Cross-Domain Offshore Wind Power Forecasting: Transfer Learning Through Meteorological Clustering and Gaussian Process Adaptation

September 5, 2025

Contents

Abstract

The green energy transition depends on efficiently deploying renewable energy sources such as offshore wind farms, where accurate power production forecasting is essential for planning, funding, and operations. Although data-driven Machine Learning (ML) models have shown strong performance in offshore Wind Power Forecasting (WPF), they often struggle to generalise across locations, requiring large, costly, site-specific datasets and full retraining for new installations [1]. This report addresses these limitations by proposing a novel framework for one-hour-ahead offshore WPF, first trained on a source wind farm with a full dataset, before applying three different Transfer Learning (TL) strategies designed to enable target wind farms to achieve accurate forecasting with limited data. Unlike computationally intensive Deep Learning (DL) approaches, this methodology employs unsupervised clustering to group meteorologically similar time-periods, within which individual Gaussian Processes (GPs) are trained for each cluster. This design allows each model to specialise in distinct weather patterns while leveraging the strengths of GPs, including uncertainty quantification, interpretability and robust performance in data-scarce settings [2]. This framework allows time-periods from target wind farms to be allocated to source clusters with similar weather patterns, enabling the transfer of knowledge between comparable meteorological conditions.

The proposed WPF framework achieved a Mean Absolute Error (MAE) of 4.14%, expressed as a % of the total power capacity of the wind farm, using only 2 years of training data. This outperforms the Informer network proposed by Wang et al. [3], which achieved 4.42% MAE despite being trained on five years of data from the same dataset. To evaluate the TL methodologies, experiments were ran on two target wind farms. The approaches demonstrated strong cross-domain generalisation, achieving MAEs of 3.62%, 3.64% and 4.01% when trained on only 50% of the available two-year dataset. Notably, these results were achieved using a standard CPU-only setup, in contrast to Wang et al.'s model, which required a high-end GPU (NVIDIA GeForce RTX 3090) for training. By leveraging this TL framework, forecasting systems can be rapidly deployed at new offshore wind farms, reducing the need for extensive site-specific datasets and supporting scalable expansion of renewable energy infrastructure.

Chapter 1

Introduction

The global energy landscape is undergoing a fundamental transformation, with offshore wind power emerging as a cornerstone of sustainable electricity generation. According to the Global Offshore Wind Report [4], worldwide offshore wind capacity exceeded 80 GW by the end of 2024, with projections reaching 254 GW by 2030. However, this rapid expansion faces a critical hurdle: only 19 countries globally have installed offshore wind turbines, with merely three additional nations expected to join by 2030. This geographic concentration highlights the fundamental challenges facing countries entering this sector: substantial technical expertise and financial investment are required, yet the lack of historical operational data hinders the deployment of accurate WPF systems complicating economic viability assessments.

Accurate WPF is essential for both existing and planned offshore wind farms, ensuring the efficient integration of this inherently variable resource into power grids and enabling optimal scheduling, reserve allocation and market operations [5]. The stochastic nature of wind resources introduces significant uncertainty into power system operations, with forecast errors directly translated to increased operational costs through unnecessary reserve procurement or reliability risks from insufficient backup capacity [6]. For offshore installations, these challenges are amplified by complex marine atmospheric dynamics, limited meteorological observations and the significant cost of deploying comprehensive data acquisition infrastructure [7]. Traditional offshore meteorological masts can cost between €10-12 million each to install offshore in Europe [8]. To address this, floating Light Detection and Ranging (LIDAR) systems have emerged as a more economical alternative. LIDAR is a remote sensing technology that uses laser pulses to measure atmospheric properties such as wind speed and direction, and in offshore wind it is increasingly deployed for resource assessments and site characterisation. These systems can reduce costs by up to 90% compared to masts [9], although they still represent a considerable investment, typically in the range of €1–4 million depending on system complexity and deployment conditions [10].

ML and DL approaches have revolutionised short-term WPF, consistently outperforming physical and statistical models through their ability to capture complex non-linear relationships in meteorological data [11]. DL architectures, particularly Long Short-Term Memory (LSTM) networks and their variants, have achieved remarkable success in modelling the temporal dependencies inherent in wind power generation [12]. However, these successes are predominantly confined to locations with extensive historical datasets. For

countries and companies evaluating the feasibility of new offshore wind farms, a circular dependency emerges: accurate forecasts are needed to secure project financing and grid connection agreements, yet such forecasts rely on operational data that only becomes available after construction.

TL offers a promising solution to this paradox by leveraging knowledge gained from data-rich source domains to improve learning in data-scarce target domains [13]. Recent applications in WPF have demonstrated that neural network models trained on established wind farms can be successfully adapted to new locations, achieving performance improvements of 6-28% compared to training from scratch [14]. However, existing TL approaches for WPF predominantly rely on complex DL architectures that present significant practical limitations. While these models excel at capturing long-range patterns, their computational requirements often limit real-time deployment [15]. Moreover, their "black-box" nature provides limited insight into the meteorological basis for their predictions, hindering acceptance by grid operators who require interpretable uncertainty quantification for risk management. The computational and interpretability challenges of DL models are particularly acute in the offshore context, where operational constraints demand robust, efficient and transparent forecasting systems. Offshore wind farm operators require not only accurate point forecasts but also reliable uncertainty estimates to manage maintenance scheduling, vessel logistics and grid stability commitments [16].

To address these challenges, this report proposes a novel forecasting framework that combines Variational Autoencoders (VAEs) with GPs. A VAE is used to learn a compact latent representation of meteorological conditions for individual time-periods, allowing these time-periods to be clustered by distinct weather characteristics. For each cluster, an individual GP is trained to model the relationship between meteorological features and power output. The probabilistic formulation of both VAEs and GPs also enables uncertainty quantification, while clustering time-periods and training separate GPs for each cluster reduces the size of individual training sets, helping to make the framework more manageable for operational offshore environments.

Crucially, this design facilitates a more interpretable form of TL. Rather than assuming geographic similarity between wind farms, the framework recognises that it is the meteorological conditions that govern wind-to-power relationships. By transferring models between clusters that share similar weather patterns, predictive knowledge can be reused across offshore sites even when they are geographically distant. For example, a cluster capturing strong westerly winds with moderate variance may emerge both in the North Sea and in the Baltic, resulting in comparable power generation patterns. Transferring a GP trained on this cluster from a data-rich site to a new offshore farm thus provides accurate forecasts with minimal local data, while maintaining interpretability through meteorologically grounded transfer.

This framework therefore addresses the dual limitations of DL approaches in WPF: it reduces computational overhead by avoiding deep parametric networks and enhances interpretability by grounding transfer in observable meteorological regimes.

The remainder of this report is organised as follows: Chapter 2 presents a comprehensive literature review, highlighting recent and innovative research in the application of

ML, DL and TL to WPF. Chapter 3 introduces the theoretical foundations of the study, focusing on VAEs, GPs and TL, which form the basis of the methodological framework. Chapter 4 details the proposed methodology, followed by Chapter 5 which presents and analyses the experimental results. Finally, Chapter 6 discusses potential avenues for future research and outlines how the presented methodology and findings could be further developed.

Chapter 2

Literature Review

Evolution of offshore Wind Power Forecasting Methods

The progression of WPF methodologies reflects the broader evolution of computational capabilities and data availability in the renewable energy sector. Early forecasting approaches were rooted in Numerical Weather Prediction (NWP) models. A well-known example is the Integrated Forecasting System (IFS) of the European Centre for Medium-Range Weather Forecasts (ECMWF) [17]. While physically grounded, these models rely on solving complex atmospheric equations and thus demand significant computational power. Moreover, NWP models often struggle to capture the fine-scale atmospheric variability typical of offshore sites [18]. These limitations motivated a shift toward lighter statistical approaches. Statistical time-series approaches, notably AutoRegressive Integrated Moving Average (ARIMA) and exponential smoothing, emerged as more computationally viable and easier to deploy alternatives. However, despite their practicality, linear time-series models often fail to capture the highly non-linear relationships between meteorological inputs (like wind speed and direction) and power output [5].

The ML revolution in offshore WPF brought methods like Random Forests (RF), Gradient Boosting Machines (GBMs) and Support Vector Machines (SVMs) to the fore. These models often significantly outperformed traditional linear statistical approaches by effectively handling the complex, non-linear relationships present in meteorological data [19]. SVMs, in particular, excelled in modelling high-dimensional input data through kernel transformations, though their performance often depended on careful feature selection and preprocessing [20].

Deep Learning Architectures for offshore Wind Power Forecasting

The emergence of DL offered a breakthrough by enabling models to capture spatio-temporal dynamics inherent in wind speed and power signals without the need for manual feature engineering. For instance, Long Short-Term Memory (LSTM) networks and their variants have consistently outperformed traditional ML models in short-term WPF, including offshore applications [21]. For more complex architectures, hybrid DL models have achieved state-of-the-art results in offshore WPF. For example, Khan et al. [22] proposed a hybrid CNN-LSTM model, which combines Convolutional Neural Networks (CNNs) for spatial feature extraction with LSTMs for temporal modelling, improving forecasting accuracy by capturing both local wind patterns and time-series dependencies.

Additionally, the attention-based Informer model, such as the one proposed by Wang et al. [3], designed for long-sequence forecasting, has demonstrated superior performance in large-scale offshore wind farms by leveraging self-attention mechanisms to focus on critical temporal patterns, achieving lower prediction errors in multi-step short-term forecasting compared to traditional LSTMs.

Limitations of Current Deep Learning Approaches

Despite these advances in accuracy, DL methods face severe limitations in real-world deployment. High computational cost and data requirements make them challenging to apply in newly commissioned wind farms where historical data is scarce. Moreover, DL models are often criticised for their limited interpretability, restricting their acceptance in safety-critical energy systems. To overcome these limitations, TL has been increasingly applied to WPF, offering a means to leverage knowledge from data rich source domains to enhance predictions in data scarce target farms [23].

Gaussian Process Regression for Wind Power Forecasting

As an alternative to DL methods, GPs offer a promising approach for WPF due to their ability to quantify uncertainty, provide interpretability and perform effectively with limited data [24]. For instance, GPs have been successfully applied to wind speed forecasting, capturing short-term variability with high accuracy [25]. In WPF, studies have leveraged GPs to generate probabilistic forecasts, with one approach applying Gaussian Process Regression with a combination of Gaussian kernels, yielding a 1.8% improvement over a neural network based model in monthly forecasting performance [26]. Another study applied GPs in solar forecasting by training individual models on clustered time-periods, similar to the methodology proposed in this report [27]. While these studies demonstrate GPs' effectiveness in renewable energy forecasting, the integration of cluster-specific GP models with TL techniques presents an unexplored opportunity to combine interpretability with cross-domain generalisation in offshore WPF.

Transfer Learning in Wind Power Forecasting

The challenge of deploying accurate forecasting systems at new wind farms with limited operational data has driven significant interest in TL for WPF. Early applications of TL to wind power focused on simple parameter transfer between similar wind farms. However, more sophisticated approaches have emerged that adapt to the unique characteristic of each site whilst preserving valuable knowledge from source domains. Islam Sajol et al. applied deep domain adaptation to WPF, demonstrating that selectively fine-tuning only the final layers of Deep Neural Networks pre-trained on source wind farms can improve forecasting accuracy by 6.14%-28.44% when transferring between sites in Germany, France and the UK [14].

Similarly, Li et al. developed a TimesNet–Gated Recurrent Unit (GRU) architecture that employs Maximum Mean Discrepancy (MMD) to quantify distributional differences between source and target domains [28]. This metric guides the selection of which network components to freeze or fine-tune, enabling more precise adaptation strategies. Their ap-

proach achieved substantial performance improvements for new wind farms.

Addressing the computational burden of DL approaches, Khan et al. proposed a hybrid framework combining VAEs with TL [29]. Their method first compresses high-dimensional wind data through pre-trained Multi Layer Perceptron (MLP) autoencoders, then fine-tunes on just 10% of target farm data. This reduced computational overhead by a factor of 72.29 compared to training from scratch, making TL more feasible for resource-constrained scenarios.

Recent advances have explored spatial transfer across multiple wind farms simultaneously. Wang et al.'s proposed Informer-based multi-location multi-modal multi-step (M3STIN) prediction model integrates Graph Attention Networks (GANs) with Informer architectures to capture both spatial correlations between farms and temporal dependencies within each site [3]. By modelling eight offshore wind farms jointly, their approach achieved Root Mean Squared Error (RMSE) reductions of 6.03-12.67% across multi-step forecasts. However, the requirement for extensive inter-farm data and complex graph structures limits its applicability to isolated offshore installations.

Beyond neural network architectures, alternative transfer mechanisms have shown promise. Wang et al. introduced TL for Concept Drift Detection-GRU (TLCDD-GRU) model, which employs online TL to adapt dynamically to concept drift [30]. Chen et al. demonstrated that knowledge distillation using teacher-student frameworks, including non-neural models, can achieve RMSE improvements of 3.3–23.9% [31]. These diverse approaches highlight that effective TL extends beyond deep neural networks, encompassing various methodologies.

Despite these advances, current TL methods for WPF face practical limitations. The TimesNet-GRU model requires extensive hyperparameter tuning across multiple parallel modules, whilst domain adaptation approaches demand careful architectural decisions about layer freezing. Moreover, these methods often rely on extensive meteorological features that may be unavailable at new offshore sites. The opacity and complexity of DL models further limits their adoption in safety-critical grid operations where interpretable uncertainty quantification is essential. These limitations motivate the exploration for approaches that not only transfer knowledge effectively but also provide interpretable and data-efficient forecasts. Methods that embed probabilistic modelling within TL can address the opacity, feature dependence and adaptability challenges of current DL-based approaches, thereby offering a more practical pathway for offshore WPF.

Chapter 3

Models and Notation

Notation

Throughout this chapter, a consistent mathematical notation is adopted to facilitate understanding. Input data vectors are denoted as $\mathbf{x} \in \mathbb{R}^n$, with corresponding output variables $y \in \mathbb{R}$. In the context of VAEs, $\mathbf{z} \in \mathbb{R}^k$ represents the latent representation vector where $k \ll n$, while ϕ and θ denote the parameters of the encoder and decoder networks, respectively. The probabilistic framework employs $p(\cdot)$ for prior distributions, $q_\phi(\mathbf{z} \mid \mathbf{x})$ for the approximate posterior distribution parametrised by the encoder, and $p_\theta(\mathbf{x}|\mathbf{z})$ for the conditional likelihood function parametrised by the decoder. Statistical parameters are represented by μ and σ^2 for mean and variance, respectively. For Gaussian Process Regression, $f(x) \sim \mathcal{GP}\big(m(\mathbf{x}),\, k(\mathbf{x}, \mathbf{x}')\big)$ denotes a GP with mean function $m(\mathbf{x})$ and kernel function $k(\mathbf{x}, \mathbf{x}')$, where the kernel defines similarity between inputs $\mathbf{x}$ and $\mathbf{x}'$ and includes hyperparameters such as signal variance σ_f^2, length scales ℓ_d for input dimension d, and noise variance σ_n^2. In the TL framework, domains are distinguished as $\mathcal{D}_S$ (source) and $\mathcal{D}_T$ (target), each comprising a feature space $\mathcal{X}$ and marginal probability distributions $p_S(\mathbf{x})$ and $p_T(\mathbf{x})$, respectively. Tasks are denoted as $\mathcal{T}_S$ and $\mathcal{T}_T$, formally defined as $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$ where $\mathcal{Y}$ represents the output space and $f(\cdot)$ is the predictive function to be learned. Loss functions are represented by $\mathcal{L}$ with appropriate subscripts (e.g., $\mathcal{L}_{\text{VAE}}$, $\mathcal{L}_{\text{task}}$), and additional notation includes the KL divergence $D_{\text{KL}}(\cdot\|\cdot)$, the regularisation parameter β in VAEs, distributional divergence measures $d(\cdot, \cdot)$, and sample weights w_i in instance based TL applications.

3.1 Variational Autoencoders

VAEs are generative models that combine the principles of autoencoders with Bayesian inference to learn structured latent representations of high-dimensional data [32]. A VAE consists of two primary components: an encoder, which compresses input data $\mathbf{x} \in \mathbb{R}^n$ into a lower-dimensional latent representation $\mathbf{z} \in \mathbb{R}^k$ where $k \ll n$, and a decoder, which reconstructs the input data from samples drawn from the latent space. Unlike traditional autoencoders that produce deterministic latent variables, VAEs adopt a probabilistic approach by modelling the latent variables as a distribution, typically a multivariate Gaussian. This allows the encoder to output parameters (mean μ and variance σ^2) of the latent distribution, enabling stochastic sampling during reconstruction [32].

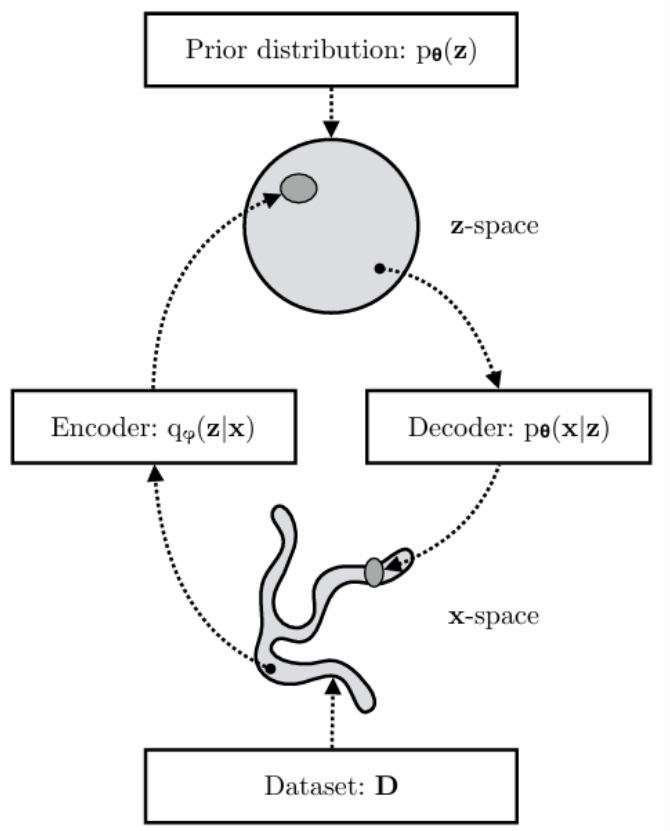


Figure 3.1: Structure of a VAE [33]

To maintain a structured and continuous latent space suitable for generative tasks, VAEs impose a regularisation constraint using the Kullback–Leibler (KL) divergence [32]. The KL divergence encourages the approximate posterior distribution

$$q_\phi(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2)),$$

produced by the encoder with parameters ϕ, where μ_j and σ_j^2 are the mean and variance of the approximate posterior for latent dimension j, to closely resemble a prior distribution $p(\mathbf{z})$, typically a standard Gaussian $\mathcal{N}(0, I)$. The KL divergence between these two Gaussian distributions has the closed-form expression:

$$D_{\mathrm{KL}}\big(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p(\mathbf{z})\big) = \frac{1}{2} \sum_{j=1}^{k} \big(\sigma_j^2 + \mu_j^2 - 1 - \log \sigma_j^2\big).$$

The decoder, parametrised by θ, models the likelihood $p_\theta(\mathbf{x} \mid \mathbf{z})$, generating reconstructed data conditioned on latent samples [32]. Training a VAE involves optimising the Evidence Lower Bound (ELBO), which balances two objectives: (i) a reconstruction term that ensures the decoded output closely matches the input data, and (ii) a KL divergence term that regularises the latent space to prevent overfitting and ensure generalisability. The loss function is expressed below, where β is a hyperparameter that controls the trade-off between reconstruction accuracy and latent space regularisation, which is adjustable to prioritise one term over the other [34]:

$$\mathcal{L}_{\mathrm{VAE}} = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big]}_{\text{Reconstruction term}} - \beta \underbrace{D_{\mathrm{KL}}\big(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z})\big)}_{\text{KL divergence term}}.$$

The reconstruction term is typically implemented as the mean squared error (for continuous data) or cross-entropy (for discrete data), while the KL divergence term can be computed analytically for Gaussian distributions, making optimisation tractable.

In the context of WPF, VAEs have proven effective for unsupervised feature extraction. The compressed latent representations learned by VAEs capture the most informative aspects of each weather time-period, enabling more efficient differentiation between distinct meteorological patterns and facilitating effective clustering. Harrou et al. [35] demonstrated the effectiveness of this approach by integrating a self-attention mechanism with a VAE, achieving superior performance compared to eight standard, non-hybrid DL and ML models.

3.2 Gaussian Process Regression

Gaussian Process Regression (GPR) offers a non-parametric Bayesian framework for modelling functions, providing inherent uncertainty quantification that is particularly valuable in applications such as WPF. A GP defines a probability distribution over an infinite collection of functions, such that any finite subset of function values follow a multivariate Gaussian distribution [36]. This section provides a concise introduction; for a more comprehensive explanation, the reader is referred to Rasmussen and Williams [24].

A GP is completely specified by a mean function $m(\mathbf{x})$, typically assumed zero, and a positive definite covariance function, or kernel, $k(\mathbf{x}, \mathbf{x}')$ [24]. A GP is denoted as:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')).$$

The kernel encodes prior assumptions about the function's properties, such as smoothness, periodicity, or trends, by defining the similarity between inputs. Different kernels impose varying notions of similarity, enabling GPs to adapt to diverse data patterns [37]. Three commonly used kernels relevant to WPF are illustrated below:

Radial Basis Function (RBF) Kernel:

Also known as the squared exponential kernel, the RBF kernel assumes infinitely differentiable functions, resulting in very smooth realisations. It models similarity via exponential decay based on the distance between inputs, with values approaching the signal variance σ_f^2 for nearby points (high correlation) and zero for distant points (low correlation). The length scales ℓ_d determine the rate of decay along each input dimension d: smaller ℓ_d permit rapid function variations, while larger values enforce smoother behaviour.

$$k_{\text{RBF}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{D}\frac{(x_d - x'_d)^2}{\ell_d^2}\right).$$

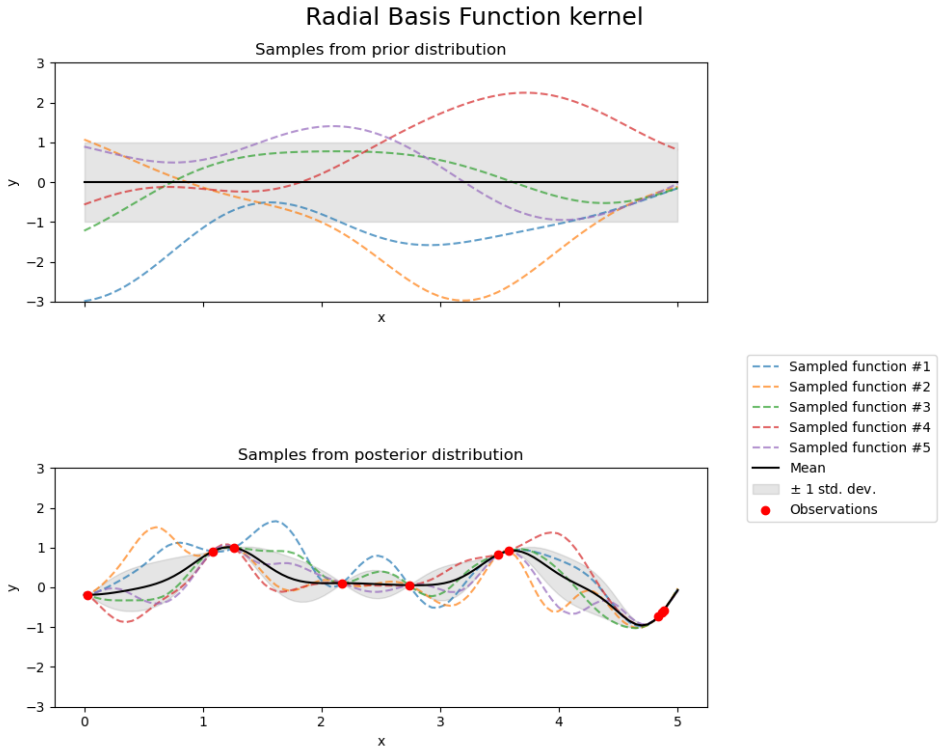


Figure 3.2: RBF Kernel Function [38]

The image for the RBF kernel in Figure 3.2 illustrates how this kernel, with its assumption of infinite differentiability, generates smooth prior samples in the top panel, reflecting a GP's ability to model continuous and predictable wind speed variations. The bottom panel demonstrates the posterior distribution, where observed data points (red circles) refine the mean function (black line) and uncertainty (shaded regions), showcasing the kernel's strength in capturing long-range correlations and stable trends, which are critical for forecasting consistent wind power output in offshore environments.

Matérn kernel with $\nu = 3/2$:

The Matérn family of kernels introduces a smoothness parameter ν that controls the differentiability of the functions. For $\nu = 3/2$, the resulting sample paths are once mean-square differentiable, producing rougher variations than the infinitely smooth RBF kernel. This makes the Matérn-3/2 kernel particularly suitable for modelling physical processes such as wind dynamics, where moderate irregularities are common. The kernel depends on the Euclidean distance $r = \|\mathbf{x} - \mathbf{x}'\|$ and an isotropic length scale ℓ, with larger ℓ enforcing smoother behaviour. As $\nu \to \infty$, the Matérn kernel converges to the RBF kernel, highlighting its role as a more flexible generalisation.

$$k_{\text{Matérn-3/2}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \left(1 + \frac{\sqrt{3}r}{\ell}\right) \exp\left(-\frac{\sqrt{3}r}{\ell}\right).$$

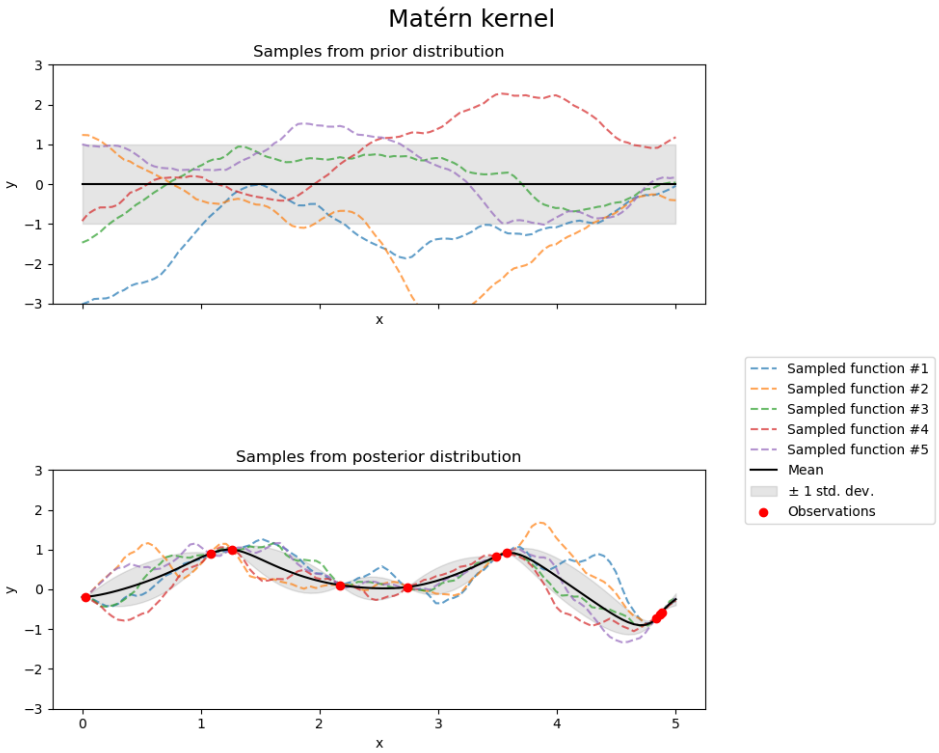


Figure 3.3: Matérn Kernel Function with $\nu = 3/2$ [38]

The Matérn kernel image in Figure 3.3 highlights its once-differentiable nature ($\nu = 3/2$) through rougher prior samples in the top panel, enabling the GP to represent the irregular, turbulent wind dynamics common in offshore settings. The bottom panel reveals the posterior distribution, where observed data (red circles) adjust the mean function (black line) and uncertainty (shaded regions), illustrating the kernel's flexibility in adapting to short-term fluctuations and enhancing forecast accuracy for volatile wind conditions.

White Noise Kernel:

The white noise kernel models independent, uncorrelated noise by contributing variance only when inputs are identical. In GPR, it is typically added to other kernels to account for measurement error or inherent stochasticity in observations. Here, σ_n^2 denotes the noise variance, and $\delta(\mathbf{x}, \mathbf{x}')$ equals 1 if $\mathbf{x} = \mathbf{x}'$ and 0 otherwise. This kernel affects only the diagonal entries of the covariance matrix, leaving different inputs uncorrelated:

$$k_{\text{white}}(\mathbf{x}, \mathbf{x}') = \sigma_n^2 \, \delta(\mathbf{x}, \mathbf{x}').$$

Kernels can also be combined to express more complex prior beliefs:

- **Addition:** $k_1 + k_2$ models a function as the sum of two independent components, each with distinct properties.
- **Multiplication:** $k_1 \times k_2$ enforces properties of both kernels simultaneously.

For example, the composite kernel

$$k_{\text{RBF}} + k_{\text{Matérn}} + k_{\text{white}}$$

represents a function as the combination of a smooth long-range trend (RBF), a moderately rough component (Matérn-3/2), and additive noise (white kernel), which is well-suited for capturing both persistent patterns and volatile fluctuations.

3.3 Transfer Learning

TL is a ML paradigm that leverages knowledge from a source domain ($\mathcal{D}_S$), typically with abundant data, to improve predictive performance in a related but data-scarce target domain ($\mathcal{D}_T$). In offshore WPF, TL can utilise rich meteorological and power output data from established wind farms to enhance forecasting accuracy at new or data-limited sites.

A task is defined as $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$, where $\mathcal{Y}$ is the output space and $f : \mathcal{X} \to \mathcal{Y}$ is the predictive function to be learned. TL is applicable in two scenarios: (i) domain adaptation, where the domains differ ($\mathcal{D}_S \neq \mathcal{D}_T$) due to distinct feature distributions (e.g., differing meteorological conditions), and (ii) task transfer, where the tasks differ ($\mathcal{T}_S \neq \mathcal{T}_T$) due to distinct predictive objectives (e.g., forecasting wind speed vs. power output). My methodology focuses on domain adaptation, where the task of forecasting wind power output remains consistent, but the feature distributions differ between the source and target wind farms. In my experiments, I investigate how forecasting accuracy evolves with incremental increases in target-domain data. In doing so, I assess the effectiveness of three TL methodologies for improving cross-domain model generalisation in offshore WPF and quantify the minimum data requirements for acceptable performance at new offshore sites.

3.3.1 TL Strategies

While TL strategies can be applied to both domain adaptation and task transfer, I focus on their application to domain adaptation. Three main strategies are commonly employed, each leveraging source domain knowledge differently.

Feature-based Transfer:
This approach learns a shared feature representation, denoted by $\phi : \mathcal{X} \to \mathcal{Z}$, which maps inputs from both source and target domains into a common latent space. The predictive function is denoted by f, while $\mathcal{L}_{\text{task}}$ represents the task-specific prediction loss evaluated on the target domain. To encourage transferability, a divergence measure $d(\cdot, \cdot)$ (e.g., Maximum Mean Discrepancy) is often introduced to quantify the discrepancy between source and target feature distributions. A trade-off parameter λ controls the balance between minimising the task loss and reducing distributional differences. Formally, the optimisation problem is expressed as:

$$\min_{\phi\in\Phi, f\in\mathcal{F}} \mathcal{L}_{\text{task}}(f, \mathcal{D}_T) + \lambda \cdot d\big(\phi(\mathcal{X}_S), \phi(\mathcal{X}_T)\big).$$

A central class of feature-based transfer methods focuses on *distribution alignment*, where the objective is to explicitly reduce discrepancies between the feature distributions of the source and target domains, thereby enabling models trained on the source domain to generalise more effectively to the target domain. In offshore WPF, such discrepancies arise naturally from differences in meteorological conditions or site-specific characteristics.

Distribution alignment can be achieved through techniques such as *moment matching*, which aligns the statistical moments (e.g., mean and variance) of the source and target distributions, or through more flexible kernel-based approaches such as Maximum Mean Discrepancy, which measures distributional differences in a reproducing kernel Hilbert space. By enforcing similarity between transformed feature distributions in the shared latent space, these methods encourage the learning of domain-invariant representations, thereby improving the transferability of predictive models. This is particularly important in offshore WPF, where variations in wind speed distributions, turbulence patterns, or topographical influences across sites can otherwise hinder model performance.

Parameter-based Transfer:
This strategy initialises the target model parameters $\boldsymbol{\theta}_T$ using parameters $\boldsymbol{\theta}_S^*$ pre-trained on the source domain, followed by fine-tuning on the target data:

$$\boldsymbol{\theta}_T^{(0)} = \boldsymbol{\theta}_S^*.$$

Fine-tuning adapts the model to target-specific patterns while retaining useful knowledge learned from the source. In WPF, parameter-based transfer is useful for transferring neural network weights or Gaussian Process hyperparameters trained on data-rich onshore wind farms to offshore sites with similar meteorological patterns, provided fine-tuning is applied to adapt to domain-specific differences.

Instance-based Transfer:
This method reweighs source domain samples based on their relevance to the target domain, effectively prioritising more transferable data. Let f denote the predictive model, and let $\ell(\cdot,\cdot)$ represent the loss function. Each source sample $\mathbf{x}_i^S$ with label y_i^S is assigned a weight w_i, which reflects its importance relative to the target domain. The target domain samples $\mathbf{x}_j^T$ with labels y_j^T are used without weighting. The weighted objective function is therefore expressed as:

$$\mathcal{L}_{\text{weighted}} = \sum_{i=1}^{n_S} w_i\, \ell\big(f(\mathbf{x}_i^S), y_i^S\big) + \sum_{j=1}^{n_T} \ell\big(f(\mathbf{x}_j^T), y_j^T\big).$$

Here, w_i is determined using strategies such as adversarial re-weighting or similarity-based scoring, ensuring that source data with meteorological or topographical similarity to the target domain is prioritised. This approach is particularly effective in offshore WPF when source and target sites share similar wind regimes but differ in local conditions.

Overall, TL provides a principled way to exploit existing offshore wind datasets, reduce data requirements for new sites, and improve generalisation of forecasting models in heterogeneous wind environments.

Chapter 4

Methodology

4.1 Dataset

This study utilises the comprehensive offshore wind farm dataset introduced by Grothe et al. [39], which provides 40 years of hourly wind speeds and synthetic power production signals for 29 major European offshore wind farms from 1980 to 2019. The dataset combines ERA5 reanalysis data with turbine-specific power curves to generate realistic production time series at hourly time intervals designed for research purposes.

ERA5 is a global atmospheric reanalysis dataset produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). It offers hourly estimates of a large number of atmospheric, land and oceanic climate variables globally and is widely used in climate research and applications due to its high temporal and spatial resolution [40].

Each wind farm dataset includes:

Features	**Units**
Power Output	MW
Wind Speed	m/s
Wind Speed at Hub Height	m/s
Wind Direction	Degrees from true north
fsr (Sea Surface Roughness)	meters
u100 (Horizontal Wind Speed Component)	m/s
v100 (Vertical Wind Speed Component)	m/s
Time	Hourly (UTC), format dd/mm/yyyy hh:mm

Table 4.1: Dataset Variables

Aligning with the goal of building a framework that does not depend on complex or costly datasets, these features are more easily attainable for new offshore wind farms, avoiding the need for expensive data collection.

For this study, wind farms were selected to evaluate the TL framework across different meteorological regimes. The goal was to test whether models trained under relatively calmer wind conditions could generalise to sites with stronger winds and greater variability. A deterministic approach was used, choosing the source wind farm with the lowest

average wind speed to assess knowledge transfer from a less energetic environment to more challenging conditions. Target sites were selected to provide a meaningful contrast: one with consistently stronger winds and frequent storms, and another with intermediate conditions influenced by a different meteorological setting. This setup allows evaluation of the TL framework's ability to handle substantial differences in climate and operating conditions. This deterministic method ensures clear contrasts but does not examine how models trained on high-wind sites transfer to lower-wind sites, highlighting a potential limitation and opening discussion on generalisability.
The wind farms selected for the analysis are:

1. **Baie de Saint-Brieuc** (*English Channel*): Located 16.3 km off the coast of Brittany, the *Baie de Saint-Brieuc* offshore wind farm comprises 62 Siemens Gamesa SG 8.0-167 turbines, each spaced approximately 1km apart, covering an area of $75\,\mathrm{km}^2$ and providing a total installed capacity of **496 MW** [41]. Influenced by both continental and Atlantic weather systems, the region experiences a highly dynamic wind climate characterised by significant inter-annual and spatial variability in wind speed and direction. [42] This complex atmospheric setting makes the site an ideal source wind farm for evaluating offshore WPF methods aimed at generalising across diverse meteorological environments.

2. **Baltic Eagle** (*Baltic Sea*): Situated 30 km North-East of Rugen Island, the *Baltic Eagle* offshore wind farm comprises 50 Vestas V174-9.5 turbines, delivering a total installed capacity of **476 MW**. [43] The broader Baltic Sea region forms a semi-enclosed basin, producing distinct meteorological conditions shaped by land-sea interactions, regional atmospheric circulation and the North Atlantic Oscillation. The area east of Rugen Island - including the Baltic Eagle site - is subject to some of the Baltic's strongest winds and highest storm frequencies, with more than twice the severe wind rate and over three times the storm freuqnecy compared to other Baltic coastal areas [44]. These complex and often extreme atmospheric dynamics provide both high energy potential and forecasting challenges, making Baltic Eagle an ideal test case for evaluating the TL methodologies.

3. **Beatrice** (*North Sea*): Located approximately 13km off the Caithness coast in the North Sea, the *Beatrice* offshore wind farm comprises 84 Siemens Gamesa 7MW turbines, yielding a total installed capacity of **588MW** [45]. Positioned in the Outer Moray Firth, the site is subjected to a dynamic and extreme marine climate, driven by its proximity to the continental shelf and exposure to prevailing westerly winds. These conditions lead to the site experiencing substantial wind speed variability, with more fluctuations in autumn and winter, and is influenced by frequent frontal systems and storm events [46]. Consequently, Beatrice serves as an optimal site for evaluating the TL methodology in an extreme weather environment.

Baie de Saint-Brieuc was selected as the source domain because it represents the site with the lowest average wind speeds across the three farms. Training the source models on a calmer and less energetic regime creates a deliberately challenging scenario for TL, where the objective is to generalise from limited and less extreme conditions to more demanding environments. The TL methodology was then applied to two target domains: Beatrice, representing a consistently high-wind, storm-exposed environment in the North Sea, and Baltic Eagle, which offers intermediate conditions within the Baltic Sea.

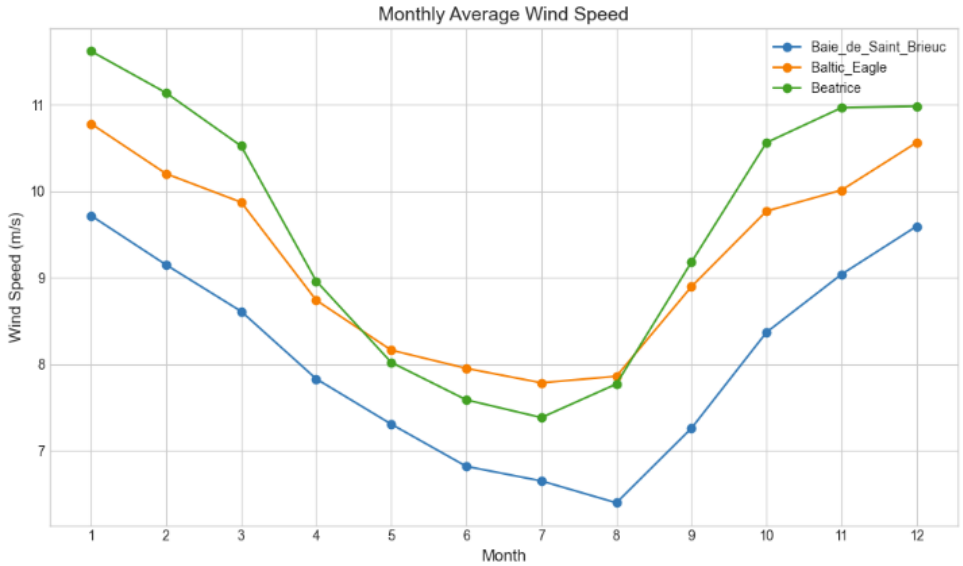


Figure 4.1: Average monthly wind speed for each wind farm over a 40 year period

4.2 Framework

This chapter presents the methodology for a novel framework for one-hour-ahead offshore WPF, along with three TL strategies designed to enhance cross-domain generalisability. The framework addresses two key challenges in the field. The first is improving one-hour-ahead forecasting performance without relying on large, complex and computationally intensive DL models, while providing reliable uncertainty quantification. The second is enabling more efficient deployment of models to new offshore wind farms by reducing the dependence on extensive site-specific datasets.

The methodology integrates three techniques: VAEs for weather pattern representation, GPs for probabilistic forecasting with uncertainty quantification and TL for cross-domain model adaptation. This combination leverages the representational power of VAEs to identify distinct meteorological regime, the flexibility of GPs to model complex non-linear relationships, and the generalisation capabilities of TL to extend models across different offshore locations.

The proposed framework operates through three interconnected stages:

- Weather Pattern Discovery: Raw meteorological time series are segmented into fixed-length time-periods, T, and encoded into low-dimensional representations using VAEs, then clustered to identify distinct weather regimes.
- Cluster-Specific Modelling: Individual GPs are trained on each weather cluster, enabling specialised forecasting for different meteorological conditions.
- TL Evaluation: Three TL strategies are assessed for their effectiveness in adapting trained models to new offshore wind farm locations with varying sizes off training sets.

Figure 4.2 illustrates the GP-VAE workflow and the relationships between these components

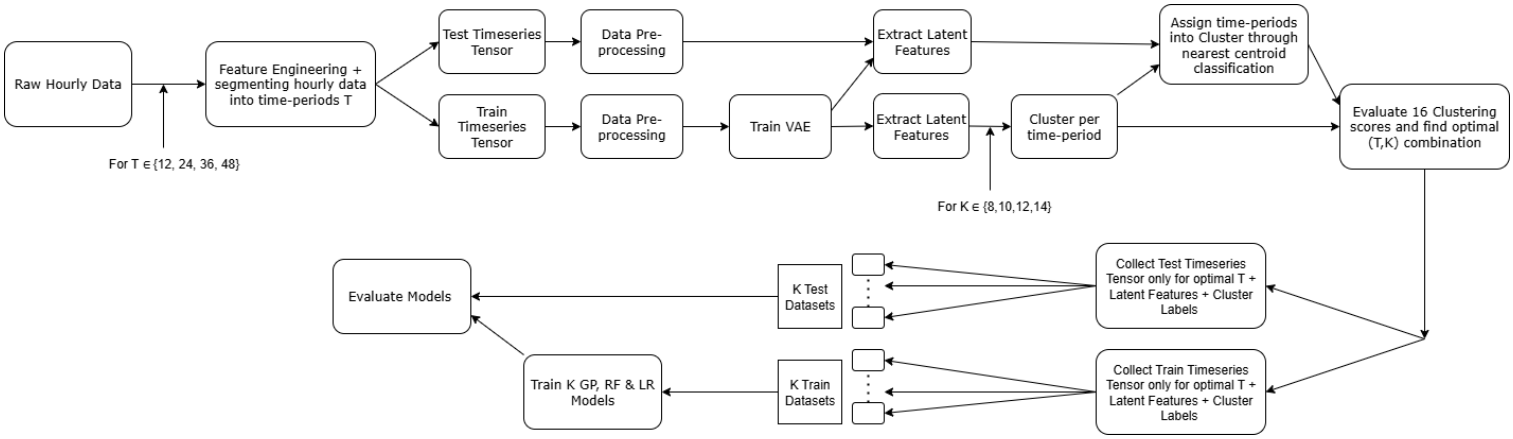


Figure 4.2: Workflow for the GP-VAE training methodology

The systematic evaluation of time-period lengths $T \in \{12, 24, 36, 48\}$ hours and cluster numbers $K \in \{8, 10, 12, 14\}$ creates a comprehensive search space of 16 combinations, ensuring optimal selection through rigorous validation.

4.2.1 Data Preprocessing & Feature Engineering

Temporal Segmentation Strategy

The foundation of the framework lies in segmenting continuous hourly meteorological data into fixed-length time-periods p_i. This segmentation serves dual purposes: providing sufficient temporal context for VAEs to capture meaningful weather patterns, and establishing natural boundaries for subsequent cluster assignment. Each time-period p_i is represented as a Matrix $\mathbf{X}_i \in \mathbb{R}^{T \times F}$, where F denotes the number of meteorological features and T represents the period length.

To determine the optimal temporal resolution to cluster weather patterns, four period length are systematically evaluated: $T \in \{12, 24, 36, 48\}$ hours. These choices span from half-day patterns that capture semi-diurnal cycles to two-day patterns that encompass longer-term meteorological transitions.

Train-Test Split, and VAE Feature Selection

Time-periods, $\mathbf{X}i$, are split into training and testing sets using a random period-based approach that maintains temporal integrity within each T-hour period. The splitting process begins by randomly shuffling all period indices using a random seed for reproducibility. These shuffled indices are then divided at a predetermined split point, with 80% allocated to training and 20% to testing. Crucially, entire time-periods are assigned to either the training or testing set, ensuring that all hours within a given T-hour period remain together in the same partition. This approach preserves the continuity of meteorological patterns within each period. For VAE training, each time-period is represented using a set of five meteorological features: wind speed and wind direction, which are the primary drivers of power generation; the horizontal and vertical wind components (u_{100}, v_{100}), which provide a vector representation of wind flow; and sea surface roughness (fsr).

Treatment of Cyclical Variables:

The VAE features are employed to cluster time-periods into distinct weather regimes, while the full dataset, including cyclical variables such as wind direction, hour of day, and month, are used to construct the training samples. These cyclical variables present challenges due to their circular nature, where values at boundaries (e.g., 0° and 359° for wind direction) represent similar conditions but appear distant numerically. To preserve this continuity, cyclical variables are transformed using sine-cosine encoding where θ represents the cyclical variable and $\theta_{\max}$ its maximum value (23 for hours, 359 for wind direction, 12 for months):

$$\theta_{\sin} = \sin\left(\frac{2\pi \cdot \theta}{\theta_{\max}}\right), \quad \theta_{\cos} = \cos\left(\frac{2\pi \cdot \theta}{\theta_{\max}}\right).$$

Normalisation:

All non-cyclical features are normalised to provide a consistent scale across variables. Sea Surface Roughness exhibits pronounced positive skew which can bias learning if left uncorrected. To address this distributional asymmetry, a logarithmic transinformation is first applied:

$$\text{fsr}_{\log} = \log(\text{fsr}).$$

Subsequently, all features undergo MinMax normalisation to the interval $[-1, 1]$:

$$x_{\text{norm}} = 2 \cdot \frac{x - x_{\min}}{x_{\max} - x_{\min}} - 1,$$

where scaling parameters are computed exclusively from the training set to avoid data leakage.

4.2.2 Weather Pattern Clustering Framework

Variational Autoencoder Implementation

The VAE provides the foundation for weather pattern clustering, compressing high-dimensional meteorological time-series into compact latent representation suitable for clustering analysis. For each time-period length T, a dedicated VAE transforms the $5 \times T$ dimensional input into an 8-dimensional latent space that captures essential meteorological characteristics while filtering noise and redundant information.

VAE Architecture

The encoder employs a 1D convolutional architecture specifically designed to capture temporal dependencies within weather sequences:

Input Layer: Accepts sequences of shape (B,C,T) where B represents batch size, C denotes channels (5 VAE features), and T indicates time-period length.

Convolutional Progression: Three 1D convolutional layers progressively expand feature channels ($5 \to 64 \to 128 \to 256$) while maintaining temporal resolution through appropriate padding. Each convolution is followed by batch normalisation and ReLU

activation to improve gradient flow and introduce non-linearity.

Temporal Aggregation: An adaptive average pooling layer fixes the sequence length to 8 time-steps regardless of input period T, enabling consistent latent dimensionality across different temporal configurations.

Latent Projection: Fully connected layers map the flattened representation to latent parameters $\mu_\phi(x)$ and $\log \sigma^2_\phi(x)$, defining the mean and log variance of the latent Gaussian distribution.

The decoder mirrors this structure with upsampling and transposed convolution operations to reconstruct the original input sequence.

Layer / Block	**Output Shape**	**Notes**
Input	(B, C, T)	B: batch size, C: channels, T: timeperiod
Encoder:		
Conv1D ($C \to 64, k = 5, \text{pad} = 2$)	$(B, 64, T)$	BatchNorm, ReLU
Conv1D ($64 \to 128, k = 3, \text{pad} = 1$)	$(B, 128, T)$	BatchNorm, ReLU
Conv1D ($128 \to 256, k = 3, \text{pad} = 1$)	$(B, 256, T)$	BatchNorm, ReLU
AdaptiveAvgPool1d(8)	$(B, 256, 8)$	Temporal length fixed to 8
Flatten & Fully Connected	$(B, 64)$	Dense projection
FC $\to \mu$, FC $\to \log \sigma^2$	$(B, 8)$	Latent mean & log variance
Decoder:		
FC layers	$(B, 256 \times 8)$	Unflatten to $(B, 256, 8)$
Upsample(T) + Conv1D blocks	(B, C, T)	Final Conv1D outputs raw reconstruction

Table 4.2: Layer-wise architecture of the TimeSeriesVAE.

VAE Training

As discussed in Chapter 3, VAE training optimises the Evidence Lower Bound (ELBO) objective, balancing reconstruction loss with latent space regularisation. To prevent posterior collapse, where the model ignores latent variables in favour of powerful decoder reconstruction, a progressive (β)-annealing schedule is implemented. Training begins with low KL weight to prioritise reconstruction accuracy, then gradually increases (β) to encourage informative latent representations.

Training hyperparameters:

- Epochs: 200
- Batch size: 32
- Optimizer: Adam (learning rate=1×10^{-4}, weight decay=1×10^{-5})
- (β)-annealing: Progressive schedule from 0.01 to 0.2

Hierarchical Clustering in Latent Space

Following VAE training, the encoder mean predictions $\mu_\phi(x)$ provide 8-dimensional representations of each time-period $\mathbf{X}i$. These latent features, having undergone dimensionality reduction while preserving essential meteorological patterns, form the basis for weather pattern identification.

Agglomerative hierarchical clustering with Ward linkage partitions the latent space into K weather clusters. Ward's method minimises within-cluster variance by defining the distance between clusters $\mathcal{C}_i$ and $\mathcal{C}_j$, where N_i and N_j represent cluster sizes, and $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ denote cluster centroids:

$$d(\mathcal{C}_i, \mathcal{C}_j) = \frac{N_i\, N_j}{N_i + N_j} \left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|^2 .$$

The systematic evaluation of $K \in \{8, 10, 12, 14\}$ clusters across $T \in \{12, 24, 36, 48\}$ hour periods generates 16 different combinations, each representing an alternative hypothesis about the optimal weather pattern segmentation.

4.2.3 Time-period and Cluster Size Optimisation

Selecting the optimal (T, K) combination requires balancing multiple competing objectives: clustering quality, meteorological coherence and practical viability for GP training. A comprehensive evaluation framework employs four complementary metrics:

- **Silhouette Score** measures how well each data point fits within its assigned cluster compared to its nearest neighbouring cluster. For each point i, let $a(i)$ represent the average distance to all other points within the same cluster. To find $b(i)$, the distance from point i to all points in every other cluster is calculated and averaged per cluster, then the minimum of these averages is taken, representing the average distance to the nearest neighbouring cluster [47],

  $$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}.$$

  Values are constrained to the range $(-1, 1)$ because the numerator $(b(i) - a(i))$ is always divided by the larger of $a(i)$ or $b(i)$. Scores near $+1$ indicate the point is much closer to its own cluster than to any other cluster, scores near 0 suggest the point lies between clusters with similar distances to both, and scores near -1 indicate the point is much closer to another cluster and is likely misclassified.

- **Davies-Bouldin Index** evaluates clustering quality by measuring how compact and separated clusters are, comparing each cluster to its most similar neighbour. For cluster k with centroid c_k and N_k points, let

  $$\sigma_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \|x_i - c_k\|$$

be the average distance from all points within a cluster to the centroid, representing the cluster's spread (compactness). The Davies-Bouldin index is then defined as:

$$DB = \frac{1}{K} \sum_{k=1}^{K} \max_{j \neq k} \left\{ \frac{\sigma_k + \sigma_j}{\|c_k - c_j\|} \right\},$$

[48] where K is the number of clusters, and $\|c_k - c_j\|$ is the Euclidean distance between centroids, measuring separation between clusters. Here, j represents each of the other clusters, and the maximum selects the cluster with the highest ratio, indicating the most similar neighbour (i.e. the cluster with the largest combined spread $\sigma_k + \sigma_j$ or smallest distance relative to cluster k). The numerator $(\sigma_k + \sigma_j)$ sums the spread of clusters k and j, indicating how scattered their points are. The denominator $\|c_k - c_j\|$ shows how far apart the clusters are. A lower ratio means compact clusters (small $(\sigma_k + \sigma_j)$) that are well-separated (large $\|c_k - c_j\|$), indicating good clustering. Thus, lower DB values reflect better clustering quality [49]. Since lower values are better but other metrics in the evaluation use higher-is-better conventions, the DB score is transformed using

$$\mathrm{DB}_{\mathrm{scaled}} = \frac{2.0 - DB}{2.0}$$

to reverse the scale while maintaining interpretability.

- **Temporal Coherence** assesses whether clusters exhibit consistent patterns in the original meteorological space. For each cluster k containing N_k time-periods, $\mathbf{X}_i \in \mathbb{R}^{T \times F}$ denote the matrix of meteorological features for the i-th time-period in the cluster. Temporal coherence is then computed as the average pairwise Pearson correlation between these matrices:

$$C_k = \frac{2}{N_k(N_k - 1)} \sum_{i=1}^{N_k} \sum_{j=i+1}^{N_k} \rho(\mathbf{x}_i, \mathbf{x}_j).$$

  This ensures that clusters represent meteorologically meaningful weather patterns rather than arbitrary latent space groupings.

- **Viability Metrics** enforce practical requirements for subsequent GP training. Coverage measures the fraction of clusters containing at least $T_{\min} = 15$ training time-periods and at least $S_{\min} = 5$ test time-periods. With k number of clusters, N_k^{train} the number of training periods in cluster k, and N_k^{test} the number of test periods. Using the indicator function $\mathbf{1}\{\cdot\}$, coverage is defined as:

$$\mathrm{Coverage} = \frac{\sum_{k=1}^{K} \mathbf{1}\{n_k^{\mathrm{train}} \geq T_{\min} \text{ and } n_k^{\mathrm{test}} \geq S_{\min}\}}{\sum_{k=1}^{K} \mathbf{1}\{n_k^{\mathrm{train}} \geq T_{\min}\}}.$$

  Values closer to 1 indicate higher practical viability.

The final clustering score integrates clustering quality, temporal coherence, and practical viability using carefully chosen weights:

$$\begin{aligned}\textbf{Composite Score} = (0.30 \times \text{Silhouette}) + (0.30 \times \text{Normalised DB}) \\ + (0.20 \times \text{Temporal Coherence}) + (0.20 \times \text{Coverage})\end{aligned}$$

This weighting prioritises cluster quality (60%) while ensuring meteorological meaningfulness and sufficient training/test coverage (40%).

4.2.4 Dataset Creation

Following the selection of the optimal combination of time-period T and cluster size K, K cluster-specific datasets are constructed by combining VAE latent representations with temporal and cyclical features.

Sliding Window Augmentation

Lagged features for the preceding one and two hours are precomputed for each hour of the dataset before segmenting into time-periods. As a result, the first two hours of the dataset are discarded since their lagged features are undefined. For subsequent hours, all time-periods have access to valid lagged information, allowing each T-hour time-period to generate T training samples without losing any hours.

For a time-period, $\mathbf{X}_i$, containing hours $\{0, 1, \ldots, T-1\}$, the samples, $\mathbf{x}_i \in \mathbb{R}^F$, are generated as follows:

Sample 1: Lagged features from hours -2 and -1 predict power at hour 0
Sample 2: Lagged features from hours -1 and 0 predict power at hour 1
$\vdots$
Sample T: Lagged features from hours T-3 and T-2 predict power at hour T-1.

This approach preserves temporal continuity within each period and maximises the effective dataset size for Gaussian Process training while maintaining causal consistency.

Training Sample Representation

Each resulting training sample $\mathbf{x}_t \in \mathbb{R}^{20}$ at hour t consists of three distinct feature groups, giving a total of 20 features.

The first group comprises the 8-dimensional VAE latent features, which remain constant across all hours within a time-period and capture a compressed representation of the prevailing meteorological regime. The second group contains 6 local temporal features, including lagged power output (P_{t-1}, P_{t-2}), lagged wind speed (WS_{t-1}, WS_{t-2}), and current weather conditions (WS_t, fsr_t), providing short-term dynamics. The final group consists of 6 cyclical features that encode temporal periodicities as sine-cosine pairs, covering hour-of-day, month, and wind direction.

4.2.5 Gaussian Process Implementation

Following the establishment of the theoretical foundation of GPR in Section 4.2 and the construction of cluster-specific datasets $\{\mathcal{D}_k\}_{k=1}^{K}$, the implementation of cluster-specific GPs trained on the source wind farm data is detailed. This approach enables each GP to specialise in the power dynamics characteristic of its assigned meteorological pattern, with separate models capturing the distinct input-output relationships that emerge under different weather regimes.

Kernel Architecture and Hyperparameter Optimisation

Each weather cluster employs a specialised GP configured with a composite kernel introduced in Section 3.2:

$$k(\mathbf{x}, \mathbf{x}') = k_{\text{RBF}}(\mathbf{x}, \mathbf{x}') + k_{\text{Matérn-3/2}}(\mathbf{x}, \mathbf{x}') + k_{\text{white}}(\mathbf{x}, \mathbf{x}')$$

Feature importance varies across weather regimes, necessitating adaptive feature weighting. Automatic Relevance Determination (ARD) assigns independent length scales ℓ_d to each of the 20 input dimensions for both the RBF and Matérn kernels, transforming isotropic kernels into anisotropic forms. For example, the RBF component is given by:

$$k_{\text{RBF}}(x, x') = \sigma_f^2 \exp\left(-\tfrac{1}{2}\sum_{d=1}^{20}\frac{(x_d - x'_d)^2}{\ell_d^2}\right).$$

Similarly, the Matérn-3/2 component uses its own set of length scales $\{\ell_d^{\text{Matérn}}\}_{d=1}^{20}$. Small length scales ($\ell_d \to 0$) indicate high relevance, requiring fine-grained modelling, while large length scales ($\ell_d \to 100$) effectively reduce the influence of a feature. This automatic feature selection adapts to each cluster's specific input–output relationships without manual intervention.

The complete hyperparameter vector comprises 43 parameters:

$$\boldsymbol{\theta} = \left\{\sigma_f^{\text{RBF}}, \{\ell_d^{\text{RBF}}\}_{d=1}^{20}, \sigma_f^{\text{Matérn}}, \{\ell_d^{\text{Matérn}}\}_{d=1}^{20}, \sigma_n^2\right\}.$$

Optimisation maximises the marginal likelihood using L-BFGS-B with ten random restarts to mitigate poor local optima in this high-dimensional parameter space. Length scales are constrained within $[0.1, 100]$ to prevent numerical instability, while noise variance is bounded in $[10^{-6}, 1]$.

Training and Inference

GP training constructs the covariance matrix K using optimised hyperparameters and computes its Cholesky decomposition L such that $K = LL^\top$. This decomposition, alongside the hyperparameters, enables efficient predictive inference for new test inputs.

For a test input x^* assigned to cluster k, the trained GP yields a predictive distribution $\mathcal{N}(\mu_*, \sigma_*^2)$ where:

$$\mu_* = k_*^\top K^{-1} y$$

$$\sigma_*^2 = k(x^*, x) - k_*^\top K^{-1} k_*.$$

Predictions are inverse-transformed to the original MW scale and constrained within the physical bounds $[0, 496]$ MW, corresponding to the total wind farm capacity. The predictive variance provides uncertainty quantification, appropriately expanding for extrapolated inputs or volatile weather conditions.

4.2.6 Baseline Models & Evaluation

To validate the effectiveness of the cluster-specific GP approach, two baseline models with contrasting characteristics are implemented: Random Forest (RF) representing an ensemble method, and Linear Regression (LR) providing a simple parametric baseline. Both models are trained on identical cluster-specific datasets $\{\mathcal{D}_k\}_{k=1}^{K}$ to ensure fair comparison, processing the same 20-dimensional input vectors.

Random Forest Regression

RF is an ensemble learning method that constructs multiple decision trees during training and outputs the mean prediction across all trees for a regression task. Each tree is trained on a bootstrap sample of the training data, and random feature subsets are considered at each split, introducing diversity that reduces over-fitting and improves generalisation.

For WPF, RF offers several advantages: it captures non-linear relationships without explicit feature engineering, naturally handles interactions between meteorological variables, and provides feature importance rankings through the mean decrease in impurity. The RF baseline is configured with 100 decision trees and a maximum depth of 15 levels. These were selected via cross-validation, balancing model complexity against the risk of over-fitting while ensuring robust performance across training folds.

The ensemble prediction averages individual tree outputs, where $B = 100$ is the number of trees and $f_b(\mathbf{x})$ represents the prediction from the b-th tree:

$$\hat{y}_{\text{RF}} = \frac{1}{B} \sum_{b=1}^{B} f_b(\mathbf{x}).$$

Unlike GPs, RF provides point predictions without inherent uncertainty quantification, though prediction intervals can be estimated through the variance across trees.

Linear Regression

LR serves as a parametric baseline, modelling power output as a linear combination of input features where $\mathbf{w} \in \mathbb{R}^{20}$ denotes the regression weights and w_0 the intercept term:

$$\hat{y}_{\text{LR}} = \mathbf{w}^T \mathbf{x} + w_0.$$

The parameters are estimated via ordinary least squares, minimising the sum of squared residuals:

$$(\mathbf{w}^*, w_0^*) = \arg\min_{\mathbf{w}, w_0} \sum_{i=1}^{n} \left(y_i - \mathbf{w}^T \mathbf{x}_i - w_0 \right)^2.$$

Although LR is limited in capturing the non-linear dynamics inherent in wind power generation, it provides a lower bound on expected performance and offers complete interpretability through its weights. Its simplicity also ensures computational efficiency and numerical stability, making it a robust baseline even with relatively small training datasets.

Evaluation Metrics

Model performance is assessed using three complementary metrics standard in the WPF literature:

- **Mean Absolute Error (MAE)** quantifies the average magnitude of prediction errors without considering direction:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

  providing an interpretable measure in MW.

- **Root Mean Square Error (RMSE)** penalises large errors more heavily through squaring:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

  also providing an interpretable measure in MW that emphasizes larger deviations.

- **Coefficient of Determination (R^2)** quantifies the proportion of variance in the observed data explained by the model. y_i denotes the observed values, $\hat{y}_i$ the predicted values, and $\bar{y}$ the mean of the observed values. Then R^2 is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  providing a normalised measure of model fit independent of scale, facilitating comparison across different wind farms and power ranges.

4.2.7 Transfer Learning Strategies

Following the establishment and optimisation of source GP models on data from Baie de Saint-Brieuc, three TL strategies were evaluated for adapting the framework to new offshore locations with limited data availability. The source wind farm, positioned in the English Channel off Brittany, experiences diverse meteorological conditions from both continental and Atlantic weather systems. This meteorological diversity ensures that the learned weather patterns encompass a broad range of conditions likely to be encountered at other offshore sites.

Transfer performance was evaluated on two contrasting target wind farms: Baltic Eagle in the southern Baltic Sea and Beatrice in the North Sea off Scotland. For each TL method, performance was incrementally evaluated using 10%, 20%, 30%, 40%, and 50% of the available target data for training, with the remainder reserved for testing.

This progressive evaluation reveals how quickly each method converges to source-level performance and identifies the minimum data requirements for operational deployment.

Method 1: Frozen VAE with GP Hyperparameter Transfer and Adaptation

The first TL strategy preserves the source VAEs learned representations by maintaining the encoder in a frozen state, transferring knowledge through initialising the cluster-specific target GP hyperparameters with values learned from the source domain. By freezing the VAE encoder, the learned latent space structure is preserved while allowing the GP models to adapt to target-specific feature-output relationships. This approach assumes that the meteorological patterns learned from the source domain are sufficiently general to represent weather dynamics at the target location.

Latent Features Extraction and Clustering:

The process begins by segmenting hourly target wind farm data into time-periods matching the source configuration. Each time-period undergoes identical preprocessing as the source data. The frozen source VAE encoder then extracts 8-dimensional latent representations, where ϕ_{source} represents the fixed encoder parameters learned from Baie de Saint-Brieuc and $\boldsymbol{\mu}_{\phi_{\text{source}}}(\mathbf{x})$ is the mean vector output by the encoder:

$$\mathbf{z}_{\text{target}}^{(i)} = \boldsymbol{\mu}_{\phi_{\text{source}}}\left(\mathbf{x}_{\text{target}}^{(i)}\right).$$

These representations maintain the semantic structure of the source latent space.

Target time-periods are assigned to source clusters through nearest centroid classification in the latent space. Centroids are computed for each source cluster in the latent space where C_k represents the set of source latent vectors assigned to cluster k:

$$\mathbf{z}_k^{\text{source}} = \frac{1}{|C_k|} \sum_{j \in C_k} \mathbf{z}_j^{\text{source}}.$$

Each target latent vector is then assigned to the cluster with the nearest centroid:

$$c_{\text{target}}^{(i)} = \arg \min_{k \in \{1,\ldots,K\}} \left\| \mathbf{z}_{\text{target}}^{(i)} - \mathbf{z}_k^{\text{source}} \right\|_2.$$

This assignment preserves the meteorological interpretation established during source training, with each cluster maintaining its characteristic weather pattern signature.

Dataset Construction and GP transfer:

Following cluster assignment, target datasets are constructed using the same sliding window approach. Each time-period generates multiple training samples, creating 20-dimensional feature vectors.

Rather than training new GPs from scratch, each target GP is initialised using the optimised hyperparameters from the corresponding source cluster:

$$\theta^{(0)}_{k,\text{target}} = \theta^{*}_{k,\text{source}}.$$

This transfer includes all 43 hyperparameters of the composite kernel. The transferred ARD length scales encode which meteorological features are most informative for each weather regime, knowledge that transfers effectively when meteorological patterns are consistent across sites.

GP Retraining and Prediction:

Starting from the transferred hyperparameter initialisation, each target GP undergoes retraining on target cluster data through marginal likelihood maximisation:

$$\boldsymbol{\theta}^{*}_{k,\text{target}} = \arg\max_{\boldsymbol{\theta}} \log p(\mathbf{y}_{k,\text{target}} \mid \mathbf{X}_{k,\text{target}}, \boldsymbol{\theta}).$$

The optimisation employs only 5 random restarts compared to 10 for source training, as the informed initialisation reduces the risk of convergence to poor local optima. This allows hyperparameters to fully adapt to target-specific feature-output relationships while benefiting from meteorologically informed starting values.

Method 2: Adaptive VAE with GP Hyperparameter Transfer and Adaptation

The second strategy extends Method 1 by allowing both the VAE encoder and GPs to adapt to the target domain while maintaining the fundamental cluster structure learned from the source. This approach acknowledges that while meteorological patterns are transferable across sites, the specific feature representations that best capture these patterns may benefit from target-specific refinement.

Pre-clustering with Source VAE:

The initial steps mirror those described in method 1: target time-periods are segmented and preprocessed identically to the source data. The frozen source VAE encoder extracts latent representations and are used for cluster assignment through nearest centroid classification, ensuring target time-periods are grouped according to the meteorological patterns identified in the source domain. This pre-determination of cluster assignments using source-space features prevents cluster drift that could arise from VAE adaptation.

VAE fine-tuning for Latent Feature Adaptation and Dataset Construction:

With cluster assignments fixed, the source VAE undergoes adaptation on the target training data. The adaptation process optimises both encoder parameters ϕ and decoder parameters θ using a β-VAE objective with reduced regularisation. The adaptation loss is:

$$\mathcal{L}_{\text{adaptation}} = \text{MSE}(\mathbf{x}_{\text{target}}, \hat{\mathbf{x}}_{\text{target}}) - 0.1 \cdot D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}_{\text{target}}) \,\|\, p(\mathbf{z})\right).$$

The reduced KL weight β allows greater flexibility for the encoder to adapt to target-specific patterns. Training proceeds for 50 epochs using the Adam optimizer with learning rate 10^{-4}.

Dataset Construction with Adapted Features

Following VAE adaptation, new latent representations are extracted using the fine-tuned encoder:

$$\mathbf{z}_{\text{target}}^{\text{adapted}} = \boldsymbol{\mu}_{\phi_{\text{adapted}}}(\mathbf{x}_{\text{target}}).$$

These adapted features are paired with the cluster assignments determined from the source-space features, not used for re-clustering. This ensures that while the feature representations evolve to capture target-specific patterns, each cluster maintains its fundamental meteorological interpretation established during source training. The sliding window approach then constructs cluster-specific datasets combining these adapted latent features with the temporal and lagged features as before.

The adaptation process introduces measurable distribution shifts in the latent space. Initial distribution shift between source and target domains in the source-space is quantified as the mean absolute difference between source and target feature centroids. Following adaptation, additional shifts in both mean and variance capture how the fine-tuning process modifies the features space to better represent target-domain patterns.

GP transfer:

The cluster-specific GPs undergo the same hyperparameter transfer and retraining process as in Method 1, initialising with source parameters and fine-tuning with 5 restarts.

Method 3: Domain Adaptation through Cluster-based Dataset alignment

The third strategy builds upon Method 2 by incorporating domain adaptation through selective dataset shifting while preserving cyclical features that represent inherent temporal patterns. This approach recognises that statistical alignment between source and target domains can improve transfer performance.

Initial Processing and Clustering:

The initial data preparation, latent feature extraction and cluster assignment follow the same methodology as in Method 1 and 2. Target data is segmented into time-periods, processed through the same preprocessing pipeline, and assigned to source clusters using the frozen source VAE encoder through nearest centroid classification in the latent space.

VAE Fine-tuning:

The VAE fine-tuning process mirrors that described in method 2, employing the same reduced KL divergence weight ($\beta = 0.1$) and training parameters (50 epochs, learning rate=10^{-4}), allowing target-specific adaptation while maintaining cluster structure.

Dataset Domain Shifting:

Following VAE adaptation, a selective dataset shifting procedure is applied to align statistical distributions between domains, allowing for the source gp models to more closely

represent the target datasets, while preserving the physical meaning of the temporal features within the dataset. This approach recognises that domain shift, where source and target data exhibit different statistical distributions, can severely impair GP performance even with fine-tuning.

In the dataset, meteorological variables such as wind speed and power output may benefit from alignment, however cyclical temporal features encode universal patterns and so no shift is applied here.

For feature-wise scaling, the transformation is applied:

$$\mathbf{X}^{(i)}_{\text{shifted}} = \begin{cases} \frac{(\mathbf{X}^{(i)}_{\text{target}} - \mu^{(i)}_{\text{target}})}{\sigma^{(i)}_{\text{target}}} \cdot \sigma^{(i)}_{\text{weighted}} + \mu^{(i)}_{\text{weighted}} & \text{if } i \notin \mathcal{C} \\ \mathbf{X}^{(i)}_{\text{target}} & \text{if } i \in \mathcal{C} \end{cases}$$

where $\mathcal{C}$ represents the set of cyclical feature indices and the weighted parameters provide conservative alignment:

$$\mu^{(i)}_{\text{weighted}} = (1 - w)\, \mu^{(i)}_{\text{target}} + w\, \mu^{(i)}_{\text{source}}$$

$$\sigma^{(i)}_{\text{weighted}} = (1 - w)\, \sigma^{(i)}_{\text{target}} + w\, \sigma^{(i)}_{\text{source}}$$

The shift weights ($w = 0.5$ for features, $w = 0.3$ for targets) are deliberately conservative, providing partial alignment rather than complete transformation. This reflects the hypothesis that moderate statistical alignment can improve GP performance without overly compromising the target domain's intrinsic characteristics.

Source Dataset Statistics and GP Training

The approach requires comprehensive characterisation of source domain statistics for each cluster, including cluster-specific means, variances, and covariance structures. This statistical profiling enables selective shifting towards appropriate reference distributions, while maintaining awareness of which features should be preserved. The rationale is that each meteorological cluster represents a distinct weather regime with characteristic feature relationships, and successful transfer requires alignment to the appropriate regime-specific statistics rather than global source statistics.

GP training then proceeds using the selectively shifted datasets, with the expectation that improved statistical alignment will enhance the transferred hyperparameters' effectiveness. The fundamental premise is that GPs trained on better-aligned data will exhibit superior predictive performance when the domain gap is reduced, while preservation of temporal structure ensures that learned weather pattern dynamics remain intact.

Prediction Inverse Transformation

The prediction workflow concludes by transforming GP outputs from the shifted space back to the original target domain:

$$\hat{y}_{\text{target}} = \frac{\hat{y}_{\text{shifted}} - \mu^{(y)}_{\text{shifted}}}{\sigma^{(y)}_{\text{shifted}}} \cdot \sigma^{(y)}_{\text{target}} + \mu^{(y)}_{\text{target}}.$$

This inverse transformation is essential for obtaining predictions in the correct target domain scale while retaining the predictive improvements gained through selective alignment. The approach thus aims to strike an optimal balance between leveraging source domain knowledge and respecting target domain characteristics.

Chapter 5

Results

To contextualise the performance of the proposed VAE- GP framework, it is necessary to first consider the results reported by Wang et al. [3] using the same offshore wind power dataset. Their Multi-location Multi-modal Multi-step Informer (M3STIN) model, published in March 2025, represents a recent benchmark for offshore WPF, employing a sophisticated DL architecture that combines Graph Attention Networks (GAT) for spatial dependency modelling with Informer modules for temporal feature extraction.

The M3STIN framework was trained on five years of historical data (2015-2019, comprising 43,800 hourly observations) from eight offshore wind farms simultaneously. An attribute graph was constructed to connect all sites with Pearson correlation coefficients greater than 0.9, while Gaussian kernel distance functions were used to determine edge weights. This design enables the GAT layers to capture spatial dependencies across the wind farm network, reflecting how meteorological patterns propagate between locations. For one-hour-ahead predictions, the best performing model proposed by Wang et al. achieved an average MAE of **3.60%** of power capacity and a coefficient of determination of $\boldsymbol{R^2 = 0.973}$ across all eight sites.

Model	RMSE	MAE	R^2
M3STIN	6.03	3.60	97.36
M3STIN[1]	6.17	3.70	97.23
M3STIN[2]	6.31	3.94	97.11
M3STIN[3]	6.29	3.86	97.13
M3STIN[4]	6.37	3.92	97.06
GAT-BiGRU	6.34	3.88	97.08
GAT-ALSTM	6.40	4.27	97.03
GCN-GRU	6.43	3.99	97.00
Informer	6.67	4.42	96.78
BiGRU	6.93	4.62	95.85
SVR	7.79	5.34	95.59

Table 5.1: Performance comparison of different models for 1-hour-ahead wind power forecasting proposed by Wang et al.

5.1 Source Model Performance

The VAE-GP framework was trained and evaluated on the Baie de Saint-Brieuc wind farm using datasets of one and two years to investigate the relationship between data availability, computational demand and model performance. This analysis established both the optimal source model configuration and demonstrates the framework's data efficiency relative to Wang et al.s DL approach. For model evaluation, the overall framework MAE is reported as a weighted average of the cluster-wise MAE values, where the weights correspond to the number of test samples in each cluster. To maintain consistency with Wang et al., the overall and cluster individual MAEs are often expressed as a percentage of the total wind farm capacity to account for differences in capacity levels.

5.1.1 One-year Training Results:

Training the VAE-GP framework on a single year of Baie de Saint-Brieuc data (2019, comprising 8,757 hourly observations) demonstrates the methodology's ability to achieve competitive performance with reduced data requirements. The optimal configuration identified through evaluation consists of **12 clusters** with **12 hour time-periods**, achieving a clustering quality score of 0.645. The individual power and weather patterns per cluster are shown in the Appendix as Figure A.1

The framework attains an average MAE of 21.66 MW (**4.37%** of capacity) with a R^2=**0.928**. Comparing with the DL baselines reported by Wang et al. in Table 5.1 reveals the proposed approach achieves superior MAE accuracy to several established methods, including the Informer model which achieved a MAE of 4.42%, while utilising 20% of the training data.

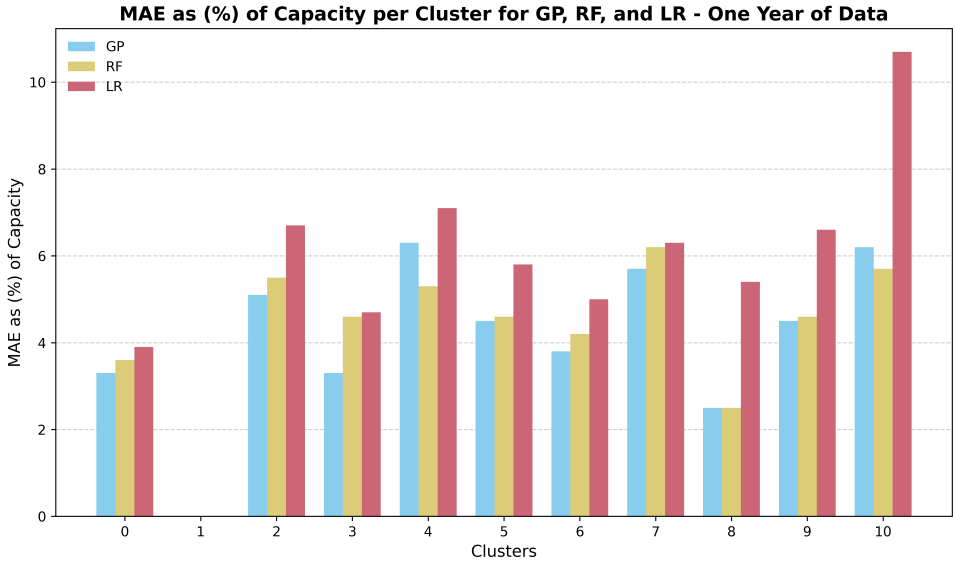


Figure 5.1: MAE of each Cluster as a % of total Capacity

Clusters 1 and 11 show no test results as no time-periods in the test set were assigned to these clusters during the evaluation period.

The cluster-specific results seen in Figure 5.1 reveals substantial performance variation across different meteorological conditions, validating the meteorological clustering approach. As shown in the cluster performance breakdown the GP model demonstrates heterogeneous accuracy across clusters, with MAE ranging from 2.46% to 6.25% of capacity. Notable performances include:

- **High-Performance Clusters (0, 3, 8):** Cluster 8 achieves exceptional accuracy with an MAE of 2.46%, largely due to its consistent maximum capacity power output, while Cluster 0 and Cluster 3 maintain MAEs of 3.33% and 3.35% respectively.

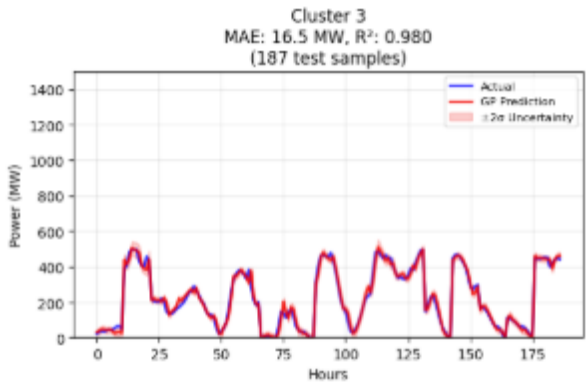


Figure 5.2: Cluster 3 forecasting performance

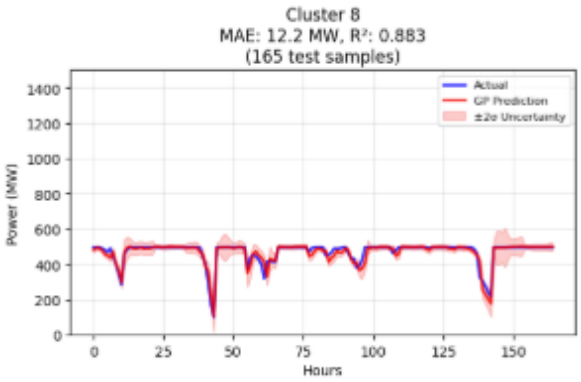


Figure 5.3: Cluster 8 forecasting performance

- **Moderate-Performance Clusters (5,6,9):** These clusters achieve MAEs of 4.48%, 3.75% and 4.48% respectively.

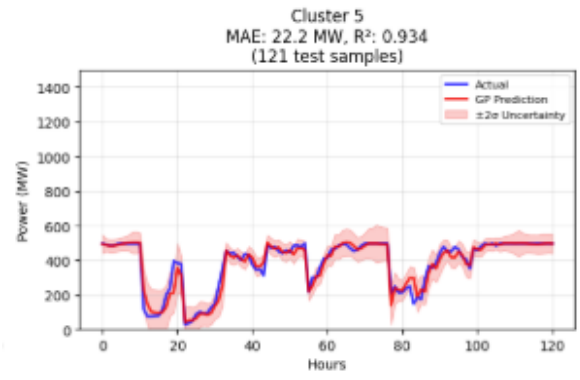


Figure 5.4: Cluster 5 forecasting performance

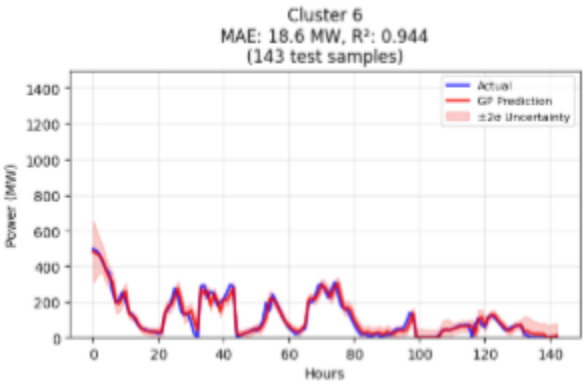


Figure 5.5: Cluster 6 forecasting performance

- **Challenging Clusters (2,4,7,10):** Cluster 4 exhibits the highest forecasting error with an MAE of 6.25%, yet achieves a relatively high R^2 of 0.939. This suggests that while the model captures the overall pattern of this cluster well, the inherent variability in this challenging weather regime makes precise predictions difficult, even with a relatively large number of training samples (616) relative to other clusters.

Cluster 10 follows closely with an MAE of 6.23% but has the fewest training samples of all clusters (187), suggesting that its performance and uncertainty bounds could improve with additional data. Clusters 2 and 7 have MAEs of 5.14% and 5.69% with 418 and 693 training samples respectively. Across the 10 clusters with testing data, the average number of training samples is 595.1.

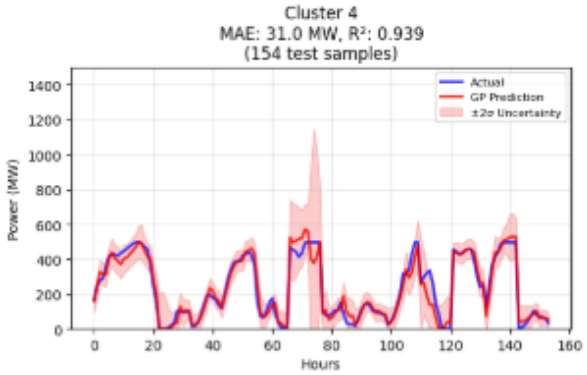


Figure 5.6: Cluster 4 forecasting performance

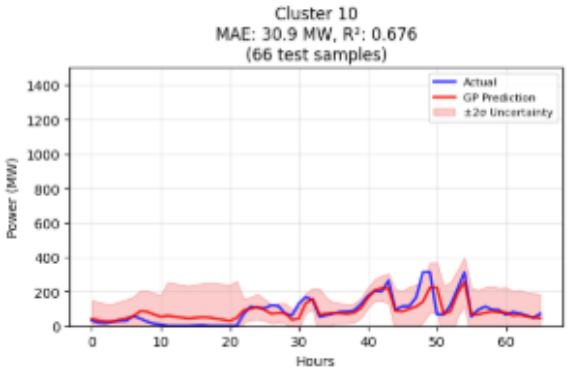


Figure 5.7: Cluster 10 forecasting performance

The per cluster-breakdown of training and testing samples as well as results for all models can be found in the Appendix as Table A.1 alongside the forecasting performance graphs per cluster Figure A.2.

The RF baseline achieves comparable overall performance with an MAE of 22.85 MW (**4.61%**) and R^2 of 0.910, with the GP models outperforming RF models in 7 out of 10 active clusters.
Linear Regression significantly underperforms across all clusters achieving an overall MAE of 29.1 MW (**5.86%**) and R^2: 0.863. The consistent superior performance of both GP and RF models over LR confirms the non-linear nature of wind power generation dynamics.

While training with one-year of data achieves promising results, the presence of substantial errors in clusters such as 4 and 10 highlights a limitation, motivating the extension of training with two years of data to enhance robustness and consistency across conditions.

5.1.2 Two-years Training Results:

Extending the training dataset to include 2018-2019 data enables more refined meteorological segmentation, with the optimal configuration shifting to **14 clusters** with **12-hour time-periods** (clustering quality score: 0.68). The additional year of data facilitates the identification of two additional distinct weather patterns that were previously merged in the one-year analysis. It is important to note that the new clusters in this run have no direct correspondence to those obtained when training with only one year of data. The individual power and weather patterns per cluster are shown in Figure A.3.

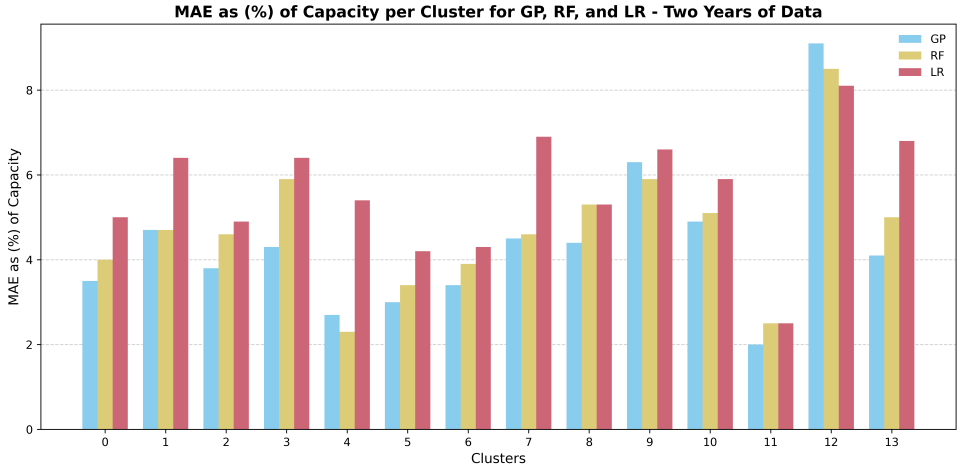


Figure 5.8: MAE of each Cluster as a % of total Capacity - Training on two years of data

The expanded dataset yields improved overall performance for the GP models, with MAE reducing to 20.5 MW (**4.14%** of capacity), representing a 5.36% improvement over the one-year analysis. However, the overall R^2 remains stable at **0.925**, primarily due to Cluster 12's anomalous performance which has an MAE of 45.2 MW (9.11%) and $R^2 = 0.765$.

The cluster-wise performance analysis in Figure 5.8 shows:

- **High-Performance Clusters (4,5,11):** These clusters achieve exceptional accuracy with MAEs of 2.72%, 3.02% and 2.04% respectively. Cluster 4 resembles cluster 8 from the one-year analysis. Cluster 5 achieves an $R^2 = 0.977$, as seen in Figure 5.10, which is particularly noteworthy given the high degree of volatility present in the data. Cluster 11 has the lowest number of training samples (253) yet its performance appears strong, likely due to the large proportion of zero-power values.

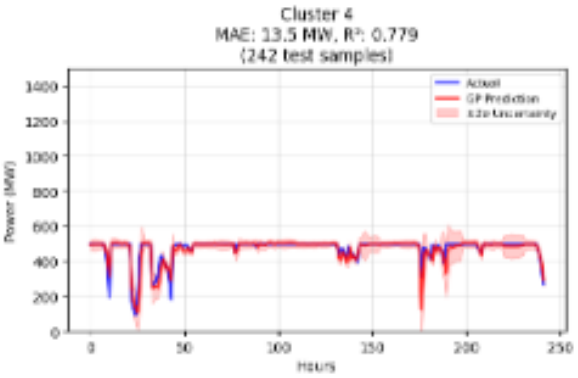


Figure 5.9: Cluster 4 forecasting performance

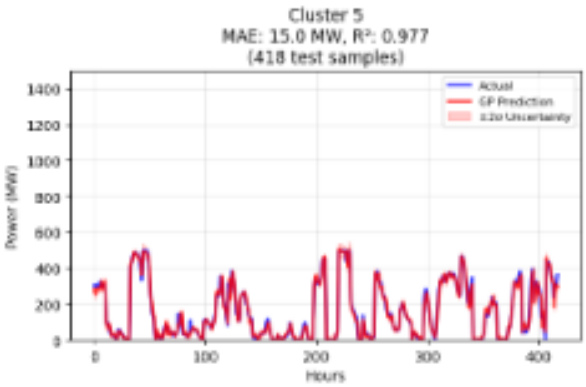


Figure 5.10: Cluster 5 forecasting performance

- **Moderate-Performance Clusters (0,1,2,3,6,7,8,10,13):** Performance from these clusters range from 3.43% to 4.86%

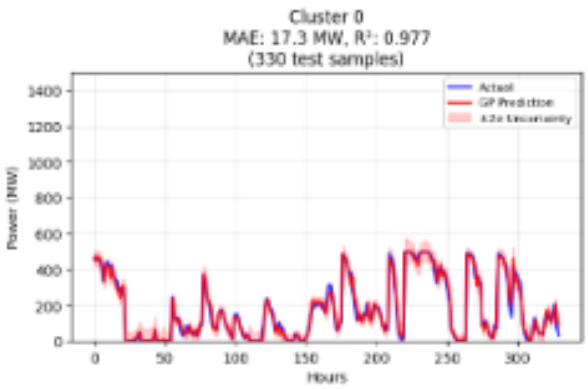


Figure 5.11: Cluster 0 forecasting performance

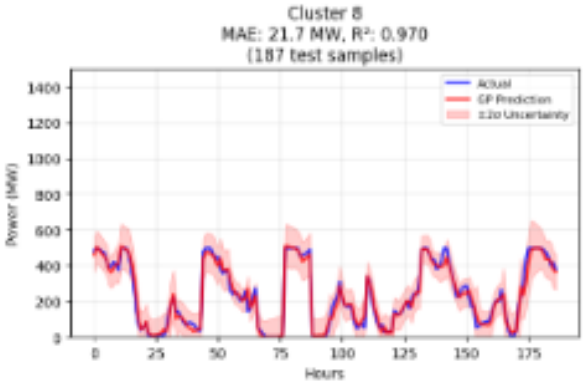


Figure 5.12: Cluster 8 forecasting performance

- **Challenging Clusters (9,12):** These clusters exhibit anomalous performance, with MAEs of 6.25% and 9.11% respectively. Cluster 12 contains only 308 training samples, low, though not the lowest, as cluster 11 has 253 samples. However, as mentioned above cluster 11 includes a large proportion of zero-power forecasts, which likely explains its comparatively stronger performance. By contrast, the limited data for cluster 12, well below the overall average of 917 samples, likely contributes to its poorer results and suggests that this cluster represents unique or extreme meteorological conditions which occur with rare frequency. In comparison, cluster 9 has 1,078 training samples.

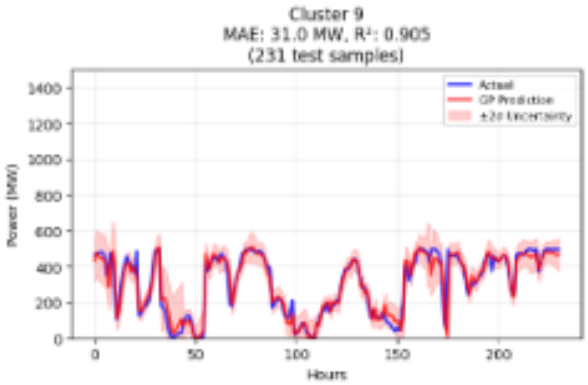


Figure 5.13: Cluster 9 forecasting performance

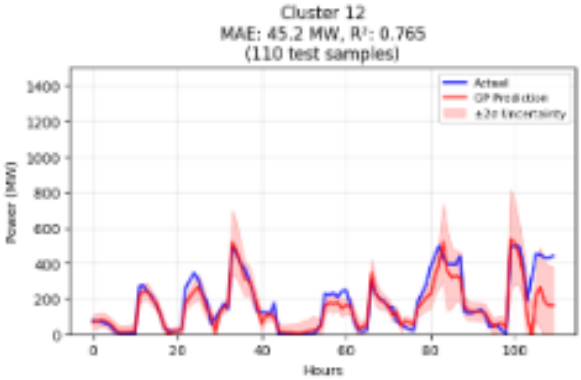


Figure 5.14: Cluster 12 forecasting performance

The cluster-breakdown of training and testing samples as well as results for all models can be found in the Appendix as Table A.2 alongside the forecasting performance graphs per cluster in Figure A.4.

While Cluster 12 exhibits poor performance, it reinforces the rationale for developing cluster-specific models for offshore WPF. The results highlight how challenging it is to predict rare and complex weather regimes with a specialised model, suggesting that a generalised model would likely perform even worse under such conditions. As shown in Figure 5.15, the power and weather profiles for Cluster 12, combined with the small number of training samples, point to the presence of transitional weather regimes. In

particular, the sharp decline in power output and wind speed around hour six is indicative of frontal passages, where high wind conditions rapidly shift to calm periods. Such transitions are notoriously difficult to forecast, and the limited representation of these events in the training data further compounds the predictive challenge.

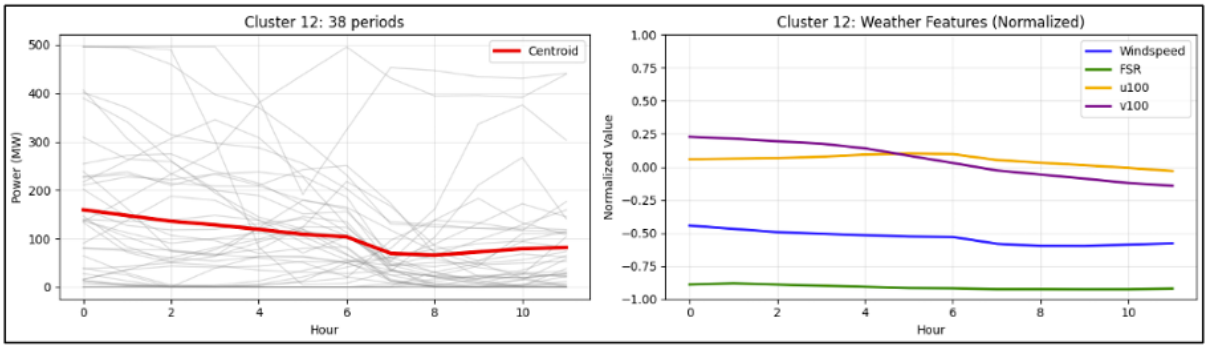


Figure 5.15: Power and Normalised Weather pattern for Cluster 12

Both baseline methods also benefit from the additional training data. The RF model improves to 22.15 MAE (**4.47%**), while LR achieves 27.26 MW MAE (**5.50%**). Despite these improvements, the GP framework maintains superior performance across 10 of the 14 clusters.

The two-year model configuration (14 clusters, 12-hour periods) provides the optimal balance between performance and computational efficiency for practical deployment. The 5.36% improvement in MAE, coupled with the enhanced meteorological resolution capturing 14 distinct weather regimes, establishes this configuration as the foundation for subsequent TL experiments to target wind farms. The slight reduction in R^2 is considered acceptable given the overall improvement in prediction accuracy and the isolated nature of the anomalous cluster performance.

5.2 Transfer Learning Performance

For each TL method, experiments were conducted using three random seeds (42, 63, 84) in Python, with MAE and R^2 results averaged across these runs per cluster. The data fractions referenced throughout this analysis (10%, 20%, 30%, 40% and 50%) are relative to the two-year dataset used for source model training, with the full dataset only representing 40% of the data volume employed in the Wang et al. study.

NB: The forecasting plots in the TL section don't display uncertainty estimates due to a coding error in the plotting function. The issue has been corrected in the notebooks; however, due to time constraints, the updated plots have not been incorporated into this report. Future runs will produce plots with uncertainty estimates.

5.2.1 Method 1 - Frozen VAE with GP Hyperparameter Transfer and Adaptation

As explained in 4.2.7, the first TL method maintains the source VAE in a frozen state whilst adapting only the GP hyperparameters from the trained source GPs to the target

wind farm data. An important thing to consider is Beatrice wind farm has a power capacity of 588 MW compared to 476 MW at Baltic Eagle. So while MAE might be higher, the MAE per capacity might be lower.

Clusters	10%		20%		30%		40%		50%	
	Beatrice	Baltic Eagle	Beatrice	Baltic Eagle	Beatrice	Baltic Eagle	Beatrice	Baltic Eagle	Beatrice	Baltic Eagle
0	34.73	37.57	21.03	30.83	18.13	27.50	17.90	21.73	16.60	21.17
1	41.50	33.47	29.27	35.33	24.73	33.27	24.47	29.90	23.50	27.77
2	14.97	21.30	15.47	15.07	15.13	13.47	14.90	12.93	14.50	12.97
3		14.70	9.60	13.70	11.70	13.13	7.63	11.63	7.20	11.53
4	23.73	23.33	24.83	19.50	21.60	19.53	22.17	18.37	21.87	16.13
5	18.70	33.43	18.07	17.73	17.87	18.30	14.20	17.33	14.90	16.57
6	11.40	20.50	10.03	14.57	9.30	15.10	9.50	12.47	8.60	11.03
7	27.17	24.43	24.00	20.70	24.17	22.03	22.97	20.93	25.23	19.90
8	40.00	16.75	26.07	15.80	26.67	15.40	20.30	13.93	18.77	13.90
9	22.07	22.83	22.27	22.00	20.67	20.67	20.27	21.83	20.30	21.03
10	40.17	38.80	25.90	23.07	20.67	21.63	19.20	19.63	18.83	17.73
11	29.30		21.57	13.05	13.87	10.93	12.37	8.00	10.53	8.50
12		23.70		23.75	76.33	22.85	60.23	16.97	41.10	16.90
13	34.87		16.17		13.60	16.47	12.33	13.97	13.27	13.27
Avg MAE	4.92%	5.73%	3.88%	4.51%	3.70%	4.41%	3.49%	4.10%	3.43%	3.82%
Across both sites		5.32%		4.19%		4.06%		3.79%		3.62%
Avg R^2	0.918	0.889	0.941	0.915	0.92	0.913	0.936	0.928	0.948	0.942

Table 5.2: Cluster-wise results at different training fractions. Cluster results given in raw MAE, while average MAE given as % of capacity

Cross-Site Performance Comparison

The TL results seen in Table 5.2 reveal clear performance differences between the two target wind farms. Beatrice wind farm consistently outperforms the Baltic Eagle site across all training fractions, with average MAE (as % of capacity) differences within the same data fraction ranging from 0.39% to 0.81%. The smallest gap is observed when 50% of target data is used. This disparity likely reflects greater meteorological similarity between Beatrice and the Baie de Saint-Brieuc source site, both of which are exposed to Atlantic weather systems and harsh marine conditions.

Data Efficiency and Benchmark Comparison

Overall the TL methodology delivers strong forecasting performance with minimal target data requirements. With only 20% of target data, Beatrice achieves 3.88% MAE and an $R^2 = 0.941$, already outperforming the fully trained source model performance of 4.14% and $R^2 = 0.925$. When averaged across both wind farms, the framework surpasses source MAE performance at just 30% data utilisation, achieving 4.06% MAE with a slightly lower R^2 of 0.917.

At 50% target data utilisation, the average MAE across both wind farms is 3.62%, which would rank second among the models reported by Wang et al., despite using only 20% of their data volume. While Wang et al.'s analysis represents results averaged across eight wind farms, these sites are all similarly located between the UK and Belgium/Netherlands as seen in figure 5.16. Although further investigations are needed to validate the proposed methodology across additional wind farm locations, the results from

two wind farms separated by much greater geographic distances (North Sea and Baltic Sea) demonstrate comparable performance, with significant data efficiency advantages.

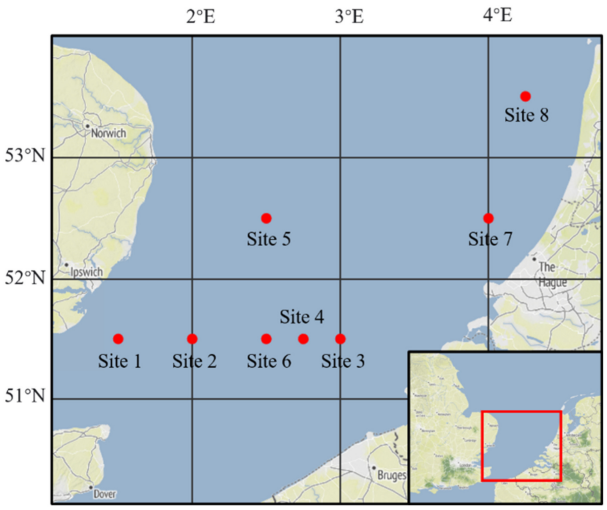


Figure 5.16: Locations of the eight wind farms used in Wang et al. analysis

Cluster-specific Transfer Dynamics

The cluster-wise breakdown reveals heterogeneous transfer performance across the different meteorological regimes. Several clusters demonstrate strong stability and require only minimal target data for effective transfer, while others exhibit higher sensitivity and variation:

Stable Transfer Clusters: Cluster 2 maintains consistent performance across both wind farms and all data fractions, with MAE values ranging from 12.93-21.30 MW. Figure 5.17 illustrates forecasting results for Baltic Eagle at 20% target data, showing how effectively the cluster generalises with limited training. The associated power and weather pattern in Figure 5.18 suggests that this cluster corresponds to calmer meteorological conditions with reduced volatility, facilitating more reliable transfer.

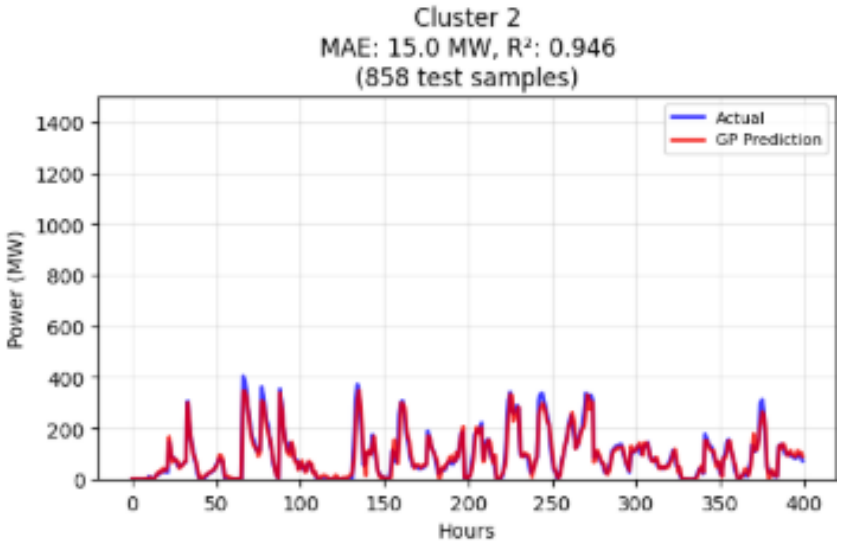


Figure 5.17: Forecasting performance for cluster 2 on Baltic Eagle with 20% of target data used (seed 84)

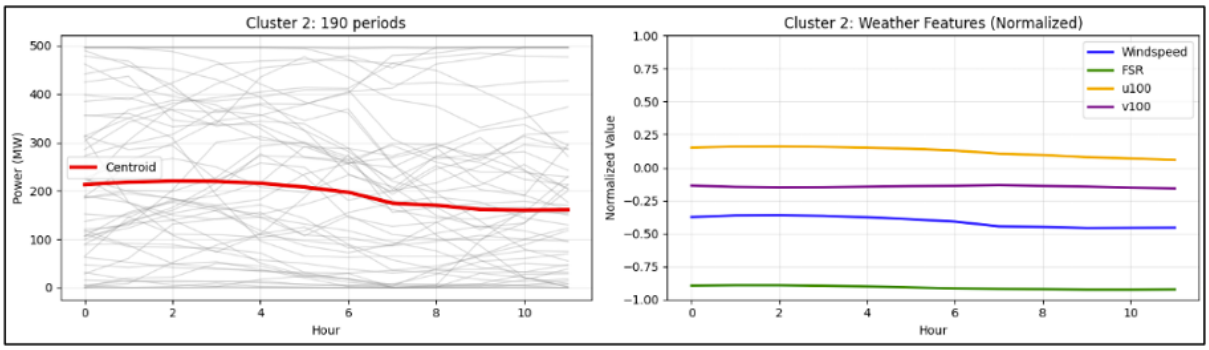


Figure 5.18: Power and Weather pattern for cluster 2

Variable Transfer Clusters: In contrast, cluster 12 exhibits the most pronounced variability, with MAEs ranging from 41.10–76.33 MW at Beatrice and 16.90–23.75 MW at Baltic Eagle. This disparity between wind farms and data fractions supports the earlier hypothesis that cluster 12 captures frontal passage events. The stronger and more volatile North Sea frontal systems at Beatrice pose greater forecasting challenges compared to the relatively sheltered Baltic environment. This is supported by Figures 5.19 and 5.20 which show the forecasting performance on cluster 12 data from both Beatrice and Baltic Eagle when using 50% of the data and a seed of 63. Beatrice clearly has more extreme power output increases, indicating more extreme weather environments.

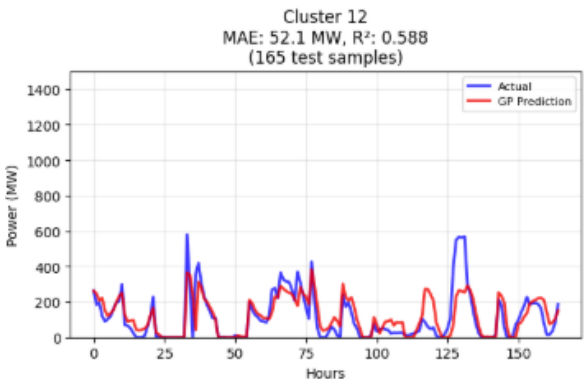


Figure 5.19: Forecasting performance for cluster 12 on Beatrice with 50% of target data used (seed 63)

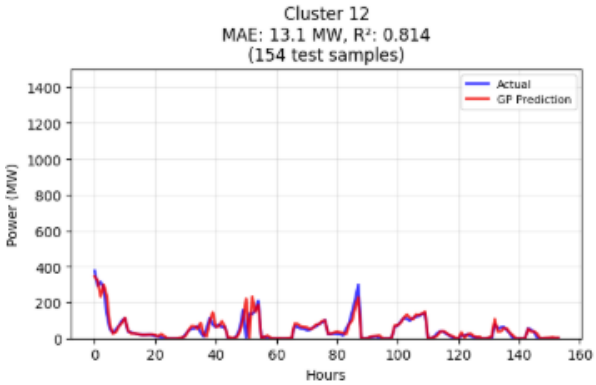


Figure 5.20: Forecasting performance for cluster 12 on Baltic Eagle with 50% of target data used (seed 63)

Progressive Improvement Clusters: Most clusters, such as 0, 1 and 10, show consistent performance improvements as more training data is introduced. The most notable gains occur between 10% and 20% utilisation, with the average MAE across both wind farms reducing from 5.32% to a respectable 4.19%, suggesting that at least one-fifth of the target dataset is necessary to enable meaningful transfer with this method.

Overall, the cluster-wise analysis demonstrates that this methodology effectively captures the underlying meteorological patterns from the source domain while adapting to site-specific characteristic at both target locations. Consistent performance improvements observed as training data increases, along with the ability to achieve competitive results using limited data, validate the approaches effectiveness.

5.2.2 Method 2 - Adaptive VAE with GP Hyperparameter Transfer and Adaptation

Method 2 extends the TL framework by fine-tuning the VAE encoder on the target data, which contributes to the features in the datasets, while maintaining fixed cluster assignments from the source VAE.

Clusters	10%		20%		30%		40%		50%	
	Beatrice	Baltic Eagle	Beatrice	Baltic Eagle	Beatrice	Baltic Eagle	Beatrice	Baltic Eagle	Beatrice	Baltic Eagle
0	36.10	31.20	21.17	24.97	18.53	30.53	18.00	21.40	17.20	21.40
1	37.23	28.63	28.47	33.57	24.50	34.50	23.93	29.10	23.40	25.87
2	14.80	21.20	15.33	14.93	15.37	13.70	15.13	13.03	15.10	13.07
3		14.75	9.60	13.57	11.73	12.70	7.73	11.57	6.05	12.37
4	23.63	20.37	25.43	18.97	22.03	17.63	21.77	17.47	22.25	18.40
5	17.50	29.30	17.80	20.63	17.47	18.20	14.63	17.13	14.65	15.95
6	12.00	17.75	9.77	15.03	9.20	15.23	9.47	13.70	9.15	13.05
7	26.40	24.07	24.33	20.33	24.50	21.43	23.40	20.40	24.95	19.65
8	40.10	16.05	26.83	15.70	26.37	14.67	20.23	13.30	18.65	13.50
9	21.80	22.47	22.70	21.80	20.73	20.43	20.43	21.70	20.85	22.10
10	39.50	43.27	25.63	24.67	19.80	20.57	19.07	18.97	18.55	19.00
11	31.35		21.80	13.55	13.93	10.83	12.87	8.07	9.70	8.50
12		23.70		23.25	54.80	23.25	65.80	17.03	34.50	19.25
13	37.67		16.20		13.10	17.63	12.53	15.20	10.90	13.00
Avg MAE	4.83%	5.47%	3.91%	4.51%	3.65%	4.33%	3.52%	4.00%	3.41%	3.87%
Across both sites		5.15%		4.21%		3.99%		3.76%		3.64%
Avg R^2	0.924	0.900	0.940	0.917	0.919	0.915	0.929	0.930	0.950	0.937

Table 5.3: Cluster-wise results at different training fractions.

Performance Comparison with Method 1:

The results in Table 5.3 show that Method 2 preserves the strong transferability of Method 1 while delivering modest improvements, particularly at the Baltic Eagle site. The familiar trend remains: Beatrice consistently outperforms Baltic Eagle across all data fractions, though the performance gap narrows to between 0.46–0.68%, with the largest difference being seen with 30% of the data.

At 10% data utilisation, Method 2 achieves slight but measurable gains over Method 1, with Beatrice improving from 4.92% to 4.83% and Baltic Eagle from 5.73% to 5.47%. Across higher fractions, the largest single-site improvement occurs at 40% data utilisation for Baltic Eagle, where the MAE decreases from 4.10% to 4.00%, underlining the overall similarity between the two methods.

Data Efficiency and Benchmark Performance:

Given the close alignment with Method 1, Method 2 also achieves competitive results compared with both the source model and the benchmarks reported by Wang et al.. At the Beatrice site, Method 2 surpasses the source model's MAE of 4.14% using just 20% of target data, achieving 3.91%. Baltic Eagle, in contrast, requires 40% of the data to match this level, reaching 4.00% MAE.

When averaged across both farms, Method 2 outperforms the source model with 30% of the target data, achieving 3.99% MAE compared with Method 1's 4.06%. At 50% utilisation, the average MAE falls is 3.64%, which would place Method 2 in second among the models reported by Wang et al, similar to method 1.

Cluster-specific Transfer Dynamics:

At the cluster level, Method 2 produces results similar to those of Method 1, confirming that most of the transfer benefit stems from initialising the target GP hyperparameters with the optimised values learned from the source site.

Cluster 12 remains a persistent challenge for Beatrice; however, at 50% data utilisation, the MAE drops from 41.10 in Method 1 to 34.50 in Method 2.

More broadly, Beatrice benefits more from increased data availability than Baltic Eagle. Between 10% and 50% data utilisation, Beatrice shows an average improvement of 11.18 MW (1.9% of capacity) per cluster, compared with only 6.35 MW (1.33% of capacity) at Baltic Eagle. This suggests greater volatility in North Sea conditions, where smaller datasets struggle to capture the variability, compared with the more stable Baltic Sea.

The most pronounced improvement for Beatrice occurs in cluster 13, where MAE decreases from 37.67 MAE (6.41%) at 10% utilisation to 10.90 MW (1.85%) at 50%. This progression is illustrated in Figures 5.21 and 5.22 using the results from seed 42 as an example. Although training with only 10% of the data provides a larger test set, the model demonstrates a markedly improved ability to capture the sudden power spikes

seen within this cluster. Correspondingly, the R^2 score improves dramatically from 0.595 to 0.930 for this seed.

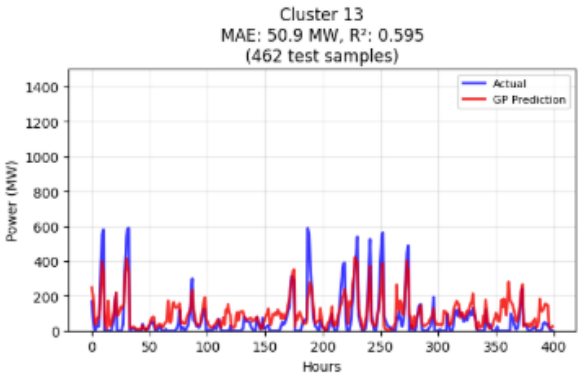


Figure 5.21: Forecasting performance for cluster 13 on Beatrice with 10% of target data used (seed 42)

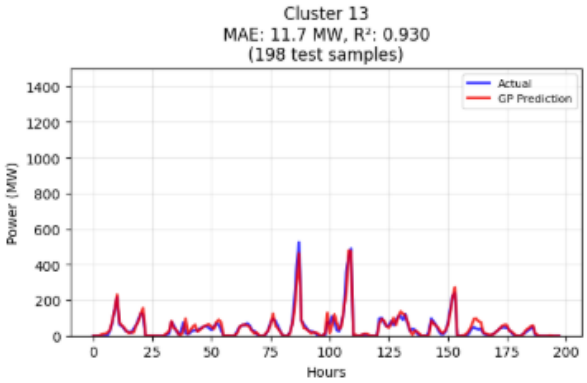


Figure 5.22: Forecasting performance for cluster 13 on Beatrice with 50% of target data used (seed 42)

While the average improvement for Baltic Eagle is more modest, cluster 10 demonstrates a notable reduction in error, decreasing from 43.27 MAE (9.09%) at 10% utilisation to 19.0 MAE (3.99%) at 50%. This suggests that although less frequent, Baltic Eagle does encounter extreme meteorological conditions that require larger volumes of data for accurate forecasting. A similar pattern is observed at Beatrice, where cluster 10 achieves a substantial improvement of 20.95 MW MAE, supporting the hypothesis that this cluster represents more extreme weather patterns that require more data. The improvement for Baltic Eagle is illustrated in Figures 5.23 and 5.24, again using the results from seed 42 as an example. It is worth noting that for seed 42, the best performance for cluster 10 was reached at 40% data utilisation, achieving 19.0 MW MAE compared to 19.5 MW MAE with 50%.

Figure 5.23: Forecasting performance for cluster 10 on Baltic Eagle with 10% of target data used (seed 42)

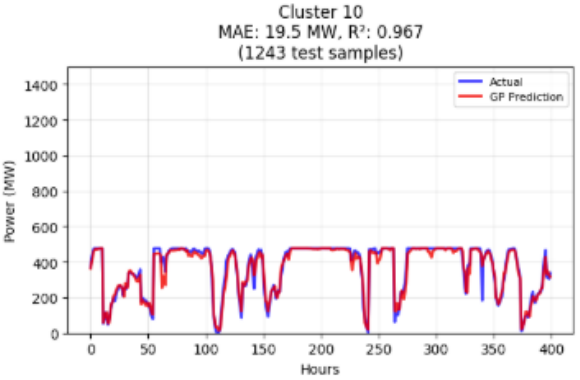


Figure 5.24: Forecasting performance for cluster 10 on Baltic Eagle with 40% of target data used (seed 42)

Overall, Method 2 confirms that fine-tuning the encoder enhances transfer learning efficiency by improving feature representation for the target farms. The combination of low MAE at small data fractions and robust performance across clusters highlights the method's effectiveness and data efficiency.

5.2.3 Method 3 - Cluster-based Dataset alignment with Adaptive VAE and GP Hyperparameter Transfer and Adaptation

Method 3 builds on the TL framework in method 2 by partially shifting the non-cyclical weather features and the power output within the constructed target datasets. The purpose of this adjustments is to align the target data distributions more closely with those of the source dataset before transferring the hyperparameters from the source GP models.

Clusters	10%		20%		30%		40%		50%	
	Beatrice	Baltic Eagle	Beatrice	Baltic Eagle	Beatrice	Baltic Eagle	Beatrice	Baltic Eagle	Beatrice	Baltic Eagle
0	29.70	38.25	31.40	31.85	22.55	32.40	19.95	23.75	21.40	24.20
1	44.40	36.90	33.40	37.30	28.85	37.20	25.70	31.45	25.95	31.10
2	15.15	25.10	15.20	17.55	14.90	15.75	15.00	14.75	15.55	14.80
3	-	14.35	8.20	15.70	9.65	14.40	9.15	13.10	10.15	12.85
4	26.85	22.75	30.45	19.95	29.40	19.25	28.25	17.75	28.80	15.20
5	31.30	31.60	40.00	22.50	36.70	19.55	30.95	20.75	26.75	17.25
6	11.30	18.15	8.45	13.05	9.65	17.00	8.65	12.45	8.40	12.70
7	26.85	24.95	26.05	20.70	24.80	23.00	23.85	21.00	27.90	19.85
8	32.60	18.25	26.55	16.60	22.90	16.20	19.40	14.10	19.50	13.70
9	24.50	28.30	23.50	22.25	22.40	22.80	21.15	22.85	21.60	22.50
10	41.00	46.65	27.20	26.25	22.00	23.00	23.80	20.65	28.00	20.00
11	28.00	-	28.10	11.60	14.95	8.95	16.50	10.95	16.50	9.75
12	-	23.10	-	25.15	72.55	22.70	62.60	20.30	28.40	19.70
13	34.30	-	16.05	-	12.50	22.40	11.40	14.60	12.80	14.05
Avg MAE	5.07%	6.18%	4.59%	4.85%	4.18%	4.73%	3.92%	4.28%	3.96%	4.06%
Across both sites		5.63%		4.72%		4.46%		4.1%		4.01%
Avg R^2	0.921	0.870	0.931	0.909	0.906	0.845	0.939	0.928	0.937	0.936

Table 5.4: Cluster-wise results at different training fractions.

Performance Comparison with Methods 1 and 2

The results in Table 5.4 show that method 3 consistently performs worse than methods 1 and 2 across all fractions of training data. Across both sites, the MAE is on average 0.4% higher than the mean performance of methods 1 and 2 for each data fraction. The biggest gap appears at 20% utilisation, where method 3 is 0.52% worse (4.72% to 4.2%), while the smallest gap is at 40%, where it is 0.32% worse. Despite adding an extra step of distribution-shifting, the method reduces rather than improves transfer effectiveness.

There are several potential reasons for this negative outcome. First, shifting weather features and power output distributions by different amounts could disrupt the natural physical relationships that the models rely on for accurate predictions. Second, the approach might have introduced artificial noise, so the features no longer represent real meteorological conditions. In WPF, where input-output relationships are so key, this could create unrealistic relationships that the models cannot learn effectively. And lastly, the shifting may have distorted the characteristics of time-periods that were originally on the boundary between clusters. These boundary periods, which were already at the edge of their assigned cluster's typical patterns, could have been pushed further away from their cluster's centre, making them poorly represented by their cluster's model even though they remain assigned to the same cluster.

Further investigation is needed to explore different shifting percentages for both weather features and power output, as the current approach of 50% for features and 30% for power may not provide optimal alignment. Testing various combinations of

these percentages could help understand how the magnitude and balance of distribution shifts affects transfer learning performance.

Data Efficiency and Benchmark Performance

Unlike methods 1 and 2, method 3 shows slower improvements compared to the source model baseline of 4.14% MAE, requiring 40% training data, to achieves 4.1% MAE across both farms. Method 2 for example, achieved 3.99% MAE with only 30% of the data. For the Baltic Eagle site, method 3 only achieves an MAE of 4.06% with 50% utilisation, significantly lower then then method 1 that achieves 3.82% in the same data utilisation.

Cluster-level Differences

Table 5.5 shows the differences in raw MAE per cluster, data fraction and wind farm between Method 3 and the average of Method 1 and 2. In only 2 of the 14 clusters does Method 3 outperform the average of method 1 and 2. The biggest difference is in cluster 5 where on average Method 3 underperform

Clusters	10%		20%		30%		40%		50%		Average MAE difference across data fractions and wind farms per cluster
	Beatrice	Baltic Eagle	Beatrice	Baltic Eagle	Beatrice	Baltic Eagle	Beatrice	Baltic Eagle	Beatrice	Baltic Eagle	
0	-5.72	3.87	10.30	3.95	4.22	3.38	2.00	2.18	4.50	2.92	3.16
1	5.03	5.85	4.53	2.85	4.23	3.32	1.50	1.95	2.52	4.28	3.61
2	0.27	3.85	-0.20	2.55	-0.35	2.17	-0.02	1.77	1.07	1.78	1.29
3		-0.38	-1.40	2.07	-2.07	1.48	1.47	1.50	2.97	0.90	0.73
4	3.17	0.90	5.32	0.72	7.58	0.67	6.28	-0.17	6.95	-1.68	2.97
5	13.20	0.23	22.07	3.32	19.03	1.30	16.53	3.52	11.65	0.77	9.16
6	-0.40	-0.98	-1.45	-1.75	0.40	1.83	-0.83	-0.63	-0.23	1.28	-0.28
7	0.07	0.70	1.88	0.18	0.47	1.27	0.67	0.33	2.82	0.22	0.86
8	-7.45	1.85	0.10	0.85	-3.62	1.17	-0.87	0.48	0.70	-0.02	-0.68
9	2.57	5.65	1.02	0.35	1.70	2.25	0.80	1.08	1.30	1.27	1.80
10	1.17	5.62	1.43	2.38	1.77	1.90	4.67	1.35	9.25	2.02	3.16
11	-2.33		6.42	-1.70	1.05	-1.93	3.88	2.92	6.13	1.23	1.74
12		-0.60		1.65	6.98	-0.35	-0.42	3.30	-10.35	2.65	0.36
13	-1.97		-0.13		-0.85	5.35	-1.03	0.02	-0.27	0.78	0.24

Table 5.5: Difference in raw MAE between Method 3 results and the average Method 1 and Method 2 results per cluster, wind farm and data fraction

This uneven impact suggests that certain weather patterns may be more sensitive to distribution shifts, with clusters like Cluster 5 being particularly affected, possibly because power relationships are more pronounced in these conditions.

While this report doesn't investigate the exact reasons why this poorer performance is seen in method 3, experiments running different levels of shifting than the 50% and 30% used in these experiments might give a better understanding as to the performance of this methodology.

Chapter 6

Next Steps

Key Research Findings

The novel offshore WPF framework developed in this study, enhanced through TL applications, achieves forecasting performance competitive with the latest methodologies while offering insights into the effectiveness of different transfer learning methodologies applied to GP models. When trained with only 50% of the target data (equivalent to one year), the three transfer learning frameworks achieve average MAEs of 3.62%, 3.64%, and 4.01% across both wind farms. The results for methods 1 and 2 are only marginally higher than the 3.60% reported by the best performing model in Wang et al. [3], outperforming the other ten models presented in their study, despite requiring just 20% of the data used in their work. This represents a significant advancement in data efficiency for offshore WPF, especially given the greater geographic diversity of the wind farms considered in this study compared to those in Wang et al.

The weather regime clustering methodology proves particularly effective in addressing the inherent variability of meteorological data. The substantial forecasting performance differences observed across clusters highlights both the potential and challenges of this approach. While specialised models excel at forecasting specific weather patterns, certain extreme conditions remain difficult to predict accurately. However, this exact variability underscores the limitations of generalised models in maintaining consistent performance across diverse temporal conditions and supports the research into specialised models.

The superior performance achieved when transferring from Baie de Saint-Brieuc wind farm to Beatrice, compared to transfers to Baltic Eagle, suggests that meteorological similarity significantly influences TL effectiveness even when sites are geographically far apart. This finding emphasises the critical importance of source-target compatibility for TL in offshore WPF applications.

Future Research

Several avenues for future research emerge from this work. However, maintaining alignment with the core objective of increasing access to forecasting tools requires balancing performance improvements with practical implementation constraints. For example, sophisticated feature engineering using additional meteorological parameters might improve forecasting accuracy, but such approaches would reduce accessibility for smaller operators or emerging offshore wind markets. Key directions for future research include:

First, expanded validation across a larger set of wind farms is essential to establish the methodology's generalisability compared to the eight wind farms used in Wang et al.'s analysis. A particularly interesting experiment would be to investigate the TL potential across different oceans, moving beyond the relatively concentrated European offshore sites examined to date.

Second, systematic investigation of source-target transferability patterns is a crucial step. Evaluating TL performance when using different source wind farms (e.g. training on Baltic Eagle and testing on Beatrice or other locations) would inform optimal source domain selection strategies and provide practical guidelines for implementation.

Third, the consistent underperformance of specific weather clusters, particularity Cluster 12 indicates substantial room for improvement in model generalisation. Future work could explore cluster-specific kernel structures tailored to individual weather regimes, allowing each GP model to optimise for its particular atmospheric conditions rather than applying uniform kernel functions, with hyperparameter tuning, across all clusters.

Fourth, enhanced clustering methodologies offer another promising research avenue. Clustering forms a foundational element of the framework, yet training on one or two years of data achieved only moderate clustering scores (0.645 and 0.68, respectively). Advanced clustering techniques and evaluation metrics could improve the separation of weather patterns and, consequently forecasting accuracy. Additionally, integrating weather data from multiple, geographically distinct wind farms during VAE training may produce more robust latent representations of atmospheric conditions, improving separation between clusters and supporting the advancement of source GP models.

In conclusion, this report introduces a novel framework for offshore WPF that enables emerging wind farms to achieve accurate forecasts with substantially reduced data requirements. The framework demonstrates promising results while outlining clear avenues for future research. Pursuing these directions could further enhance forecasting performance, supporting the expansion of wind power as a key component of the global energy market beyond the 19 countries it currently serves.

Appendix A

Appendix

Cluster	Train	Test	MAE			R-Squared		
			GP	RF	LR	GP	RF	LR
0	715	198	16.6	17.9	19.2	0.942	0.935	0.927
1	330	0	-	-	-	-	-	-
2	418	121	25.5	27.5	33.0	0.949	0.938	0.917
3	770	187	16.5	22.6	23.1	0.980	0.967	0.965
4	616	154	31.0	26.5	35.0	0.939	0.958	0.906
5	704	121	22.2	22.7	29.0	0.934	0.935	0.916
6	561	143	18.6	21.0	24.9	0.944	0.923	0.916
7	693	198	28.2	30.6	31.2	0.919	0.910	0.903
8	594	165	12.2	12.6	26.6	0.883	0.814	0.637
9	693	176	22.2	22.7	32.9	0.964	0.959	0.943
10	187	66	30.9	28.4	53.2	0.676	0.537	0.223
11	132	0	-	-	-	-	-	-
Weighted Avg	534.4	127.4	21.7	22.8	29.1	0.928	0.910	0.864

Table A.1: Training and Testing Data Samples per Cluster along with MAE and R-Squared values for GP, RF, and LR models - One year of Training Data.

Cluster	Train	Test	MAE			R-Squared		
			GP	RF	LR	GP	RF	LR
0	1320	330	17.3	19.8	24.8	0.977	0.967	0.958
1	1397	264	23.3	23.3	31.8	0.932	0.927	0.883
2	1694	396	19.0	23.0	24.3	0.962	0.945	0.944
3	495	121	21.3	29.1	31.5	0.936	0.832	0.832
4	638	242	13.5	11.2	26.8	0.779	0.756	0.416
5	1914	418	15.0	17.1	20.7	0.977	0.971	0.960
6	704	187	17.0	19.2	21.1	0.903	0.882	0.869
7	1485	374	22.4	22.7	34.4	0.950	0.945	0.915
8	484	187	21.7	26.3	26.2	0.970	0.957	0.960
9	1078	231	31.0	29.4	32.9	0.905	0.908	0.889
10	748	187	24.1	25.5	29.1	0.956	0.952	0.944
11	253	110	10.1	12.5	12.5	0.889	0.817	0.834
12	308	110	45.2	42.4	40.2	0.765	0.852	0.867
13	319	55	20.5	24.9	33.6	0.665	0.574	0.626
Weighted Avg	916.9	229.4	20.5	22.1	27.3	0.925	0.912	0.876

Table A.2: Training and Testing Data Samples per Cluster along with MAE and R-Squared values for GP, RF, and LR models- Two years of Training Data

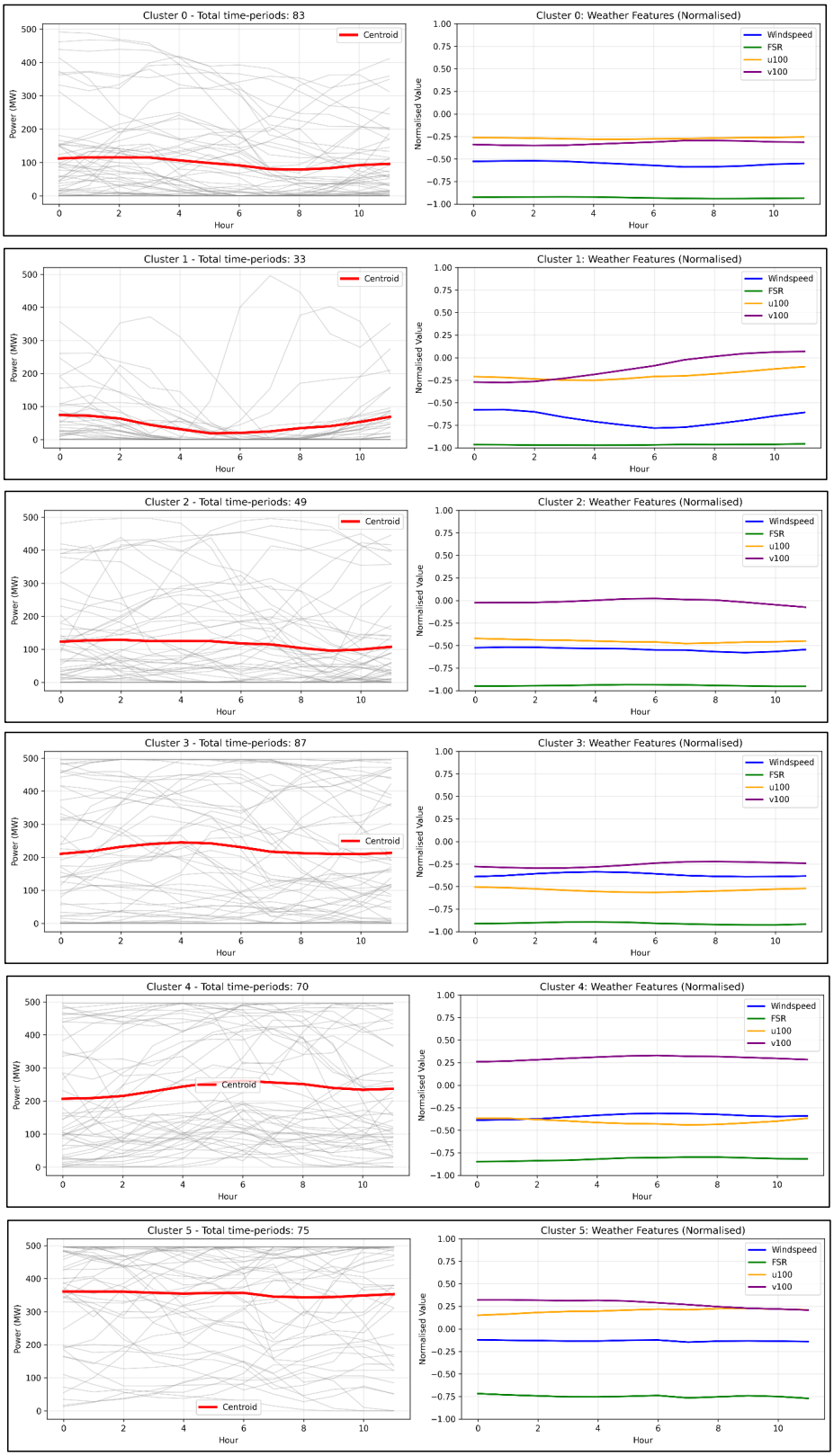


Figure A.1: Power and Weather patterns per cluster using one year of training data

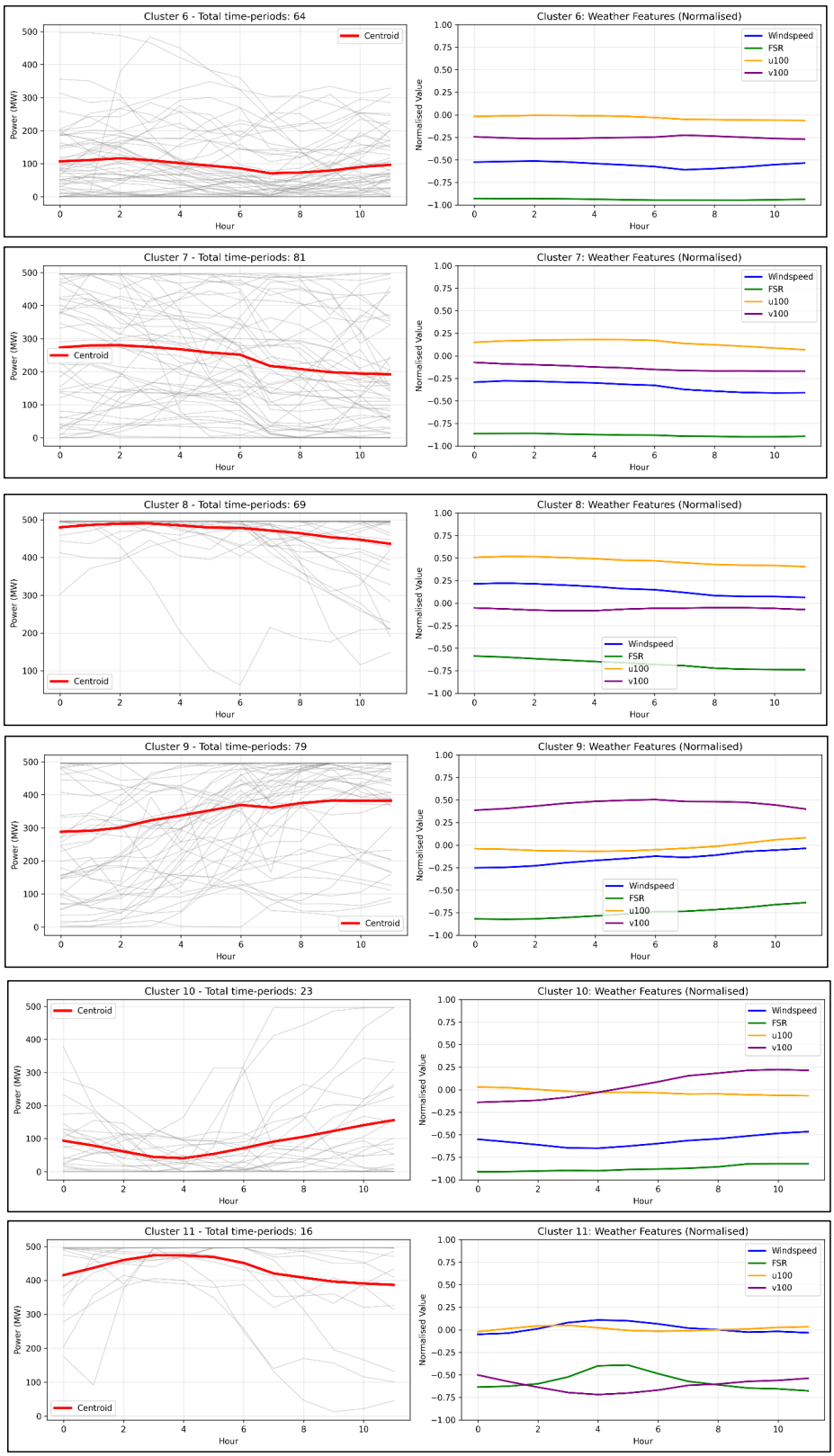


Figure A.1: Power and Weather patterns per cluster using one year of training data (continued)

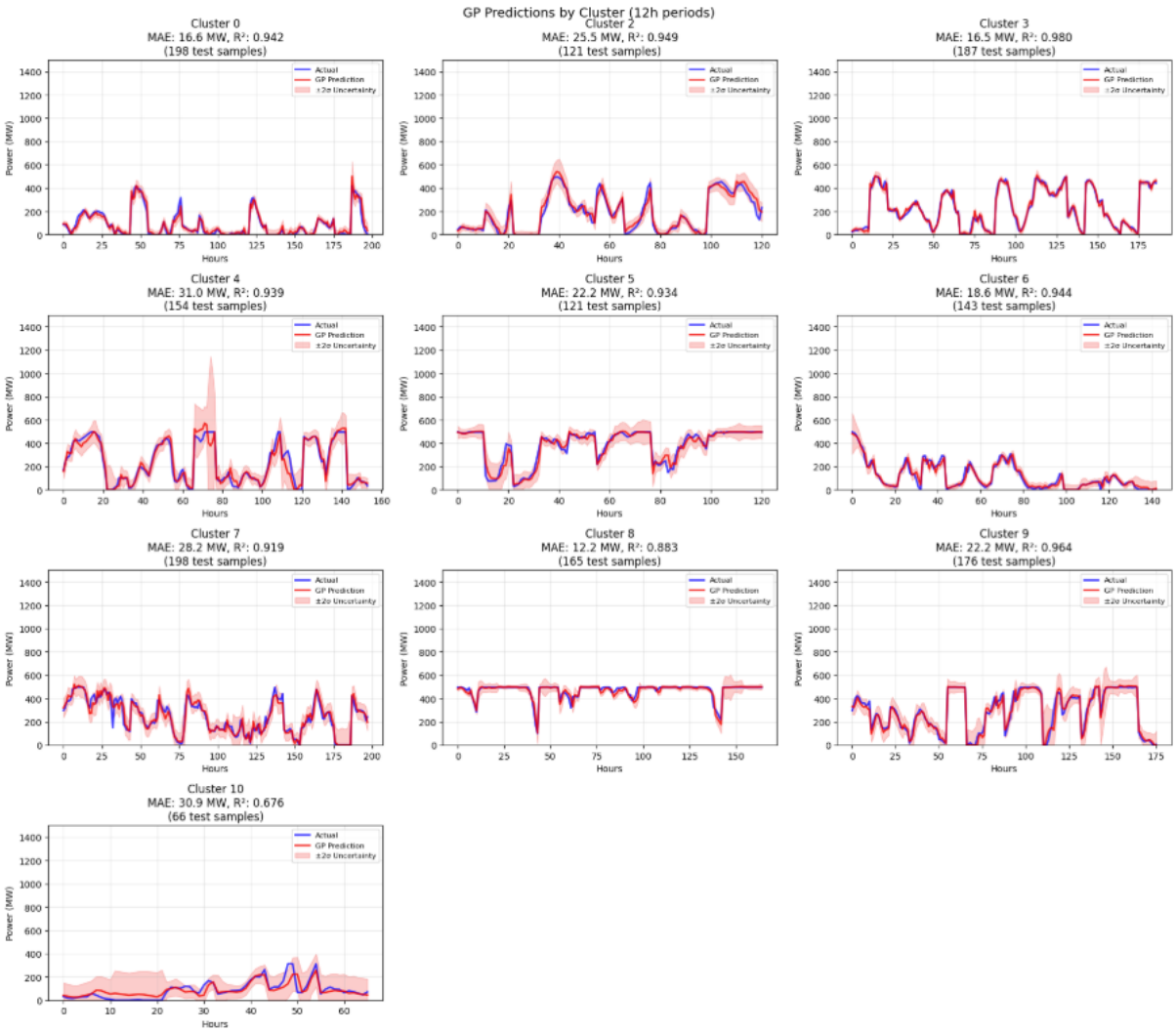
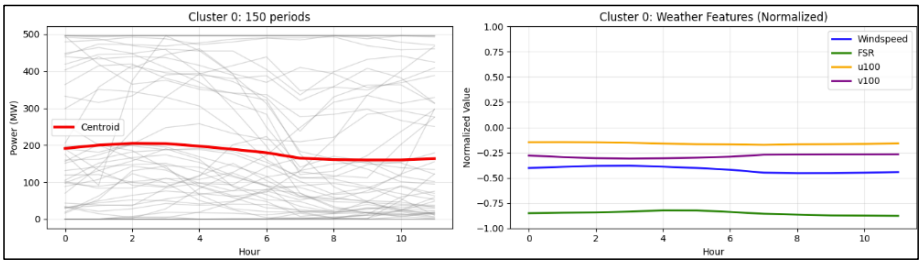
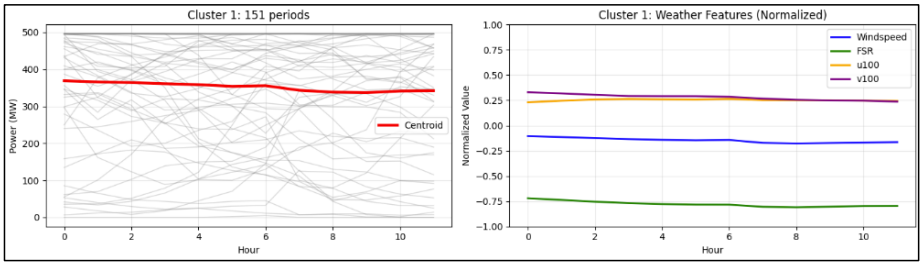
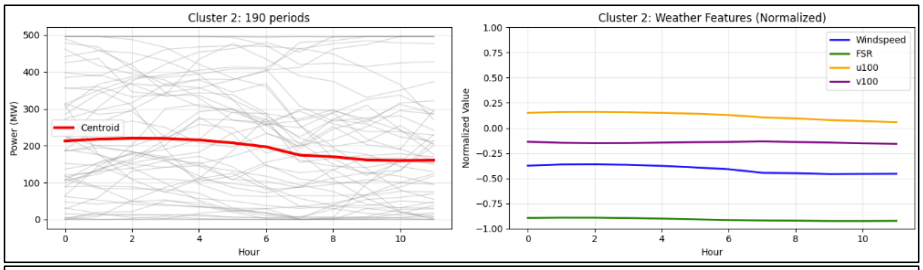
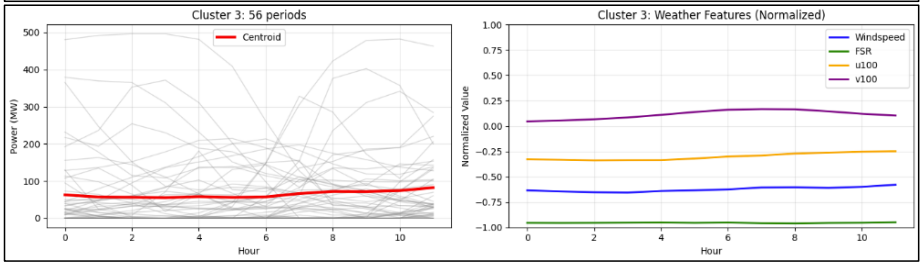


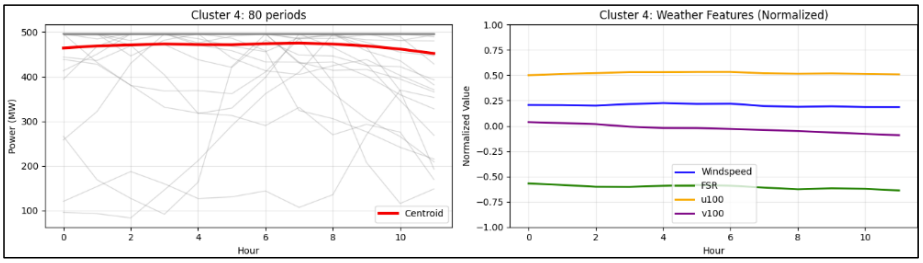
Figure A.2: Forecasts for each Cluster using one year of training data for GP models

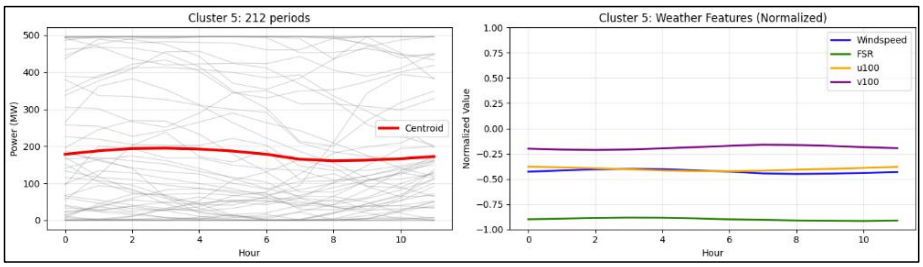

Cluster 0: 150 periods
Power (MW)
Hour
Centroid
Cluster 0: Weather Features (Normalized)
Normalized Value
Windspeed
FSR
u100
v100

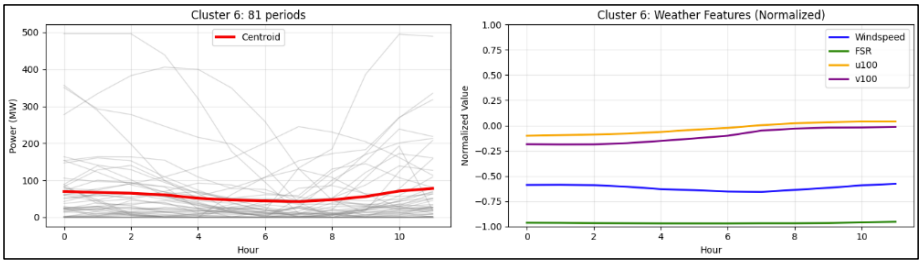

Cluster 1: 151 periods
Power (MW)
Hour
Centroid
Cluster 1: Weather Features (Normalized)
Normalized Value
Windspeed
FSR
u100
v100

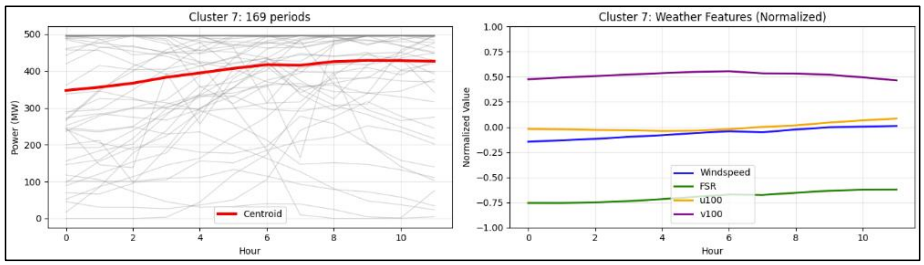

Cluster 2: 190 periods
Power (MW)
Hour
Centroid
Cluster 2: Weather Features (Normalized)
Normalized Value
Windspeed
FSR
u100
v100

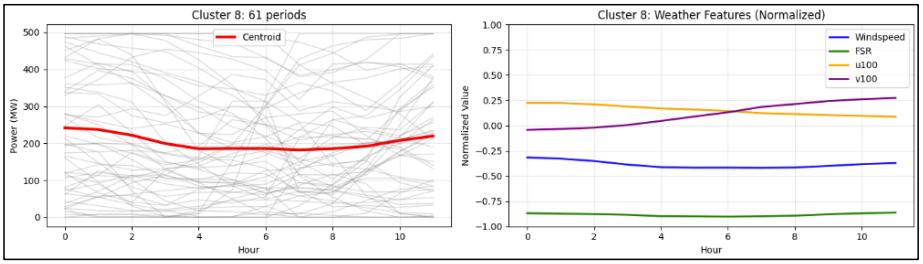

Cluster 3: 56 periods
Power (MW)
Hour
Centroid
Cluster 3: Weather Features (Normalized)
Normalized Value
Windspeed
FSR
u100
v100

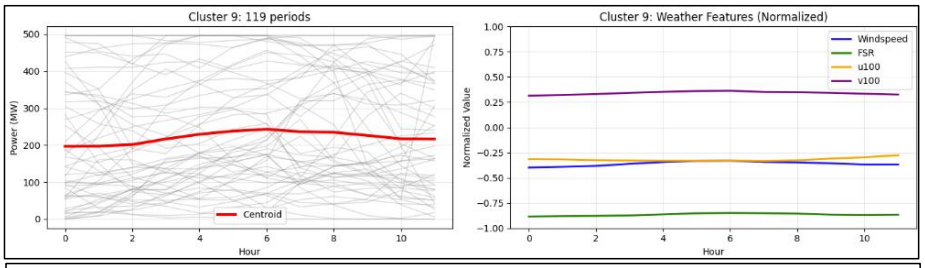

Cluster 4: 80 periods
Power (MW)
Hour
Centroid
Cluster 4: Weather Features (Normalized)
Normalized Value
Windspeed
FSR
u100
v100

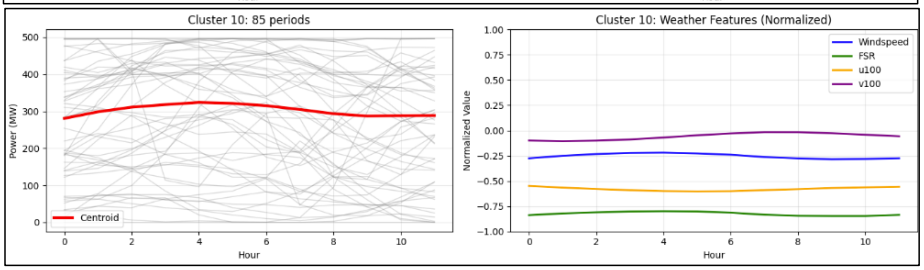

Cluster 5: 212 periods
Power (MW)
Hour
Centroid
Cluster 5: Weather Features (Normalized)
Normalized Value
Windspeed
FSR
u100
v100

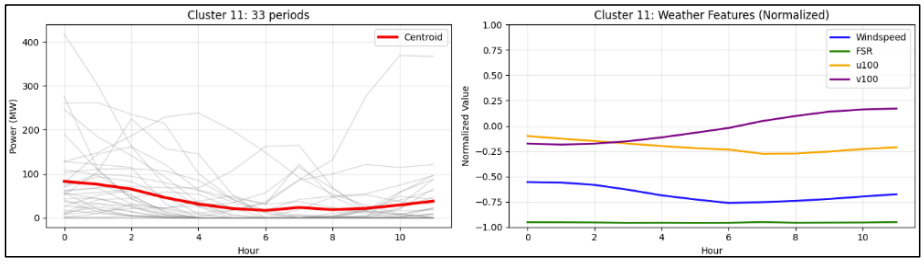

Cluster 6: 81 periods
Centroid
Power (MW)
Hour
Cluster 6: Weather Features (Normalized)
Windspeed
FSR
u100
v100
Normalized Value
Hour


Cluster 7: 169 periods
Centroid
Power (MW)
Hour
Cluster 7: Weather Features (Normalized)
Windspeed
FSR
u100
v100
Normalized Value
Hour


Cluster 8: 61 periods
Centroid
Power (MW)
Hour
Cluster 8: Weather Features (Normalized)
Windspeed
FSR
u100
v100
Normalized Value
Hour


Cluster 9: 119 periods
Centroid
Power (MW)
Hour
Cluster 9: Weather Features (Normalized)
Windspeed
FSR
u100
v100
Normalized Value
Hour


Cluster 10: 85 periods
Centroid
Power (MW)
Hour
Cluster 10: Weather Features (Normalized)
Windspeed
FSR
u100
v100
Normalized Value
Hour


Cluster 11: 33 periods
Centroid
Power (MW)
Hour
Cluster 11: Weather Features (Normalized)
Windspeed
FSR
u100
v100
Normalized Value
Hour

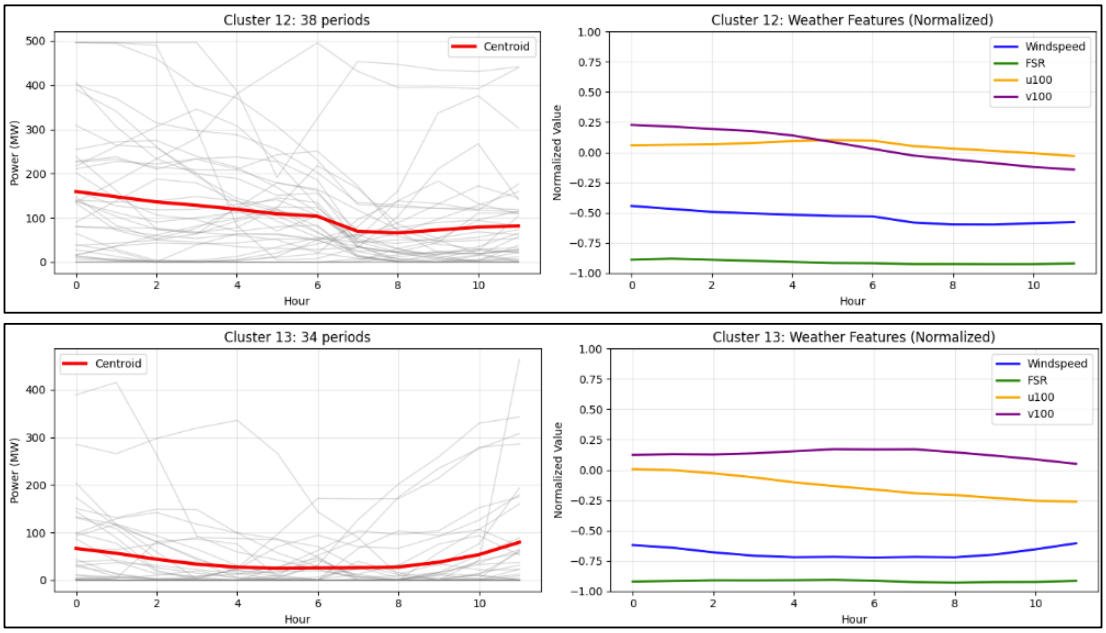


Figure A.3: Power and Weather patterns per cluster using two years of training data

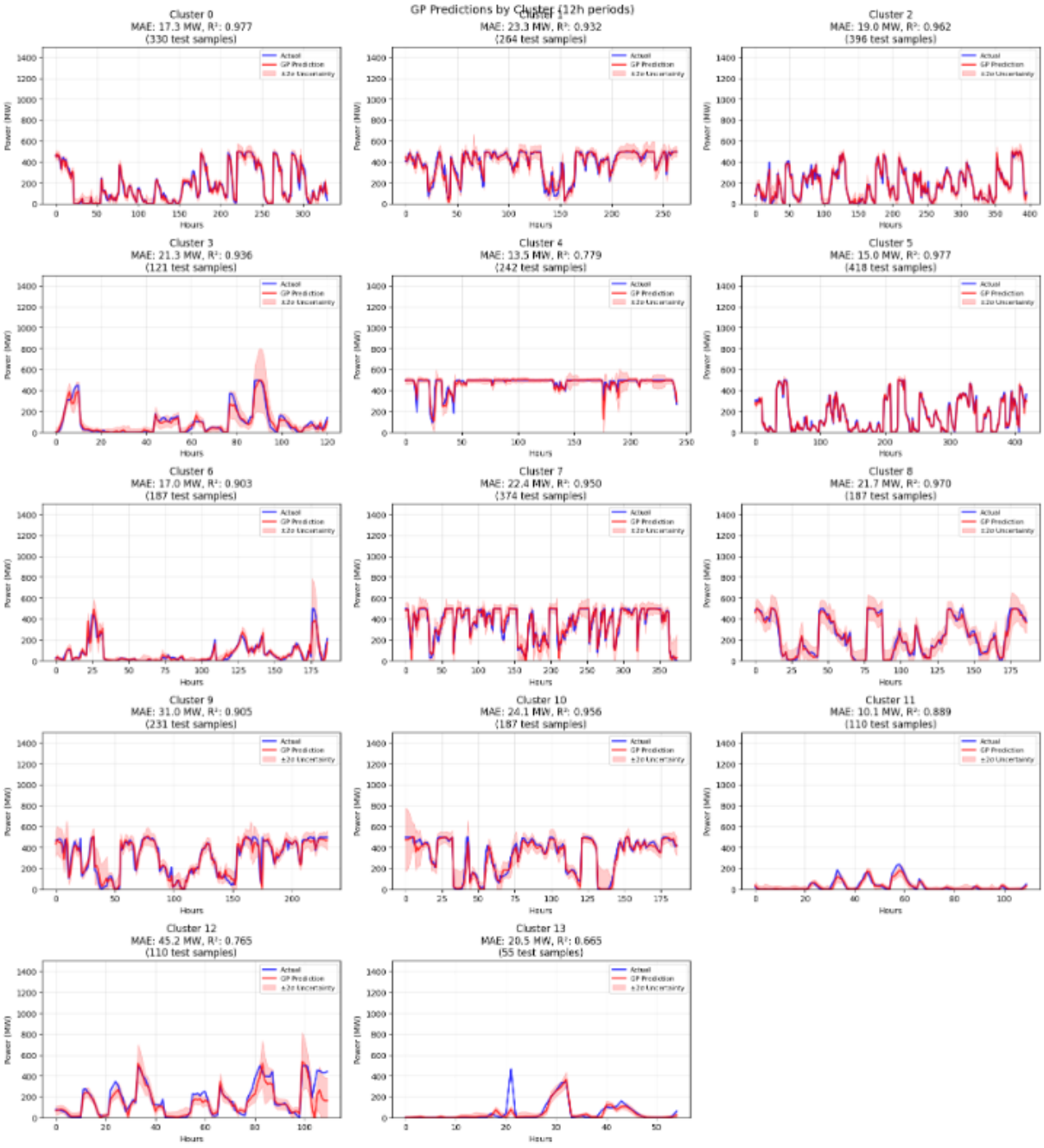


Figure A.4: Forecasts of the GP models for each cluster using two years of data

Bibliography

[1] Navid Zehtabiyan-Rezaie, Alexandros Iosifidis, and Mahdi Abkar. Physics-guided machine learning for wind-farm power prediction: Toward interpretability and generalizability. *PRX Energy*, 2(1):013009, 2023.

[2] Abdulelah Alkesaiberi, Fouzi Harrou, and Ying Sun. Efficient wind power prediction using machine learning methods: A comparative study. *Energies*, 15(7):2327, 2022.

[3] Zhongrui Wang, Chunbo Wang, Liang Chen, Min Yu, and Wenteng Yuan. Short-term offshore wind power multi-location multi-modal multi-step prediction model based on informer (m3stin). *Energy*, 322:135616, 2025.

[4] Global Wind Energy Council. Global Offshore Wind Report 2025. Technical report, Global Wind Energy Council (GWEC), 2025. URL `https://www.gwec.net/reports/globaloffshorewindreport`. Accessed: 2025-08-28.

[5] Shahram Hanifi, Xiaolei Liu, Zi Lin, and Saeid Lotfian. A critical review of wind power forecasting methods—past, present and future. *Energies*, 13(15):3764, 2020.

[6] A. Botterud, Z. Zhou, J. Wang, R. J. Bessa, H. Keko, J. Mendes, J. Sumaili, and V. Miranda. Use of wind power forecasting in operations decisions. Technical report, Argonne National Laboratory, September 2011.

[7] U.S. Department of Energy. Wind forecast improvement project will monitor weather, ocean, and wildlife data near active and proposed offshore wind farms off the east coast. `https://www.energy.gov/eere/wind/articles/wind-forecast-improvement-project-will-monitor-weather-ocean-and-wildlife-data-ne` February 2024. Accessed: 2025-08-28.

[8] Wind Power Monthly. Do we still need met masts? *Wind Power Monthly*, March 2018. URL `https://www.windpowermonthly.com/article/1458018/need-met-masts`. Accessed: 2025-08-28.

[9] The Carbon Trust. Floating lidars prove their worth. Offshore Wind Biz, May 2017. URL `https://www.offshorewind.biz/2017/05/30/the-carbon-trust-floating-lidars-prove-their-worth/`. Accessed: 2025-08-28.

[10] Leice LiDARs. How much does a floating lidar cost? `https://www.leicelidars.com/info/how-much-does-a-floating-lidar-cost--90995875.html`, January 2024. Accessed: 2025-08-28.

[11] Zongxu Liu, Hui Guo, Yingshuai Zhang, and Zongliang Zuo. A comprehensive review of wind power prediction based on machine learning: Models, applications, and challenges. *Energies*, 18(2):350, 2025.

[12] Mohammed AA Al-qaness, Ahmed A Ewees, Ahmad O Aseeri, and Mohamed Abd Elaziz. Wind power forecasting using optimized lstm by attraction–repulsion optimization algorithm. *Ain Shams Engineering Journal*, 15(12):103150, 2024.

[13] Elissaios Sarmas, Nikos Dimitropoulos, Vangelis Marinakis, Zoi Mylona, and Haris Doukas. Transfer learning strategies for solar power forecasting under data scarcity. *Scientific Reports*, 12(1):14643, 2022.

[14] Md Saiful Islam Sajol, Md Shazid Islam, ASM Jahid Hasan, Md Saydur Rahman, and Jubair Yusuf. Wind power prediction across different locations using deep domain adaptive learning. In *2024 6th Global Power, Energy and Communication Conference (GPECOM)*, pages 518–523. IEEE, 2024.

[15] Ahmet Durap. Explainable deep learning techniques for wind speed forecasting in coastal areas: Integrating model configuration, regularization, early stopping, and shap analysis. *Neural Computing and Applications*, 37(1):1–23, 2025. doi: 10.1007/s00521-025-11433-w. URL `https://doi.org/10.1007/s00521-025-11433-w`. Accessed: 2025-08-28.

[16] Jie Yan, Corinna Möhrlen, Tuhfe Göçmen, Mark Kelly, Arne Wessel, and Gregor Giebel. Uncovering wind power forecasting uncertainty sources and their propagation through the whole modelling chain. *Renewable and Sustainable Energy Reviews*, 165: 112519, 2022.

[17] Peter Kalverla, Gert-Jan Steeneveld, Reinder Ronda, and Albert AM Holtslag. Evaluation of three mainstream numerical weather prediction models with observations from meteorological mast ijmuiden at the north sea. *Wind Energy*, 22(1):34–48, 2019.

[18] Ioannis K Bazionis and Pavlos S Georgilakis. Review of deterministic and probabilistic wind power forecasting: Models, methods, and future research. *Electricity*, 2 (1):13–47, 2021.

[19] Yang Yang, Hao Lou, Jinran Wu, Shaotong Zhang, and Shangce Gao. A survey on wind power forecasting with machine learning approaches. *Neural Computing and Applications*, 36(21):12753–12773, 2024.

[20] Alireza Zendehboudi, M Abdul Baseer, and R Saidur. Application of support vector machine models for forecasting solar and wind energy resources: A review. *Journal of cleaner production*, 199:272–285, 2018.

[21] Yu Sun, Qibo Zhou, Li Sun, Liping Sun, Jichuan Kang, and He Li. Cnn–lstm–am: A power prediction model for offshore wind turbines. *Ocean Engineering*, 301:117598, 2024.

[22] Mansoor Khan, Essam A Al-Ammar, Muhammad Rashid Naeem, Wonsuk Ko, Hyeong-Jin Choi, and Hyun-Koo Kang. Forecasting renewable energy for environmental resilience through computational intelligence. *Plos one*, 16(8):e0256381, 2021.

[23] Sen Wang, Wenjie Zhang, Yonghui Sun, Anupam Trivedi, CY Chung, and Dipti Srinivasan. Wind power forecasting in the presence of data scarcity: A very short-term conditional probabilistic modeling framework. *Energy*, 291:130305, 2024.

[24] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[25] Xiaoqian Jiang, Bing Dong, Le Xie, and Latanya Sweeney. Adaptive gaussian process for short-term wind speed forecasting. In *ECAI 2010*, pages 661–666. IOS Press, 2010.

[26] Yimei Wang, Peng Song, Hui Liu, and Linlin Wu. A probabilistic wind power forecasting approach based on gaussian process regression. In *2020 IEEE/IAS Industrial and Commercial Power System Asia (I&CPS Asia)*, pages 1363–1368. IEEE, 2020.

[27] Fatemeh Najibi, Dimitra Apostolopoulou, and Eduardo Alonso. Enhanced performance gaussian process regression for probabilistic short-term solar output forecast. *International Journal of Electrical Power & Energy Systems*, 130:106916, 2021.

[28] Dan Li, Yue Hu, Baohua Yang, Zeren Fang, Yunyan Liang, and Shuai He. A novel transfer learning strategy for wind power prediction based on timesnet-gru architecture. *Journal of Renewable and Sustainable Energy*, 16(3), 2024.

[29] Mansoor Khan, Muhammad Rashid Naeem, Essam A Al-Ammar, Wonsuk Ko, Hamsakutty Vettikalladi, and Irfan Ahmad. Power forecasting of regional wind farms via variational auto-encoder and deep hybrid transfer learning. *Electronics*, 11(2):206, 2022.

[30] Pingfan Wang, Nanlin Jin, Duncan Davies, and Wai Lok Woo. Model-centric transfer learning framework for concept drift detection. *Knowledge-Based Systems*, 275: 110705, 2023.

[31] Hao Chen. Knowledge distillation with error-correcting transfer learning for wind power prediction. *arXiv preprint arXiv:2204.00649*, 2022.

[32] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[33] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[34] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

[35] Fouzi Harrou, Abdelkader Dairi, Abdelhakim Dorbane, and Ying Sun. Enhancing wind power prediction with self-attentive variational autoencoders: A comparative study. *Results in Engineering*, 23:102504, 2024.

[36] Christopher Williams and Carl Rasmussen. Gaussian processes for regression. *Advances in neural information processing systems*, 8, 1995.

[37] Andrew Wilson and Ryan Adams. Gaussian process kernels for pattern discovery and extrapolation. In *International conference on machine learning*, pages 1067–1075. PMLR, 2013.

[38] scikit-learn developers. Illustration of prior and posterior gaussian process for different kernels. `https://scikit-learn.org/stable/auto_examples/gaussian_process/plot_gpr_prior_posterior.html`, 2024. Accessed: August 30, 2025.

[39] Oliver Grothe, Fabian Kächele, and Mira Watermeyer. Analyzing europe's biggest offshore wind farms: a data set with 40 years of hourly wind speeds and electricity production. *Energies*, 15(5):1700, 2022.

[40] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146 (730):1999–2049, 2020.

[41] Iberdrola. Saint-brieuc offshore wind farm. `https://www.iberdrola.com/about-us/what-we-do/offshore-wind-energy/saint-brieuc-offshore-wind-farm`, 2024. Accessed: 29 August 2025.

[42] Abderrahim Bentamy and Denis Croize-Fillon. Spatial and temporal characteristics of wind and wind power off the coasts of brittany. *Renewable Energy*, 66:670–679, 2014.

[43] Iberdrola. Baltic eagle offshore wind farm. `https://www.iberdrola.com/about-us/what-we-do/offshore-wind-energy/baltic-eagle-offshore-wind-farm`, 2025. Accessed: 2025-08-29.

[44] Ewa Dąbrowska and Mateusz Torbicki. Forecast of hydro–meteorological changes in southern baltic sea. *Water*, 16(8):1151, 2024.

[45] Beatrice Offshore Windfarm Ltd. Beatrice offshore wind farm. `https://www.beatricewind.com/`, 2025. Accessed: 2025-08-29.

[46] Claire L Vincent, Pierre Pinson, and Gregor Giebela. Wind fluctuations over the north sea. *International Journal of Climatology*, 31(11):1584–1595, 2011.

[47] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[48] Davies Dl. A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell*, 1: 224–227, 1979.

[49] Sara Abreu, Fátima Rodrigues, and João Pereira. Clustering of renewable energy assets to optimize resource allocation and operational strategies. *Journal of Intelligent Information Systems*, pages 1–23, 2025.