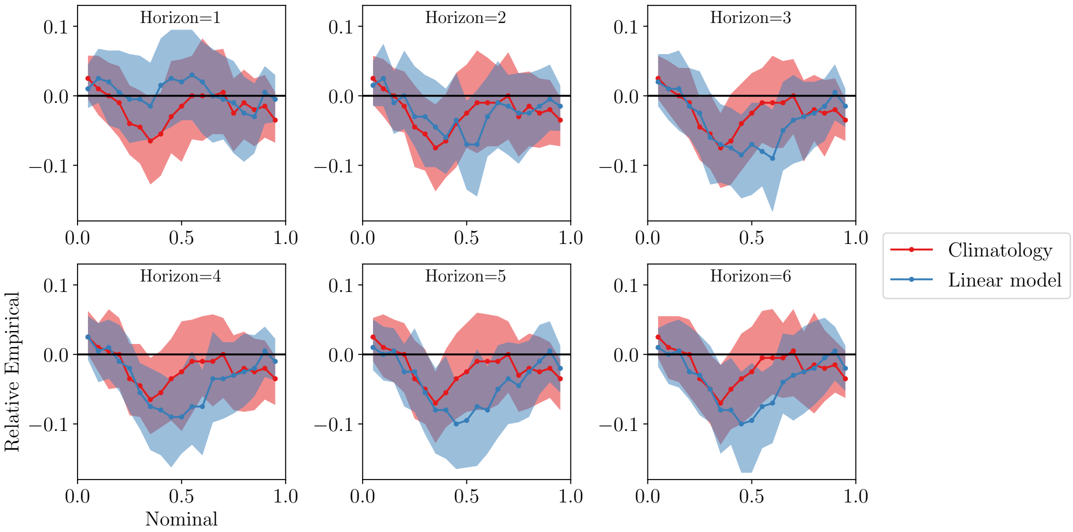


Figure 5.16: Reliability of variability forecast at WF1. Intervals show 95% bootstrapped confidence bands.

ure 5.20. μ_t, σ_t, ν and τ are the BEINF parameters estimated through GAMLSS model fitting following Equations 5.14 – 5.17 and are then converted to p_0, p_1, $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ using the relations in Equation 5.18.

Figure 5.21 shows the various possible scenarios, given three turbines need to be repaired. The expected cost from not hiring the crane includes the lost energy for the week plus a terminal penalty given by the average energy loss in subsequent weeks before the next opportunity to repair all remaining turbines. If a crane is hired, there are several different numbers of turbines that could get fixed depending on the actual weather conditions that week. The end cost of a scenario where n turbines are fixed and m turbines remain down is the cost of energy for the portion of the week before the n turbines are fixed, plus the lost energy for the entire week for the m turbines that aren't fixed, plus the cost of the crane, plus a final terminal penalty of the average energy loss in subsequent weeks before the next opportunity to repair the m remaining turbines. The overall expected cost of hiring the crane is the sum of these scenarios, weighted by their probabilities of occurrence. These probabilities of occurrence are derived from the forecast distribution of x, the number of useful hours in the week. Then a decision on whether to hire a crane may be made, based on which option has

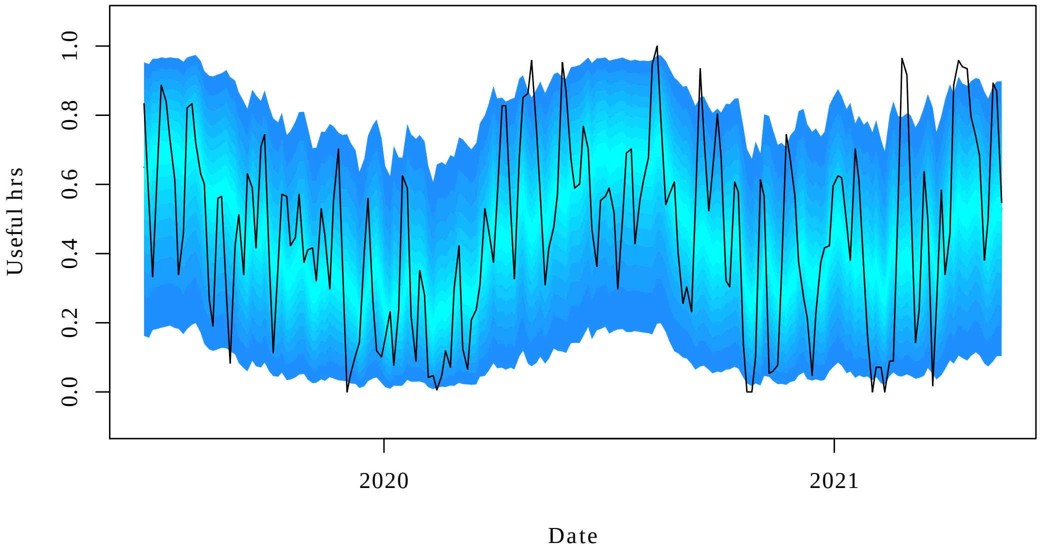


Figure 5.17: Fan plot of 2 weeks ahead weather window index forecasts

the lowest expected cost given the forecast. This decision is sensitive to two costs: the cost of crane hire, and the cost of lost power production, which depends on electricity price. The expected costs of hiring or not hiring are plotted in Figure 5.22. All points below the $y = x$ line represent times where hiring a crane is cheaper than not hiring one.

At very high electricity prices, the expected cost of lost energy is always greater than the expected costs incurred if a crane is hired (the curve in Figure 5.22 is shifted so it lies entirely below the $y = x$ line), so the decision is always to hire the crane. Conversely at very low electricity prices, the cost of lost energy is much less than the cost of the crane and so it is never beneficial to hire the crane. In reality, there may be other contractual penalties for failure to repair any turbines but this is not included in the cost loss model used here, as the financial cost of this may be linked to maintenance of the whole wind farm across a whole season rather than individual maintenance decisions. For the cost-loss model in this example, for a given crane hire cost, there exists a range of electricity prices where the decision to hire or not hire depends on the forecast number of useful

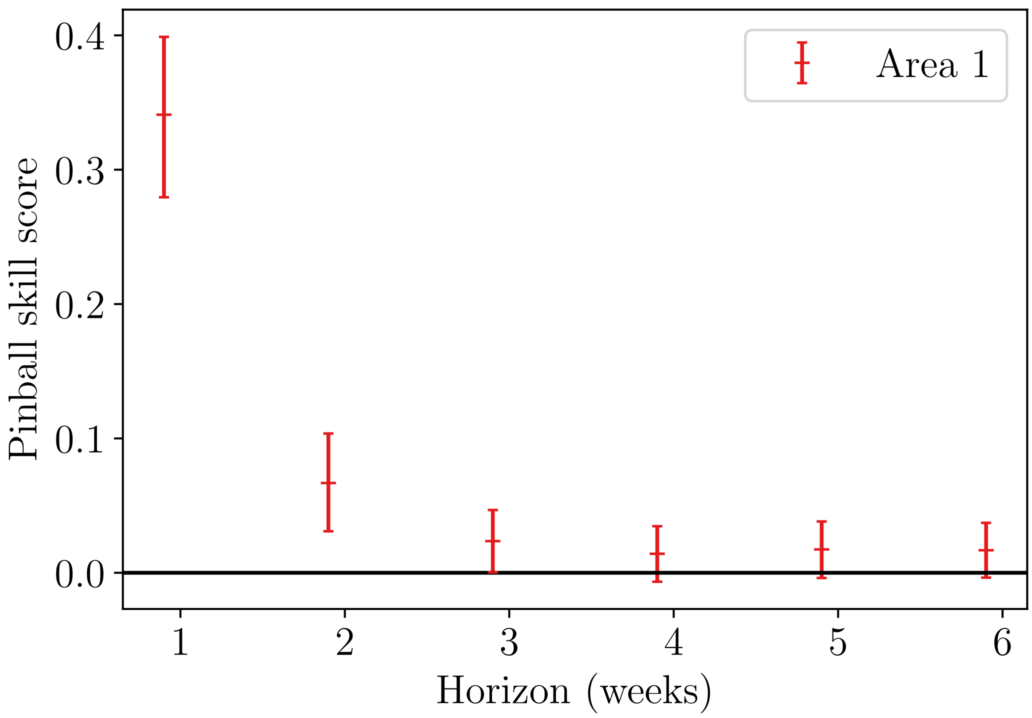


Figure 5.18: Pinball skill score of weather window index relative to climatology. for area 1.Error bars show 95% bootstrapped confidence intervals.

hours in the week (equally, for a set electricity price there would be a range of crane hire prices where this also holds). The range of electricity prices where the decision to hire or not hire a crane is sensitive to the forecast number of useful hours is shown in Figure 5.23. Further analysis is based on WF3, where the range of electricity prices is the greatest.

The true cost of crane hire decisions is evaluated by comparison to what the actual weather would have allowed. The number and timings of actual weather windows in the ERA5 site wind speed data give the first available times each turbine may be fixed and the corresponding actual lost energy before those times. Then the 'actual' cost of hiring or not hiring can be calculated. The total cost calculated assumes there are always either 2 or 3 turbines down, across all 200 weeks evaluated. However, in reality this scenario will only apply for a portion of weeks, and the more relevant metrics are the number of turbines fixed and the average cost per turbine. These metrics are plotted in Figures 5.24 and 5.25 respectively. As might be expected, the total number

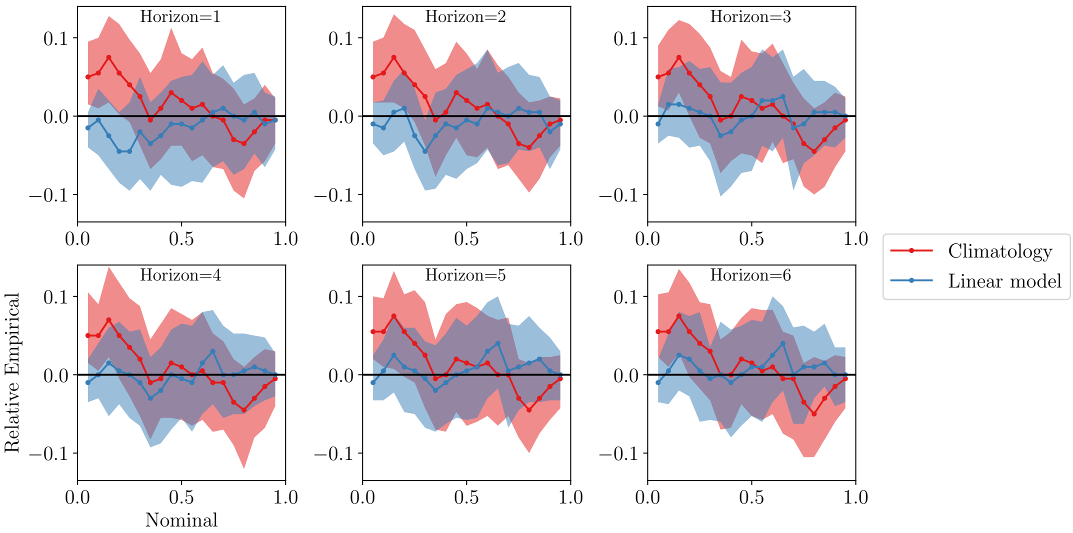


Figure 5.19: Reliability of weather window index forecast at WF1. Intervals show 95% bootstrapped confidence bands.

of turbines fixed increases with electricity price as the decision to hire a crane is made a higher proportion of the time when the cost of lost energy is higher. There is no clear difference in number of turbines fixed between decisions based on climatology or the S2S index forecasts. Again there is little difference between climatology or S2S forecasts in cost per turbine, apart from at the lowest electricity price considered where no turbines were fixed under climatology. Above £60/MWh for 3 turbines or £90/MWh for 2 turbines, the cost per turbine is relatively stable. Comparing the decisions made at each timestamp by the climatology or S2S index forecasts, both forecasts made the same hiring decision at 189 of the 200 weeks analysed, which perhaps explains why the costs associated with both forecasts are so similar.

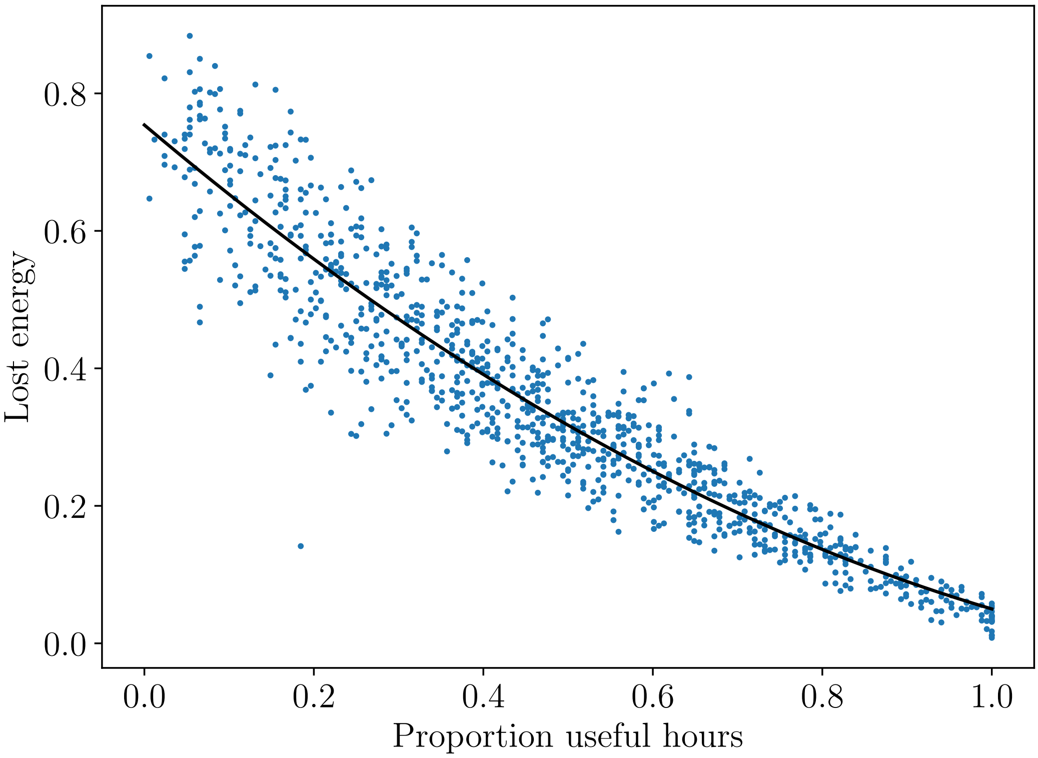


Figure 5.20: Relationship between number of hours in a weather window and weekly total lost energy at WF1. Both axes are normalised by their maximum (i.e. number of hours in a week and maximum weekly energy output).

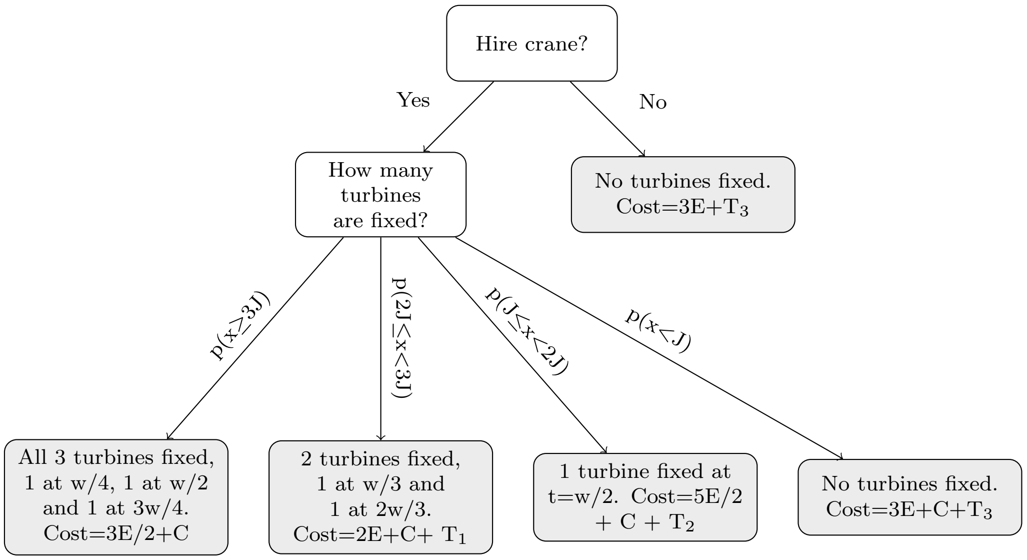

Hire crane?
Yes
No
How many turbines are fixed?
No turbines fixed. Cost=3E+T3
p(x≥3J)
p(2J≤x<3J)
p(J≤x<2J)
p(x<J)
All 3 turbines fixed, 1 at w/4, 1 at w/2 and 1 at 3w/4. Cost=3E/2+C
2 turbines fixed, 1 at w/3 and 1 at 2w/3. Cost=2E+C+ T1
1 turbine fixed at t=w/2. Cost=5E/2 + C + T2
No turbines fixed. Cost=3E+C+T3

Figure 5.21: Flowchart of crane hire decision and resulting costs, assuming three turbines are broken. The week starts at time t=0 and ends at time t=w. E is the expected cost of lost energy for one turbine for the whole week, given the weather window index forecast (see Equations 5.20 - 5.23). C denotes the cost of crane hire for the week, J denotes the job time for repair of one turbine and T_m is a terminal cost representing the average energy loss in subsequent weeks before the next opportunity to repair the m unfixed turbines. x is the number of useful hours. Where one or more turbines are repaired, the times of repair are assumed to be divided evenly throughout the week.

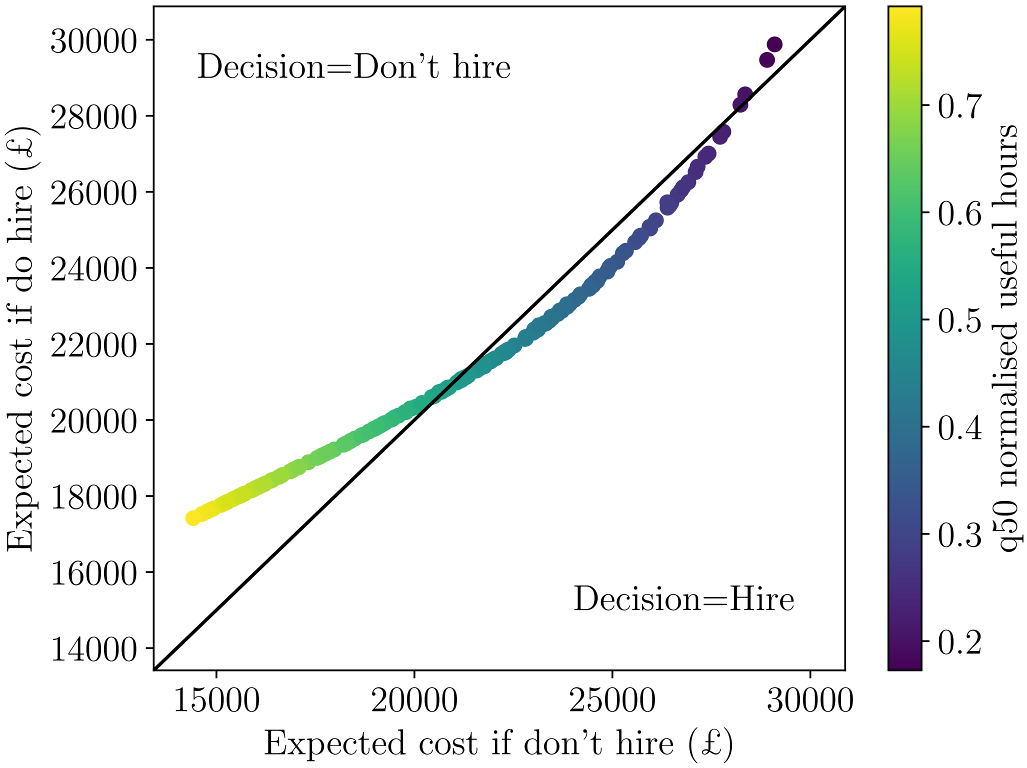


Figure 5.22: Expected cost of hiring crane, vs expected cost of not hiring crane, for 2-week ahead forecasts of number of available hours in the week. Colour shows the q50 forecast value of normalised number of hours within a weather window.

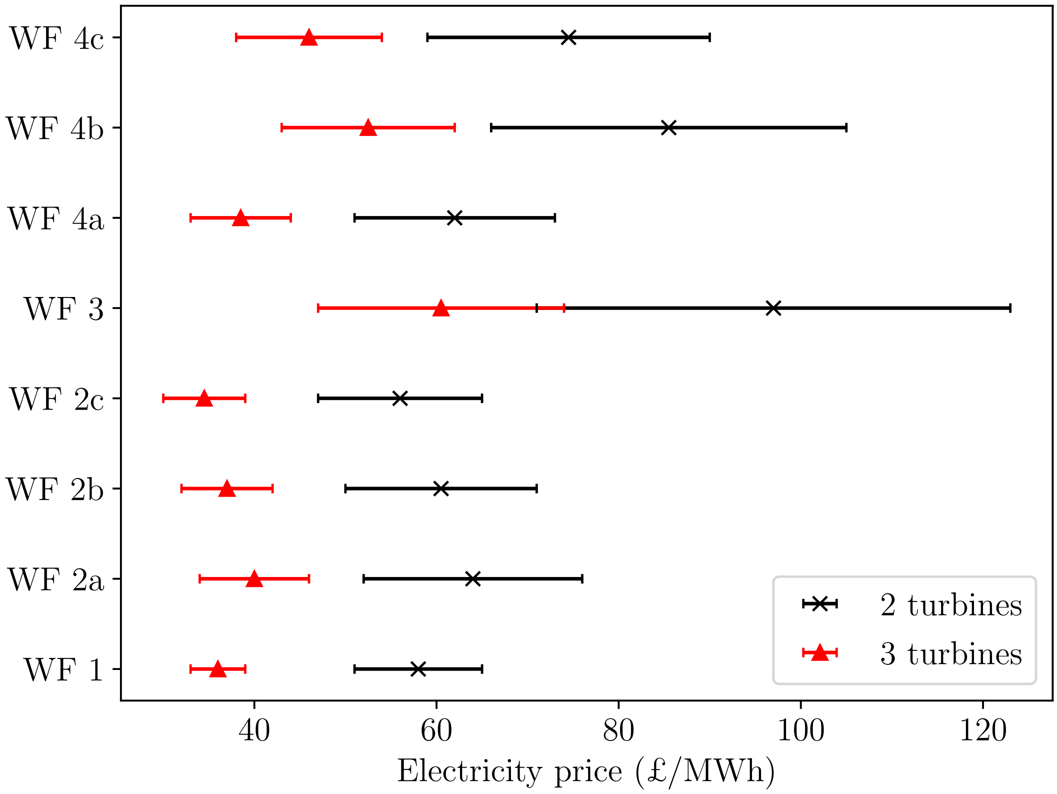


Figure 5.23: Range of electricity prices at each site where crane hire decision is sensitive to the forecast number of weather window hours, for either 2 or 3 turbines down. Marker shows the median price and whiskers show the minimum and maximum electricity prices where the hire decision is sensitive to the forecast.

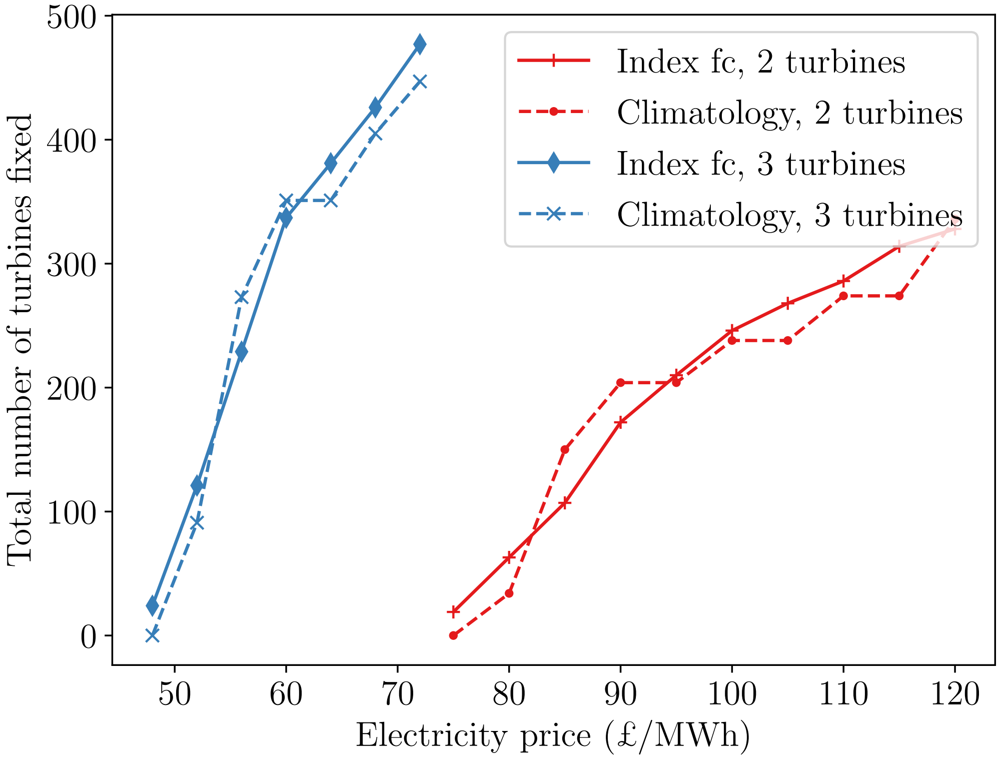


Figure 5.24: Number of turbines fixed when index forecasts or climatology forecasts are used to make crane hiring decision, dependent on electricity price, at WF3.

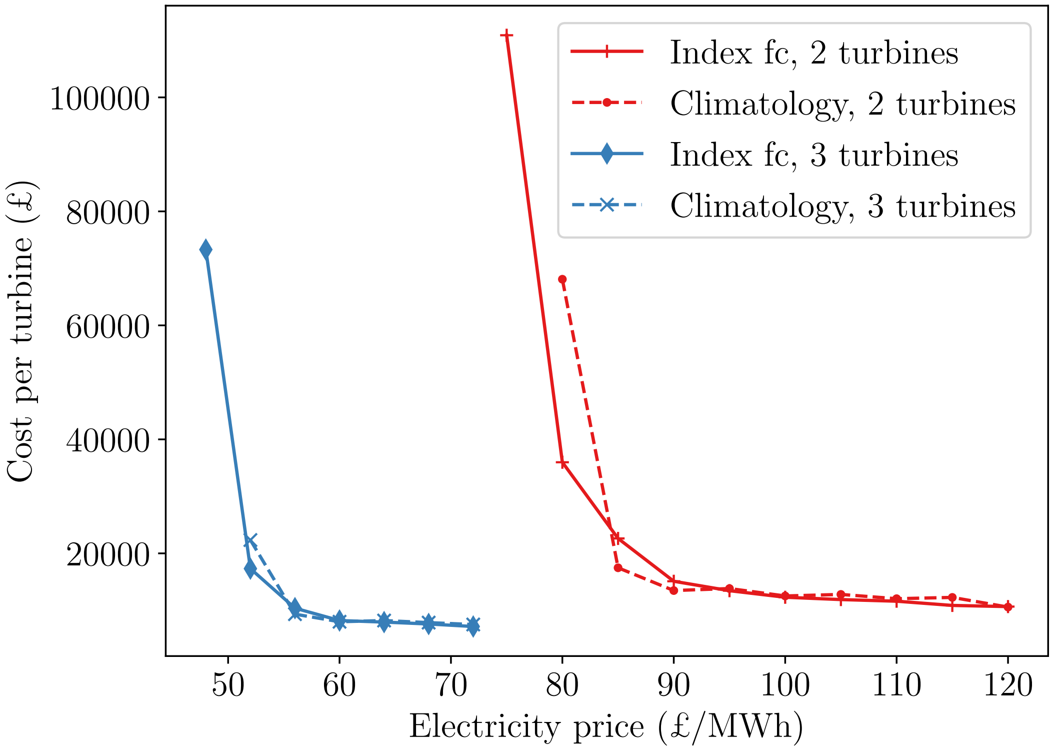


Figure 5.25: Cost of crane hire and lost energy when index forecasts or climatology forecasts are used to make hiring decision, for a range of electricity prices, at WF3.

5.4 Conclusions

Forecasts of weekly mean wind speed, weekly wind speed variation and number of useful work hours within the week have been shown to be skilful on S2S timescales. While this has been demonstrated for sites in Scotland, the methodology presented could be applied to any location. Skill would have to be checked on a site-by-site basis but it is hoped that the presence of skill shown in this case study would also be found at other locations. Future improvements in the accuracy, resolution and relevant variables available (e.g. 100m wind speed) from S2S forecast products may be expected to further increase the skill of these forecasts. Mean wind speed forecasts showed skill out to three weeks ahead. Variability forecasts had lower skill for one or two week horizons, but retain fair skill out to 6 weeks ahead. A new 'weather window index' of the number of useful hours in a week was proposed to inform activity-specific decision making, for example crane hire. This new useful-hours metric could be applied for any problem where an activity depends on a weather condition threshold and requires a minimum amount of time, such as crew transfers for offshore wind which depends on significant wave height. These weather window index forecasts showed very good skill one week ahead and fair skill out to 6 weeks. A cost-loss model was implemented to determine the economic benefit of forecasts, but found little difference between the decisions made under climatology and the decisions made by the S2S forecasts. The availability of historical S2S forecasts of many important weather variables through the S2S database would allow users interested in this methodology to test its viability and potential for better decision-making within their sector before investing in production of a live system.

Training forecasts on reanalysis data corrected to the site showed no significant worsening in forecast skill compared to forecasts trained on a long measured time series of MIDAS data. When correcting reanalysis data to site measurements, accounting for seasonal variations improves the corrected time series by 0.8%. Due to the lack of 100m wind speed variable in the ECMWF S2S forecasts, several weather variables related to 100m wind speed were used in place. Performing one principal component analysis on all these variables together had little effect on the final forecast skill but improved computation times by decreasing the number of model inputs. A linear model with regularisation was used to generate site-specific index forecasts at weekly resolution; including standard deviation, min and max of all principal components (the input features to this model) as well as mean gave no improvement in mean wind speed forecast but was included for variability forecasts. Index forecasts were generated for each ensemble member, before EMOS was applied to correct for the underdispersion seen in the ensemble.

Original contributions of this work include:

- A comparison between forecasts generated with a complete measured time series and those using reanalysis data corrected with a limited history of site data. This bridges the gap between common methods for desk-based studies and those necessary to apply models to real world sites.

- Determination of the skill of S2S forecasts across three different metrics that are relevant for maintenance planning.

- Implementation of a cost-loss model and investigation of the sensitivity of hiring decisions to electricity price

To further extend this work, it would be beneficial to investigate the performance of other models to produce the index forecasts, especially nonlinear models such as GAMLSS or gradient boosted trees that can also learn interactions between input features. Besides the simple linear model used here or a polynomial regression approach [230], no other types of model have been explored for this purpose. Further feature engineering could also prove beneficial. Weather window index forecasts were only generated for one wind speed threshold and weather window length. The safe limit for work in the nacelle is much higher and therefore ‘unsafe’ times will form a much smaller proportion of the dataset in this case; it would be interesting to explore how this affects the skill of the weather window forecasts and how this translates to decision making. Additionally, whole site servicing campaigns require many individual but possibly interdependent decisions to be made. Allowing sharing of resources between sites also adds complexity as well as opportunities for cost reduction that were not considered here. Whole-campaign planning would also lend itself more easily to the inclusion of contractual penalties in the cost-loss model. The calculation of expected lost energy could also be improved by modelling the relationship between useful hours and lost energy (Figure 5.20) as heteroscedastic, although a careful choice of distribution would be needed to evaluate the subsequent integration exactly without numeric integration.

Chapter 6

Conclusions

Chapters 3, 4 and 5 have investigated current problems faced in operational wind power forecasting and suggested possible solutions. Three distinct problems were identified, linked to important properties of a useful and useable forecast.Firstly, the problem of how to deal with missing data both in training and in the live running of a forecast model was explored, as a forecasting system that doesn't consider data quality will not always function to produce sensible forceasts. Secondly, the problem of how to improve forecasting accuracy of power forecasts around times of ramps was investigated through a novel forecast combination approach. Finally, forecasts on subseasonal-to-seasonal timescales were tailored to time and wind speed limits specific to certain maintenance activities to provide relevant forecasts for these use-cases, namely crane use.

Missing data

In the work on missing data, the properties of missing data in real SCADA time series were found, before the effect of various missing data scenarios on forecast skill were simulated through case studies. Real wind power data is shown to have typical median levels of missing data of 2.70% for the power variable and 1.57% for wind speed. However, some sites may display levels up to 36%, greatly reducing forecast skill. Data is Missing Not at Random (MNAR), meaning care must be taken to use an appropriate missing data technique. The impact of missing data on wind power forecasts in an autoregressive framework has been demonstrated, with the most appropriate mitigation methods identified. The key results are summarised:

- Missing training data can have a significant impact on results if not dealt with appropriately; multiple imputation is found to be the best of the methods considered here to compensate for this

- If inputs to an operational forecast model are missing, retraining the model without these inputs results in better performance than filling the missing values using a regression model based on available inputs

- Forecast error improves across all sites when more sites are included in the model, with particular improvement at sites that are missing forecast input data; therefore, spatio-temporal models including a greater number of sites are generally more robust to missing data

- Forecasts continue to worsen with increasing length of missing period, but the largest proportion of the loss of forecast skill comes from missing the most recent information

- When a subset of sites have a short historic dataset available, ERA5 reanalysis data scaled to site location and hub height and passed through a power curve provides a good substitute. Where applicable and only one or two sites have short datasets, Balancing Mechanism data may be used.

While these results are from case studies using a Vector Autoregressive (VAR) forecasting model, future work could extend this to other models. It is expected the results would be similar, as the change in forecast skill is likely related more to the loss of information from the missing variable(s) than the modelling framework itself. In summary, awareness of the properties of missing data, its potential impact on model performance and use of suitable mitigation techniques is essential to realise that model's full potential.

Limitations

This study presented results based on just one forecasting method, the VAR model and at a single group of sites. While it is expected the results would generalise to other models and sites, the exact performance of the proposed missing data methods may vary. It is assumed that all instances of missing data are captured by excluding curtailments and site-wide maintenance activities. However, the maintenance logs vary in their detail and accuracy and so capturing the exact start and end of these may not be possible. Other instances where data is not explicitly missing, but data quality may be affected, were not considered but in reality this would have an impact on forecast performance. For example, if one turbine's sensors become stuck on a single value but this is not flagged and only the whole site power is checked for data quality issues. This case study relies on a method where multiple sites are forecast within one model, which allows inter-site dependencies to be exploited in the missing data mitigation methods. In reality, many forecasters will have access to a very limited set of sites, generally only the sites belonging to the same operator, which will limit the improvement gained from the methods outlined here.

Future research questions

Do other forecasting methods show the same level of improvement using the missing data methods as the VAR case study in this thesis does? While the relative performance of the different missing data methods is expected to remain the same using different forecasting methods and different datasets it would be valuable to demonstrate this.

Quantifying the effect of distance between sites: As an extension to case 1, where a random subset of sites were included in the forecast model, further work is needed to explicitly test the benefit of nearby sites and if distance between sites, or some other measure of similarity, can anticipate the 'added value' of inclusion of the other site. This could also help to demonstrate a use-case for data sharing and data markets.

The work set out in Chapter 3 only analyses the effect of missing data

on deterministic forecasts. Extending these results and methods to probabilistic forecasts would be valuable future work.

Forecast combination

Four individual model forecasts were created, three based on recent measurements and one on Numerical Weather Prediction (NWP) forecasts. None of these models showed good calibration before forecast combination. Four benchmark combination methods were employed: both constrained and unconstrained Optimal Linear Pool (a linear weighted combination), and a beta-transformed Optimal Linear Pool (OLP) combined by both quantile and power. A new nonlinear forecast combination method based on a Gradient Boosted Machine (GBM) is proposed, including a variant that explicitly takes forecasts of ramp rate as well as individual power forecasts into the combination model. All forecast combination models showed higher skill scores (on both Mean Absolute Error (MAE) and Pinball score) than the best individual model for horizons 1–3 hours ahead, with the greatest improvement for 2 hours ahead where the skill of individual models was most similar. The benefit of forecast combination was less clear 4–6 hours ahead where the individual GBM model performed much better than the other individual models.

The proposed combination approach based on a lightgbm model with ramp features gave significantly better skill than all other combination models 1 hour ahead (and up to 3 hours ahead at some locations), but the simpler linear combinations outperformed it at longer (4–6 hours) horizons.

Time points were identified as either a ramp or non-ramp, and the ability of the forecast to predict these assessed. It was found that the lightgbm-ramps model correctly forecasts the highest proportion of true ramps 2 or more hours ahead at 6 of the 10 locations tested, but also has a tendency to over-preedict ramps at non-ramp times. Overall, there is no one model that is consistently better at forecasting ramps across all locations and horizons.

The choice of combination model will depend on the individual site characteristics and the specifics of the wider problem (for example, if there is a greater penalty for missing a ramp than forecasting a false positive). A larger dataset would be needed with significance tests employed to fully understand differences between models. Integrating these forecasts with the financial consequences of system actions would better define their usefulness and limitations.

Limitations

In the case study for forecast combination presented in this thesis, none of the individual models were calibrated. It is possible that forecast combination methods would perform differently (possibly better) with more reliable individual forecasts as inputs. The amount of benefit gained for employing a forecast combination approach will depend on the diversity of the individual models used and how complementary they are to each other. There is also a requirement for relatively large amounts of training data, and that forecast input training data from different sources covers at least some of the same time period to allow for both training of the individual forecast models while also producing out of sample forecasts to use to train the forecast combination step. No one best approach was found in this work; testing of different forecast combination approaches and the conditions under which each are preferred would be necessary but time consuming to operationalise. Similarly, setting up and producing forecasts from multiple different models especially with different data sources is more time consuming and expensive than relying on one model, so the benefits would have to be clear.

Future research questions

The cost function for action taken to mitigate ramps in power can be asymmetric. How can this be incorporated with forecasts to provide the best information for decision makers? As an example, for a downward ramp in power the consequences of procuring too much reserve, while expensive, are less severe than the consequences of customer disconnection in the worst case where the grid can't meet demand. This contextual knowledge must be incorporated into decisions alongside the forecasts. Actions taken when upwards ramps are expected are also different from downwards ramps so these must be treated differently.

Modelling of the most extreme ramps: The most extreme ramps (with the largest change in power or over a very short time) are the most difficult conditions under which to manage the power system. As such, advance warning of these times would be beneficial. Rare or extreme events like this require a different approach to forecasting, for example using extreme value theory, than forecasts of average quantities.

Subseasonal-to-seasonal forecasting

Forecasts of weekly mean wind speed, weekly wind speed variation, and number of useful work hours within the week have been shown to be skilful on S2S timescales. Mean wind speed forecasts showed skill out to three weeks ahead. Variability forecasts had lower skill for one or two week horizons, but retain fair skill out to 6 weeks ahead. A new 'weather window index' of the number of useful hours in a week was proposed to inform activity-specific decision making, for example crane hire. These weather window index forecasts showed very good skill one week ahead and fair skill out to 6 weeks. A cost-loss model was implemented to determine the economic benefit of forecasts, but found little difference between the decisions made under climatology and the decisions made by the S2S forecasts.

Training forecasts on reanalysis data corrected to the site showed no significant worsening in forecast skill compared to forecasts trained on a long measured time series of MIDAS data. When correcting reanalysis data to site measurements, accounting for seasonal variations improves the corrected time series by 0.8%. Due to the lack of 100m wind speed variable in the European Centre for Medium-range Weather Forecasts (ECMWF) S2S forecasts, several weather variables related to 100m wind speed were used in place. Performing one principal component analysis on all these variables together had little effect on the final forecast skill but improved computation times by decreasing the number of model inputs. A linear model with regularisation was used to generate site-specific index forecasts at weekly resolution; including standard deviation, min and max of all principal components (the input features to this model) as well as mean gave no improvement in mean wind speed forecast but was included for variability forecasts. Index forecasts were generated for each ensemble member, before Ensemble Model Output Statistics (EMOS) was applied to correct for the underdispersion seen in the ensemble.

Original contributions of this work include:

- A comparison between forecasts generated with a complete measured time series and those using reanalysis data corrected with a limited history of site data. This bridges the gap between common methods for desk-based studies and those necessary to apply models to real world sites.
- Determination of the skill of S2S forecasts across three different metrics that are relevant for maintenance planning.
- Implementation of a cost-loss model and investigation of the sensitivity of hiring decisions to electricity price.

To further extend this work, it would be beneficial to investigate the performance of other models to produce the index forecasts, especially nonlinear models such as Generalised Additive Model of Location, Scale and Shape (GAMLSS) or gradient boosted trees that can also learn interactions between input features. Besides the simple linear model used here or a polynomial regression approach [230], no other types of model have been explored for this purpose. Further feature engineering could also prove beneficial. Weather window index forecasts were only generated for one wind speed threshold and weather window length. The safe limit for work in the hub is much higher and therefore 'unsafe' times will form a much smaller proportion of the dataset in this case; it would be interesting to explore how this affects the skill of the weather window forecasts and how this translates to decision making. Additionally, whole site servicing campaigns require many individual but possibly interdependent decisions to be made. Allowing sharing of resources between sites also adds complexity as well as opportunities for cost reduction that were not considered here. Whole-campaign planning would also lend itself more easily to the inclusion of contractual penalties in the cost-loss model. The calculation of expected lost energy could also be improved by modelling the relationship between useful hours and lost energy as heteroscedastic, although a careful choice of distribution would be needed to evaluate the subsequent integration exactly without numeric integration.

Limitations

Forecasts for maintenance decisions were produced for a small set of sites in a fairly limited geographical area, due to data availability. However, this limits the confidence in the generalisation of these results to other sites outwith the study area. It was found the decision to hire or not hire a crane was only sensitive to the forecasts for a range of electricity prices; outside this range the decision was always the same. Therefore there may be some sites where electricity price is outside this range and so the forecasts would not aid decision making. The cost loss model presented does not represent whole-year penalties, only the penalty due to lost energy before a turbine is fixed. This type of long term penalty would be more complex to integrate for making single decisions. In addition, the cost loss model assumes a constant known electricity price. Whilst this may be the case for sites with subsidies or a Power Purchase Agreement, increasingly new subsidy-free sites would be exposed to varying electricity prices. In reality, the decision whether to hire equipment or not, and for which exact days, will be revised closer to the booking date. However, this revision decision is asymmetrical: it would be possible to cancel a booking made earlier but much less likely to be able to put in a booking at short notice when the earlier decision on Subseasonal-to-Seasonal (S2S) timescales had been to not book equipment. This is currently not accounted for in the S2S decision making methodology.

Future research questions

Refining of index model step and future improvements to S2S forecast products: Only a linear model was used to go from principal components on a large geographical scale to a site-specific index forecast in this work. Future work could explore the use of more sophisticated models that include nonlinear relationships for example. It is also anticipated that future developments in S2S products from providers such as ECMWF may also increase skill. It is hoped that increasing interest from the energy community in this field may also spur provision of further relevant weather variables such as 100 m wind speed which may also be used in future work to improve forecast skill for wind energy applications.

Application of a similar methodology to other decisions: Here only the use of S2S forecasts for equipment hire decisions have been tested with a cost-loss model. The useful hours metric used here could be extended to other applications with different cost-loss functions.

Fully representing all penalties in cost-loss modelling: Developing a method to include long-term penalties in the cost-loss based decision methodology laid out in this work would be valuable for use cases such as annual servicing campaigns across many turbines.

6.1 Final remarks

This thesis has identified three areas within forecasting for wind energy where it was felt further development of methodologies would be of benefit. A review of recent work in the field was undertaken, where parallels were drawn between wind and solar literatures and recommendations for good practice in forecast development and testing identified. Chapter 3 gives a guide to suitable strategies for dealing with missing data when implementing a live forecasting system. This has already informed decisions made in the design of a new very short-term forecasting model at Natural Power and it is hoped it will draw attention to this aspect of forecasting which is often not given a lot of attention. A novel forecast combination approach has been presented in Chapter 4 to improve very short-term forecasts at times of ramps. While this work did show slight improvements in both accuracy of power forecasts and identification of ramps an hour or two ahead, it is clear this is a difficult task that requires further refining. Finally, a new metric for use in wind turbine equipment hire decisions for maintenance tasks was proposed. This work demonstrates the potential of S2S forecasts for this new application and encourage both interest from industry in using such a forecast product and also from academia to develop future improvement to forecast accuracy and applicability on these timescales. The research outputs listed in the next section serve as a record of this research for others to refer to and build on.

Future improvements in forecast performance may come from greater availability of a wider range of data sources such as satellite data and sky-imaging systems for solar. Higher resolution NWP using improved atmospheric models will also improve renewable energy forecasts. Wider use of probabilistic forecasts would benefit the energy sector and forecasts for new ancillary services will be required. While the focus of this thesis has been on forecasts for wind energy applications, similar approaches could be applied for solar too - although normalisation of solar to remove diurnal variations is an additional challenge.

6.2 Research outputs

Papers

R. Tawn, J. Browell and D. McMillan, “Subseasonal-to-Seasonal forecasting for wind turbine maintenance scheduling,” *Wind,* accepted May 2022.

R. Tawn and J. Browell, “A review of very short-term wind and solar power forecasting,” *Renewable and Sustainable Energy Reviews*, vol. 153, p. 111758, Jan. 2022

R. Tawn, J. Browell, and I. Dinwoodie, “Missing data in wind farm time series: Properties and effect on forecasts,” *Electric Power Systems Research*, vol. 189, p. 106640, Dec. 2020

J. Browell, C. Gilbert, R. Tawn, and L. May, “Quantile Combination for the EEM20 Wind Power Forecasting Competition,” in *2020 17th International Conference on the European Energy Market (EEM).* Stockholm, Sweden: IEEE, Sep. 2020, pp. 1–6

Presentations

R. Tawn and J. Browell, “Forecast combination and adaptation for improved ramp forecasts,” in *Wind Energy Science Conference,* online, May 2021.

J. Browell, C. Gilbert, R. Tawn and L. May, "Quantile combination for the EEM wind power forecasting competition," in *17th International Conference on the European Energy Market,* online, Sep. 2020.

R. Tawn and J. Browell, "Missing data in wind farm time series: Properties and effect on forecasts," in *XXI Power Systems Computation Conference,* online, May 2020.

Bibliography

[1] Intergovernmental Panel on Climate Change, *Climate Change 2014 Mitigation of Climate Change: Working Group III Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge: Cambridge University Press, 2014.

[2] R. Yamartino, "A comparison of several single-pass estimators of the standard deviation of wind direction," *Journal of Applied Meteorology and Climatology*, vol. 23, pp. 1362–1366, 1984.

[3] R. Tawn and J. Browell, "A review of very short-term wind and solar power forecasting," *Renewable and Sustainable Energy Reviews*, vol. 153, p. 111758, Jan. 2022.

[4] R. Tawn, J. Browell, and I. Dinwoodie, "Missing data in wind farm time series: Properties and effect on forecasts," *Electric Power Systems Research*, vol. 189, p. 106640, Dec. 2020.

[5] J. Browell, C. Gilbert, R. Tawn, and L. May, "Quantile Combination for the EEM20 Wind Power Forecasting Competition," in *2020 17th International Conference on the European Energy Market (EEM).* Stockholm, Sweden: IEEE, Sep. 2020, pp. 1–6.

[6] S. Arrhenius, "On the influence of carbonic acid in the air upon the temperature of the ground," *Philosophical Magazine and Journal of Science*, vol. 41, no. 5, p. 22, 1896.

[7] "Supplemental data of the Global Carbon Budget 2021," 2021.

[8] "State of the Climate: Global Climate Report for 2020," NOAA National Centers for Environmental Information, Tech. Rep., 2021.

[9] V. Masson-Delmotte, P. Zhai, H. O. Pörtner, D. Roberts, J. Skea, P. R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J. B. R. Matthews, Y. Chen, X. Zhou, M. I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, and T. Waterfield, "Summary for policymakers. In: Global warming of 1.5°C . A special report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty," IPCC, Tech. Rep., 2018.

[10] "UK Energy in Brief 2020," UK Government, Tech. Rep., 2020.

[11] D. Harvey, "Hinckley nuclear power station on track for 2026 opening," bbc.co.uk/news/uk-england-somerset-58724732, Sep. 2021, Accessed 16/01/2022.

[12] Low Carbon Contracts Company, "CfD Register," lowcarboncontracts.uk/cfd-register/, Accessed 16/01/2021.

[13] J. Lelieveld, K. Klingmüller, A. Pozzer, R. T. Burnett, A. Haines, and V. Ramanathan, "Effects of fossil fuel and total anthropogenic emission removal on public health and climate," *Proceedings of the National Academy of Sciences*, vol. 116, no. 15, pp. 7192–7197, Apr. 2019.

[14] "Review of Maritime Transport 2019," United Nations Conference on Trade and Development, Tech. Rep., 2020.

[15] T. Smith, E. O'Keeffe, L. Aldous, S. Parker, C. Raucci, M. Traut, J. J. Corbett, J. J. Winebreak, J.-P. Jalkanen, L. Johansson, B. Anderson, A. Agrawal, S. Ettinger, S. Ng, S. Hanayama, J. Faber, D. Nelissen, M. 't Hoen, D. Lee, S. Chesworth, and A. Pandey, "Third IMO Greenhouse Gas Study 2014," International Maritime Organisation, Tech. Rep., 2014.

[16] R. K. Tiwary, "Environmental impact of coal mining on water regime and its management," *Water, Air and Soil Pollution*, vol. 132, pp. 185–199, 2001.

[17] J. Beyer, H. C. Trannum, T. Bakke, P. V. Hodson, and T. K. Collier, "Environmental effects of the Deepwater Horizon oil spill: A review," *Marine Pollution Bulletin*, vol. 110, no. 1, pp. 28–51, Sep. 2016.

[18] "COP26: Thousands march for Glasgow's biggest protest," bbc.co.uk/news/uk-scotland-59185007, Nov. 2021, Accessed 16/01/2022.

[19] "Cambo oil field development off Shetland to be paused," bbc.co.uk/news/uk-scotland-59608521, Dec. 2021, Accessed 16/01/2022.

[20] P. Pinson, "Wind energy: Forecasting challenges for its operational management," *Statistical Science*, vol. 28, no. 4, pp. 564–585, Nov. 2013.

[21] J. Wang, Y. Song, F. Liu, and R. Hou, "Analysis and application of forecasting models in wind power integration: A review of multi-step-ahead wind speed forecasting models," *Renewable and Sustainable Energy Reviews*, vol. 60, pp. 960–981, Jul. 2016.

[22] G. Kariniotakis, *Renewable Energy Forecasting: From Models to Applications*. Woodhead publishing, 2017.

[23] P. Pinson, C. Chevallier, and G. N. Kariniotakis, "Trading Wind Generation From Short-Term Probabilistic Forecasts of Wind Power," *IEEE Transactions on Power Systems*, vol. 22, no. 3, pp. 1148–1156, Aug. 2007.

[24] H. Holttinen, J. Kiviluoma, N. Helistö, T. Levy, N. Menemenlis, L. Jun, N. Cutululis, M. Koivisto, K. Das, A. Orths, P. Börre Eriksen, N. Emmanuel, J.-Y. Bourmand, J. Dobschinski, C. Pellinger, S. von Roon, A. Guminski, and others, "Design and operation of energy systems with large amounts of variable generation," IEA TCP Wind Task 25, FI, Tech. Rep., 2021.

[25] D. Pagnani, F. Blaabjerg, C. L. Bak, F. M. Faria da Silva, Ł. H. Kocewiak, and J. Hjerrild, "Offshore Wind Farm Black Start Service Integration: Review and Outlook of Ongoing Research," *Energies*, vol. 13, no. 23, p. 6286, Nov. 2020.

[26] T. Corcut, "Open letter on dynamic parameters and other information submitted by generators in the balancing mechanism," Sep. 2020.

[27] J. Wohland, D. Brayshaw, H. Bloomfield, and M. Wild, "European multidecadal solar variability badly captured in all centennial reanalyses except CERA20C," *Environmental Research Letters*, vol. 15, no. 10, p. 104021, Sep. 2020.

[28] J. Wohland, N. E. Omrani, N. Keenlyside, and D. Witthaut, "Significant multidecadal variability in German wind energy generation," *Wind Energy Science*, vol. 4, no. 3, pp. 515–526, Sep. 2019.

[29] D. Hdidouan and I. Staffell, "The impact of climate change on the levelised cost of wind energy," *Renewable Energy*, vol. 101, pp. 575–592, Feb. 2017.

[30] World Meteorological Organisation, https://www.wmo.int/pages/prog/www/DPS/gdps.html, Accessed 09/11/2020.

[31] M. Baldini, "The Danish perspective on forecasting and integration of renewables in power systems," Danish Energy Agency, Tech. Rep., Aug. 2020.

[32] G. Kariniotakis, S. Camal, and Smart4RES team, "Smart4RES: Improved weather modelling and forecasting dedicated to renewable energy applications," in *EGU General Assembly*, online, 2021, p. 2.

[33] C. W. Hansen, W. F. Holmgren, A. Tuohy, J. Sharp, A. T. Lorenzo, L. J. Boeman, and A. Golnas, "The solar forecast arbiter: An open source evaluation framework for solar forecasting," in *2019 IEEE 46th Photovoltaic Specialists Conference (PVSC)*. Chicago, IL, USA: IEEE, Jun. 2019, pp. 2452–2457.

[34] T. Hong, P. Pinson, Y. Wang, R. Weron, D. Yang, and H. Zareipour, "Energy forecasting: A review and outlook," *IEEE Open Acces Journal of Power and Energy*, pp. 376–388, 2020.

[35] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. Martinez-de-Pison, and F. Antonanzas-Torres, “Review of photovoltaic power forecasting,” *Solar Energy*, vol. 136, pp. 78–111, Oct. 2016.

[36] K. Doubleday, V. Van Scyoc Hernandez, and B.-M. Hodge, “Benchmark probabilistic solar forecasts: Characteristics and recommendations,” *Solar Energy*, vol. 206, pp. 52–67, Aug. 2020.

[37] R. H. Inman, H. T. Pedro, and C. F. Coimbra, “Solar forecasting methods for renewable energy integration,” *Progress in Energy and Combustion Science*, vol. 39, no. 6, pp. 535–576, Dec. 2013.

[38] S. Sobri, S. Koohi-Kamali, and N. A. Rahim, “Solar photovoltaic generation forecasting methods: A review,” *Energy Conversion and Management*, vol. 156, pp. 459–497, Jan. 2018.

[39] R. Ahmed, V. Sreeram, Y. Mishra, and M. Arif, “A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization,” *Renewable and Sustainable Energy Reviews*, vol. 124, p. 109792, May 2020.

[40] D. Yang, J. Kleissl, C. A. Gueymard, H. T. Pedro, and C. F. Coimbra, “History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining,” *Solar Energy*, vol. 168, pp. 60–101, Jul. 2018.

[41] J. Jung and R. P. Broadwater, “Current status and future advances for wind speed and power forecasting,” *Renewable and Sustainable Energy Reviews*, vol. 31, pp. 762–777, Mar. 2014.

[42] G. Giebel and G. Kariniotakis, “Wind power forecasting—a review of the state of the art,” in *Renewable Energy Forecasting*. Elsevier, 2017, pp. 59–109.

[43] A. M. Foley, P. G. Leahy, A. Marvuglia, and E. J. McKeogh, “Current methods and advances in forecasting of wind power generation,” *Renewable Energy*, vol. 37, no. 1, pp. 1–8, Jan. 2012.

[44] A. Tascikaraoglu and M. Uzunoglu, “A review of combined approaches for prediction of short-term wind speed and power,” *Renewable and Sustainable Energy Reviews*, vol. 34, pp. 243–254, Jun. 2014.

[45] M. Santhosh, C. Venkaiah, and D. M. Vinod Kumar, “Current advances and approaches in wind speed and wind power forecasting for improved renewable energy integration: A review,” *Engineering Reports*, vol. 2, no. 6, Jun. 2020.

[46] C. Sweeney, R. J. Bessa, J. Browell, and P. Pinson, “The future of forecasting for renewable energy,” *WIREs Energy and Environment*, Sep. 2019.

[47] F. Barbieri, S. Rajakaruna, and A. Ghosh, “Very short term photovoltaic power forecasting with cloud modeling: A review,” *Renewable and Sustainable Energy Reviews*, vol. 75, pp. 242–263, 2017.

[48] M. Guermoui, F. Melgani, K. Gairaa, and M. L. Mekhalfi, “A comprehensive review of hybrid models for solar radiation forecasting,” *Journal of Cleaner Production*, vol. 258, p. 120357, Jun. 2020.

[49] T. Jensen, T. Fowler, B. Brown, J. Lazo, and S. E. Haupt, “Metrics for evaluation of solar energy forecasts,” NCAR, Tech. Rep. NCAR/TN-527+STR, Jun. 2016.

[50] Y. Zhang, J. Wang, and X. Wang, “Review on probabilistic forecasting of wind power generation,” *Renewable and Sustainable Energy Reviews*, vol. 32, pp. 255–270, Apr. 2014.

[51] M. S. Santos, R. C. Pereira, A. F. Costa, J. P. Soares, J. Santos, and P. H. Abreu, “Generating synthetic missing data: A review by missing mechanism,” *IEEE Access*, pp. 11 651–11 667, 2019.

[52] P. Lauret, M. David, and P. Pinson, “Verification of solar irradiance probabilistic forecasts,” *Solar Energy*, vol. 194, pp. 254–271, Dec. 2019.

[53] UK Government, “Solar Voltaics Deployment May 2020,” Accessed 28/07/2020.

[54] ——, “Renewable Energy Planning Database,” Accessed 28/07/2020.

[55] I. Neher, T. Buchmann, S. Crewell, B. Pospichal, and S. Meilinger, "Impact of atmospheric aerosols on solar power," *Meteorologische Zeitschrift*, vol. 28, no. 4, pp. 305–321, Nov. 2019.

[56] B. Stridh, "Evaluation of economical benefit of cleaning of soiling and snow in PV plants at three European locations," in *2012 38th IEEE Photovoltaic Specialists Conference*. Austin, TX, USA: IEEE, Jun. 2012, pp. 1448–1451.

[57] H. T. Pedro, R. H. Inman, and C. F. Coimbra, "Mathematical methods for optimized solar forecasting," in *Renewable Energy Forecasting*. Elsevier, 2017, pp. 111–152.

[58] H. Yang, B. Kurtz, D. Nguyen, B. Urquhart, C. W. Chow, M. Ghonima, and J. Kleissl, "Solar irradiance forecasting using a ground-based sky imager developed at UC San Diego," *Solar Energy*, vol. 103, pp. 502–524, May 2014.

[59] F. Wang, Z. Zhen, C. Liu, Z. Mi, B.-M. Hodge, M. Shafie-khah, and J. P. Catalão, "Image phase shift invariance based cloud motion displacement vector calculation method for ultra-short-term solar PV power forecasting," *Energy Conversion and Management*, vol. 157, pp. 123–135, Feb. 2018.

[60] T. Schmidt, J. Kalisch, E. Lorenz, and D. Heinemann, "Evaluating the spatio-temporal performance of sky-imager-based solar irradiance analysis and forecasts," *Atmospheric Chemistry and Physics*, vol. 16, no. 5, pp. 3399–3412, Mar. 2016.

[61] J. Lago, K. De Brabandere, F. De Ridder, and B. De Schutter, "Short-term forecasting of solar irradiance without local telemetry: A generalized model using satellite data," *Solar Energy*, vol. 173, pp. 566–577, Oct. 2018.

[62] T. M. Harty, W. F. Holmgren, A. T. Lorenzo, and M. Morzfeld, "Intra-hour cloud index forecasting with data assimilation," *Solar Energy*, vol. 185, pp. 270–282, Jun. 2019.

[63] K. Bellinguer, R. Girard, G. Bontron, and G. Kariniotakis, “Short-term forecasting of photovoltaic generation based on conditioned learning of geopotential fields,” in *2020 55th International Universities Power Engineering Conference (UPEC)*. Torino, Italy: IEEE, Sep. 2020, pp. 1–6.

[64] T. Carriere, C. Vernay, S. Pitaval, and G. Kariniotakis, “A novel approach for seamless probabilistic photovoltaic power forecasting covering multiple time frames,” *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2281–2292, May 2020.

[65] M. Fliess, C. Join, and C. Voyant, “Prediction bands for solar energy: New short-term time series forecasting techniques,” *Solar Energy*, vol. 166, pp. 519–528, May 2018.

[66] H. Lee, N.-W. Kim, J.-G. Lee, and B.-T. Lee, “Uncertainty-aware forecast interval for hourly PV power output,” *IET Renewable Power Generation*, vol. 13, no. 14, pp. 2656–2664, Oct. 2019.

[67] F. Golestaneh, P. Pinson, and H. B. Gooi, “Very short-term nonparametric probabilistic forecasting of renewable energy generation— with application to solar energy,” *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 3850–3863, Sep. 2016.

[68] H. Sheng, J. Xiao, Y. Cheng, Q. Ni, and S. Wang, “Short-term solar power forecasting based on weighted gaussian process regression,” *IEEE Transactions on Industrial Electronics*, vol. 65, no. 1, pp. 300–308, Jan. 2018.

[69] I. Majumder, P. Dash, and R. Bisoi, “Variational mode decomposition based low rank robust kernel extreme learning machine for solar irradiation forecasting,” *Energy Conversion and Management*, vol. 171, pp. 787–806, Sep. 2018.

[70] C. Huang, L. Wang, and L. L. Lai, “Data-driven short-term solar irradiance forecasting based on information of neighboring sites,” *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9918–9927, Dec. 2019.

[71] M. Rana, I. Koprinska, and V. G. Agelidis, "Univariate and multivariate methods for very short-term solar photovoltaic power forecasting," *Energy Conversion and Management*, vol. 121, pp. 380–390, Aug. 2016.

[72] B. Sivaneasan, C. Yu, and K. Goh, "Solar forecasting using ANN with fuzzy logic pre-processing," *Energy Procedia*, vol. 143, pp. 727–732, Dec. 2017.

[73] X. Luo, J. Sun, L. Wang, W. Wang, W. Zhao, J. Wu, J.-H. Wang, and Z. Zhang, "Short-term wind speed forecasting via stacked extreme learning machine with generalized correntropy," *IEEE Transations on Industrial Informatics*, vol. 14, no. 11, pp. 4963–4971, Nov. 2018.

[74] P. Tang, D. Chen, and Y. Hou, "Entropy method combined with extreme learning machine method for the short-term photovoltaic power generation forecasting," *Chaos, Solitons & Fractals*, vol. 89, pp. 243–248, Aug. 2016.

[75] M. Abuella and B. Chowdhury, "Random forest ensemble of support vector regression models for solar power forecasting," in *2017 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)*. North America: IEEE, 2017, pp. 1–5.

[76] A. T. Eseye, J. Zhang, and D. Zheng, "Short-term photovoltaic solar power forecasting using a hybrid wavelet-PSO-SVM model based on SCADA and meteorological information," *Renewable Energy*, vol. 118, pp. 357–367, Apr. 2018.

[77] C. Li, Z. Xiao, X. Xia, W. Zou, and C. Zhang, "A hybrid model based on synchronous optimisation for multi-step short-term wind speed forecasting," *Applied Energy*, vol. 215, pp. 131–144, Apr. 2018.

[78] X. G. Agoua, R. Girard, and G. Kariniotakis, "Short-term spatio-temporal forecasting of photovoltaic power production," *IEEE Transactions on Sustainable Energy*, vol. 9, no. 2, pp. 538–546, Apr. 2018.

[79] S. E. Haupt, B. Kosović, T. Jensen, J. K. Lazo, J. A. Lee, P. A. Jiménez, J. Cowie, G. Wiener, T. C. McCandless, M. Rogers, S. Miller, M. Sengupta,

Y. Xie, L. Hinkelman, P. Kalb, and J. Heiser, "Building the Sun4Cast system: Improvements in solar power forecasting," *Bulletin of the American Meteorological Society*, vol. 99, no. 1, pp. 121–136, Jan. 2018.

[80] J. A. Lee, S. E. Haupt, P. A. Jiménez, M. A. Rogers, S. D. Miller, and T. C. McCandless, "Solar irradiance nowcasting case studies near Sacramento," *Journal of Applied Meteorology and Climatology*, vol. 56, no. 1, pp. 85–108, Jan. 2017.

[81] M. Méchali, R. Barthelmie, S. Frandsen, L. Jensen, and P.-E. Réthoré, "Wake effects at Horns Rev and their influence on energy production," in *Proc. European Wind Energy Conference*, 2006, p. 10.

[82] F. Lamraoui, G. Fortin, R. Benoit, J. Perron, and C. Masson, "Atmospheric icing impact on wind turbine production," *Cold Regions Science and Technology*, vol. 100, pp. 36–49, Apr. 2014.

[83] W. Michael, "Long term performance of wind farms," http://www.ewea.org/events/workshops/wp-content/uploads/2014/12/Tech14a2-1-Wilkinson.pdf, Accessed 09/11/2020.

[84] H. Zhou and Z. Wang, "A multiple-model based adaptive control algorithm for very-short term wind power forecasting," in *2016 IEEE International Conference on Power System Technology (POWERCON).* Wollongong, Australia: IEEE, Sep. 2016, pp. 1–6.

[85] L. Cavalcante, R. J. Bessa, M. Reis, and J. Browell, "LASSO vector autoregression structures for very short-term wind power forecasting," *Wind Energy*, vol. 20, no. 4, pp. 657–675, Apr. 2017.

[86] J. W. Messner and P. Pinson, "Online adaptive LASSO estimation in vector autoregressive models for high dimensional wind power forecasting," *International Journal of Forecasting*, vol. 35, no. 4, pp. 1485–1498, 2019.

[87] J. Dowell and P. Pinson, "Very-short-term probabilistic wind power forecasts by sparse vector autoregression," *IEEE Transactions on Smart Grid*, pp. 763–770, 2015.

[88] Y. Zhao, L. Ye, P. Pinson, Y. Tang, and P. Lu, "Correlation-constrained and sparsity-controlled vector autoregressive model for spatio-temporal wind power forecasting," *IEEE Transactions on Power Systems*, vol. 33, no. 5, pp. 5029–5040, Sep. 2018.

[89] J. Browell, D. R. Drew, and K. Philippopoulos, "Improved very short-term spatio-temporal wind forecasting using atmospheric regimes," *Wind Energy*, pp. 968–979, May 2018.

[90] C. Gonçalves, R. J. Bessa, and P. Pinson, "A critical overview of privacy-preserving approaches for collaborative forecasting," *International Journal of Forecasting*, vol. 37, no. 1, pp. 322–342, 2021.

[91] Y. Zhang and J. Wang, "A distributed approach for wind power probabilistic forecasting considering spatio-temporal correlation without direct access to off-site information," *IEEE Transactions on Power Systems*, vol. 33, no. 5, pp. 5714–5726, Sep. 2018.

[92] A. Chaudhary, A. Sharma, A. Kumar, K. Dikshit, and N. Kumar, "Short term wind power forecasting using machine learning techniques," *Journal of Statistics and Management Systems*, vol. 23, no. 1, pp. 145–156, Jan. 2020.

[93] M. Landry, "Probabilistic gradient boosting machines for GEFCom2014 wind forecasting," *International Journal of Forecasting*, pp. 1061–1066, 2016.

[94] S. Jiang, R. Fang, L. Wang, and C. Peng, "Very short-term wind power forecasting based on SVM-Markov," in *Proc. Int. Conference Advances Energy Environ. Chem. Eng.* Changska, China: Atlantis Press, 2015, pp. 130–134.

[95] J. He and J. Xu, “Ultra-short-term wind speed forecasting based on support vector machine with combined kernel function and similar data,” *Journal on Wireless Communications and Networking*, vol. 2019, no. 248, Dec. 2019.

[96] J. Wang, J. Heng, L. Xiao, and C. Wang, “Research and application of a combined model based on multi-objective optimization for multi-step ahead wind speed forecasting,” *Energy*, vol. 125, pp. 591–613, Apr. 2017.

[97] P. Du, J. Wang, Z. Guo, and W. Yang, “Research and application of a novel hybrid forecasting system based on multi-objective optimization for wind speed forecasting,” *Energy Conversion and Management*, vol. 150, pp. 90–107, Oct. 2017.

[98] P. Du, J. Wang, W. Yang, and T. Niu, “A novel hybrid model for short-term wind power forecasting,” *Applied Soft Computing*, vol. 80, pp. 93–106, Jul. 2019.

[99] M. Khodayar, O. Kaynak, and M. E. Khodayar, “Rough deep neural architecture for short-term wind speed forecasting,” *IEEE Transactions on Industrial Informatics*, vol. 13, no. 6, pp. 2770–2779, Dec. 2017.

[100] NREL, “Western Wind data set,” www.nrel.gov/grid/western-wind-data.html, Accessed 2021-02-04.

[101] F. Rodríguez, A. M. Florez-Tapia, L. Fontán, and A. Galarza, “Very short-term wind power density forecasting through artificial neural networks for microgrid control,” *Renewable Energy*, vol. 145, pp. 1517–1527, Jan. 2020.

[102] M. Khodayar and J. Wang, “Spatio-temporal graph deep neural network for short-term wind speed forecasting,” *IEEE Transactions on Sustainable Energy*, vol. 10, no. 2, pp. 670–681, Apr. 2019.

[103] M. A. Hossain, R. K. Chakrabortty, S. Elsawah, and M. J. Ryan, “Very short-term forecasting of wind power generation using hybrid deep learning model,” *Journal of Cleaner Production*, vol. 296, p. 126564, May 2021.

[104] X. Ma, Y. Jin, and Q. Dong, “A generalized dynamic fuzzy neural network based on singular spectrum analysis optimized by brain storm optimization for short-term wind speed forecasting,” *Applied Soft Computing*, vol. 54, pp. 296–312, May 2017.

[105] Z. Zhang, H. Qin, Y. Liu, Y. Wang, L. Yao, Q. Li, J. Li, and S. Pei, “Long short-term memory network based on neighborhood gates for processing complex causality in wind speed prediction,” *Energy Convers. Manag.*, vol. 192, pp. 37–51, Jul. 2019.

[106] J. Chen, G.-Q. Zeng, W. Zhou, W. Du, and K.-D. Lu, “Wind speed forecasting using nonlinear-learning ensemble of deep learning time series prediction and extremal optimization,” *Energy Conversion and Management*, vol. 165, pp. 681–695, Jun. 2018.

[107] H. Liu, X.-W. Mi, and Y.-F. Li, “Wind speed forecasting method based on deep learning strategy using empirical wavelet transform, long short term memory neural network and Elman neural network,” *Energy Conversion and Management*, vol. 156, pp. 498–514, Jan. 2018.

[108] P. Lu, L. Ye, B. Sun, C. Zhang, Y. Zhao, and J. Teng, “A new hybrid prediction method of ultra-short-term wind power forecasting based on EEMD-PE and LSSVM optimized by the GSA,” *Energies*, vol. 11, no. 4, p. 697, Mar. 2018.

[109] J. Zhang, Y. Wei, Z.-F. Tan, W. Ke, and W. Tian, “A hybrid method for short-term wind speed forecasting,” *Sustainability*, vol. 9, no. 4, p. 596, Apr. 2017.

[110] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, “The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis,” *Proceedings of the Royal Society of London A*, vol. 454, no. 1971, pp. 903–995, 1996.

[111] Z. Liu, P. Jiang, L. Zhang, and X. Niu, “A combined forecasting model for time series: Application to short-term wind speed forecasting,” *Applied Energy*, vol. 259, p. 114137, Feb. 2020.

[112] W. Zhang, Z. Qu, K. Zhang, W. Mao, Y. Ma, and X. Fan, "A combined model based on CEEMDAN and modified flower pollination algorithm for wind speed forecasting," *Energy Conversion and Management*, vol. 136, pp. 439–451, Mar. 2017.

[113] Aasim, S. Singh, and A. Mohapatra, "Repeated wavelet transform based ARIMA model for very short-term wind speed forecasting," *Renewable Energy*, vol. 136, pp. 758–768, Jun. 2019.

[114] C. Zhang, J. Zhou, C. Li, W. Fu, and T. Peng, "A compound structure of ELM based on feature selection and parameter optimization using hybrid backtracking search algorithm for wind speed forecasting," *Energy Conversion and Management*, vol. 143, pp. 360–376, Jul. 2017.

[115] S. G. Hall and J. Mitchell, "Combining density forecasts," *International Journal of Forecasting*, vol. 23, no. 1, pp. 1–13, Jan. 2007.

[116] F. X. Diebold, "Forecast combination and encompassing: Reconciling two divergent literatures," *International Journal of Forecasting*, vol. 5, pp. 589–592, 1989.

[117] J. Shi, Z. Ding, W.-J. Lee, Y. Yang, Y. Liu, and M. Zhang, "Hybrid forecasting model for very-short term wind power forecasting based on grey relational analysis and wind speed distribution features," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 521–526, Jan. 2014.

[118] L. Xiao, J. Wang, Y. Dong, and J. Wu, "Combined forecasting models for wind energy forecasting: A case study in China," *Renewable and Sustainable Energy Reviews*, vol. 44, pp. 271–288, Apr. 2015.

[119] A. Zameer, J. Arshad, A. Khan, and M. A. Z. Raja, "Intelligent and robust prediction of short term wind power using genetic programming based ensemble of neural networks," *Energy Conversion and Management*, vol. 134, pp. 361–372, Feb. 2017.

[120] T. Ouyang, X. Zha, and L. Qin, “A combined multivariate model for wind power prediction,” *Energy Conversion and Management*, vol. 144, pp. 361–373, Jul. 2017.

[121] Y. Lin, M. Yang, C. Wan, J. Wang, and Y. Song, “A multi-model combination approach for probabilistic wind power forecasting,” *IEEE Transations on Sustainable Energy*, vol. 10, no. 1, pp. 226–237, Jan. 2019.

[122] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, “Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond,” *International Journal of Forecasting*, vol. 32, no. 3, pp. 896–913, Jul. 2016.

[123] C. Feng, M. Cui, B.-M. Hodge, and J. Zhang, “A data-driven multi-model methodology with deep feature selection for short-term wind forecasting,” *Applied Energy*, vol. 190, pp. 1245–1257, Mar. 2017.

[124] P. Li, X. Wang, and J. Yang, “Short-term wind power forecasting based on two-stage attention mechanism,” *IET Renewable Power Generation*, vol. 14, no. 2, pp. 297–304, Feb. 2020.

[125] P. Jiang, R. Li, and H. Li, “Multi-objective algorithm for the design of prediction intervals for wind power forecasting model,” *Applied Mathematical Modelling*, vol. 67, pp. 101–122, Mar. 2019.

[126] H.-Z. Wang, G.-Q. Li, G.-B. Wang, J.-C. Peng, H. Jiang, and Y.-T. Liu, “Deep learning based ensemble approach for probabilistic wind power forecasting,” *Applied Energy*, vol. 188, pp. 56–70, Feb. 2017.

[127] A. Carpinone, M. Giorgio, R. Langella, and A. Testa, “Markov chain modeling for very-short-term wind power forecasting,” *Electric Power Systems Research*, vol. 122, pp. 152–158, May 2015.

[128] A. Bracale and P. De Falco, "An advanced Bayesian method for short-term probabilistic forecasting of the generation of wind power," *Energies*, vol. 8, no. 9, pp. 10 293–10 314, Sep. 2015.

[129] Z. Zhang, L. Ye, H. Qin, Y. Liu, C. Wang, X. Yu, X. Yin, and J. Li, "Wind speed prediction method using shared weight long short-term memory network and Gaussian process regression," *Appl. Energy*, vol. 247, pp. 270–284, Aug. 2019.

[130] C. Gilbert, J. Browell, and D. McMillan, "Leveraging turbine-level data for improved probabilistic wind power forecasting," *IEEE Transactions on Sustainable Energy*, vol. 11, no. 3, pp. 1152–1160, Jul. 2020.

[131] P. Jiang, Y. Wang, and J. Wang, "Short-term wind speed forecasting using a hybrid model," *Energy*, vol. 119, pp. 561–577, Jan. 2017.

[132] A. A. Ezzat, "Turbine-specific short-term wind speed forecasting considering within-farm wind field dependencies and fluctuations," *Applied Energy*, vol. 269, p. 115034, Jul. 2020.

[133] Y. Ding, *Data Science for Wind Energy*, 1st ed. Chapman and Hall/CRC, 2019.

[134] R. Yu, J. Gao, M. Yu, W. Lu, T. Xu, M. Zhao, J. Zhang, R. Zhang, and Z. Zhang, "LSTM-EFG for wind power forecasting based on sequential correlation features," *Future Generation Computer Systems*, vol. 93, pp. 33–42, Apr. 2019.

[135] I. Würth, L. Valldecabres, E. Simon, C. Möhrlen, B. Uzunoğlu, C. Gilbert, G. Giebel, D. Schlipf, and A. Kaifel, "Minute-Scale Forecasting of Wind Power—Results from the Collaborative Workshop of IEA Wind Task 32 and 36," *Energies*, vol. 12, no. 4, p. 712, Feb. 2019.

[136] L. Valldecabres, N. Nygaard, L. Vera-Tudela, L. von Bremen, and M. Kühn, "On the use of dual-doppler radar measurements for very short-term wind power forecasts," *Remote Sensing*, vol. 10, no. 11, p. 1701, Oct. 2018.

[137] L. Valldecabres, A. Peña, M. Courtney, L. von Bremen, and M. Kühn, “Very short-term forecast of near-coastal flow using scanning lidars,” *Wind Energy Science*, vol. 3, no. 1, pp. 313–327, May 2018.

[138] S. G. Benjamin, S. S. Weygandt, J. M. Brown, M. Hu, C. R. Alexander, T. G. Smirnova, J. B. Olson, E. P. James, D. C. Dowell, G. A. Grell, H. Lin, S. E. Peckham, T. L. Smith, W. R. Moninger, J. S. Kenyon, and G. S. Manikin, “A North American hourly assimilation and model forecast cycle: The rapid refresh,” *Monthly Weather Review*, vol. 144, no. 4, pp. 1669–1694, Apr. 2016.

[139] M. Milan, B. Macpherson, R. Tubbs, G. Dow, G. Inverarity, M. Mittermaier, G. Halloran, G. Kelly, D. Li, A. Maycock, T. Payne, C. Piccolo, L. Stewart, and M. Wlasak, “Hourly 4D-Var in the Met Office UKV operational forecast model,” *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 728, pp. 1281–1301, Apr. 2020.

[140] S. Hagelin, J. Son, R. Swinbank, A. McCabe, N. Roberts, and W. Tennant, “The Met Office convective-scale ensemble, MOGREPS-UK,” *Quarterly Journal of the Royal Meteorological Society*, vol. 143, no. 708, pp. 2846–2861, Oct. 2017.

[141] J. R. Lawson, J. S. Kain, N. Yussouf, D. C. Dowell, D. M. Wheatley, K. H. Knopfmeier, and T. A. Jones, “Advancing from convection-allowing NWP to warn-on-forecast: Evidence of progress,” *Weather Forecasting*, vol. 33, no. 2, pp. 599–607, Apr. 2018.

[142] J. W. Messner, P. Pinson, J. Browell, M. B. Bjerregård, and I. Schicker, “Evaluation of wind power forecasts—An up-to-date view,” *Wind Energy*, vol. 23, no. 6, pp. 1461–1481, Jun. 2020.

[143] M. E. Torres, M. A. Colominas, G. Schlotthauer, and P. Flandrin, “A complete ensemble empirical mode decomposition with adaptive noise,” in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Prague, Czech Republic: IEEE, May 2011, pp. 4144–4147.

[144] P. Chen, K. K. Berthelsen, B. Bak-Jensen, and Z. Chen, "Markov model of wind power time series using Bayesian inference of transition matrix," in *2009 35th Annual Conference of IEEE Industrial Electronics.* Porto, Portugal: IEEE, Nov. 2009, pp. 627–632.

[145] C. Möhrlen and J. Zack, "Recommended practices for selecting renewable power forecasting solutions," IEA Wind Task 36, Tech. Rep., Aug. 2019.

[146] B. Efron, "Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods," *Biometrika*, vol. 68, no. 3, pp. 589–599, 1981.

[147] F. X. Diebold and R. S. Mariano, "Comparing predictive accuracy," *Journal of Business & Economic Statistics*, vol. 13, no. 3, pp. 253–263, Jul. 1995.

[148] P. Pinson, "Introducing distributed learning approaches in wind power forecasting," in *2016 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS).* Beijing: IEEE, Oct. 2016, pp. 1–6.

[149] C. Gonçalves, P. Pinson, and R. J. Bessa, "Towards data markets in renewable energy forecasting," *IEEE Transations on Sustainable Energy*, vol. 12, no. 1, pp. 533–542, Jan. 2021.

[150] J. Salmon and P. Taylor, "Errors and uncertainties associated with missing wind data and short records: Uncertainties and missing data," *Wind Energy*, vol. 17, no. 7, pp. 1111–1118, Jul. 2014.

[151] A. Coville, A. Siddiqui, and K.-O. Vogstad, "The effect of missing data on wind resource estimation," *Energy*, vol. 36, no. 7, pp. 4505–4517, Jul. 2011.

[152] Y. Hu, Y. Qiao, J. Liu, and H. Zhu, "Adaptive confidence boundary modelling of wind turbine power curve using SCADA data and its application," *IEEE Transactions on Sustainable Energy*, vol. 10, no. 3, pp. 1330–1341, Jul. 2019.

[153] S. H. Hosseini, C. Y. Tang, and J. N. Jiang, "Calibration of a Wind Farm Wind Speed Model With Incomplete Wind Data," *IEEE Transactions on Sustainable Energy*, vol. 5, no. 1, pp. 343–350, Jan. 2014.

[154] M. Martinez-Luengo, M. Shafiee, and A. Kolios, “Data management for structural integrity assessment of offshore wind turbine support structures: Data cleansing and missing data imputation,” *Ocean Engineering*, vol. 173, pp. 867–883, Feb. 2019.

[155] R. Razavi-Far and M. Saif, “Imputation of missing data for diagnosing sensor faults in a wind turbine,” in *2015 IEEE International Conference on Systems, Man, and Cybernetics.* Kowloon Tong, Hong Kong: IEEE, Oct. 2015, pp. 99–104.

[156] R. Becker and D. Thrän, “Completion of wind turbine data sets for wind integration studies applying random forests and k-nearest neighbors,” *Applied Energy*, vol. 208, pp. 252–262, Dec. 2017.

[157] Y. Mao and M. Jian, “Data completing of missing wind power data based on adaptive BP neural network,” in *2016 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS).* Beijing: IEEE, Oct. 2016, pp. 1–6.

[158] B. Lotfi, M. Mourad, M. B. Najiba, and E. Mohamed, “Treatment methodology of erroneous and missing data in wind farm dataset,” in *Conference on Systems, Signals & Devices.* Sousse: IEEE, Mar. 2011, pp. 1–6.

[159] J. G. Ibrahim and G. Molenberghs, “Missing data methods in longitudinal studies: A review,” *TEST*, vol. 18, no. 1, pp. 1–43, May 2009.

[160] D. B. Rubin, “Inference and missing data,” *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[161] N. J. Horton and K. P. Kleinman, “Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models,” *The American Statistician*, vol. 61, no. 1, pp. 79–90, Feb. 2007.

[162] J. G. Ibrahim, M.-H. Chen, S. R. Lipsitz, and A. H. Herring, “Missing-data methods for generalized linear models: A comparative review,” *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 332–346, Mar. 2005.

[163] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society. Series B*, vol. 39, no. 1, 1977.

[164] R. J. A. Little, “A test of missing completely at random for multivariate data with missing values,” *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1198–1202, Dec. 1988.

[165] R. Little, “Regression with missing X’s: A review,” *Journal of the American Statistical Association*, vol. 87, no. 420, 1992.

[166] T. D. Pigott, “A review of methods for missing data,” *Educational Research and Evaluation*, vol. 7, no. 4, pp. 353–383, Dec. 2001.

[167] G. Shmueli, “To explain or to predict?” *Statistical Science*, vol. 25, no. 3, pp. 289–310, 2010.

[168] Y. Ding and J. S. Simonoff, “An investigation of missing data methods for classification trees applied to binary response data,” *Journal of Machine Learning Research*, vol. 11, pp. 131–170, 2010.

[169] S. R. Seaman and I. R. White, “Review of inverse probability weighting for dealing with missing data,” *Statistical Methods in Medical Research*, vol. 22, no. 3, pp. 278–295, Jun. 2013.

[170] R. Little and D. B. Rubin, *Statistical Analysis of Missing Data*, 2nd ed. Wiley, 2002.

[171] J. Honaker and G. King, “What to do about missing values in time-series cross-section data,” *American Journal of Political Science*, vol. 54, no. 2, pp. 561–581, Apr. 2010.

[172] A. R. T. Donders, G. J. van der Heijden, T. Stijnen, and K. G. Moons, “Review: A gentle introduction to imputation of missing values,” *Journal of Clinical Epidemiology*, vol. 59, no. 10, pp. 1087–1091, Oct. 2006.

[173] M. Nakai and W. Ke, “Review of the methods for handling missing data in longitudinal data analysis,” *International Journal of Mathematical Analysis*, vol. 5, no. 1, pp. 1–13, 2011.

[174] D. Schunk, “A Markov chain Monte Carlo algorithm for multiple imputation in large surveys,” *Advances in Statistical Analysis*, vol. 92, no. 1, pp. 101–114, Feb. 2008.

[175] P. D. Allison, “Multiple imputation for missing data: A cautionary tale,” *Sociological Methods & Research*, vol. 28, no. 3, 2000.

[176] S. Gavankar and S. Sawarkar, “Decision tree: Review of techniques for missing values at training, testing and compatibility,” in *2015 3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS).* Kota Kinabalu, Malaysia: IEEE, Dec. 2015, pp. 122–126.

[177] F. Bashir and H.-L. Wei, “Handling missing data in multivariate time series using a vector autoregressive model based imputation (VAR-IM) algorithm: Part I: VAR-IM algorithm versus traditional methods,” in *2016 24th Mediterranean Conference on Control and Automation (MED).* Athens, Greece: IEEE, Jun. 2016, pp. 611–616.

[178] S. Zhang, Y. Hao, M. Wang, and J. H. Chow, “Multi-channel missing data recovery by exploiting the low-rank hankel structures,” in *2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP).* Curacao: IEEE, Dec. 2017, pp. 1–5.

[179] P. Gao, M. Wang, J. H. Chow, M. Berger, and L. M. Seversky, “Missing data recovery for high-dimensional signals with nonlinear low-dimensional structures,” *IEEE Transactions on Signal Processing*, vol. 65, no. 20, pp. 5421–5436, Oct. 2017.

[180] I. Pratama, A. E. Permanasari, I. Ardiyanto, and R. Indrayani, "A review of missing values handling methods on time-series data," in *2016 International Conference on Information Technology Systems and Innovation (ICITSI).* Bandung - Bali, Indonesia: IEEE, Oct. 2016, pp. 1–6.

[181] P. Liu and L. Lei, "Missing data treatment methods and NBI model," in *Sixth International Conference on Intelligent Systems Design and Applications*, vol. 1. Jian, China: IEEE, Oct. 2006, pp. 633–638.

[182] Q. Yu, Y. Miche, E. Eirola, M. van Heeswijk, E. Séverin, and A. Lendasse, "Regularized extreme learning machine for regression with missing data," *Neurocomputing*, vol. 102, pp. 45–51, Feb. 2013.

[183] F. M. Bianchi, L. Livi, K. Ø. Mikalsen, M. Kampffmeyer, and R. Jenssen, "Learning representations for multivariate time series with missing data," *Pattern Recognition*, vol. 96, p. 106973, May 2018.

[184] C. M. St. Martin, J. K. Lundquist, A. Clifton, G. S. Poulos, and S. J. Schreck, "Atmospheric turbulence affects wind turbine nacelle transfer functions," *Wind Energy Science*, vol. 2, no. 1, pp. 295–306, Jun. 2017.

[185] G. James, D. Witten, T. Hastie, and R. Tibshirani, Eds., *An Introduction to Statistical Learning: With Applications in R*, ser. Springer Texts in Statistics. New York: Springer, 2013, no. 103.

[186] J. Olauson, "ERA5: The new champion of wind power modelling?" *Renewable Energy*, vol. 126, pp. 322–331, Oct. 2018.

[187] T. Burton, T. Jenkins, D. Sharpe, and E. Bossanyi, *Wind Energy Handbook*, 2nd ed. Wiley, 2011.

[188] I. Staffell and S. Pfenninger, "Using bias-corrected reanalysis to simulate current and future wind power output," *Energy*, vol. 114, pp. 1224–1239, Nov. 2016.

[189] M. Abuella and B. Chowdhury, "Forecasting of solar power ramp events: A post-processing approach," *Renewable Energy*, vol. 133, pp. 1380–1392, Apr. 2019.

[190] C. Gallego-Castillo, "A review on the recent history of wind power ramp forecasting," *Renewable and Sustainable Energy Reviews*, p. 10, 2015.

[191] M. Cui, V. Krishnan, B.-M. Hodge, and J. Zhang, "A copula-based conditional probabilistic forecast model for wind power ramps," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 3870–3882, Jul. 2019.

[192] C. J. White, H. Carlsen, A. W. Robertson, R. J. T. Klein, J. K. Lazo, A. Kumar, F. Vitart, E. Coughlan de Perez, A. J. Ray, V. Murray, S. Bharwani, D. MacLeod, R. James, L. Fleming, A. P. Morse, B. Eggen, R. Graham, E. Kjellström, and others, "Potential applications of subseasonal-to-seasonal (S2S) predictions," *Meteorological Applications*, vol. 24, no. 3, pp. 315–325, Apr. 2017.

[193] D. K. Barrow and N. Kourentzes, "Distributions of forecasting errors of forecast combinations: Implications for inventory management," *International Journal of Production Economics*, vol. 177, pp. 24–33, Jul. 2016.

[194] G. Elliott and A. Timmermann, "Optimal forecast combinations under general loss functions and forecast error distributions," *Journal of Econometrics*, vol. 122, no. 1, pp. 47–79, Sep. 2004.

[195] R. F. Bordley, "Linear combination of forecasts with an intercept: A Bayesian approach," *Journal of Forecasting*, vol. 5, no. 4, pp. 243–249, Oct. 1986.

[196] J. M. Bates and C. W. J. Granger, "The combination of forecasts," *Operational Research Quarterly*, vol. 20, no. 4, p. 19, 1969.

[197] R. Ranjan and T. Gneiting, "Combining probability forecasts," *Journal of the Royal Statistical Society. Series B*, vol. 72, p. 21, 2010.

[198] L. L. Pauwels and A. L. Vasnev, "A note on the estimation of optimal weights for density forecast combinations," *International Journal of Forecasting*, vol. 32, no. 2, pp. 391–397, Apr. 2016.

[199] G. Claeskens, J. R. Magnus, A. L. Vasnev, and W. Wang, "The forecast combination puzzle: A simple theoretical explanation," *International Journal of Forecasting*, vol. 32, no. 3, pp. 754–762, Jul. 2016.

[200] J. Smith and K. F. Wallis, "A simple explanation of the forecast combination puzzle," *Oxford Bulletin of Economics and Statistics*, vol. 71, no. 3, pp. 331–355, Jun. 2009.

[201] S. M. Blanc and T. Setzer, "When to choose the simple average in forecast combination," *Journal of Business Research*, vol. 69, no. 10, pp. 3951–3962, Oct. 2016.

[202] A. Garratt, J. Mitchell, S. P. Vahey, and E. C. Wakerly, "Real-time inflation forecast densities from ensemble Phillips curves," *The North American Journal of Economics and Finance*, vol. 22, no. 1, pp. 77–87, Jan. 2011.

[203] V. Genre, G. Kenny, A. Meyler, and A. Timmermann, "Combining expert forecasts: Can anything beat the simple average?" *International Journal of Forecasting*, vol. 29, no. 1, pp. 108–121, Jan. 2013.

[204] E. N. Coulson, "Forecast combination in a dynamic setting," *Journal of Forecasting*, vol. 12, no. 1, pp. 63–67, Jan. 1993.

[205] C. Lemke and B. Gabrys, "Meta-learning for time series forecasting and forecast combination," *Neurocomputing*, vol. 73, no. 10-12, pp. 2006–2016, Jun. 2010.

[206] S. G. Hall and J. Mitchell, "Density forecast combination," 2004.

[207] M. H. Pesaran and A. Pick, "Forecast combination across estimation windows," *Journal of Business & Economic Statistics*, vol. 29, no. 2, pp. 307–318, Apr. 2011.

[208] W. P. Gaglianone and L. R. Lima, "Constructing optimal density forecasts from point forecast combinations: Optimal density forecasts," *Journal of Applied Econometrics*, vol. 29, no. 5, pp. 736–757, Aug. 2014.

[209] K. Lahiri, H. Peng, and Y. Zhao, "Testing the value of probability forecasts for calibrated combining," *International Journal of Forecasting*, vol. 31, no. 1, pp. 113–129, Jan. 2015.

[210] J. Browell and C. Gilbert, "ProbCast: Open-source production, evaluation and visualisation of probabilistic forecasts," in *2020 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*. Liege, Belgium: IEEE, Aug. 2020, pp. 1–6.

[211] C. Gonçalves, L. Cavalcante, M. Brito, R. J. Bessa, and J. Gama, "Forecasting conditional extreme quantiles for wind energy," *Electric Power Systems Research*, vol. 190, p. 106636, Jan. 2021.

[212] S. Wood, *Generalized Additive Models An Introduction with R*, 2nd ed. Chapman and Hall/CRC, Jun. 2017.

[213] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017, p. 9.

[214] C. Ferriera, J. Gama, L. Matias, A. Botterud, and J. Wang, "A survey on wind power ramp forecasting," Argonne National Laboratory, Tech. Rep., 2010.

[215] A. Soret, V. Torralba, N. Cortesi, I. Christel, L. Palma, A. Manrique-Suñén, L. Lledó, N. González-Reviriego, and F. J. Doblas-Reyes, "Sub-seasonal to seasonal climate predictions for wind energy forecasting," *Journal of Physics: Conference Series*, vol. 1222, p. 012009, May 2019.

[216] L. Lledó, V. Torralba, A. Soret, J. Ramon, and F. Doblas-Reyes, "Seasonal forecasts of wind power generation," *Renewable Energy*, vol. 143, pp. 91–100, Dec. 2019.

[217] C. J. White, D. I. V. Domeisen, N. Acharya, E. A. Adefisan, M. L. Anderson, S. Aura, A. A. Balogun, D. Bertram, S. Bluhm, D. J. Brayshaw, J. Browell,

D. Büeler, A. Charlton-Perez, X. Chourio, I. Christel, C. A. S. Coelho, M. J. DeFlorio, L. Delle Monache, and others, "Advances in the application and utility of subseasonal-to-seasonal predictions," *Bulletin of the American Meteorological Society*, vol. Early online release, Nov. 2021.

[218] L. Ferranti, L. Magnusson, F. Vitart, and D. S. Richardson, "How far in advance can we predict changes in large-scale flow leading to severe cold conditions over Europe?" *Quarterly Journal of the Royal Meteorological Society*, vol. 144, no. 715, pp. 1788–1802, Jul. 2018.

[219] D. Specq and L. Batté, "Do subseasonal forecasts take advantage of Madden-Julian oscillation windows of opportunity?" *Atmospheric Science Letters*, vol. 23, no. 4, Apr. 2022.

[220] M. Maier-Gerber, A. H. Fink, M. Riemer, E. Schoemer, C. Fischer, and B. Schulz, "Statistical-Dynamical Forecasting of Sub-Seasonal North Atlantic Tropical Cyclone Occurrence," *Weather and Forecasting*, Oct. 2021.

[221] A. Khalid, T. Miesse, E. Erfani, S. Thomas, C. Ferreira, K. Pegion, N. Burls, and J. Manganello, "Evaluating storm surge predictability on subseasonal timescales for flood forecasting applications: A case study for Hurricane Isabel and Katrina," *Weather and Climate Extremes*, vol. 34, p. 100378, Dec. 2021.

[222] World Meteorological Organisation, "Subseasonal-to-seasonal prediction project," s2sprediction.net, Accessed 09/12/2021.

[223] A. Soret, "Subseasonal-to-seasonal forecasting for energy," s2s4e.eu, Accessed 09/12/2021.

[224] A. Orlov, J. Sillmann, and I. Vigo, "Better seasonal forecasts for the renewable energy industry," *Nature Energy*, vol. 5, no. 2, pp. 108–110, Feb. 2020.

[225] R. Beerli, H. Wernli, and C. M. Grams, "Does the lower stratosphere provide predictability for month-ahead wind electricity generation in Europe?" *Quarterly Journal of the Royal Meteorological Society*, vol. 143, pp. 3025–3036, Oct. 2017.

[226] C. I. Garfinkel, C. Schwartz, D. I. V. Domeisen, S.-W. Son, A. H. Butler, and I. P. White, "Extratropical Atmospheric Predictability From the Quasi-Biennial Oscillation in Subseasonal Forecast Models," *Journal of Geophysical Research: Atmospheres*, vol. 123, no. 15, pp. 7855–7866, Aug. 2018.

[227] T. Jung, M. A. Kasper, T. Semmler, and S. Serrar, "Arctic influence on subseasonal midlatitude prediction," *Geophysical Research Letters*, vol. 41, no. 10, pp. 3676–3680, May 2014.

[228] L. Lledó and F. J. Doblas-Reyes, "Predicting daily mean wind speed in Europe weeks ahead from MJO status," *Monthly Weather Review*, vol. 148, no. 8, pp. 3413–3426, Aug. 2020.

[229] L. Lledó, I. Cionni, V. Torralba, P.-A. Bretonnière, and M. Samsó, "Seasonal prediction of Euro-Atlantic teleconnections from multiple systems," *Environmental Research Letters*, vol. 15, no. 7, p. 074009, Jun. 2020.

[230] B. Alonzo, P. Tankov, P. Drobinski, and R. Plougonven, "Probabilistic wind forecasting up to three months ahead using ensemble predictions for geopotential height," *International Journal of Forecasting*, vol. 36, no. 2, pp. 515–530, Apr. 2020.

[231] K. J. Lynch, D. J. Brayshaw, and A. Charlton-Perez, "Verification of European Subseasonal Wind Speed Forecasts," *Monthly Weather Review*, vol. 142, no. 8, pp. 2978–2990, Aug. 2014.

[232] S. Bayo-Besteiro, M. García-Rodríguez, X. Labandeira, and J. A. Añel, "Seasonal and subseasonal wind power characterization and forecasting for the Iberian Peninsula and the Canary Islands," *International Journal of Climatology*, vol. 42, pp. 2601–2613, Aug. 2021.

[233] H. C. Bloomfield, D. J. Brayshaw, and A. J. Charlton-Perez, "Characterizing the winter meteorological drivers of the European electricity system using targeted circulation types," *Meteorological Applications*, vol. 27, no. 1, Jan. 2020.

[234] N. Cortesi, V. Torralba, N. González-Reviriego, A. Soret, and F. J. Doblas-Reyes, “Characterization of European wind speed variability using weather regimes,” *Climate Dynamics*, vol. 53, no. 7-8, pp. 4961–4976, Oct. 2019.

[235] P. L. M. Gonzalez, D. J. Brayshaw, and F. Ziel, “A new approach to extended-range multimodel forecasting: Sequential learning algorithms,” *Quarterly Journal of the Royal Meteorological Society*, vol. 147, no. 741, pp. 4269–4282, Oct. 2021.

[236] J. Hwang, P. Orenstein, J. Cohen, K. Pfeiffer, and L. Mackey, “Improving Sub-seasonal Forecasting in the Western U.S. with Machine Learning,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* Anchorage AK USA: ACM, Jul. 2019, pp. 2325–2335.

[237] H. C. Bloomfield, D. J. Brayshaw, P. L. M. Gonzalez, and A. Charlton-Perez, “Sub-seasonal forecasts of demand, wind power and solar power generation for 28 European countries,” *Earth System Science Data*, vol. 13, pp. 2259–2274, May 2021.

[238] K. van der Wiel, H. C. Bloomfield, R. W. Lee, L. P. Stoop, R. Blackport, J. A. Screen, and F. M. Selten, “The influence of weather regimes on European renewable energy production and demand,” *Environmental Research Letters*, vol. 14, no. 9, p. 094010, Sep. 2019.

[239] Y. Dalgic, I. Lazakis, I. Dinwoodie, D. McMillan, and M. Revie, “Advanced logistics planning for offshore wind farm operation and maintenance activities,” *Ocean Engineering*, vol. 101, pp. 211–226, Jun. 2015.

[240] I. Lazakis and S. Khan, “An optimization framework for daily route planning and scheduling of maintenance vessel activities in offshore wind farms,” *Ocean Engineering*, vol. 225, p. 108752, Apr. 2021.

[241] M. Shafiee, “Maintenance logistics organization for offshore wind energy: Current progress and future perspectives,” *Renewable Energy*, vol. 77, pp. 182–193, May 2015.

[242] J. V. Taboada, V. Diaz-Casas, and X. Yu, “Reliability and maintenance management analysis on offshore wind turbines (OWTs),” *Energies*, vol. 14, no. 22, p. 7662, Nov. 2021.

[243] C. A. Irawan, D. Ouelhadj, D. Jones, M. Stålhane, and I. B. Sperstad, “Optimisation of maintenance routing and scheduling for offshore wind farms,” *European Journal of Operational Research*, vol. 256, no. 1, pp. 76–89, Jan. 2017.

[244] E. Barlow, D. Tezcaner Öztürk, M. Revie, E. Boulougouris, A. H. Day, and K. Akartunalı, “Exploring the impact of innovative developments to the installation process for an offshore wind farm,” *Ocean Engineering*, vol. 109, pp. 623–634, Nov. 2015.

[245] G. Ji, W. Wu, and B. Zhang, “Robust generation maintenance scheduling considering wind power and forced outages,” *IET Renewable Power Generation*, vol. 10, no. 5, pp. 634–641, May 2016.

[246] Q. Yu, M. Patriksson, and S. Sagitov, “Optimal scheduling of the next preventive maintenance activity for a wind farm,” Design methods, reliability and uncertainty modelling, Preprint, Dec. 2020.

[247] E. Barlow, T. Bedford, M. Revie, J. Tan, and L. Walls, “A performance-centred approach to optimising maintenance of complex systems,” *European Journal of Operational Research*, vol. 292, no. 2, pp. 579–595, Jul. 2021.

[248] J. C. Pelajo, L. E. Brandão, L. L. Gomes, and M. C. Klotzle, “Wind farm generation forecast and optimal maintenance schedule model,” *Wind Energy*, vol. 22, no. 12, pp. 1872–1890, Dec. 2019.

[249] J. Browell, I. Dinwoodie, and D. McMillan, “Forecasting for day-ahead offshore maintenance scheduling under uncertainty,” in *Risk, Reliability and Safety: Innovating Theory and Practice*, L. Walls, M. Revie, and T. Bedford, Eds. CRC Press, Sep. 2016, pp. 1137–1144.

[250] F. Vitart, C. Ardilouze, A. Bonet, A. Brookshaw, M. Chen, C. Codorean, M. Déqué, L. Ferranti, E. Fucile, M. Fuentes, H. Hendon, J. Hodgson, H.-S. Kang, A. Kumar, H. Lin, G. Liu, X. Liu, P. Malguzzi, I. Mallas, M. Manoussakis, D. Mastrangelo, C. MacLachlan, P. McLean, A. Minami, R. Mladek, T. Nakazawa, S. Najm, Y. Nie, M. Rixen, A. W. Robertson, P. Ruti, C. Sun, Y. Takaya, M. Tolstykh, F. Venuti, D. Waliser, S. Woolnough, T. Wu, D.-J. Won, H. Xiao, R. Zaripov, and L. Zhang, "The subseasonal to seasonal (S2S) prediction project database," *Bulletin of the American Meteorological Society*, vol. 98, no. 1, pp. 163–173, Jan. 2017.

[251] Met Office, "Met Office integrated data archive system (MIDAS) land and marine surface stations data (1853-current)." 2012, Accessed 08/2021.

[252] J. Ramon, L. Lledó, V. Torralba, A. Soret, and F. J. Doblas-Reyes, "What global reanalysis best represents near-surface winds?" *Quarterly Journal of the Royal Meteorological Society*, vol. 145, no. 724, pp. 3236–3251, Oct. 2019.

[253] N. Schuhen, T. L. Thorarinsdottir, and T. Gneiting, "Ensemble model output statistics for wind vectors," *Monthly Weather Review*, vol. 140, no. 10, pp. 3204–3219, Oct. 2012.

[254] R. A. Rigby and D. M. Stasinopoulos, "Generalized additive models for location, scale and shape," *Applied Statistics*, vol. 54, no. 3, pp. 507–554, 2005.

[255] M. Scheuerer, "Probabilistic quantitative precipitation forecasting using ensemble model output statistics: Probabilistic precipitation forecasting using EMOS," *Quarterly Journal of the Royal Meteorological Society*, vol. 140, no. 680, pp. 1086–1096, Apr. 2014.

[256] G. Chang, H. Lu, Y. Chang, and Y. Lee, "An improved neural network-based approach for short-term wind speed and power forecast," *Renewable Energy*, vol. 105, pp. 301–311, May 2017.

[257] A. Dupré, P. Drobinski, B. Alonzo, J. Badosa, C. Briard, and R. Plougonven, "Sub-hourly forecasting of wind speed and wind energy," *Renewable Energy*, vol. 145, pp. 2373–2379, Jan. 2020.

[258] W. Fu, K. Wang, J. Tan, and K. Zhang, "A composite framework coupling multiple feature selection, compound prediction models and novel hybrid swarm optimizer-based synchronization optimization strategy for multi-step ahead short-term wind speed forecasting," *Energy Conversion and Management*, vol. 205, p. 112461, Feb. 2020.

[259] Y. Noorollahi, M. A. Jokar, and A. Kalhor, "Using artificial neural networks for temporal and spatial wind speed forecasting in Iran," *Energy Conversion and Management*, vol. 115, pp. 17–25, May 2016.

[260] L. Ye, Y. Zhao, C. Zeng, and C. Zhang, "Short-term wind power prediction based on spatial model," *Renewable Energy*, vol. 101, pp. 1067–1074, Feb. 2017.

Appendix A

Summary of papers

Table A.1: Table of papers reviewed for the literature review in Chapter 2 and broad categorisation of types of methods used. Forecast horizons are given in this table when specified in the paper but otherwise left blank.

Paper	Horizon	Solar	Wind	Probabilistic	ML	Statistical	Decomposition	hybrid/combination	turbine-level	image based
Abuella (2017) [75]	1 day	✓			✓			✓		
Agoua (2018) [78]	0–6hrs	✓				✓				
Bellinguer (2020) [63]	0-6hrs	✓				✓				✓
Carriere (2020) [64]		✓								✓
Eseye (2018) [76]	3–24hrs	✓			✓		✓	✓		
Fliess (2019) [65]	1–60 min	✓		✓						
Golestaneh (2016) [67]		✓		✓	✓					
Harty (2019) [62]	15–60min	✓						✓		✓
Huang (2019) [70]	30–120min	✓			✓					
Lago (2018) [61]		✓			✓					✓
Lee (2016) [80]	15min–6hrs	✓		✓	✓			✓		✓
Lee (2019) [66]		✓		✓	✓					

Paper	Horizon	Solar	Wind	Probabilistic	ML	Statistical	Decomposition	hybrid/combination	turbine-level	image based
Li (2018) [77]		✓			✓		✓			
Luo (2018) [73]	5s; 10–60min	✓			✓					
Majumder (2018) [69]	15min–1 day	✓			✓		✓			
Rana (2016) [71]	5-60min	✓			✓					
Schmidt (2016) [60]	≤25min	✓			✓					✓
Sheng (2018) [68]	5min	✓			✓					
Sivaneasan (2017) [72]	m̃inutes	✓			✓					
Tang (2016) [74]		✓			✓					
Wang (2018) [59]		✓								✓
Aasim (2019) [113]	10 min		✓			✓	✓			
Bracale (2015) [128]	≤24hrs		✓	✓		✓				
Browell (2018) [89]	1–6hrs		✓			✓				
Carpinone (2015) [127]	10min		✓	✓						
Cavalcante (2017) [85]			✓			✓				
Chang (2017) [256]			✓		✓					
Chaudhary (2020) [92]			✓		✓					
Chen (2018) [106]	10min, 1hr		✓		✓	✓				
Dowell (2016) [87]	5min		✓			✓				
Du (2017) [97]	10–60min		✓		✓					
Du (2019) [98]			✓		✓		✓			
Dupre (2020) [257]	10–170min		✓			✓				
Ezzat (2020) [132]	1–12hrs		✓	✓		✓			✓	
Feng (2017) [123]	1hr		✓	✓	✓			✓		
Fu (2020) [258]	10min–1hr		✓		✓		✓	✓		
Gilbert (2020) [130]	≤48hrs		✓	✓	✓				✓	

Paper	Horizon	Solar	Wind	Probabilistic	ML	Statistical	Decomposition	hybrid/combination	turbine-level	image based
He (2019) [95]			✓		✓					
Hong (2016) [122]			✓	✓						
Hossain (2021) [103]	10–45min		✓		✓					
Jiang (2015) [94]	10min		✓		✓					
Jiang (2017) [131]			✓		✓					
Jiang (2019) [125]	10–60min		✓	✓	✓			✓		
Khodayar (2017) [99]	10min–3hrs		✓		✓					
Khodayar (2019) [102]	1–24hrs		✓		✓					
Lin (2019) [121]	≤24hrs		✓	✓		✓		✓		
Li (2020) [124]			✓	✓	✓					
Liu (2020) [111]	10–30min		✓	✓	✓	✓	✓	✓		
Liu (2018) [107]			✓		✓		✓	✓		
Lu (2018) [108]	15min–1hr		✓		✓		✓			
Ma (2017) [104]	10–60min		✓		✓					
Messner (2019) [86]	≤1hr		✓			✓				
Noorollahi (2016) [259]			✓							
Ouyang (2017) [120]	1–10hr		✓					✓		
Rodriguez (2020) [101]	10min		✓	✓	✓					
Shi (2014) [117]	15min		✓		✓			✓		
Valledecabres (2018) [136]	5min		✓	✓					✓	✓
Wang (2017) [126]	1hr		✓	✓	✓		✓			
Wang (2017) [96]	10–30min		✓		✓			✓		
Wurth (2019) [135]	1hr		✓						✓	
Xiao (2015) [118]			✓					✓		
Ye (2017) [260]			✓							

Paper	Horizon	Solar	Wind	Probabilistic	ML	Statistical	Decomposition	hybrid/combination	turbine-level	image based
Yu (2019) [134]	90min		✓		✓				✓	
Zameer (2017) [119]	1hr		✓		✓			✓		
Zhang (2017) [112]			✓		✓		✓	✓		
Zhang (2017) [114]			✓		✓		✓			
Zhang (2017) [109]	3–24hrs		✓		✓	✓	✓	✓		
Zhang (2018) [91]	≤24hrs		✓	✓		✓				
Zhang (2019) [105]			✓		✓					
Zhang (2019) [129]			✓	✓	✓					
Zhao (2018) [88]	≤90min		✓			✓				
Zhou (2016) [84]	15min–4hrs		✓			✓		✓		
totals (% of studies)		29%	71%	26%	64%	24%	18%	25%	7%	11%

Appendix B

Bayesian Markov chain derivation

For a Bayesian approach, forecasts are generated by integrating over the whole distribution of values for the transition matrix parameters:

$$p(y|x) = \int p(y|\Theta)p(\Theta|x)d\Theta \tag{B.1}$$

with

$$p(\Theta|x) \propto p(\Theta)p(x|\Theta) \tag{B.2}$$

by Bayes theorem. $p(\Theta)$ is the prior distribution, and $p(x|\Theta)$ is the likelihood function (eq. 2.5). $p(y|\Theta)$ is the multiplication of the (observed) forecast input state by the transition matrix.

A Dirichlet prior is chosen [144] as it is the conjugate prior of the multinomial distribution (i.e. it takes the same form as the likelihood function) and incorporates the constraint that the row-wise sum of transition matrix probabilities is one. The dirichlet distribution for row l of the transition matrix is defined as

$$p(\Theta) = f(\theta_{l1}, ..., \theta_{lK}; \alpha_1, ..., \alpha_K) = \frac{1}{B(\alpha)} \prod_{j=1}^{K} \theta_{lj}^{\alpha_j - 1} \quad ; \tag{B.3}$$

$\sum_{j=1}^{K} \theta_{lj} = 1$ and all $\theta_{lj} \geq 0$. Ignoring normalisation constants and substituting Equations 2.5 and B.3 into eq. B.1, assuming we know the previous forecast state is l,

$$p(y|x) \propto \int \begin{pmatrix} \theta_{l1} \\ \theta_{l2} \\ ... \\ \theta_{l(K-1)} \\ \theta_{lK} \end{pmatrix} \prod_{i}^{K} \prod_{j}^{K} \theta_{ij}^{N_{ij}+\alpha_{ij}-1} d\Theta \quad . \tag{B.4}$$

For an arbitrary forecast input state, the integral in Equation B.1 would be over all K^2 elements of the transition matrix; however, if we know the input state is l and set the transition matrix up directly for the desired forecast horizon (rather than iteratively forecasting multiple steps ahead), we need only integrate over elements of Θ in row l (since all other elements are not contained in the vector element of the integral, they integrate to a constant and therefore can be ignored as long as the state probabilities are normalised after the integration). Also including the constraint that the row-wise sum of elements in Θ is equal to one, Equation (B.4) reduces to the integral over $K-1$ parameters:

$$p(y|x) \propto \int \begin{pmatrix} \theta_{l1} \\ \theta_{l2} \\ ... \\ \theta_{l(K-1)} \\ 1-\theta_{l1}-\theta_{l2}...-\theta_{l(K-1)} \end{pmatrix} \times$$
$$\theta_{l1}^{N_{l1}+\alpha_{l1}-1} \times \theta_{l2}^{N_{l2}+\alpha_{l2}-1} ...(1-\theta_{l1}-\theta_{l2}...-\theta_{l(K-1)})^{N_{lK}+\alpha_{lK}-1} \; d\theta_{l1}d\theta_{l2}...d\theta_{l(K-1)} \quad . \tag{B.5}$$

Each element in the vector integration is of the general form

$$\int \theta_{l1}^{a} \times \theta_{l1}^{b} ... \times (1-\theta_{l1}-\theta_{l2}...-\theta_{l(K-1)})^{j} \; d\theta_{l1}d\theta_{l2}...d\theta_{l(K-1)} \quad . \tag{B.6}$$

Each parameter may then be integrated over. The following example purely considers the integration over θ_{l1} and in the case where K=4, but the integrals over all other parameters and for other values of K follow the same process. The integration limits are from 0 to $(1-\theta_{l2}-\theta_{l3})$ since the sum of all parameters must be one. We can make the substitution $\theta_{l1} = (1-\theta_{l2}-\theta_{l3}) \times u$ where all θ_{lj} parameters yet to be integrated are included on the RHS. This leads to

$$\frac{\theta_{l2}^{b}\ \theta_{l3}^{c}}{1-\theta_{l2}-\theta_{l3}} \int_{0}^{u=1} [(1-\theta_{l2}-\theta_{l3})u]^{a} \times [(1-\theta_{l2}-\theta_{l3})(1-u)]^{d}\, du \tag{B.7}$$

$$= \theta_{l2}^{b}\ \theta_{l3}^{c}\ (1-\theta_{l2}-\theta_{l3})^{a+b-1} \int_{0}^{u=1} u^{a}(1-u)^{d} du \quad . \tag{B.8}$$

Now this is in the same form as the definition of the Beta distribution, so we get

$$= \theta_{l2}^{b}\ \theta_{l3}^{c}\ (1-\theta_{l2}-\theta_{l3})^{a+d-1}\ B(a-1, d-1) \tag{B.9}$$

which can then be further integrated over the remaining θ_{l2},θ_{l3}. Writing the beta function in terms of gamma functions gives the result for the general integral in Equation (B.6):

$$\frac{\Gamma(a-1)\Gamma(b-1)\Gamma(c-1)...\Gamma(j-1)}{\Gamma(a+b+c+...+j-(K-1)} \quad . \tag{B.10}$$

Combining this with the vector in Equation B.5, using the recursion relation $\Gamma(a) = (a-1)\Gamma(a-1)$ and cancelling common factors between the vector elements, the unnormalised state vector resulting from the integration is

$$p(y|x) \propto \begin{pmatrix} N_{l1}+\alpha_{l1}-1 \\ N_{l2}+\alpha_{l2}-1 \\ N_{l3}+\alpha_{l4}-1 \\ ... \\ N_{lK}+\alpha_{lK}-1 \end{pmatrix} \tag{B.11}$$

and this then just has to be normalised so that the elements sum to 1 (i.e. the total probability across all states is one). To specify the K^2 prior values needed we utilise the assumption that transitions to more similar, or closer, power levels are more likely than large jumps in power between time steps. Then the priors are constrained by

$$\alpha_{lj} = K - |l - j| \quad . \tag{B.12}$$

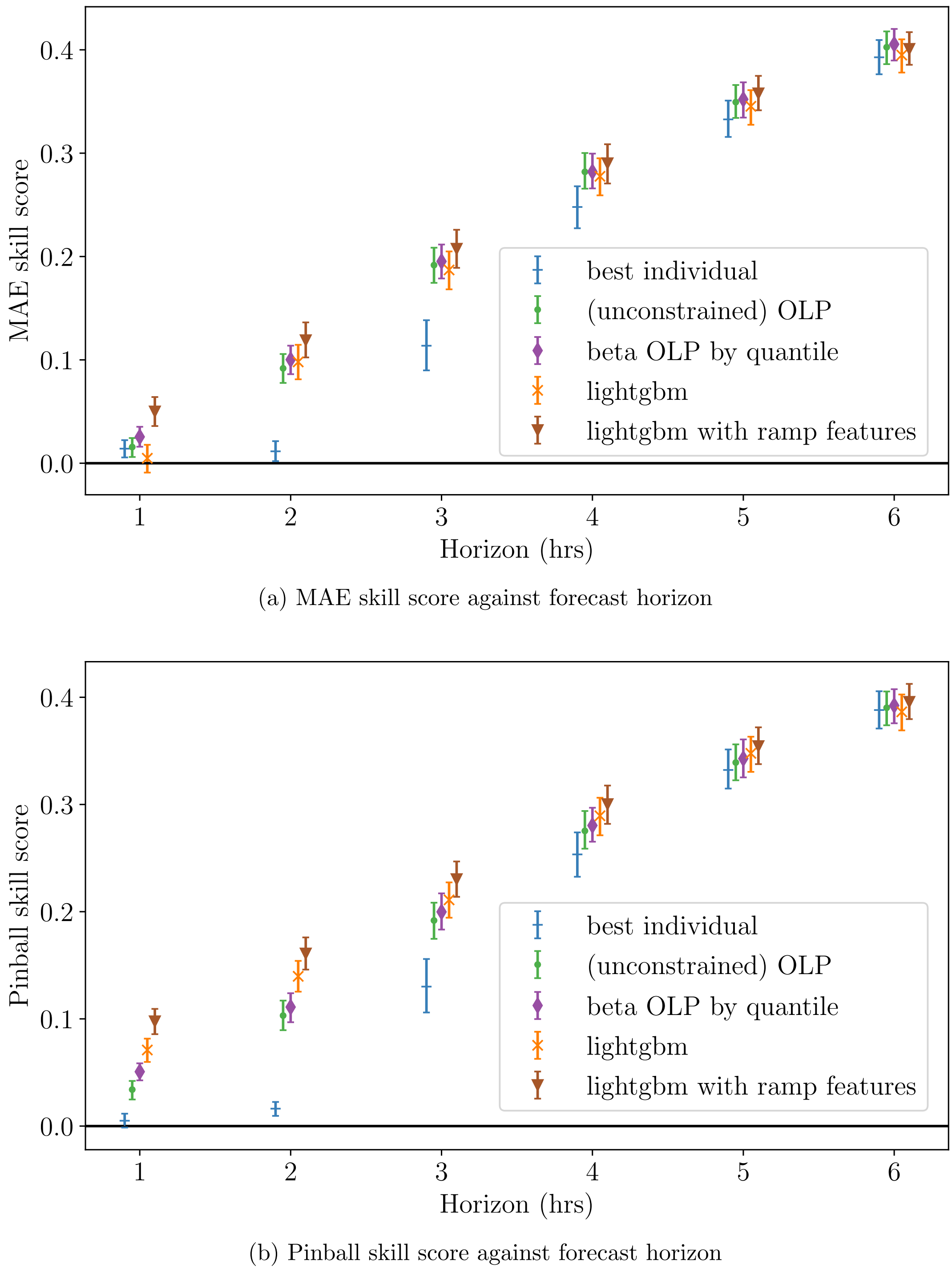
These don't need to be normalised as α_{lj} is combined with transition counts in Equation B.11 before the whole vector is normalised. This constraint on the priors defines the relative size of each α_{lj} to the others, but we can still optimise the relative size of the priors to the counts as a whole by introducing a 'scaling factor' c:

$$p(y|x) \propto \begin{pmatrix} cN_{l1} + \alpha_{l1} - 1 \\ cN_{l2} + \alpha_{l2} - 1 \\ cN_{l3} + \alpha_{l4} - 1 \\ ... \\ cN_{lK} + \alpha_{lK} - 1 \end{pmatrix} \quad . \tag{B.13}$$

This means we can control the importance of the observed transition counts, relative to the priors, with just one parameter c.

Appendix C

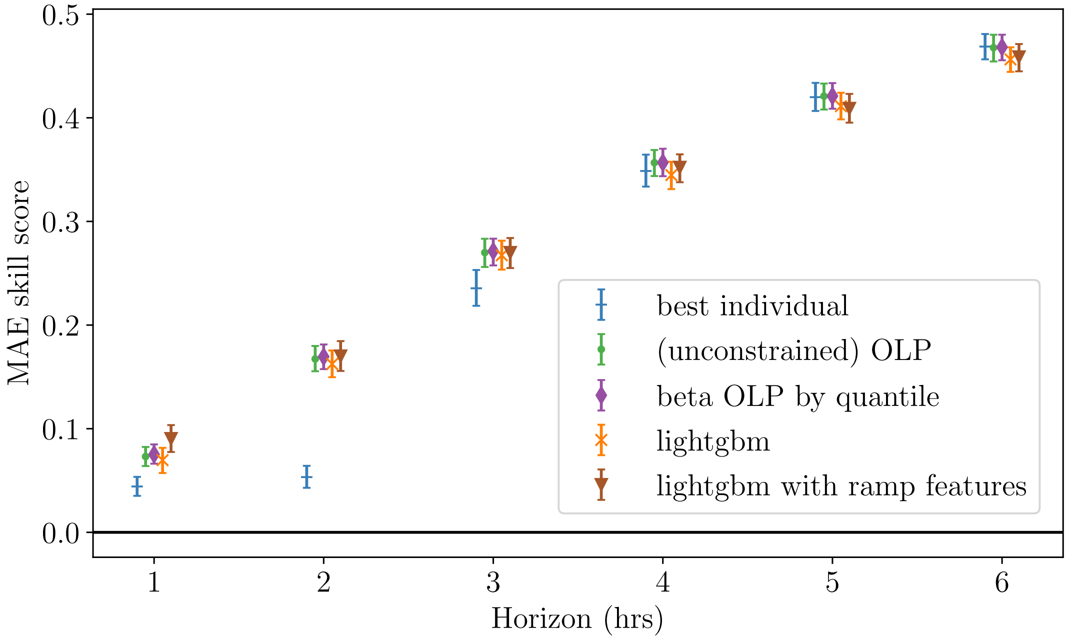
Forecast combination evaluation at all zones



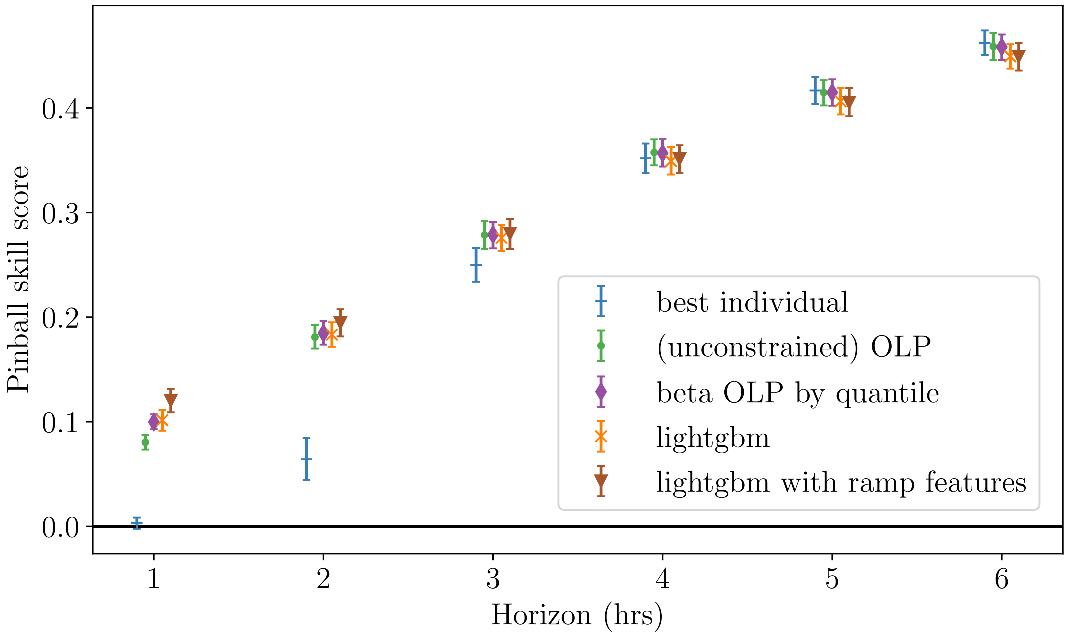
(a) MAE skill score against forecast horizon

(b) Pinball skill score against forecast horizon

Figure C.1: Skill scores of combination models at zone 2, relative to probabilistic persistence model. The 95% interval of bootstrap samples is shown. Positive values indicate improvement over persistence. Constrained OLP is omitted as it is outperformed by unconstrained OLP for every zone and horizon.

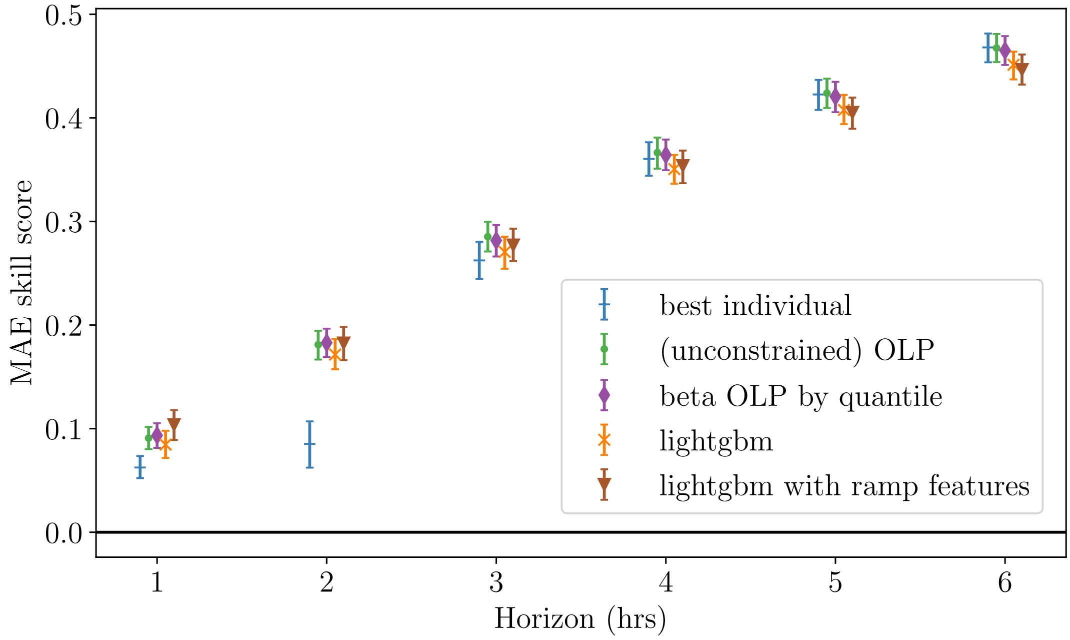


(a) MAE skill score against forecast horizon

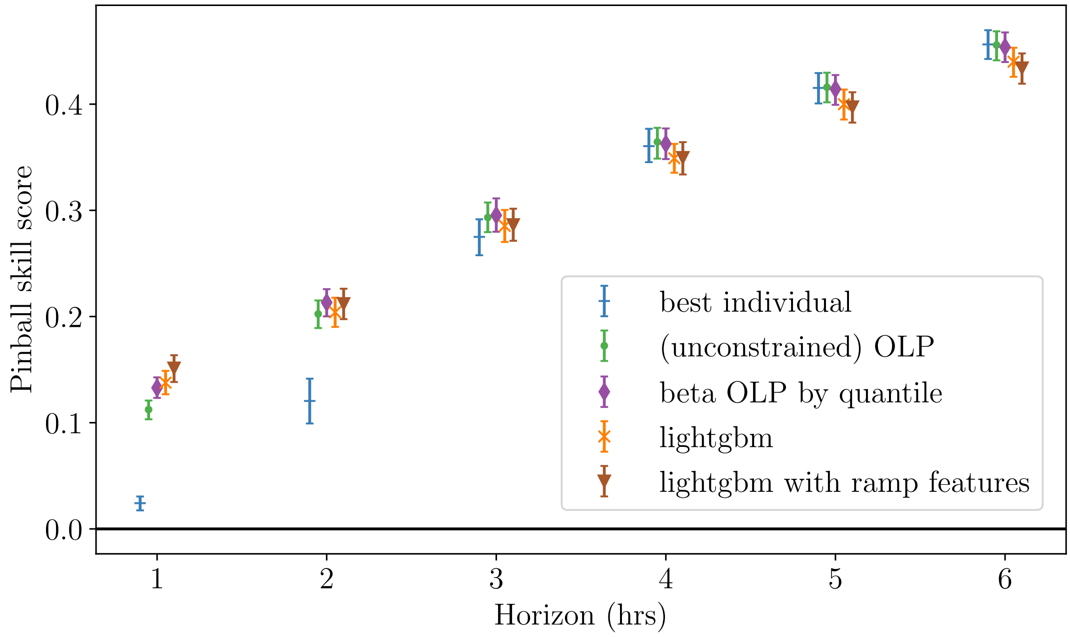


(b) Pinball skill score against forecast horizon

Figure C.2: Skill scores of combination models at zone 3, relative to probabilistic persistence model

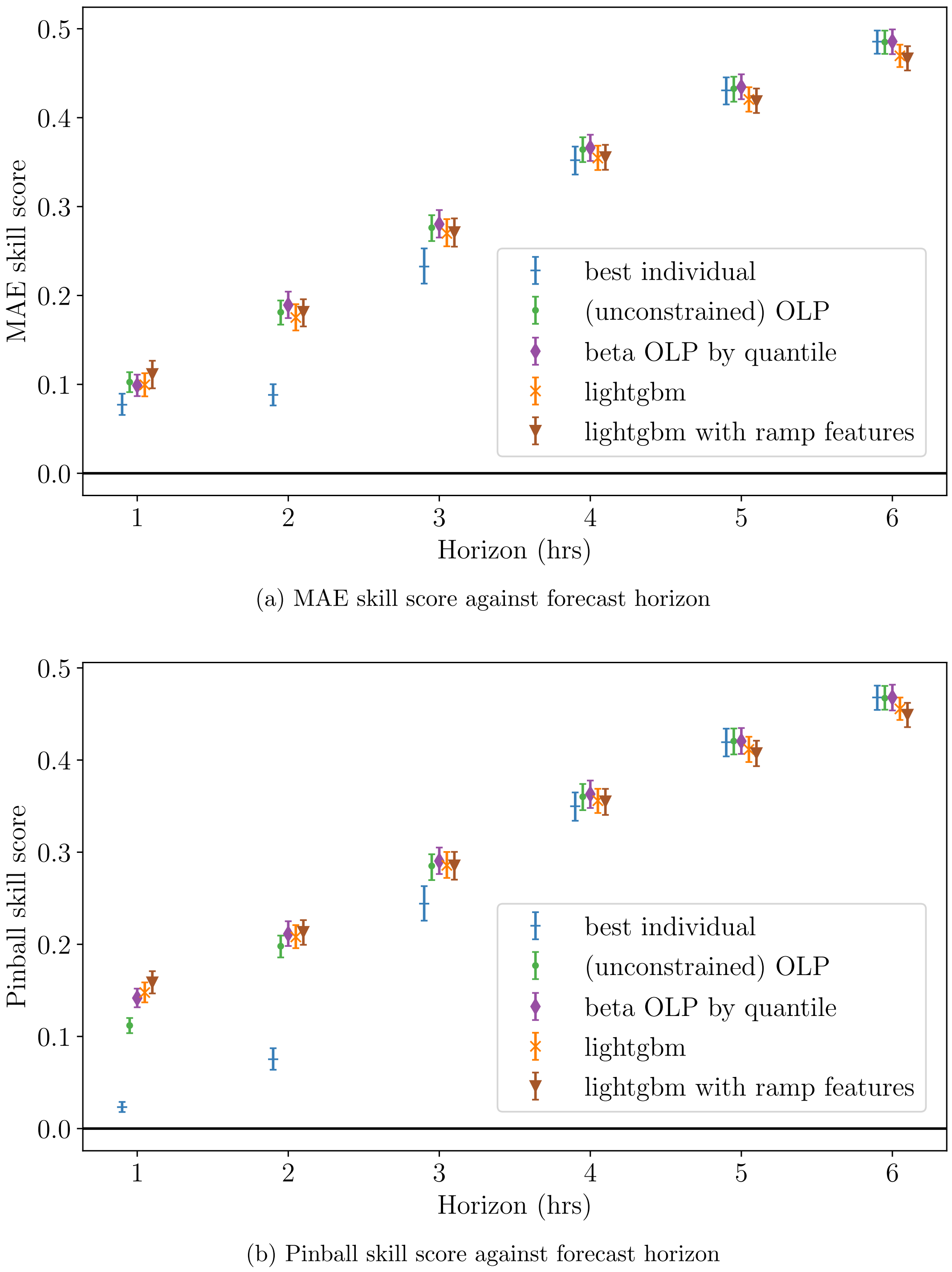


(a) MAE skill score against forecast horizon



(b) Pinball skill score against forecast horizon

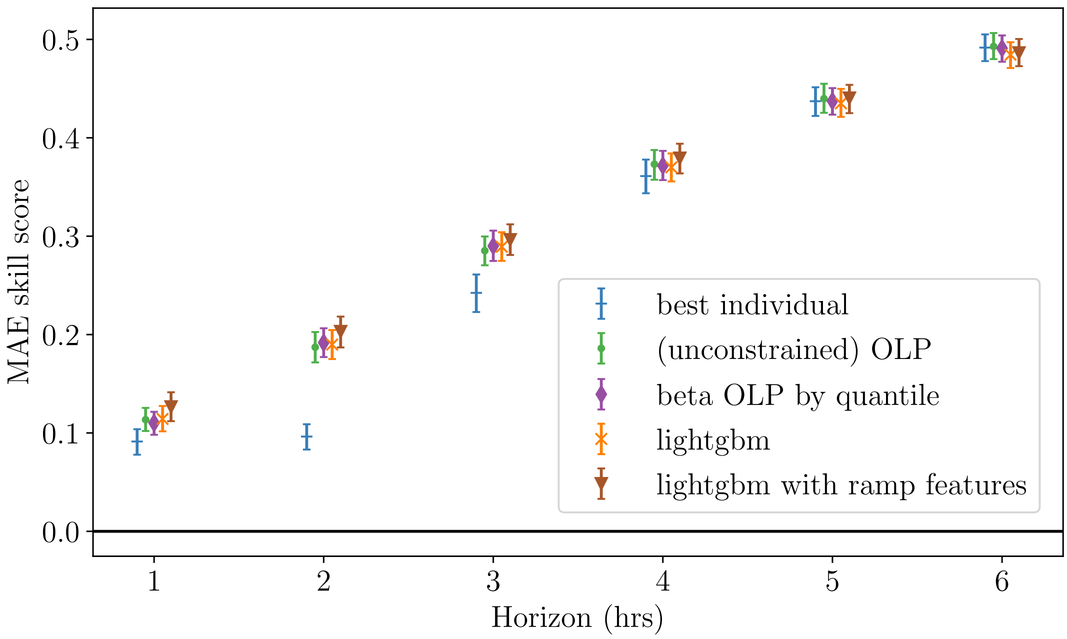
Figure C.3: Skill scores of combination models at zone 4, relative to probabilistic persistence model



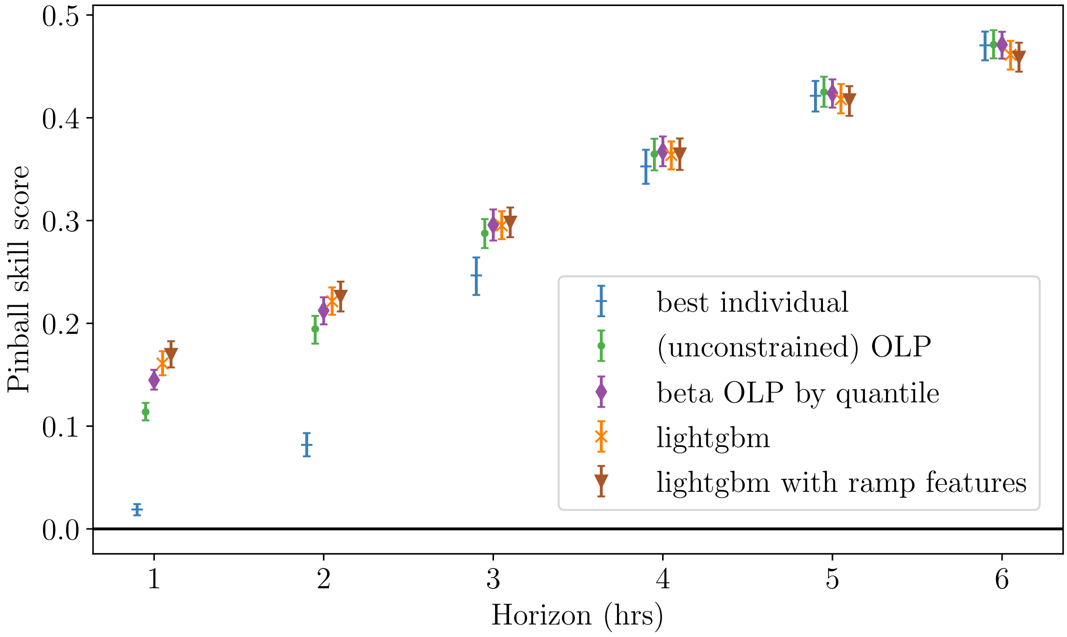
(a) MAE skill score against forecast horizon

(b) Pinball skill score against forecast horizon

Figure C.4: Skill scores of combination models at zone 5, relative to probabilistic persistence model

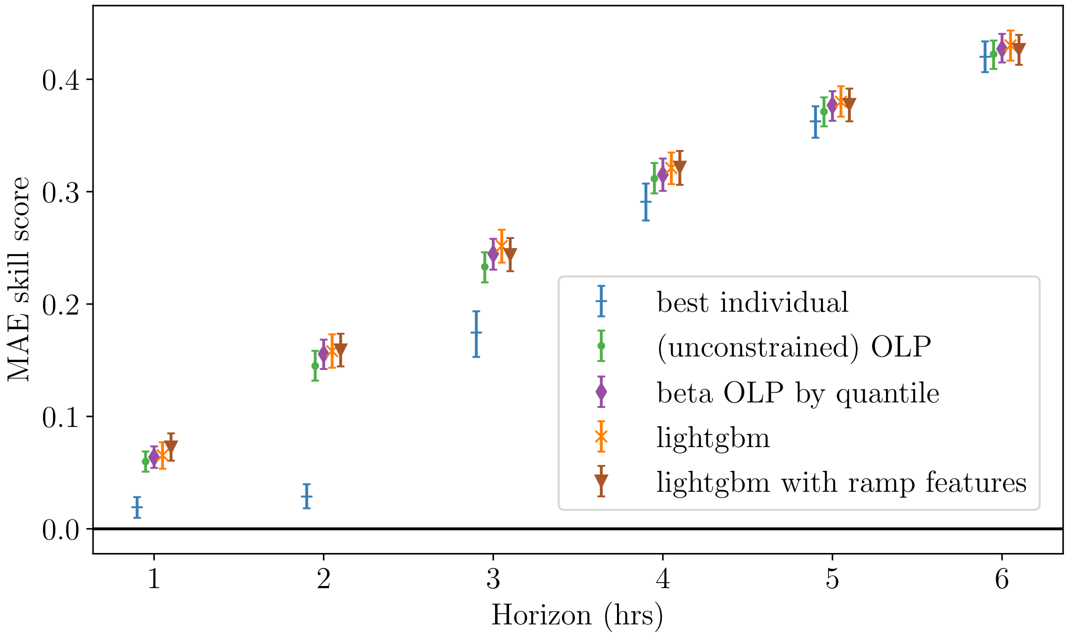


(a) MAE skill score against forecast horizon

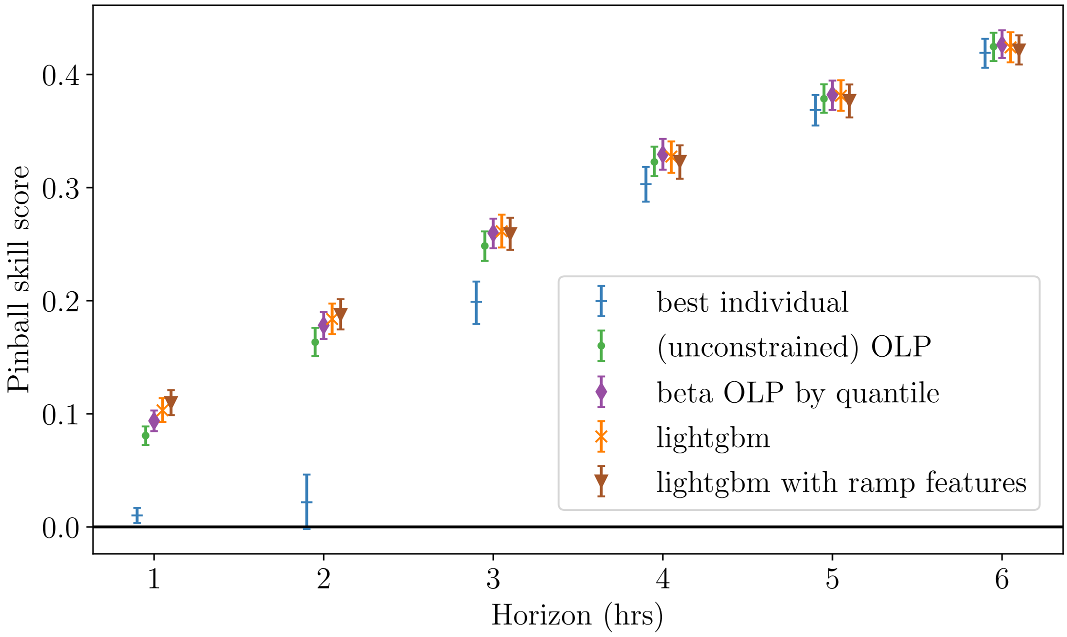


(b) Pinball skill score against forecast horizon

Figure C.5: Skill scores of combination models at zone 6, relative to probabilistic persistence model

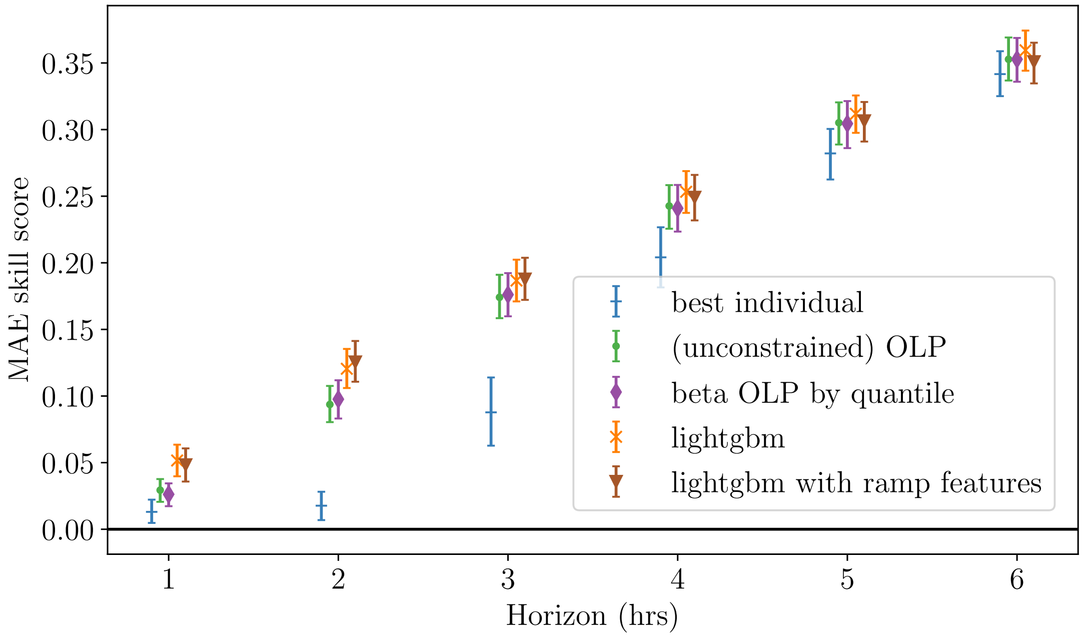


(a) MAE skill score against forecast horizon

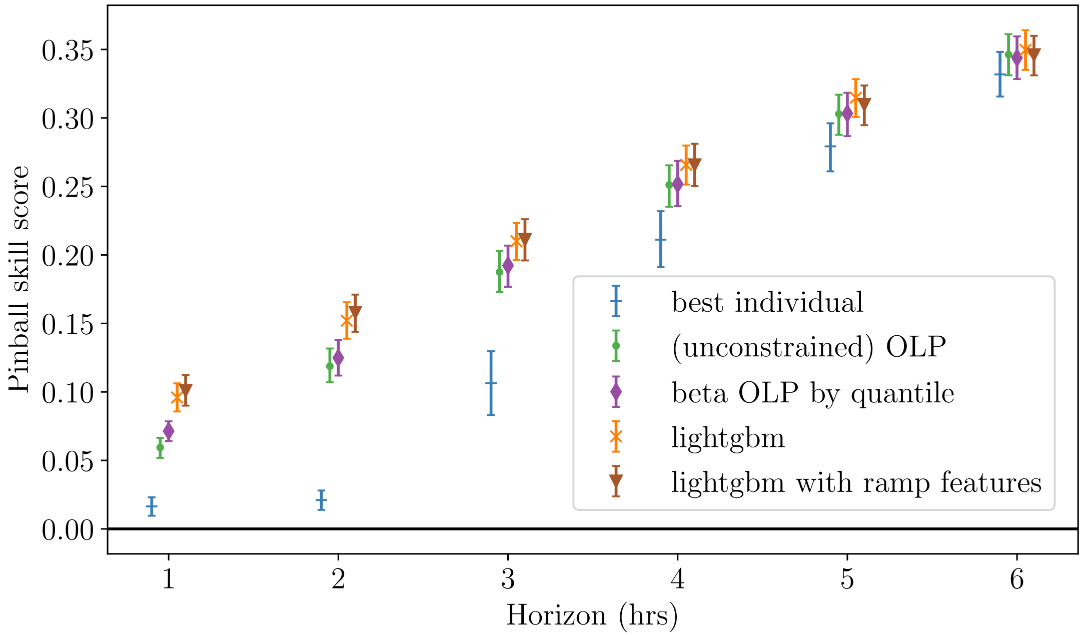


(b) Pinball skill score against forecast horizon

Figure C.6: Skill scores of combination models at zone 7, relative to probabilistic persistence model

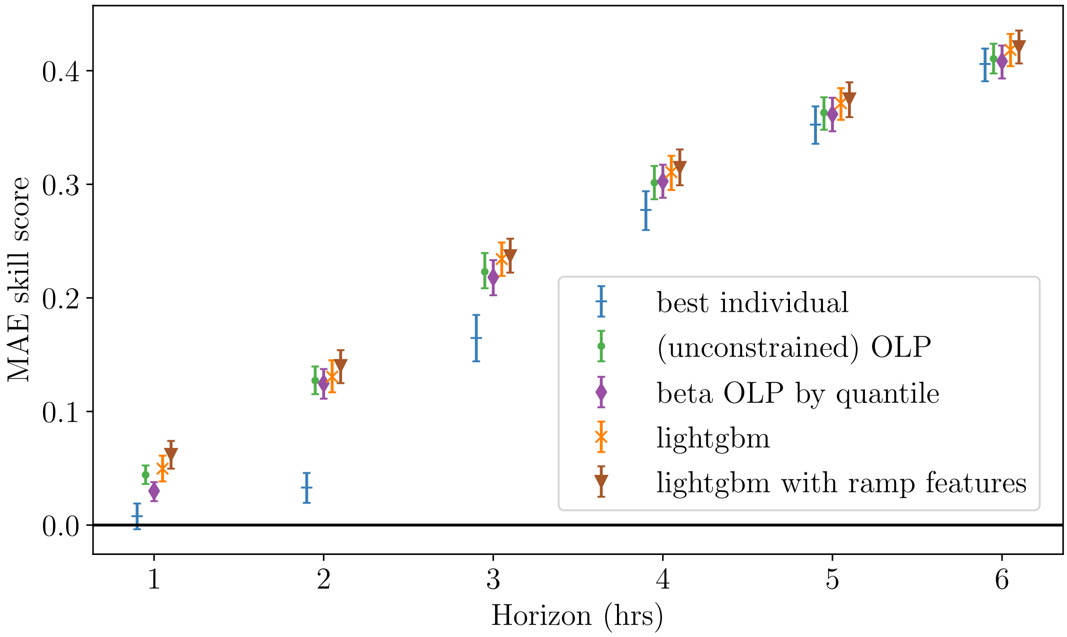


(a) MAE skill score against forecast horizon

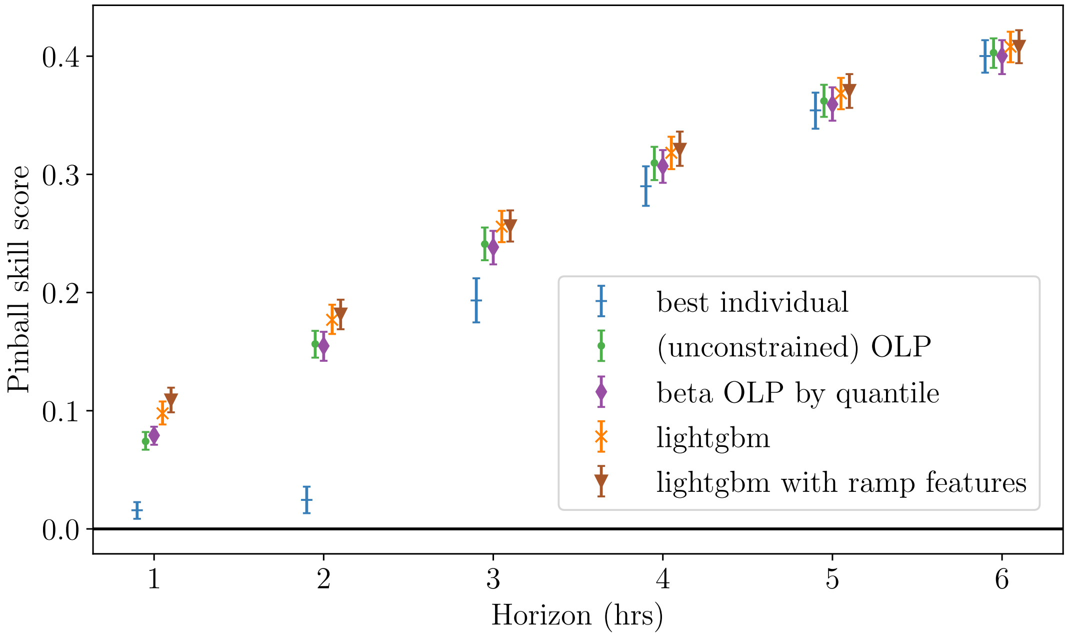


(b) Pinball skill score against forecast horizon

Figure C.7: Skill scores of combination models at zone 8, relative to probabilistic persistence model

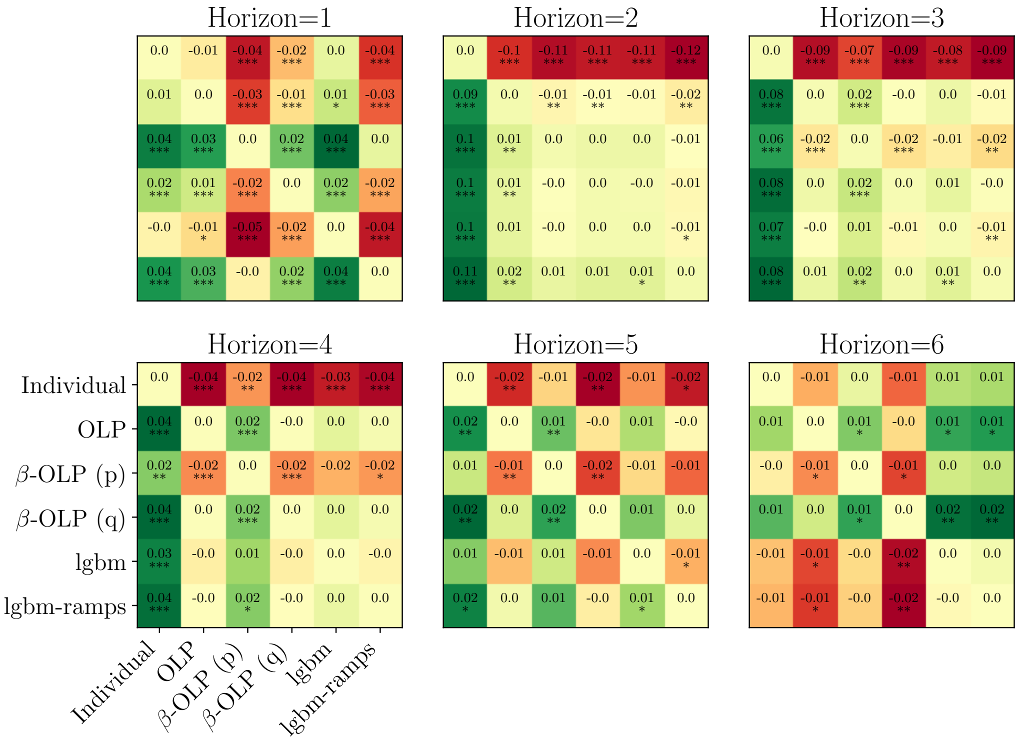


(a) MAE skill score against forecast horizon

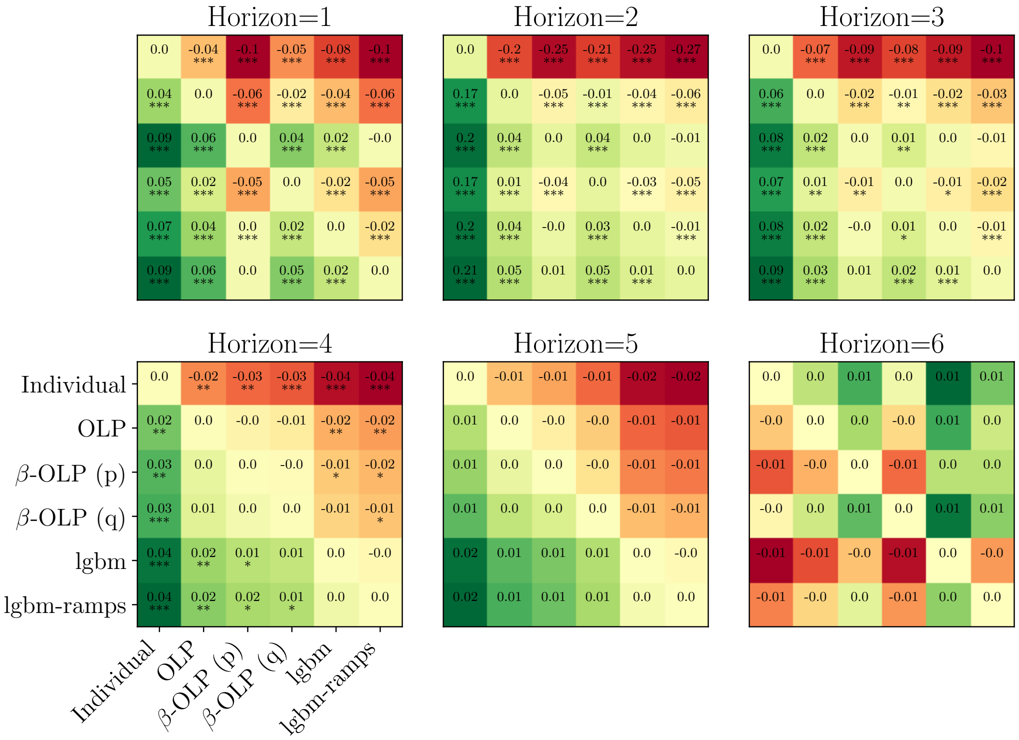


(b) Pinball skill score against forecast horizon

Figure C.8: Skill scores of combination models at zone 9, relative to probabilistic persistence model

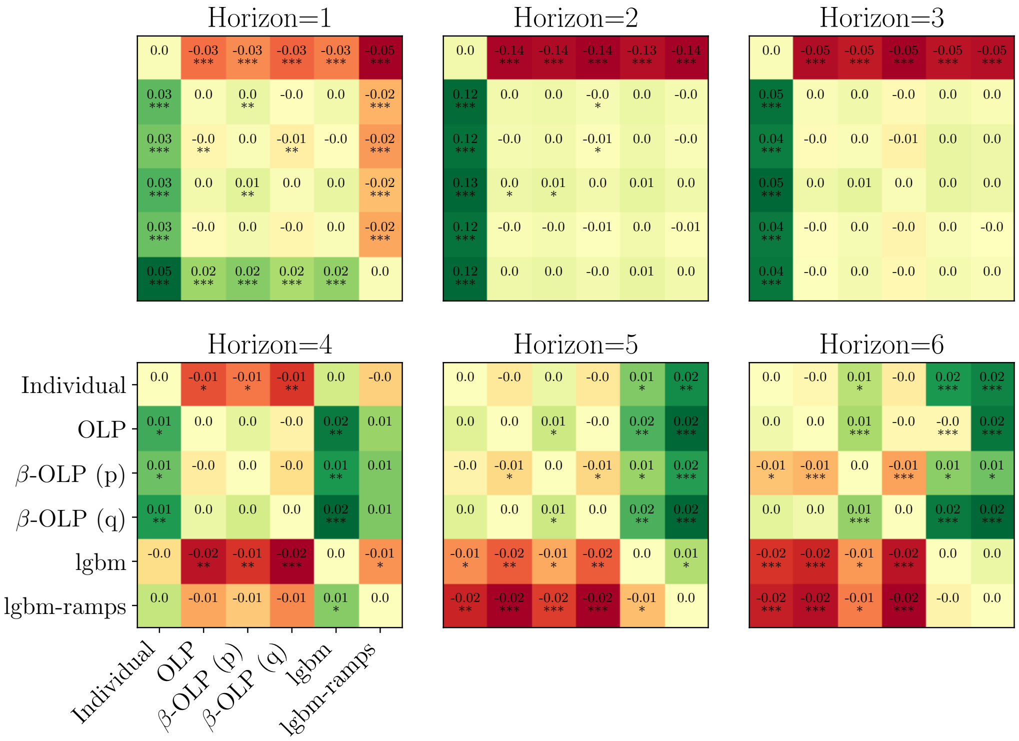


(a) Matrix of MAE skill scores. *(Figure continues onto next page)*

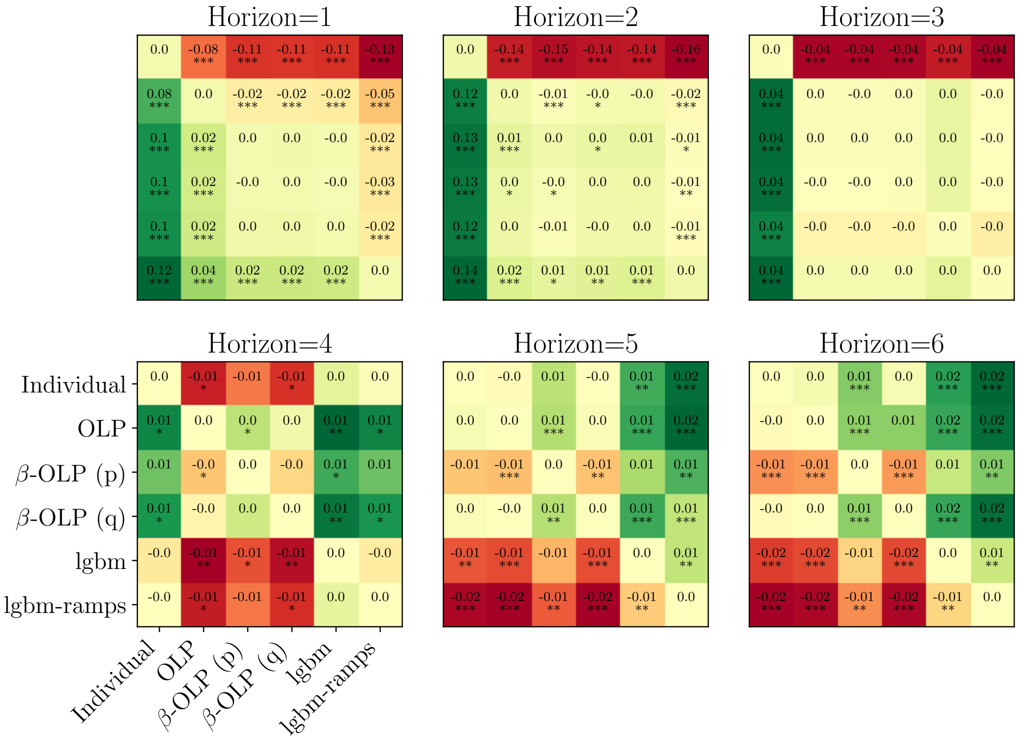


(b) Matrix of Pinball skill scores

Figure C.9: Matrices of skill scores for each pair of forecast combination methods for zone 2. Asterisks denote significance level of Diebold-Mariano test values: *=0.05; **=0.01; ***=0.001.

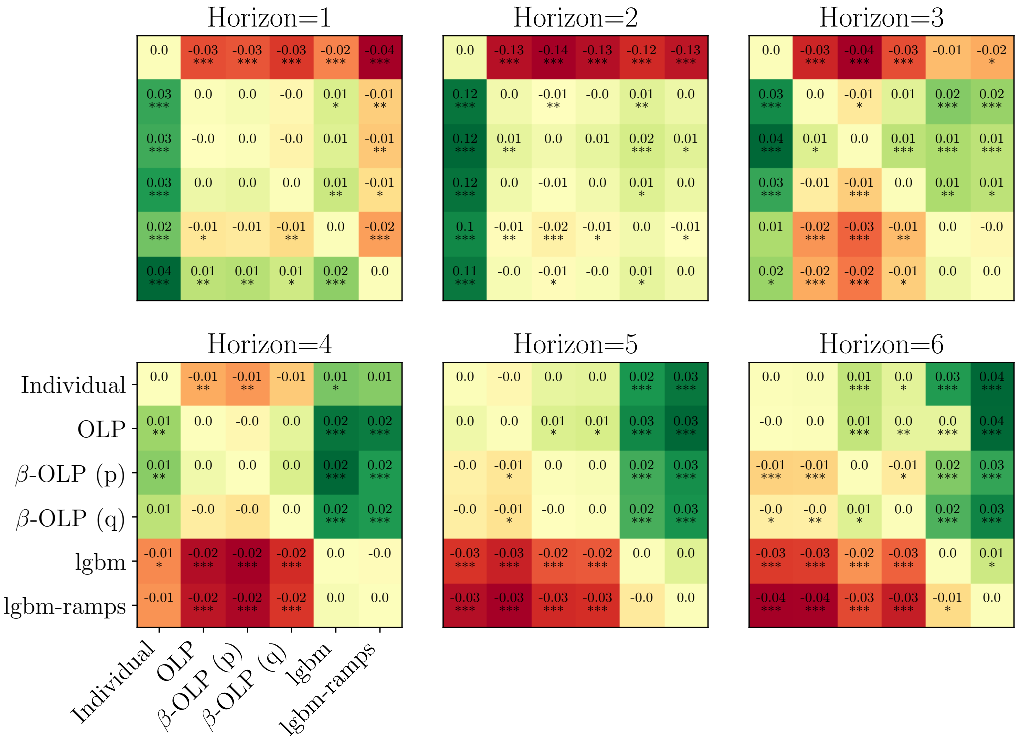


(a) Matrix of MAE skill scores. *(Figure continues onto next page)*

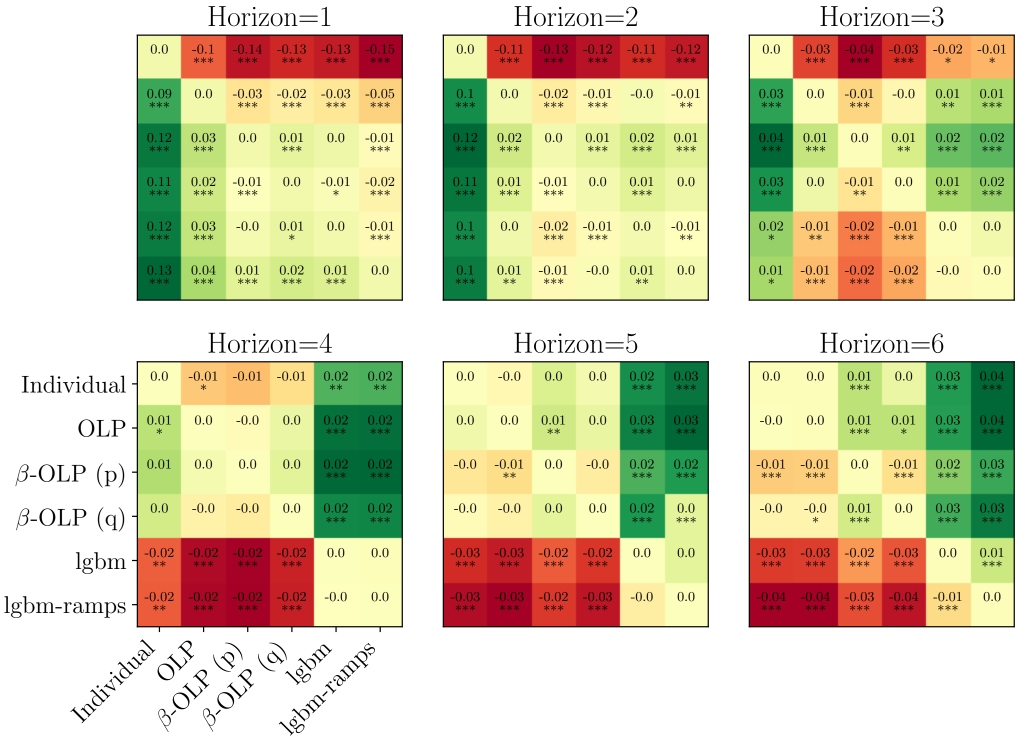


(b) Matrix of Pinball skill scores

Figure C.10: Matrices of skill scores for each pair of forecast combination methods for zone 3. Asterisks denote significance level of Diebold-Mariano test values: *=0.05; **=0.01; ***=0.001.



(a) Matrix of MAE skill scores. *(Figure continues onto next page)*



Horizon=1

	Individual	OLP	β-OLP (p)	β-OLP (q)	lgbm	lgbm-ramps
Individual	0.0	-0.1 ***	-0.14 ***	-0.13 ***	-0.13 ***	-0.15 ***
OLP	0.09 ***	0.0	-0.03 ***	-0.02 ***	-0.03 ***	-0.05 ***
β-OLP (p)	0.12 ***	0.03 ***	0.0	0.01 ***	0.0	-0.01 ***
β-OLP (q)	0.11 ***	0.02 ***	-0.01 ***	0.0	-0.01 *	-0.02 ***
lgbm	0.12 ***	0.03 ***	-0.0	0.01 *	0.0	-0.01 ***
lgbm-ramps	0.13 ***	0.04 ***	0.01 ***	0.02 ***	0.01 ***	0.0

Horizon=2

	Individual	OLP	β-OLP (p)	β-OLP (q)	lgbm	lgbm-ramps
Individual	0.0	-0.11 ***	-0.13 ***	-0.12 ***	-0.11 ***	-0.12 ***
OLP	0.1 ***	0.0	-0.02 ***	-0.01 ***	-0.0	-0.01 **
β-OLP (p)	0.12 ***	0.02 ***	0.0	0.01 ***	0.02 ***	0.01 ***
β-OLP (q)	0.11 ***	0.01 ***	-0.01 ***	0.0	0.01 ***	0.0
lgbm	0.1 ***	0.0	-0.02 ***	-0.01 ***	0.0	-0.01 **
lgbm-ramps	0.1 ***	0.01 **	-0.01 ***	-0.0	0.01 **	0.0

Horizon=3

	Individual	OLP	β-OLP (p)	β-OLP (q)	lgbm	lgbm-ramps
Individual	0.0	-0.03 ***	-0.04 ***	-0.03 ***	-0.02 *	-0.01 *
OLP	0.03 ***	0.0	-0.01 ***	-0.0	0.01 **	0.01 ***
β-OLP (p)	0.04 ***	0.01 ***	0.0	0.01 **	0.02 ***	0.02 ***
β-OLP (q)	0.03 ***	0.0	-0.01 **	0.0	0.01 ***	0.02 ***
lgbm	0.02 *	-0.01 **	-0.02 ***	-0.01 ***	0.0	0.0
lgbm-ramps	0.01 *	-0.01 ***	-0.02 ***	-0.02 ***	-0.0	0.0

Horizon=4

	Individual	OLP	β-OLP (p)	β-OLP (q)	lgbm	lgbm-ramps
Individual	0.0	-0.01 *	-0.01	-0.01	0.02 **	0.02 **
OLP	0.01 *	0.0	-0.0	0.0	0.02 ***	0.02 ***
β-OLP (p)	0.01	0.0	0.0	0.0	0.02 ***	0.02 ***
β-OLP (q)	0.0	-0.0	-0.0	0.0	0.02 ***	0.02 ***
lgbm	-0.02 **	-0.02 ***	-0.02 ***	-0.02 ***	0.0	0.0
lgbm-ramps	-0.02 **	-0.02 ***	-0.02 ***	-0.02 ***	-0.0	0.0

Horizon=5

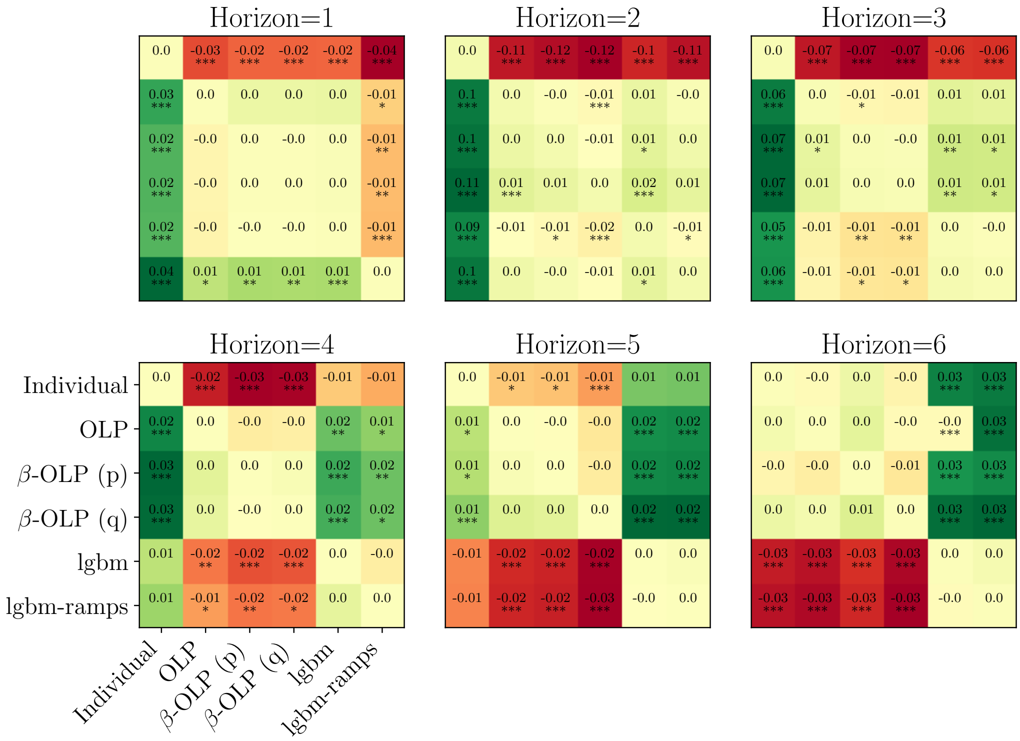
	Individual	OLP	β-OLP (p)	β-OLP (q)	lgbm	lgbm-ramps
Individual	0.0	-0.0	0.0	0.0	0.02 ***	0.03 ***
OLP	0.0	0.0	0.01 **	0.0	0.03 ***	0.03 ***
β-OLP (p)	-0.0	-0.01 **	0.0	-0.0	0.02 ***	0.02 ***
β-OLP (q)	-0.0	-0.0	0.0	0.0	0.02 ***	0.0 ***
lgbm	-0.03 ***	-0.03 ***	-0.02 ***	-0.02 ***	0.0	0.0
lgbm-ramps	-0.03 ***	-0.03 ***	-0.02 ***	-0.03 ***	-0.0	0.0

Horizon=6

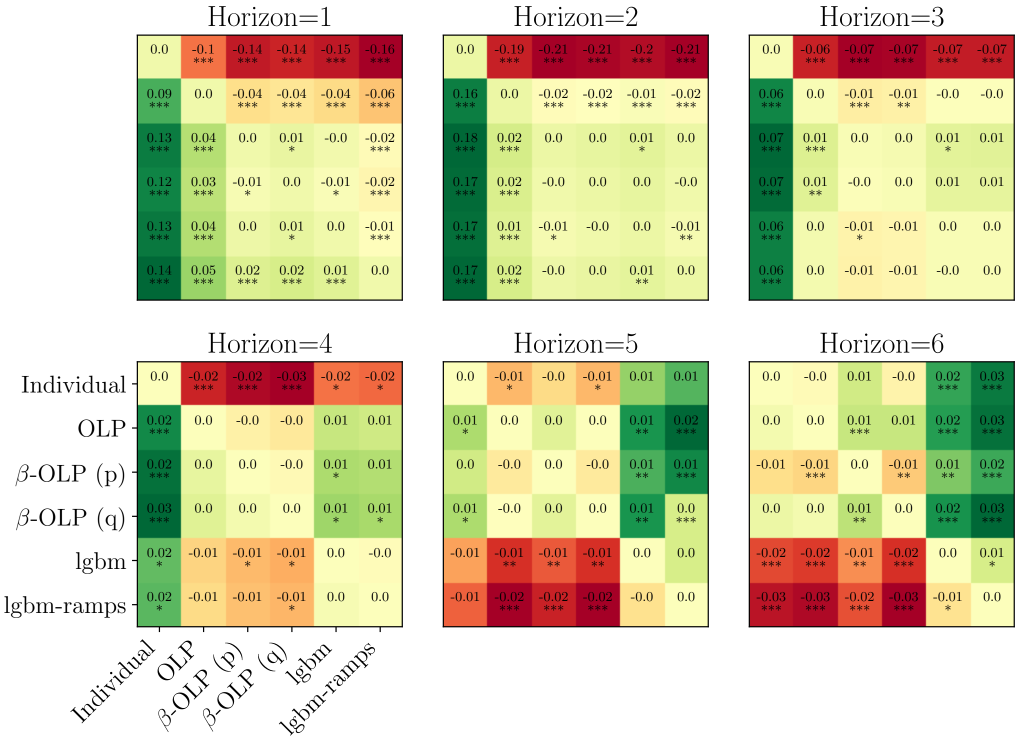
	Individual	OLP	β-OLP (p)	β-OLP (q)	lgbm	lgbm-ramps
Individual	0.0	0.0	0.01 ***	0.0	0.03 ***	0.04 ***
OLP	-0.0	0.0	0.01 ***	0.01 *	0.03 ***	0.04 ***
β-OLP (p)	-0.01 ***	-0.01 ***	0.0	-0.01 ***	0.02 ***	0.03 ***
β-OLP (q)	-0.0	-0.0 *	0.01 ***	0.0	0.03 ***	0.03 ***
lgbm	-0.03 ***	-0.03 ***	-0.02 ***	-0.03 ***	0.0	0.01 ***
lgbm-ramps	-0.04 ***	-0.04 ***	-0.03 ***	-0.04 ***	-0.01 ***	0.0

(b) Matrix of Pinball skill scores

Figure C.11: Matrices of skill scores for each pair of forecast combination methods for zone 4. Asterisks denote significance level of Diebold-Mariano test values: *=0.05; **=0.01; ***=0.001.

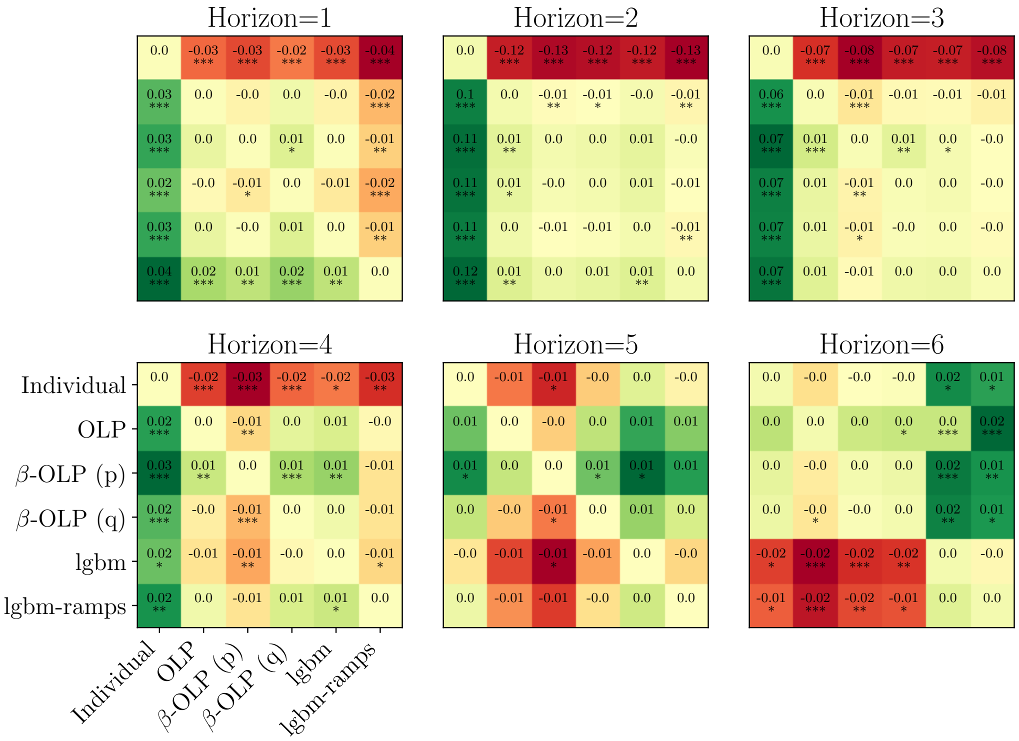


(a) Matrix of MAE skill scores. *(Figure continues onto next page)*

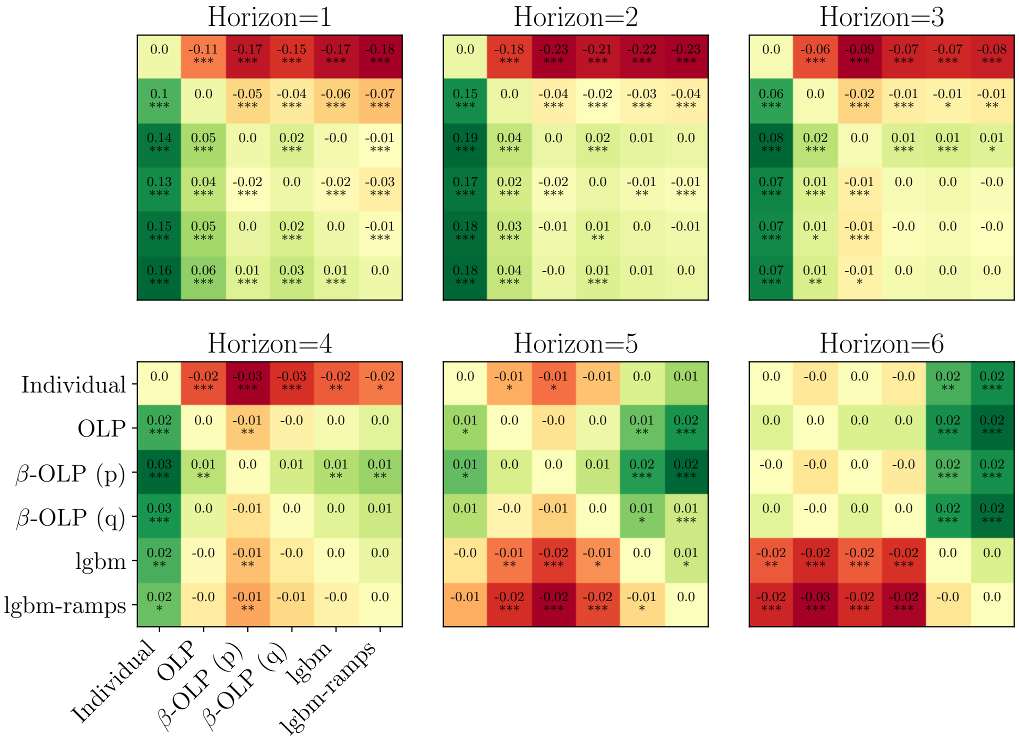


(b) Matrix of Pinball skill scores

Figure C.12: Matrices of skill scores for each pair of forecast combination methods for zone 5. Asterisks denote significance level of Diebold-Mariano test values: *=0.05; **=0.01; ***=0.001.

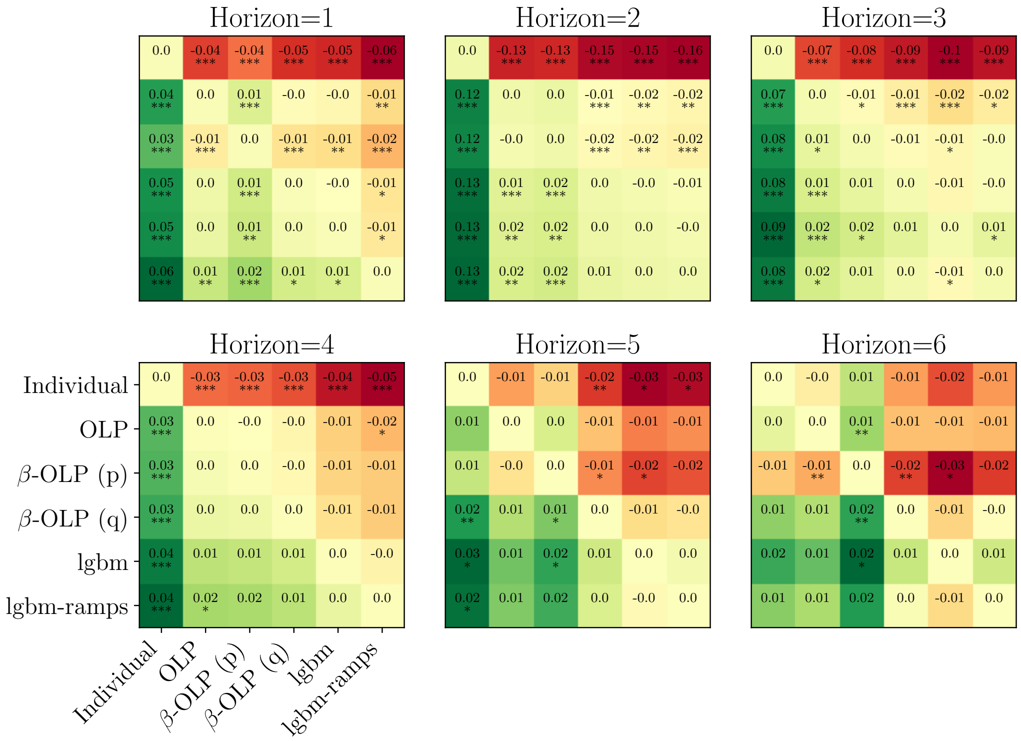


(a) Matrix of MAE skill scores. *(Figure continues onto next page)*

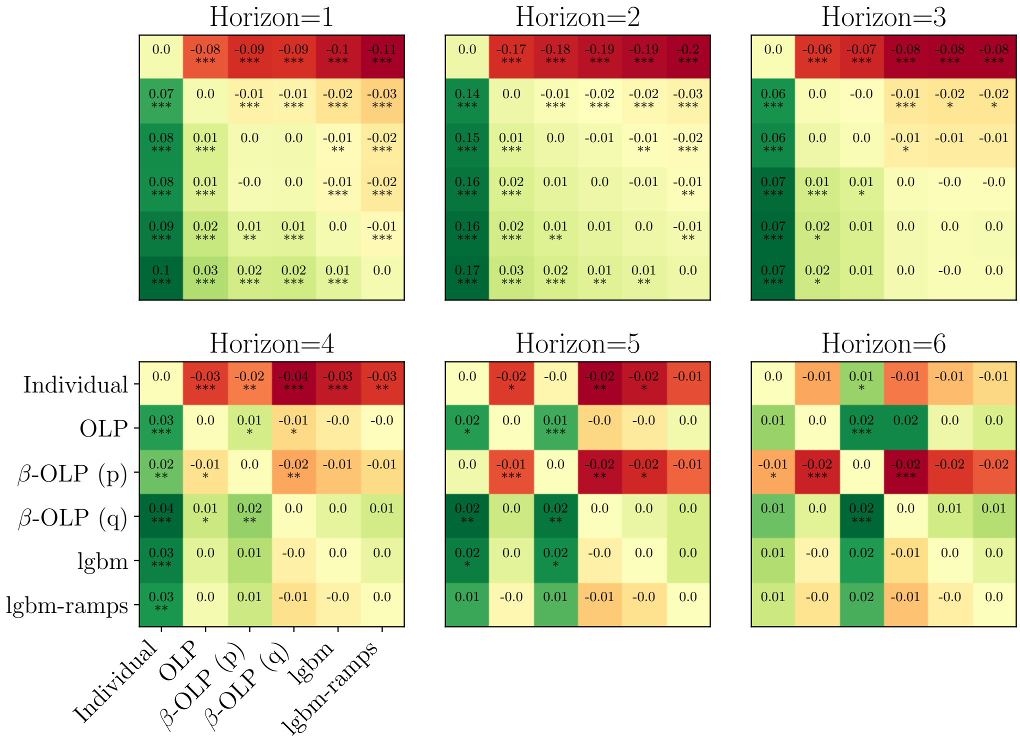


(b) Matrix of Pinball skill scores

Figure C.13: Matrices of skill scores for each pair of forecast combination methods for zone 6. Asterisks denote significance level of Diebold-Mariano test values: *=0.05; **=0.01; ***=0.001.

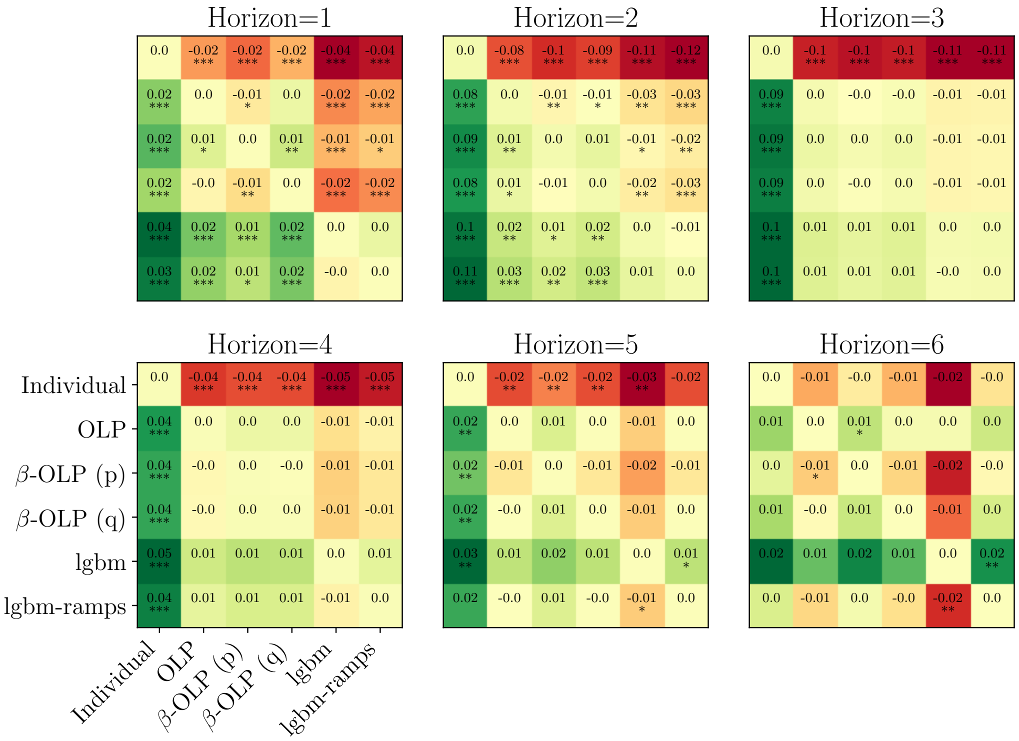


(a) Matrix of MAE skill scores. *(Figure continues onto next page)*

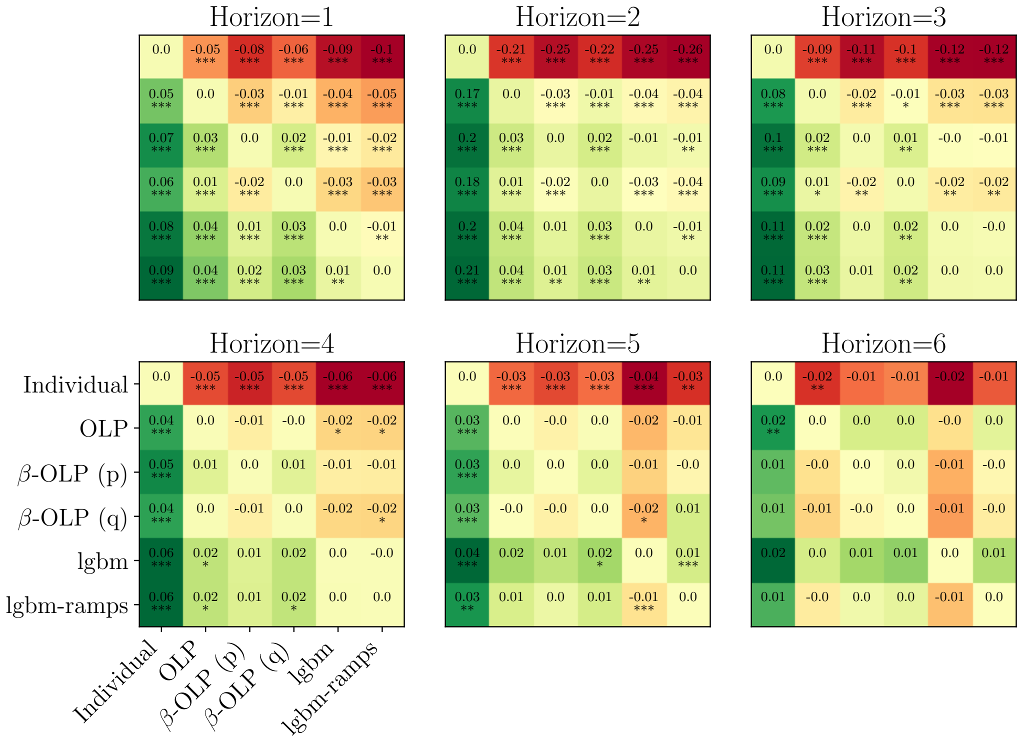


(b) Matrix of Pinball skill scores

Figure C.14: Matrices of skill scores for each pair of forecast combination methods for zone 7. Asterisks denote significance level of Diebold-Mariano test values: *=0.05; **=0.01; ***=0.001.

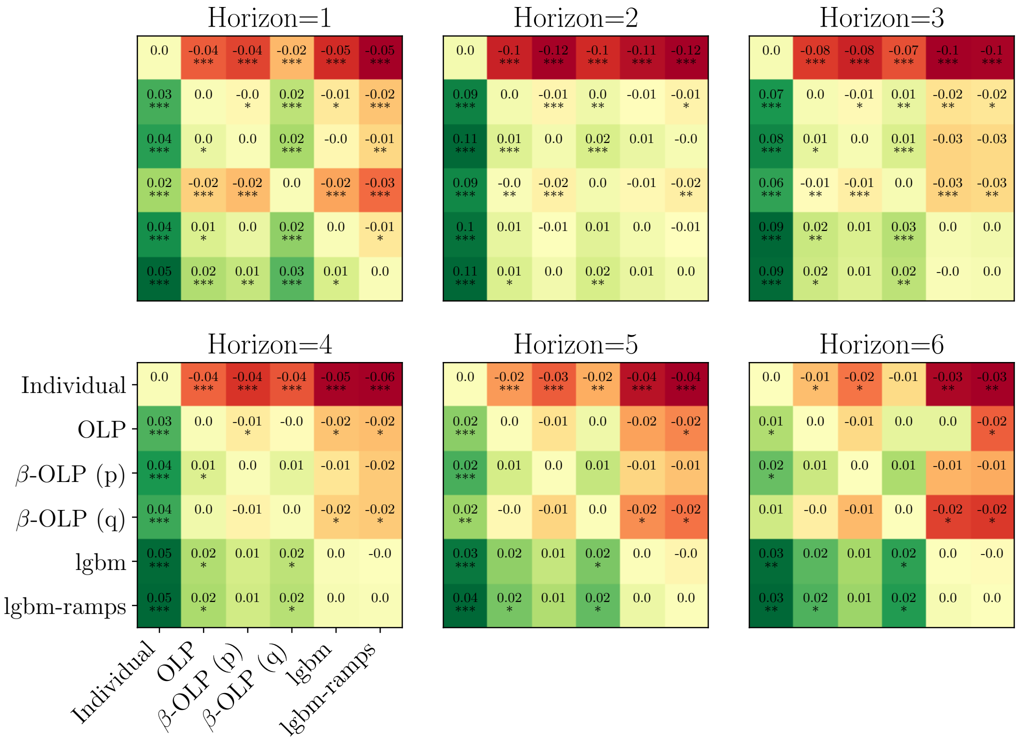


(a) Matrix of MAE skill scores. *(Figure continues onto next page)*

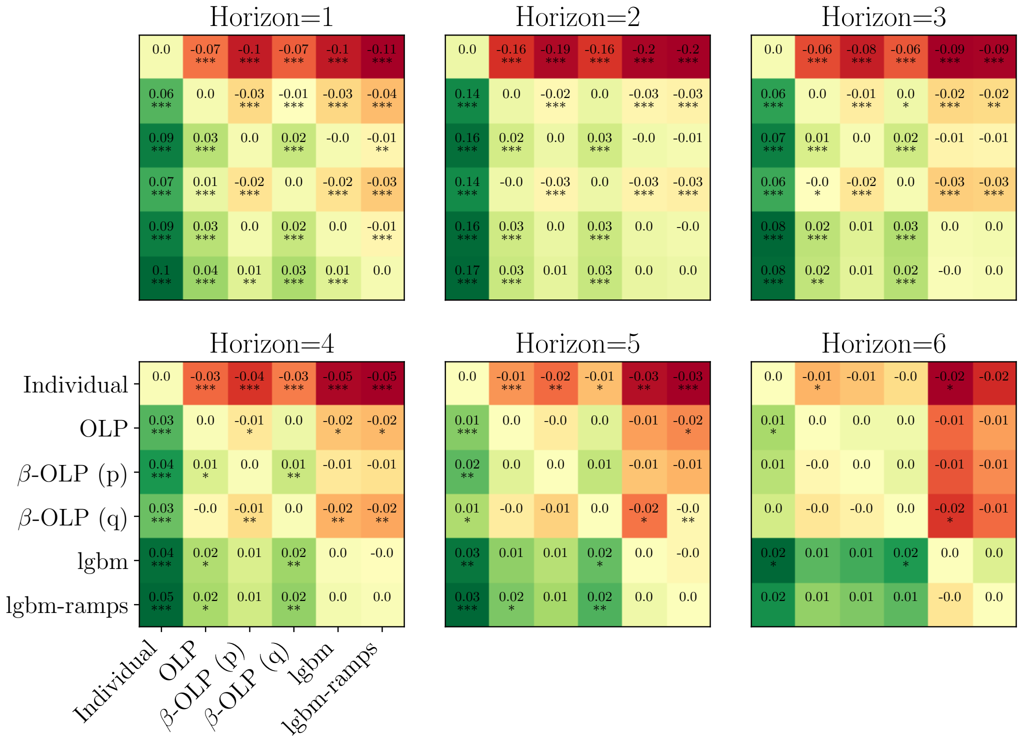


(b) Matrix of Pinball skill scores

Figure C.15: Matrices of skill scores for each pair of forecast combination methods for zone 8. Asterisks denote significance level of Diebold-Mariano test values: *=0.05; **=0.01; ***=0.001.

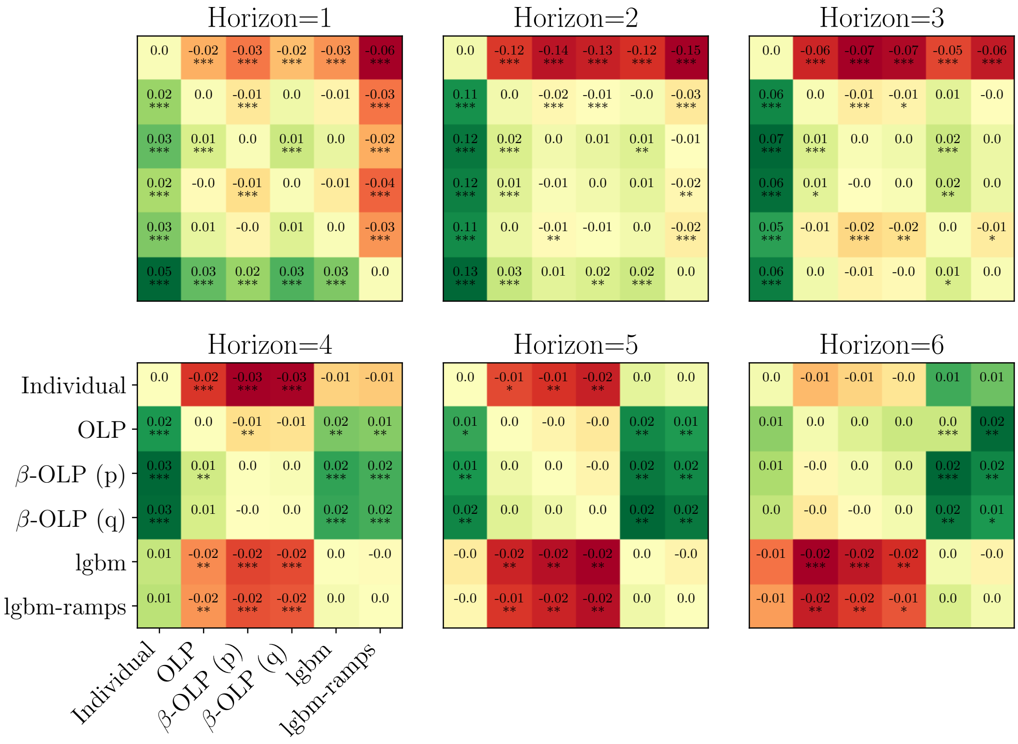


(a) Matrix of MAE skill scores. *(Figure continues onto next page)*

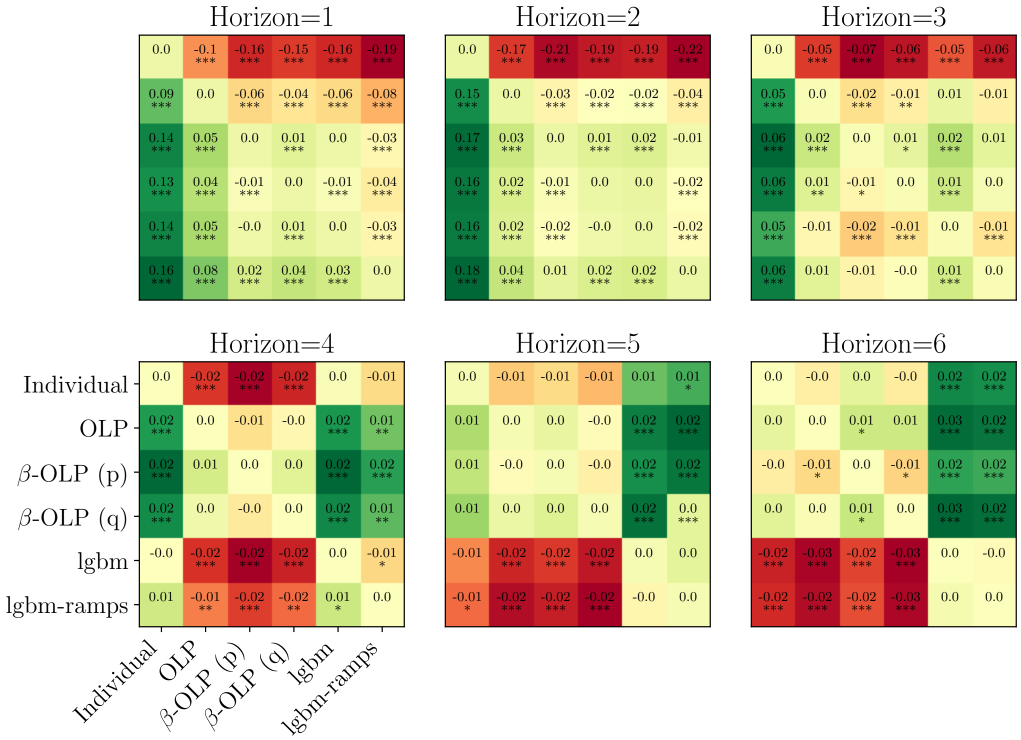


(b) Matrix of Pinball skill scores

Figure C.16: Matrices of skill scores for each pair of forecast combination methods for zone 9. Asterisks denote significance level of Diebold-Mariano test values: *=0.05; **=0.01; ***=0.001.



(a) Matrix of MAE skill scores. *(Figure continues onto next page)*



(b) Matrix of Pinball skill scores

Figure C.17: Matrices of skill scores for each pair of forecast combination methods for zone 10. Asterisks denote significance level of Diebold-Mariano test values: *=0.05; **=0.01; ***=0.001.

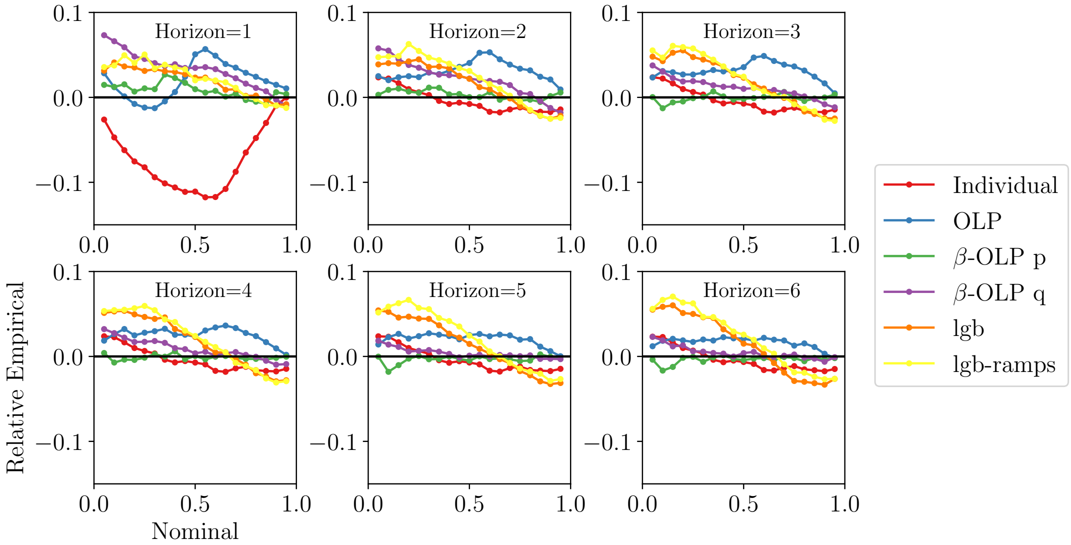


Figure C.18: Relative reliability for the combined model forecasts at zone 2. A relative empirical frequency of zero represents ideal reliability.

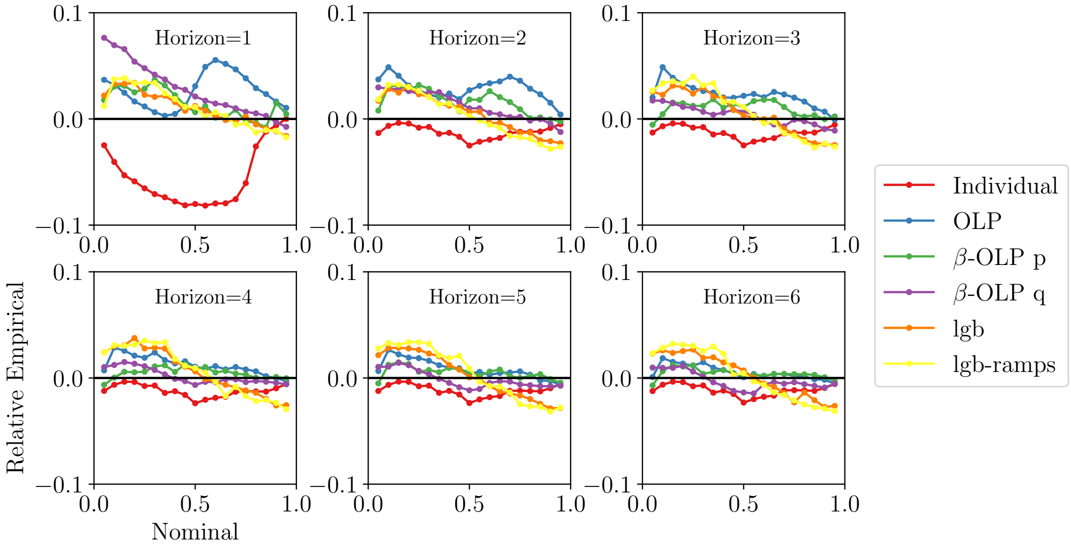


Figure C.19: Relative reliability for the combined model forecasts at zone 3.

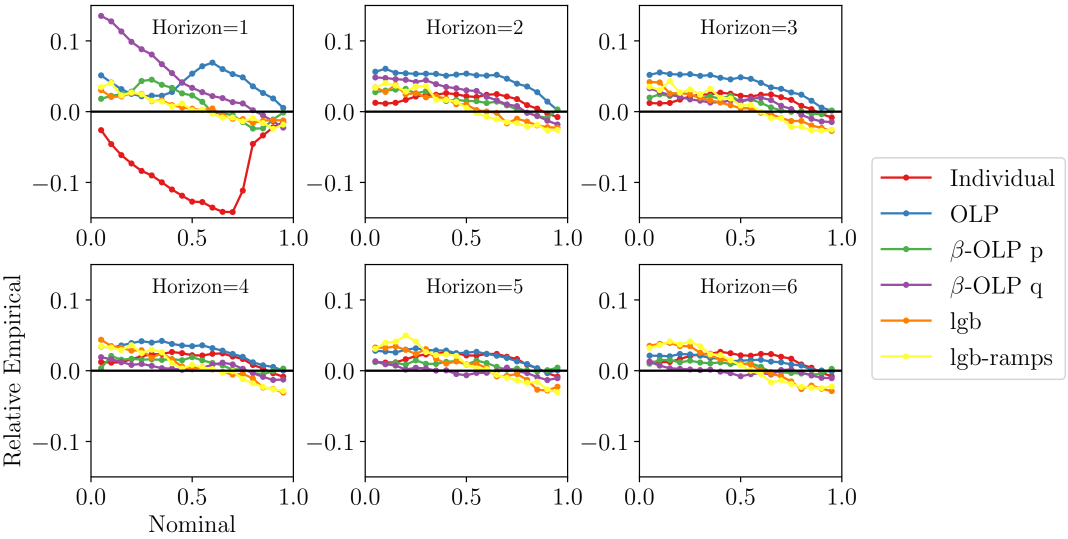


Figure C.20: Relative reliability for the combined model forecasts at zone 4.

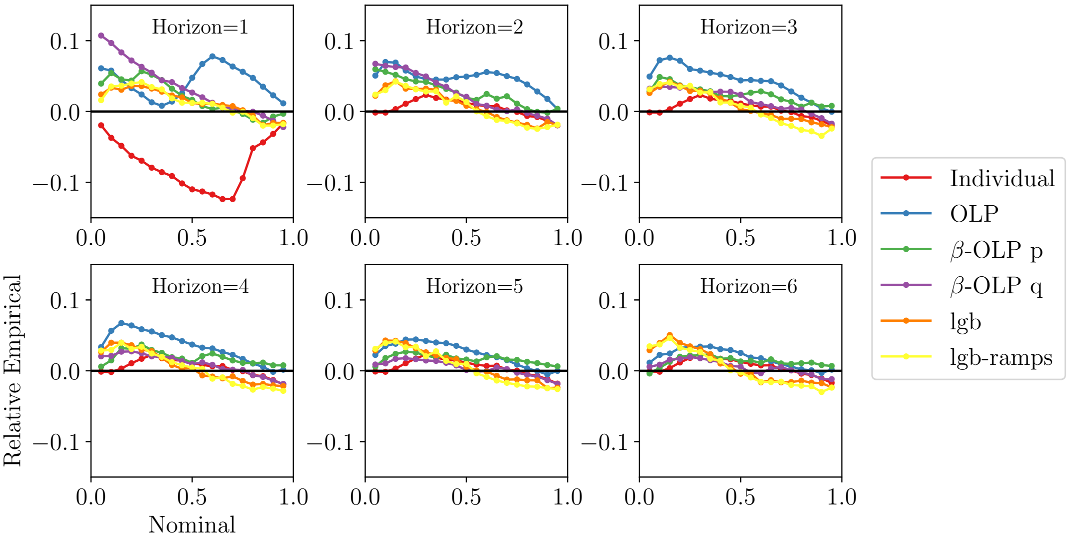


Figure C.21: Relative reliability for the combined model forecasts at zone 5.

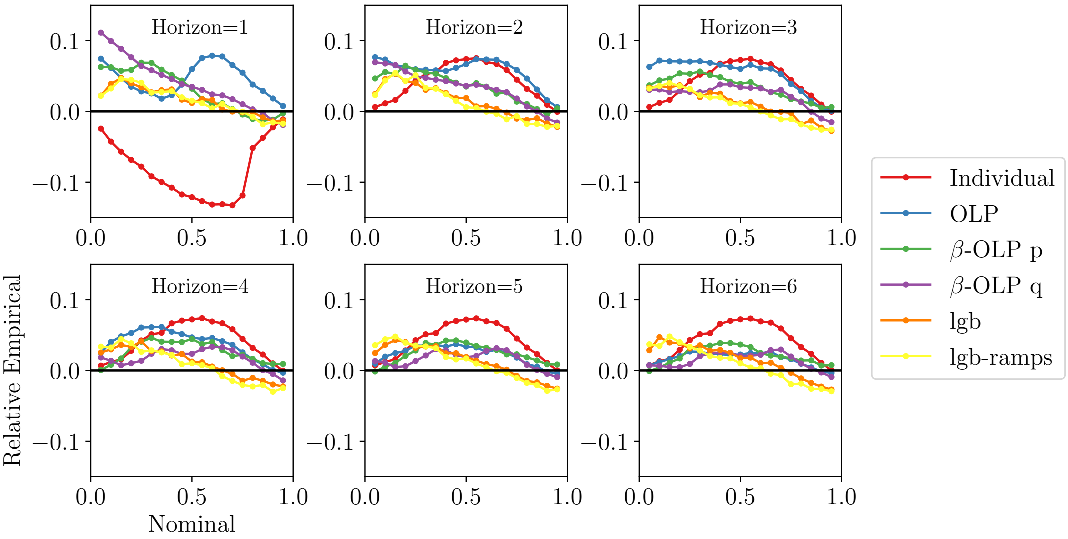


Figure C.22: Relative reliability for the combined model forecasts at zone 6.

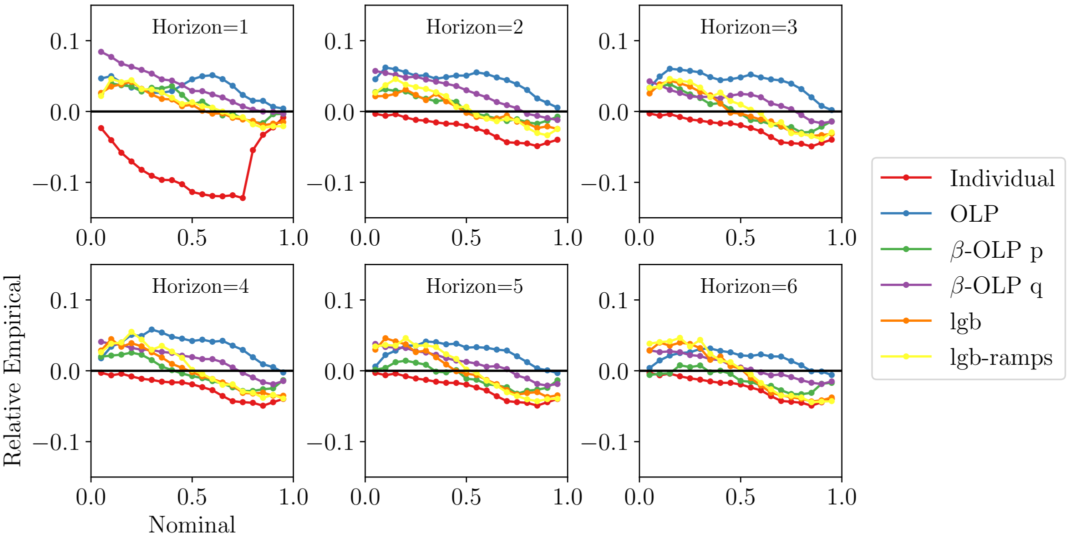


Figure C.23: Relative reliability for the combined model forecasts at zone 7.

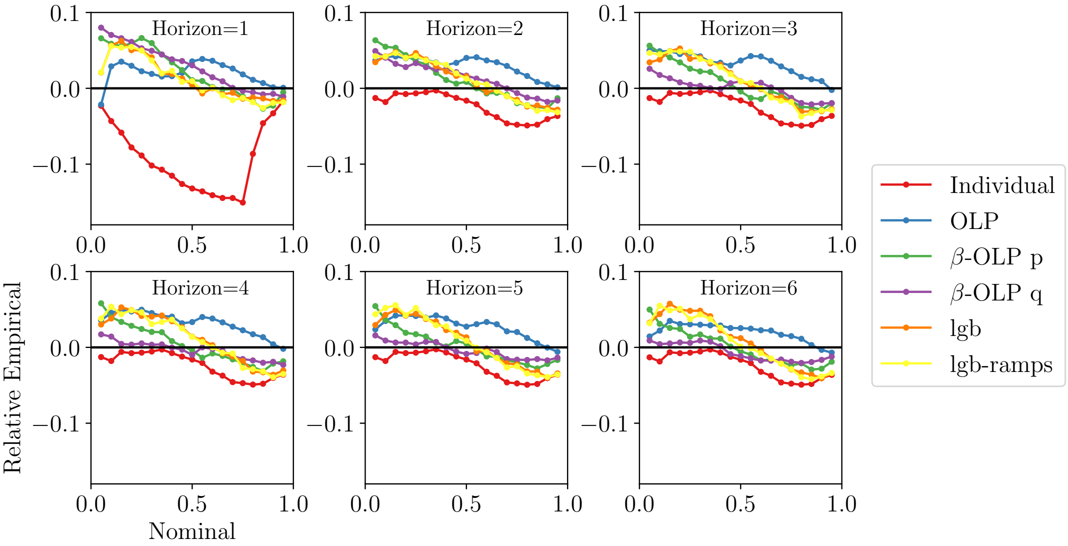


Figure C.24: Relative reliability for the combined model forecasts at zone 8.

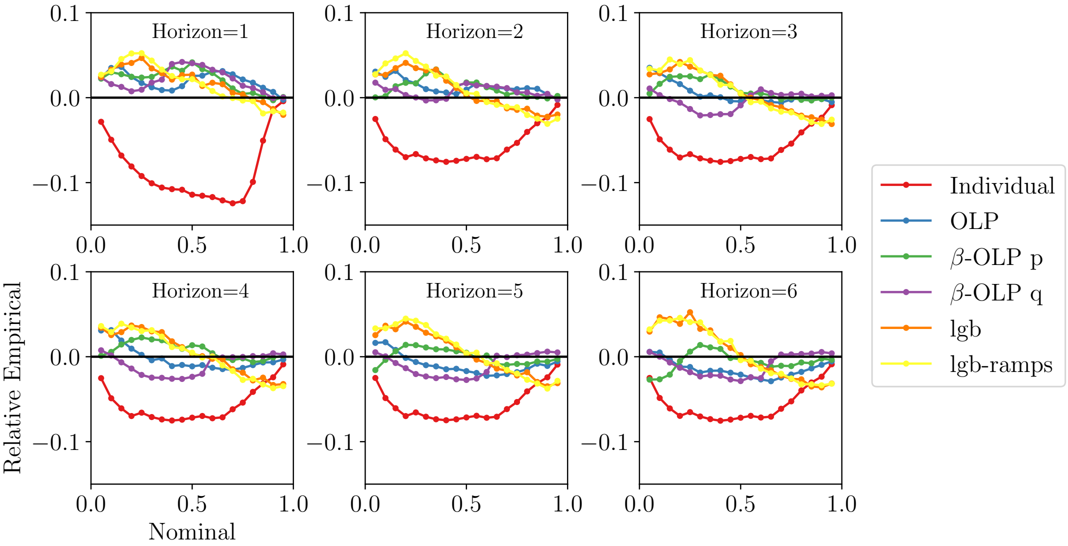


Figure C.25: Relative reliability for the combined model forecasts at zone 9.

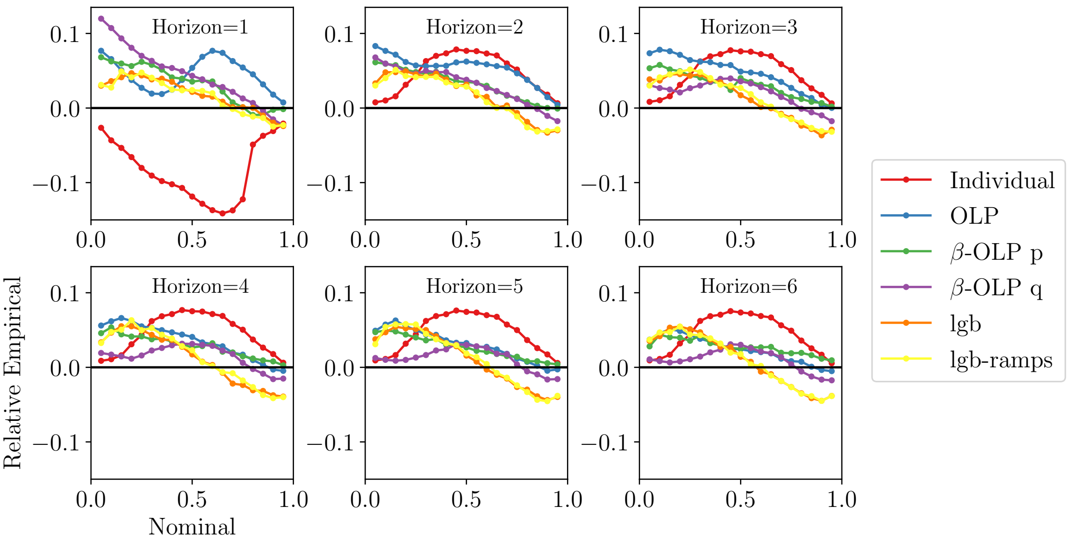


Figure C.26: Relative reliability for the combined model forecasts at zone 10.

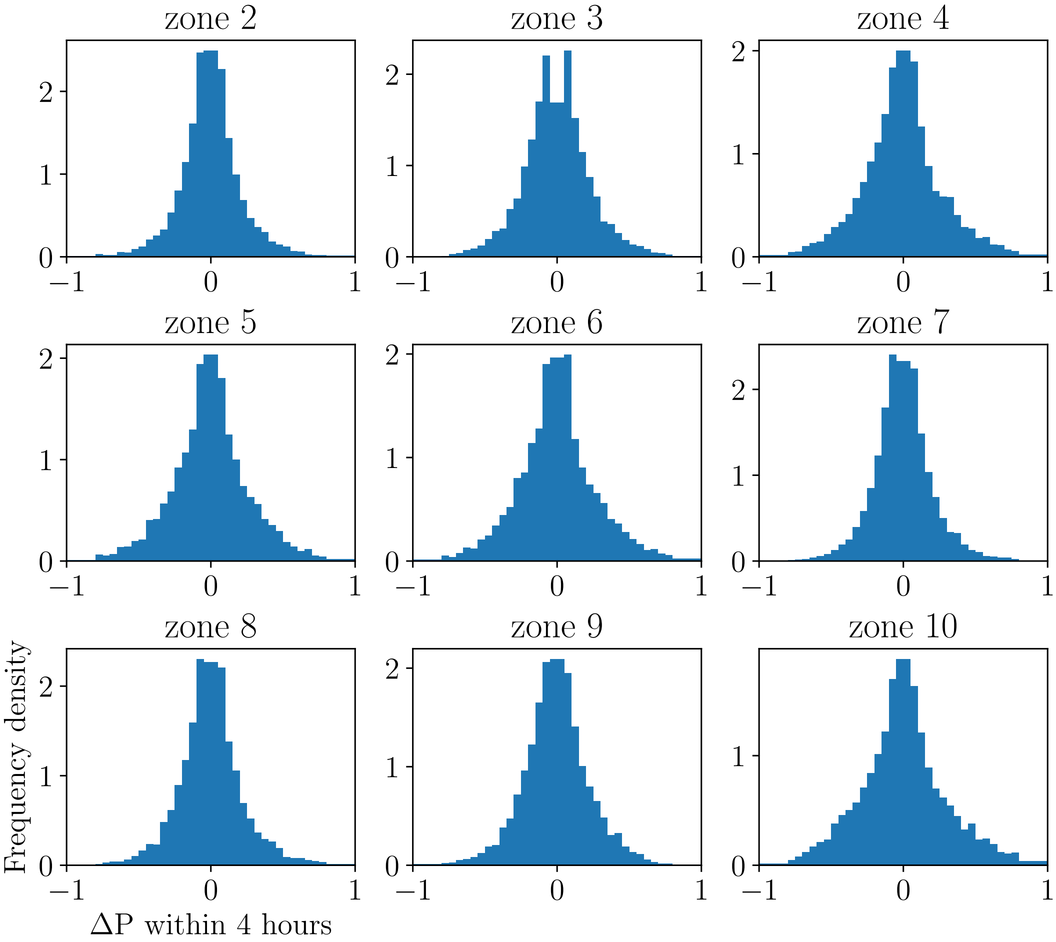


Figure C.27: Distribution of ramp magnitudes over a 4 hour window for zones 2-10.

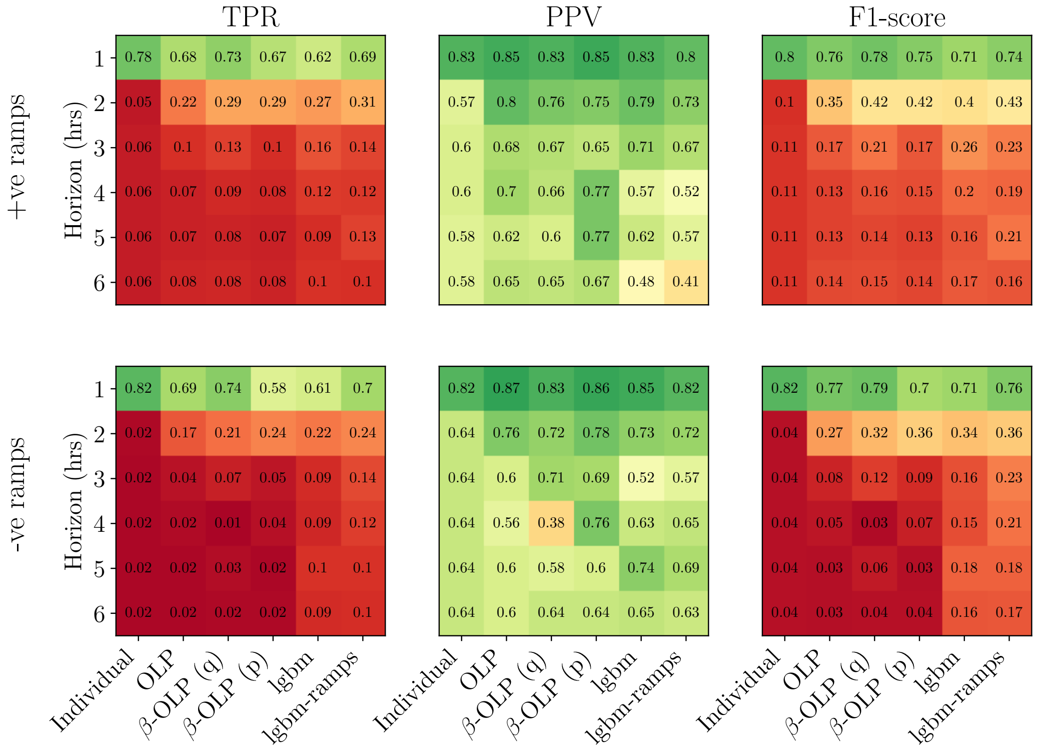


Figure C.28: Confusion matrix scores for all combination models and forecast horizons for zone 2.

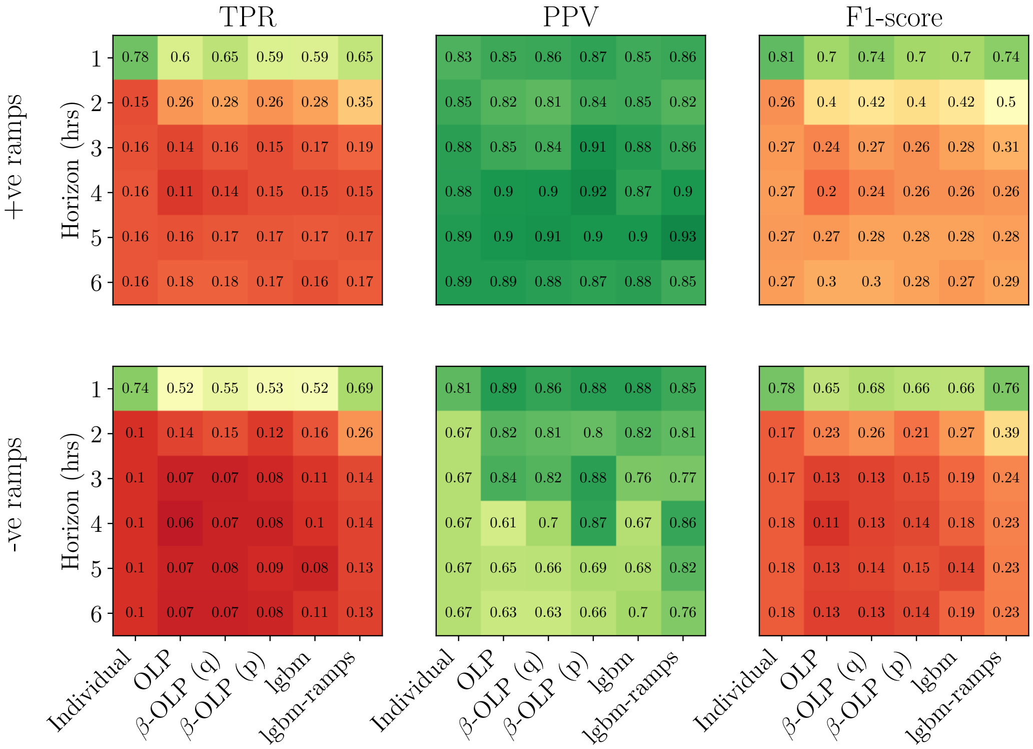


Figure C.29: Confusion matrix scores for all combination models and forecast horizons for zone 3.

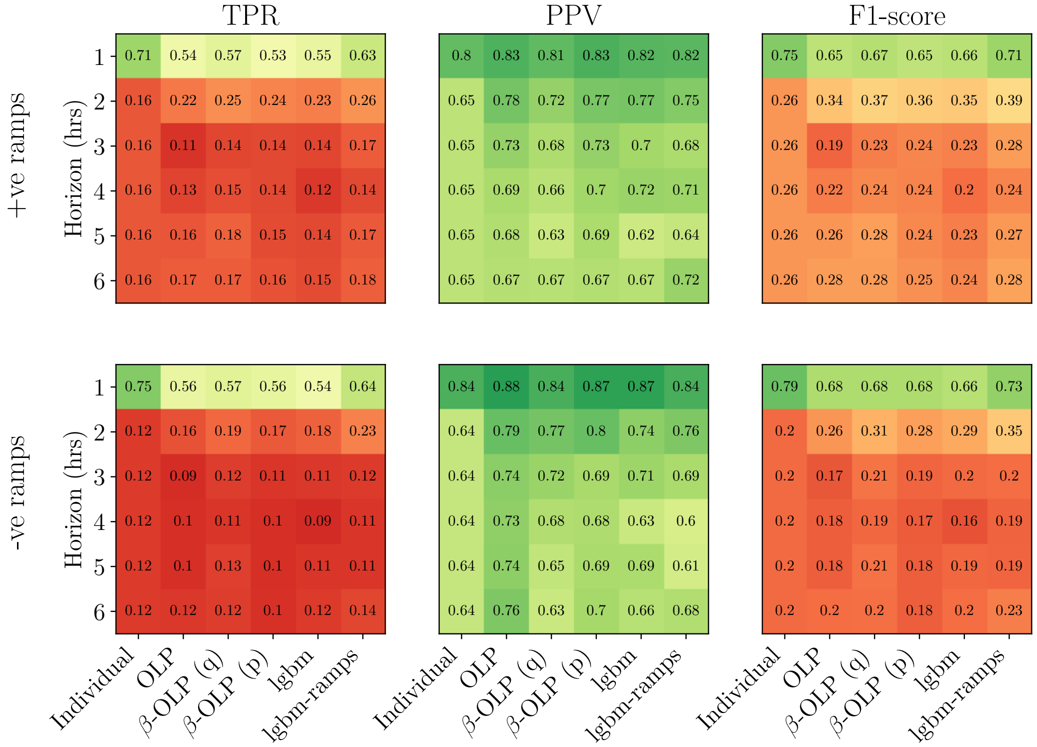


Figure C.30: Confusion matrix scores for all combination models and forecast horizons for zone 4.

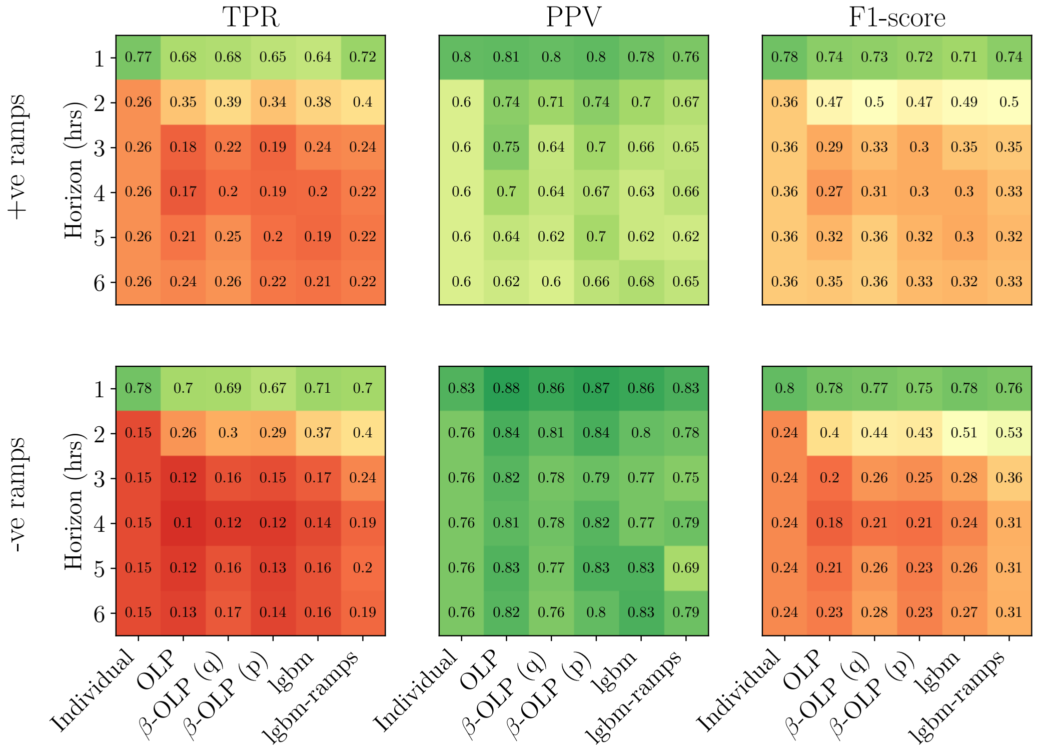


Figure C.31: Confusion matrix scores for all combination models and forecast horizons for zone 5.

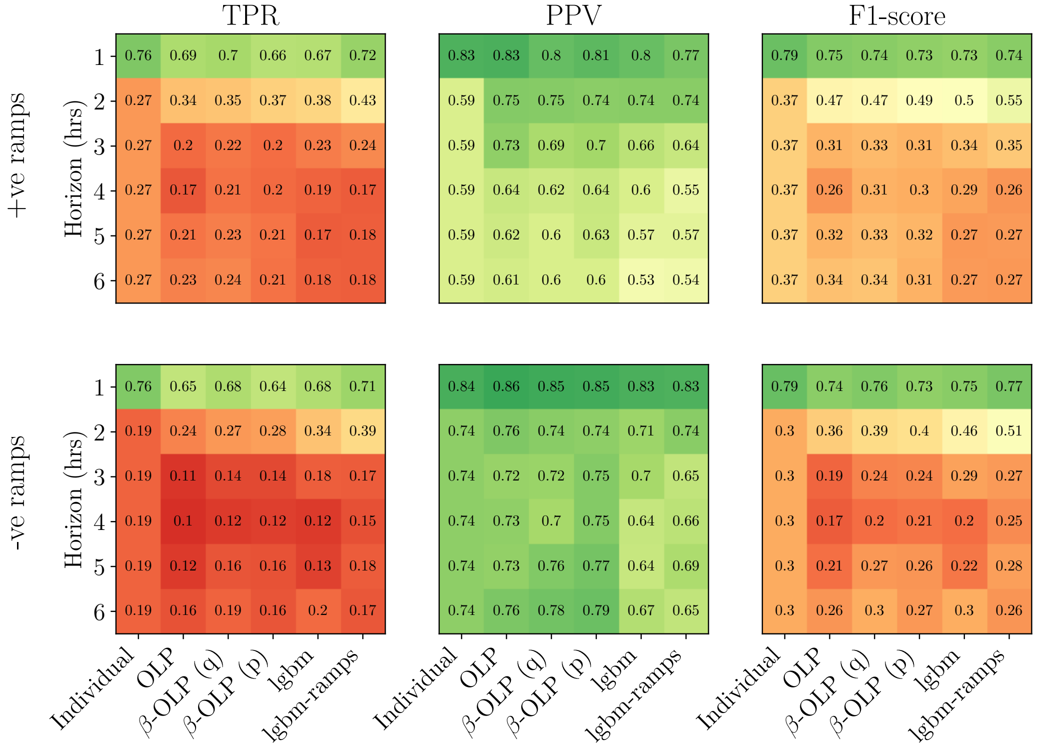


Figure C.32: Confusion matrix scores for all combination models and forecast horizons for zone 6.

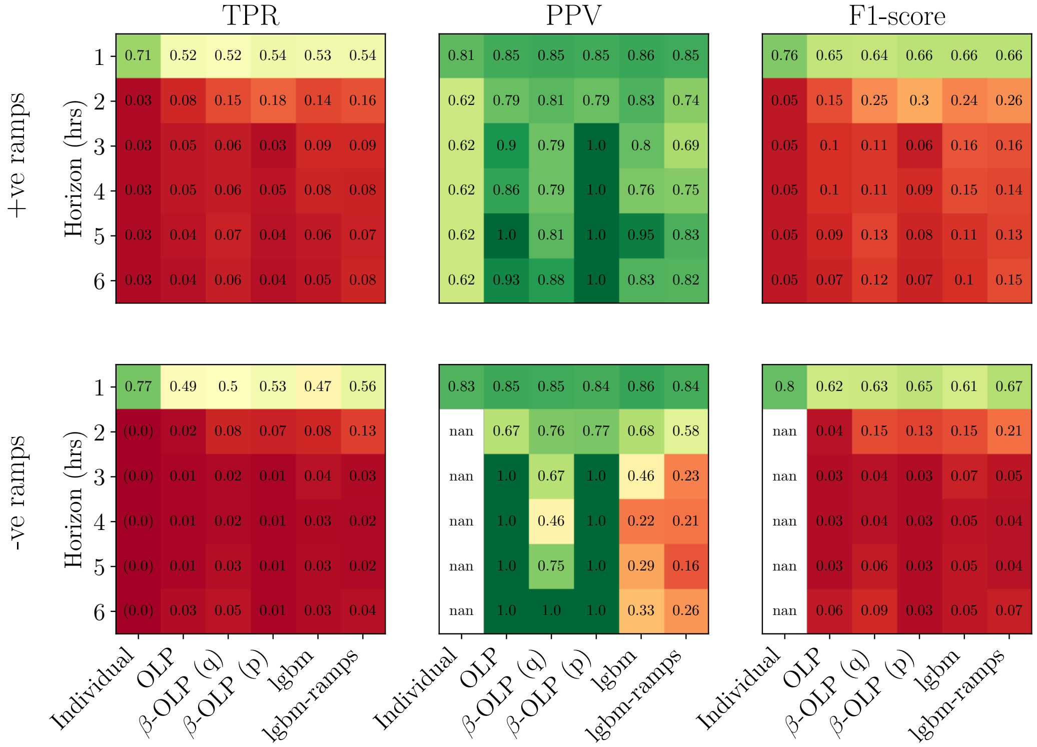


Figure C.33: Confusion matrix score for all combination models and forecast horizons for zone 7.

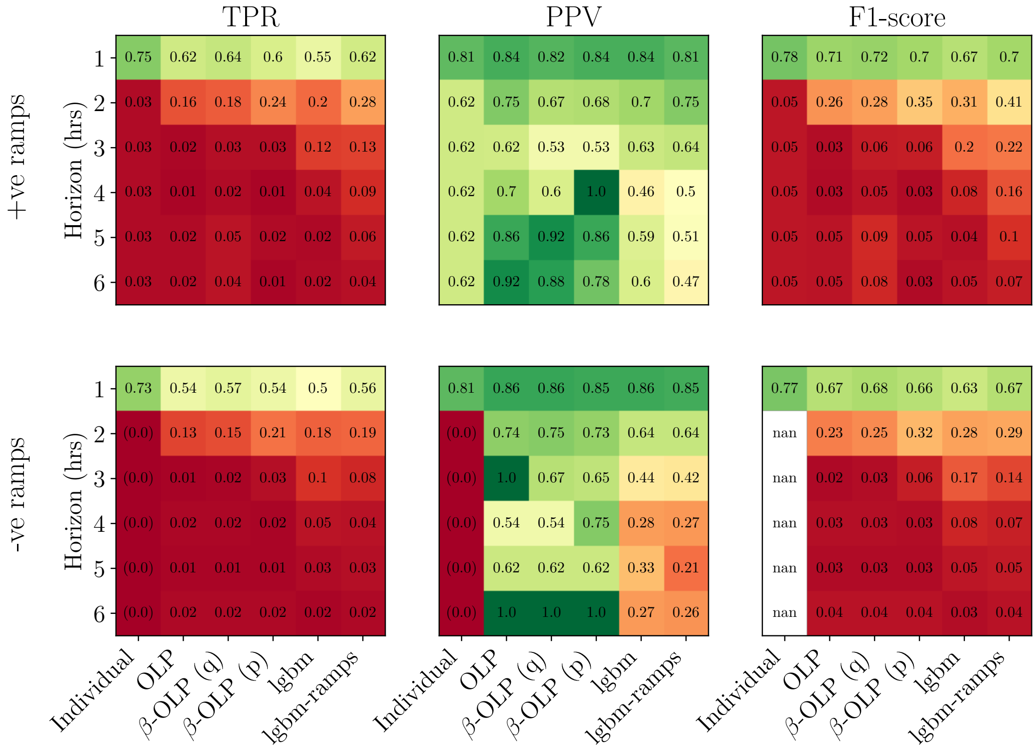


Figure C.34: Confusion matrix scores for all combination models and forecast horizons for zone 8.

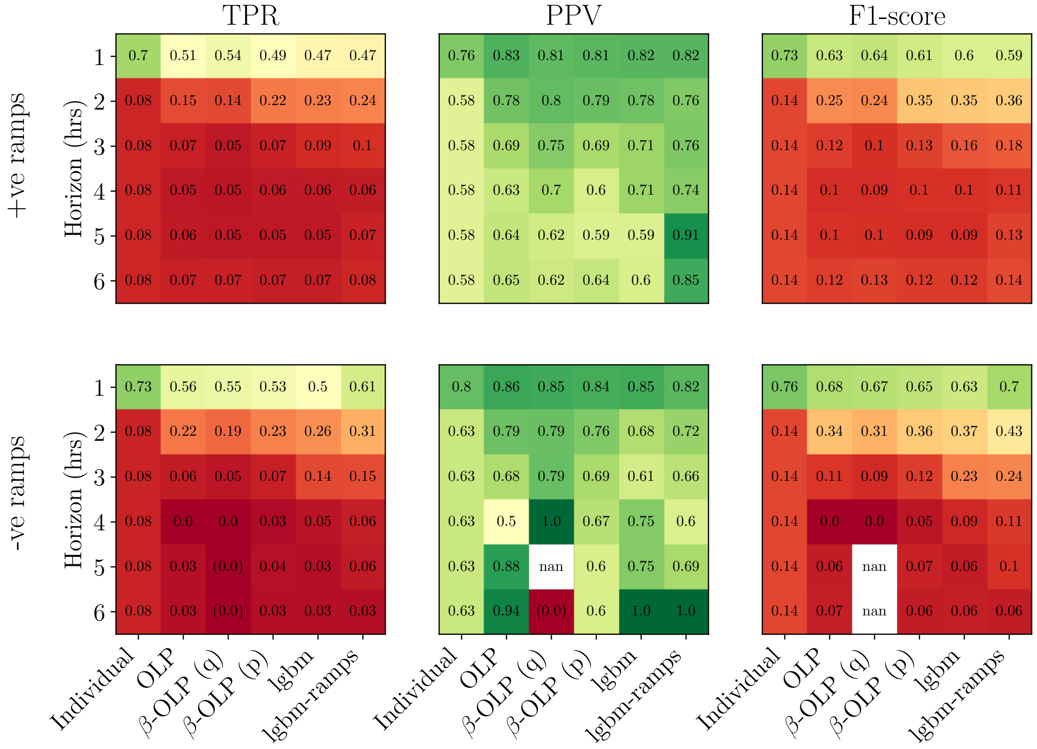


Figure C.35: Confusion matrix scores for all combination models and forecast horizons for zone 9.

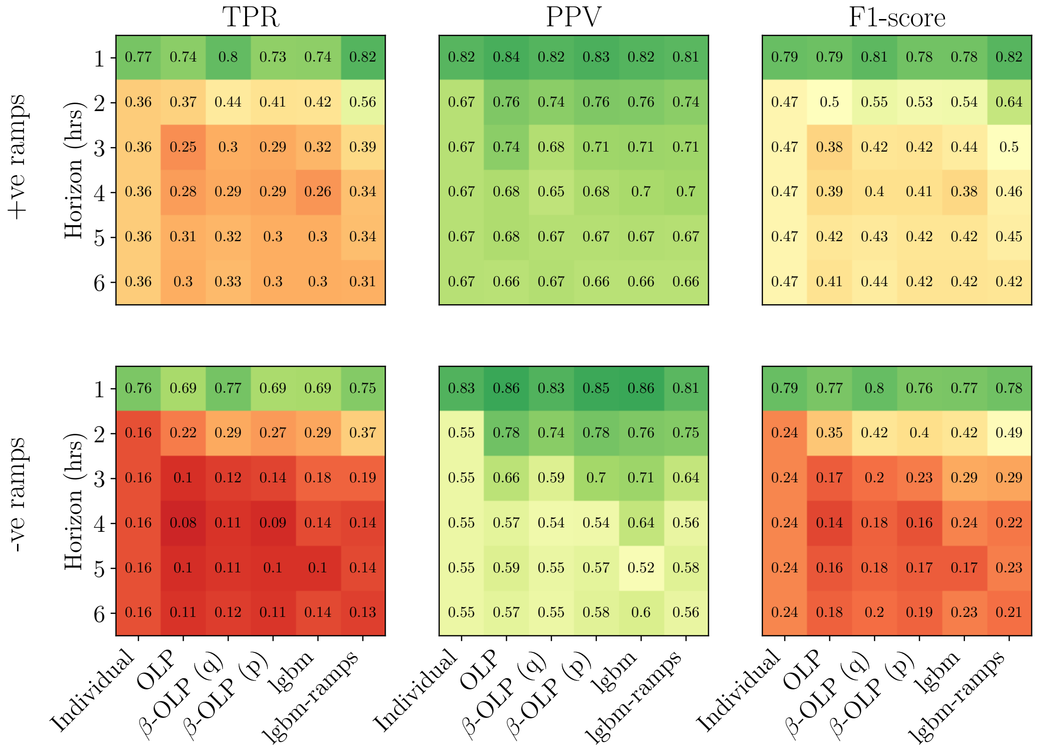


Figure C.36: Confusion matrix scores for all combination models and forecast horizons for zone 10.

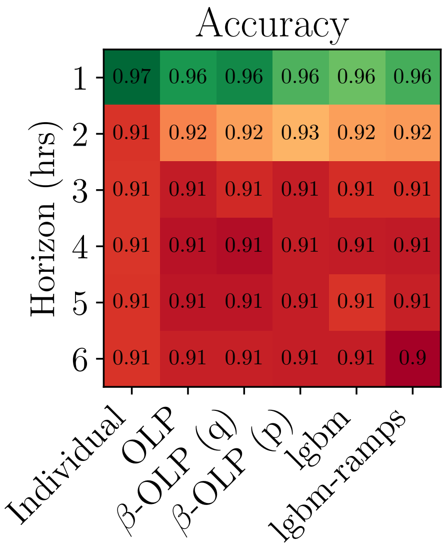


Figure C.37: Accuracy scores for all combination models and forecast horizons for zone 2

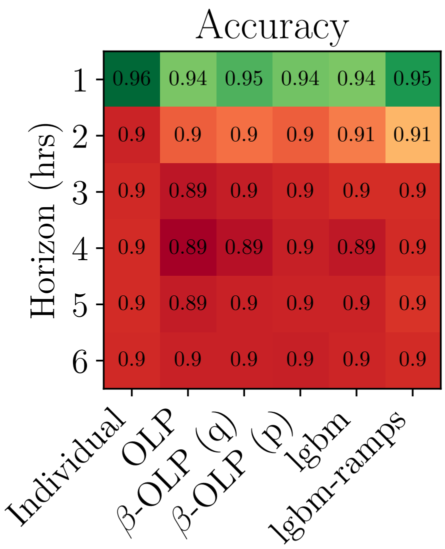


Figure C.38: Accuracy scores for all combination models and forecast horizons for zone 3

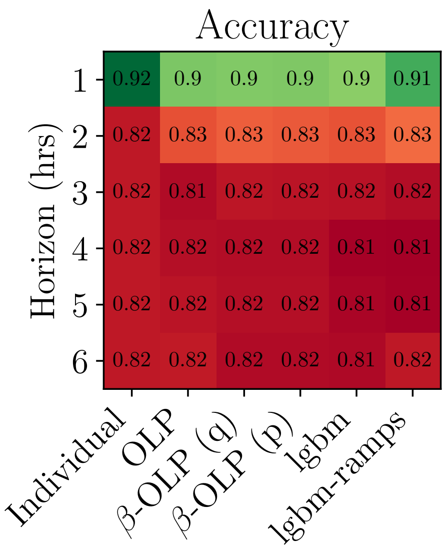


Figure C.39: Accuracy scores for all combination models and forecast horizons for zone 4

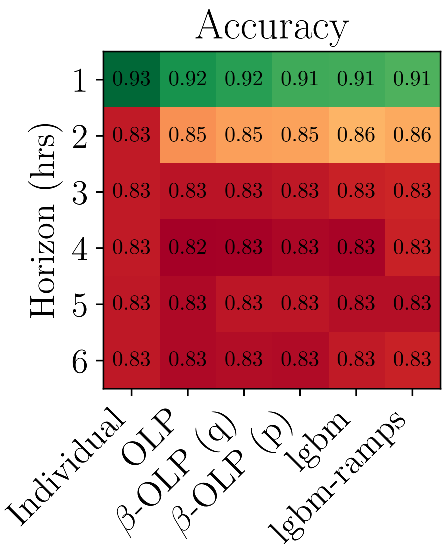


Figure C.40: Accuracy scores for all combination models and forecast horizons for zone 5

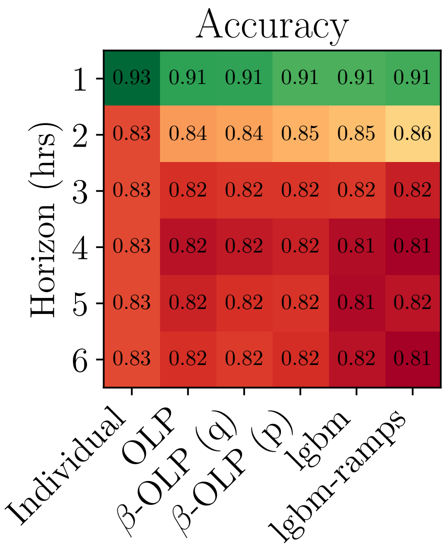


Figure C.41: Accuracy scores for all combination models and forecast horizons for zone 6

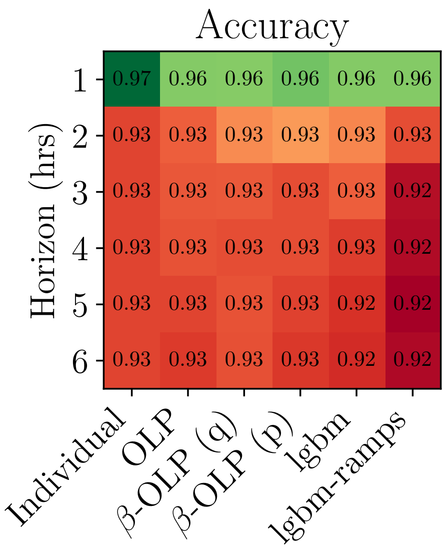


Figure C.42: Accuracy scores for all combination models and forecast horizons for zone 7

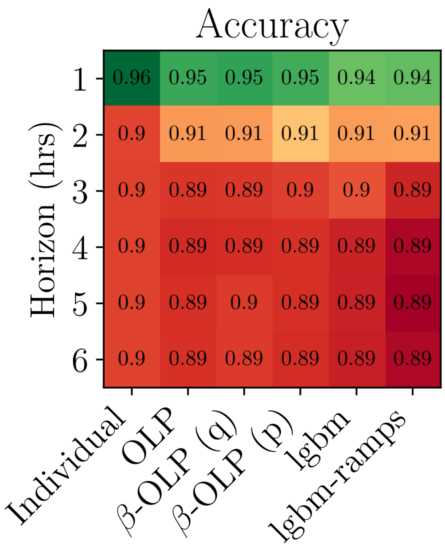


Figure C.43: Accuracy scores for all combination models and forecast horizons for zone 8

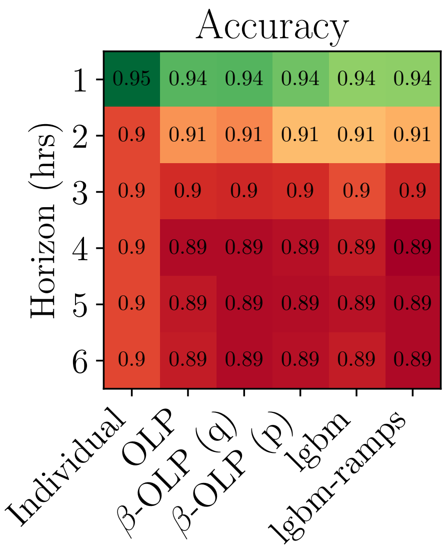


Figure C.44: Accuracy scores for all combination models and forecast horizons for zone 9

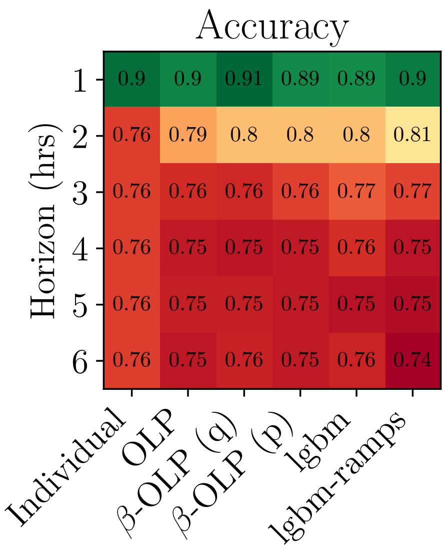


Figure C.45: Accuracy scores for all combination models and forecast horizons for zone 10

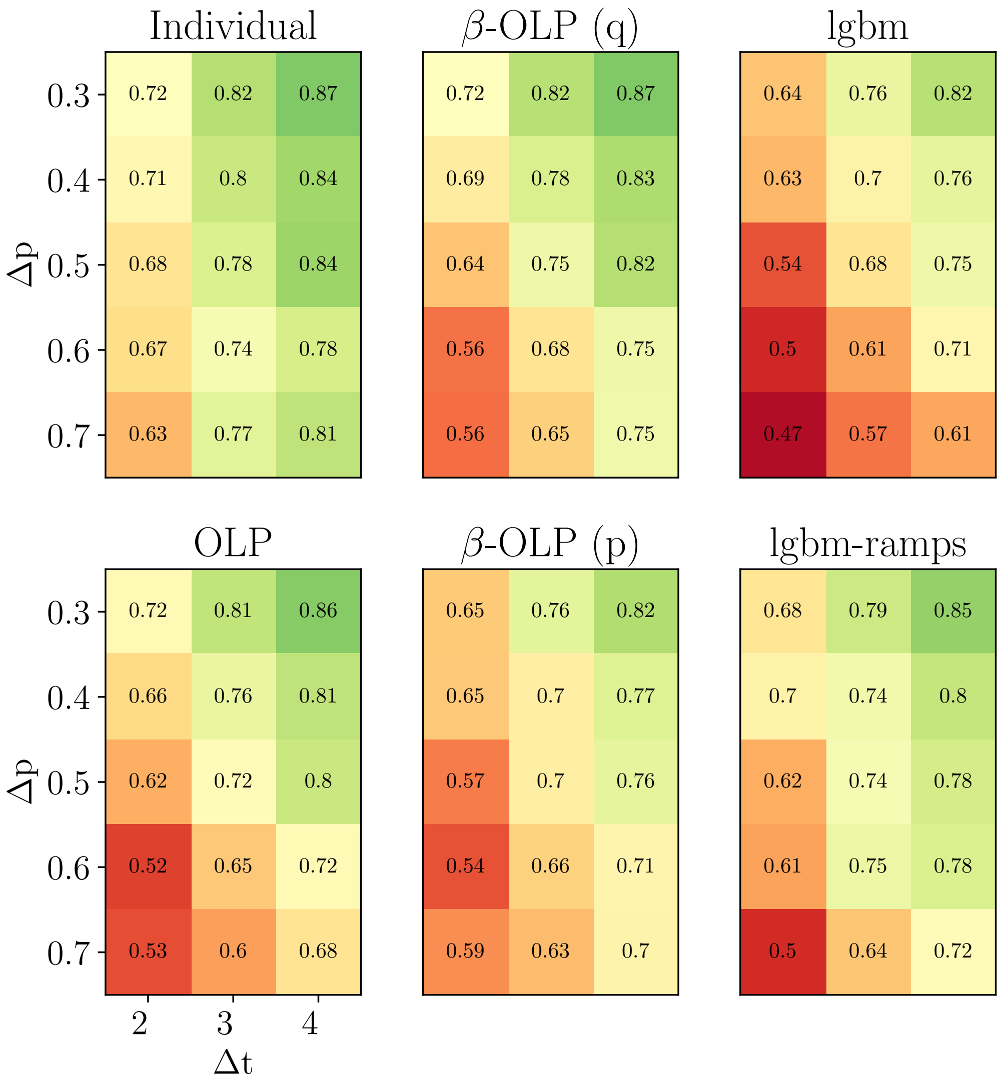


Figure C.46: F1 scores for varying ramp definitions, for all combination models at zone 2. Values are for a 1 hour ahead forecast.

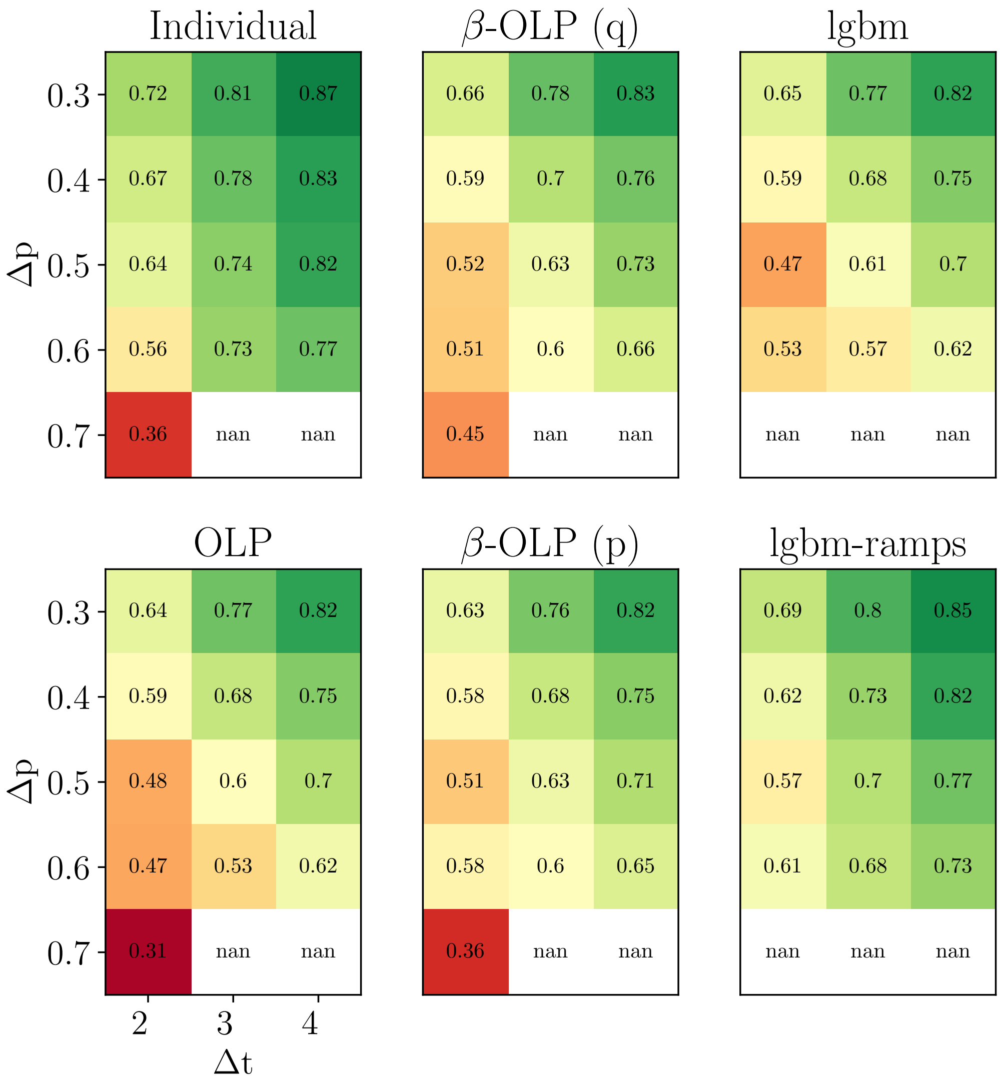


Figure C.47: F1 scores for varying ramp definitions, for all combination models at zone 3. Values are for a 1 hour ahead forecast.

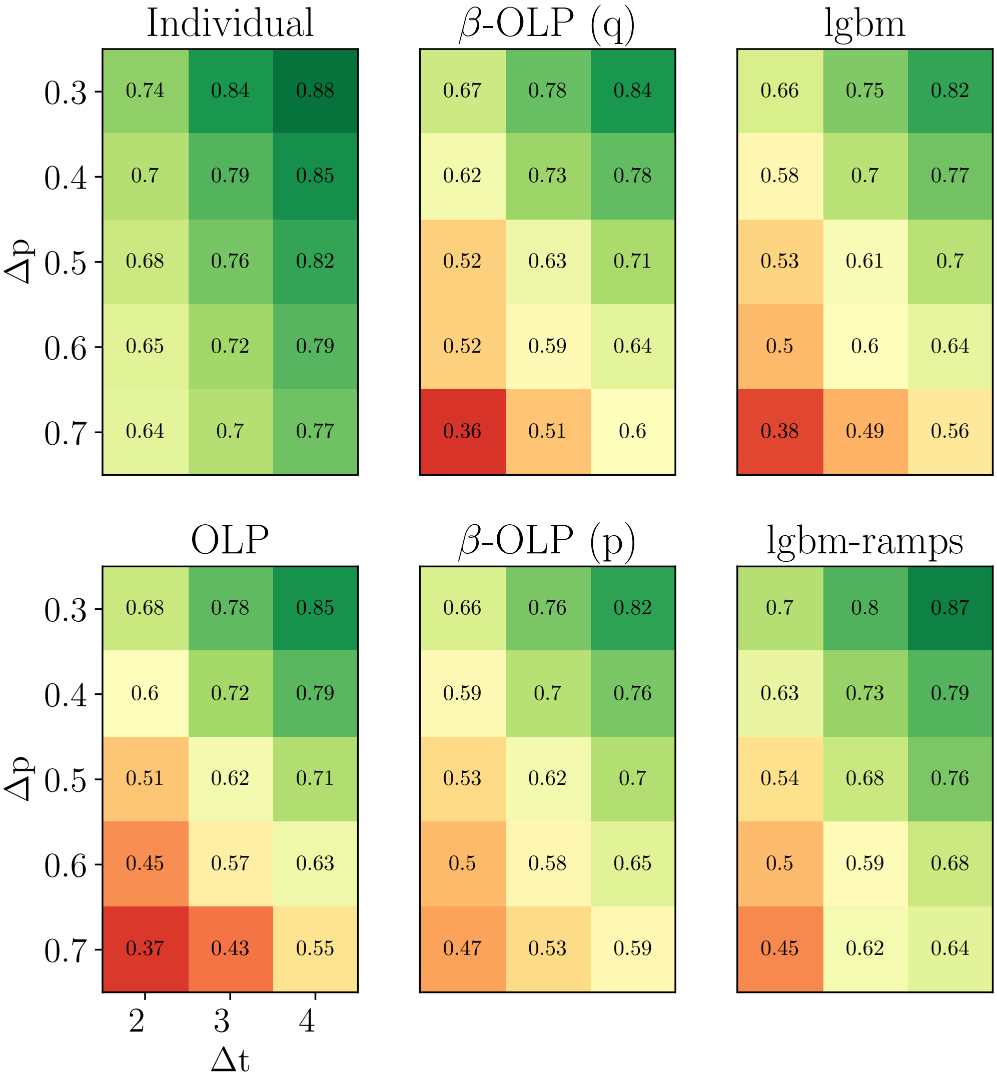


Figure C.48: F1 scores for varying ramp definitions, for all combination models at zone 4. Values are for a 1 hour ahead forecast.

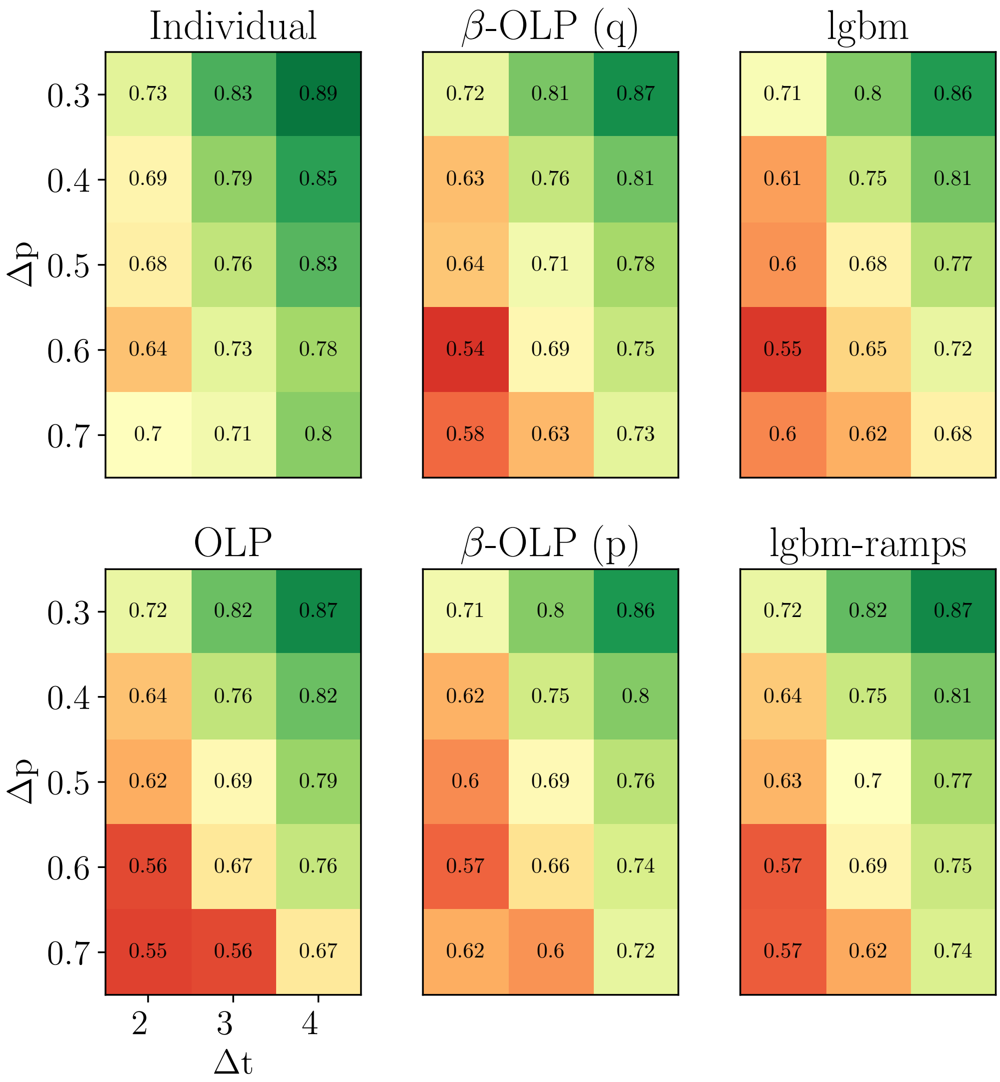


Figure C.49: F1 scores for varying ramp definitions, for all combination models at zone 5. Values are for a 1 hour ahead forecast.

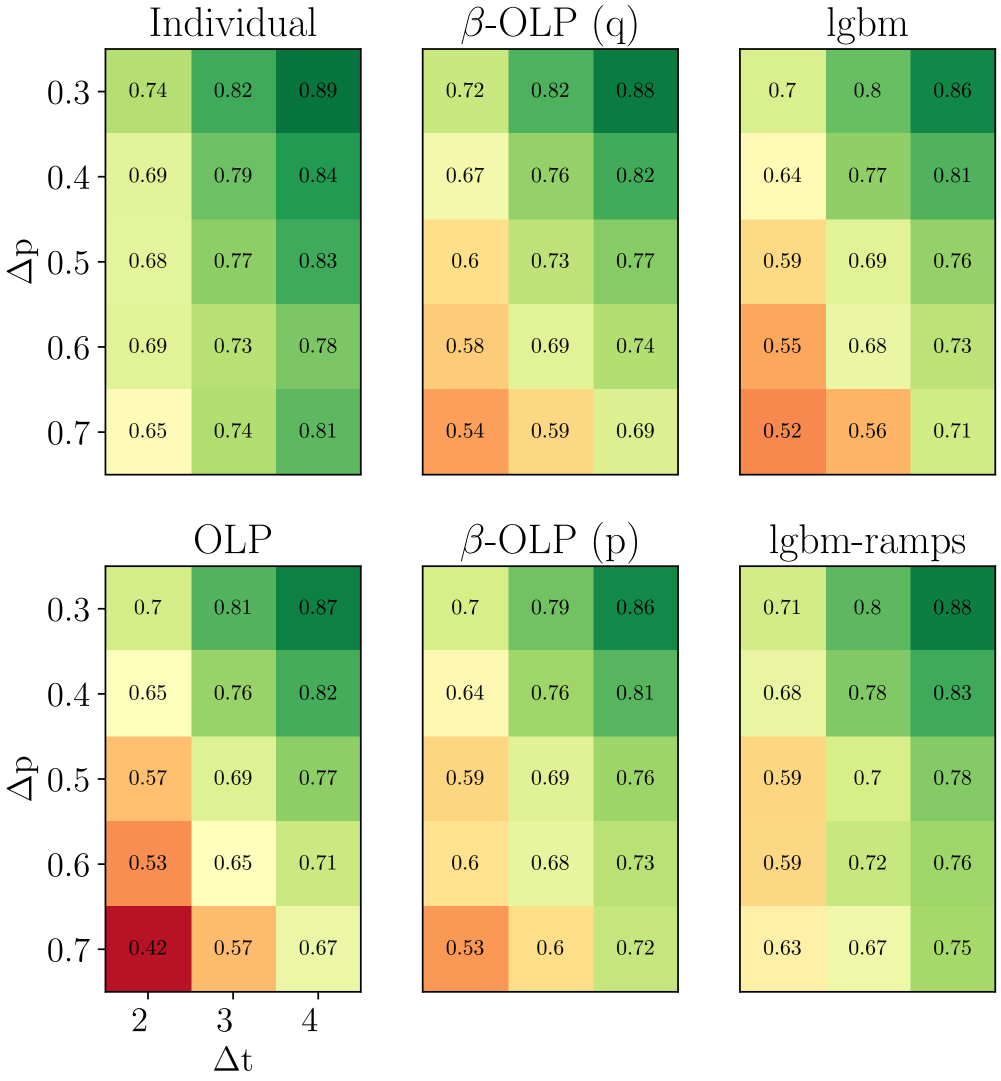


Figure C.50: F1 scores for varying ramp definitions, for all combination models at zone 6. Values are for a 1 hour ahead forecast.

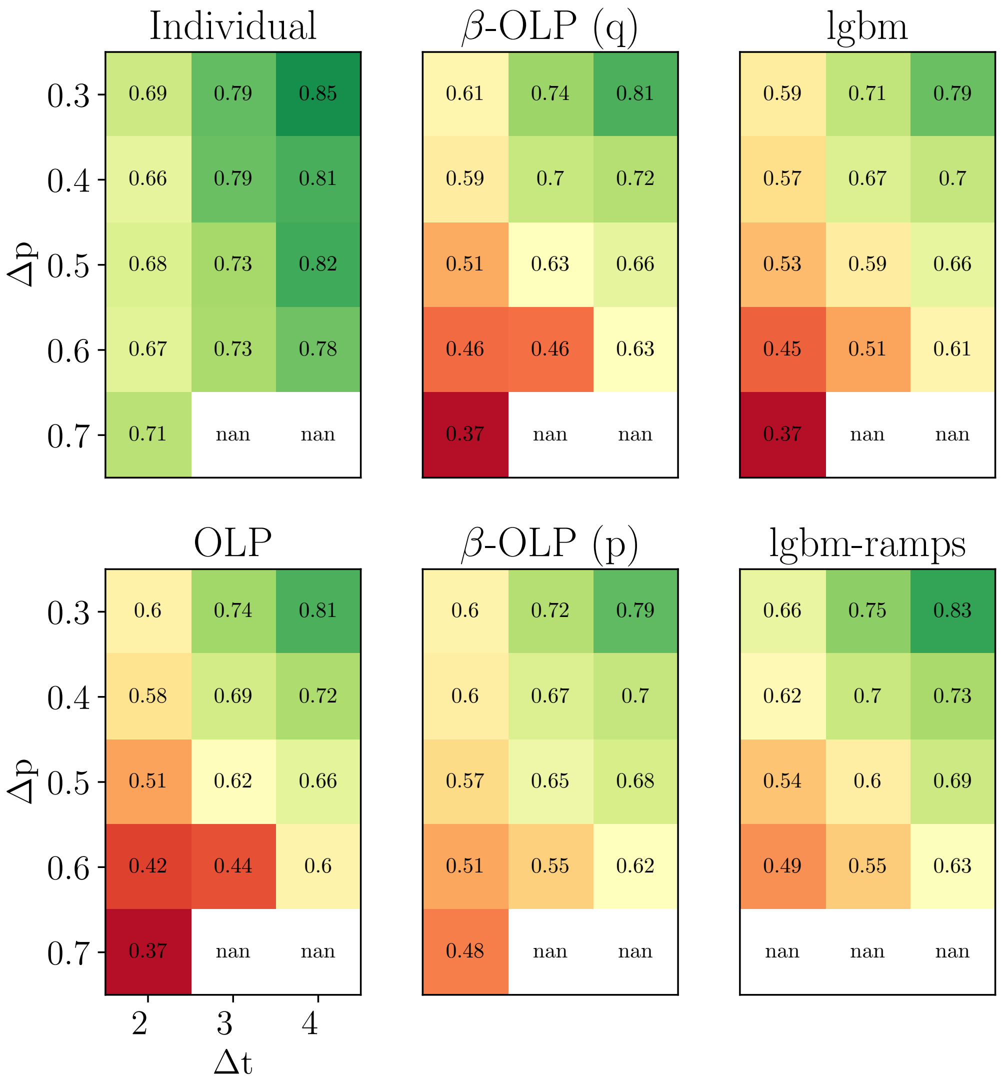


Figure C.51: F1 scores for varying ramp definitions, for all combination models at zone 7. Values are for a 1 hour ahead forecast.

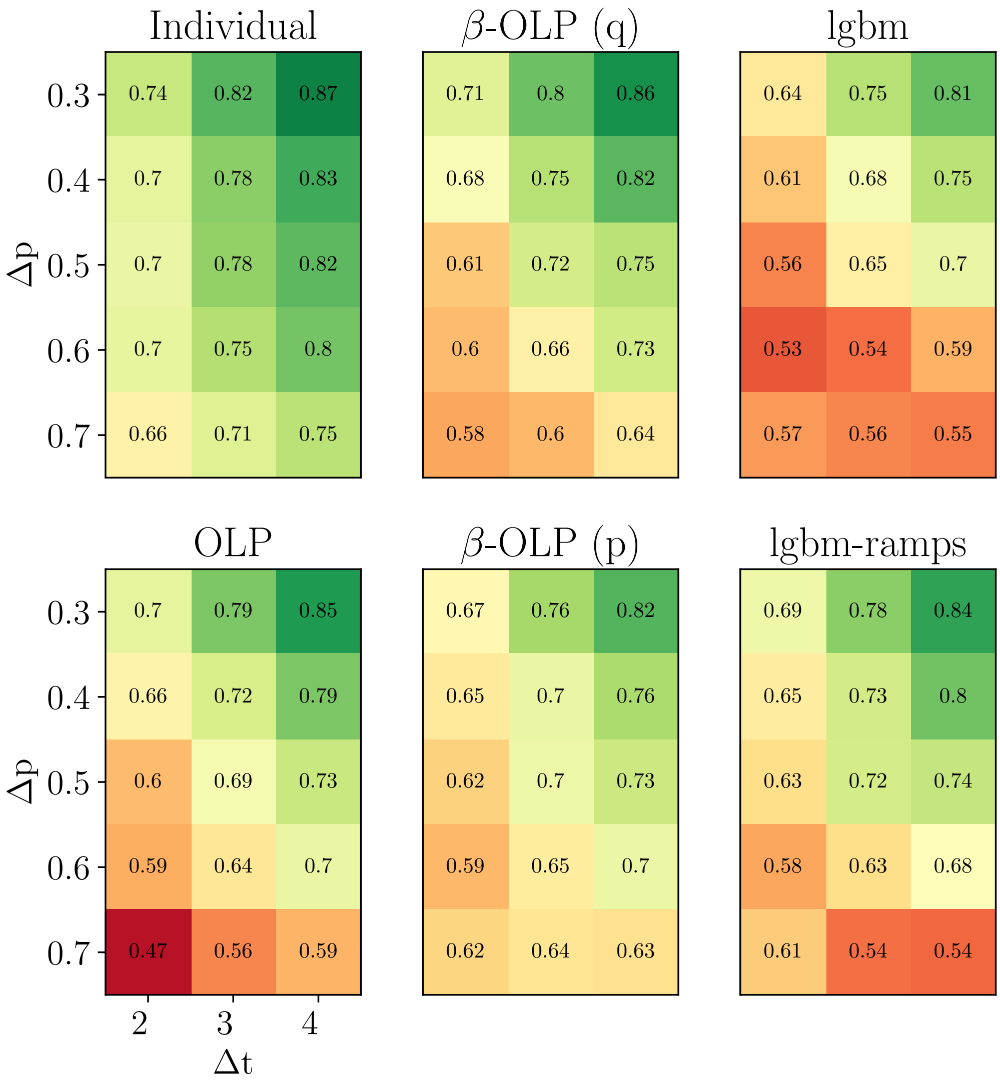


Figure C.52: F1 scores for varying ramp definitions, for all combination models at zone 8. Values are for a 1 hour ahead forecast.

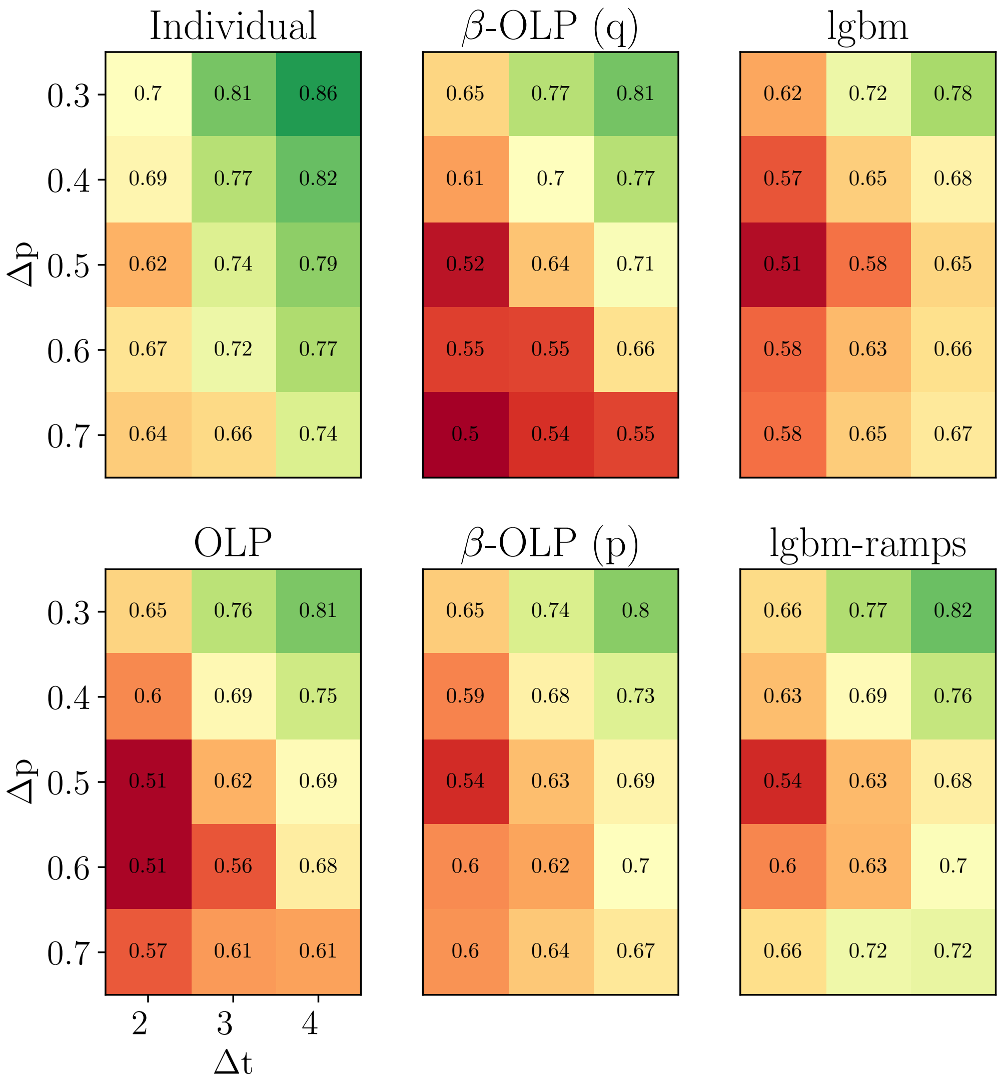


Figure C.53: F1 scores for varying ramp definitions, for all combination models at zone 9. Values are for a 1 hour ahead forecast.

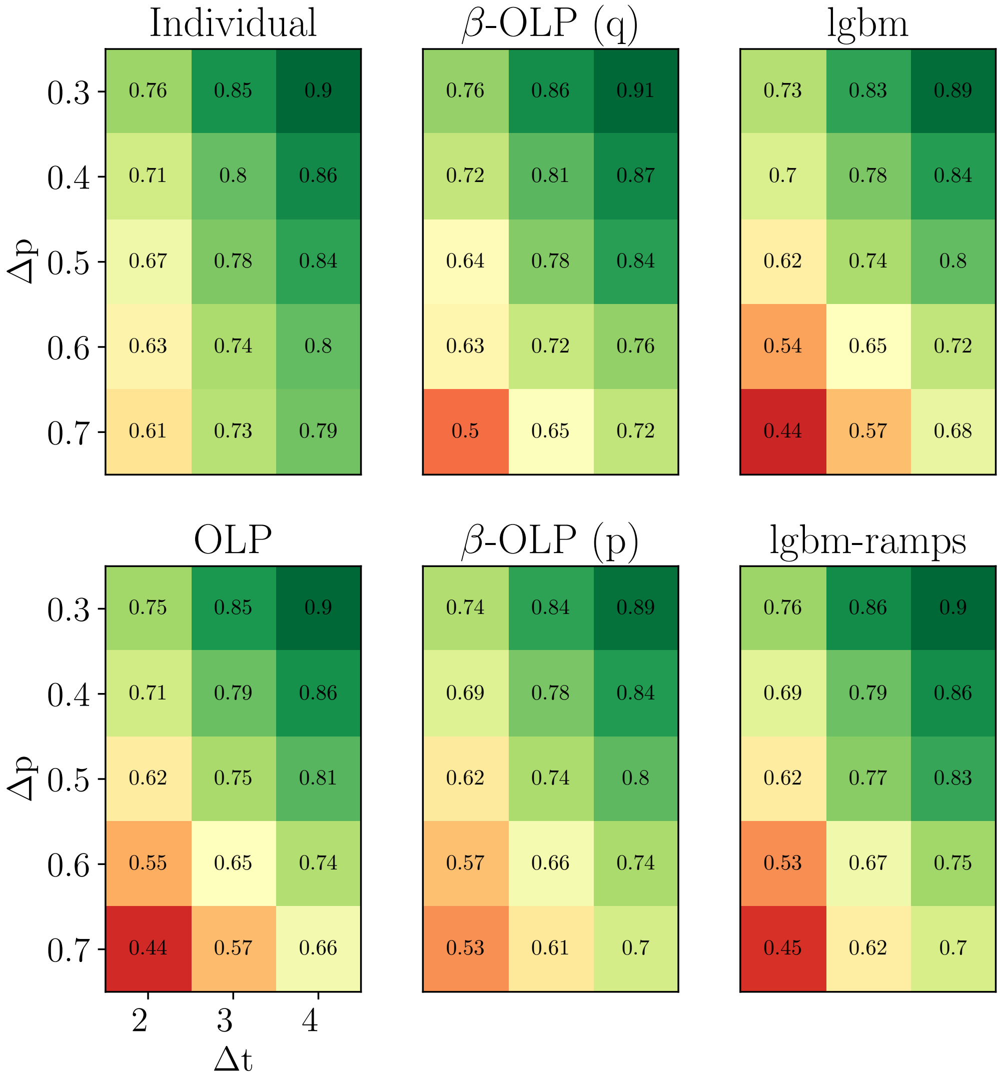


Figure C.54: F1 scores for varying ramp definitions, for all combination models at zone 10. Values are for a 1 hour ahead forecast.

Appendix D

S2S Appendix

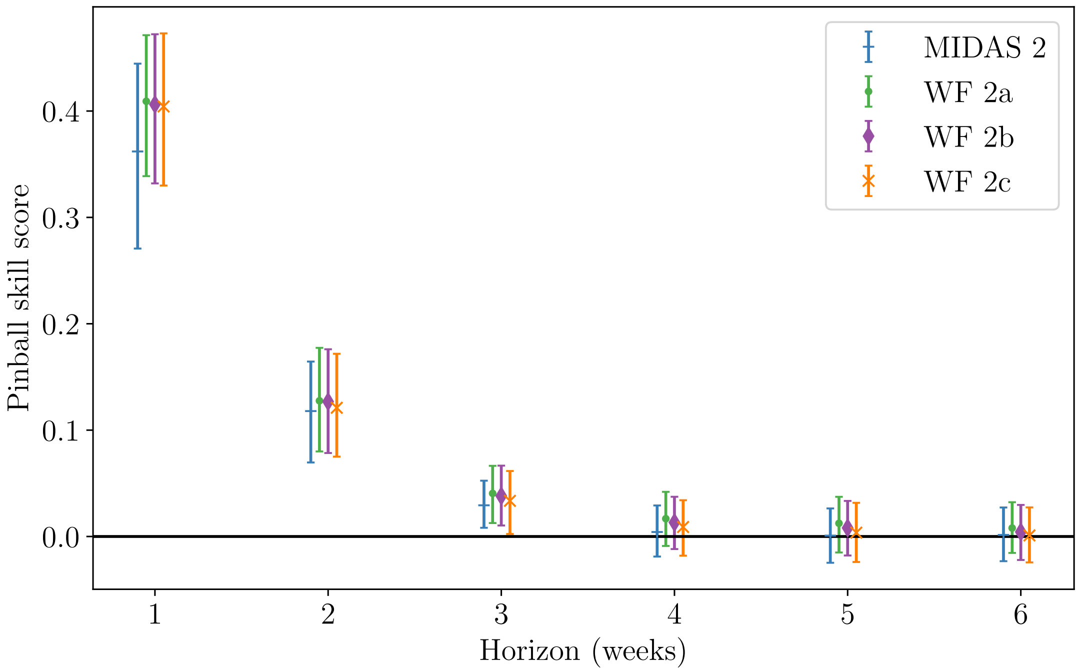


Figure D.1: Pinball skill score of weekly mean wind speed index relative to climatology. The forecasts labelled ‘WF 2’ are based on corrected ERA 5 data for that wind farm. Error bars show 95% bootstrapped confidence intervals.

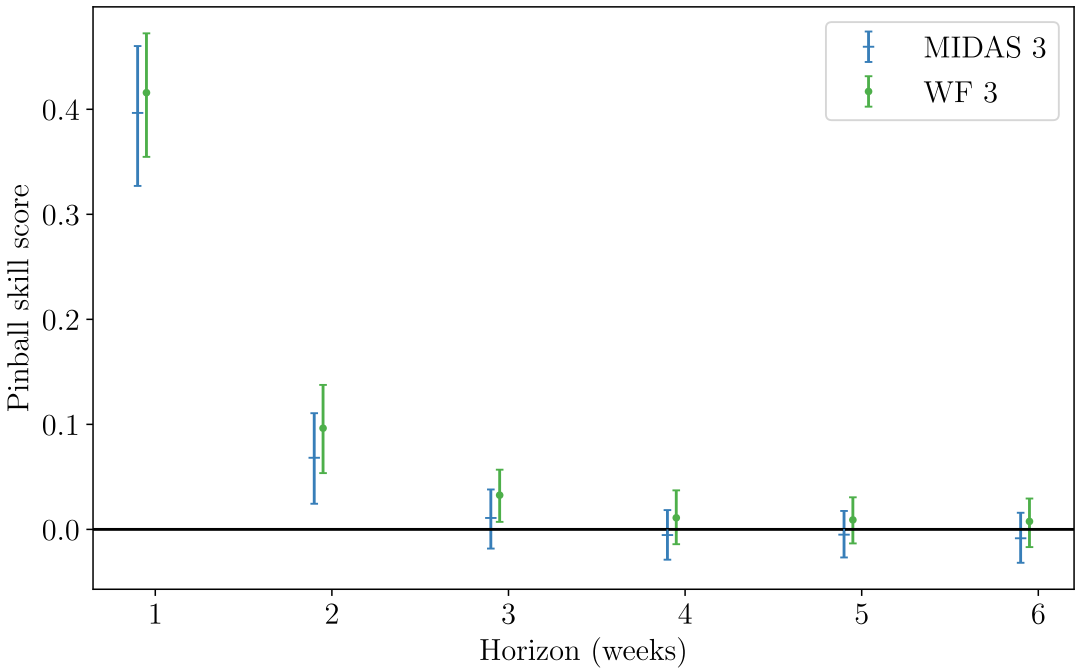


Figure D.2: Pinball skill score of weekly mean wind speed index relative to climatology. The forecasts labelled ‘WF 3’ are based on corrected ERA 5 data for that wind farm. Error bars show 95% bootstrapped confidence intervals.

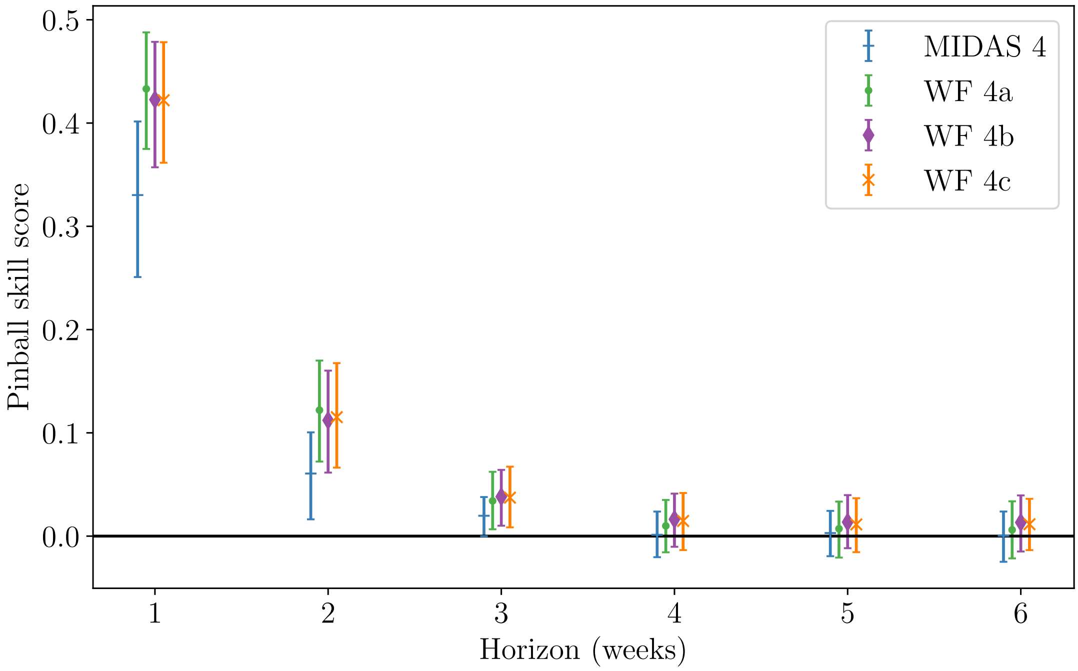


Figure D.3: Pinball skill score of weekly mean wind speed index relative to climatology. The forecasts labelled ‘WF 4’ are based on corrected ERA 5 data for that wind farm. Error bars show 95% bootstrapped confidence intervals.

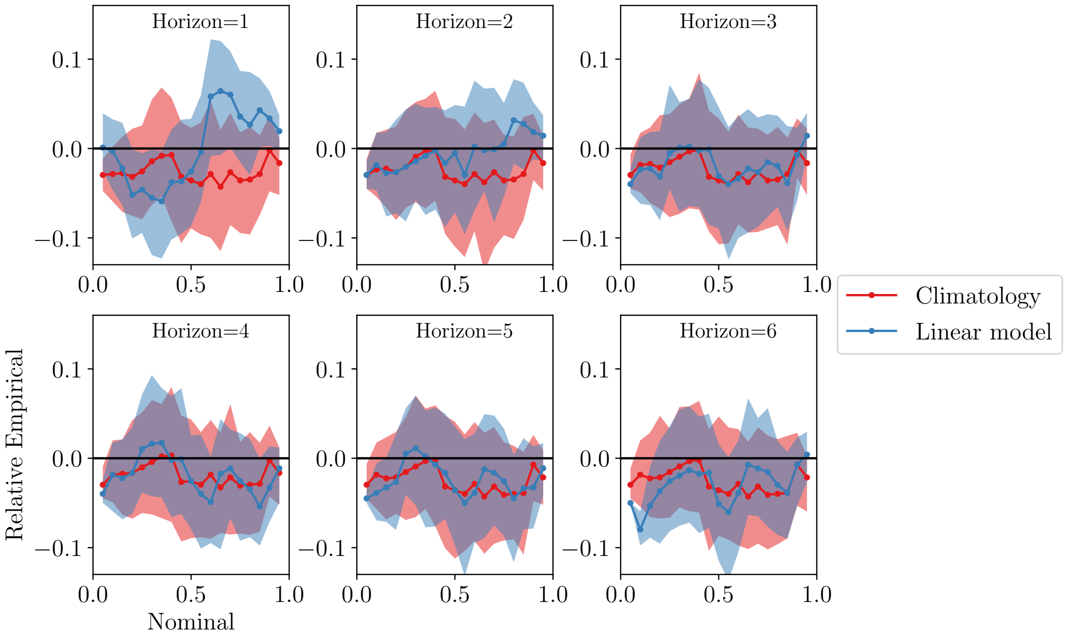


Figure D.4: Reliability of weekly mean wind speed forecast at MIDAS 2. Intervals show 95% bootstrapped confidence bands.

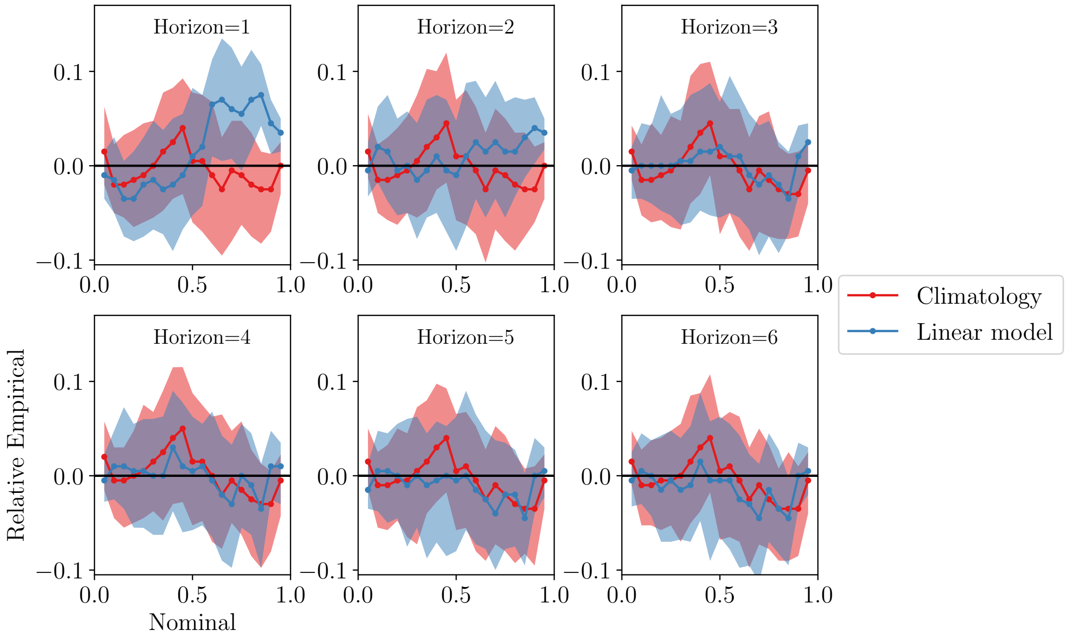


Figure D.5: Reliability of weekly mean wind speed forecast at MIDAS 3. Intervals show 95% bootstrapped confidence bands.

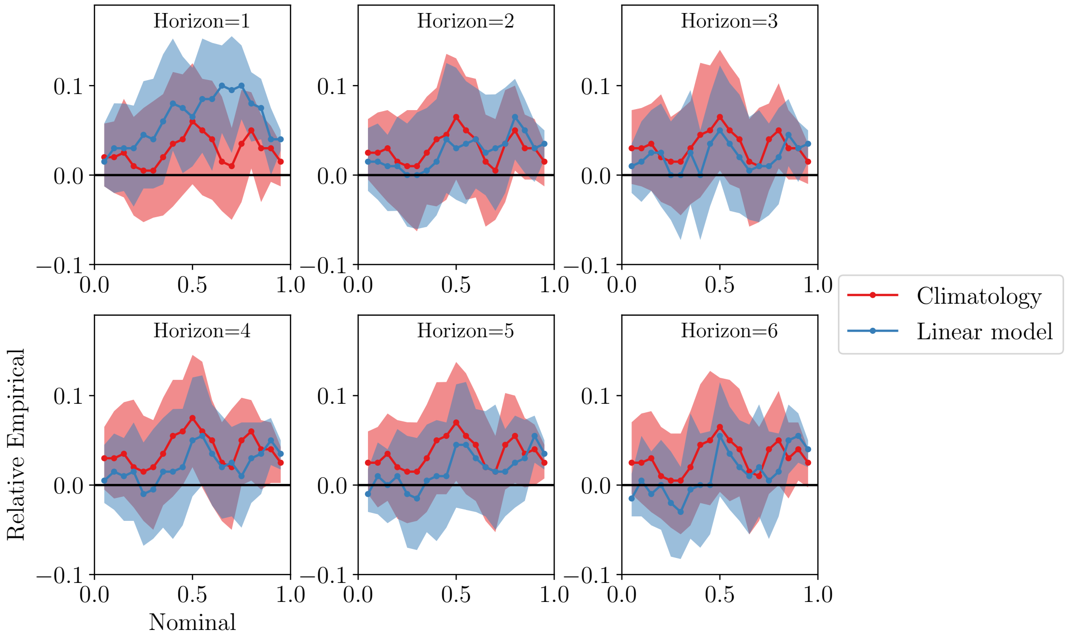


Figure D.6: Reliability of weekly mean wind speed forecast at MIDAS 4. Intervals show 95% bootstrapped confidence bands.

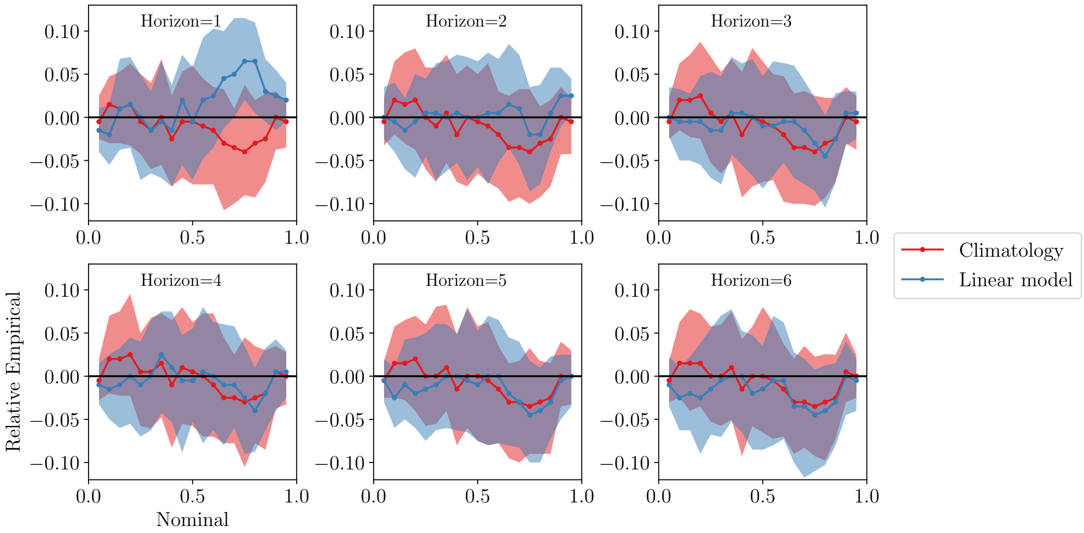


Figure D.7: Reliability of weekly mean wind speed forecast at WF2a. Intervals show 95% bootstrapped confidence bands.

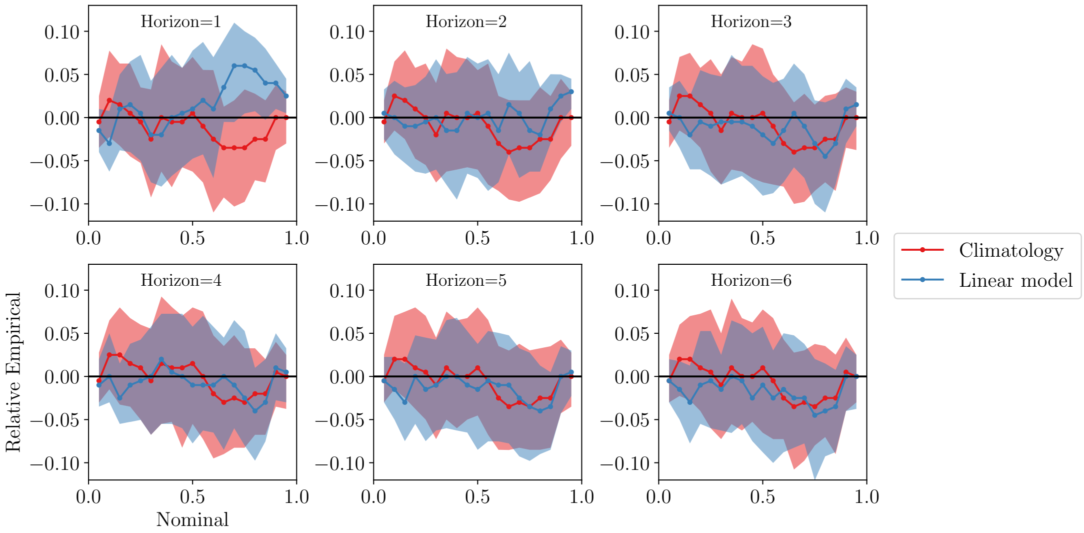


Figure D.8: Reliability of weekly mean wind speed forecast at WF2b. Intervals show 95% bootstrapped confidence bands.

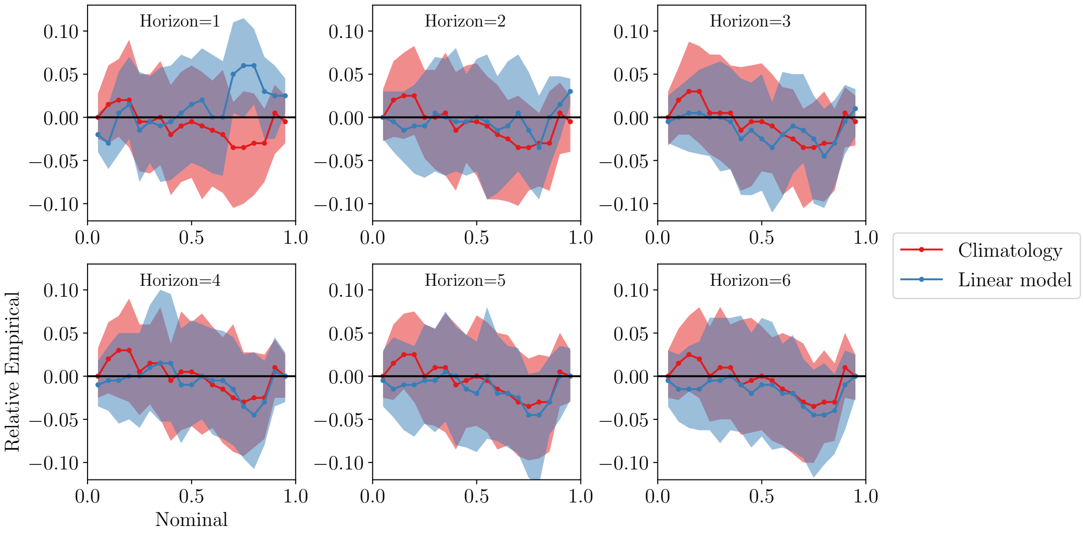


Figure D.9: Reliability of weekly mean wind speed forecast at WF2c. Intervals show 95% bootstrapped confidence bands.

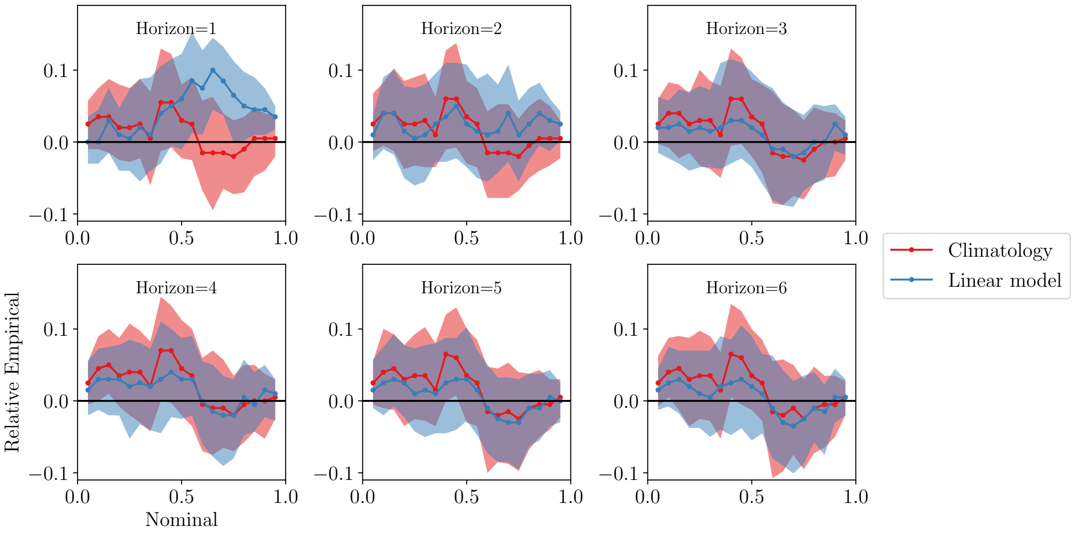


Figure D.10: Reliability of weekly mean wind speed forecast at WF3. Intervals show 95% bootstrapped confidence bands.

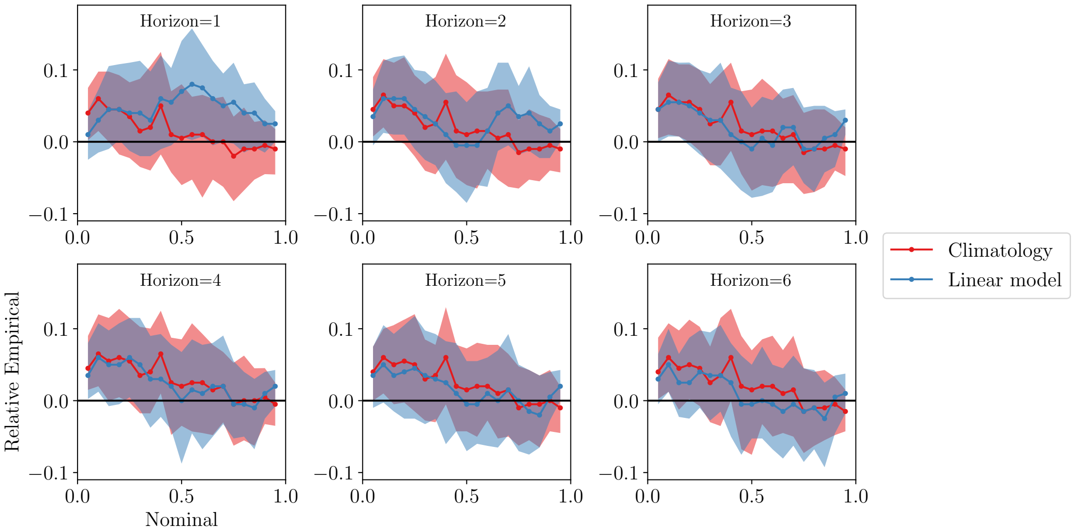


Figure D.11: Reliability of weekly mean wind speed forecast at WF4a. Intervals show 95% bootstrapped confidence bands.

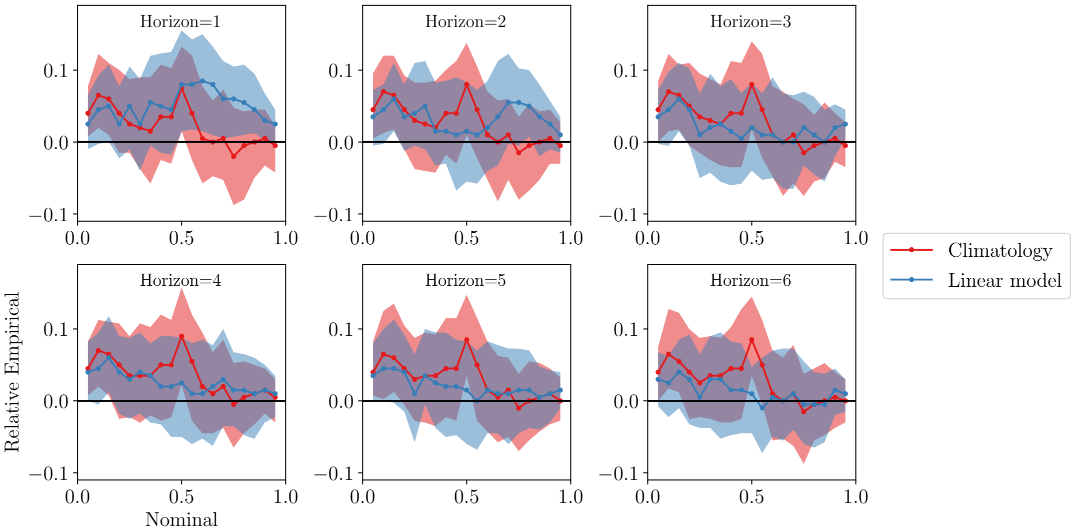


Figure D.12: Reliability of weekly mean wind speed forecast at WF4b. Intervals show 95% bootstrapped confidence bands.

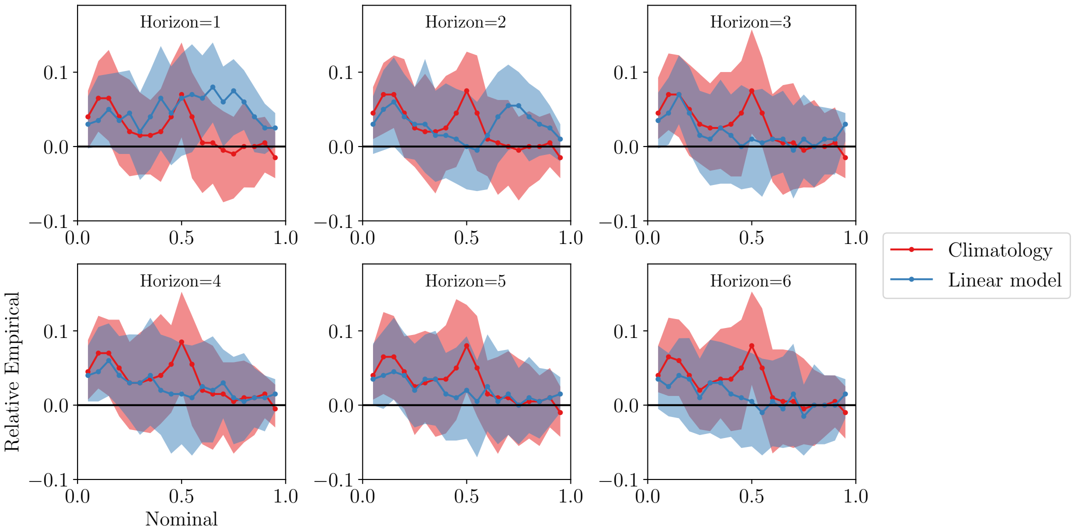


Figure D.13: Reliability of weekly mean wind speed forecast at WF4c. Intervals show 95% bootstrapped confidence bands.

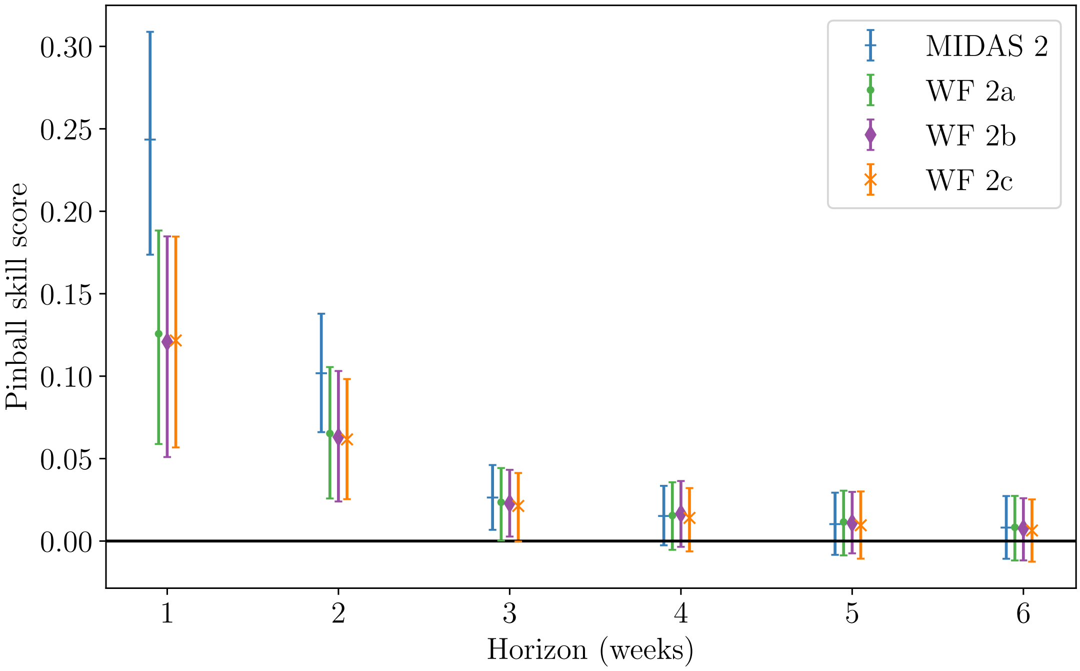


Figure D.14: Pinball skill score of weekly variability (weekly standard deviation of hourly wind speeds) relative to climatology. The forecasts labelled ‘WF 2’ are based on corrected ERA 5 data for that wind farm. Error bars show 95% bootstrapped confidence intervals.

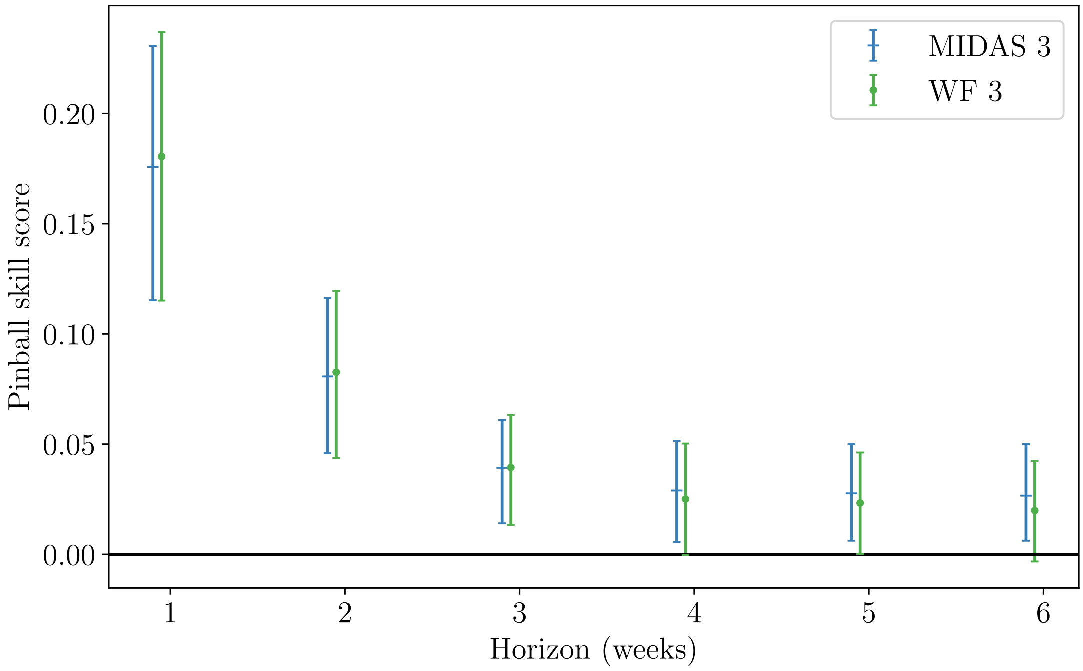


Figure D.15: Pinball skill score of weekly variability (weekly standard deviation of hourly wind speeds) relative to climatology. The forecasts labelled ‘WF 3’ are based on corrected ERA 5 data for that wind farm. Error bars show 95% bootstrapped confidence intervals.

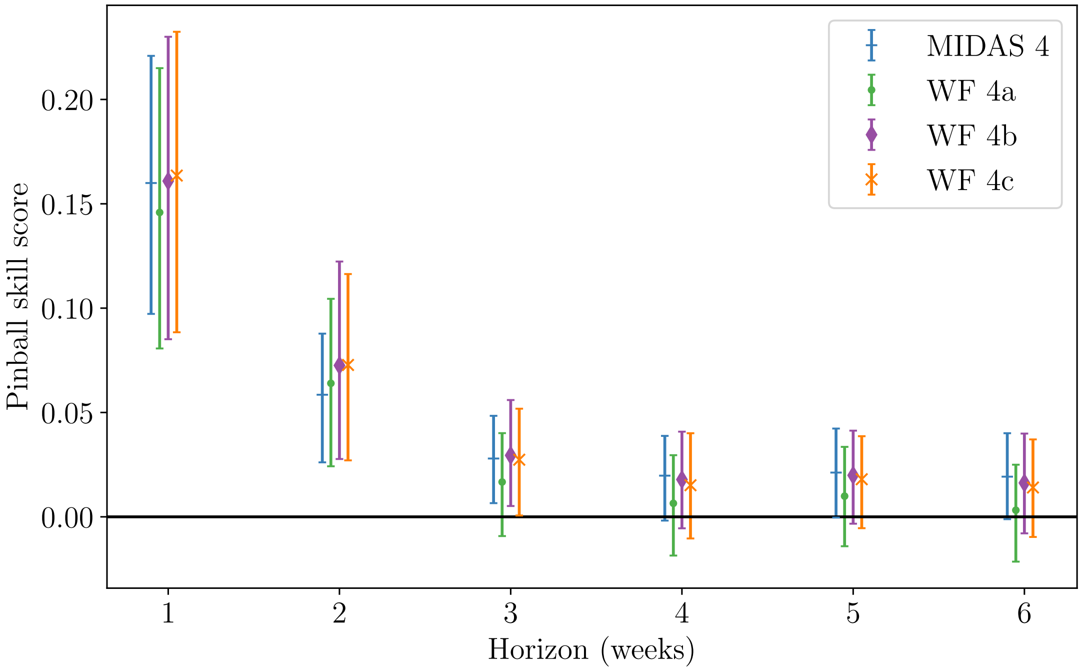


Figure D.16: Pinball skill score of weekly variability (weekly standard deviation of hourly wind speeds) relative to climatology. The forecasts labelled ‘WF 4’ are based on corrected ERA 5 data for that wind farm. Error bars show 95% bootstrapped confidence intervals.

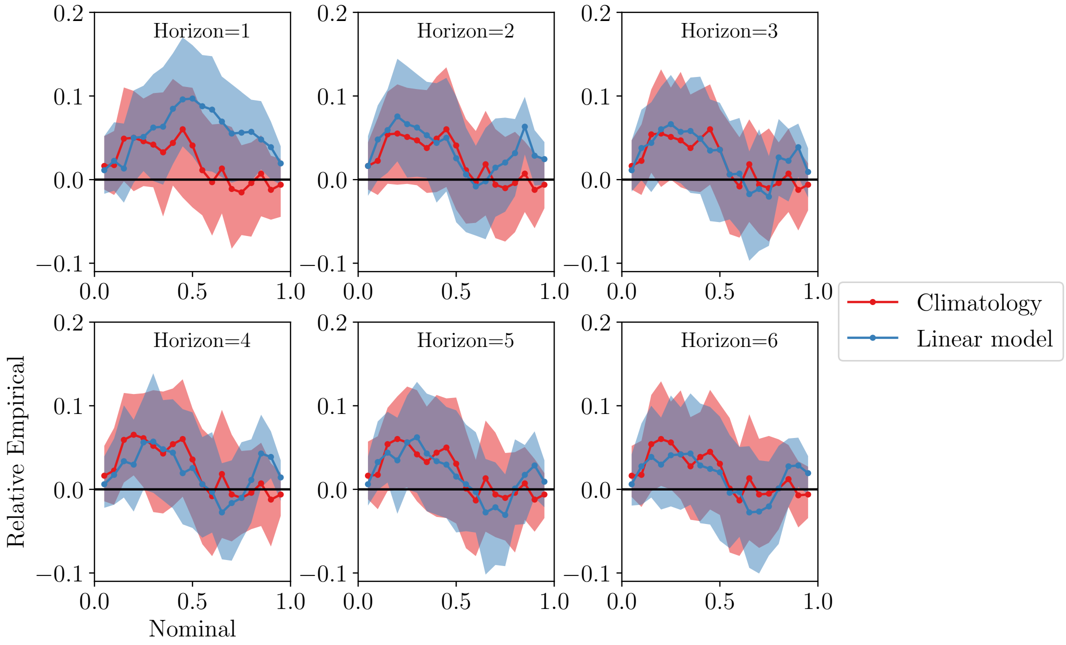


Figure D.17: Reliability of variability forecast at MIDAS 2. Intervals show 95% bootstrapped confidence bands.

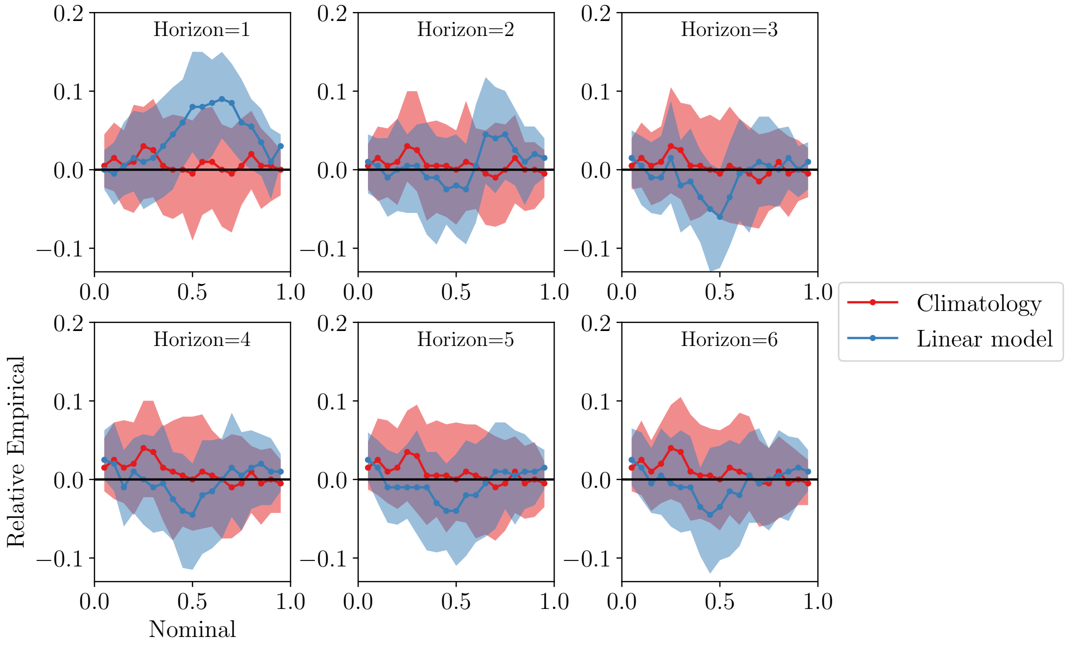


Figure D.18: Reliability of variability forecast at MIDAS 3. Intervals show 95% bootstrapped confidence bands.

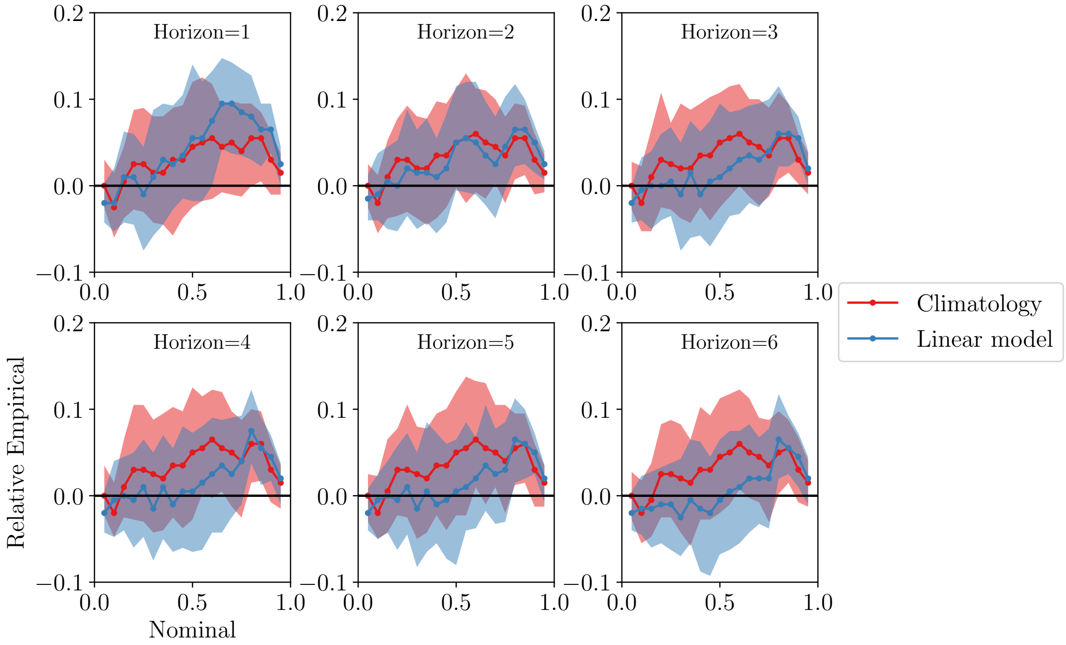


Figure D.19: Reliability of variability forecast at MIDAS 4. Intervals show 95% bootstrapped confidence bands.

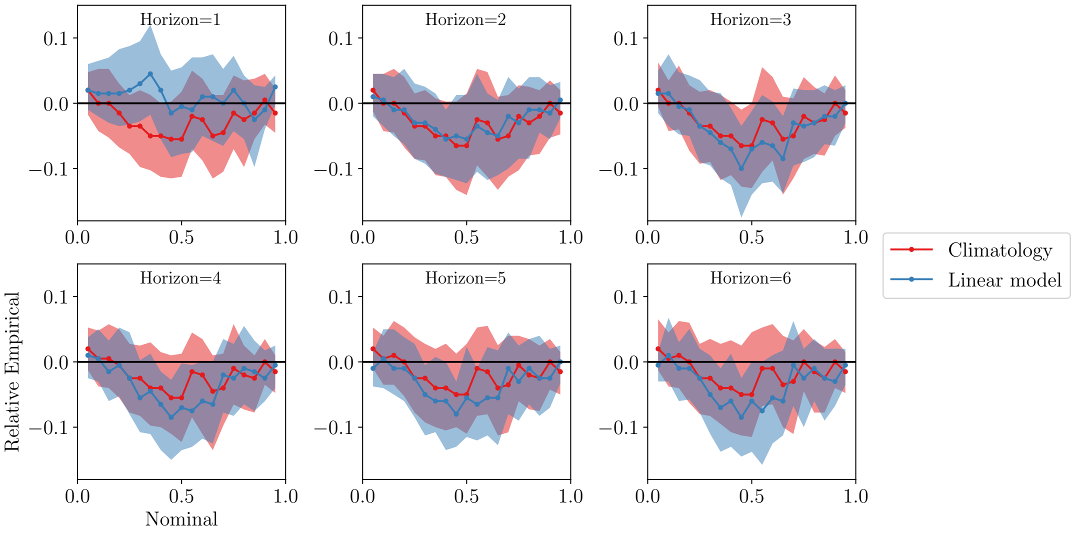


Figure D.20: Reliability of variability forecast at WF2a. Intervals show 95% bootstrapped confidence bands.

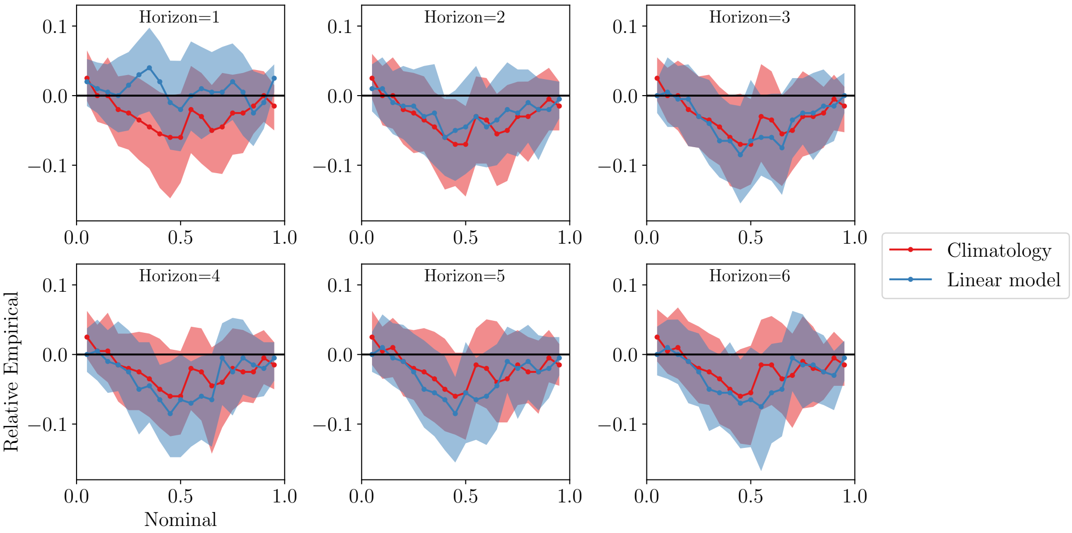


Figure D.21: Reliability of variability forecast at WF2b. Intervals show 95% bootstrapped confidence bands.

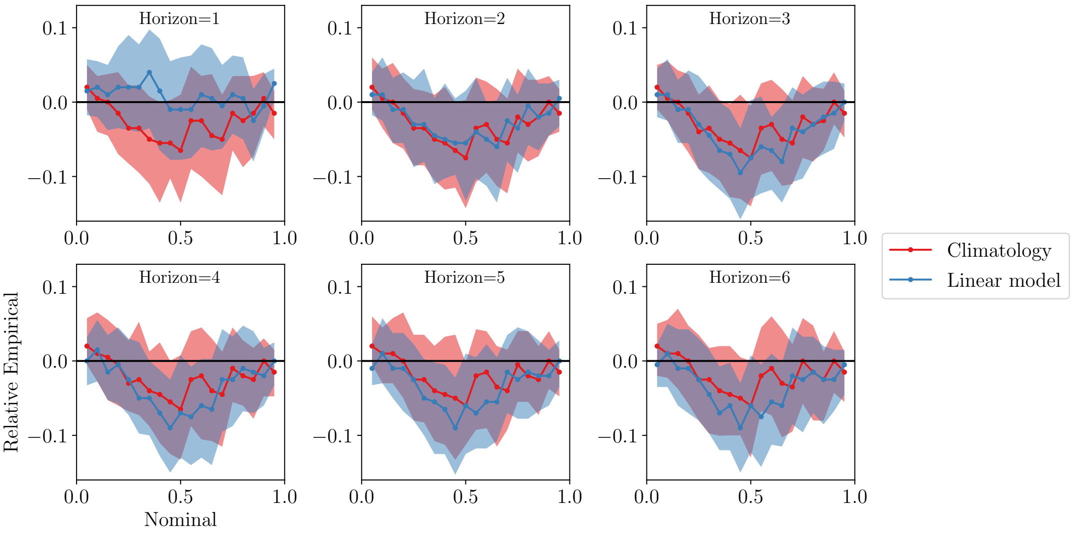


Figure D.22: Reliability of variability forecast at WF2c. Intervals show 95% bootstrapped confidence bands.

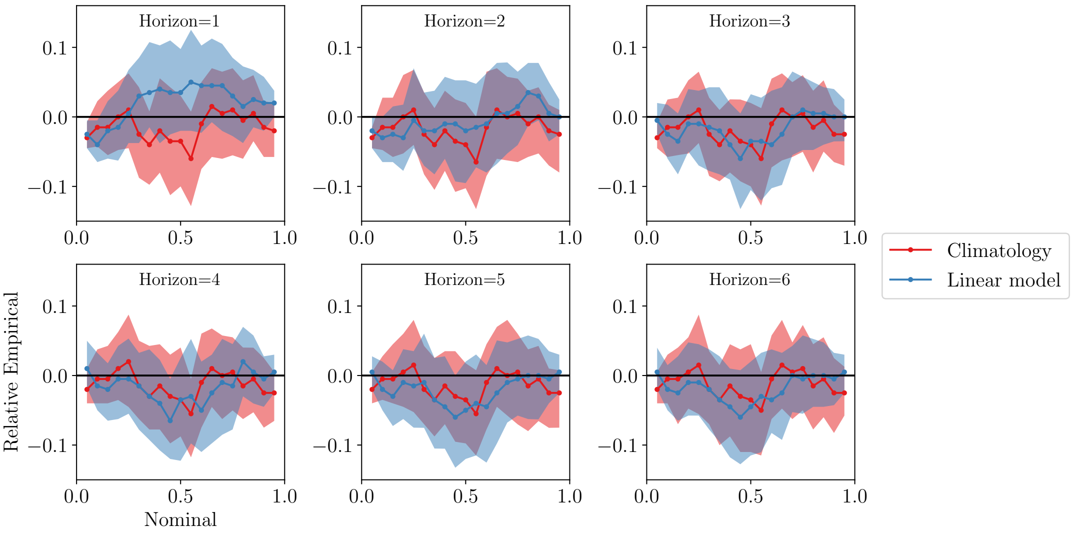


Figure D.23: Reliability of variability forecast at WF3. Intervals show 95% bootstrapped confidence bands.

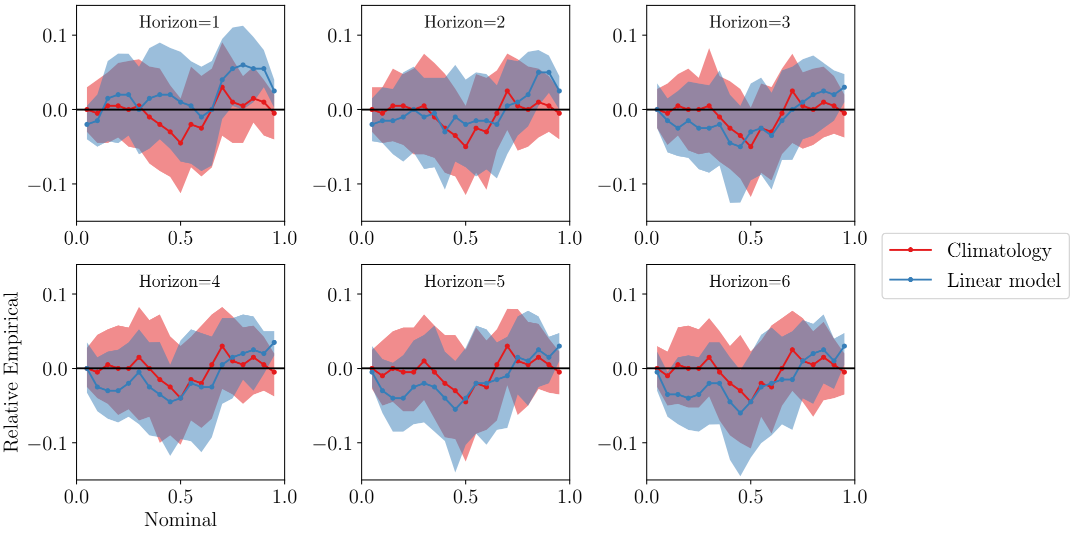


Figure D.24: Reliability of variability forecast at WF4a. Intervals show 95% bootstrapped confidence bands.

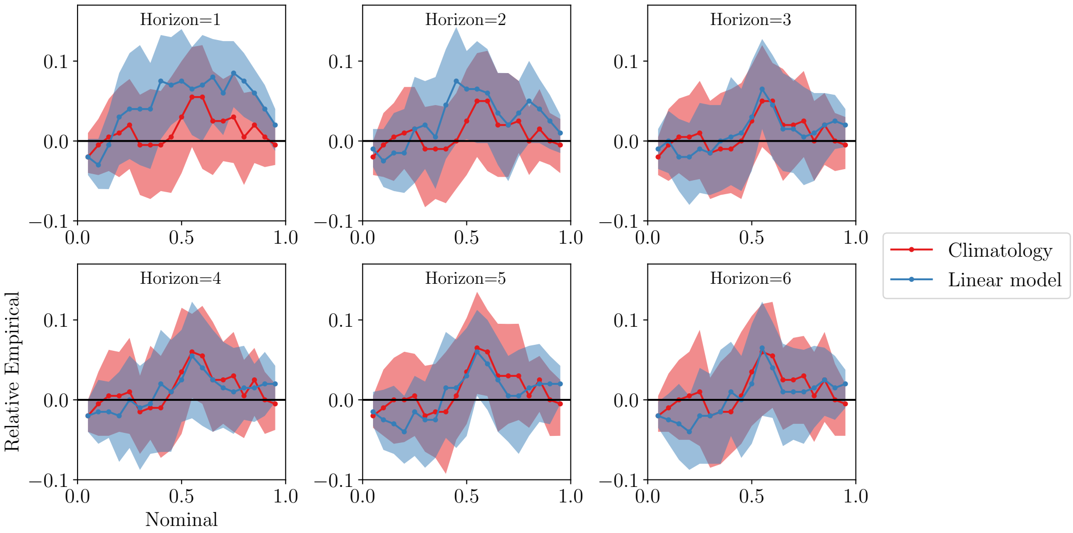


Figure D.25: Reliability of variability forecast at WF4b. Intervals show 95% bootstrapped confidence bands.

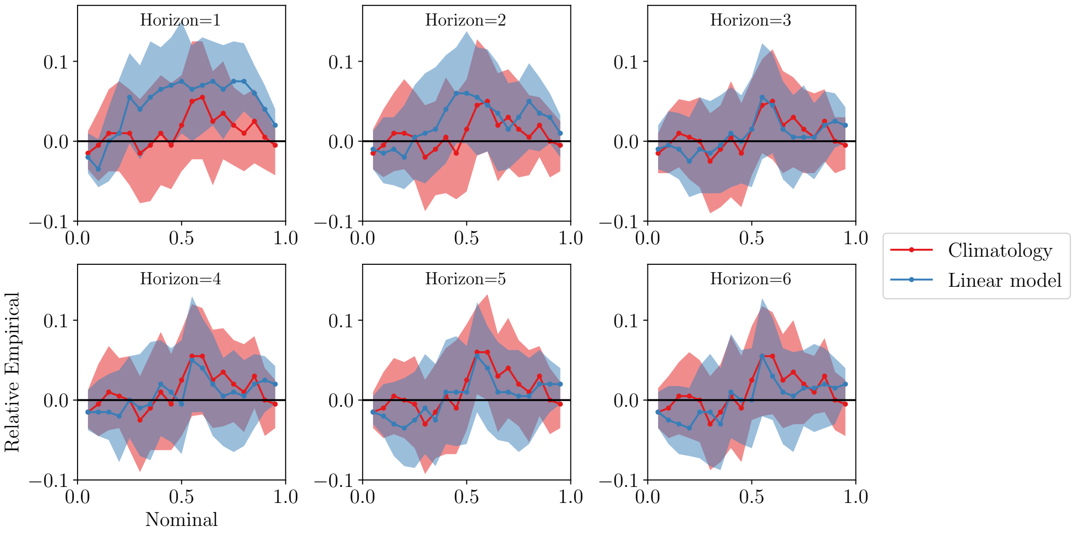


Figure D.26: Reliability of variability forecast at WF4c. Intervals show 95% bootstrapped confidence bands.

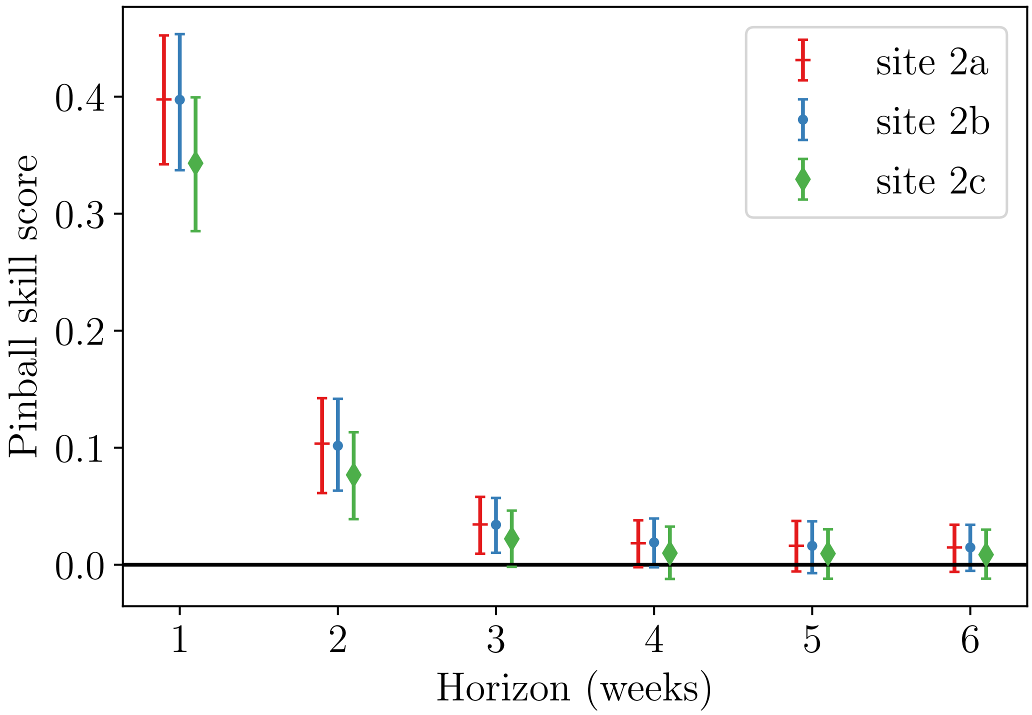


Figure D.27: Pinball skill score of weather window index relative to climatology. for area 2. Error bars show 95% bootstrapped confidence intervals.

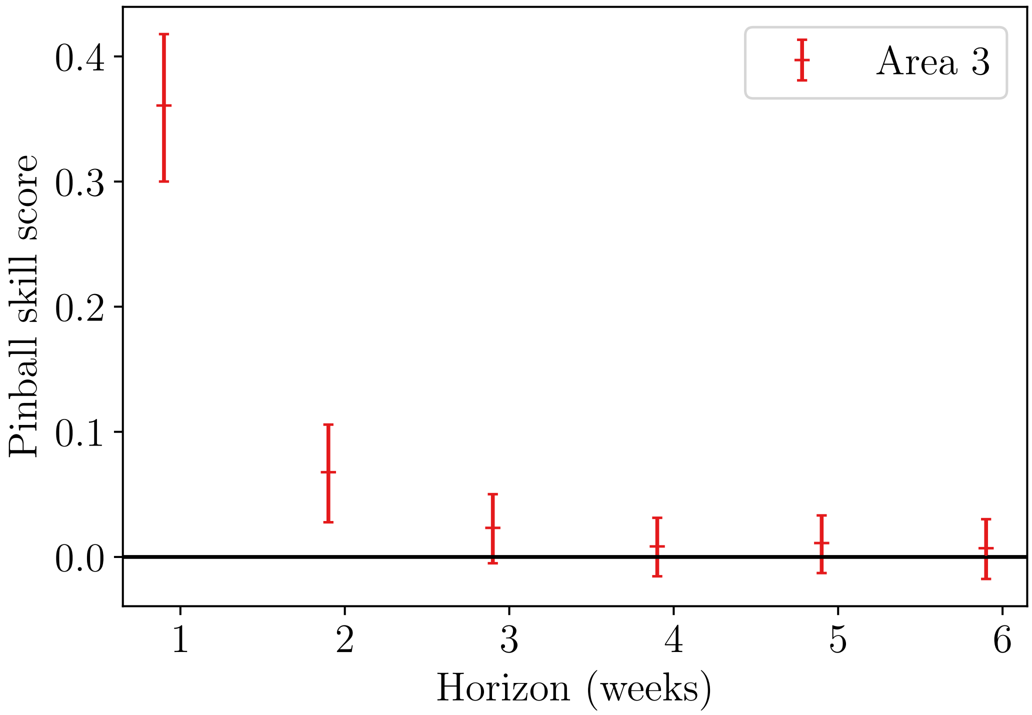


Figure D.28: Pinball skill score of weather window index relative to climatology. for area 3. Error bars show 95% bootstrapped confidence intervals.

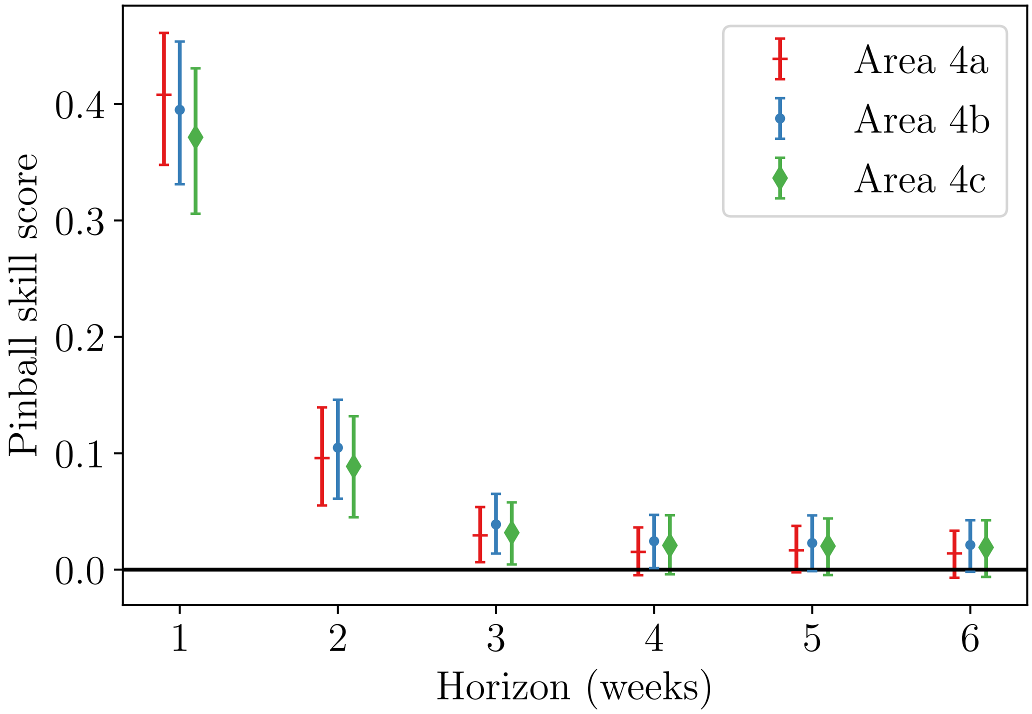


Figure D.29: Pinball skill score of weather window index relative to climatology. for area 4. Error bars show 95% bootstrapped confidence intervals.

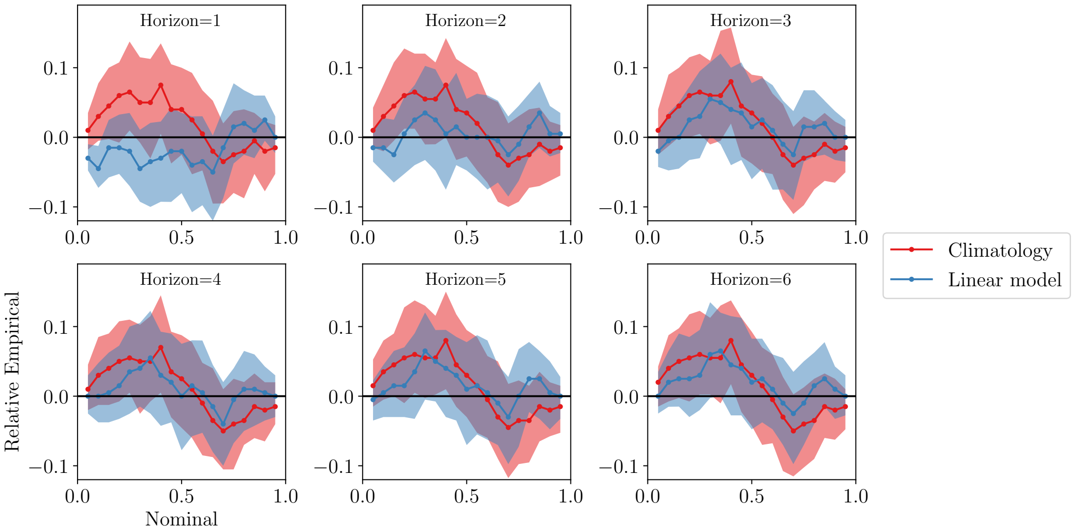


Figure D.30: Reliability of weather window index forecast at WF2a. Intervals show 95% bootstrapped confidence bands.

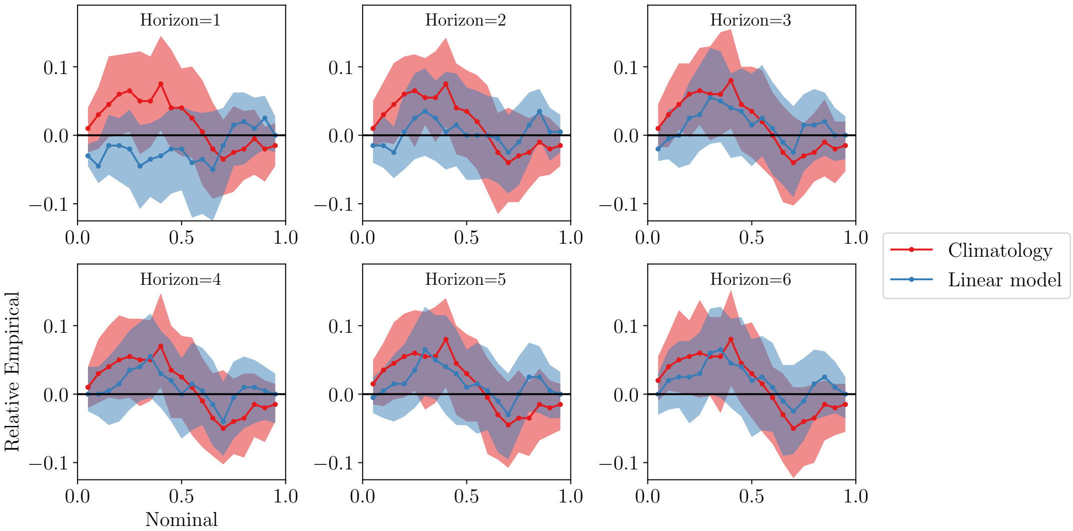


Figure D.31: Reliability of weather window index forecast at WF2b. Intervals show 95% bootstrapped confidence bands.

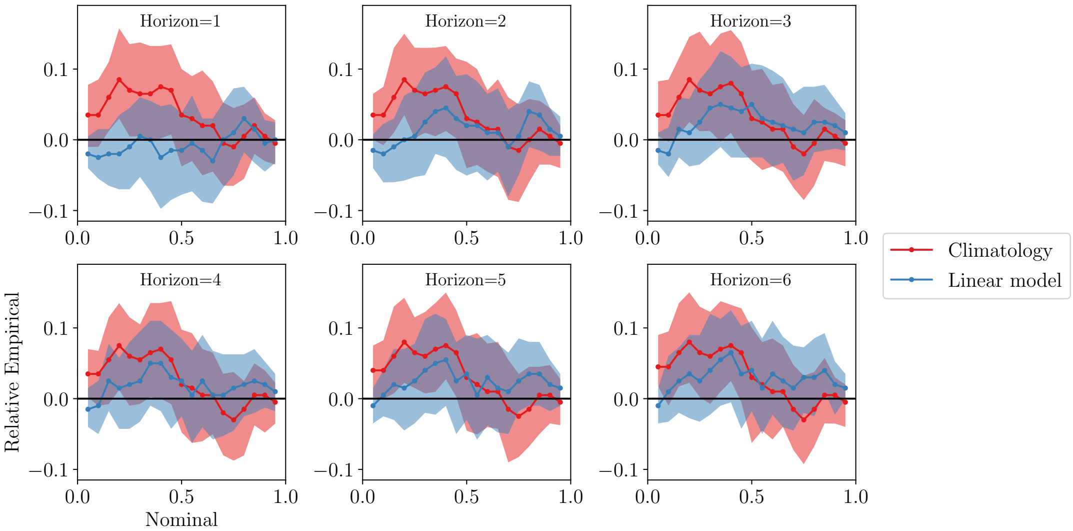


Figure D.32: Reliability of weather window index forecast at WF2c. Intervals show 95% bootstrapped confidence bands.

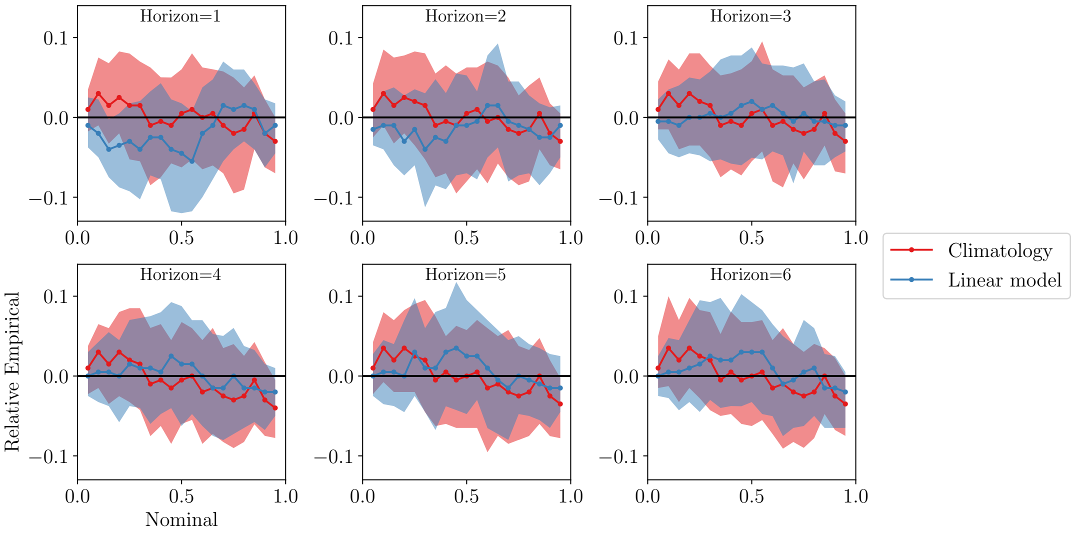


Figure D.33: Reliability of weather window index forecast at WF3. Intervals show 95% bootstrapped confidence bands.

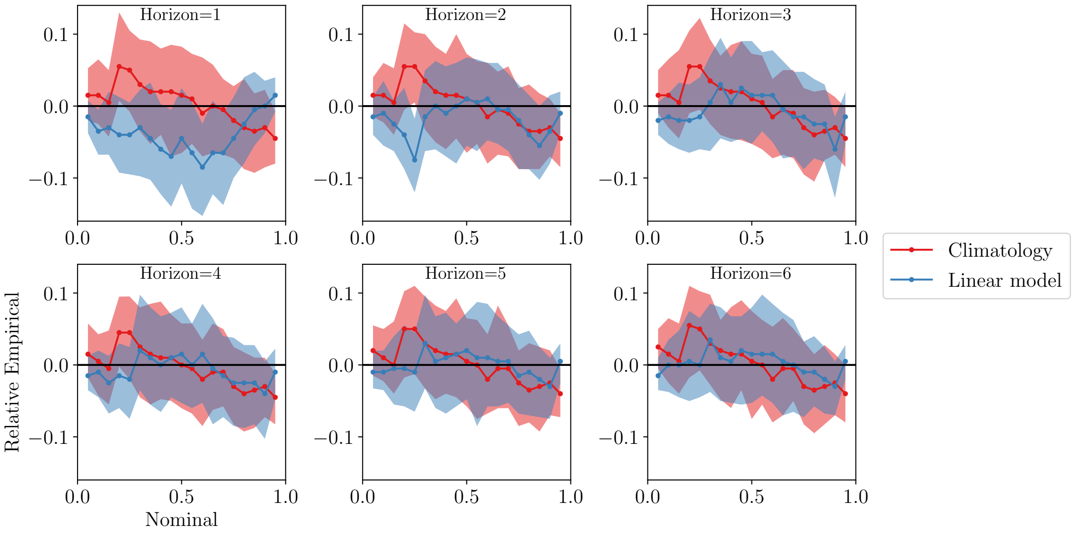


Figure D.34: Reliability of weather window index forecast at WF4a. Intervals show 95% bootstrapped confidence bands.

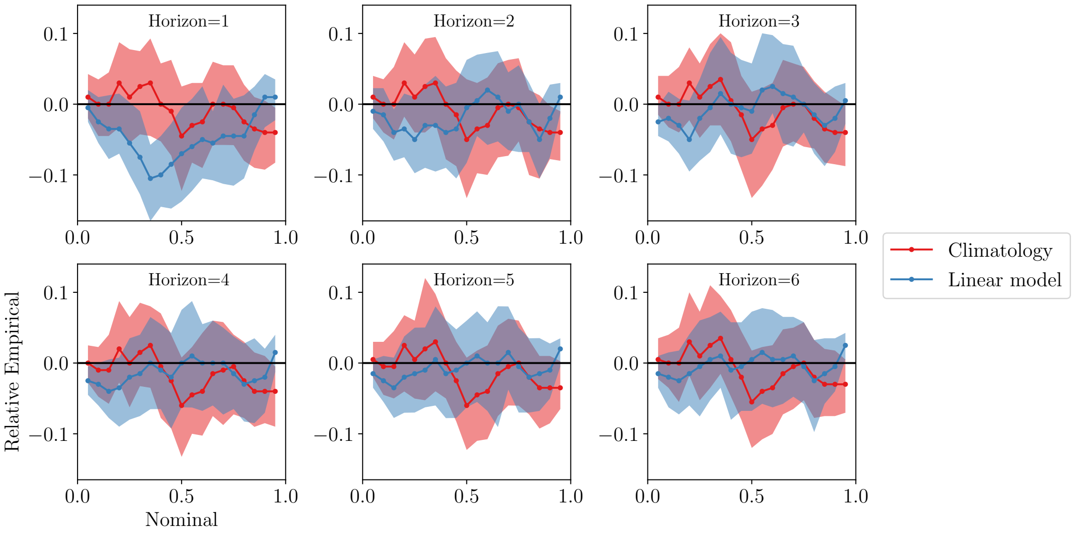


Figure D.35: Reliability of weather window index forecast at WF4b. Intervals show 95% bootstrapped confidence bands.

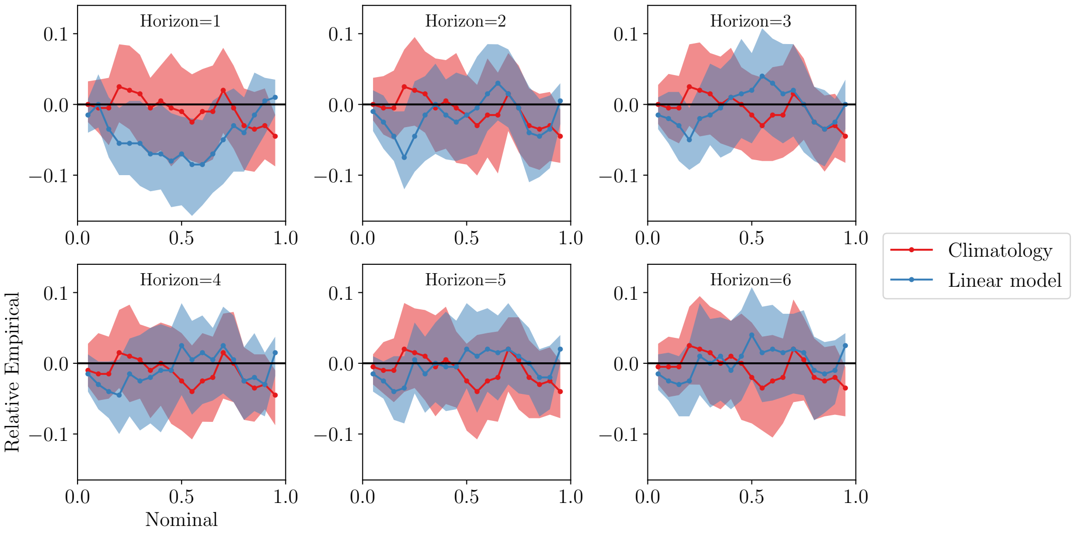


Figure D.36: Reliability of weather window index forecast at WF4c. Intervals show 95% bootstrapped confidence bands.