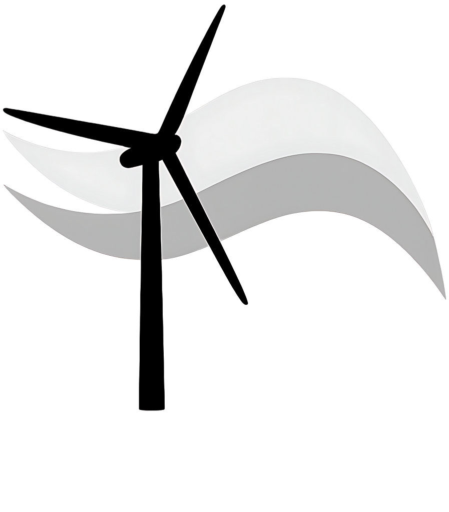
Department of Statistical Science

The Certainty of Uncertainty: Adaptive Conformal Prediction for Wind Farm Forecasting

Carl Somers

Supervised by Dr F. Javier Rubio & Domna Ladopoulou



Submitted in partial fulfilment of the requirements for the degree of MSc in Statistics

September 2025

Abstract

Reliable power forecasting lies at the heart of the transition to renewable energy. Currently the UK relies on wind energy for approximately 30% of its power needs. However, replacing readily dispatchable fossil fuels with inherently variable wind power means it is critical for grid operators to have accurate forecasts, and a clear understanding of their uncertainty. Yet, existing probabilistic forecasting methods often rely on strong distributional assumptions, lacklustre coverage guarantees, or methods computationally prohibitive for real-time applications. This thesis addresses this gap by applying and evaluating online Adaptive Conformal Inference (ACI) as a distribution-free alternative for wind power forecasting.

We integrate ACI and its adaptations with a state-of-the-art Spatio-Temporal Graph Convolutional Network to explicitly model the complex spatio-temporal dependencies between turbines. We validate our methodology through a range of synthetic experiments and case study on real operational data from Kelmarsh Wind Farm. Our results demonstrate the benefits of ACI-based methods for consistently valid and efficient coverage even under distributional shift, offering to turn any point predictor into a probabilistic model.

Acknowledgements

This thesis has been both challenging and rewarding, but impossible without the help of those around me. My thanks go first to Javier for his guidance and unwavering support throughout, and to Domna for her patience and generosity with her time. I am also grateful to my family and friends who have kept me sane - and even motivated my interest in wind farms, and (lest he feel forgotten) to a dog back home in Ireland, who endured my work-from-home routine long before this MSc began.

Contents

Abbreviations

ACE	Average Coverage Error
ACI	Adaptive Conformal Inference
AgACI	Online Expert Aggregation ACI
ARIMA	Auto-Regressive Integrated Moving Average
CP	Conformal Prediction
DtACI	Dynamically Tuned ACI
GCN	Graph Convolutional Network
GRU	Gated Recurrent Unit
LGBM	Light Gradient Boosting Machine
MAE	Mean Absolute Error
NWP	Numerical Weather Prediction
OSSCP	Online Sequential Split Conformal Prediction
QR	Quantile Regression
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
SCADA	Supervisory Control and Data Acquisition
STGCN	Spatio-Temporal Graph Convolutional Network
TCN	Temporal Convolutional Network
UQ	Uncertainty Quantification
WS	Winkler Score

1 Introduction

> "All models are wrong, but some are useful."
>
> *George E. P. Box*

Wind energy has rapidly become one of the worlds fastest-growing global energy sources, generating over 2,300 terawatt-hours annually, with countries like the UK often exceeding a third of their power from wind (International Energy Agency, 2024; National Grid Electricity System Operator (ESO), 2025). This growth introduces a fundamental challenge whereby wind's inherent variability creates supply-demand mismatches that strain grids designed for dispatchable fossil fuel generation. Operators must commit reserves, schedule units and set market prices hours ahead, yet face significant penalties for any deviation from their generation schedule (Chen et al., 2024). Consequently, mere point forecasts are insufficient and they require probabilistic outputs that capture the inherent uncertainty of modern power generation.

Most existing uncertainty-quantification methods (e.g Quantile Regression (QR), Bayesian Neural Networks, Gaussian Processes) either rely on assumptions or requirements unattainable in real-world online forecasting, limiting the underlying models that can be used. Conformal Prediction (CP) offers a principled alternative by providing finite-sample, distribution-free prediction intervals with a user-specified α to provide theoretical $1 - \alpha$ coverage guarantees for any base predictor (Vovk et al., 2005; Shafer & Vovk, 2008).

Its model-agnostic property has led to a rapid rise in popularity and proven success across diverse fields, from image classification to survival analysis, making it a compelling framework for the challenge of wind power forecasting (Angelopoulos & Bates, 2021). However, standard CP assumes *exchangeability*, an assumption fundamentally violated by wind power time series, which exhibit strong temporal autocorrelation, seasonal patterns, non-stationarity, and even spatial dependencies between turbines. Naively applying CP can yield wide or miscalibrated intervals.

Recent research tackles this by adapting CP under distributional shift, notably, in the online case through Adaptive Conformal Inference (ACI) and its variants AgACI and DtACI. These algorithms adjust interval width on the fly and retain valid coverage without retraining of the underlying point forecasting model. In this thesis we exploit these advances for wind-farm forecasting. Our contributions are:

1. Novel application of ACI to wind farm power forecasting.
2. Integrating a complex spatio-temporal deep-learning forecaster (STGCN).
3. Comprehensive evaluation under both controlled synthetic scenarios and a real case study (Kelmarsh Wind Farm), demonstrating reliable coverage under extreme exchangeability violations.

This thesis aims to deliver theoretically valid, operationally sharp prediction intervals to close the key gap between conformal theory and the practical needs of a renewable dominated power grid. In line with our goal, our full code implementation is available on GitHub[1].

The remainder of our paper is structured as follows. We first review the relevant literature spanning wind power forecasting, and conformal prediction. We detail our proposed methodology, followed by a rigorous evaluation on both synthetic and real-world data, before concluding with a discussion of our work.

[1] https://github.com/24223499/conformal_forecasting

2 Related Work

This section provides a comprehensive review of the literature pertinent to wind energy forecasting and uncertainty quantification. We specifically focus on the evolution and application of conformal prediction, its various extensions, and their utility in enhancing the reliability of wind energy predictions.

2.1 Wind Power Forecasting

Early deterministic wind power forecasting efforts relied predominantly on physical modelling approaches, which simulate atmospheric dynamics through Numerical Weather Prediction (NWP) models with power curve transformations. These methods achieve reasonable accuracy for day-ahead forecasts but struggle with high-resolution, temporal prediction at turbine levels due to computational costs and error accumulation from meteorological inputs.

Statistical models such as Auto-Regressive Integrated Moving Average (ARIMA) provided computationally efficient alternatives, capturing temporal dependencies through established time series frameworks (Kavasseri & Seetharaman, 2009). However, these linear models assume stationarity and Gaussian errors - assumptions frequently violated in wind data due to seasonal variations, weather regime changes and the non-linear wind speed-power (WS/P) relationship[1].

The limitations of traditional approaches catalysed the adoption of machine learning techniques, offering flexibility to capture complex non-linear relationships without explicit physical modelling. Support Vector Machines (SVMs) emerged as early leaders, demonstrating robust performance through kernel methods and resilience to overfitting (Zendehboudi et al., 2018). Ensemble methods, particularly Random Forests and gradient boosting techniques like XGBoost, subsequently excelled in computationally efficient forecasting, handling mixed meteorological and temporal features effectively.

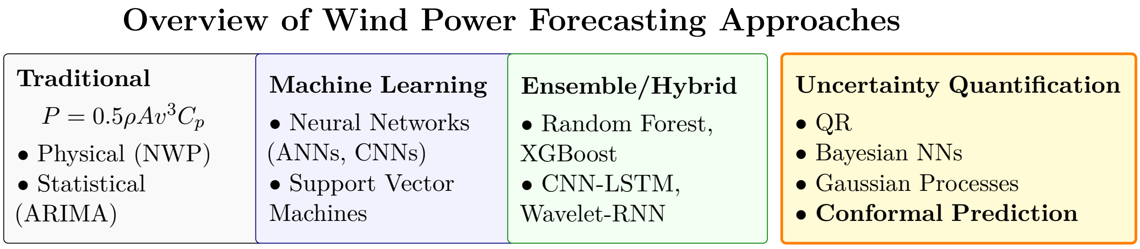
Deep learning fundamentally transformed wind forecasting through architectures designed for temporal sequence modelling. Long Short-Term Memory (LSTM) networks addressed limitations in modelling long-range dependence through gating mechanisms, while Gated Recurrent Unit (GRU) offer computational efficiency with comparable performance. Recent research increasingly favours hybrid approaches combining multiple architectures, such as CNN-LSTM for spatial-temporal patterns, or wavelet transform preprocessing with recurrent networks for multi-scale decomposition (Y. Wang et al., 2021).

Despite significant advances in point forecasting accuracy, Uncertainty Quantification

[1] Theoretically, power output is given by: $P = 0.5\rho A v^3 C_p$ where P is power output, v is wind speed, C_p is a performance coefficient, and A is the rotor swept area. However, this relationship degrades over time, new turbines may match the manufacturer's power curve, but older ones often fall short.

(UQ) remains inadequately addressed for operational requirements. Given wind's inherent variability, deterministic forecasts provide insufficient information for optimal grid decision-making under uncertainty.

Recent research has shifted toward probabilistic forecasting approaches. QR methods enable prediction interval construction while maintaining computational efficiency, but require separate models for each quantile (Cui et al., 2023). Bayesian neural networks provide principled UQ through prior distributions on weights, but typically require computationally expensive inference procedures such as variational inference or Markov Chain Monte Carlo (MCMC) (Zou et al., 2022). In contrast, Gaussian Process models provide flexible, kernel-based uncertainty estimates, but encounter significant scalability challenges (e.g training time & cost) with large spatio-temporal datasets (Ladopoulou et al., 2025). These existing probabilistic approaches often require strong distributional assumptions about forecast errors, lack coverage guarantees, or impose expensive training procedures impractical for real-time operations. Conformal prediction directly addresses these limitations.



Overview of Wind Power Forecasting Approaches

Traditional	**Machine Learning**	**Ensemble/Hybrid**	**Uncertainty Quantification**
$P = 0.5\rho A v^3 C_p$ • Physical (NWP) • Statistical (ARIMA)	• Neural Networks (ANNs, CNNs) • Support Vector Machines	• Random Forest, XGBoost • CNN-LSTM, Wavelet-RNN	• QR • Bayesian NNs • Gaussian Processes • **Conformal Prediction**

Figure 2.1: Taxonomy of wind power prediction (Deterministic vs Probabilistic)

2.2 Fundamentals of Conformal Prediction

Conformal prediction, a distribution-free framework for uncertainty quantification, originated from Vladimir Vovk[2], Alexander Gammerman, and Vladimir Vapnik's collaboration at Royal Holloway in the mid-1990s, building upon Vapnik's transduction principles (Vovk et al., 2005; Saunders et al., 1998). Unlike traditional methods requiring strong distributional assumptions, this framework provides prediction sets with guaranteed $1-\alpha$ coverage under the weaker exchangeability assumption, making it applicable to a broader class of problems while maintaining finite-sample validity (Shafer & Vovk, 2008; J. Lei et al., 2018).

The theoretical foundations for CP rest on the assumption of *exchangeability*, which requires the joint distribution of observations remains invariant under permutations. This is often considered a much weaker assumption than traditional independence and identically distributed (iid) assumption, making conformal a much broader class of problem. Under exchangeability, conformal prediction provides marginal coverage guarantees for successive examples sampled independently from the true distribution ensuring the prediction intervals contain the true values with the specified nominal

[2] Vovk was one of the last PhD students of Andrei Kolmogorov, widely regarded as a founding figure in modern probability theory.

coverage probability. This sets out that we can create a predictive set, $\hat{C}_\alpha$, such that:

$$\mathbb{P}(Y^{(n+1)} \in \hat{C}_\alpha(X^{(n+1)})) \geq 1 - \alpha \tag{2.1}$$

Definition 1 (Exchangeability) *A sequence of random variables is exchangeable if its joint probability distribution remains unchanged under any permutation of its indices. A finite sequence $x_1, x_2, \ldots, x_n$ is exchangeable if for any permutation π of $\{1, 2, \ldots, n\}$:*

$$\mathbb{P}(x_1, x_2, \ldots, x_n) = \mathbb{P}(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(n)})$$

The core framework centres on conformity scores, which measure how well a new observation conforms (or not conforms) to the previously observed patterns. For a regression problem with training data $(x_1, y_1), \ldots, (x_n, y_n)$ and a new test point x_{n+1}, the conformity score typically quantifies the disagreement between the predicted value $\hat{y}_{n+1}$ and the true value y_{n+1}. Common choices include the absolute residual $|y_{n+1} - \hat{y}_{n+1}|$ or normalised variants that account for heteroskedasticity.

Definition 2 (Nonconformity Measure) *Given a feature–label pair (x, y) and a predictive model $\hat{f}$, a nonconformity measure/score is a function $s(x, y)$ that quantifies how atypical y is with respect to x and $\hat{f}$. In regression, a common choice is the absolute residual:*

$$S_t = s(x_t, y_t) = |y_t - \hat{f}(x_t)|.$$

The original conformal prediction procedure works by computing conformity scores for all training examples and the test example for each possible test label. A p-value for the candidate label is calculated as the proportion of training examples with conformity scores larger or equal to the test conformity score (Vovk et al., 2005). The prediction region at significance level α comprises all candidate labels with p-values exceeding α, ensuring that the true label is excluded from the prediction region with probability of, at most, α (Shafer & Vovk, 2008). However, this is computationally intensive so a more popular approach is inductive or split conformal prediction (SCP), which uses a calibration set rather than the whole dataset. This is demonstrated simply in Figure 2.2:

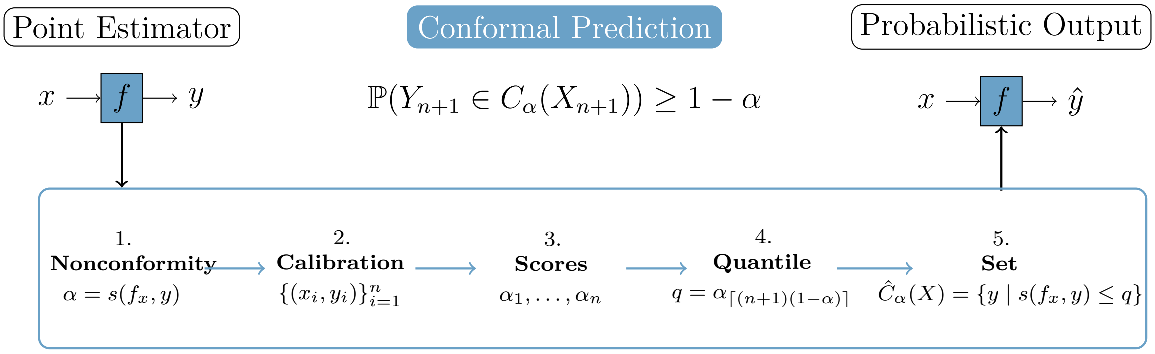

Point Estimator
Conformal Prediction
Probabilistic Output
x → f → y
$\mathbb{P}(Y_{n+1} \in C_\alpha(X_{n+1})) \geq 1 - \alpha$
x → f → $\hat{y}$
1. Nonconformity $\alpha = s(f_x, y)$
2. Calibration $\{(x_i, y_i)\}_{i=1}^n$
3. Scores $\alpha_1, \ldots, \alpha_n$
4. Quantile $q = \alpha_{\lceil (n+1)(1-\alpha) \rceil}$
5. Set $\hat{C}_\alpha(X) = \{y \mid s(f_x, y) \leq q\}$

Figure 2.2: Split Conformal Prediction Framework (Manokhin, 2023)

The validity of conformal prediction regions has two key properties: marginal coverage and conditional coverage (J. Lei et al., 2018; Romano et al., 2020). Marginal coverage ensures prediction regions contain true values with the specified probability across the entire test distribution. Conditional coverage, a stronger requirement, demands the coverage probability holds for any subset of the input space. While conformal prediction guarantees marginal coverage under exchangeability, achieving conditional coverage requires further assumptions or adaptive procedures (Gibbs et al., 2025).

Recent developments have significantly broadened conformal prediction beyond the classical exchangeability framework to handle diverse applied challenges. These include methods for temporal dependencies in stationary time series (Chernozhukov et al., 2018; Xu & Xie, 2023), panel and cross-sectional time series (Lin et al., 2022), distribution shifts including covariate shift (Tibshirani et al., 2019; L. Lei & Candès, 2021; Kasa et al., 2024) and label shift (Podkopaev & Ramdas, 2021), online adaptive inference (Gibbs & Candes, 2021; Zaffran et al., 2022), spatial and spatio-temporal prediction (Mao et al., 2024), and general frameworks for re-weighting non-identically distributed data (Barber et al., 2023). Additionally, computational advances have enabled conformal methods for high-dimensional settings, deep learning application, and multi-output prediction tasks (Angelopoulos & Bates, 2021).

2.3 Conformal Prediction for Time Series

Time series pose a distinct challenge for non-exchangeable conformal prediction due to temporal dependence, non-stationarity, and heteroskedasticity. The literature has developed along two main paths: one aims to restore approximate exchangeability via tailored data partitioning; the other embraces its violation through adaptive, online methods.

The theoretical foundation for CP under dependence was established by (Chernozhukov et al., 2018), who demonstrated that split CP (SCP) methods retain asymptotic coverage validity under weak dependence conditions, including strong mixing and martingale difference sequences.

Theorem 1 (Adapted from Chernozhukov et al. (2018)) *Under the following conditions:*

- *The data $\{Z_t = (x_t, y_t)\}_{t=1}^{T_0+T_1}$ form a weakly dependent process (e.g. strongly mixing with summable mixing coefficients).*
- *The conformity score function $S(\cdot)$ satisfies regularity conditions (measurability, bounded density).*
- *The number of permutation blocks $K \to \infty$ as $T_0 \to \infty$.*

Then the split conformal prediction method achieves asymptotic time-averaged coverage:

$$\lim_{T_0\to\infty} \frac{1}{T_1} \sum_{t=T_0+1}^{T_0+T_1} \mathbb{P}\Big(Y^{(t)} \in \widehat{\mathcal{C}}_\alpha^t(X^{(t)})\Big) = 1 - \alpha.$$

When data are exchangeable (e.g. i.i.d.), this method achieves exact finite-sample coverage for each prediction.

However, these asymptotic results only recover the finite-sample guarantees in the iid case, leaving practitioners with methods that exhibit poor transient behaviour, or fail under distributional shift. Usually, conformal time series relaxes (2.1) to:

$$\frac{1}{T_1}\sum_{t=T_0+1}^{T_0+T_1} \mathbb{I}\left\{Y^{(t)} \in \widehat{\mathcal{C}}_\alpha^t\left(X^{(t)}\right)\right\} \approx 1-\alpha. \tag{2.2}$$

Restoring Approximate Exchangeability

The first strand of research seeks to restore approximate-exchangeability through careful temporal calibration. J. Lei et al. (2018) pioneered this approach by applying split conformal prediction with temporal data partitioning, where the calibration set consists of recent observations, maintaining distributional similarity with the test period under local stationarity assumptions, resulting in Online Sequential Split Conformal Prediction (OSSCP). While this method initially lacked theoretical justifications, Oliveira et al. (2024) recently provided refined analysis showing under strict stationarity, the coverage deficit decreases at a rate of $O(\frac{1}{\sqrt{n}})$ for a calibration window size n.

This framework spawned several extensions. Several applied papers developed rolling window approaches that continuously update calibration sets (See Zaffran et al. (2022), Wisniewski et al. (2020), and Kath and Ziel (2021)), while Tibshirani et al. (2019) proposed sequential methods with exponential down-weighting of historical observation. However, both require careful tuning of hyperparmeters and assume local exchangeability that may fail catastrophically during regime changes, a limitation that has proven difficult to address systematically.

A popular related strand of work uses ensemble procedures to improve robustness under temporal dependence. Jackknife+ employs leave-one-out residuals and assumes exchangeable data for valid coverage (Barber et al., 2021), while its extension, NexCP, calibrates weights via cross-validation to handle the non-exchangeable settings (Barber et al., 2023). EnbPI (Xu & Xie, 2021) trains ensembles on bootstrap samples or rolling windows to address temporal structure directly. The underlying intuition is that ensemble diversity may mitigate dependence effects. These methods accommodate model uncertainty and integrate with black-box forecasting algorithms, but their coverage guarantees under strong dependence remain heuristic.

Beyond Exchangeability Through Adaptation

Recognising the fundamental limitations of restoring exchangeability, the second research direction explicitly abandons this assumption in favour of online learning mechanisms. The seminal ACI algorithm of Gibbs and Candes (2021) reformulates coverage control as an online optimisation problem, updating miscoverage rates via gradient descent:

$$\alpha_{t+1} = \alpha_t + \gamma\left(\alpha - \mathbb{I}_{y_t \notin \hat{C}_t(x_t)}\right)$$

where α_t is the adaptive α coverage parameter. ACI provides finite-sample bounds on time-averaged miscoverage without distributional assumptions in a realistic online manner, but its performance critically depends on the learning rate, γ, which governs a fundamental trade-off between adaption speed and stability.

Subsequent work sought to address ACI's hyper-parameter sensitivity. Zaffran et al. (2022) developed Online Expert Aggregation ACI (AgACI), maintaining multiple predictors with different learning rates and combining the outputs via exponential weighting. While massively reducing sensitivity, AgACI increases computational complexity by a factor K and introduces meta-parameters for aggregation. Gibbs and Candès (2024) proposed Dynamically Tuned ACI (DtACI), adapting learning rates based on local miscoverage volatility, though this shifts rather than eliminates the tuning burden through second-order adaptation parameters.

Specialised Developments

Recent work has focused on task-specific adaptations of general conformal frameworks. Multi-step forecasting extensions reveal fundamental trade-offs between per-step and joint coverage guarantees (X. Wang & Hyndman, 2024; Sun & Yu, 2023). Change-point detection methods have been developed to handle abrupt regime changes where traditional approaches fail (Vovk et al., 2021). Additionally, specialised techniques for deep learning models address the unique challenges of high-dimensional neural network outputs (Lee et al., 2024a).

2.4 Conformal Prediction in Energy Forecasting

Given the inherent volatility of electricity markets, price forecasting was one among early applications of conformal prediction for time series forecasting, as traditional econometric models fail to adapt to distributional shift and extreme pricing events. Introduced by Kath and Ziel (2021) to short-term electricity price forecasting, they found normalised conformal prediction provided better coverage than the conventional QR methods. More recently, O'Connor et al. (2024) demonstrated that time-series–adapted CP methods (EnbPI and sequential CP) produce precise, reliable intervals, even during extreme re-pricing events.

Researchers subsequently extended these methods to renewable energy forecasting, with applications spanning both solar and wind power prediction often used for testing novel conformal methods (Xu & Xie, 2023, 2021; Lee et al., 2024b). For solar forecasting, Renkema et al. (2024) demonstrated conformal prediction variants (weighted CP, Mondrian binning) consistently deliver well-calibrated prediction intervals, achieving approximately 14% improvement in interval over QR baselines.

Wind energy has seen an extensive application of conformal prediction techniques. For short-term wind speed forecasting, Althoff et al. (2023) applied Conformal Predictive Distribution Systems (CPDS) and NexCP to MET Norway forecasts, achieving narrower prediction intervals and valid coverage compared to QR forests. Zuege et al. (2025) further enhanced this using a hybrid Light Gradient Boosting Machine (LGBM) with EnbPI, achieving valid coverage intervals for wind speed forecasts across different geographical regions including Germany and Brazil.

Using turbine-level data, Jonkers et al. (2024) combined convolutional neural networks (CNN) with Split-CPDS for day-ahead regional wind farm forecasting, incorporating spatial farm structure to capture inter-turbine dependencies and achieving improved coverage validity compared to state-of-the-art methods.

Noteably underdeveloped from the literature is the application of ACI methods to wind energy forecasting, despite its online nature being a natural fit, and its ability to deal with distributional shift. This work contributes to the literature in two ways: 1) the novel application of adaptive conformal inference methods to wind energy; 2) the combination of both global NWP and turbine level predictions.

3 Preliminary Material

In this chapter we develop the basic background of wind energy, online prediction and the mathematical problem formulation readily used throughout the thesis.

3.1 Wind Energy Background

NWP models form the backbone of traditional wind and weather forecasting, solving atmospheric dynamics through discretised versions of the Navier-Stokes equations at spatial resolutions of 10-25km. However, these forecasts are often too coarse for wind farm or turbine specific conditions and too computationally expensive to run at local resolutions. Recent advances in deep learning has shifted the focus to models such as Google DeepMind's GraphCast (Lam et al., 2023) and Microsoft's Aurora (Bodnar et al., 2025) achieving accuracy comparable or superior to traditional NWP at a fraction of the computational cost. However, these models produce black-box point forecasts without associated uncertainty estimates, limiting their operational value for wind farms.

Wind-to-Power Conversion Physics

The theoretical relationship between wind and the power generated is governed by:

$$P_{wind} = \frac{1}{2}\rho A v^3 C_p, \tag{3.1}$$

where P_{wind} is power output, v is wind speed, C_p is a performance coefficient (with a maximum theoretical value $16/27 \approx 0.593$ by Betz's limit), and A is the rotor swept area equal to $\pi \times$ radius. This cubic dependence on wind speed means that small errors in wind speed forecasts can translate to disproportionately large errors in predicted power output. An example power-curve is given in Figure 3.1.

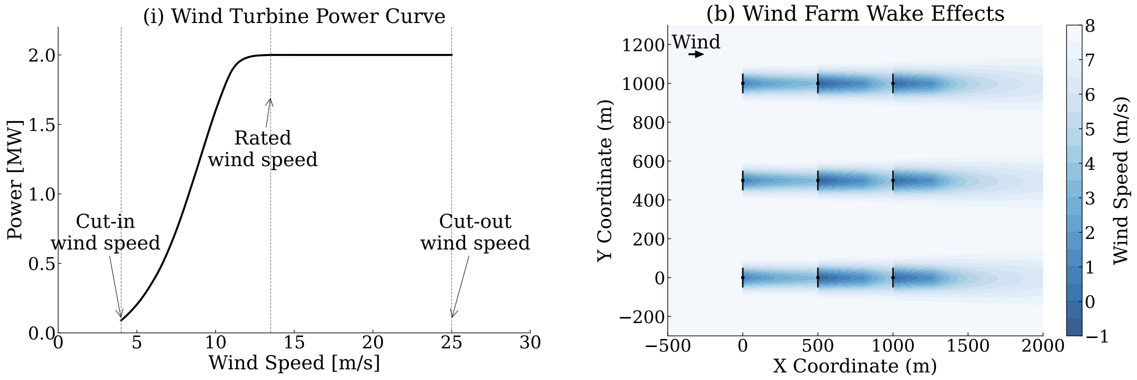


Figure 3.1: (i) Example Power Curve from manufacturer specifications. (ii) Example wake effects in a 3 × 3 wind farm under a westerly wind, the black lines representing turbines from aerial wind farm view.

In practice, the manufacturer-specified power curve is derived under standardised test conditions (see International Electrotechnical Commission, 2005), while operational turbines are affected by gradual mechanical wear, terrain, turbulence and even wake effects. Wake effects arise when upstream turbines reduce both wind speed and stability downstream, ultimately influencing performance and even farm layout. Over time, such factors shift the effective power curve, and are manifested through complex spatio-temporal statistical dynamics. Supervisory Control and Data Acquisition (SCADA) systems record high-frequency (10-minute) operational measurements, providing the detailed turbine-level data required to capture these dynamics.

3.2 Online Prediction Framework

Operational wind power forecasting is inherently sequential. Meteorological and SCADA data arrive continuously, and require immediate forecast updates, often without retraining often referred to as *online*. Grid operators require forecasts across multiple horizons (Table 3.1), we focus on short-term predictions due to its importance for wind farm operations.

Classification	**Horizon**	**Application**
Ultra-short-term	$<$ 30 min	Real-time grid scheduling and wind farm control.
Short-term	30 min to several hours	Day-ahead market operations and daily grid scheduling.
Mid-term	Days to months	Medium-term market transactions and maintenance planning.
Long-term	Years	Wind farm site selection and long-term grid planning.

Table 3.1: Classification of wind power prediction horizons (Y. Wang et al., 2021).

The data stream in this setting is non-exchangeable, with temporal autocorrelation from atmospheric persistence, cyclical patterns due to seasonality, spatial dependencies through wake effects and even abrupt weather regime shifts or turbine degradation. These evolving statistical properties over time result in distributional shift, requiring both the forecasting model and the conformal method to adapt online. These are demonstrated by Figure 3.2.

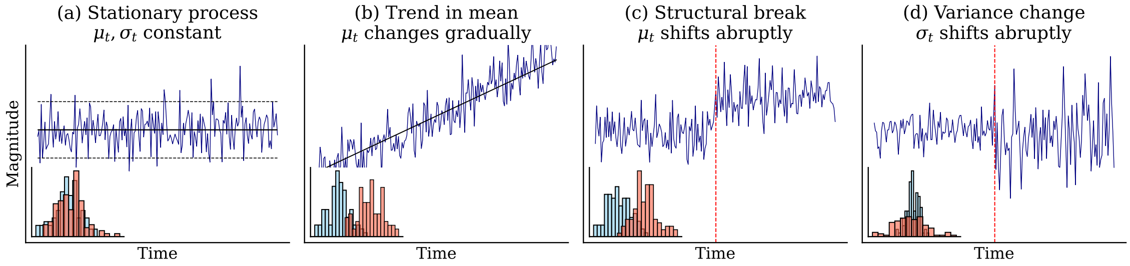


Figure 3.2: Example stationary distribution (a), and common violations via distributional shift (b,c,d). Where μ_t, σ_t represent the mean and variance at time t.

3.3 Mathematical Problem Statement

We consider the online forecasting problem for wind farm power generation. Let (x_t, y_t) denote the process observed sequentially over time $t \in \mathbb{N}$ where $x_t \in \mathbb{R}^d$ comprises both the NWP-derived features, SCADA turbine-level measurements available at time t, and fixed spatial meta-data, and where $y_t \in \mathbb{R}^1$ the realised aggregate wind farm power output in kilowatts (kW). This feature vector admits a natural decomposition as $x_t = (x_t^{(\mathrm{NWP})}, x_t^{(\mathrm{SCADA})}, x^{(\mathrm{SPATIAL})})$.

The stochastic sequence $\{(x_t, y_t)\}_{t=1}^{\infty}$ is inherently non-exchangeable. Dependence arising from the mechanisms characteristic of wind power production, temporal dependence from wind persistence implies $\mathrm{Cov}(y_t, y_{t+h}) \neq 0$ for horizon $h > 0$, non-stationarity and spatial correlation between turbines resulting in multivariate dependence even after farm-level aggregation. Consequently, the standard exchangeability assumption of conformal prediction is violated necessitating new inferential tools.

Given the historical information $\mathcal{F}_{t-1} = \sigma((x_s, y_s)_{s=1}^{t-1})$ a measurable predictor $\hat{f}_t : \mathbb{R}^d \to \mathbb{R}$ produces a point estimate $\hat{y}_t = \hat{f}_t(x_t)$, where $\hat{f}_t$ may be updated adaptively on a learning schedule using recent data to account for distributional shifts. The prediction errors $R_t = |y_t - \hat{y}_t|$ form a stochastic process whose distribution is both time varying and serially dependent. The objective is to construct, for a fixed, predefined miscoverage rate $\alpha \in (0, 1)$ and for each time t, a prediction interval $\hat{C}_t(x_t) \subset \mathbb{R}$ of the form

$$\hat{C}_t(x_t) = [\hat{y}_t - \hat{q}_t,\ \hat{y}_t + \hat{q}_t] \ ,$$

where $\hat{q}_t$ is an estimated quantile of the conditional residual distribution at time t, such that the long-run empirical coverage probability satisfies

$$\lim_{T\to\infty} \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \mathbb{I}\{y_t \in \hat{C}_t(x_t)\} \geq 1 - \alpha \quad \text{a.s.}$$

This relaxes the finite-sample validity of classic conformal prediction to an asymptotic coverage guarantee that expands to arbitrary distributional shift resulting in non-exchangeability. Conditional on this validity, we further require the interval to be efficient by minimising

$$\mathbb{E}[\mathrm{width}(\hat{C}_t(x_t))] = 2\mathbb{E}[\hat{q}_t] \ ,$$

since overly conservative intervals, while valid, have limited operational value for grid management.

Formally, the mathematical problem is to design an adaptive conformal inference procedure that, when applied to forecasts $\hat{y}_t = \hat{f}_t(x_t)$ derived from NWP-SCADA features representative of state-of-the-art spatio-temporal modelling complexity, produces a sequence of prediction intervals $\hat{C}_t(x_t)$ that achieve $1-\alpha$ coverage, minimising expected interval width and remaining implementable under an operational, online forecasting system.

4 Methodology

In this section we introduce Adaptive Conformal Inference (ACI) and its extensions, along with our forecasting model within our framework. We consider an online framework where we observe a sequence input-output pairs $(x_t, y_t) \in \mathbb{R}^d \times \mathbb{R}$. After seeing an initial batch of T_0 observations, our goal is to predict y_t and construct a valid prediction interval $\hat{C}_t(x_t)$ for each new input x_t, using only the data observed up to time $t-1$.

4.1 Adaptive Conformal Inference (ACI)

As established in section 2.2, standard CP constructs prediction sets with finite-sample coverage guarantees under exchangeability assumptions, using the standard nonconformity scores, $S_t = |y_t - \hat{y}_t|$, to form prediction intervals via empirical quantiles. However, wind power time series violates this assumption due to autocorrelation and non-stationarity. Under such violations and a result of distributional shift, standard CP may yield miscalibrated intervals with poor transient behaviour. Furthermore, beyond usual forecasting difficulties, energy prediction is an inherently variable, online task with varying sequential data. This motivates the use of an adaptive method that updates the prediction sets in real time.

Adaptive Conformal Inference (ACI), proposed by Gibbs and Candes (2021), reformulates the coverage as an online optimisation task. Rather than maintaining a fixed α, ACI dynamically adjusts the quantile level α_t through a simple update rule which tracks the target miscoverage rate:

$$\alpha_{t+1} = \alpha_t + \gamma(\alpha - \mathbb{I}_{yt \notin \hat{C}_t(x_t)}) \tag{4.1}$$

where $\gamma > 0$ is the learning rate governing adaption speed, and the $\mathbb{I}_{yt \notin \hat{C}_t(x_t)}$ indicator flag for miscoverage. Simply, when the true value y_t falls outside the prediction set $\hat{C}_t(x_t)$, the algorithm increases α_t^* to widen future intervals, and conversely tightens them when coverage is excessive. The full algorithm is given in Appendix 1.

Implicit Assumptions and Limitations

While ACI removes the explicit statistical assumption of exchangeability, it introduces a crucial implicit operational assumption - that the dynamics of the distribution are *trackable*. The online learning structure means performance is governed by the speed and nature of distributional shift relative to its fixed learning rate (γ). This assumption can lead to failure in several scenarios, from abrupt shifts which cause a sudden change in non-conformity scores to fast oscillations in the distributional shift, resulting in unstable calibration as the algorithm fails to quickly adapt. This exposes a fundamental trade-off whereby larger learning rates allow faster adaptation to distributional shifts, but can also lead to greater instability in the quantile level α_t, potentially resulting in over- or under-coverage introducing interval inefficiency. Conversely, smaller γ-values promote stability but adapt more slowly to changes.

For highly non-stationary series, such as wind energy, this trade-off becomes critical, making fixed γ ACI suboptimal in practice.

Theoretical Guarantees

ACI provides asymptotic validity guarantees that hold regardless of the underlying data distribution. Specifically, this method ensures the long-term average miscoverage rate converges to the target level:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{I}_{y_t \notin \hat{C}_t(x_t)} \to \alpha \quad \text{as} \quad T \to \infty \quad a.s. \tag{4.2}$$

This property holds regardless of exchangeability assumptions. However, this convergence rate is fundamentally governed by the learning rate, γ, leading to the bound (Zaffran et al., 2022):

$$\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{I}_{y_t \notin \hat{C}_t(x_t)} - \alpha\right| \leq \frac{2}{\gamma T} \tag{4.3}$$

While Gibbs and Candes (2021) establish asymptotic coverage guarantees for fixed learning rates, they do not provide explicit guidance on optimal γ schedules. To improve convergence, Zaffran et al. (2022) propose a decaying schedule $\gamma_t = t^{-1/2}$, achieving $O(T^{-1/2})$ rates, though this adapts slowly to abrupt shifts. To overcome these limitations, several extensions have been developed, including ensemble-based methods like AgACI and dynamically-tuned algorithms like DtACI.

Online Expert Aggregation ACI (AgACI)

AgACI, introduced by Zaffran et al. (2022), enhances ACI by running parallel instances (experts) with a grid of learning rates, then aggregating their outputs to form robust prediction intervals. This mitigates the risk of suboptimal fixed-γ choices. For each time step, upper and lower quantiles are aggregated separately using a variant of Bernstein Online Aggregation (BOA) (see Wintenberger (2017)), an online experts algorithm that weights contributions based on cumulative performance. These are then updated to minimise regret, favouring experts that historically perform well under distributional shifts.

Performance is measured using pinball loss, a standard loss function for evaluating quantile forecasts. For a target quantile $\tau \in (0,1)$, and a prediction error $\epsilon = Y - \hat{q}_\tau$, the pinball loss is:

$$\rho_\tau(\epsilon) = \begin{cases} \tau\epsilon & \text{if } \epsilon \geq 0 \\ (1-\tau)(-\epsilon) & \text{if } \epsilon < 0 \end{cases} \tag{4.4}$$

This asymmetric loss function allows for penalisation of under-prediction and over-prediction differently, making it ideal for tracking specific quantiles. This is demonstrated by Figure 4.1 and the full algorithm is given in Appendix 2.

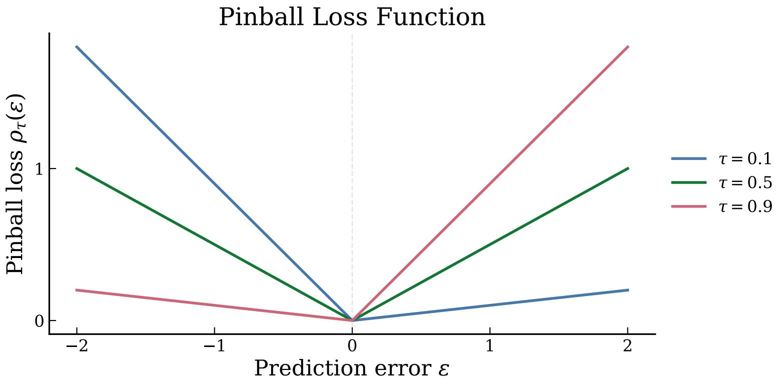


Figure 4.1: Example of Pinball Loss

Theoretical Guarantees

Unlike ACI, AgACI does not provide formal theoretical coverage guarantees due to its separate treatment of upper and lower prediction bounds. As such, the method is best viewed as an empirical improvement over fixed-γ ACI, offering practical robustness without formal statistical guarantees.

Dynamically Tuned ACI (DtACI)

DtACI, developed by the original method authors Gibbs and Candès (2024), builds on ACI by introducing a meta-learner that dynamically selects the best γ from a candidate set at each step, rather than aggregating outputs. This switching uses an online exponential re-weighted scheme to track which γ minimises cumulative miscoverage regret, enabling seamless transitions during abrupt changes.

At each time step, DtACI selects the expert i with probability $p_t^i = w_t^i / \sum_{j=1}^k w_t^j$, making the chosen expert sovereign for that prediction. Similarly to ACI this uses a pinball loss. However, unlike AgACI's aggregation approach, DtACI ensures that the prediction set $\hat{C}_t(x_t)$ at any time is generated by a single ACI algorithm for both upper and lower bounds, preserving formal theoretical guarantees. The full algorithm is given in Appendix 3.

Theoretical Guarantees

DtACI provides theoretical guarantees finite-sample dynamic regret bounds that control performance over any local interval (Gibbs & Candès, 2024). The key result bounds the algorithm's performance relative to the best expert, or learning rate, in hindsight over any interval $I = [r, s]$.

Assuming a grid of learning rates $\gamma_1 < \cdots < \gamma_k$ satisfying $\frac{\gamma_{i+1}}{\gamma_i} \leq 2$ and $\gamma_k \geq \sqrt{1 + \frac{1}{|I|}}$, DtACI satisfies the regret bound:

$$\frac{1}{|I|}\sum_{t=r}^{s} \mathbb{E}[\ell(\beta_t, \alpha_t)] - \frac{1}{|I|}\sum_{t=r}^{s} \ell(\beta_t, \alpha_t^*) = O\left(\sqrt{\frac{\log |I|}{|I|}}\right) + O\left(\sqrt{\frac{V_I}{|I|}}\right), \tag{4.5}$$

where ℓ is the pinball loss and $V_I = \sum_{t=r+1}^{s} |\alpha_t^* - \alpha_{t-1}^*|$ quantifies the variation in the optimal threshold over the interval. Therefore, this algorithm selects approximately the best possible rate with hindsight of any given period I, a powerful property which ensures that DtACI adapts to local distribution shifts.

Under a mild regularity assumption, whereby the conformity score β_t has a density bounded below by $p > 0$, this regret bound controls the squared error in the adaptive quantile level (α^*). As:

$$\frac{1}{|I|}\sum_{t=r}^{s}\frac{p\mathbb{E}[(\alpha_t - \alpha_t^*)^2]}{2} = O\left(\sqrt{\frac{\log|I|}{|I|}} + \sqrt{\frac{V_I}{|I|}}\right). \tag{4.6}$$

Therefore, when the algorithm's parameters decay appropriately over time, DtACI recovers the same asymptotic guarantee as ACI in Equation (4.2). Together showing, DtACI preserves the distribution-free validity of ACI while offering sharper finite-sample guarantees in dynamic, non-stationary series such as wind.

4.2 Point Forecasting Models

To provide a comprehensive evaluation of the conformal methods, we implement three distinct forecasting models: a classical statistical baseline (ARIMA), a widely applied tree-based probabilistic model (LightGBM with quantile regression), and a state-of-the-art deep learning model designed for spatio-temporal problems (STGCN). Each model is chosen to represent different paradigms for wind power forecasting.

Auto-Regressive Integrated Moving Average (ARIMA)

The ARIMA model is the quintessential statistical forecasting model. This univariate model captures both autocorrelation in the series (AR), and its errors (MA) in a stationary series. This is given in full by:

$$\hat{y}_t^d = c + \phi_1 y_{t-1}^d + \cdots + \phi_p y_{t-p}^d + \theta_1\epsilon_{t-1}\cdots + \theta_q\epsilon_{t-q} + \epsilon_t \tag{4.7}$$

where p, q represents the order of the AR and MA respectively, and d the degree of differencing required for stationarity. While traditionally p is selected via autocorrelation function (ACF), q using partial ACF (PACF) and d using a stationarity test, commonly Augmented Dickey-Fuller (ADF). To use in an online setting we instead automate the selection of its parameters, testing a grid of parameters, selecting the best model using Akaike Information Criterion (AIC). This is a standard approach across the literature and industry, therefore we use this baseline model to benchmark our point-forecasts.

Probabilistic LightGBM with Quantile Regression

To generate and demonstrate a robust probabilistic baseline, we employ a Light Gradient Boosting Machine (LGBM) an efficient, and widely utilised implementation of gradient-boosted decision trees synonymous with tabular data within a quantile regression. The core of this approach is gradient boosting, an ensemble technique which stacks weakly learning decision trees to correct errors of preceding models. For a given differentiable loss function L_i

$$r_{im} = -\left[\frac{\partial L(y_i, F(x))}{\partial F(x)}\right]_{F(x)=F_{m-1}(x)}$$

The model is then updated as $F_m(x) = F_{m-1}(x) + \nu h_m(x)$ where ν is the learning rate.

To achieve probabilistic forecasts, we go a step further, replacing the mean squared error loss with the pinball loss function described in Figure 4.1, enabling a quantile regression. Here the gradient of the pinball loss with respect to the prediction $\hat{y}_\tau$ is the step function. Allowing a new tree to be trained at each boosting step, effectively pushing the prediction towards the true τ-th quantile of the conditional distribution. For a given quantile, and horizons, we train separate models with 6 distinct models trained at each refit. While computationally intensive, it is a common approach used across industry to prevent error-accumulation which plague recursive approaches into the future. However, as the LGBM does not inherently deal with time series, a rich set of features with lags and rolling windows is constructed.

Spatio-Temporal Graph Convolutional Network (STGCN)

While forecasting wind power is largely dominated with simpler statistical models or light-weight machine learning alternatives, computational developments have expanded scope for sophisticated deep learning architectures as operators try squeeze performance out of models which better represent the data. One such popular enhancement is the Spatio-Temporal Graph Convolutional Network (STGCN), allowing for spatio-temporal representations of the wind farm.

Graph Convolutional Network (GCN)

The Graph Convolutional Network (GCN), introduced by Kipf (2016), builds on the convolutional approach ubiquitous to the Convolutional Neural Network (CNN) (LeCun et al., 1989) to non-Euclidean domains - as demonstrated in Figure 4.2. While CNNs operate on regular grid-structured data, GCNs generalise convolution to graph-structured data whereby learned filters are applied across non-Euclidean domains. This allows the application to arbitrary spatial components, such as wind farms.

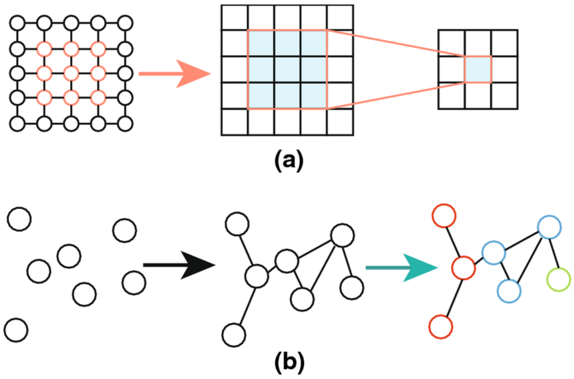


Figure 4.2: (a) Euclidean CNN vs (b) Non-Euclidean GCN structure

In a GCN, we represent the wind farm spatial structure as an undirected graph,

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}) \ ,$$

where the set of nodes, $\mathcal{V}$, represents the individual turbines, and the set of edges $\mathcal{E}$ encodes the spatial relationship between them. This takes as an input a feature matrix $X \in \mathbb{R}^{N\times D}$ and a representative spatial description typically in the form of an adjacency matrix $A \in \mathbb{R}^{N\times N}$. The output is node level $Z \in \mathbb{R}^{N\times F}$.

The adjacency matrix is constructed via a k-nearest neighbour approach with Gaussian kernel weighting. For turbine coordinates, pairwise Euclidean distances D_{ij} are computed and the edge weights defined as

$$W_{ij} = \exp\left(-\frac{D_{ij}^2}{2\sigma^2}\right) \tag{4.8}$$

Where σ is the bandwidth parameter set to the median of non-zero pairwise distances to ensure an adequately adaptive scale. We fix $k = 2$, to balance local connectivity and computational feasibility. A smaller k can discard relevant spatial dependencies, while larger k can enforce spurious relationships and is computationally demanding. The normalised adjacency matrix is defined as:

$$\tilde{A} = (D+I)^{-\frac{1}{2}}(A+I)(D+I)^{-\frac{1}{2}} \tag{4.9}$$

where D is the diagonal degree matrix with $D_{ii} = \sum_{i=1}^{n} \tilde{a}_{ij}$, and I the identity, allowing for numerical stability and preventing feature magnitudes from exploding during propagation. Every GCN layer can be written as a non-linear function:

$$H^{(\ell+1)} = GCN(H^{(\ell)}) = \sigma[\tilde{A}H^{(\ell)}W^{(\ell)}]\,, \tag{4.10}$$

where $H^{(0)} = X$, $W^{(\ell)}$ is the trainable weight matrix, and $\sigma(\cdot)$ the non-linear activation.

Temporal Convolutional Network (TCN)

While traditional approaches to temporal modelling, such as Recurrent Neural Network (RNN)s process data sequentially, this process is slow and difficult to learn long range dependence. The Temporal Convolutional Network (TCN), formalised later by Bai et al. (2018), apply 1D convolutions across the time axis, allowing it to be easily parallelised, and capture more intricate temporal structures.

The architecture relies on two key principles. First, causal convolutions ensure the output at time t depends only on inputs from the past $(x_t, x_{t-1}, x_{t-2}\ldots)$ preventing information leakage from from future time steps. Secondly, dilated convolutions efficiently capture the data's long-term dependencies, allowing it to take sequences of any length. For a 1D sequence, $x \in \mathbb{R}^n$, the TCN operation F with filter f of length k the dilated convolution calculates:

$$F(x)_t = \sum_{i=0}^{k-1} f(i)\cdot x_{t-d\cdot i}\ ,$$

here d the dilation factor, while $x_{t-d\cdot i}$ accounts for the direction into the past. By stacking layers with exponentially increasing dilation factors ($d = 2\ell$ for layer ℓ), growing the receptive field to efficiently model long-range temporal patterns (seasonality, wind persistence etc.). Finally, the residual connections within each block facilitate stable gradient flows within this deep architecture.

STGCN Hybrid

Introduced by Yu et al. (2017), STGCN merges both the spatial (GCN) and temporal (TCN) in a increasingly popular hybrid architecture. As shown in Figure 4.3 this model is comprised of two Spatio-Temporal Convolution Blocks. Each block integrates these components through a *sandwich* structure of temporal and spatial graph convolutions. Effectively, this structure allows the model to learn patterns across both space and time simultaneously. Here we use the specification from the original paper.

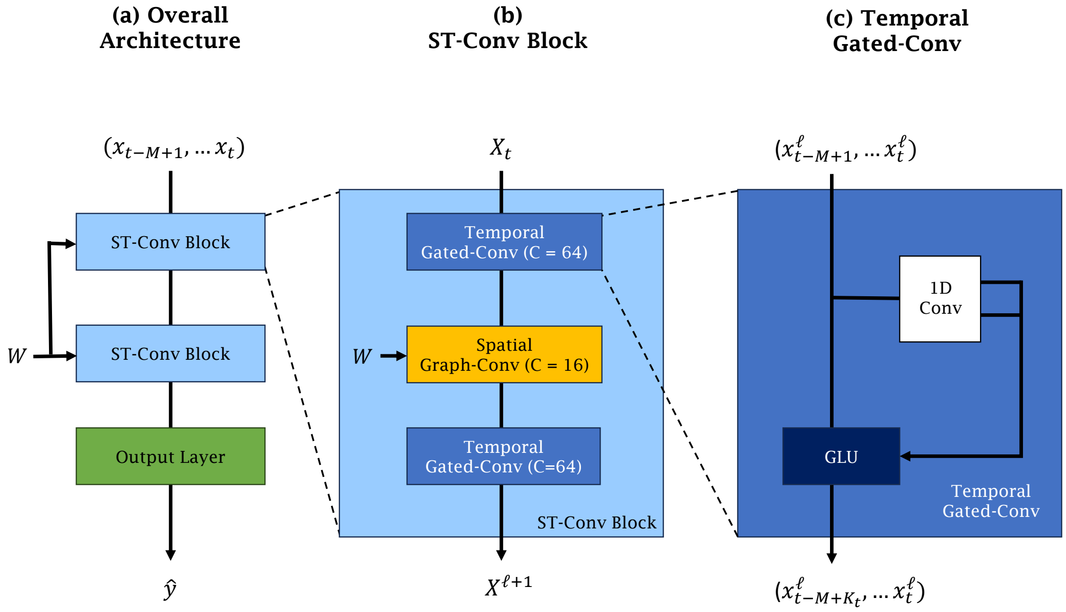


Figure 4.3: (a) overall STGCN model architecture; (b) details of core Spatio-Temporal Convolution Blocks; (c) the temporal unit (TCN) with M look-back window (Yu et al., 2017).

4.3 Evaluation

To properly understand our forecasting model and CP performance we adopt the following evaluation methods.

Point Estimates

We assess the accuracy of our point estimates with two standard metrics. Root Mean Squared Error (RMSE), given by Equation (4.11), is popular as it heavily punishes large deviations, however this results in sensitivity to outliers, especially in asymmetric series.

$$RMSE(\hat{y}, y) = \sqrt{\frac{1}{n}(y - \hat{y})^2} \tag{4.11}$$

For this reason we also use Mean Absolute Error (MAE) given by Equation (4.12), more robust to outliers and helpful for skewed series, like wind power, due to its focus on median errors.

$$MAE(\hat{y}, y) = \frac{1}{n}|y - \hat{y}| \tag{4.12}$$

Uncertainty Quantification

Prediction interval coverage probability (PICP), or mean coverage, details the proportion of observations contained within the prediction intervals, while mean interval width (MW) quantifies the efficiency or sharpness of these intervals:

$$\text{PICP} = \frac{1}{n}\sum_{t=1}^{n} \mathbb{I}\left\{Y^{(t)} \in \widehat{\mathcal{C}}_t\left(x_t\right)\right\} \qquad \text{MW}(\hat{C}_t(x_t)) = \frac{1}{n}\sum_{t=1}^{n}\left(\hat{u}_t - \hat{\ell}_t\right) \tag{4.13}$$

where $\hat{u}_t$ and $\hat{\ell}_t$ are the upper and lower bounds of $\hat{C}_t(x_t)$. Coverage is of absolute importance and directly measures the validity of Theorem 1, whereby valid conformal prediction should maintain coverage close to the nominal level $1 - \alpha$. However, interval efficiency or sharpness is equally crucial, as a model can always achieve strong coverage with infinitely wide intervals. The mean coverage fails to describe how far off the violations are, therefore it is useful to also observe the Average Coverage Error (ACE) - especially in high stakes energy forecasting.

The Winkler Score (WS) is a popular scoring method to jointly evaluate both coverage and sharpness:

$$\text{WS}(\hat{\ell}_t, \hat{u}_t, y_t) = (\hat{u}_t - \hat{\ell}_t) + \frac{2}{\alpha}(\hat{\ell}_t - y_t)\mathbb{I}\{y_t < \hat{\ell}_t\} + \frac{2}{\alpha}(y_t - \hat{u}_t)\mathbb{I}\{y_t > \hat{u}_t\} \tag{4.14}$$

This score combines a simple width penalty with a miscoverage penalty proportionate to how far the observation lies from the bound, scaled by $\frac{2}{\alpha}$. This dual penalty structure enforces dual mandate of narrow intervals and valid coverage, and is therefore indicative a more skilful forecast.

5 Synthetic Experiments

To better characterise the behaviour of each ACI approach, we test these methods on synthetic data with a known data generating process (DGP), allowing for rigorous performance assessment under controlled conditions (Morris et al., 2019). Our analysis centers on three realistic sources of non-exchangeability typical in forecasting wind power: temporal dependence, abrupt regime changes, and spatial correlations.

5.1 Scenarios

Each experiment consisted of 500 independent simulations, each generating 800 observations. The first half of each sequence served as a warmup set, while the remaining 50% was used to evaluate online conformal prediction performance. The forecasting model used is a Exponential Moving Average (EMA) with noise:

$$\hat{y}_t = \lambda y_{t-1} + (1-\lambda) y_{t-1} + \epsilon_t, \tag{5.1}$$

where λ is the smoothing factor and ϵ_t represents the noise. The additional stochastic component helps to better replicate an imperfect prediction model, with realistically variable point estimates to better evaluate the conformal methods. We employ this simple model to isolate the performance of the conformal methods under clear model misspecification.

Performance is assessed via empirical coverage and median interval width. Three ACI variants are evaluated: standard ACI with $\gamma \in 0.01, 0.05$, AgACI using a grid containing from 0.001 to 0.5 in 0.01 steps, and DtACI with the same grid and memory parameter $I = 50$. As in Zaffran et al. (2022), we benchmark versus OSSCP, a natural adaptation of SCP to an online setting via rolling windows. All methods target 90% coverage ($\alpha = 0.1$) using the standard absolute residual non-conformity scores and standardised alternatives. All experiments were parallelised across a 64-core AMD EPYC 7702 with 32GB RAM.

Temporal Dependence

Temporal dependence is one of the core motivation of ACI methods, due to its violation of exchangeability. However, we investigate how well this truly works in practice. We do this by testing 3 standard 1-lag DGPs; Auto Regressive (AR), Moving Average (MA), and Auto-Regressive Moving Average (ARMA). These are defined as:

$$\begin{aligned} \text{AR(1)}: &\quad y_t = \phi y_{t-1} + \epsilon_t \\ \text{MA(1)}: &\quad y_t = \epsilon_t + \theta \epsilon_{t-1} \\ \text{ARMA(1,1)}: &\quad y_t = \phi y_{t-1} + \epsilon_t + \theta \epsilon_{t-1} \end{aligned} \tag{5.2}$$

where $\epsilon_t \sim$ i.i.d. $\mathcal{N}(0, \sigma^2)$ denotes white noise. We vary the dependence parameters $\phi, \theta \in [0.1, 0.5, 0.8, 0.9, 0.95]$ to examine performance under increasing temporality.

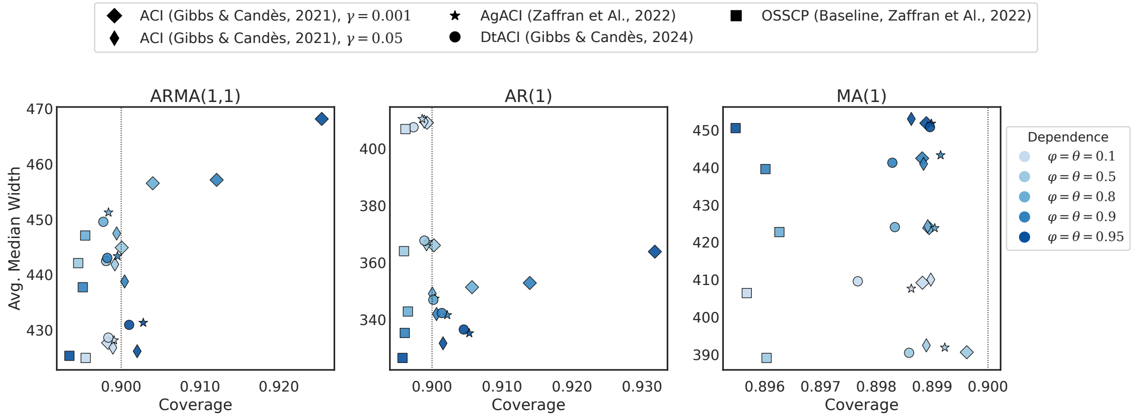


Figure 5.1: Comparison of coverage and sharpness (width) of ACI methods under increasing temporal dependence (ϕ, θ) across several DGPs. 90% target coverage.

Regime Changes

While temporal dependence is a major complication for CP methods, ACI's primary motivation lies in its ability to handle distributional shifts and regime changes. This is particularly relevant to wind power forecasting, where output experiences rapid transitions between weather states, from low production calm regimes to storms.

We evaluate the ACI methods under regime-switching using an AR(1) model with a regime-dependent distribution, mimicking non-stationarity and real world distributional shift as:

$$y_t = \max(0, \phi y_{t-1} + (1-\phi)\mu_{r_t} + \epsilon_t)$$

Where $r_t \in [1, 2, 3]$ represents the current regime (calm, normal, stormy), fixed temporal dependence $\phi = 0.8$, and $\epsilon_t \sim \mathcal{N}(0, \sigma_{r_t}^2)$ with regime-specific variance. The probability of regime shift ($p_{shift} \in [0.005, 0.01, 0.02]$) is a key parameter as it corresponds to the per-step probability of regime change, how often regime changes are expected with durations of 200, 100, 50 time steps respectively. The three regimes are distributed as:

1. **Calm** $(r = 0)$: $\mu_0 = 5, \quad \sigma_0 = 10$ (low mean, low variance)
2. **Normal** $(r = 1)$: $\mu_1 = 500$, $\sigma_1 = 100$ (moderate mean, moderate variance)
3. **Stormy** $(r = 2)$: $\mu_2 = 1000$, $\sigma_2 = 200$ (high mean, high variance)

Spatial Correlation

A key difficulty of forecasting wind farms are the strong correlations across turbines, not only temporally but spatially. Due to wake effects and terrain differences, each turbine generates a different power depending on the location of the wind. We therefore simulate a correlated multi-variate wind-farm, whereby more complicated turbine level correlations are at play. To do so we use the `FLORIS` package (National Renewable Energy Laboratory (NREL), 2025).

The DGP uses Gaussian-Curl-Hybrid wake physics to model the configuration of the wind farm, which allow for physical asymmetry in wake effects comparing to the

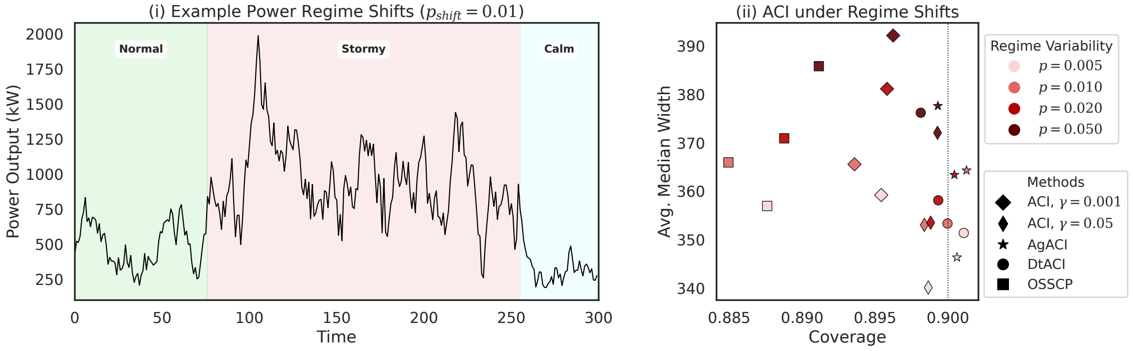


Figure 5.2: (i) Simulated wind farm power from a regime-switching AR(1) process under calm, normal, and stormy. (ii) ACI average empirical coverage and median interval width as a function of regime change frequency (p_{shift}). 90% target coverage.

standard Gaussian model. For simplicity and to show the extent of it's impact we chose a squared 3 × 3 wind farm - with our full configuration in Appendix A.1. Our experiment tests three wind directional scenarios, each testing more complicated wake effects. The first of which tests 4 cardinal directions $[0°, 90°, 180°, 270°]$ and 8 directions containing $[0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°]$. To standardise the effects themselves wind speed is fixed at 8m/s with turbine intensity of 0.06.

Each individual turbine is forecast as Equation (5.1), then aggregated to the farm level for a univariate conformal prediction, capturing these complex multivariate dependencies in a single series. Noise is added to each turbine's base power output allowing us to represent this forecast error aggregation when summing to farm level output.

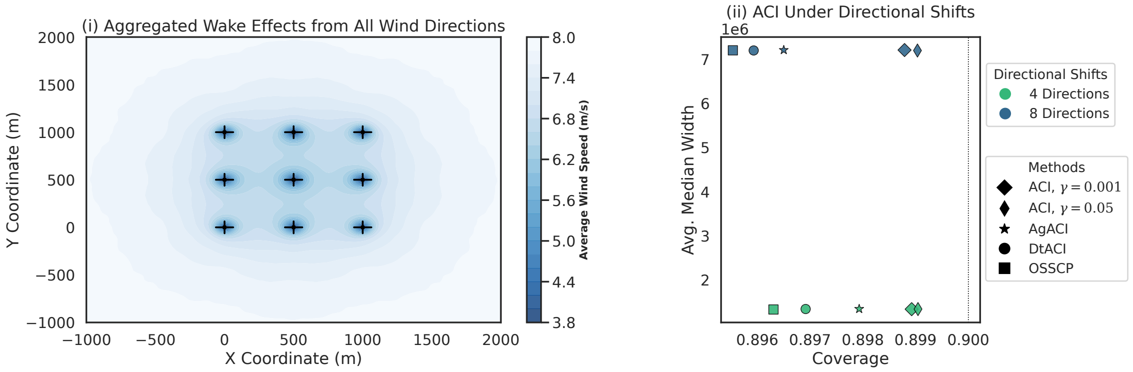


Figure 5.3: (i) Simulated wake effects of a 3 × 3 windfarm from all directions (5° increments). (ii) Average empirical coverage and median interval width of ACI methods under shifting directions & wake effects. 90% coverage target.

5.2 Results

Table 5.1 offers a summary of the most extreme violations of exchangeability, while the full results are in the Appendix.

Table 5.1: Summary of CP method performance under most extreme exchangeability violations ($T = 800, N = 500$). Wind farm width in 10^6. Coverage target of 90%.

Scenario	**Method**	**Coverage**	**Median Width**	**Time (s)**
Autocorrelation (AR(1), $\rho = 0.9$)	ACI (0.001)	0.914 (0.012)	352.9 (12.3)	10.2
	ACI (0.05)	0.901 (0.002)	342.0 (16.8)	10.3
	AgACI	0.902 (0.005)	341.6 (14.9)	115.6
	DtACI	0.901 (0.005)	342.3 (15.1)	41.2
	OSSCP	0.896 (0.010)	335.3 (14.2)	2.7
Regime Switch ($p_{\text{switch}} = 0.02$)	ACI (0.001)	0.896 (0.037)	381.1 (52.9)	10.2
	ACI (0.05)	0.899 (0.008)	353.5 (76.9)	10.5
	AgACI	0.900 (0.013)	363.4 (71.2)	115.6
	DtACI	0.899 (0.010)	358.1 (76.7)	41.0
	OSSCP	0.889 (0.028)	370.9 (69.6)	2.8
Wind Farm (8-directions)	ACI (0.001)	0.899 (0.014)	7.20 (0.07)	21.9
	ACI (0.05)	0.899 (0.003)	7.19 (0.09)	22.1
	AgACI	0.897 (0.007)	7.20 (0.09)	129.0
	DtACI	0.896 (0.006)	7.19 (0.09)	53.2
	OSSCP	0.896 (0.011)	7.20 (0.09)	13.5

Across all synthetic experiments, the ACI variants maintained marginal coverage approximate to the nominal 90% level ($\alpha = 0.1$), validating the theoretical guarantees in Equation 4.2 under controlled exchangeability violations. In contrast, OSSCP baseline systematically under-covered as it failed to adapt to distributional shifts. Under conditions of pure temporal dependence, all ACI methods achieved valid coverage, with interval sharpness improving with autocorrelation strength in AR(1) and ARMA(1,1) processes. This suggests that greater predictability in the series leads to smaller residuals, and more efficient intervals.

The regime-switching experiment offered the clearest performance differentiation, simulating the abrupt distributional shifts characterised by the wind. In these volatile settings, DtACI demonstrated sharper intervals without sacrificing coverage, due to its meta-learned capability, it can dynamically switch to a more aggressive learning rate (γ) under abrupt error distribution shift. Conversely, standard ACI with $\gamma = 0.001$ produced overly conservative intervals, often $15 - 20\%$ wider than other methods. Once again, OSSCP failed to maintain valid coverage. Finally, in our spatial correlation experiment, all ACI methods demonstrate robustness against complex dependencies introduced by the simulated wake effects.

Collectively, these simulation results demonstrate that while any ACI method is preferable to its non-adaptive counterpart, the advanced AgACI and DtACI algorithms provide more robust and efficient performance due to insensitivity to γ-parametrisation, albeit at a significantly higher computational cost.

6 Case Study: Kelmarsh Wind Farm

While the simulations offered insight into several forms of exchangeability violations, it is still a controlled setting. To demonstrate the practical value of ACI for uncertainty quantification, we extend our analysis to a real-world online application to Kelmarsh Wind Farm.

6.1 Data

This study integrates two complementary data sources, firstly high-frequency Supervisory Control and Data Acquisition (SCADA) measurements from Kelmarsh wind farm in Northamptonshire, UK, and numerical weather prediction (NWP) outputs from the National Centre for Environmental Prediction's (NCEP) Global Forecast System (GFS). Together these provide a multivariate, spatio-temporal dataset which spans January 2022 until January 2024 at hourly resolution.

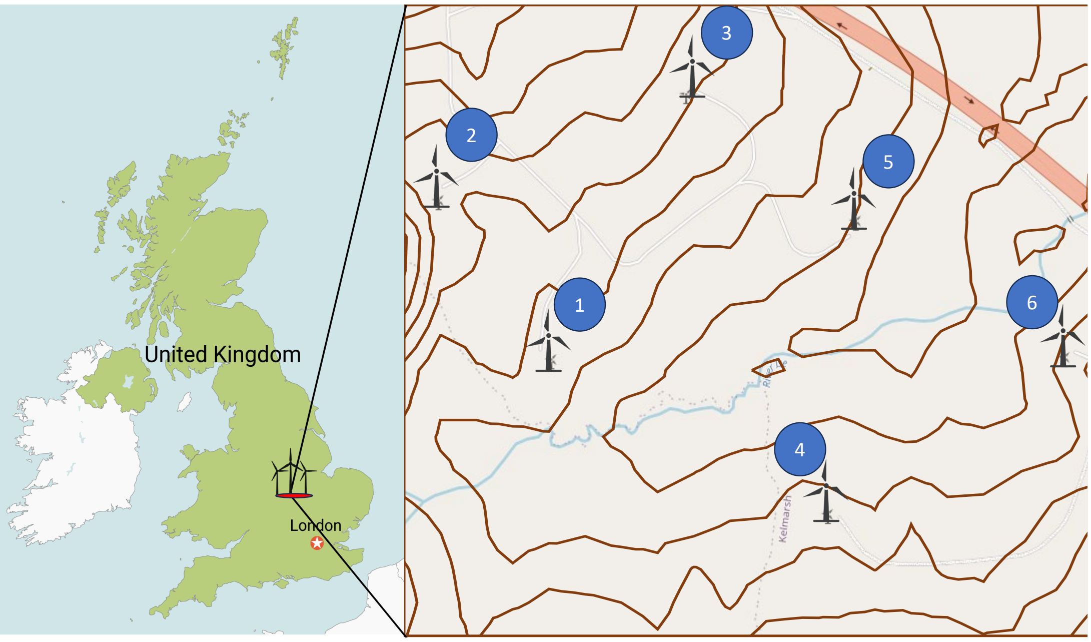


Figure 6.1: Contour map of Kelmarsh wind farm. Each contour line represents 5 m elevation gain. Turbine markers (T01-T06) highlight the relative spatial layout.

Kelmarsh SCADA

This open source dataset, provided by Plumley (2022), collected from 6 Senvion MM92 turbines with combined rated capacity of 12.3 MW (12,300 kW). Measurements were originally at a 10-minute resolution, with each record containing the distributional statistics (minimum, maximum, mean, standard deviation) over the interval, furthermore another dataset provides detailed logs of each turbines maintenance, curtailment and even technical difficulties. Turbine specifications are detailed in Appendix A.7, while a contour map of the farm topography is shown in Figure 6.1, illustrating modest but non-negligible variation in terrain.

The SCADA data provides key, operational turbine level measurements: power output (kW), wind speed (m/s), and wind direction (degrees). Turbine condition measurements encompass ambient conditions at the converter, nacelle temperature, and rotor bearing temperature (°C). Operational parameters include blade pitch angle measurements. Finally, spatial coordinates (latitude, longitude, elevation) are recorded for each turbine location.

NCEP GFS Historical Forecasts

To provide meteorological context, we retrieved historical forecasts from the NCEP Global Forecast System (GFS), accessed via OpenMeteo's historical forecast API (Zippenfenig, 2023). The GFS outputs deterministic forecasts at hourly granularity over a 0.11° latitude-longitude grid (corresponding to $\sim$ 13km resolution). These forecasts are available at 6-hour update cycles. Leveraging this external NWP data demonstrates the flexibility of our approach and ensures the methodology remains wind-farm agnostic and globally applicable.

From here, we extracted several key NWP variables, motivated by their importance in short-term wind power forecasting. Namely surface pressure (hPa), wind speed (m/s), wind direction (degrees), and temperature (°C) all at 80m height, closest to turbine hub height (78.5m). Additional variables include 10m wind gusts (m/s) and precipitation measurements (mm) as in Zjavka (2015).

6.2 Preprocessing

As this is real-world data, it requires rigorous preprocessing. This is necessary to ensure models are trained on data that is representative of operational data while avoiding curtailment, sensor errors or extreme outliers.

Circular Direction Encoding

Important features such as wind direction (predicted & observed) and turbine blade angles are inherently circular, 0° and 360° correspond to the same orientation. To avoid these discontinuities in the model - we apply sine-cosine encoding:

$$\theta_{\sin} = \sin\!\left(\tfrac{\pi}{180}\theta\right), \quad \theta_{\cos} = \cos\!\left(\tfrac{\pi}{180}\theta\right),$$

mapping these angular values into smooth two-dimensional representations on the unit circle. This transformation is common across wind forecasting literature.

Wind/Power Curve

Real world power curves are highly variable, mostly due to the factors discussed in 3.1. However, beyond these normal variations, we must remove abnormalities to evaluate our approach under usual operational conditions. As in Ladopoulou et al. (2025), we took a three step cleaning approach: 1) we matched the SCADA with the logs, removing any events detailing stoppage, warnings or curtailment; 2) we removed any with precisely 0 wind; 3) finally, we removed those that were de-rated - whereby they were at 100% of potential power. This preprocessing process removes $\sim 40\%$ of raw samples. Figure 6.2 compares the empirical power curve before and after cleaning. The cleaned dataset better reflects ordinary, unconstrained operational conditions.

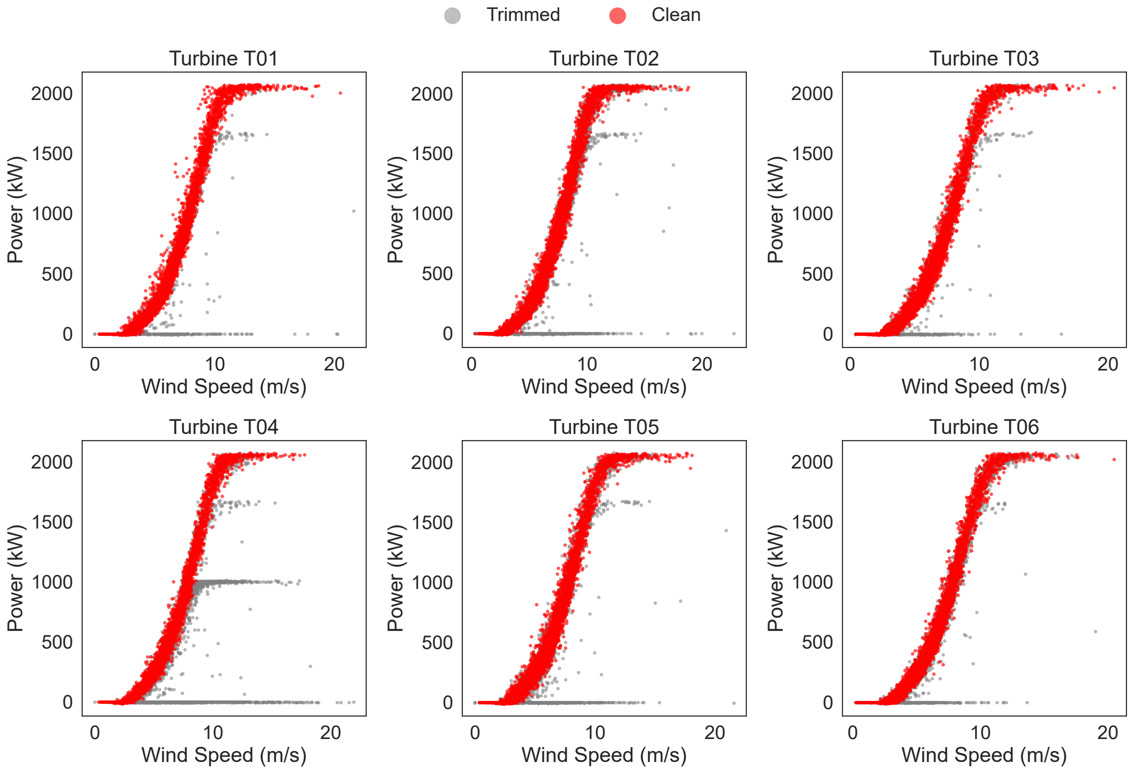


Figure 6.2: Processed Power Curves

Resampling & Missing Values

Although the SCADA data is provided at 10-minute intervals, we upsampled to an hourly resolution by averaging within each hour. This coarser granularity of GFS forecasts removing issues arising from mixed-frequency data fusion, and mimicking the real-time arrival of sequential data. Remaining missing values are imputed using linear interpolation. While hours during which all turbines were inactive are explicitly set to zero farm output to preserve physical consistency. This resulted in the data dimensions in Table 6.1.

Table 6.1: Comparison of raw and cleaned dataset shapes for each turbine.

Turbine	Raw Shape	Cleaned Shape	Reduction (%)
T01	(104,483, 318)	(26,304, 19)	74.82
T02	(104,967, 318)	(26,304, 19)	74.94
T03	(104,754, 318)	(26,304, 19)	74.89
T04	(102,756, 318)	(26,304, 19)	74.40
T05	(104,671, 318)	(26,304, 19)	74.87
T06	(104,561, 318)	(26,304, 19)	74.84

6.3 Modelling

To mimic real-world operational forecasting we adopt an online evaluation scheme characterised by its weekly re-training schedule, and expanding training window. We use the first full year of data (January 2022 - January 2023) for initial model training and hyper-parameter tuning, specifically we used a held out validation set and grid-based selecting using RMSE to select the hyper-parameters (see Appendix A.8). These were then fixed for consistency and before an online forecast evaluation on the subsequent year (January 2023 - January 2024) with weekly model retraining (every 168 hours). At each forecast time t, models are trained/updated using all data up to t and produce predictions for horizon $t + 1, \ldots, t + 6$, aligning with the NCEP GFS forecast.

This protocol reflects real world operational practice, where models are periodically retrained on a schedule, for computational feasibility. This also prevents any data leakage while maximising the evaluation period, similar to traditional cross-validation. Figure 6.3 illustrates this rolling retraining and evaluation process.

To apply our conformal methods, we expand to the multi-step case as in X. Wang and Hyndman (2024), whereby each horizon is grouped into it's own series, running per-step online evaluations. This prevents the difficulty presented with increasingly complex residuals, which result in conservative intervals that systematically undercover at larger steps. We used 6-months of out-of-sample forecasts residuals to calibrate our conformal methods, with the remaining 6-months used for online interval evaluation.

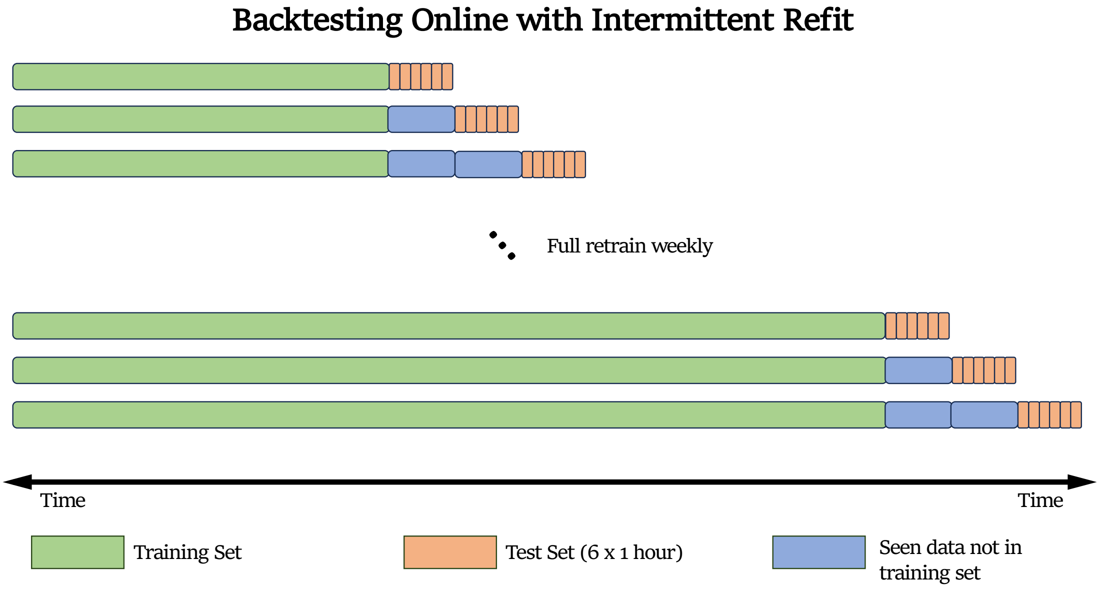

Backtesting Online with Intermittent Refit
Full retrain weekly
Time
Time
Training Set
Test Set (6 x 1 hour)
Seen data not in training set

Figure 6.3: Online evaluation protocol. Models are retrained weekly using an expanding window of past data and then evaluated on the following 6-hour horizon.

6.4 Results

The out-of-sample performance of the point forecasting models is presented in Table 6.2. These evaluation metrics offer a clear hierarchy - we find that hybrid STGCN significantly outperform the statistical ARIMA and LGBM, confirming its ability to capture the complex spatio-temporality inherent to wind farms. This is expected, as both benchmarks were strictly univariate with past lags of only wind farm output. Note, the LGBM-QR is used as a point-forecast by taking the median (or $\tau = 0.5$ pinball loss). As expected, the forecast horizon degrades as the horizon increases, clear by the increasing MAE and RMSE from $\hat{y}_{t+1}$ to $\hat{y}_{t+6}$. Critically and clearly demonstrating model uncertainty in multi-step scenarios.

Table 6.2: Point Forecast Evaluation by Horizons (in Megawatts)

Model	**t+1**		**t+2**		**t+3**		**t+4**		**t+5**		**t+6**	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
ARIMA	2.39	3.35	2.39	3.35	2.41	3.35	2.40	3.31	2.43	3.34	2.43	3.33
LGBM	1.52	2.20	1.60	2.28	1.67	2.36	1.73	2.40	1.77	2.45	1.81	2.51
STGCN	0.85	1.20	1.04	1.43	1.20	1.68	1.29	1.78	1.41	1.93	1.53	2.06

Probabilistic Forecasts

Using the point predictions from the ARIMA and STGCN models, we apply the conformal methods in a multi-step setting, we compare their performance against both a non-adaptive baseline (OSSCP) and a traditional probabilistic model LGBM-QR. The comprehensive results are summarised in Table 6.3.

Table 6.3: Probabilistic forecast evaluation.

Model	Method	Coverage	Width	WS	ACE	Bias Up (%)
STGCN	ACI ($\gamma = 0.001$)	0.897	5330	7392	0.006	62.0
	ACI ($\gamma = 0.05$)	0.897	5240	6695	0.040	59.8
	DtACI	0.893	5137	6699	0.030	57.9
	AgACI	0.891	5112	6725	0.025	58.4
	OSSCP	0.871	4802	7470	0.029	57.8
ARIMA	ACI ($\gamma = 0.001$)	0.874	10566	14540	0.019	41.1
	ACI ($\gamma = 0.05$)	0.894	9690	11089	0.074	66.5
	DtACI	0.891	9964	11673	0.058	68.0
	AgACI	0.891	9896	11740	0.049	64.7
	OSSCP	0.839	10086	14617	0.061	47.7
LGBM-QR	ACI ($\gamma = 0.001$)	0.894	7478	10404	0.006	60.7
	ACI ($\gamma = 0.05$)	0.895	7418	9158	0.044	66.4
	DtACI	0.893	7295	9259	0.038	64.7
	AgACI	0.891	7199	9249	0.033	64.7
	OSSCP	0.863	6767	10373	0.037	60.6
	Quantile Regression	0.695	3767	10959	0.205	54.1

The primary finding supports that the advanced STGCN model, when augmented with conformal uncertainty bands, provides the most skilful predictions, achieving the lowest Winkler Scores. While a well-tuned ACI can perform exceptionally well (e.g $\gamma = 0.05$) the AgACI and DtACI variants offer greater stability, as evidenced from their consistently lower ACE. Furthermore, the LGBM-QR baseline highlights a severe miscoverage of 69.5% reinforcing the premise of this research. Traditional probabilistic models can be dangerously miscalibrated under real-world distributional shift common in wind power. This empirical failure clearly justifies the adaptive approach of ACI, which maintained target coverage even when wrapped around the same underlying residuals.

Finally, a consistent pattern emerged across all models - a bias towards upper miscoverage. This is a result of the use of symmetric distribution intervals for wind power which is physically lower bound at zero. This is an important observation as it suggests that, while ACI provides valid marginal coverage, its intervals can be much more efficient through the development of asymmetric or physically constrained methods. This is further illustrated through the November and December 2023 time-series evaluation snapshot in Appendix A.1, demonstrating physically impossible intervals after abrupt regime changes.

There are two clear takeaways from Figure 6.4. Firstly, it demonstrates the performance and trade-off between coverage and width of our ACI methods. We observe a clear diversion in performance between the simple baseline ARIMA and the STGCN hybrid. While the overarching coverage performance is similar, with adaptive conformal methods providing the approximately the targeted empirical coverage, while

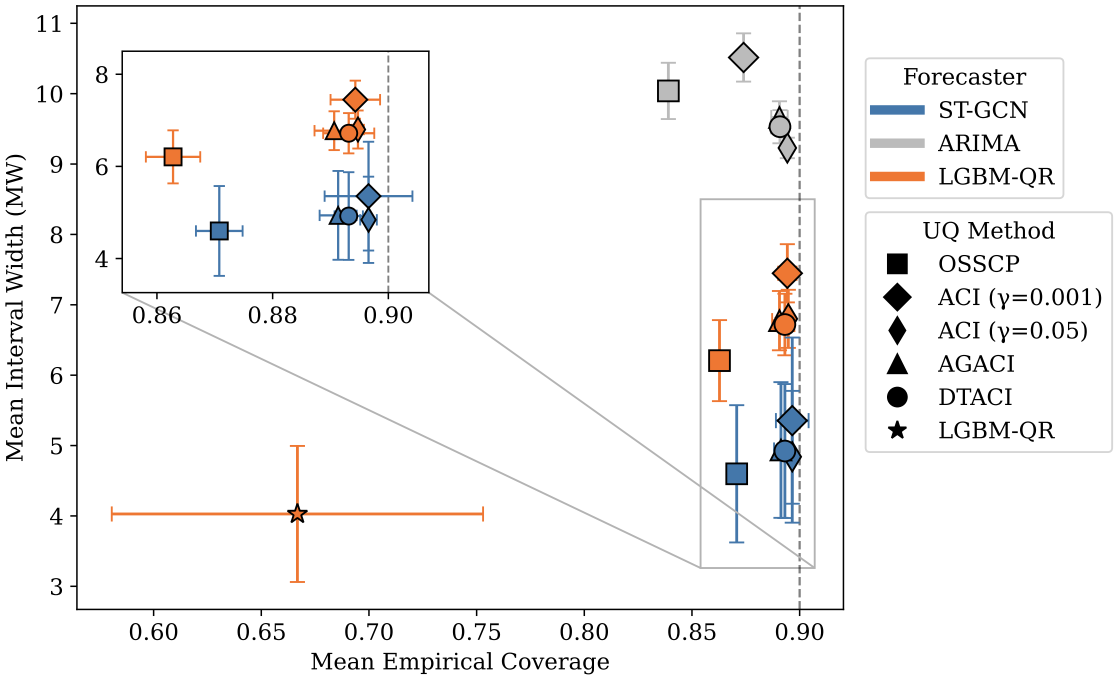


Figure 6.4: Final mean coverage and width of predictions. Dashed line at 90% coverage target. Poorly performing forecasters use conservative intervals to maintain valid coverage.

non-adaptive OSSCP and LGBM-QR methods significantly under-covered. However, it is within the width that this performance is more stark, to maintain the desired coverage, our algorithms require extremely conservative intervals. For the univariate ARIMA, which had poor point estimates across horizons, this is especially obvious. Furthermore, the error bands for the STGCN highlight the difficulty in coverage with larger forecast errors. Motivating a further look at per-horizon coverage.

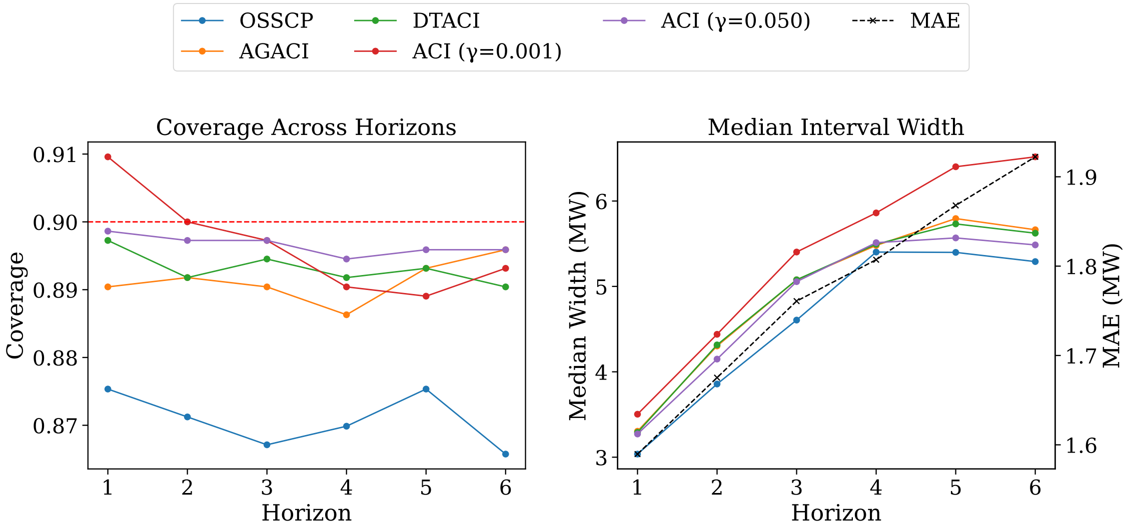


Figure 6.5: Multi-step width/coverage performance versus STGCN point estimates

The per-horizon analysis in Figure 6.5 reveals a systematic decay in coverage as the lead time increases. While all (strongly) adaptive methods provide approximate coverage for $\hat{y}_{t+1}$ and $\hat{y}_{t+6}$ the widths are forced to become increasingly conservative, with performance deteriorating at approximately the same rate as the underlying point predictors MAE. This is critical for computationally heavy models, where training per horizon is not operationally feasible, and provides strong motivation for multi-step aware ACI (X. Wang & Hyndman, 2024).

7 Conclusion

This thesis confronted the challenge of reliable uncertainty quantification in wind energy by developing a framework that bridges the gap between deep learning and theoretical statistics. We successfully demonstrated that by integrating state-of-the-art Spatio-Temporal Graph Convolutional Network with online Adaptive Conformal Inference (ACI), it is possible to produce distribution-free, theoretically valid prediction intervals, even for in a highly volatile, non-stationary system. Our findings provide evidence that this model-agnostic approach offers a practical solution for grid operators who require trustworthy probabilistic forecasts for real-time, high-stakes decision making.

Our findings, drawn from both rigorous synthetic experiments and a real-world case study, provide clear evidence for the efficiency of this approach. The synthetic experiments confirmed that ACI and its variants successfully maintain nominal coverage even under acute violations of the exchangeability assumption, adapting dynamically where non-adaptive baselines fail. The Kelmarsh wind farm case study translated the same controlled robustness into practical value. The superior accuracy of our hybrid STGCN point-forecasting model translated directly into sharper, more operationally useful intervals. This demonstrates an important result, while ACI guarantees validity, the efficiency of its intervals is tethered to the quality of the point estimates. This provides a clear pathway to turning advances in deterministic forecasting directly into quantified uncertainty, a critical development as more forecasters turn to complicated physics-informed, graphical or even foundation models for accurate predictions.

Looking forward, this research opens up several avenues for future research. The path to real-world deployment requires addressing and pushing the frontiers of methodology, validation and implementation. Methodologically, while we relied on a simplified aggregation using a standard non-conformity score, future work should move beyond univariate aggregation (sum of turbines to farm level) and explore multivariate conformal methods, multi-step specific algorithms (such as those by X. Wang and Hyndman (2024)), and physics-informed non-conformity scores. These advances would better capture the rich time-series information and result in sharper, more operationally useful intervals. Empirically, the framework must be validated at scale on both larger wind farms and benchmarked against more advanced industry-standard probabilistic techniques. Finally, practical implementation will hinge on optimising the computational overhead of advanced ACI variants into real-time operations.

In conclusion, this work establishes ACI as a powerful alternative for managing uncertainty in renewable energy systems. By providing operators with well-calibrated intervals free from unrealistic assumptions, this framework offers a tangible path to improving grid stability, and ultimately accelerating the reliable integration of renewable energy. As George E.P. Box once said, “All models are wrong, but some are useful.” We have demonstrated adaptive conformal inference helps make deterministic models reliably useful.

Bibliography

International Energy Agency. (2024, December). Renewables 2024. https://www.iea.org/energy-system/renewables/wind

National Grid Electricity System Operator (ESO). (2025, January). Britain's Electricity Explained: 2024 Review. https://www.nationalgrideso.com/

Chen, H., Li, J., O'Leary, N., & Shao, J. (2024). Examining the drivers of the imbalance price: Insights from the balancing mechanism in the united kingdom. *Journal of Environmental Management*, *371*, 123239.

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world* (Vol. 29). Springer.

Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, *9*(3).

Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Kavasseri, R. G., & Seetharaman, K. (2009). Day-ahead wind speed forecasting using f-arima models. *Renewable energy*, *34*(5), 1388–1393.

Zendehboudi, A., Baseer, M. A., & Saidur, R. (2018). Application of support vector machine models for forecasting solar and wind energy resources: A review. *Journal of Cleaner Production*, *199*, 272–285.

Wang, Y., Zou, R., Liu, F., Zhang, L., & Liu, Q. (2021). A review of wind speed and wind power forecasting with deep neural networks. *Applied Energy*, *304*, 117766.

Cui, W., Wan, C., & Song, Y. (2023). Ensemble deep learning-based non-crossing quantile regression for nonparametric probabilistic forecasting of wind power generation. *IEEE Transactions on Power Systems*, *38*(4), 3163–3178. https://doi.org/10.1109/TPWRS.2022.3202236

Zou, M., Holjevac, N., aković, J., Kuzle, I., Langella, R., Giorgio, V. D., & Djokic, S. Z. (2022). Bayesian cnn-bilstm and vine-gmcm based probabilistic forecasting of hour-ahead wind farm power outputs. *IEEE Transactions on Sustainable Energy*, *13*(2), 1169–1187. https://doi.org/10.1109/TSTE.2022.3148718

Ladopoulou, D., Hong, D. M., & Dellaportas, P. (2025). Probabilistic wind power forecasting via non-stationary gaussian processes. *arXiv preprint arXiv:2505.09026*.

Saunders, C., Gammerman, A., & Vovk, V. (1998). Ridge regression learning algorithm in dual variables.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, *113*(523), 1094–1111.

Manokhin, V. (2023). *Practical guide to applied conformal prediction in python*. Packt Publishing Ltd., Birmingham.

Romano, Y., Sesia, M., & Candes, E. (2020). Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, *33*, 3581–3591.

Gibbs, I., Cherian, J. J., & Candès, E. J. (2025). Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, qkaf008. https://doi.org/10.1093/jrsssb/qkaf008

Chernozhukov, V., Wüthrich, K., & Yinchu, Z. (2018). Exact and robust conformal inference methods for predictive machine learning with dependent data. *Conference On learning theory*, 732–749.

Xu, C., & Xie, Y. (2023). Conformal prediction for time series. *IEEE transactions on pattern analysis and machine intelligence*, *45*(10), 11575–11587.

Lin, Z., Trivedi, S., & Sun, J. (2022). Conformal prediction intervals with temporal dependence. *arXiv preprint arXiv:2205.12940*.

Tibshirani, R. J., Foygel Barber, R., Candes, E., & Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in neural information processing systems*, *32*.

Lei, L., & Candès, E. J. (2021). Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *83*(5), 911–938.

Kasa, K., Zhang, Z., Yang, H., & Taylor, G. W. (2024). Adapting conformal prediction to distribution shifts without labels. *arXiv preprint arXiv:2406.01416*.

Podkopaev, A., & Ramdas, A. (2021). Distribution-free uncertainty quantification for classification under label shift. *Uncertainty in artificial intelligence*, 844–853.

Gibbs, I., & Candes, E. (2021). Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, *34*, 1660–1672.

Zaffran, M., Féron, O., Goude, Y., Josse, J., & Dieuleveut, A. (2022). Adaptive conformal predictions for time series. *International Conference on Machine Learning*, 25834–25866.

Mao, H., Martin, R., & Reich, B. J. (2024). Valid model-free spatial prediction. *Journal of the American Statistical Association*, *119*(546), 904–914.

Barber, R. F., Candes, E. J., Ramdas, A., & Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. *The Annals of Statistics*, *51*(2), 816–845.

Oliveira, R. I., Orenstein, P., Ramos, T., & Romano, J. V. (2024). Split conformal prediction and non-exchangeable data. *Journal of Machine Learning Research*, *25*(225), 1–38.

Wisniewski, W., Lindsay, D., & Lindsay, S. (2020). Application of conformal prediction interval estimations to market makers' net positions. *Conformal and probabilistic prediction and applications*, 285–301.

Kath, C., & Ziel, F. (2021). Conformal prediction interval estimation and applications to day-ahead and intraday power markets. *International Journal of Forecasting*, *37*(2), 777–799.

Barber, R. F., Candes, E. J., Ramdas, A., & Tibshirani, R. J. (2021). Predictive inference with the jackknife+. *The Annals of Statistics*, *49*(1), 486–507.

Xu, C., & Xie, Y. (2021). Conformal prediction interval for dynamic time-series. *International Conference on Machine Learning*, 11559–11569.

Gibbs, I., & Candès, E. J. (2024). Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, *25*(162), 1–36.

Wang, X., & Hyndman, R. J. (2024). Online conformal inference for multi-step time series forecasting. *arXiv preprint arXiv:2410.13115*.

Sun, S. H., & Yu, R. (2023). Copula conformal prediction for multi-step time series prediction. *The Twelfth International Conference on Learning Representations*.

Vovk, V., Petej, I., Nouretdinov, I., Ahlberg, E., Carlsson, L., & Gammerman, A. (2021). Retrain or not retrain: Conformal test martingales for change-point detection. *Conformal and Probabilistic Prediction and Applications*, 191–210.

Lee, J., Xu, C., & Xie, Y. (2024a). Transformer conformal prediction for time series. *arXiv preprint arXiv:2406.05332*.

O'Connor, C., Prestwich, S., & Visentin, A. (2024). Conformal prediction techniques for electricity price forecasting. *International Workshop on Advanced Analytics and Learning on Temporal Data*, 1–17.

Lee, J., Xu, C., & Xie, Y. (2024b). Kernel-based optimally weighted conformal prediction intervals. *arXiv preprint arXiv:2405.16828*.

Renkema, Y., Brinkel, N., & Alskaif, T. (2024). Conformal prediction for stochastic decision-making of pv power in electricity markets. *Electric Power Systems Research*, *234*, 110750.

Althoff, S., Szabadv'ary, J. H., Anderson, J., & Carlsson, L. (2023). Evaluation of conformal-based probabilistic forecasting methods for short-term wind speed forecasting. *Conformal and Probabilistic Prediction with Applications*, 100–115.

Zuege, C. V., Stefenon, S. F., Yamaguchi, C. K., Mariani, V. C., Gonzalez, G. V., & dos Santos Coelho, L. (2025). Wind speed forecasting approach using conformal prediction and feature importance selection. *International Journal of Electrical Power & Energy Systems*, *168*, 110700.

Jonkers, J., Avendano, D. N., Van Wallendael, G., & Van Hoecke, S. (2024). A novel day-ahead regional and probabilistic wind power forecasting framework using deep cnns and conformalized regression forests. *Applied Energy*, *361*, 122900.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al. (2023). Learning skillful medium-range global weather forecasting. *Science*, *382*(6677), 1416–1421.

Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Allen, A., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J. A., Dong, H., et al. (2025). A foundation model for the earth system. *Nature*, 1–8.

International Electrotechnical Commission, I. (2005). 12-1: Power performance measurements of electricity producing wind turbines. *British Standard, IEC*, 61400–12.

Wintenberger, O. (2017). Optimal learning with bernstein online aggregation. *Machine Learning*, *106*(1), 119–141.

Kipf, T. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551.

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Yu, B., Yin, H., & Zhu, Z. (2017). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, *38*(11), 2074–2102.

National Renewable Energy Laboratory (NREL). (2025). *FLORIS: FLOw Redirection and Induction in Steady State* [Software package]. Golden, CO, USA. https://github.com/NREL/floris

Plumley, C. (2022). Kelmarsh wind farm data; 2022.

Zippenfenig, P. (2023). Open-meteo. com weather api. In *Computer software*. Zenodo.

Zjavka, L. (2015). Wind speed forecast correction models using polynomial neural networks. *Renewable Energy*, *83*(100), 998–1006. https://doi.org/10.1016/j.renene.2015.04.054

A Appendix

A.1 Algorithms

Algorithm 1 Adaptive Conformal Inference (ACI)

```
Input: starting value $\theta_1$, learning rate $\gamma > 0$.
for $t = 1, 2, \ldots, T$ do
    Output: prediction interval $\hat{C}_t(X_t) = \{y \in \mathbb{R} : |y - \hat{Y}_t| \leq \hat{\alpha}_{t-1}\}$.
    Observe $y_t$.
    Evaluate $\mathrm{err}_t = \mathbb{I}\{y_t \notin \hat{C}_t(X_t)\}$.
    Update $\alpha_{t+1} = \alpha_t + \gamma(\alpha - \mathrm{err}_t)$.
end for
```

Intuition: ACI adaptively adjusts the width of prediction intervals to maintain the desired coverage, widening when intervals miss the target and shrinking when they over-cover.

Algorithm 2 Aggregated Adaptive Conformal Inference (AgACI)

```
Input: learning rates $(\gamma_k)_{k=1}^{K}$, starting value $\theta_1$.
Initialise BOAs: $B_\ell \leftarrow \mathrm{BOA}(\alpha \leftarrow (1-\alpha)/2)$, $B_u \leftarrow \mathrm{BOA}(\alpha \leftarrow 1 - (1-\alpha)/2)$.
for $k = 1, \ldots, K$ do
    Initialise $A_k \leftarrow \mathrm{ACI}(\alpha, \gamma_k, \theta_1)$.
end for
for $t = 1, \ldots, T$ do
    for $k = 1, \ldots, K$ do
        Retrieve interval $[\ell_t^k, u_t^k]$ from $A_k$.
    end for
    Compute $\tilde{\ell}_t \leftarrow B_\ell(\{\ell_t^k\}_{k=1}^{K})$,    $\tilde{u}_t \leftarrow B_u(\{u_t^k\}_{k=1}^{K})$.
    Output interval $[\tilde{\ell}_t, \tilde{u}_t]$ and observe $y_t$.
    for $k = 1, \ldots, K$ do
        Update $A_k$ with $y_t$.
    end for
    Update $B_\ell$ and $B_u$ with $y_t$.
end for
```

Intuition: AgACI aggregates many experts with different learning rates (γs) then individually combines them to get upper and lower bounds via BOA online aggregation. Reduced dependence on γ selection.

Algorithm 3 Dynamically-tuned Adaptive Conformal Inference (DtACI)

```
Input: $\theta_1$, $(\gamma_k)_{k=1}^{K}$, $\sigma$, $\eta$.
for $k = 1, \ldots, K$ do
    $A_k \leftarrow \text{ACI}(\alpha, \gamma_k, \theta_1)$.
end for
for $t = 1, \ldots, T$ do
    Normalise $p_t^k \leftarrow p_t^k / \sum_i p_t^i$.
    $\theta_t \leftarrow \sum_k \theta_t^k p_t^k$; output $\hat{C}_t(\theta_t)$.
    Observe $y_t$, compute $r_t$.
    $\bar{w}_t^k \leftarrow p_t^k e^{-\eta L_\alpha(\theta_t^k, r_t)}$, $\bar{W}_t \leftarrow \sum_i \bar{w}_t^i$.
    $p_{t+1}^k \leftarrow (1-\sigma)\frac{\bar{w}_t^k}{\bar{W}_t} + \frac{\sigma}{K}$.
    $err_t \leftarrow \mathbb{I}[y_t \notin \hat{C}_t(\theta_t)]$.
    for $k = 1, \ldots, K$ do
        Update $A_k$ with $y_t$; get $\theta_{t+1}^k$.
    end for
end for
```

Intuition: DtACI dynamically combines multiple ACI *learners* with different γs based on past performance, adaptively tuning the intervals to correct over- or under-coverage within a specified window I.

A.2 Simulation Study

Table A.1: Key `FLORIS v4` settings for simulation replication.

Component	Settings
Version / Model	FLORIS v4, GCH
Solver	`turbine_grid` (3 points)
Layout	9 turbines (3×3 grid, 500 m spacing)
Turbine	NREL 5MW
Flow Field	$U = 8$ m/s, Dir=270°, TI=0.1, Shear=0.12, $\rho = 1.225$ kg/m^3

Wake Models	Velocity=gauss; Deflection=gauss;
	Turbulence=crespo_hernandez; Combo=sosfs

Wake Options	Sec. steering=Yes; Yaw recovery=Yes; Active mixing=No; Transverse=Yes
Parameters	Gauss: $\alpha = 0.58$, $\beta = 0.077$, $k_a = 0.38$, $k_b = 0.004$ Crespo-Hernandez: init=0.1, const=0.5, $a_i = 0.8$, down=-0.32

Table A.2: Autocorrelation Experiment Results for AR(1) Model.

Method	**Param = 0.1**	**Param = 0.5**	**Param = 0.8**	**Param = 0.9**	**Param = 0.95**
Cell format: Coverage (Std. Dev.) / Median Length (Std. Dev.)					
ACI (0.001)	0.899 (0.013) 409.0 (14.9)	0.900 (0.013) 366.0 (12.3)	0.906 (0.012) 351.4 (11.6)	0.914 (0.012) 352.9 (12.3)	0.932 (0.011) 363.8 (14.2)
ACI (0.05)	0.899 (0.003) 409.4 (19.0)	0.899 (0.003) 366.4 (17.7)	0.900 (0.003) 349.1 (16.7)	0.901 (0.002) 342.0 (16.8)	0.902 (0.003) 331.7 (18.0)
AgACI	0.899 (0.006) 410.3 (17.4)	0.899 (0.006) 367.2 (16.7)	0.900 (0.006) 347.4 (14.6)	0.902 (0.005) 341.6 (14.9)	0.905 (0.006) 335.3 (16.3)
DtACI	0.897 (0.006) 407.4 (17.8)	0.899 (0.005) 367.7 (15.4)	0.900 (0.006) 346.9 (13.4)	0.901 (0.005) 342.3 (15.1)	0.905 (0.005) 336.5 (15.0)
OSSCP	0.896 (0.010) 406.8 (18.6)	0.896 (0.010) 364.0 (15.7)	0.897 (0.009) 342.9 (15.1)	0.896 (0.010) 335.3 (14.2)	0.896 (0.011) 326.6 (15.9)

Table A.3: Autocorrelation Experiment Results for MA(1) Model.

Method	**Param = 0.1**	**Param = 0.5**	**Param = 0.8**	**Param = 0.9**	**Param = 0.95**
Cell format: Coverage (Std. Dev.) / Median Length (Std. Dev.)					
ACI (0.001)	0.899 (0.013) 409.1 (14.3)	0.900 (0.013) 390.5 (14.5)	0.899 (0.014) 423.8 (16.1)	0.899 (0.014) 442.4 (17.4)	0.899 (0.014) 451.9 (16.4)
ACI (0.05)	0.899 (0.003) 410.0 (19.1)	0.899 (0.003) 392.4 (19.4)	0.899 (0.003) 424.3 (21.1)	0.899 (0.003) 441.0 (23.6)	0.899 (0.003) 453.0 (23.1)
AgACI	0.899 (0.006) 407.6 (18.8)	0.899 (0.006) 391.8 (16.6)	0.899 (0.006) 423.8 (20.6)	0.899 (0.006) 443.3 (21.4)	0.899 (0.006) 451.7 (20.1)
DtACI	0.898 (0.006) 409.5 (19.0)	0.899 (0.006) 390.4 (17.0)	0.898 (0.006) 424.0 (20.3)	0.898 (0.006) 441.3 (19.7)	0.899 (0.005) 450.8 (20.6)
OSSCP	0.896 (0.010) 406.4 (18.6)	0.896 (0.010) 389.0 (18.1)	0.896 (0.011) 422.7 (20.6)	0.896 (0.011) 439.6 (21.8)	0.895 (0.011) 450.5 (21.7)

Table A.4: Autocorrelation Experiment Results for ARMA(1,1) Model.

Method	**Param = 0.1**	**Param = 0.5**	**Param = 0.8**	**Param = 0.9**	**Param = 0.95**
Cell format: Coverage (Std. Dev.) / Median Length (Std. Dev.)					
ACI (0.001)	0.898 (0.014) 427.6 (16.2)	0.900 (0.015) 444.9 (19.1)	0.904 (0.014) 456.5 (17.3)	0.912 (0.014) 457.1 (18.0)	0.925 (0.015) 468.1 (22.8)
ACI (0.05)	0.899 (0.003) 426.8 (21.5)	0.899 (0.003) 441.8 (22.1)	0.899 (0.003) 447.4 (23.9)	0.900 (0.003) 438.7 (26.3)	0.902 (0.004) 426.1 (31.1)
AgACI	0.899 (0.006) 428.1 (19.3)	0.898 (0.006) 442.4 (21.5)	0.898 (0.006) 451.2 (21.7)	0.900 (0.007) 443.3 (24.1)	0.903 (0.007) 431.3 (31.0)
DtACI	0.898 (0.006) 428.6 (20.4)	0.898 (0.006) 442.4 (21.3)	0.898 (0.006) 449.5 (21.9)	0.898 (0.007) 443.0 (24.4)	0.901 (0.007) 430.9 (32.2)
OSSCP	0.896 (0.011) 424.9 (19.8)	0.895 (0.011) 442.1 (22.8)	0.895 (0.010) 447.1 (23.2)	0.895 (0.011) 437.7 (22.9)	0.893 (0.014) 425.3 (30.0)

Table A.5: Regime Switching Experiment Results.

Method	**Prob = 0.005**	**Prob = 0.01**	**Prob = 0.02**	**Prob = 0.05**
Cell format: Coverage (Std. Dev.) / Median Length (Std. Dev.)				
ACI (0.001)	0.895 (0.052) 359.2 (76.5)	0.894 (0.048) 365.5 (73.4)	0.896 (0.037) 381.1 (52.9)	0.896 (0.025) 392.1 (37.6)
ACI (0.05)	0.899 (0.010) 340.1 (110.6)	0.898 (0.009) 353.0 (101.7)	0.899 (0.008) 353.5 (76.9)	0.899 (0.006) 372.0 (54.0)
AgACI	0.901 (0.016) 346.4 (105.6)	0.901 (0.016) 364.3 (96.9)	0.900 (0.013) 363.4 (71.2)	0.899 (0.010) 377.6 (49.3)
DtACI	0.901 (0.012) 351.4 (110.6)	0.900 (0.011) 353.3 (96.6)	0.899 (0.010) 358.1 (76.7)	0.898 (0.009) 376.2 (50.0)
OSSCP	0.888 (0.032) 356.9 (109.3)	0.885 (0.034) 365.9 (94.1)	0.889 (0.028) 370.9 (69.6)	0.891 (0.020) 385.8 (49.9)

Table A.6: Wind Farm Experiment Results.

Method	**4 Wind Dirs**	**8 Wind Dirs**
Cell format:	*Coverage (Std. Dev.) / Median Length (Std. Dev.)*	
ACI (0.001)	0.899 (0.014) 1.337e6 (5.02e4)	0.899 (0.014) 7.203e6 (7.23e4)
ACI (0.05)	0.899 (0.003) 1.339e6 (6.49e4)	0.899 (0.003) 7.192e6 (9.07e4)
AgACI	0.898 (0.006) 1.343e6 (5.92e4)	0.897 (0.007) 7.203e6 (8.73e4)
DtACI	0.897 (0.006) 1.340e6 (6.03e4)	0.896 (0.006) 7.193e6 (8.60e4)
OSSCP	0.896 (0.010) 1.329e6 (6.19e4)	0.896 (0.011) 7.196e6 (8.85e4)

A.3 Case Study

Table A.7: Specifications for the six turbines at Kelmarsh Wind Farm, UK. All turbines are Senvion MM92 models with a rated power of 2050 kW and a rotor diameter of 92 m. Commercial operations began in 15/04/2016.

Turbine ID	**Hub Height (m)**	**Coordinates (Lat, Lon)**	**Elevation (m)**
Kelmarsh 1	78.5	$(52.4006, -0.9471)$	145.598
Kelmarsh 2	78.5	$(52.4026, -0.9495)$	156.577
Kelmarsh 3	68.5	$(52.4038, -0.9442)$	153.477
Kelmarsh 4	78.5	$(52.3988, -0.9412)$	146.313
Kelmarsh 5	78.5	$(52.4023, -0.9405)$	142.901
Kelmarsh 6	68.5	$(52.4007, -0.9361)$	135.039

Table A.8: Forecasting model configurations.

Model	Parameters	Training
STGCN	$n_{his} = 48$, hidden=64, $k = 3$, blocks=2, dropout=0.2, act=GLU	50 epochs, lr=10^{-3}, batch=32
LGBM	300 estimators, depth=7, lr=0.05, min_child=20, objective=quantile	quantiles [0.05, 0.5, 0.95]
ARIMA	$p, q \in [0, 5]$, rolling window=168	retrain weekly

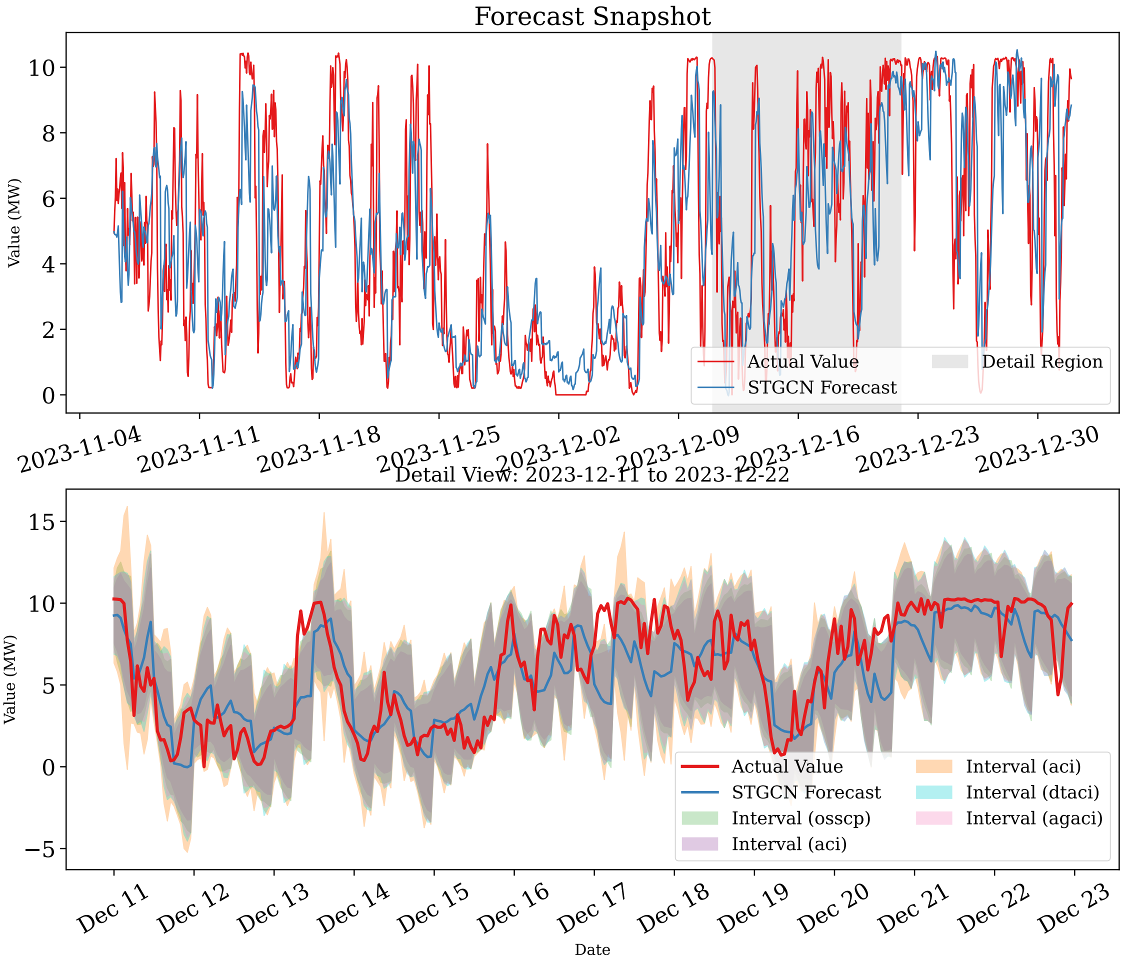


Figure A.1: Snapshot of point estimate wrapped via conformal methods for probabilistic bounds in real-world regime-shifts.