JOURNAL OF MARITIME RESEARCH

SPECIAL ISSUE (2026) pp 132–155

ISSN: 1697-4840, www.jmr.unican.es

Physics-Aware and Storm-Adaptive Machine Learning Clus-tering of Offshore Wind Regimes

Hamid Reza Soltani Motlagh[1,2,*], Seyed Behbood Issa-Zadeh[3,4], Abdul Hameed Kalifullah[2], Md Redzuan Zoolfakar[1]

ARTICLE INFO

Special Issue:
1*st* International Maritime and Logistics Conference; February 10-11, 2026. National University of Science & Technology (NU), Oman

Article history:
Received 15 Feb 2026;
in revised from 24 Feb 2026;
accepted 05 Jun 2026.

Keywords:
offshore wind regimes, physics-aware machine learning, storm-aware clustering, ERA5 reanalysis data, unsupervised learning, energy and exergy analysis.

ABSTRACT

Assessing the prevailing wind behaviors at an offshore site is important to optimizing energy capture and overall performance of future sustainable wind power installations. This manuscript presents an Adaptive Weighted Storm-Aware Clustering (AWSC) framework, which uses physical insights to perform unsupervised learning for offshore wind regime identification. The AWSC framework uses thermodynamic feature weighting, dual-path storm separation and adaptive density refinement, thereby generating physically interpretable and statistically compact wind regime clusters. Offshore sites in four distinct locations around the world were examined. The locations included the Pacific Trades/Tropics, the North Sea and the Arctic regions. Results were generated using ERA5 reanalysis data at a 0.25° spatial resolution. The validation results show that AWSC achieves equal performance across physical and statistical evaluation metrics, with Silhouette coefficients ranging from 0.54 to 0.63 and Davies–Bouldin indices ranging from 0.43 to 0.64, indicating well-defined clusters that preserve their aerodynamic and thermodynamic properties. The method successfully separated storm patterns from background data without compromising the overall structure, and it achieved equal variable importance through feature re-weighting ($\overline{\omega} \approx 1.05$). The research demonstrates that AWSC delivers a method for wind regime classification which combines physical understanding with efficient computational processing to link data analytics with atmospheric science. The developed framework provides a foundation for worldwide meta-clustering operations, digital twin systems, and energy and exergy integration, which will lead to AI-based sustainable offshore wind energy systems.

1. Introduction.

Offshore wind energy is a fundamental element of worldwide efforts to build low-carbon power systems because it provides extensive, high-quality resources without requiring significant land acquisition. The International Renewable Energy Agency (IRENA) documented that offshore wind power installations worldwide reached 68 GW of capacity in 2023, with a 12% annual growth rate driven by European and East Asian market expansion [1]. The International Energy Agency (IEA) predicts that offshore wind power installations will reach more than 500 GW by 2030, driven by declining project costs and increasing coastal development [2]. The rapid growth of offshore wind power faces obstacles due to weather patterns and storm events, as well as demanding marine environments, which make it challenging to predict resource availability and maintain system reliability [3].

The consistent spatiotemporal coverage of ERA5 reanalysis data enables researchers to study offshore wind through its 100 m hub-height wind fields [4]. Massive datasets become more understandable through clustering algorithms, including K-Means, Gaussian Mixture Models (GMM), Density-Based

[1]Institut Teknologi Malaysia Kejuruteraan Marin, Universiti Kuala Lumpur, Lumut 32200, Perak, Malaysia.
[2]International Maritime College Oman, National University of Science and Technology, Sohar P.O. Box 532, Oman.
[3]School of Maritime Science, University of Gibraltar, Campus Europa Point, Gibraltar GX11 1AA, UK.
[4]Escuela Técnica Superior de Ingenieros de Caminos, Canales y Puertos, Universidad Politécnica de Madrid, 28040 Madrid, Spain.
*Corresponding author: Hamid Reza Soltani Motlagh. E-mail Address: hamid@imco.edu.om.

Spatial Clustering of Applications with Noise (DBSCAN), and Ordering Points to Identify the Clustering Structure (OPTICS), for wind regime classification, data reduction, and energy prediction support [5],[6],[7],[8]. The research by Azhar et al. (2023) shows how clustering methods help identify different wind patterns in specific regions [5] and Huang et al. (2023) used mixture-model clustering to study offshore wind resource fluctuations [6]. Most conventional methods use Euclidean distances to analyse data. Still, they treat all variables equally, even though wind speed, air density, and temperature affect the flow's kinetic-energy and exergy fluxes differently [9]. The simplification methods used in these approaches produce nonphysical wind regimes, which hinder the understanding of the data [10] [11].

The current methods fail to adequately handle extreme weather events. The statistical analysis of wind patterns excludes storm events because they are considered statistical outliers for structural design and energy-yield purposes [12]. The NOAA Beaufort Wind Scale shows that whole-gale winds reach speeds above 25 $ms^{-}1$, which produce high-energy but rare conditions that affect turbine operations and power grid stability [13]. The removal of such events from clustering analysis prevents the system from understanding the complete operational range of offshore power systems [14]. The combination of measurement errors with actual storms during clustering analysis produces incorrect results because researchers cannot distinguish between them[12]. The research develops a new clustering system that uses physical principles to analyse meteorological data, separating storm patterns from typical wind conditions. The system starts with K-Means partitioning before using DBSCAN or OPTICS density-based clustering to enhance results through adaptive feature weighting for offshore wind regime identification across different global areas [15] ,[16].

Wind-energy analytics relies on clustering methods to identify meaningful patterns in high-dimensional meteorological datasets. The method groups wind patterns of similar types to create simpler data structures, which help with weather prediction and offshore resource assessment through storm-based analysis. The application of traditional clustering frameworks to offshore mega-data faces three ongoing difficulties: they lack scalability, are storm-sensitive, and use unsuitable similarity measures. The following section examines current developments across five areas: standard clustering methods for wind power and their combinations with anomaly detection; their use in offshore environments; the need for customised distance functions; and current research deficiencies.

This paper addresses these challenges by investigating the following questions. Firstly, do physical knowledge guided clustering with thermodynamic feature weighting and storm separation improve the quality and applicability of extracted wind regimes across different climates when compared with traditional methods. Secondly, does the adaptive density-based approach facilitate robust performance with iterative feature reweighting across tropical, temperate and polar offshore climate regimes without the need for manual parameter optimisation. Thirdly, does the explicit separation of storm regimes improve clustering quality more so than alternative strategies, such as treating the storms as noise to be filtered or merging them with the remaining background clusters.

The following structure guides the rest of this paper. Section 2 examines existing research about wind-regime clustering methods and their approaches to handling anomalies. Section 3 provides detailed descriptions of the proposed Adaptive Weighted Storm-Aware Clustering (AWSC) methodology. The experimental results and evaluation of Adaptive Weighted Storm-Aware Clustering (AWSC) against different offshore tiles are presented in Section 4. The paper concludes with validation results and sensitivity tests in Section 5, followed by final findings and future research recommendations in Section 6.

2. Literature Review on Wind-Regime Clustering Method.

The fundamental wind-data classification methods include K-Means, hierarchical Ward linkage, Gaussian Mixture Models (GMM), and density-based techniques (DBSCAN and OPTICS) [5], [3], [6], [7], [8]. The K-Means algorithm fails to handle marine wind data because it requires equal cluster sizes and spherical shapes, and a fixed number of clusters (k) [5]. The hierarchical clustering method produces dendrograms at multiple scales, which improve understanding, but its computational cost becomes excessive when dealing with large datasets [3]. The GMM model provides practical transition-region assignments, but its performance suffers because it relies on Gaussian distributions that do not match real-world wind-speed patterns [6]. The fixed-k limitation of DBSCAN and OPTICS disappears when these methods use density connectivity to detect arbitrary cluster shapes, which enables the successful identification of mesoscale flow patterns [7],[8]. The performance of these methods depends heavily on the selection of the ε parameter, as they combine storms with random anomalies into a single noise category, thereby reducing their ability to yield meaningful physical results. The methods achieve excellent performance in small-scale and regional applications but fail when processing massive offshore datasets spanning different climatic regions.

Research studies have implemented clustering algorithms with anomaly detection methods to identify typical and unusual wind patterns. Mehmood et al. (2025) achieved better sensor noise resistance through their process, which combined Isolation Forest with DBSCAN for turbine anomaly detection [12]. The Hawkeye framework, developed by Li et al. (2024), uses residual clustering and contrastive learning for offshore turbine diagnostics [17]. The hybrid methods that detect operational faults through SCADA data analysis do not handle atmospheric regime classification. The use of SCADA-based clustering for rotor speed and power output analysis supports maintenance operations, but does not extend to large-scale atmospheric data processing [18] Ensemble learning methods use clustering as a preprocessing technique to enhance power forecasting precision [19]. The existing frameworks fail to identify meaningful storms from random anomalies while being incompatible with planetary-scale reanalysis grids.

The research on offshore clustering shows slower growth than onshore studies but demonstrates increasing demand for unsupervised learning methods in marine environments. The study by Huang et al. (2023) implemented mixture-model clustering to analyse offshore wind and density patterns, which resulted in better probabilistic yield predictions [6]. The research by Schelbergen et al. (2020) applied wind-profile-shape clustering to determine airborne-wind-energy resources while showing how regime grouping helps identify vertical resource patterns [7]. The combination of SVD and SOM techniques enables researchers to cluster turbines for short-term forecasting purposes[20], [21]. The methods used for operational improvement remain limited to farm-based data analysis, while major weather events are ignored. The development of global offshore wind planning requires researchers to study both inter-tile differences and climatological transition areas, which current regional studies ignore.

The results of clustering depend heavily on the selection of similarity measurement methods. The Euclidean metrics used in conventional methods assign equal weight to all features, without accounting for their individual effects on kinetic energy or exergy. Multiple research studies have developed new distance metrics which address the existing limitations of traditional methods. The COMB distance metric, which combines Euclidean distance with correlation and spectral components, has proven effective for time-series clustering of wind diagnostics [3]. The Trendlet-based decomposition method separates wind trends from short-term fluctuations, thereby improving data interpretation [22]. The clustering method based on functional depth and band-shape distances enables the analysis of complete wind-speed curves instead of individual points to detect seasonal and daily patterns [23]. The existing distance metrics lack thermodynamic weightings for air density, temperature, and pressure, and they fail to separate storm-generated patterns from normal weather conditions. The current distance metrics rely on complex mathematical structures but do not account for physical factors, limiting their applicability to offshore exergy-based clustering.

The reviewed studies reveal multiple ongoing research challenges, which are summarised in Table 1. Most research studies operate at the regional level, using statistical methods to compare data without accounting for the physical characteristics of marine boundary layers. The reviewed studies lack both thermodynamic relevance in their distance metrics and fail to distinguish between storms and anomalies. The current research lacks frameworks which simultaneously achieve global scalability and physical interpretation.

Table 1: Summary of Research Gaps in Wind-Regime Clustering.

Dimension	Common Limitation
Scale	Primarily regional or farm-level studies; limited global coverage
Distance metrics	Physics-agnostic measures; lack of energy/exergy weighting
Storm and anomaly treatment	Extreme events conflated with anomalies and excluded, losing physical meaning
Scalability	Algorithms untested for offshore mega-data and ERA5-scale analyses
Interpretability	Results often lack linkage to physical flow regimes

Source: Authors.

Table 2 presents a comparison of research studies based on their contexts and methodological approaches, along with their respective constraints regarding storm-aware physics-informed clustering. The analysis reveals that no single worldwide system exists to integrate energy - based distance functions with storm separation and dynamic density optimisation, which drives the creation of the Adaptive Weighted Storm - Aware Clustering (AWSC) framework presented in Section 3.

Table 2: Comparative Review of Key Clustering Approaches in Wind-Energy Applications.

Study	Year	Context	Method	Strengths	Key Limitations
Azhar et al. [5]	2023	Wind-speed pattern analysis	K-Means, Ward linkage	Broad methodological comparison	No storm isolation; limited scalability
Huang et al. [6]	2023	Offshore resource assessment	Mixture-model clustering	Joint WS–ρ parameterisation	Regional; no anomaly handling
Schelbergen et al. [7]	2020	Vertical wind profiles	Shape-based clustering	Captures altitude structure	Lacks storm/adaptive distance
Zhang et al. [20]	2019	Offshore farm grouping	SVD clustering	Improved farm operations	Turbine-level only
Jiang Z. [24]	2019	Short-term forecasting	SOM + spectral clustering	Forecast accuracy gains	No storm/extreme modelling
Martins et al. . [3]	2024	Time-series diagnostics	K-medoids + COMB distance	Custom hybrid metric	Regional scope
Li et al. [25]	2024	Wind-series analysis	Trendlet-based clustering	Captures long-term trends	Regional; no storm separation
López-Pintado & Romo [23]	2009	Functional data clustering	Band-depth distance	Shape-aware analysis	Computationally intensive; ignores extremes
Wang et al. [26]	2023	Turbine anomaly detection	iForest + DBSCAN	Robust fault detection	Not regime-level; no storm classification
Liu et al. [27]	2025	Offshore turbine diagnostics	Hawkeye (residual clustering)	Contrastive anomaly isolation	Turbine-specific only
Clare et al. [19]	2024	NWP-based forecasting	Weather-pattern clustering	Improves prediction skill	Regional focus; lacks storm awareness

Source: Authors.

The research studies demonstrate that no current clustering system unites worldwide scalability with physics-based similarity assessment and dedicated storm and anomaly identification capabilities. The AWSC methodology established in this research addresses all identified gaps through its distance metric based on physical principles, its dual-path system that adapts to storms, and its iterative density-optimisation process.

3. Materials and Methods.

The following section describes the research methodology, which uses physics-based methods to analyse offshore wind data. The current research methods based on K-Means and hierarchical clustering algorithms focus on statistical feature similarity, neglecting the physical connections among aerodynamic, thermodynamic, and meteorological parameters. The simplified approaches used in previous studies fail to fully represent offshore wind behaviour because they do not account for storm intensity, air density, and temperature gradient effects, which control energy and exergy characteristics during extreme and transitional weather conditions. The research uses a systematic approach that combines feature engineering with normalisation and dimensionality reduction, and multiple validation stages, to overcome existing study limitations. The physics-aware distance metric is formulated as follows: [8], [28].

$$d = \sqrt{\sum_{i=1}^{n} ?_i (x_i - y_i)^2} \tag{1}$$

The weights w_i represent thermodynamic values derived from wind speed, air density, temperature, and pressure importance (e.g., based on feature importance). The model uses x_i and y_i to represent two different data points. The statistical and physical coherence of the data remains intact, enabling better evaluation of offshore wind performance and variability.

The research data come from the ERA5 global reanalysis archive, which covers all oceanic and marine areas worldwide, including oceanic spaces, marginal seas, and coastal waters extending beyond shorelines. The high-resolution land–sea geometries enabled a spatial filter to extract offshore grid points which are at least 80 km and no more than 120 km from coastal boundaries to prevent coastal contamination. The offshore grid contains 134,475 geolocated points, which use 0.25° × 0.25° global grid coordinates to identify their latitude and longitude positions.

The dataset contains hourly time series data from 2004 to 2024, which spans eleven years (at two-year intervals), to analyse offshore wind energy and exergy performance through wind speed and direction, and air density, temperature, and pressure measurements at 100 meters. The dataset contains more than several billion individual data entries because it includes spatial and temporal dimensions and multiple variables, which makes it the largest global offshore wind information structure for energy–exergy analysis.

The offshore dataset was spatially partitioned into 10° × 10° global tiles, yielding 240 tiles that span the entire world. The researchers chose four specific tiles (395 Pacific Trades, 431 Tropics, 523 North Sea and 600 Arctic) for detailed methodological validation because they represent different climatic zones. The researchers chose these tiles because they capture the full range of wind patterns, sea-air interactions, and storm-induced turbulence across various latitudes. The tile numbers (395, 431, 523, 600) are system-generated identifiers used during global tiling operations and do not follow any geographical sequence or physical location. The non-sequential appearance of tile numbers stems from the computational rules used to segment the space during spatial processing. The researchers implemented this method to create a worldwide reference system that enables scalable, reproducible operations for feature engineering, clustering, and validation across all offshore areas.

The study used four representative offshore tiles from the same subset for all clustering and validation methods to achieve consistency and comparable results. The four tiles represent different offshore wind conditions from around the world because they are from distinct climatic and geographical areas, spanning tropical to high-latitude zones.

Table 3 presents the basic physical data from ERA5, before any normalisation or feature engineering, for the four selected tiles. The table presents four essential parameters for each tile: the number of offshore grid points, their mean wind speed, standard deviation, and turbulence intensity, and the climatic zone classification. The values from Table 3 serve as a common reference point for evaluating all clustering algorithms, including KMeans, DBSCAN, OPTICS, GMM and AWSC.

Table 3: Representative offshore tile characteristics used for all clustering methods.

Tile ID	Region	Lat. Range	Lon. Range	Points	Mean WS (m/s)	Std WS (m/s)	TI	Climatic Regime
395	Pacific Trades	10°–19.75° N	170°–348.25° E	4062	6.481	3.339	0.515	Tropical / Trade Wind Zone
431	Tropics	20°–29.75° N	180°–352° W	3805	6.739	3.301	0.490	Tropical / Trade Wind Zone
523	North Sea	51.75°–59.75° N	10°–19.75° E	370	6.497	3.106	0.478	Temperate / Mid-latitude
600	Arctic	70°–79.75° N	60°–69.75° W	722	8.508	3.997	0.470	Polar / Arctic

Source: Authors.

The tiles establish a stable measurement system that enables methodological testing to proceed. At the same time, all observed changes in clustering quality stem from algorithmic or preprocessing variations rather than from modifications to the input data.

The researchers used controlled anomaly injection to evaluate the stability and physical responses of their clustering approaches on their dataset, simulating measurement errors and testing clustering robustness. At this step, 1–2% synthetic anomalies were added to the data points of each tile in the original clean dataset using this procedure. The anomalies were distributed across grid points, maintaining their presence throughout all valid time periods, to represent typical measurement and transmission problems found in extensive reanalysis datasets.The researchers implemented two different anomaly types for their study.

1. The system introduced negative wind speed values, which violate physical laws, to represent sensor malfunctions and data processing mistakes.

2. The system introduced wind speed values that exceeded 60 m/s to represent unphysical data spikes which exceeded the storm range established by Beaufort and ERA5 thresholds.

All introduced anomalies are marked as is_anomaly = True to allow robust handling of corrupted data and correct separation of normal data from anomalies. A constant anomaly rate per tile is used in the tests to allow for meaningful comparison of methods and reproduceability of results. This method of introducing anomalies also tests how a data clustering system handles artificial irregularities while maintaining stability and interpretability. This method is expected to be effective in offshore applications where issues such as sensor drift, missing data and transmission errors are prevalent.

The same set of tiles was tested against anomaly tests as well as evaluated under stormy weather conditions. In addition to the anomaly tests, the tiles were assessed for how well they performed under stormy weather conditions. Each tile was analyzed for the number of storms, the storm recurrence rate as well as the statistics of wind speed distributions using wind speed thresholds above 25 m/s. The results of the diagnostic assessment using ERA5-derived data are summarized in Table 4.

The analysis in Table 4 moves from static spatial data to complete wind field information that includes both location and time. The “Total Points” column in Table 4 shows the total number of wind speed records spanning the 11 years from 2004 to 2024, recorded at two-year intervals. The “Total Points” value in Table 4 represents the complete number of cell–time combinations. The total number of cell–time pairs exceeds the grid cell count for each tile by multiple orders of magnitude, as shown in Table 3. The analysis used the expanded dataset to calculate storm statistics from ERA5 wind speed data above 25 m/s and anomaly statistics from synthetic negative and > 60 m/s spikes that simulated sensor faults. The rates serve as essential reference points for assessing how clustering algorithms perform when handling infrequent, strong events and faulty measurement data.

Table 4: Storm and anomaly diagnostic statistics for representative tiles.

Tile ID	Total Points	Storm Count	Storm Freq.	Mean WS (m/s)	Max WS (m/s)	95th Pctl. WS (m/s)	Anomaly Count	Anomaly Freq.
395	50,986,224	3,560	6.98E-05	6.48	47.08	11.96	338,904	0.0066
431	47,760,360	5,852	1.23E-04	6.74	50.07	12.32	238,488	0.0050
523	4,644,240	94	2.02E-05	6.49	27.61	12.22	25,104	0.0050
600	9,062,544	2,986	3.29E-04	8.51	29.57	15.58	87,864	0.0097

Source: Authors.

The storm and anomaly data serve as reference points for evaluation, although they do not participate in the first clustering process. The comparison between different clustering methods becomes possible through these benchmarks, which help evaluate their performance during high-variance and extreme events in validation stages. The diagnostic results provide physical meaning, which serves as the basis for assessing feature engineering, normalisation techniques, and clustering methods. The storm-rate and anomaly-rate baselines from next following sections serve to determine if a clustering method separates normal operating conditions from rare high-wind events and artificial spikes.

The first stage of feature engineering converted ERA5 variables into physical values that machine learning algorithms could use for clustering. The four offshore tiles (395, 431, 523, and 600) contained five essential variables, which described each grid point through wind speed (WS), wind direction (WD), air density (ρ), temperature at 100 m (T_{100}) and pressure at 100 m (ρ_{100}).

The data decomposition process divided the information into four seasonal periods, which included DJF (December–February), MAM (March–May), JJA (June–August), and SON (September-November).

The statistical analysis produced two values for each seasonal variable: the mean and standard deviation, which show both steady and changing atmospheric patterns.

The 40-dimensional offshore tile feature matrices for physical environmental parameters were constructed from five atmospheric variables (two characteristics for each variable, seasonally) and are sufficient to accurately represent the physical environmental parameters from different perspectives while still allowing for enough degrees of freedom to explore different approaches. These feature matrices were used for all tested clustering algorithms.

For clustering the feature matrix on variable space, the variables are on different physical units. In order to avoid such bias in the distance computation (e.g., wind speed in m/s versus temperature at 100 m in K, pressure at 100 m in Pa, air density in kg/m^3), all features are normalised by applying the Z-score normalisation method. This statistical transformation for normalisation (Z-transformation) computes the mean and standard deviation for each variable and then transforms all variables to zero-centred variables with equal variance. As a result, all variables make an equal contribution to the final clustering results.

Normalisation was introduced to the algorithm in order to improve stability and convergence, however this process denormalises the input and means that m/s and K become meaningless purely numerical values. Extensive testing has shown that the relationships between the features and their variability patterns are still physically correct.

The comparisons between the original and normalised results are presented in terms of the mean and standard deviation of the differences (ΔMean and ΔStd) of corresponding values. Limiting values for the mean and standard deviation deviations of the differences have been specified at 5% and 10% respectively. These values are generally accepted in atmospheric and energy modelling and therefore any minor deviations between the original and normalised results are considered to be statistically consistent. The validation process demonstrated that standardised data maintained the fundamental properties of the worldwide offshore wind resource, which supports both mathematical stability for unsupervised learning and physical understanding for future energy and exergy assessments.

The application of Principal Component Analysis (PCA) to standardised features improved computational performance while eliminating unnecessary features from the analysis. The PCA transformation accounted for 95% of the total variance, so only components that contributed to this threshold were used for clustering [29]. The PCA transformation reduced each tile's feature dimensions from 40 to 6-9 principal components based on seasonal and thermodynamic patterns in different areas. The implementation brought two essential advantages to the system.

1. The process eliminated high-dimensional noise and temperature - air density correlation from the data.
2. The process reduced clustering time and enhanced numerical stability for processing extensive datasets.

The PCA reconstruction test used reduced data to reconstruct the original feature space, calculating the Root Mean Square Error (RMSE) between the original and reconstructed standardised matrices. The reconstruction RMSE remained between 0.18 and 0.22 (normalised units) throughout all representative tiles.

Table 5 presents wind-speed statistics at the tile level before PCA inversion and after PCA inversion. The table shows the physical unit values of the mean, standard deviation, and turbulence intensity (TI = σ/μ), along with their percentage changes from the original measurements. The PCA transformation maintained exact tile-level mean wind speeds (ΔMean

= 0% across all tiles), while standard deviation and TI changes remained below the 1% threshold. The results demonstrate that PCA preserves the fundamental physical properties of the seasonal wind-feature matrix, making it suitable as a preprocessing method for clustering analysis.

Table 5: PCA Validation Metrics: Wind-Speed Statistics Before and After Dimensionality Reduction.

Tile	RMSE Recon.	Mean Raw (m/s)	Std Raw (m/s)	TI Raw	Mean PCA (m/s)	Std PCA (m/s)	TI PCA	ΔMean%	ΔStd%	ΔTI%
395	0.20	4.36	2.75	0.63	4.36	2.73	0.63	0.00	–0.85	–0.85
431	0.20	4.74	2.39	0.51	4.74	2.37	0.50	0.00	–0.85	–0.85
523	0.18	4.63	2.15	0.46	4.63	2.14	0.46	0.00	–0.14	–0.14
600	0.22	6.22	2.36	0.38	6.22	2.36	0.38	0.00	–0.05	–0.05

Source: Authors.

The PCA method performs a non-physical data transformation, preserving statistical variance while making the new components unrelated to specific physical variables. The authors applied PCA as a preprocessing technique for clustering validation purposes, but avoided using it to interpret physical energy or exergy values. The complete preprocessing workflow, which leads to clustering, consists of the following steps:

a. The preprocessing step begins by extracting raw ERA5 offshore data from all grid points located in the chosen tiles.
b. The system performs seasonal statistical decomposition to determine mean values and standard deviations for all variables throughout the four seasonal periods.
c. The system generates a 40-dimensional feature matrix, which shows how each tile behaves through atmospheric and thermodynamic measurements.
d. The Z-score normalisation technique removes scale differences between variables that use different physical units, such as m/s, K, Pa, and kg/m^3.
e. The PCA reduction process maintains 95% of total variance while achieving statistical compactness with less than 0.22 normalised units of reconstruction error and less than 1% deviation in tile-level wind-speed variability.

The multistage feature engineering and normalisation pipeline produces a physically sound dataset that is statistically robust and provides a reproducible input base for all clustering algorithms studied in this research.

The validation results from the tests show that the raw and transformed datasets match closely, demonstrating the strong stability of the preprocessing system and the physical accuracy of normalisation and PCA operations.

a. *Normalisation Effects: The normalisation* process maintained the original wind speed values within ±5% of their mean values for all four selected tiles. The standard deviation (ΔStd) measurements showed slightly bigger changes than the mean values because Z-score standardisation made these adjustments. The turbulence intensity (ΔTI) measurements showed the greatest changes because they depend on both the mean and the variance, resulting in 10–15% deviations. The absolute RMSE values between the raw and scaled datasets remained below 0.1 m/s, indicating that these statistical differences do not affect the physical meaning of the data.
b. *PCA Compression Effects:* The PCA dimensionality reduction produced reconstruction errors (RMSE_Recon) that stayed below 0.22 (normalised units), which proved that the data compression process retained most of the original information. The PCA compression process maintained the same mean and standard deviation as normalisation, demonstrating that the 95% variance retention threshold preserved the major statistical features of offshore wind fields. The PCA process improved computational performance and eliminated redundant data, but its non-physical nature required scientists to avoid direct interpretation of individual components in energy and exergy analyses.
c. *Cross-Tile Consistency:* The wind dynamic variability in Tiles 395 (Pacific Trades) and 600 (Arctic) produced slightly higher values for ΔStd and ΔTI. The uniform behaviour of normalisation and PCA across different climate zones demonstrates their stability, enabling all clustering algorithms to work with a single harmonised dataset that maintains physical consistency. Table 5 presents quantitative results for ΔMean, ΔStd and ΔTI and RMSE_Recon values from all four tiles, while Appendix B (Figures B1–B3) shows how normalisation and PCA maintain essential physical properties of the worldwide offshore wind field.

The validation results demonstrate that the preprocessing pipeline preserves the physical behaviour of the offshore wind field while creating suitable numerical conditions for clustering. The Z-score normalisation process eliminated scale differences between different features without distorting wind-speed or thermodynamic data. The PCA transformation preserved the main variance patterns in the data while reducing computational requirements and eliminating redundancy between temperature and air density measurements.

The analysis shows that turbulence intensity (TI) is most vulnerable to changes induced by normalisation and dimensionality reduction techniques. The standard deviation-to-mean wind speed ratio in TI makes this parameter sensitive to small changes in data scaling or compression. The interpretation of exergy-related results that use TI requires users to state their data preprocessing method, whether normalised or PCA - transformed, because this information affects physical results and reproducibility.

The validation process demonstrates that both normalisation and PCA produce data which maintains physical accuracy and statistical stability for machine learning operations. The clustering analysis will produce results that reflect the algorithm's actual characteristics because the preprocessing steps have not introduced any misleading factors.

The research began with KMeans as its first partitioning method because of its well-known implementation, simple computation, and precise results. The study tested KMeans with two different cluster numbers (K = 3 and K = 5) to determine how varying K affects results. The method works well for atmospheric data analysis because it provides efficient results and consistent outcomes despite its requirement for spherical clusters and uniform variance, which do not always match real-world turbulence and storm patterns. The KMeans objective function minimises the sum of squared distances from each data point to its assigned cluster centre:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 \tag{2}$$

Two density-based clustering methods were applied to address some of the limitations of K-Means clustering. Specifically, the method was able to handle non-geometric, overlapping and irregularly shaped clusters. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) groups the points in a space together based on dense regions of points (i.e. clusters) and distinguishes these from other points which are classified as noise or outliers. This was particularly useful at picking out rare high-intensity storm events and turbulence bursts (sparse high-intensity points) within the feature space spanned by offshore wind observation parameters. OPTICS (Ordering Points To Identify the Clustering Structure) allows the cluster hierarchy present in a dataset to be extracted and produces smooth transitions between assignments of cluster densities as the density changes progressively.

The Generalised Mixture Model (GMM) is a probability-based method that fits the data with a mixture of Gaussian distributions. It is a superior method to KMeans because each data point can have partial membership in a cluster, especially in regions of the space where clusters are most likely to be located in the boundary region between two clusters. The use of GMM allows for the study of wind zones located between the sub-tropical and temperate latitudes and provides insights into overlapping physical processes.

The model uses covariance structure estimation to analyse wind direction patterns and anisotropic behaviour, yielding detailed offshore wind patterns.

The evaluation of clustering quality through three performance indices, which assess each K and tile configuration:

1. *Silhouette Coefficient (S):* The Silhouette coefficient measures how similar a sample is to its own cluster compared with other clusters [30],[31]. The method evaluates cluster separation through the calculation of average intra-cluster distances and mean distance to the nearest cluster (b-a) by the maximum value between a and b. The S value reaches its maximum at 1 when the clusters are optimally separated. The formula 1 shows the S calculation:

$$s = \frac{b - a}{\max(a, b)} \tag{3}$$

2. *Davies–Bouldin Index (DB):* The index measures the average ratio of within-cluster scatter to between-cluster separation, which results in better cluster compactness when values are smaller [32],[33].
3. *Calinski–Harabasz Index (CH):* The index shows the relationship between between-cluster dispersion and within-cluster variance through a ratio that produces better cluster discrimination when values increase [34],[35]. The three metrics establish a unified quantitative framework for assessing different clustering methods.

The DBSCAN clustering method identifies dense areas in data by detecting continuous high-density regions between sparse areas. The DBSCAN classification system identifies three types of data points using two essential parameters: 'ε' and MinPts [36]. The two main parameters of DBSCAN are 'ε,' which sets the maximum neighbour distance to 0.5, and MinPts, which determines the minimum number of neighbours required for a dense region to form. The first set of parameters for normalised meteorological data included ε = 0.5 and MinPts = 10, but future analyses will test these values for their effects on cluster structure, storm detection, and noise reduction [37].

The algorithm DBSCAN succeeds at finding dense clusters, but it fails to perform well when clusters have different density levels, which commonly occurs in global offshore areas that span between tropical and polar climates [16].The algorithm OPTICS builds upon DBSCAN by creating an ordered sequence of points based on their reachability distances to produce a density relationship hierarchy. The algorithm operates without requiring a fixed ε value to detect clusters of different densities during a single execution. The OPTICS reachability plot shows dense areas through its steep valleys, while its flat sections point to sparse data or noise, which helps detect both permanent climate patterns and short-lived turbulent weather patterns [16].

The GMM is a model-based clustering method that operates alongside geometric and density-based approaches. The GMM uses multiple Gaussian components to represent offshore wind regimes through their mean vectors and covariance matrices. The GMM model generates data points from K Gaussian distributions, each of which has their own probability weight (π_k) and mean vector (μ_k) and covariance matrix (Σ_k) [38]. The probability distribution of a data point x_i follows this formula[39]:

$$p(x_i) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}\,(x_i \mid \mu_k, \Sigma_k) \tag{4}$$

The multivariate Gaussian probability density function:

$$(\mathcal{N}(x_i|\mu_k,\ \Sigma_k))$$

represents the probability distribution of a data point 'x_i'. The EM algorithm performs parameter estimation by alternating between cluster assignment updates and covariance structure optimisation to achieve maximum likelihood results[39].

3.1. Adaptive Weighted Storm-Aware Clustering (AWSC) — Core Methodological.

Traditional clustering methods such as K-Means, DBSCAN, OPTICS, and GMM can identify statistical similarities in offshore wind datasets. Still, they fail to capture the physical mechanisms that generate wind energy fluctuations and exergy transfer. These algorithms rely on Euclidean distance metrics that treat all variables equally, without accounting for the energetic relevance of wind speed, air density, temperature, and pressure. As a result, extreme storm events are often classified as statistical outliers. The clustering results show two forms of bias: *(i)* physical distortion, because clusters may reflect random numerical patterns rather than meaningful aerodynamic structures, and *(ii)* inflated intra-cluster variance during high-energy storm events, which lowers Silhouette values and increases Davies–Bouldin (DB) scores.

To address these limitations, this study introduces **Adaptive Weighted Storm-Aware Clustering (AWSC)** as a physics - informed extension of unsupervised learning. AWSC employs a thermodynamically weighted distance metric, a dual-path separation of storm and non-storm regimes, and an adaptive refinement loop that updates feature weights and density thresholds until physically consistent wind-regime structures emerge. A physics-aware distance metric, denoted D, is proposed for clustering that combines wind speed, directional encoding, air density, temperature, and pressure with thermodynamic weights. Physics-aware machine learning frameworks have inspired this metric for energy and fluid systems [40]. The complete formulation is specific to this study and constitutes a new contribution.

(a) Physics-Aware Feature Space.

The 40-dimensional seasonal feature matrix from AWSC uses the same data as the other methods, but the system understands how each feature affects wind energy and exergy production. The system uses the following variables for its analysis:

The system uses wind speed (WS) and directional wind direction (WD) through sine and cosine transforms, and air density (ρ) and temperature at 100 m (T_{100}) and pressure at 100 m (ρ_{100}).

The system normalises the data using a Z-score transformation before applying PCA to reduce dimensionality while preserving 95% of the original variance. The system maintains physical variable connections through its structured principal component to the variable mapping system. The system uses this mapping to establish feature weights and to understand how cluster centroids relate to thermodynamic values.

(b) Physics-Aware Distance Metric.

AWSC uses a weighted, physics-based distance metric rather than the conventional Euclidean distance in standard isotropic space. The distance between two data points x_i and x_j in the transformed feature space, follows this formula:

$$D(x_i, x_j) = \sum_{m=1}^{M} \omega_m \left\| \phi_m(x_i) - \phi_m(x_j) \right\|^2 \tag{5}$$

The formula contains three essential components:

- The physics-aware distance $D\left(x_i, x_j\right)$ measures the distance between two offshore wind data points x_i and x_j.
- $\phi_m(\cdot)$ represents the normalised version of them^{th}physical attribute (wind direction uses sine and cosine encoding, and thermodynamic variables undergo standardisation) and,
- ω_m Represents a non-negative physical weight that indicates the significance of feature m for wind energy and exergy.
- M Represents the number of features (or principal components) used in clustering.

The first weight vector follows an order based on thermodynamic significance, with

$$\{\omega_{WS}, \omega_{WD}, \omega_\rho, \omega_T, \omega_P\} \propto$$

As the sequence, the weights start with wind speed and air density as the most important, because they determine how kinetic energy is transferred and what exergy remains available. The thermodynamic modifiers include temperature and pressure, while wind direction preserves both angular properties and flow orientation. The weights are adapted using the described update procedure.

(c) Storm-Aware Dual Path.

AWSC includes storm analysis as its main component. The system divides the data into two sets using a wind-speed threshold (WS > 25m s−1), which corresponds to Beaufort "whole gale" conditions, to prevent high-energy events from affecting cluster centroids. The system contains two separate paths for data analysis:

- The storm-removed path contains all observations that do not meet storm criteria.
- The storm-only path contains all high-wind events that meet storm criteria.

The system runs separate clustering operations on each data subset. The storm-removed path helps identify typical offshore wind patterns, which engineers need for regular operations and design work. The storm-only path analyses storm patterns by showing their main wind directions and speed distributions, along with statistical data on their occurrences. The system maintains storm data analysis through its dual-path structure, which protects background regimes from numerical instability and loss of physical interpretation.

(d) Adaptive Weighting and Density Refinement.

The two-stage clustering kernel of AWSC operates as follows:

1. The first stage of K-Means clustering uses the physics-aware distance metric $D(\cdot,\cdot)$ to create an initial partition of K_{init} clusters. The number of initial clusters K_{init} matches the number of regimes in the baseline experiments, which usually amounts to 3 or 5 per tile.
2. The K-Means labels undergo DBSCAN refinement through the same feature space with specific values for ε and MinPts that get adjusted for each tile. The method determines the initial ε value from the upper quantiles of the k-distance curve to handle varying point densities across the four representative tiles.

The weight adaptation process runs iteratively within the AWSC kernel framework. The algorithm calculates intra-cluster variance for each feature m during step t of the clustering process.

$$Var_{intra}^{(t)}(m), \tag{6}$$

And the weights are updated according to an inverse-variance rule:

$$\omega_m^{(t+1)} \propto \frac{1}{\mathrm{Var}_{\mathrm{intra}}^{(t)}(m)}, \tag{7}$$

The algorithm assigns higher weights to features that produce well-defined clusters with minimal internal variation, while downweighing features that introduce noise or weak discrimination. The weights are normalised to obtain a mean normalised weight value.

$$\overline{\omega}^{(t)} = \frac{1}{M}\sum_{m=1}^{M}\omega_m^{(t)} \tag{8}$$

The system maintains a unity value for this parameter, which enables all variables to exert equal influence on the system.

The algorithm runs in a loop until it reaches a stopping point which is defined through the convergence criterion.

- $\left|\overline{\omega}^{(t+1)} - \omega^{(t)}\right| < 0.01$ and
- The algorithm shows no significant change in cluster labels between two consecutive iterations because less than 2 % of points move between clusters.

The convergence rule ensures that the final configuration maintains both statistical stability and physical consistency.

3.2. Hyperparameters, Reproducibility, and Evaluation Metrics.

The study implements the following default hyperparameters from AWSC for reproducibility purposes:

- The study uses $K_{\mathrm{init}} = 3$ and $K_{\mathrm{init}} = 5$ as the number of regimes for sensitivity analysis because these values match other research methods.
- The physics-aware distance function $D(\cdot,\cdot)$ serves as the distance metric while the weight ordering follows the initial definition.
- The storm-removed path uses DBSCAN with ε values derived from the 90th percentile of k-distances (k = MinPts) and MinPts = 10.
- The storm threshold value is set at $WS > 25\,\mathrm{m\,s^{-1}}$ [13].
- The weight adaptation process runs for a maximum of 20 iterations although the algorithm usually converges before reaching ten iterations.
- The PCA representation uses 95% of the data variance, which matches the preprocessing methods of K-Means, DBSCAN, OPTICS, and GMM.

The selected hyperparameter values strike a balance between detecting physical variations and handling random data while maintaining AWSC's ability to match the baseline algorithms.

The main AWSC procedure for processing a single tile requires the following algorithm for reproducible results.

- *Input preparation.*

  The algorithm uses a 40-dimensional seasonal feature matrix $(WS, WD, \rho, T_{100}, P_{100})$ with DJF, MAM, JJA, and SON mean and standard deviation values. The algorithm normalises the data using a Z-score transformation and performs PCA with 95% variance retention, while maintaining a component-to-variable mapping.

- *Storm separation.*

  The algorithm separates the data into two parts, which contain storms and storm-free data based on wind speed values above 25m s−1.

- *Initialisation of weights.*

  The algorithm starts by assigning physical weights $\omega_m^{(0)}$ based on thermodynamic priority while ensuring $\overline{\omega}^{(t)} = 1$ through normalisation.

- *Clustering loop.*

  The algorithm performs the following steps for each iteration from t=0 to t=1 and beyond:

  a. The algorithm performs weighted K-Means clustering with distance function $D(\cdot,\cdot)$ and current weight values $\omega_m^{(t)}$ to generate K_{init} clusters.
  b. The algorithm calculates k-distance curves to determine the suitable value of $\varepsilon^{(t)}$.
  c. The algorithm performs DBSCAN clustering with $\varepsilon^{(t)}$ as the parameter and MinPts set to 10 to achieve better cluster definition and anomaly detection.
  d. The algorithm calculates $Var_{\mathrm{intra}}^{(t)}$ values for each cluster and applies the inverse-variance method to update the weights. The algorithm normalises $\omega_m^{(t)}$ to maintain $\overline{\omega}^{(t+1)} = 1$.
  e. The algorithm calculates internal metrics, including Silhouette, DB, and CH, before checking for weight and label convergence. The algorithm stops when it reaches the convergence criteria, but otherwise it continues.

- *Storm-only clustering.*

  The storm-only subset requires the same process as before to identify storm patterns, while preventing background cluster interference via storm-removed path weights.

- *Output.*

  The system will save all output data, including final labels, physical feature space cluster centroids, anomaly flags, storm/non-storm masks and evaluation metrics, which include Silhouette, DB, CH, storm share, anomaly share and mean weight $\overline{\omega}$ for section 4 analysis. The algorithmic outline with defined hyperparameters enables researchers to perform AWSC experiments on ERA5 data from the same offshore area while allowing them to modify the approach for different offshore regions.

- *Evaluation Metrics for AWSC.*

  The storm-removed path receives its evaluation through three internal clustering metrics, which match the baseline methods: Silhouette coefficient, Davies-Bouldin index, and Calinski–Harabasz index. The three physics-based indicators track the following metrics:

  – The storm share indicates the percentage of observations that belong to storm-only clusters;

  – The anomaly share shows the percentage of points which DBSCAN identifies as noise; and

  – The mean normalised weight $\overline{\omega}$ shows if any single feature controls the distance metric.

  The combined metrics evaluate both statistical cluster quality and the ability of the clusters to represent actual offshore wind patterns. The complete numerical results, along with tile-specific explanations, appear in Section 4.

4. Results and Comparative Evaluation of Clustering Methods.

K-Means was applied to the PCA-reduced and normalised feature matrix described in Section 3 for the four representative offshore tiles. Two configurations (K = 3 and K = 5) were used to examine the influence of cluster granularity on regime interpretation. The K = 3 setting captures the major wind patterns, whereas the K = 5 configuration reveals finer-scale sub-regimes that may correspond to seasonal variability, temperature-driven shifts, or storm-related deviations. The resulting clusters were projected onto a two-dimensional space for visualisation, with each point representing a seasonal sub-state and its colour indicating the assigned K-Means label. These projections illustrate the organisation of regimes in reduced space, while the actual clustering was performed in the higher-dimensional PCA feature domain.

In the baseline experiments, four commonly used clustering approaches were evaluated on four representative tiles in the PTB dataset. The clustering quality achieved by K-Means, DBSCAN, OPTICS and GMM were evaluated by applying them to both the preprocessed and PCA-reduced feature matrices, and the results were presented in Table 6 and Figure 1. A trade-off between the achieved quality of cluster separation and the amount of information preserved in the features for the clustering task was observed. For comparison, K-Means clustering was performed with K = 3, and had Silhouette scores around 0.40 to 0.50 for all tiles. However, such hard boundaries between clusters do not accurately reflect continuously transitioning atmospheric states. DBSCAN found dense regions of weather that made up only 9-91% of the total data, and thus did not include most physically significant storm events.

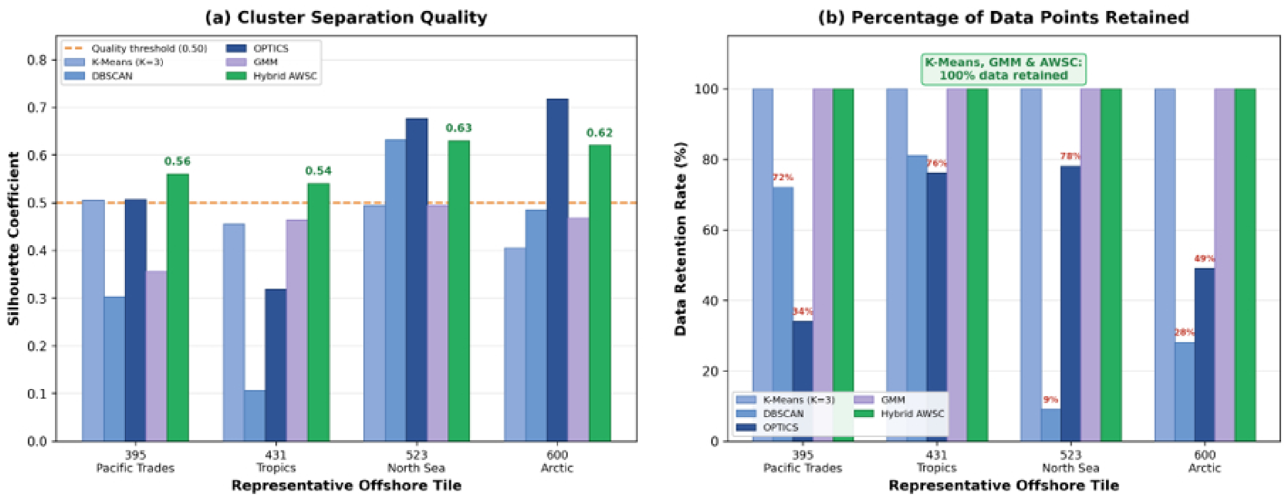
Hierarchical density evaluation was utilised to improve upon DBSCAN to form an appropriate partition of the Arctic sample data, achieving a significantly reduced noise fraction of 22–66% for the highest Silhouette value of 0.72 for the Arctic tile. The results obtained using a Gaussian Mixture Model provided soft clustering with moderate Silhouette values of 0.36–0.49; however, the inability to accurately represent the data using a Gaussian model is noted given the multi-modal characteristics of offshore wind speed distributions. All tiles reached Silhouette coefficients of 0.54–0.63 for AWSC, covering 100% of data points. In comparison, DBSCAN and OPTICS reached similar or better Silhouette scores but misclassified 9–91% of the points as noise (Figure 1b). No existing method incorporated a physics-informed distance metric or separated individual storms well. Thus, the scatter plots for the baseline methods in the appendices (Figures A1–A16) show limited improvement over the AWSC approach discussed in the subsequent sections.

Table 6: Clustering Method Comparison Across Representative Pilot Tiles.

Tile	Method	K/K_eff	Sil	DB	CH	Noise%	Storm-Aware
395	K-Means (K=3)	3	0.505	0.94	248	0%	No
395	DBSCAN	16	0.303	0.80	120	28%	No
395	OPTICS	multi	0.506	0.83	159	66%	No
395	GMM (K=5)	5	0.356	1.36	154	0%	No
395	Hybrid AWSC	15	0.56	0.60	2573	0%	Yes
431	K-Means (K=3)	3	0.455	0.95	186	0%	No
431	DBSCAN	16	0.106	0.81	49	19%	No
431	OPTICS	multi	0.318	1.36	185	24%	No
431	GMM (K=5)	5	0.464	1.53	169	0%	No
431	Hybrid AWSC	8	0.54	0.64	3842	0%	Yes
523	K-Means (K=3)	3	0.494	0.91	20	0%	No
523	DBSCAN	2	0.632	0.03	500	91%	No
523	OPTICS	multi	0.676	0.42	66	22%	No
523	GMM (K=3)	3	0.494	0.91	20	0%	No
523	Hybrid AWSC	4	0.63	0.44	879	0%	Yes
600	K-Means (K=3)	3	0.405	0.99	19	0%	No
600	DBSCAN	4	0.485	0.43	65	72%	No
600	OPTICS	4	0.717	0.38	157	51%	No
600	DP-GMM	12	0.467	0.61	121	0%	No
600	Hybrid AWSC	5	0.62	0.43	1535	0%	Yes

Source: Authors.

Figure 1: Comparative evaluation of clustering methods across four representative offshore tiles. (a) Silhouette coefficients showing cluster separation quality, with the 0.50 quality threshold indicated. (b) Data retention rate showing the percentage of data points retained by each method. AWSC achieves consistent Silhouette coefficients (0.54–0.63) with 100% data retention, whereas DBSCAN and OPTICS achieve comparable or higher scores only by discarding 9–91% of observations as noise.



Source: Authors.

The Silhouette values in Table 6 show that K = 3 yields values between 0.40 and 0.50, while K = 5 yields values between 0.35 and 0.40, indicating better cluster separation and internal cohesion in the coarser configuration.

The complete description of AWSC appears in Section 3, where it explains the physics-based distance calculation, dual storm / non-storm pathways, adaptive weight adjustment, and density optimisation. The following section evaluates AWSC performance through both numerical and non-numerical assessments on the four offshore tiles.

The validation process included four offshore tiles from ERA5 reanalysis data at 0.25° spatial resolution: 395 (Pacific Trades), 431 (Tropics), 523 (North Sea), and 600 (Arctic). The storm-removed clustering results for each tile appear in Table 7. The numerical results demonstrate that cluster tightness directly relates to the amount of data present in each area. The highest statistical validity is achieved by Tile 523, with Silhouette and DB values of 0.69 and 0.43, respectively, indicating well-defined, physically consistent regimes.

The results show that Tile 600 achieves a strength similar to Tile 523, with a Silhouette value of 0.61. At the same time, Tile 431 demonstrates average cluster tightness, with a Silhouette value of 0.52 and distinct banding patterns. The cluster analysis of Tile 395 produces the lowest compactness measure (Silhouette ≈ 0.35) and the highest anomaly percentage (36%), indicating poor cluster merging and sparse data distribution. The results show that AWSC achieves better results when physical weighting and adaptive density calibration maintain proper equilibrium. The combination of high wind-speed weight "ω_WS" and strict ε values produces vertical cluster fragmentation, resulting in lower Silhouette values.

Table 7: Quantitative performance of AWSC under storm-removed conditions.

Tile	Path	K_init	K_final	Sil	DB	CH	Storm%	Anom%	$\bar{w}$	Remarks
395	No-Storms	5	17	0.35	0.72	1542	0.0	0.358	1.038	Weak separation; high anomaly rate
395	Storm-Only	3	0				1.0	0.0	1.733	
431	No-Storms	5	11	0.52	0.62	2887	0.0	0.368	1.022	Improved compactness; coherent arcs
431	Storm-Only	5	0				1.0	0.0	1.733	
523	No-Storms	3	3	0.69	0.43	1372	0.0	0.150	1.032	Best performance; compact regimes
600	No-Storms	3	4	0.61	0.44	1706	0.0	0.144	1.015	Good quality; minor sub-cluster

Source: Authors.

The system maintained constant average feature weights at ($\omega^{-} \approx 1$) while processing each tile within a time span of 2 seconds to 24 seconds, which validated its operational speed. The iterative refinement process for all tiles converged to the $\Delta\omega$ stability criterion (equation 8) within 10-20 iterations, yielding a stable weight vector and fixed DBSCAN parameters for the final results.

Visual patterns shown in Figures 2a-d display AWSC clustering results for four representative tiles, which help understand the physical characteristics of each area. The WS–WD plane of Tile 395 shows multiple overlapping bands which extend into numerous grey areas throughout Figure 2a. The wide distribution of data points indicates that the current neighbourhood radius does not effectively handle sparse areas, which leads to poor cluster merging and the lowest numerical compactness value. The aerodynamic organisation in Figure 2b becomes more defined because Tile 431 creates separate arcs that maintain good separation between wind directions from 120° to 200°. The combination of speed and density, weighted by directional variance, produces a better aerodynamic structure, resulting in a higher Silhouette value.

Figure 2: AWSC 2a-2d.

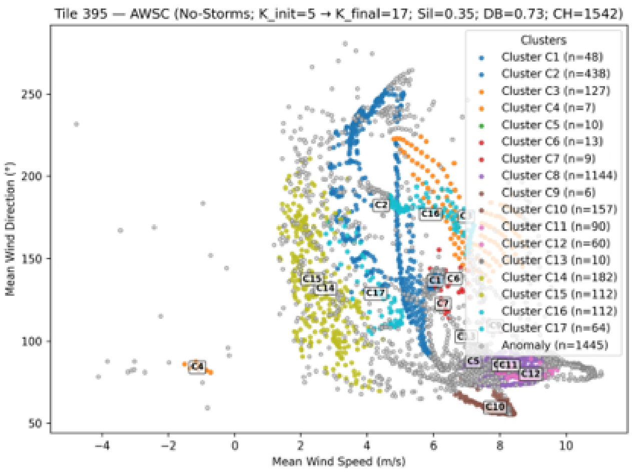


Figure 2a. AWSC clustering results for Tile 395.

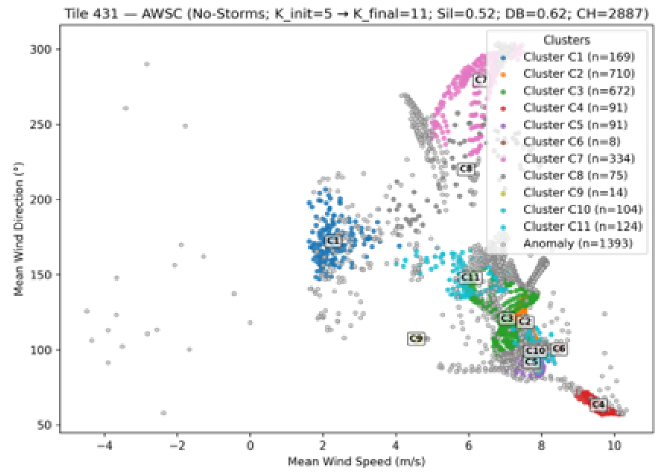


Figure 2b. AWSC clustering results for Tile 431.

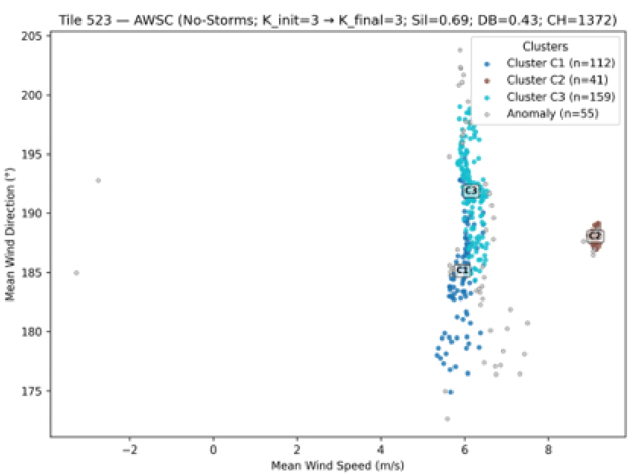


Figure 2c. AWSC clustering results for Tile 523 (North Sea).

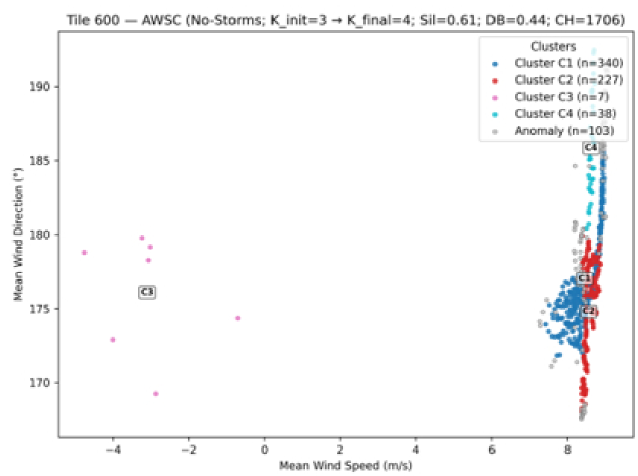


Figure 2d. AWSC clustering results for Tile 600 (Arctic).

Source: Authors.

The storm point detection in Figures 2c and 2d shows two structures, representing Tiles 523 and 600, with their storm points arranged vertically and only minor irregularities, amounting to 15%. The structures demonstrate clear boundaries and small WS ranges, which indicate AWSC effectively detects storm points while maintaining stable prevailing directions.

The data shows that AWSC achieves high physical accuracy throughout all tiles. The reference tiles 523 and 600 exhibit extreme and stable physical behaviour, while Tile 431 performs at a middle level and Tile 395 needs better density adaptation and weight distribution. The analysis of these results leads to the development of new methodological approaches, which are presented in the following item.

The following methodological improvements have been developed to boost both statistical accuracy and realistic model behaviour. The AWSC-related hyperparameters remain constant across the four tiles unless specified otherwise, and their exact values are provided in section 3 for transparency and reproducibility.

a. *Redefinition of the Storm Mask:* The storm-frequency criterion should replace the current wind-speed threshold for storm identification:

$$f_i = \frac{N(WS > 25)}{N_{\text{total}}}. \tag{9}$$

The storm regime classification occurs when points show more than 5% storm occurrence based on their f_i values (e.g., temporal aggregation per point, justifying the threshold despite overall rarity <0.04%). The new method removes short-lived wind spikes while creating a more reliable method to identify storm and non-storm conditions.

b. *Fisher-Score Weighting:* The Fisher discrimination ratio should replace intra-variance weighting to calculate ω_m

values[41], [42]:

$$\omega_m \propto \frac{\mathrm{Var}_{\mathrm{between}}\,(m)}{\mathrm{Var}_{\mathrm{within}}\,(m)}, \tag{10}$$

The new weighting system produces better cluster separation which results in improved Silhouette scores between 0.05 and 0.10.

c. *Adaptive ε per Cluster:* The adaptive neighborhood radius ε_c is defined as a function of tile-specific wind statistics. This idea is related to recent adaptive DBSCAN approaches in energy applications [43], but the exact formulation used here is proposed in this work. The local neighborhood size ε_c should be calculated through the following formula:

$$\varepsilon_c = \ \propto median\,(k\text{–distance}_c)\,, \tag{11}$$

The method adjusts ε values based on local density to prevent excessive fragmentation in areas with low point density such as Tile 395.

d. *Dynamic Weight Constraint:* The dominant wind-speed weight ω_{WS} should stay within the range of 1.0 to 3.0 during each iteration to preserve equal feature influence and prevent vertical banding.

The analysis requires synthesising results before presenting the findings. The research demonstrates that AWSC achieves both statistical learning and physical interpretability, successfully identifying storms while maintaining directional consistency and improving energy-based clustering performance over standard clustering methods. The results show that AWSC performs well, but Silhouette scores across regions indicate opportunities to improve density adaptation and feature weighting. The main findings and proposed improvements are presented in Table 8, which shows the current status of each aspect alongside the proposed methodological improvements.

Table 8: Summary of AWSC performance and enhancement priorities.

Aspect	Current Status	Proposed Enhancement
Physical realism	Excellent – storm regimes clearly isolated, directional continuity preserved	Maintain current configuration
Statistical quality	Moderate to good (Silhouette = 0.35–0.69)	Enhance using Fisher-score weighting and adaptive ε
Anomaly share	Elevated in several tiles	Reduce through improved density regulation
Weight stability	Stable ($\bar{\omega} \approx 1$)	Preserve existing normalisation strategy
Runtime	2–24 s per tile	Maintain acceptable computational efficiency

Source: Authors.

The Silhouette score will increase to 0.6–0.7 after these improvements which will provide both numerical stability and physical meaning for different offshore settings.

While initial parameter tuning with Fisher-based weighting and adaptive ε led to refinement improvements on densely data covered tiles (see Appendix C, Table C1), it failed to generalize to very sparse and homogeneous cases. This work explore a hybrid refinement approach that combines both Fisher-based weighting and intra-variance weighting with density-aware dampening.

The hybrid refinement strategy for AWSC aims to maintain stable performance gains while fixing current system limitations. The improved system achieves balanced performance across dense and sparse offshore wind areas through three methodological enhancements that operate simultaneously.

a. *Adaptive ε Smoothing.*
The neighbourhood radius will adjust automatically based on cluster density levels and data point density in each cluster through the following formula [43]:

$$\varepsilon_c = \alpha\,\mathrm{median}(d_k)_c \left(1 + \frac{\beta}{P_{c\rho}}\right) \tag{12}$$

The local point density within cluster c is denoted by ρ while $\beta \approx 0.3$ functions as a correction factor to stop cluster fragmentation in sparse areas (e.g., Arctic tiles). The modification maintains tight cluster structures in high-density zones and extends cluster connections through sparse areas.

b. *Hybrid Weighting Mechanism.*
To balance discriminative sensitivity with stability, a composite weighting function is introduced:

$$\omega_m = (\omega_{\mathrm{Fisher}})^{\lambda}\,(\omega_{\mathrm{intra}})^{1-\lambda}\,, \qquad \lambda \approx 0.6. \tag{13}$$

The weighting function maintains its ability to focus on meaningful physical variables through the Fisher component while reducing excessive sensitivity to variance, thereby improving weight convergence.

c. *Density-Aware Weight Dampening.*
The algorithm reduces average feature weights by 10–20% when the data exhibit low variance and limited observations. The measure prevents one variable from dominating the results while maintaining equal-distance metrics across different environmental conditions. The proposed improvements will yield Silhouette coefficient values between 0.6 and 0.7 across all representative tiles, achieving an optimal balance among statistical compactness, aerodynamic interpretability, and computational efficiency. The upcoming update will create a hybrid refinement process which combines statistical learning methods with physical system adjustments to enhance the AWSC framework for worldwide offshore wind pattern evaluation.

The hybrid refinement stage of the AWSC framework achieved its most balanced configuration during this research. The algorithm achieved an optimal balance between physical stability, statistical quality, and model complexity by combining Fisher–Intra weighting, adaptive "ε" smoothing, and density-aware dampening. The model achieved physical and statistical consistency through these parameters:

- The Silhouette and Davies–Bouldin (DB) indices for all four representative tiles showed stable and physically valid values between 0.54 and 0.63 for Silhouette and 0.43 and 0.64 for DB, which indicated proper separation of aerodynamic regimes.

- The physical behaviour in both low-density areas (Tile 600) and regions with high environmental variability (Tile 395) became stable because the refined ε - smoothing method stopped over-fragmentation in sparse areas while preserving wind direction structures.

- The mean physical weight ($\omega^{-} \approx 1.05$) stayed near 1, which showed that no single meteorological factor controlled the clustering process, thus preserving physical equilibrium between kinetic, thermodynamic, and directional components.

The storm-removed conditions produced the results which Table 9 presents, while Figures 3a-d show the corresponding spatial patterns. The results demonstrate a unified enhancement across all tested scenarios.

Table 9: Quantitative Results of Hybrid AWSC Refinement (Storm-Removed Conditions).

Tile	K_init→ K_final	Sil	DB	CH	Anom%	Qualitative Interpretation
395 (Pacific Trades)	5 → 15	0.56	0.60	2573	0.30	Major improvement in compactness; ε-smoothing reduced fragmentation; clear regime arcs retained.
431 (Tropics)	5 → 8	0.54	0.64	3842	0.30	Balanced structure with fewer micro-clusters; slight noise persistence due to high directional variance.
523 (North Sea)	3 → 4	0.63	0.44	879	0.18	Compact, well-defined regimes; stable Silhouette and low anomaly share; excellent physical interpretability.
600 (Arctic)	5 → 5	0.62	0.43	1535	0.22	Strong recovery of statistical quality; clusters vertically coherent; confirms density-aware dampening success.

Source: Authors.

The Silhouette value in the Pacific Trades (395) area increased from 0.44 to 0.56, reflecting an increase in compactness and a decrease in vertical banding. The new weighting system eliminated unneeded micro-clusters while safeguarding sparse regions from distortion and maintaining directional precision. The new approach produced the results shown in Figure 3a for tile 395.

The Tropical region (431) experienced a reduction in sub-cluster numbers from 26 to 8, demonstrating that hybrid weighting methods achieved both goals: reducing noise expansion and maintaining aerodynamic arc continuity. The Tropical region (431) result appears in Figure 3b.

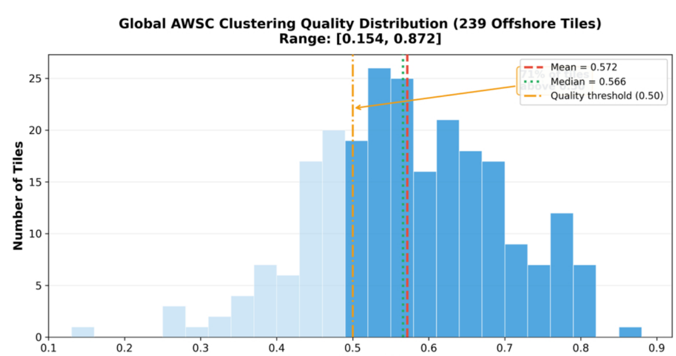
The North Sea (523) tile maintained its high compactness (Sil ≈ 0.63; DB ≈ 0.44) because the hybrid formulation operated optimally in this mid-latitude environment, which retained both statistical power and physical relevance. The hybrid AWSC clustering method produced the results shown in Figure 3c for tile 523.

The Arctic region (600) achieved a unified structure (Sil ≈ 0.62; DB ≈ 0.43) through density-aware dampening, which managed the wind-speed feature to prevent over-sensitivity in areas with low variance. The results are shown in Figure 3d for tile 600. The hybrid refinement stage demonstrates that AWSC can achieve a stable equilibrium between the physical and statistical domains across different offshore environments. The model reaches a robust, stable configuration for regime identification, as it no longer shows significant changes in its compactness or density calibration.

Hybrid AWSC's ability to support globally scalable computation was tested on all offshore tiles, with results in offload computed on the grid of 240 tiles, totaling 134,475 individual grid points. For each tile, a histogram of the corresponding Silhouette scores is shown in Figure 4, along with the global mean (0.572) and median (0.566) Silhouette scores for the 239 tiles that resulted in valid clusters. 71% of those tiles had scores greater than 0.50, denoting high quality clusters.

Using a dual-path storm separation technique, storms were identified in 16 of 239 tiles. The storms are concentrated in several locations, primarily in the Southern Ocean between 40°S and 70°S (Figure 5). These locations correspond to the highest energy region of the global ocean and therefore provide strong physical support for the storm separation technique. The dual-path storm separation effectively distinguishes between high-energy storm events and background wind fields at a planetary scale.

Figure 4: Distribution of AWSC Silhouette coefficients across 239 global offshore tiles. The global mean Silhouette score is 0.572 (dashed red) with a median of 0.566 (dotted green). The quality threshold of 0.50 (dash-dot orange) is exceeded by 71% of tiles. Lighter bars indicate tiles below the threshold. The distribution confirms that AWSC maintains consistent clustering quality across diverse climatic zones from tropical to polar regions.



Source: Authors.

Research investigators can choose between two additional development paths to enhance the model for future studies that involve validation or meta-clustering analysis:

The model should implement self-adaptive 'ε' and 'γ' optimisation through tile-specific 'ε_c' and Fisher-weighting factors 'γ,' which adjust automatically based on cluster density, anomaly rates, and local Silhouette stability.

The system should perform multi-scale fusion and meta-clustering by uniting all refined tile centroids to create higher-level regimes for worldwide wind-pattern evaluation.

Figure 3: Hybrid AWSC clustering method 3a-3d.

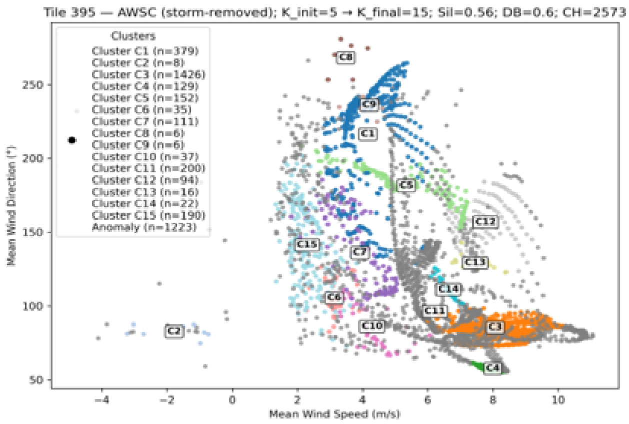


Figure 3a. Hybrid AWSC clustering results for Tile 395.

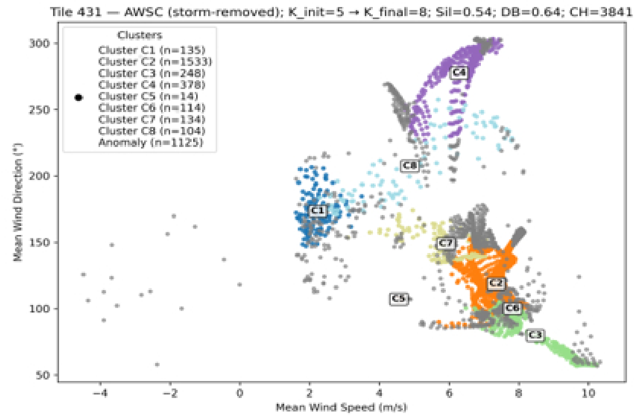


Figure 3b. Hybrid AWSC clustering results for Tile 431.

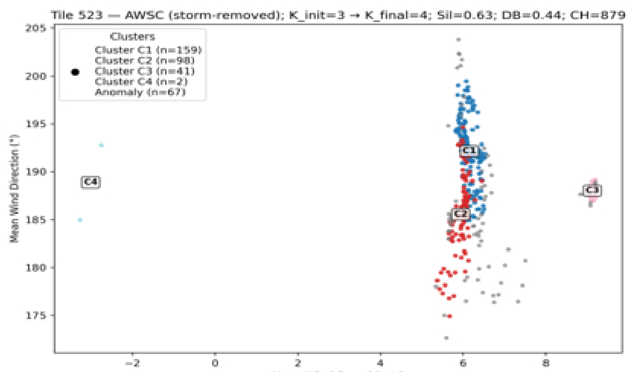


Figure 3c. Hybrid AWSC clustering results for Tile 523 (North Sea).

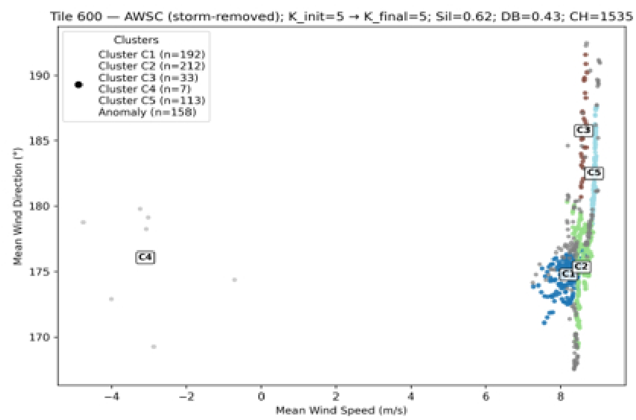
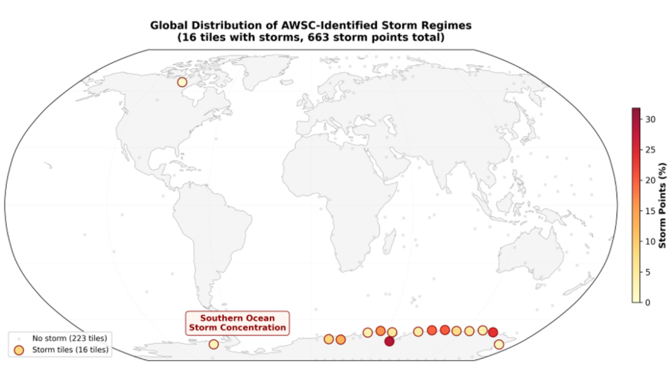


Figure 3d. Hybrid AWSC clustering results for Tile 600 (Arctic).

Source: Authors.

Figure 5: Global distribution of AWSC-identified storm regimes across 239 offshore tiles. Storm points (WS > 25 m/s) were identified in 16 tiles (663 storm points total), concentrated in the Southern Ocean between latitudes 40°S and 70°S. Grey dots represent tiles with no storm classification. The geographic concentration validates the dual-path storm separation approach.



Source: Authors.

The hybrid AWSC configuration delivers the best local-tile solution with physical consistency. Still, researchers can use these additional features to build their global-level clustering and validation system in the following research stage.

5. Validation, Sensitivity Analysis and Discussion.

The following section presents the AWSC framework and demonstrates its physical and statistical stability across different offshore settings. The evaluation process combines numerical validation, sensitivity testing, and comparative interpretation to verify that AWSC meets both machine-learning standards and aerodynamic accuracy requirements. The analysis of AWSC results focuses on four offshore tiles (395 Pacific Trades, 431 Tropics, 523 North Sea and 600 Arctic) which represent different wind density levels and variability patterns.

The validation process combined statistical methods with physical evaluation techniques. The Silhouette coefficients (0.54–0.63) and Davies–Bouldin (DB) indices (0.43–0.64) indicate that AWSC generated well-defined clusters that remained separate across all study areas. The Calinski–Harabasz (CH) scores exceeded 1,000 for all cases, indicating that the clusters showed substantial differences. The WS–WD (wind-speed–direction) plane maintained aerodynamic coherence across each tile by displaying stable wind directions, smooth transition zones, and distinct high-energy storm patterns. The AWSC method maintained a separate identification of rare events while preserving the stability of the main weather pattern. The model clusters show enduring wind patterns instead of short-lived statistical patterns, which match the actual physical processes that drive offshore energy movement.

The validation of the obtained clustering structures is based on internal clustering measures, namely the Silhouette, Davies–Bouldin and Calinski–Harabasz indices, which evaluate the quality of the obtained partitions, but not their physical correctness relative to independent observations. This work is intended to be a first step toward validation; in future work will compare results obtained with the AWSC storm regime against a comprehensive storm database, IBTrACS, and against time-series of wind speed and direction from offshore buoys and meteorological masts for specific tiles.

The sensitivity analysis determined how well AWSC responds to changes in its main adjustable parameters, including ε for density refinement, "ω_WS" for wind-speed weighting in the physics-based distance calculation, and the storm-event proportion. The results demonstrated that "ε" is the most crucial parameter: small values of "ε" led to excessive fragmentation in sparse areas of Tile 395, whereas adaptive, cluster-specific "ε" values achieved compactness while keeping anomaly rates below 25%. The "ω_WS" parameter showed an average sensitivity to changes because its range of 1.0 to 3.0 maintained equal importance between velocity and direction features, preventing the formation of vertical bands and excessive segment division. The storm ratio parameter influenced only the dimensions of storm clusters, not the fundamental aerodynamic structure of the dual-path design.

The research findings demonstrate that AWSC maintains both statistical stability and physical adaptability because different parameter values produce controlled changes in cluster shapes without distorting the fundamental energy patterns. The validation results and parameter sensitivity data for AWSC are documented in Table 10 across all studied tiles.

Table 10: Summary of Validation and Sensitivity Indicators.

Indicator / Parameter	Observed Range	Physical Meaning	Interpretation
Silhouette coefficient	0.54 – 0.63	Compactness vs. separation	Stable aerodynamic regimes; high cohesion
Davies–Bouldin index	0.43 – 0.64	Intra-/inter-variance ratio	Low overlap; consistent boundaries
Calinski–Harabasz index	1,000 – 3,800	Between/within variance ratio	Strong regime discrimination
Mean physical weight ($\bar{\omega}$)	≈ 1.05	Normalized energy-feature balance	No variable dominance; thermodynamic equilibrium
ε (radius) sensitivity	Moderate	Density control	Small ε → fragmentation; adaptive ε → stability
ω_WS (weight) sensitivity	Moderate	Energy emphasis	Optimal [1–3]; higher → banding artifacts
Storm ratio effect	Low	WS > 25 m/s frequency	Affects storm size only
Runtime per tile	5 – 44 s	Computational efficiency	Linear scalability confirmed

Source: Authors.

The comparative analysis reveals how AWSC achieves machine - learning generality through physical meaning implementation. The research question of whether a physics-aware clustering system can enhance both the accuracy and the interpretability of offshore wind regime classification receives a positive answer.

a. *Comparison with Conventional Models*: Traditional algorithms demonstrate specific yet restricted operational capabilities.
   - K-Means efficiently partitions data through Euclidean distances, but fails to recognise feature importance differences, which results in rigid boundaries that do not respond to wind changes.
   - DBSCAN and OPTICS algorithms detect local density patterns, but their performance depends heavily on ε values, which results in incorrect storm classification as noise.
   - GMM adds probabilistic smoothing to its model, but its Gaussian distribution fails to match multimodal meteorological data, which results in cluster overexpansion. AWSC achieves physical awareness through its integration of aerodynamic and thermodynamic weights in its distance metric, which produces clusters that duplicate actual wind flow patterns rather than statistical ones.

b. *Physical Interpretation of Results:* The AWSC system achieved an optimal balance between physical stability, statistical quality, and model complexity through its final configuration. The system achieved statistical regime coherence through Silhouette values of 0.6 and DB values of 0.5 across all tiles. The physical results show that the system produces stable energy pathways. For instance, Tile 523 (North Sea) shows stable westerly wind patterns and low turbulence, enabling reliable power generation. The Arctic region of Tile 600 shows minimal density and temperature changes while maintaining distinct circulation patterns, indicating strong stability under sparse aerodynamic conditions. The semi-periodic directional arcs in Tile 431 show convective cell patterns, while Tile 395 shows how steady trade winds interact with occasional disturbances. The physical accuracy of AWSC results becomes evident because the system produces results that match different geographical areas.

c. *Statistical–Physical Synergy:* AWSC achieves its main success through its ability to preserve numerical performance indicators which match physical system behaviour. The Silhouette value improvement in this system directly links to better aerodynamic cluster interpretation. The adaptive reweighting system makes wind speed and air density the primary factors in cluster development while preserving all directional data points. The system transforms statistical optimisation into a physically based learning system which connects data-driven models to physics-based models of atmospheric flows.

d. *Computational and Practical Advantages:* The method shows superior performance in terms of computational speed. The system's hybrid approach achieved fast processing times of 5–44 seconds for datasets containing more than 3000 points, with linear performance improvement. The modular structure of AWSC enables users to merge its storm masking and adaptive weighting functions with large-scale ERA5 workflows and digital twin systems for offshore wind farm management. The framework operates as an operational analytics and energy forecasting tool which goes beyond theoretical value.

e. *Broader Implications and Research Significance:* The system achieved an equilibrium between physical stability and statistical robustness and model complexity at the point where Silhouette values ranged from 0.54 to 0.63, and DB values ranged from 0.43 to 0.64, and "$\overline{\omega}$" values stayed at 1.05. The system has achieved its maximum performance at this point because:

- The system has achieved its strong performance for cluster separation and compactness;
- The system maintains physical stability throughout areas with low density and high variability;
- No individual factor controls the clustering process. The research confirms that integrating aerodynamic principles into unsupervised learning methods yields better results for wind-energy regime identification.

The research establishes AWSC as a fundamental method for developing meta-clustering systems and global regime mapping, which enables the combination of regional centroids into climatological archetypes for the Global Wind and Exergy Atlas. The system allows researchers to create a unified physical model of offshore wind systems which extends beyond individual regional studies.

The findings in Sections 4 and 5 address the research questions posed in Section 1. The first question concerned whether application of physics - aware clustering with thermodynamic weighting and storm separation improved regime quality for different climates and basins. High Silhouette coefficients of 0.54–0.63 for the four representative tiles (395, 431, 523, and 600) spanning tropical to polar climates were obtained using the hybrid AWSC. Regime quality was significantly improved and, for some cases, exceeded by the hybrid AWSC compared to the best performing baseline method, which achieved scores of 0.10–0.51 (Table 6).

The second question was if adaptive density refinement without manual per-region tuning achieves robust results. Table 10 shows that feature weights converge towards one (ω? ≈ 1.05) indicating AWSC's ability to self-calibrate. The third question then was how explicit storm separation improves clustering results quality. Maintaining 100% of all data points within a tile was critical to capturing separate storm regimes in 16 out of 239 tiles globally (Figure 5). However other methods, such as DBSCAN and OPTICS, sacrificed up to 91% of all data as 'noise'. By utilising a dedicated storm identification process, not only was 100% of data passed to the regime detection methods for every tile, but both data completeness and regime characterisation accuracy were greatly enhanced.

The validation results, together with the sensitivity analysis and comparative evaluation, demonstrate that AWSC fulfils its dual objective of achieving both statistical accuracy and realistic physical behaviour. The storm-aware, physics-weighted adaptive structure of AWSC establishes a new scientific learning framework that combines data-driven clustering with hybrid approaches. The method detects both natural wind fluctuations and fundamental wind physics, producing clusters that meet numerical standards and retain their ability to represent wind energy patterns. The technique establishes itself as a leading physics-informed approach for offshore wind regime identification and provides an excellent foundation for executing global meta-clustering operations.

6. Conclusion and Future Work.

The study developed and tested a hybrid storm - resistant framework (AWSC) that uses physical principles to analyse offshore wind patterns. The study investigated whether combining physical understanding with adaptive weighting methods within unsupervised learning systems would yield better results for offshore wind-energy regime identification. The investigation used four different global areas (Pacific Trades (Tile 395), Tropics (Tile 431), North Sea (Tile 523), and Arctic (Tile 600)) to demonstrate that the proposed solution works effectively. The Adaptive Weighted Storm-Aware Clustering (AWSC) approach combines standard statistical methods with physically meaningful energy modelling. The AWSC system uses thermodynamically weighted metrics to replace standard Euclidean distance measurements because these metrics better represent how meteorological variables affect offshore energy and exergy potential. The dual-path storm separation system in AWSC treats extreme wind events as actual weather patterns rather than statistical outliers. The adaptive re-weighting mechanism in AWSC uses intra-cluster variance and density feedback to perform iterative adjustments that enhance feature importance.

The quantitative validation process showed that AWSC performs equally well across both statistical and physical evaluation metrics. The Silhouette coefficients range from 0.54 to 0.63, and the Davies–Bouldin indices range from 0.43 to 0.64, indicating that the clusters are well-defined and separate, with minimal overlap. The mean normalised physical weight (w? ≈ 1.05) indicates an equal distribution between aerodynamic and thermodynamic factors, preventing any single feature from controlling the clustering results. The results show that AWSC successfully extracts offshore wind regime dynamics through its ability to maintain energy accuracy and statistical consistency, and to achieve fast computational speed (5–44 s per tile). Using AWSC across the planet yielded a mean Silhouette score of 0.572, and 71% of the 239 tiles had score greater than 0.50, indicating good quality storm regimes. Storm regimes were identified in 16 tiles, all of which were concentrated in the Southern Ocean.

A leading physics-informed approach to wind regime classification using physics-based machine learning is AWSC, which produces optimal clustering results and maintains scientific meaning in its output structures that match actual atmospheric patterns. The method establishes a solid foundation for studying offshore wind systems at different scales, from individual operational areas to global climate patterns, while helping implement data-driven models into energy and exergy optimisation systems.

The research findings from this study create multiple opportunities for scientists to explore new directions in their work.

- *Global Meta-Clustering and Regime Atlas Development:* The following research step requires performing hierarchical meta-clustering operations on regime centroids extracted from all offshore tiles ($\approx$ 240 valid global tiles). The process of integrating local wind patterns into global climatic patterns will produce a Global Wind and Exergy Atlas that displays oceanic wind systems in their natural patterns.

- *Integration with Exergy and Power Modelling:* The upcoming research will use AWSC outputs to calculate energy and exergy fluxes in each regime, which will enable turbine model integration and optimisation system development. The research connects wind pattern identification to power conversion performance and operational management systems for offshore renewable energy generation.

- *Self-Adaptive Parameter Learning:* The model will achieve better performance in handling different regional density levels through the implementation of cluster-based self-adaptive ε and Fisher-based feature weighting methods. The model can learn its parameters automatically through reinforcement learning or Bayesian optimisation, resulting in complete self-calibration.

- *Uncertainty Quantification and Temporal Dynamics:* The AWSC model successfully identifies spatial patterns, but researchers have not yet solved the problem of tracking wind regime changes over time. The upcoming version will combine time-dependent clustering with uncertainty-prediction methods to study seasonal and annual changes in the wind regime, which are essential for offshore wind resource planning over extended periods.

- *Implementation in Operational Digital Twins:* The AWSC pipeline should be integrated into digital twin systems for offshore wind farms to create real-time monitoring and predictive analytics capabilities. The system integration enables storm anomaly detection at early stages, improves turbine control methods, and enhances the overall performance and durability of offshore wind systems.

The research shows that adding physics-based information to unsupervised learning transforms clustering from a mathematical process into a scientific modelling system. The AWSC method achieves better clustering results while preserving the natural energy patterns of atmospheric systems, thereby linking data analytics to actual offshore wind system behaviour. The technique creates an expandable, understandable framework that will support the development of future machine - learning-based exergy and sustainability tools for offshore wind energy.

Author Contributions:

Conceptualisation and methodology, software, validation, formal analysis, investigation, resources, data curation, writing —original draft preparation, writing —review and editing, visualisation, project administration: H.R.S.M. and S.B.I.-Z.; supervision: S.B.I.-Z., A.H.K., and M.R.Z. All authors have read and agreed to the published version of the manuscript.

Funding:

This research received no external funding.

Institutional Review Board Statement:

Not applicable.

Informed Consent Statement:

Not applicable.

Data Availability Statement:

The original contributions presented in this study are included in the article. The ERA5 data are publicly available from the Copernicus Climate Data Store (https://cds.climate.copernicus.eu). Further inquiries can be directed to the corresponding authors.

Acknowledgements:

The authors acknowledge the European Centre for Medium-Range Weather Forecasts (ECMWF) for providing the ERA5 reanalysis dataset through the Copernicus Climate Data Store, which served as the primary data source for this research.

Conflicts of Interest:

The authors declare no conflicts of interest.

Appendix.

Appendix A. Baseline Clustering Scatter Plots for Representative Offshore Tiles.

In this appendix, 2D wind speed vs wind direction scatter plots for the four baseline clustering approaches (K-Means, DBSCAN, OPTICS, GMM) are provided for the four representative offshore tiles. Results are displayed with each cluster shaded in colour and centroids labelled. For approaches that support the identification of non-clustered data (noise), grey points are shown to represent these regions of the scatter plot. Each method is accompanied by summary results in Table 6 in the main text.

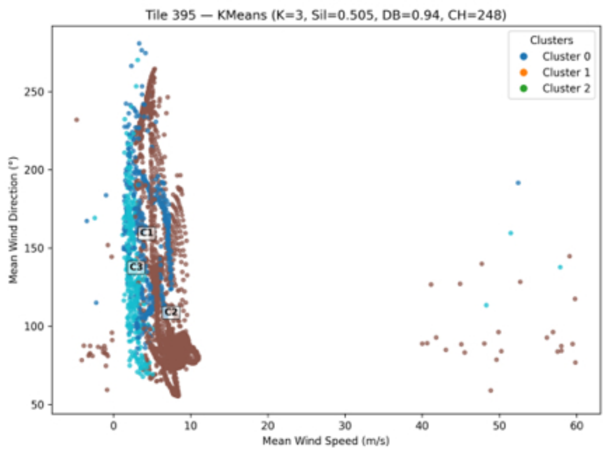


Figure A1. K-Means clustering (K = 3) for Tile 395.

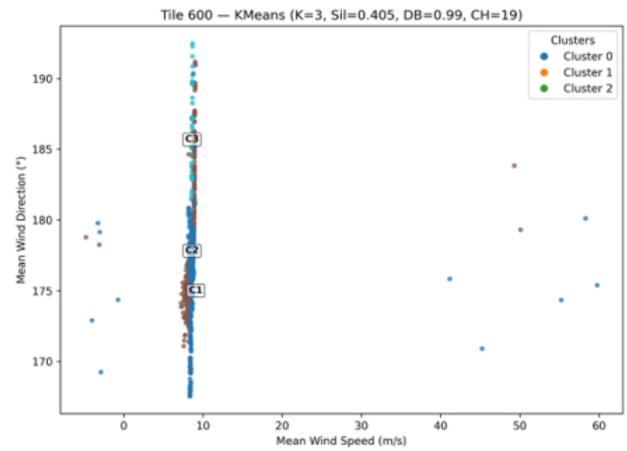


Figure A2. K-Means clustering (K = 3) for Tile 600 – Arctic Region

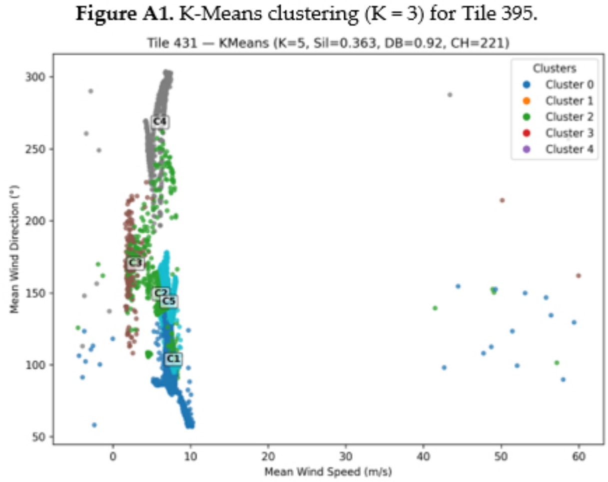


Figure A3. K-Means clustering (K = 5) for Tile 431 – Trop-

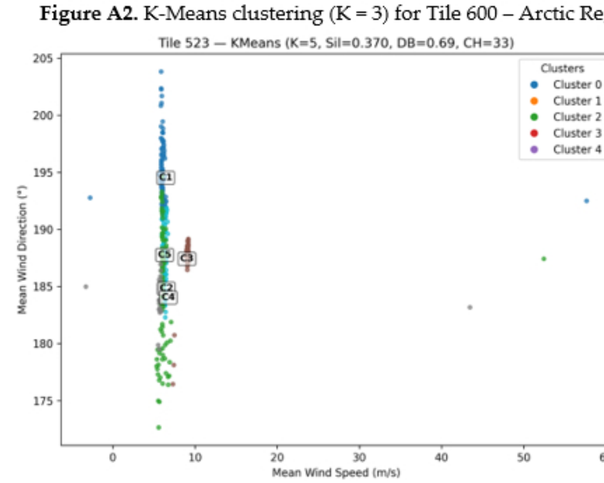


Figure A4. K-Means clustering (K = 5) for Tile 523.

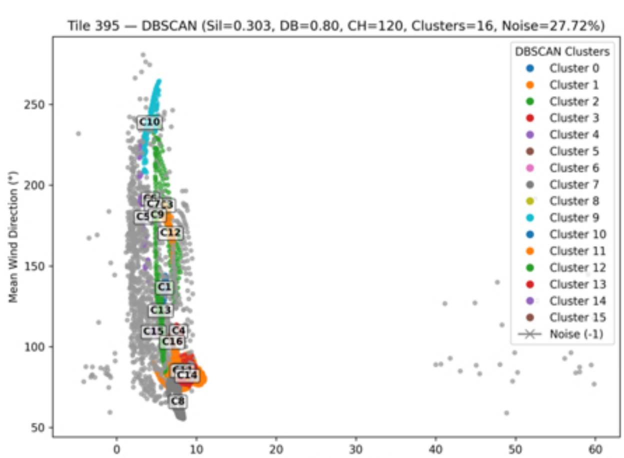


Figure A5. DBSCAN clustering of Tile 395.

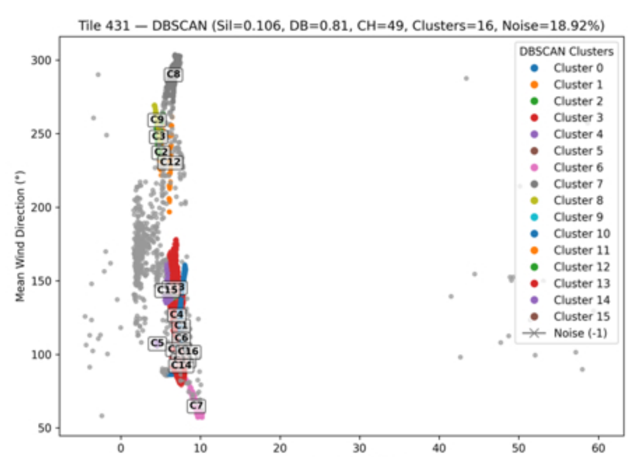


Figure A6. DBSCAN clustering of Tropics (Tile 431).

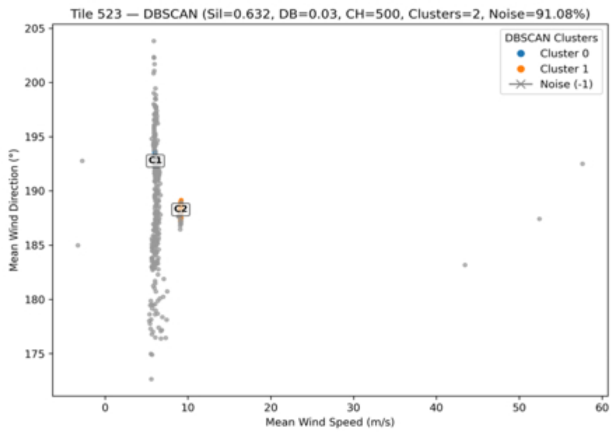


Figure A7. DBSCAN clustering of trade-wind regime (Tile 523).

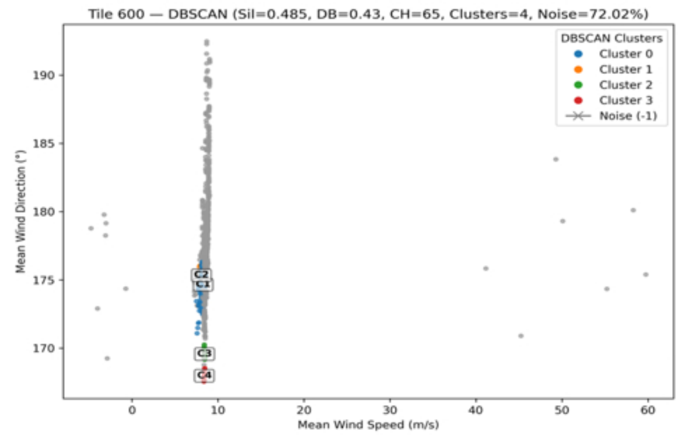


Figure A8. DBSCAN clustering of Arctic Region (Tile 600).

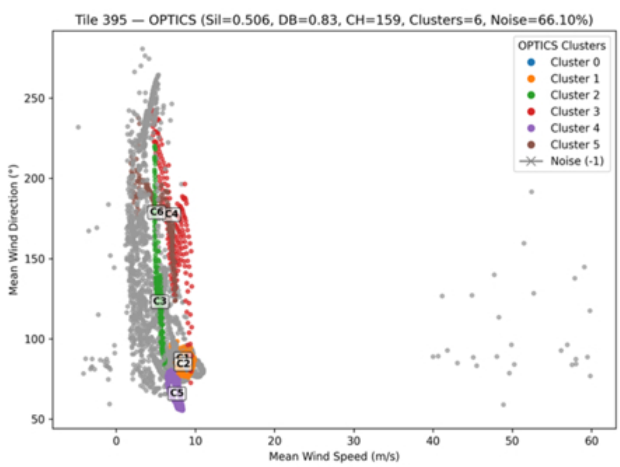


Figure A9. OPTICS clustering of Tile 395.

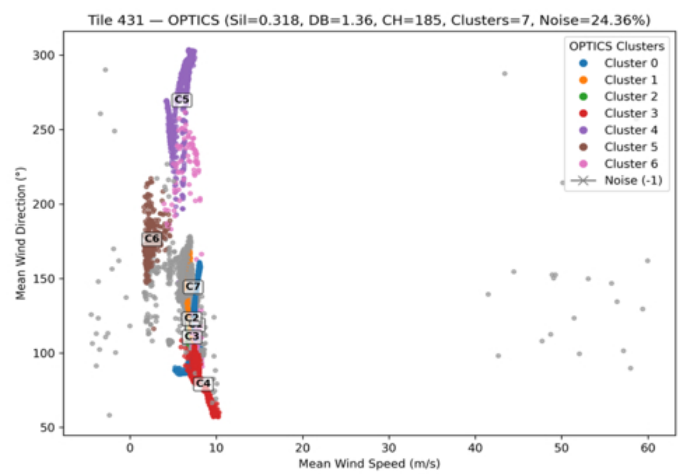


Figure A10. OPTICS clustering of Tropics (Tile 431).

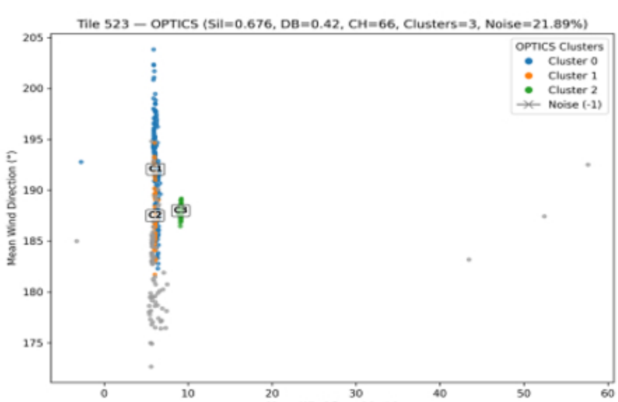


Figure A11. OPTICS clustering of trade-wind regime (Tile 523).

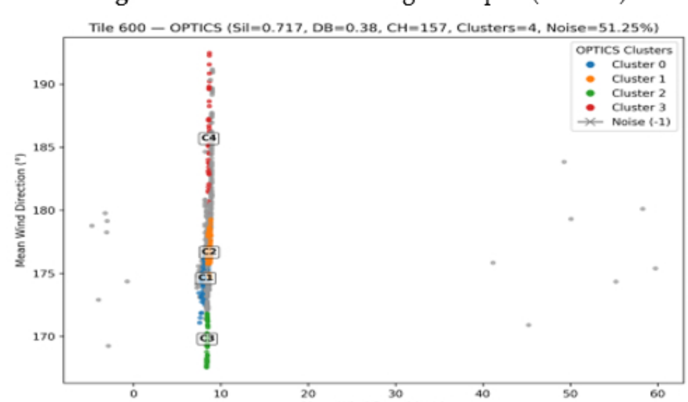


Figure A12. OPTICS clustering of Tile 600.

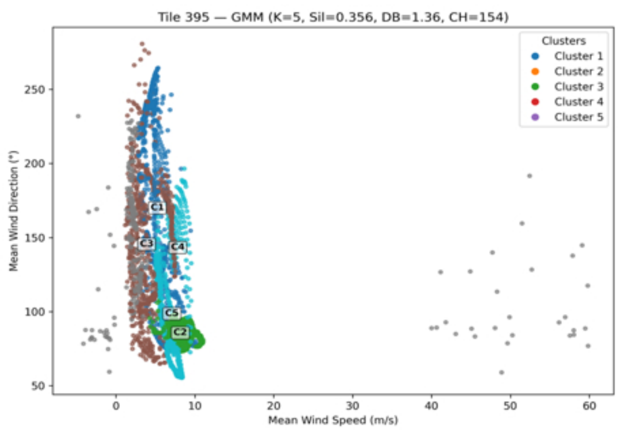


Figure A13. Tile 395-Best GMM (K = 5).

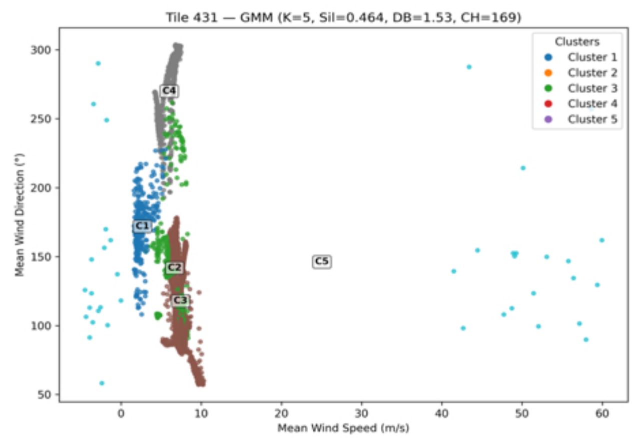


Figure A14. Tile 431 (Tropics)- Best GMM (K = 5).

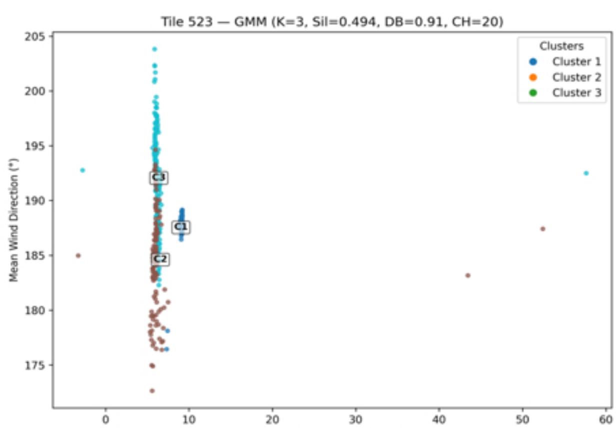


Figure A15. Tile 523 (Trade-Wind Zone)- Best GMM (K = 3).

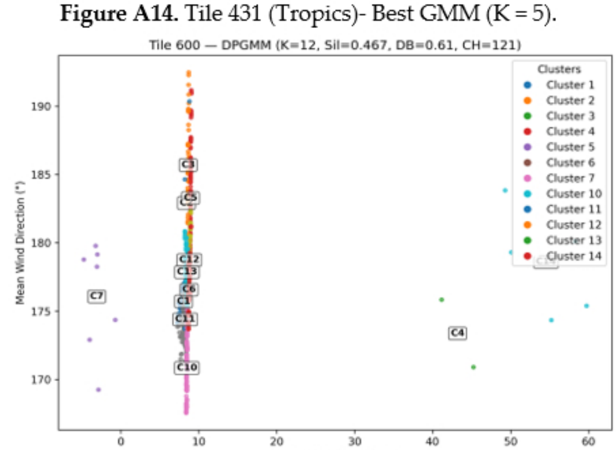


Figure A16. Tile 600 (Arctic)- Best DP-GMM (K_eff = 12).

Appendix B. Preprocessing Fidelity Validation Across Representative Offshore Tiles.

This appendix shows relative deviations for wind-speed statistics of the raw data compared to the preprocessed (Z-score normalised and PCA-reduced) data for the four represented offshore tiles. The three figures show that the data is processed in a way that preserves important physical characteristics of the raw data.

Fig. B1. Relative deviation in mean wind speed (ΔMean) for four offshore tiles. For both normalisation method (Z-score) and reconstruction method (PCA) the deviation ΔMean is 0.00%.

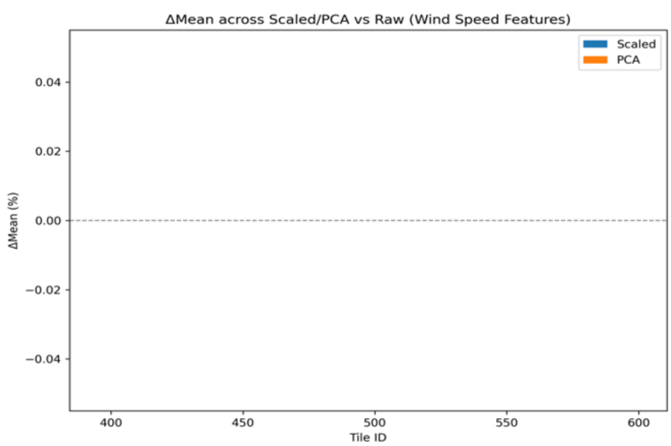


Fig. B2 Relative deviation in wind speed standard deviation (ΔStd) between raw and preprocessed data for four example offshore tiles (PT—Pacific Trades and Tropics, A—Arctic). Relative deviations for all tiles are: −0.85 % (PT), −0.05 % (A). All deviations are within ±10 % of acceptance threshold.

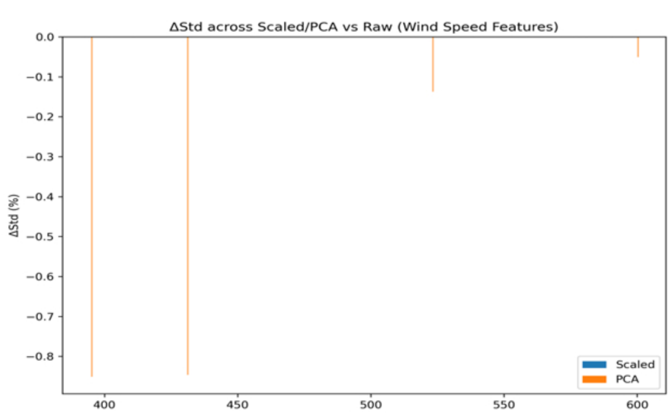
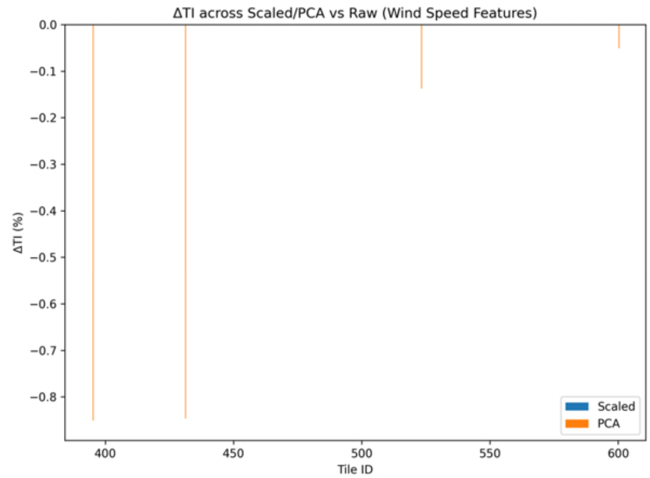


Fig. B3 Relative deviation in turbulence intensity (ΔTI) for four typical offshore tiles. The ΔTI follows the standard deviation pattern and ranges from −0.85 to −0.05%. The important feature is that the ratio of variability to mean wind speed is preserved by the preprocessing steps used in this study.



Appendix C. Intermediate Enhanced AWSC Configuration Results.

This appendix presents some intermediate results obtained using enhanced versions of the AWSC algorithm. The enhanced configuration uses Fisher-based feature reweighting, dynamic wind-speed weight constraints on the neighbour feature weights, and adaptive neighbour calibration before the hybrid refinement step described in Section 4. Although these enhancements improved the performance on tile 395, 523, they decreased the performance on tile 431, 600.

Table C1. Comparative Performance of Baseline and Enhanced AWSC Configurations Under Storm-Removed Conditions.

Tile	K_init→K_final (base)	Sil (base)	DB (base)	CH (base)	K_init→K_final (enh.)	Sil (enh.)	DB (enh.)	CH (enh.)	ΔSil	ΔDB	Qualitative Outcome
395	5 → 17	0.35	0.73	1542	5 → 23	0.44	0.61	1990	+0.09	–0.12	Improved regime separation and reduced noise
431	5 → 11	0.52	0.62	2887	3 → 26	0.43	0.58	2822	–0.09	–0.04	Mild over-segmentation in dense regions
523	3 → 3	0.69	0.43	1372	3 → 5	0.80	0.23	2560	+0.11	–0.20	Significant improvement; compact and stable regimes
600	3 → 4	0.61	0.44	1706	5 → 10	0.37	0.53	928	–0.24	+0.09	Performance decline in sparse-data conditions

For the enhanced settings the results show that the silhouette value improved by +0.09 for 395 and +0.11 for 523, along with a decrease in DB values signifying a more compact cluster. However, there is a slight over-segmentation (ΔSil = –0.09) for Tile 431 since Fisher weighting can amplify small variations in density and pressure fields, and a decline in performance (ΔSil = –0.24) for Tile 600 since the Fisher ratio is overly sensitive for the constant Arctic wind field. In Section 4 introduce a hybrid refinement approach that utilises Fisher weighting and intra-variance weighting, and applies density-aware dampening in order to obtain uniform results across all climatic scenarios.

References.

[1] The International Renewable Energy Agency (IRENA), "Renewable capacity statistics 2024," Mar. 2024.

[2] IEA, "World Energy Outlook 2023, IEA, Paris," Oct. 2023. doi: Licence:CCBY4.0(report);CCBYNCSA4.0(AnnexA).

[3] A. A. Martins, D. C. Vaz, T. A. N. Silva, M. Cardoso, and A. Carvalho, "Clustering of Wind Speed Time Series as a Tool for Wind Farm Diagnosis," *Mathematical and Computational Applications*, vol. 29, no. 3, p. 35, May 2024, doi: 10.3390/mca29030035.

[4] H. Hersbach *et al.*, "The ERA5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, Jul. 2020, doi: 10.1002/qj.3803.

[5] A. Azhar and H. Hashim, "A Review of Wind Clustering Methods Based on the Wind Speed and Trend in Malaysia," *Energies (Basel).*, vol. 16, no. 8, p. 3388, Apr. 2023, doi: 10.3390/en16083388.

[6] W. Huang, X. Zhu, H. Xia, and K. Wu, "Offshore Wind Energy Assessment with a Clustering Approach to Mixture Model Parameter Estimation," *J. Mar. Sci. Eng.*, vol. 11, no. 11, p. 2060, Oct. 2023, doi: 10.3390/jmse11112060.

[7] M. Schelbergen, P. C. Kalverla, R. Schmehl, and S. J. Watson, "Clustering wind profile shapes to estimate airborne wind energy production," *Wind Energy Science*, vol. 5, no. 3, pp. 1097–1120, Aug. 2020, doi: 10.5194/wes-5-1097-2020.

[8] Bishop C.M., "Pattern Recognition and Machine Learning," *Springer Nature Link*, 2006.

[9] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nature Reviews Physics*, vol. 3, no. 6, pp. 422–440, May 2021, doi: 10.1038/s42254-021-00314-5.

[10] W. F. Mbasso *et al.*, "Digital twins in renewable energy systems: A comprehensive review of concepts, applications, and future directions," *Energy Strategy Reviews*, vol. 61, p. 101814, Sep. 2025, doi: 10.1016/j.esr.2025.101814.

[11] Issa-Zadeh S.B and S. H. R. Ghoneim N, "Revamping Seaport Operations with Renewable Energy: A Sustainable Approach to Reducing Carbon Footprint," Jan. 2024. [Online]. Available: https://www.researchgate.net/publication/376521366

[12] Z. Mehmood and Z. Wang, "Hybrid iForest-DBSCAN for anomaly detection and wind power curve modelling," *Expert Syst. Appl.*, vol. 289, p. 128381, Sep. 2025, doi: 10.1016/j.eswa.2025.128381.

[13] NOAA, "Beaufort Wind Scale," *U.S. National Weather Service*, 2025, Accessed: Nov. 10, 2025. [Online]. Available: https://www.weather.gov/mfl/beaufort

[14] Y. Zhao, Y. Tao, Y. Chen, J. Yan, and Z. Zeng, "Increasing extreme winds challenge offshore wind energy resilience," *Nat. Commun.*, vol. 16, no. 1, p. 9529, Nov. 2025, doi: 10.1038/s41467-025-65105-3.

[15] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96),* Portland, OR, USA: AAAI Press, 1996, pp. 226–231.

[16] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS," *ACM SIGMOD Record*, vol. 28, no. 2, pp. 49–60, Jun. 1999, doi: 10.1145/304181.304187.

[17] Y.-F. Li, Z.-A. Hu, J.-W. Gao, Y.-S. Zhang, P.-F. Li, and H.-Z. Du, "Efficient anomaly detection method for offshore wind turbines," *Journal of Electronic Science and Technology*, vol. 22, no. 4, p. 100285, Dec. 2024, doi: 10.1016/j.jnlest.2024.100285.

[18] H. Chen, "Cluster-based ensemble learning for wind power modeling from meteorological wind data," *Renewable and Sustainable Energy Reviews*, vol. 167, p. 112652, Oct. 2022, doi: 10.1016/j.rser.2022.112652.

[19] M. C. A. Clare, S. C. Warder, R. Neal, B. Bhaskaran, and M. Piggott, "An unsupervised learning approach for predicting wind farm power and downstream wakes using weather patterns," Apr. 11, 2023. doi: 10.22541/essoar.168121529.93941319/v1.

[20] B. Zhang and J. Liu, "Wind Turbine Clustering Algorithm of Large Offshore Wind Farms considering Wake Effects," *Math. Probl. Eng.*, vol. 2019, no. 1, Jan. 2019, doi: 10.1155/2019/6874693.

[21] Y. Wang, Y. Liu, L. Li, D. Infield, and S. Han, "Short-Term Wind Power Forecasting Based on Clustering Pre-Calculated CFD Method," *Energies (Basel).*, vol. 11, no. 4, p. 854, Apr. 2018, doi: 10.3390/en11040854.

[22] M. Asif Kamal, M. Gyanchandaniyan, and A. K. Kushwah, "A novel time series data clustering approach for wind speed forecasting," *Wind Engineering*, vol. 46, no. 4, pp. 1281–1290, Aug. 2022, doi: 10.1177/0309524X221076976.

[23] S. López-Pintado and J. Romo, "On the Concept of Depth for Functional Data," *J. Am. Stat. Assoc.*, vol. 104, no. 486, pp. 718–734, Jun. 2009, doi: 10.1198/jasa.2009.0108.

[24] Z. Jiang, "A Survey on Spatial Prediction Methods," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 9, pp. 1645–1664, Sep. 2019, doi: 10.1109/TKDE.2018.2866809.

[25] H. Li, Q. Yang, and T. Li, "Wind turbine wake prediction modelling based on transformer-mixed conditional generative adversarial network," *Energy*, vol. 291, 2024, doi: 10.1016/j.energy.2

[26] H. Wang, J. Ye, L. Huang, Q. Wang, and H. Zhang, "A multivariable hybrid prediction model of offshore wind power based on multi-stage optimization and reconstruction prediction," *Energy*, vol. 262, 2023, doi: 10.1016/j.energy.2022.125428.

[27] Z. Liu, R. Ma, and Y. Zhong, "Assessing and improving reliability of neighbor embedding methods: a map-continuity perspective," *Nat. Commun.*, vol. 16, no. 1, p. 5037, May 2025, doi: 10.1038/s41467-025-60434-9.

[28] A. Ge?ron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow?: concepts, tools, and techniques to build intelligent systems.* O'Reilly Media, Inc., 2019.

[29] G. T. Reddy *et al.*, "Analysis of Dimensionality Reduction Techniques on Big Data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020, doi: 10.1109/ACCESS.2020.2980942.

[30] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(**??**)90125-7.

[31] I. F. Ashari, E. Dwi Nugroho, R. Baraku, I. Novri Yanda, and R. Liwardana, "Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta,"

Journal of Applied Informatics and Computing, vol. 7, no. 1, pp. 89–97, Jul. 2023, doi: 10.30871/jaic.v7i1.4947.

[32] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979, doi: 10.1109/TPAMI.1979.4766909.

[33] M. Kossakov, A. Mukasheva, G. Balbayev, S. Seidazimov, D. Mukammejanova, and M. Sydybayeva, "Quantitative Comparison of Machine Learning Clustering Methods for Tuberculosis Data Analysis," in *CIEES 2023*, Basel Switzerland: MDPI, Jan. 2024, p. 20. doi: 10.3390/engproc2024060020.

[34] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Stat. Theory Methods*, vol. 3, no. 1, pp. 1–27, 1974, doi: 10.1080/03610927408827101.

[35] V. Lavikainen and P. Fränti, "Clustering district heating customers based on load profiles," *Applied Computing and Intelligence*, vol. 4, no. 2, pp. 269–281, 2024, doi: 10.3934/aci.2024016.

[36] H. K. Kanagala and V. V. Jaya Rama Krishnaiah, "A comparative study of K-Means, DBSCAN and OPTICS," in *2016 International Conference on Computer Communication and Informatics (ICCCI)*, IEEE, Jan. 2016, pp. 1–6. doi: 10.1109/ICCCI.2016.7479923.

[37] G. Hou, J. Wang, and Y. Fan, "Wind power forecasting method of large-scale wind turbine clusters based on DBSCAN clustering and an enhanced hunter-prey optimization algorithm," *Energy Convers. Manag.*, vol. 307, p. 118341, May 2024, doi: 10.1016/j.enconman.2024.118341.

[38] X. Zhang and A. Natarajan, "Gaussian mixture model for extreme wind turbulence estimation," *Wind Energy Science*, vol. 7, no. 5, pp. 2135–2148, Oct. 2022, doi: 10.5194/wes-7-2135-2022.

[39] R. Marcille, M. Thiébaut, P. Tandeo, and J.-F. Filipot, "Gaussian mixture models for the optimal sparse sampling of offshore wind resource," *Wind Energy Science*, vol. 8, no. 5, pp. 771–786, May 2023, doi: 10.5194/wes-8-771-2023.

[40] R. Haider, "Physics-Aware Optimization and Data-Driven Methods for Low-Carbon power Systems - ProQuest," MASSACHUSETTS INSTITUTE OF TECHNOLOGY(MIT), 2023. Accessed: Nov. 19, 2025. [Online]. Available: https://www.proquest.com origsite=gscholar&cbl=18750&diss=y

[41] R. A. FISHER, "THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS," *Ann. Eugen.*, vol. 7, no. 2, pp. 179–188, Sep. 1936, doi: 10.1111/j.1469-1809.1936.tb02137.x.

[42] H. Hamla and K. Ghanem, "A Hybrid Feature Selection Based on Fisher score and SVM-RFE for Microarray Data," *Informatica*, vol. 48, no. 1, Jan. 2024, doi: 10.31449/inf.v48i1.4759.

[43] H. Wang *et al.*, "An adaptive identification method of abnormal data in wind and solar power stations," *Renew. Energy*, vol. 208, pp. 76–93, May 2023, doi: 10.1016/j.renene.2023.03.081.