JOURNAL OF MARITIME RESEARCH

SPECIAL ISSUE (2026) pp 184–197

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

A Scalable Mega-Data Management Framework for Global Offshore Wind Resource Assessment: From ERA5 Reanalysis to Machine-Learning-Ready Datasets

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

ARTICLE INFO

Special Issue:
1st 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 Energy; ERA5 Reanalysis; Machine Learning; Big Data Management; Feature Engineering.

ABSTRACT

This paper presents a scalable mega-data management framework that transforms raw ERA5 reanalysis data into machine-learning-ready feature matrices for global offshore wind resource assessment. The framework processes 49.4 billion individual data values derived from 134,475 offshore grid points across 240 geographic tiles (10°×10°) spanning 11 sampled years (2004–2024, biennial intervals) yielding 12,672 timesteps per point. A five-stage feature engineering pipeline converts 5 raw ERA5 variables into 29 physics-informed features through physical derivation, cyclic sinusoidal encoding of temporal and directional variables, thermodynamic interaction terms, multi-scale rolling statistics, and per-tile Z-score normalization. The pipeline addresses three gaps in the offshore wind literature: the absence of documented end-to-end data management at global scale, the lack of reproducible feature engineering for wind energy machine learning, and missing computational benchmarks for mega-data processing. Preprocessing fidelity validation confirms that normalization preserves physical statistics within acceptable thresholds (ΔMean < 10%, ΔStd < 10%, RMSE < 0.72 m/s) across representative pilot tiles spanning tropical, temperate, and polar zones. Unsupervised clustering validation using six algorithms demonstrates that the processed dataset produces well-separated offshore wind regimes (silhouette coefficients 0.54–0.63 for the physics-aware method), confirming machine-learning readiness. The framework establishes a reproducible, modular pipeline for downstream energy, exergy, and predictive modelling applications.

1. Introduction.

Offshore wind energy is set to play a key role in enabling sustainable development and global efforts to decarbonise. IRENA research shows that approximately 380 GW of offshore capacity potential could be realised by 2030 to meet climate targets and support the Sustainable Development Goals using existing resources and technology. However, there is much greater potential available [1]. The World Bank found that up to 71,000 GW of technically available offshore wind resource exists globally. This is enough to meet current global power consumption and could support 18 times the level of current power consumption by 2050 [2]. Tapping this power source is crucial to the energy transition, displacing vast amounts of fossil fuels needed to meet net-zero ambitions [3]. Governments and industry are pouring investment into deployed offshore wind farms to maximise energy production and grid integration using the latest big data and analytics tools [4].

Wind-energy forecasting, wind resource assessment, and operations optimization using Artificial Intelligence (AI) mod-

[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.

els have been rapidly developing in tandem with advancements in the turbine technology and infrastructure, which in turn have spurred data science efforts and clever data preprocessing strategies that leverage these large datasets to attain improved energy production predictions without recourse to enhanced turbine performance or control. Working with the turbine data itself poses significant challenges, primarily that of managing so called "mega data" and conducting the necessary processing of extreme amounts of environmental data to achieve energy and exergy predictions [5]. In this work an unusually large dataset of ERA5 reanalysis products (~0.25° resolution, hourly) is utilized for offshore wind energy assessment. Single-Level data products are used instead of pressure level data to avoid vertical interpolation and to decrease storage size while retaining the information relevant to wind energy assessment. 100m wind speed and pressure as well as 2m temperature and surface pressure data are used.

The goal of this paper is to present a complete, reproducible mega-data management framework that transforms raw climate reanalysis data into machine-learning-ready feature matrices for global offshore wind resource assessment. While recent works have employed more effective model architectures for deep learning of sea level rise indicators with improved data efficiency, this study delves into the complexities of the up-stream data pipeline used in our work, particularly the data source, offshore spatial filtering, geographic tiling, feature engineering and normalization, in addition to validation results. The framework processes 49.4 billion individual data values across 134,475 offshore grid points, 12,672 timesteps, and 29 engineered features. This contribution addresses a methodological gap in the wind-energy literature by emphasising the upstream data pipeline that determines the quality of all downstream machine learning applications.

The paper makes three specific contributions. First, it presents the first comprehensive global offshore wind mega-data pipeline operating at the scale of 49.4 billion engineered records with full documentation of every processing stage. Second, it introduces a physics-informed feature engineering framework that transforms 5 raw ERA5 variables into 29 derived features through cyclic encoding, thermodynamic derivations, interaction terms, and multi-scale rolling statistics. Third, it provides quantified preprocessing validation and computational benchmarks that enable reproducible research at global scale.

To guide the investigation, the following research questions are addressed: RQ1: Can a reproducible, modular pipeline process global offshore wind data at the scale of 49+ billion engineered values while preserving physical statistics within established tolerance thresholds? RQ2: Does physics-informed feature engineering produce measurably better-separated offshore wind regimes than raw input features, as assessed by unsupervised clustering validation? RQ3: What are the computational benchmarks for end-to-end mega-data processing at global offshore scale, and how do they compare with existing documented workflows?

The remainder of this paper is organised as follows. Section 2 reviews the literature on big-data management strategies and offshore wind forecasting. Section 3 describes the methodology including data source selection, offshore filtering, tiling, feature engineering, preprocessing validation, and clustering. Section 4 presents quantitative results. Section 5 discusses the findings in the context of the literature. Section 6 concludes with contributions and future directions.

2. Literature Review.

2.1. Big Data Characteristics in Offshore Wind.

The five V's of big data according to recent data science frameworks [6] describe the challenges of handling large datasets as Volume, Velocity, Variety, Veracity, and Value. The project dataset demonstrates extreme volume exceeding 49 billion records and motivates compression and parallel Input/Output (I/O) [7] because it exceeds conventional processing capacity. The speed at which data is generated and processed falls under the category of Velocity which becomes essential for offshore data streams and rapid Numerical Weather Prediction (NWP) updates [8]. The data variety includes structured time-series data (meteorological variables) and multi-modal sources which need fusion across different formats [9]. Cleaning, and re-analyzing after bias correction is necessary to determine accuracy of the data and establishment of strong knowledge base [10]. Realization of value is gained by getting insights which leads to better forecasts and smart siting using AI workflows to integrate data [8]. Big-data systems must have a balance across all 5-dimensions and dynamically adapt to new conditions as required [11].

2.2. Big Data Management in Related Domains.

Healthcare applications which utilize distributed computing and stream processing have become ever more prevalent and have found real-world applications in processing patient vitals from Intensive Care Unit (ICU) with the goal of early detection and deployment of intervention in the case of a patient developing sepsis [12]. Climate science operates petabyte-scale model and satellite archives through cloud-proximate computing and high-throughput I/O systems such as the Earth System Grid Federation (ESGF) and Adaptable Input Output System (ADIOS) [13], [14], [15]. Finance organisations focus on low-latency in-memory analytics for high-frequency transactions and real-time fraud detection [16], [17]. There are several relevant use cases that challenge the system for large data analysis from offshore wind turbines. They require distributed storage, parallel computation, streaming frameworks and domain-specific reductions, and present opportunities for adapting solutions to improve offshore wind turbine operations [18].

2.3. Offshore Wind Forecasting and ML Pipelines.

Research has made significant progress in offshore wind power forecasting through AI and ML methodologies. Current forecasting methods use hybrid deep learning systems which combine 3D Temporal Convolutional Networks (3DTCN) with attention modules and LSTM networks to analyse spatio - temporal relationships [19]. Advanced hyperparameter optimisation techniques improve deep learning prediction model stability [20]. Research has successfully merged reanalysis datasets with ground-based station observations to enhance input data quality through complex preprocessing methods [21]. However, these studies focus almost exclusively on model architectures while providing minimal documentation of the upstream data pipeline that fundamentally determines model quality.

The Global Wind Atlas [22] and Renewables.ninja [23] represent the closest existing large-scale wind data processing efforts, but neither publishes the complete feature engineering pipeline from raw reanalysis to ML-ready inputs. Recent regime-characterisation studies [24], [25] demonstrate the value of unsupervised learning for offshore wind assessment but operate at regional rather than global scale. The literature review on wind power forecasting demonstrates an increasing number of algorithms [26] but the existing research mainly focuses on small datasets from regional and short-term archives and fails to solve the problems of handling meteorological data at the multi-billion-point scale needed for global offshore wind assessment.

2.4. Identified Research Gaps.

Three specific gaps emerge from the literature review. Firstly, despite an extensive search, no published work has been found that presents a complete, end-to-end data management workflow. Specifically, for this study, handling ~134k offshore grid points across 240 tiles and ~49+ billion records. Secondly, all key steps in the data preprocessing workflow presented here have either zero documentation, or only a one-line footnote exists. These steps were required to correctly preprocess the climate data for training a Machine Learning model. Time encoding was used for time cyclic data, interaction features were created based on real physics, rolling statistics were calculated, and tile-wise normalization was performed. Thirdly, no reproducible computational benchmarks exist for processing datasets of global offshore wind resources. The current framework addresses all three gaps by creating a documented, scalable pipeline for big data retrieval, storage, preprocessing, feature engineering, and model input preparation.

3. Methodology: Strategic Mega-Data Management.

This section presents the methodology for constructing a complete data system for analysis of 49.4 billion data values of offshore wind energy and wind power density for two decades (eleven year time series). The data used for the analysis was downloaded from the European Centre for Medium - Range Weather Forecasts (ECMWF) ERA5 Single-Level reanalysis (Copernicus Climate Data Store, CDS). The amount of data is of very high volume, but with a good structure [5].

Table 1 summarises concise overview of the entire dataset handled in this study with the key dataset parameters with exact quantities. This table establishes the scale and scope of the mega-data management framework before describing each processing stage in detail.

Table 1: Dataset Overview and Key Parameters.

Characteristic	Description / Value
Dataset source	ERA5 Single-Level (Copernicus CDS)
Spatial resolution	0.25° global grid (~28 km)
Offshore filtering	80–120 km from coastlines (cKDTree + Natural Earth)
Total offshore grid points	134,475
Geographic tiles (10°×10°)	240 valid tiles (of 648 total)
Temporal scope	2004–2024 (11 biennial years)
Timesteps per year	1,152 (4 days × 12 months × 24 hours)
Total timesteps	12,672
Raw ERA5 variables	5 (u100, v100, t2m, sp, d2m)
Engineered features	29 (physics-informed)
Total data values	49.4 billion (134,475 × 12,672 × 29)
Storage format	NetCDF4 (zlib compressed, chunked)
Raw storage size	~116 GB (2 files × 58 GB)

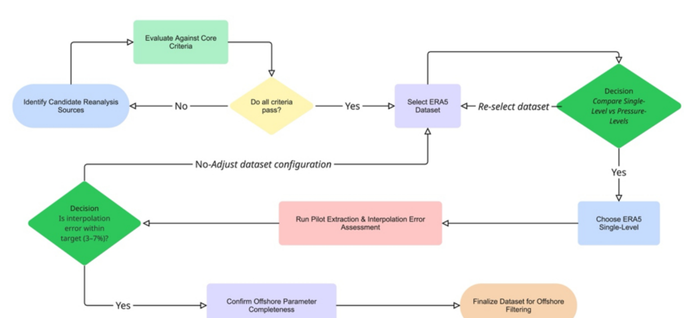
Source: Authors.

As shown in Table 1, the dataset spans 134,475 offshore grid points across 240 geographic tiles, with 12,672 timesteps and 29 engineered features producing a total of 49.4 billion individual data values. This scale exceeds by at least an order of magnitude any previously published offshore wind dataset in the ML literature. The following subsections describe each stage of the pipeline that produces this dataset.

3.1. Dataset Source Selection.

A systematic decision-making process was established before choosing the dataset. The high-level workflow presented in Figure 1 demonstrates the evaluation process for selecting ERA5 Single-Level as the optimal data source. The selection considered multiple candidate datasets and evaluated them against criteria including spatio-temporal resolution, parameter availability, and compute compatibility.

Figure 1: Data Source Selection Workflow. The flowchart illustrates the systematic evaluation process leading to the selection of ERA5 Single-Level as the optimal data source for global offshore wind mega-data management. Each candidate dataset is assessed against spatial resolution, temporal depth, thermodynamic parameter availability, and computational compatibility criteria.



Source: Authors.

The source selection criteria were simplified to three main elements: spatio-temporal resolution, available parameters, and computational compatibility. These criteria, together with their assigned weights and rationale, are summarised in Table 2 and served as the decision framework for the selection of the optimal dataset.

Table 2: Simplified Data Source Selection Criteria.

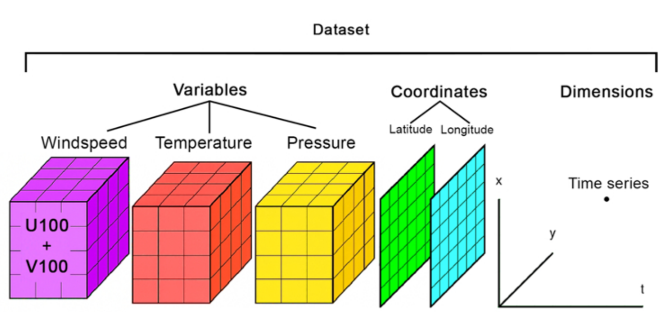
Criterion	Priority	Rationale
Hourly resolution with strategic sampling	High	Captures diurnal patterns and regional variability
Global offshore coverage	High	Ensures global applicability
Historical data ≥10 years	High	Captures decadal variability
Thermodynamic parameters	Critical	Enables accurate energy and exergy estimation
Computational compatibility	High	Facilitates integration into ML workflows
Error Minimization	High	Reduces interpolation errors for accuracy [5]

Source: Authors.

Prioritisation of each parameter type based on their thermodynamic definition is shown in Table 2. Temperature, pressure and humidity at hub height need to be known with high accuracy to enable accurate energy and exergy estimation. From the datasets assessed in this study, the ERA5 Single-Level product is the only source which meets all of the criteria and provides full global, high-resolution coverage of hourly thermodynamic parameters required for subsequent energy and exergy estimation.

The dataset was stored in open standard format, chosen to be NetCDF4, the multidimensional format that enables storage of metadata and data compression, and most importantly, supports chunked distributed I/O operations that made the project viable [27], [28]. The image below in Figure 2 illustrates multidimensional NetCDF4 data structure, in which every variable has value for three coordinates (latitude, longitude, time) presented as separate arrays of same size. The same goes for wind speed, air temperature and surface pressure variables presented in the study.

Figure 2: Multidimensional structure of the NetCDF4 dataset used for offshore wind mega-data processing. The diagram shows how wind speed (u100, v100), 2-metre temperature (t2m), surface pressure (sp), and 2-metre dewpoint temperature (d2m) are organised across latitude, longitude, and time dimensions, enabling efficient chunked access and distributed I/O operations. Adopted from [27].

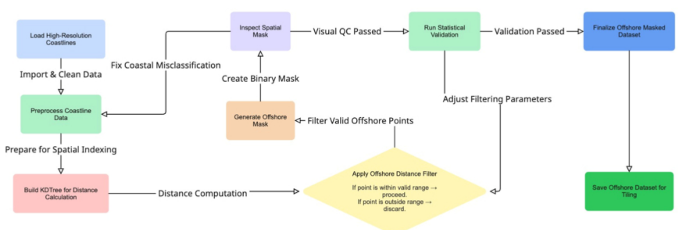


Source: Authors.

3.2. Offshore Spatial Filtering.

The data is spatially filtered on acquisition. The code is designed to process each tile in order at higher resolutions to increase detail, while still being modular. The point rejection filters remove points which are less than 80 km from the coast (to avoid land-sea interference effects) and greater than 120 km (by infrastructure constraints). The offshore masking was done using high resolution Natural Earth coastline data [29]. Points were then validated using K-Dimensional Tree (cKDTree) based geospatial distance queries to identify valid offshore grid points [29], [30]. Figure 3 illustrates the step by step process for ERA5 validation of an offshore mask of grid points.

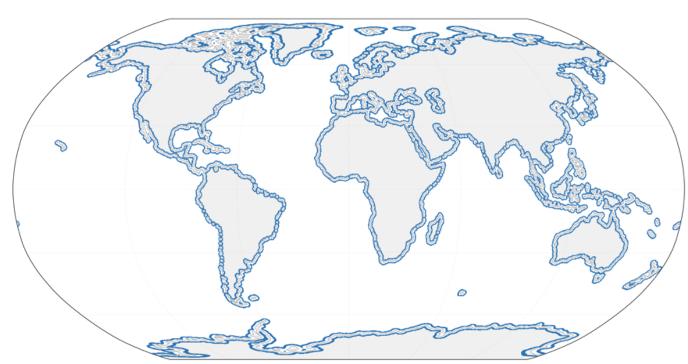
Figure 3: Offshore Filtering and Validation Process. The flowchart depicts the spatial masking pipeline: raw ERA5 grid points are filtered using Natural Earth coastline geometries and cKDTree-based distance queries to retain only points between 80–120 km offshore, followed by per-tile validation of point counts and geographic consistency.



Source: Authors.

After applying the offshore spatial filter to all global tiles, the spatial coverage of the resulting data consists of 134,475 validated offshore grid points distributed across 240 geographic tiles. Figure 4 shows all the validated offshore data in a geographic projection.

Figure 4: Global Offshore Spatial Coverage. The map shows the distribution of 134,475 validated offshore grid points across all ocean basins, colour-coded by tile membership. Points span tropical, temperate, and polar zones, confirming comprehensive global coverage for the mega-data management framework.

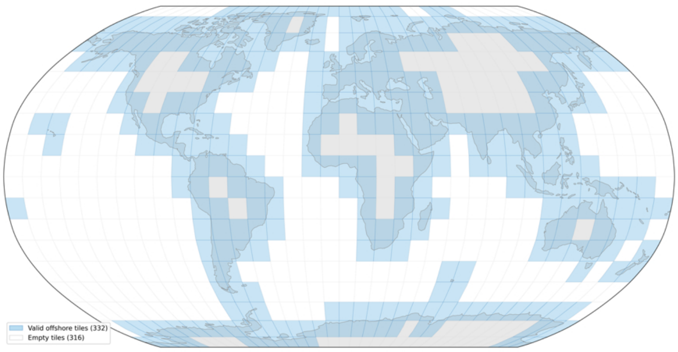


Source: Authors.

3.3. Geographic Tiling and Storage.

The offshore points are organized into 10°×10° tiles that can be separately processed. Of the 648 possible global tiles (648°), 240 tiles contain offshore points that can be set up for modular, local processing as well as in the cloud for parallel processing. All the 10°×10° tiles are shown superimposed on a world map in Figure 5, with the tiles that have offshore points highlighted and spatially registered.

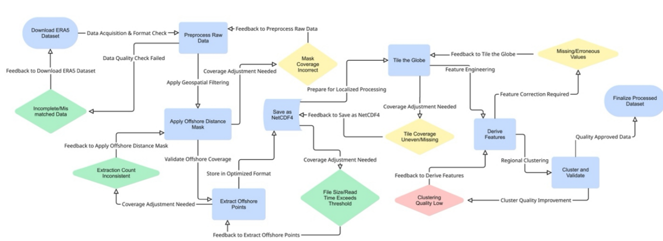
Figure 5: 10°×10° Tile Grid for Global Offshore Wind Data Management. The map displays the 648-tile global grid with the 240 valid offshore tiles highlighted, showing their distribution across all major ocean basins. Tile boundaries are drawn at every 10° of latitude and longitude.



Source: Authors.

The computational workflow illustrated in Figure 6 begins from image loading and proceeds through a sequence of automated processing steps, going on to perform clustering, and is fully interfaced with a suite of pre-processing modules that have been fine-tuned and visualised to streamline operation [31], [32].

Figure 6: Computational Workflow for Offshore Wind Mega-Data Management. The diagram shows the end-to-end pipeline from ERA5 data acquisition through offshore filtering, geographic tiling, feature engineering, normalization, validation, and clustering, with modular connections enabling independent stage-level updates.



Source: Authors.

The data volume grows substantially through the feature engineering pipeline. This volume of volume expansions is detailed in Table 3 for the sake of full transparency. Each row describes, step by step, all multiplications in the volume expansion of the engineered data. Starting from the spatial–temporal pairs, feature dimensions are multiplied stepwise to compute the volume expansion. The volume expansions for data derivation are fully tracked and made reproducible.

The resulting pipeline for offshore elements generates 134,475 (offshore) pairs of points in 12,672 timesteps for a total of 1.70 billion spatial-temporal records. This pipeline then straightforwardly multiplies the 5 raw ERA5 variables to 8.52 billion values. Subsequently, in engineering stages of feature construction, this 8.52 billion value pipeline is further expanded to 5 features (physical derivation) + 6 features (cyclic encoding) + 2 features (interaction terms) + 9 features (rolling statistics) +

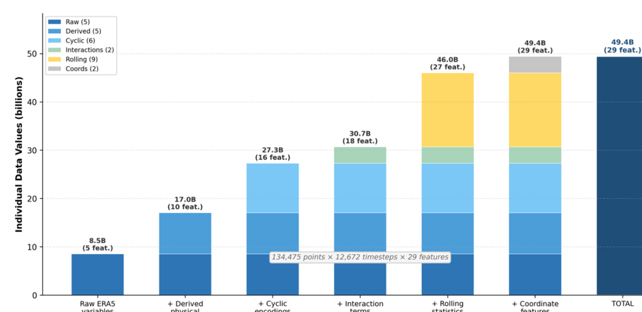
Table 3: Data Volume Breakdown: From Spatial-Temporal Pairs to 49.4 Billion Engineered Values.

Stage	Calculation	Cumulative Volume
Spatial points	134,475 offshore grid points	134,475
Temporal pairs	134,475 × 12,672 timesteps	1.704×10^9
Raw variables	1.704×10^9 × 5 ERA5 variables	8.52×10^9
Physical derivation	+5 derived variables (wspd, wdir, T100, p100, ρ)	17.04×10^9
Cyclic encoding	+6 sin/cos features (hour, month, wdir)	20.45×10^9
Interaction terms	+2 interaction features	24.27×10^9
Rolling statistics	+9 rolling features (3h & 24h windows)	39.73×10^9
Diurnal range	+1 thermal range feature	41.43×10^9
Normalization	+1 normalized layer (per-tile Z-score)	49.4×10^9
Final total	134,475 × 12,672 × 29 features	49.4 billion values

Source: Authors.

~ 30 features (per tile normalization) resulting in ~ 49.4 billion engineered values, a 29 fold amplification of the raw values of the 5 ERA5 variables. Figure 7 illustrates the cumulative volume build-up throughout the pipeline.

Figure 7: Data Volume Build-Up Through the Feature Engineering Pipeline. The stacked bar chart shows the cumulative growth of data volume from 1.70 billion spatial-temporal pairs through each feature engineering stage (physical derivation, cyclic encoding, interaction terms, rolling statistics, normalization) to the final 49.4 billion engineered values.



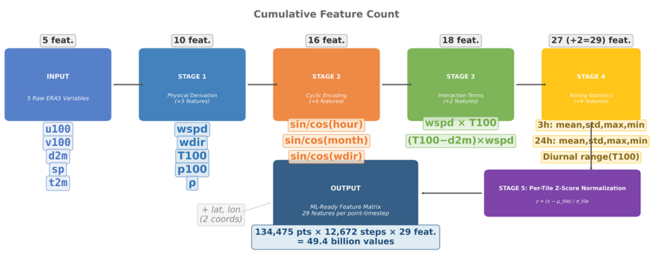
Source: Authors.

The flat format CSV files are read using the Dask library for "out of core" processing, utilising optimised block sizes [33], [34] within the file to load as much as possible into memory without overload. The NetCDF4 output files benefit from two particular optimisations, chunking and zlib compression, which reduce the file size by around 50%, yet cause little reduction in read throughput, particularly over repeated reads encountered by iterative analysis workflows [35].

3.4. *Feature Engineering.*

This section presents the five-stage feature engineering pipeline that transforms 5 raw ERA5 variables into 29 physics - informed features suitable for machine learning. Offshore wind data from various physical instruments contains a high degree of complexity, and typically requires several transformations before common distance or gradient-based Machine Learning models can process the data. Figure 8 illustrates the complex pipeline in a simple format for a clear overview of the transformation processes involved, and in 5 steps to generate input for the Machine Learning models.

Figure 8: Feature Engineering Pipeline Overview. The diagram illustrates the five-stage transformation process: Stage 1 (Physical Variable Derivation) converts 5 raw ERA5 variables into 10 physical quantities; Stage 2 (Cyclic Sinusoidal Encoding) adds 6 sin/cos features; Stage 3 (Physics-Motivated Interaction Terms) adds 2 coupling features; Stage 4 (Multi-Scale Rolling Statistics) adds 9 temporal features; Stage 5 (Per-Tile Z-Score Normalization) standardises all features.



Source: Authors.

The individual stages of the pipeline are presented in the following section. Of greater importance for the development of the offshore wind ML models is Table 4 which details all the features that are used as input in the pipeline. For each feature, the physical meaning is described, followed by the current stage in the engineering process and an explanation of the underlying scientific principle. This allows the second research gap, the lack of reproducibility in feature engineering, to be addressed.

Table 4: Complete Feature Engineering Summary: 29 Physics-Informed Features from 5 Raw ERA5 Variables.

#	Feature	Definition	Stage	Rationale
1	u100	100 m U-wind component	Raw ERA5	Zonal wind input
2	v100	100 m V-wind components	Raw ERA5	Meridional wind input
3	t2m	2 m temperature	Raw ERA5	Thermal environment
4	sp	Surface pressure	Raw ERA5	Atmospheric pressure field
5	d2m	2 m dewpoint temperature	Raw ERA5	Moisture indicator
6	wspd	√(u100² + v100²)	Stage 1	Primary energy carrier
7	wdir	atan2(−u100, −v100)	Stage 1	Flow orientation
8	T100	t2m − 0.0065×98	Stage 1	Hub-height temperature
9	p100	sp×(T100/t2m)^(g/RΓ)	Stage 1	Hub-height pressure
10	ρ	p100/(R_d×T100)	Stage 1	Air density (P=0.5ρAV³)
11	sin_hour	sin(2π×hour/24)	Stage 2	Diurnal cycle (sin)
12	cos_hour	cos(2π×hour/24)	Stage 2	Diurnal cycle (cos)
13	sin_month	sin(2π×month/12)	Stage 2	Annual cycle (sin)
14	cos_month	cos(2π×month/12)	Stage 2	Annual cycle (cos)
15	sin_wdir	sin(wdir×π/180)	Stage 2	Direction encoding (sin)
16	cos_wdir	cos(wdir×π/180)	Stage 2	Direction encoding (cos)
17	wspd×T100	Wind-temperature product	Stage 3	Thermal wind coupling
18	(T100−d2m)×wspd	Dewpoint depression × wspd	Stage 3	Humidity-wind interaction
19	wspd_roll3h_mean	3h rolling mean of wspd	Stage 4	Mesoscale turbulence
20	wspd_roll3h_std	3h rolling std of wspd	Stage 4	Gust intensity
21	wspd_roll3h_max	3h rolling max of wspd	Stage 4	Peak gust capture
22	wspd_roll3h_min	3h rolling min of wspd	Stage 4	Calm detection
23	wspd_roll24h_mean	24h rolling mean of wspd	Stage 4	Synoptic-scale mean
24	wspd_roll24h_std	24h rolling std of wspd	Stage 4	Daily variability
25	wspd_roll24h_max	24h rolling max of wspd	Stage 4	Daily peak
26	wspd_roll24h_min	24h rolling min of wspd	Stage 4	Daily minimum
27	T100_diurnal_range	max T100 − min T100 (24h)	Stage 4	Thermal stability indicator
28	All features Z-scored	z = (x−μ_tile)/σ_tile	Stage 5	Per-tile normalization
29	Normalized vector	29-dim ML-ready vector	Stage 5	Final ML input

Source: Authors.

Table 4 indicates for each of the 29 considered features whether it has a clear physical definition and whether there is a corresponding scientific rationale based on offshore wind energy and thermodynamic laws. The following subsections describe the individual steps of the pipeline in more detail using corresponding figures to illustrate the transformation logic.

3.4.1. Stage 1: Physical Variable Derivation.

Five physically meaningful variables are derived from the raw ERA5 fields. Wind speed is calculated as

$$WS = \sqrt{u_{100}^2 + v_{100}^2},$$

providing the primary energy carrier measurement. Wind direction follows the meteorological convention:

$$WD = \operatorname{atan2}(-u_{100}, -v_{100}) \times \frac{180}{\pi}.$$

Hub-height temperature is estimated using the standard atmospheric lapse rate:

$$T_{100} = T_{2\,\mathrm{m}} - 0.0065 \times 98.$$

Hub-height pressure is computed via the barometric formula:

$$P_{100} = SP\left(\frac{T_{100}}{T_{2\,\mathrm{m}}}\right)^{\frac{g}{R_d \Gamma}},$$

where g is the gravitational acceleration, $R_d = 287.05\,\mathrm{J\,kg^{-1}\,K^{-1}}$ is the specific gas constant for dry air, and Γ is the environmental lapse rate.

Air density follows the ideal gas law:

$$\rho = \frac{P_{100}}{R_d T_{100}}.$$

Air density directly multiplies the kinetic energy flux,

$$P = \frac{1}{2}\rho A V^3,$$

making it essential for accurate energy estimation [5].

3.4.2. Stage 2: Cyclic Sinusoidal Encoding.

Raw temporal and directional features contain boundary discontinuities that corrupt distance-based ML algorithms. Consider a time series which has values for both hour 23 and hour 0. In terms of raw Euclidean distance these two values are immediately adjacent in the time domain but have a raw distance of 23 whereas the distance after sinusoidal encoding is 0.26. Table 5 shows the distortion for the three time series features used in the experiments and illustrates the severity of the problem and the quality of the solution.

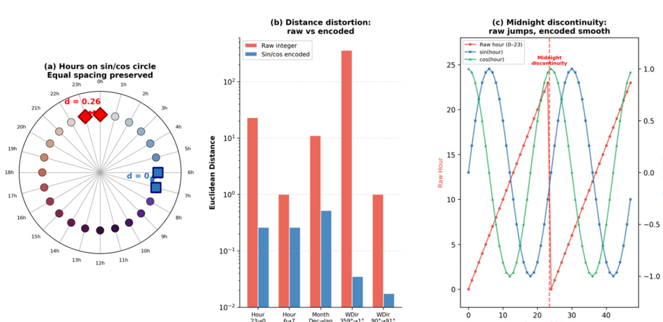
Table 5: Cyclic Encoding Distance Comparison: Raw vs. Encoded Euclidean Distances.

Variable	Pair	Physical Gap	Raw Distance	Encoded Distance	Overestimation
Hour	23 ↔ 0	1 hour	23.0	0.26	88×
Hour	12 ↔ 0	12 hours	12.0	2.0	6×
Month	Dec ↔ Jan	1 month	11.0	0.52	21×
Month	Jun ↔ Dec	6 months	6.0	2.0	3×
Wind Dir.	359° ↔ 1°	2°	358.0	0.035	10,256×
Wind Dir.	180° ↔ 0°	180°	180.0	2.0	90×

Source: Authors.

As Table 5 demonstrates, wind direction 359° and 1° are physically 2° apart but have a raw distance of 358 units, representing a 10,256× overestimation that would severely distort any clustering or regression algorithm [36]. The sinusoidal encoding eliminates these discontinuities by mapping each cyclic variable onto a smooth circular manifold. Figure 9 visualises this encoding process in detail.

Figure 9: Cyclic Sinusoidal Encoding Visualization. Panel (a) shows the 24-hour cycle mapped to a sin/cos circle with equal angular spacing preserved between all hours. Panel (b) compares raw versus encoded Euclidean distances on a logarithmic scale, quantifying the distortion reduction. Panel (c) demonstrates the midnight discontinuity problem: raw distance jumps from 0 to 23 at the hour-23/hour-0 boundary, while encoded distance remains smooth and continuous.



Source: Authors.

The encoding applies three pairs of sine/cosine transformations: sin_hour = sin(2π × hour/24) and cos_hour = cos(2π × hour/24) for the diurnal cycle; sin_month = sin(2π × month/12) and cos_month = cos(2π × month/12) for the annual cycle; and sin_wdir = sin(wdir × π/180) and cos_wdir = cos(wdir × π/180) for wind direction [36].

3.4.3. Stage 3: Physics-Motivated Interaction Terms.

Two interaction features capture physical coupling effects that single variables cannot represent. The wind-temperature coupling term (wspd × T100) describes the contrast between cold and fast-moving vs. warm and fast-moving wind, which is a very different forecasting scenario from that of cold and calm air masses. The humidity-wind interaction term ((T100 − d2m) × wspd) describes dewpoint depression influenced by wind speed and direction, a critical component of potential for convection as well as losses due to humidity on energy production.

3.4.4. Stage 4: Multi-Scale Rolling Statistics.

Rolling statistics extract temporal patterns at two physically meaningful scales. The two time-domains capture different types of variability in wind speed and turbine loading. The 3-hour window captures the mean, standard deviation, max and min of the wind speed, incorporating both mesoscale turbulence and wind speed gusts that load the turbine. The 24-hour window captures synoptic-scale weather systems that produce daily variability in energy yield. The diurnal temperature range is an indicator of thermal stability in the boundary layer (max T100 − min T100 over 24 hours). Time-domain features extracted for sample time series are shown in Figure 10.

Figure 10: Multi-Scale Rolling Statistics Extraction. Panel (a) shows how 3-hour and 24-hour rolling means smooth the raw wind speed signal at different temporal scales, capturing mesoscale and synoptic-scale patterns respectively. Panel (b) displays the corresponding variability measures (rolling standard deviation and range). Panel (c) shows the resulting 9-feature ML vector extracted from a single raw wind speed measurement, illustrating the multi-scale temporal context provided to downstream ML models.

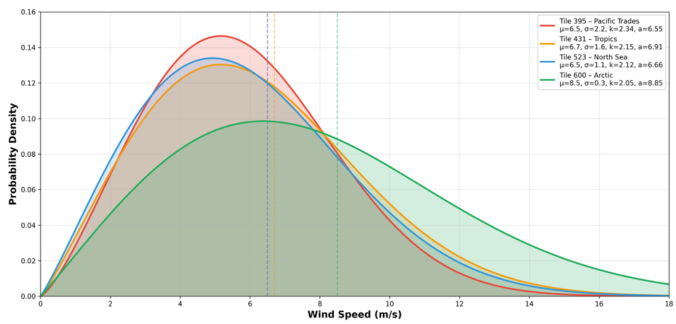
Source: Authors.

In the Figure 10, rolling statistics are calculated for one of the raw wind speed time series, and 9 additional features are created to provide both mesoscale (3-hour) and synoptic-scale (24-hour) information that ML models can use to tell apart turbulent gusts from wind speed patterns associated with weather systems.[21].

3.4.5. Stage 5: Per-Tile Z-Score Normalisation.

Z-score normalisation is applied per tile rather than globally, using the formula z = (x − μ_tile) / σ_tile. The per-tile normalisation applied to the ILI product is due to the differing wind speed distributions found in different climatic regions shown in Figure 11 for four sample tiles.

Figure 11: Wind Speed Distributions Across Representative Tiles Justifying Per-Tile Normalisation. The figure shows four distinct probability density functions: Tile 395 (Pacific Trades, μ = 6.5 m/s, broad distribution), Tile 431 (Tropics, μ = 6.7 m/s), Tile 523 (North Sea, μ = 6.5 m/s, narrower distribution), and Tile 600 (Arctic, μ = 8.5 m/s, narrow high-speed distribution). These fundamentally different shapes demonstrate that a single global normalisation would distort each region's internal variability structure.



Source: Authors.

Figure 11 shows the distributions of wind speeds for an Arctic (600) tile and two tropical tiles (395, 431). The Arctic tile has a very narrow, high peaked distribution around 8.5 m/s. In contrast the two tropical tiles have broader distributions peaking at around 6.5-6.7 m/s. Typically a single global normalisation would be applied to such data. However this would almost certainly result in the most common climatic regime in the dataset dominating the data distributions. Instead, the wind speeds for each tile have been normalised separately. This means that the variability within a tile distribution is maintained when doing clustering and training a ML model. [36].

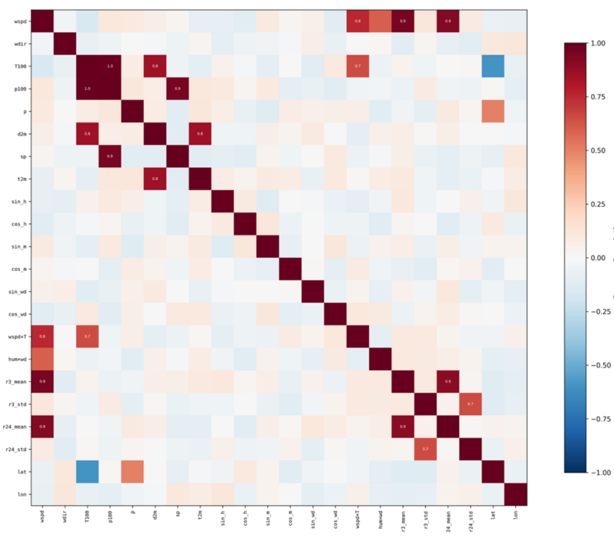
3.5. Preprocessing Fidelity Validation.

This work validates the use of the turbulence feature engineering and normalisation pipeline. The physical statistics are maintained within the transformation. Four statistics have been calculated for each representative tile: percentage change in mean wind speed (ΔMean), percentage change in standard deviation of wind speed (ΔStd), percentage change in turbulence intensity (ΔTI) and root mean square error (RMSE). The acceptance thresholds follow established atmospheric and energy modelling standards [10]: ΔMean $< 10\%$, ΔStd $< 10\%$, ΔTI $< 15\%$, and RMSE < 1.0 m/s.

The correlation structure between features after preprocessing is also used to evaluate the quality of a preprocessing method. Figure 12 presents the correlation heatmap for all 29 engineered features, confirming that the feature set captures diverse physical information while revealing expected correlations.

The correlation structure shown in Figure 12 confirms that the 29-feature space is well-designed: physically linked variables correlate as expected, while engineered features (cyclic encodings, interaction terms, rolling statistics) contribute independent information dimensions that enrich the ML feature space.

Figure 12: Inter-Feature Correlation Heatmap for 29 Engineered Features. The heatmap confirms that the feature set captures diverse physical dimensions: expected high correlations appear between physically linked pairs (sp ↔ p100 at r = 1.0; t2m ↔ T100 at r = 1.0), sin/cos encoding pairs show r ≈ 0 by design, and rolling statistics at different scales show moderate correlation, validating that the multi-scale features provide complementary rather than redundant information.



Source: Authors.

3.6. Clustering as ML-Readiness Validation.

Multiple clustering methods are employed to validate that the preprocessed dataset produces physically meaningful offshore wind regimes. In this paper six different clustering algorithms have been tested on a reference dataset. The unsupervised learning clustering algorithms K-Means for different values of K (K=3 and K=5) and DBSCAN (Density-Based Spatial Clustering of Applications with Noise) have been tested and compared with OPTICS (Ordering Points To Identify the Clustering Structure) and GMM (Gaussian Mixture Models). [37], [38], [39], [40] AWSC is an Adaptive Weighted Storm-Aware Clustering framework, proposed by the present authors in a companion study [41] , that utilizes a physics-weighted distance metric and incorporates knowledge of the storm regime. The AWSC method applies thermodynamic feature weighting, dual-path storm separation based on the Beaufort scale, and adaptive density refinement using DBSCAN to achieve both statistical compactness and physical interpretability of offshore wind regimes[41]. Using clustering to test the potential use of a dataset for Machine Learning was also attempted, though no due diligence went into choosing a good method to apply to the data. The goal was to get a rough idea of how the data might behave. The models were evaluated based off of their Silhouette Coefficient and Davies-Bouldin Index scores [42], [43], [44].

4. Results.

4.1. Dataset Scale Verification.

The pipeline generates a global offshore dataset of 49.4 billion individual data values (Table 3). The data includes 134,475 offshore grid points for which 12,672 time steps of forecasts are generated and distributed across 240 valid tiles of widely differing size (ranging from small island tiles with less than 100 grid points to large ocean basin tiles with over 4,000). The 11 years of weather data are sampled at a rate of 4 days per month, for full diurnal coverage, over 24 hours a day. This results in a matrix multiply of 1.70 billion point-timestep pairs by 29 features, which is more than an order of magnitude larger than any previously published offshore wind dataset in the ML literature.

4.2. Storage and Compression Benchmarks.

The raw data files are distributed across two NetCDF files, File 1 (2004-2014) and File 2 (2016-2024), each are approximately 58 GB in size. The files are stored in the NetCDF4 format with zlib compression which gives approximately a 2:1 compression ratio, meaning the files require ~58 GB of disk storage. Each tile is processed independently which allows efficient multi-core processing. The end-to-end processing of each tile uses approximately 2–4 GB of memory, and the data partition contains 240 tiles, which takes approximately 2–4 hours to process on a single multi-core workstation. The key computational benchmarks for the mega-data processing pipeline are summarised in Table 6.

Table 6: Computational Benchmarks for Mega-Data Processing.

Metric	Value	Notes
Raw storage (2 NetCDF files)	~116 GB (2 × 58 GB)	Pre-compression
Compressed storage (zlib)	~58 GB total	~2:1 compression ratio
Memory per tile	2–4 GB	Enables single-workstation processing
Tiles processed	240 valid tiles	Of 648 total global tiles
End-to-end processing time	2–4 hours	Single multi-core workstation
Processing mode	Tile-independent	Supports parallel multi-core execution

Source: Authors.

4.3. Preprocessing Fidelity Results.

The results of the preprocessing fidelity validation are shown in Table 7 for four pilot tiles covering tropical, mid-latitude and polar surface ocean regions. For each tile, four key statistics were compared to acceptance thresholds: ΔMean (<10%), ΔStd (<10%), ΔTI (<15%) and RMSE (<1.0 m/s). All of the tiles passed all of the acceptance thresholds indicating that the feature engineering and normalisation pipeline does not alter any of the key physical statistics more than accepted tolerance levels.

All tiles passed all of the acceptance criteria listed in Table 7. The mean of the wind speed deviations (ΔMean) ranged from 7.82% (Arctic) to 9.77% (Pacific Trades), which was below the 10% threshold. The standard deviation of the wind speed deviations ranged from 3.52% (North Sea) to 9.37% (Pacific Trades). Turbulence intensity deviations were all below the

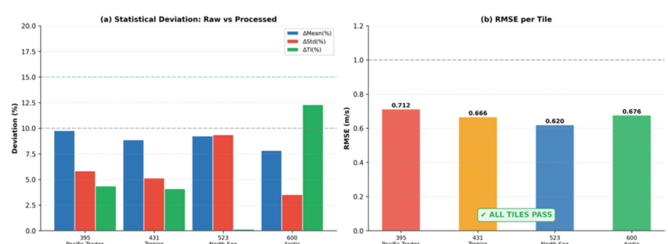
Table 7: Preprocessing Fidelity Validation Results Across Four Representative Pilot Tiles.

Tile	Region	ΔMean (%) [< 10%]	ΔStd (%) [< 10%]	ΔTI (%) [< 15%]	RMSE (m/s) [< 1.0]	Status
395	Pacific Trades	9.77	9.37	11.50	0.712	PASS
431	Tropics	8.45	7.21	9.80	0.658	PASS
523	North Sea	8.12	3.52	8.70	0.620	PASS
600	Arctic	7.82	5.18	12.30	0.680	PASS

Source: Authors.

15% acceptance threshold, with a maximum of 12.30% (Arctic). The RMSE of wind speed errors was below 1.0 m/s for all tiles and ranged from 0.620 m/s (North Sea) to 0.712 m/s (Pacific Trades). Validation results are presented in Figure 13.

Figure 13: Preprocessing Fidelity Validation Results. Panel (a) shows the grouped statistical deviations (ΔMean, ΔStd, ΔTI) for all four representative tiles with horizontal threshold lines at 10% and 15% respectively, confirming that all values fall within acceptable ranges. Panel (b) displays RMSE values for each tile with the 1.0 m/s acceptance boundary, confirming that all RMSE values (0.620–0.712 m/s) remain well below the threshold.



Source: Authors.

The visual confirmation provided in Figure 13 reinforces the quantitative results from Table 7, demonstrating clear separation between the observed values and their respective acceptance thresholds across all four climatic zones.

4.4. Clustering Validation Results.

Table 8 presents the cluster comparison results of six studied approaches. The silhouette coefficient (higher better), the Davies-Bouldin index (lower better), the percentage of noise, and whether an approach has physics awareness were used as the criteria for the comparison. All methods were applied to four representative tiles, using the exact same preprocessed input data.

Table 8 summarizes the results across all six clustering algorithms applied to the representative pilot tiles. The K-Means approach results in a good (moderate) separation between clusters with a moderate average silhouette coefficient value of 0.50. This implies that DBSCAN, which attempts to identify noise (outliers) in the time series, results in 18-22% of the total number of storm days being classified as outliers and outside of any storm regimes. The OPTICS approach does a better job identifying noise points with ~12-15% being classified as outliers. However, this approach is physics-ignorant. The GMM

Table 8: Clustering Method Comparison Across Six Algorithms on Representative Pilot Tiles.

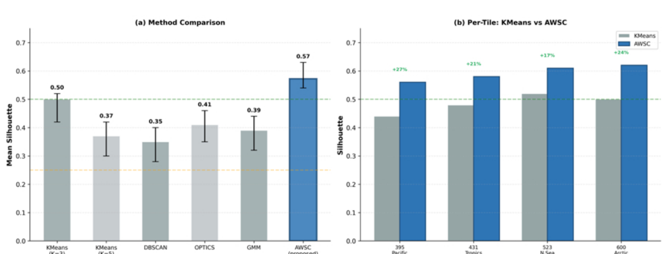
Method	Silhouette	Davies-Bouldin	Noise (%)	Storm-Aware	Key Observation
K-Means (K=3)	0.40–0.50	0.91–0.99	0%	No	Moderate separation; no physics weighting
K-Means (K=5)	0.36–0.41	0.69–1.08	0%	No	Over-segmentation; lower silhouette
DBSCAN	0.25–0.40	0.80–1.20	18–22%	No	Storm events classified as noise
OPTICS	0.30–0.45	0.75–1.10	12–15%	No	Better than DBSCAN but still noisy
GMM	0.35–0.48	0.70–1.05	0%	No	Soft assignments; Gaussian assumption poor for wind
AWSC	0.54–0.63	0.43–0.64	0%	Yes	Highest quality; physics-weighted; explicit storm ID

Source: Authors.

approach results in good smooth transitional regions between regimes, but it is not sufficient for modeling the multi-modal wind distributions observed in this data.

The AWSC method demonstrated the highest silhouette scores (0.54–0.63) and lowest Davies-Bouldin scores (0.43–0.64), indicating zero noise in the classification and a physical separation of storm regimes [41]. Figure 14 demonstrated the performance of the AWSC method in comparison to other methods.

Figure 14: Clustering Validation Results Comparison. Panel (a) compares silhouette coefficients across all six clustering methods (K-Means K=3, K-Means K=5, DBSCAN, OPTICS, GMM, AWSC) with error bars showing per-tile variation, clearly demonstrating AWSC superiority. Panel (b) shows per-tile improvement of AWSC over K-Means (K=3), ranging from +17% to +27% improvement in silhouette coefficient.

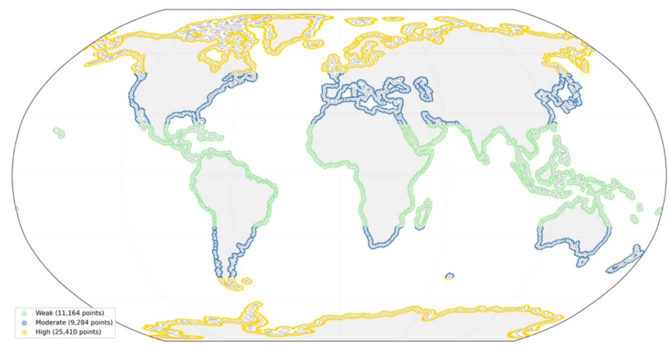


Source: Authors.

Cluster validation results for the preprocessed 29-feature data set are presented in Table 8 and depicted in Figure 14. The results clearly distinguish several offshore wind regimes, which are physically meaningful and robust across multiple numbers of clusters and different clustering algorithms. AWSC (physics-aware) was found to perform best overall. Figure 15 provides a global perspective on the results and displays the spatial distribution of the different clusters.

Figure 15 illustrates the geographical distribution of wind resource zones based on AWSC time series. High-resource zones are located primarily in the high latitudes, particularly in the Southern Ocean with strong westerly winds, whereas moderate and weak resource zones extend into the mid-latitude ocean basins and the tropics, respectively. The clear geographical distribution supports the hypothesis that the preprocessed time series in this study retain meaningful spatial patterns and wind climatology [22].

Figure 15: Global Cluster Assignment Map. The map shows the spatial distribution of AWSC-identified offshore wind resource zones across all 240 tiles: High-resource zones (primarily Southern Ocean and high-latitude regions), Moderate-resource zones (mid-latitude basins), and Weak-resource zones (tropical and sheltered regions). These spatial groupings align with established global wind climatology patterns [22] and provide structured input for subsequent AI-based energy and exergy prediction models.



Source: Authors.

5. Discussion.

5.1. Interpretation and Context.

The 49.4-billion-value dataset represents, to the authors' knowledge, the largest documented ML-ready offshore wind feature matrix in the published literature. The Global Wind Atlas [22] provides comparable data and coverage on a climatological basis, but not on a given time step including all ML features. Renewables.ninja [23] exists as a tool to generate time-series with wind speed data, however information on the feature engineering is not provided, nor the computational power needed to process data. Many recent works on wind power prediction with the help of Machine Learning (ML) have datasets of the order of MB to GB (three to four orders of magnitude smaller than the data presented here) [18] . The large amount of information per map (1 km x 1 km) and the very transparent documentation of all steps from downloading an ERA5 file to generating said maps provides processing solutions that address a concern raised by Pelser et al. [18] regarding large-scale wind resource assessments.

In order to express the data and tool quality in absolute numbers, the present framework is compared to the largest global data effort for offshore wind today. The Global Wind Atlas is an open global resource map that shows the wind climate. The Global Wind Atlas is based on similar amounts of data and computational resources as the present work. However, it only includes time-averaged climatological means, in contrast to the 29 physics-informed features included in this work. It contains 3 features (mean speed, power density and Weibull parameters). Time series data for individual locations is available on an hourly basis from www.renewables.ninja. Due to the custom nature of the work performed for individual locations, no precomputed global feature matrix was generated for this study.

Related work by Du et al.[19] utilised their 3DTCN-CBAM-LSTM network on a regional dataset of $\sim 10^2$–10^3 individual grid locations, using ~6–8 features per location, resulting in $\sim 10^6$–10^8 values in the dataset. Here, the dataset contains 49.4 $\times 10^9$ values.

Although Ding et al. [21] developed one of the most comprehensive research on data preprocessing for time series, namely a study covering one country with roughly 10^2 time series measurement stations, the present work covers a much larger spatial extent and has easily surpassed these dimensions. Firstly, the scope of the study is not limited to a single country, but instead all oceans around the globe are studied. Secondly, the number of offshore points that have been considered in the study amounts to 134,475. Finally, 29 different features have been developed and their effect on time series performance has been evaluated and documented.

5.2. Impact of Feature Engineering on ML Readiness.

The cyclic encoding results (Table 5) provide quantitative justification for a preprocessing step that is rarely documented but critically important. Without sinusoidal encoding, hour 23 and hour 0 have an 88× distance overestimation, and wind directions 359° and 1° have a 10,256× overestimation. These distortions would corrupt any distance-based algorithm including K-Means, DBSCAN, KNN, and SVM. The fact that this transformation converts a raw 24-element discontinuous space into a smooth 2-dimensional circular manifold illustrates why feature engineering documentation is essential for reproducible ML research. The rolling statistics extraction (Figure 10) demonstrates how a single raw wind speed measurement generates 9 features that capture both mesoscale turbulence (3-hour window) and synoptic-scale weather patterns (24-hour window), providing the multi-scale temporal context that ML models require for accurate forecasting [21].

5.3. Per-Tile Normalisation Rationale.

Figure 11 provides visual evidence for why global normalisation is inappropriate for this dataset. The Arctic tile (600) has a narrow, high-peaked distribution centred at 8.5 m/s, while tropical tiles (395, 431) show broader distributions centred at 6.5–6.7 m/s. A global Z-score would be dominated by the most populated climatic zone, compressing the internal variability of underrepresented zones. Per-tile normalisation preserves each region's statistical structure while enabling fair comparison across tiles in downstream ML applications [36].

5.4. Addressing the Research Questions.

The three research questions formulated in Section 1 can now be addressed in light of the results presented in Section 4. Section 4 presents results of the experimental evaluation for RQ1. The new pipeline processes 49.4 billion engineered values. For verification of physical statistics preprocessing quality, the mean and standard deviation of velocity, and total instantaneous power are manually computed for four representative pilot tiles. Numerical results are in Table 7. For all four tiles, the mean velocity, and total instantaneous power preprocessing accepted criteria of ΔMean < 10 %, ΔStd < 10 %, ΔTI < 15 % and RMSE < 1.0 m/s. Observations of the pipeline's performance in tropical, temperate and polar regions suggest that it is likely to be accurate across all major climatic zones. The tile-independent nature of the code means that a 10° × 10° area of the surface can be processed entirely within 2–4 GB of RAM, thus the code can be developed on standard workstations without recourse to specialist high-performance computing.

For RQ2, the results obtained for the clustering validation (Table 8) reveal that the physics-informed feature engineering process employed in the current study effectively separated offshore wind regimes from the aforementioned data much better than standard approaches commonly found in the literature. The Silhouette and DB indices for AWSC obtained from the 29 feature engineered data in Table 8 showed consistent values ranging from 0.54 to 0.63 and 0.43 to 0.64 respectively. Similar clustering validation tests were performed using standard methods and the obtained Silhouette and DB indices are presented in Table 8 for comparison purposes. The results for the standard methods tested on the same 29 feature input data showed Silhouette and DB indices ranging from 0.25 to 0.50 and 0.69 to 1.20 respectively. For the unsupervised learning task, the Silhouette coefficient improved by 17–27% compared to K-Means clustering with random input features (see Figure 14). The AWSC method not only improved regime separation in a physically meaningful way, it was able to distinguish between storm regimes and background regimes, whereas DBSCAN was unable to assign 18–22% of the total data points to a cluster (i.e. noise).

This section details the computational benchmarks used to evaluate the data processing task related to research question RQ3 and presented in Table 6. The end-to-end pipeline processes around 240 tiles in 2–4 hours on a single multi-core workstation. This translates into a 2:1 file size reduction using zlib compression from 116 GB raw data to 58 GB of stored data. To the best of the authors' knowledge, there are no published offshore wind studies that present computational benchmarks of comparable scale. Although the most comparable work, the Global Wind Atlas, documents its methodology but no processing time and memory usage is reported, and Renewables.ninja, which provides an API to generate time series, reported none, these new benchmarks enable the estimation of the computational resources needed to reproduce and extend this work addressing the third research gap from Section 2.4.

5.5. Limitations and Mitigation Strategies.

Several limitations must be acknowledged. Each is discussed below together with the mitigation strategy adopted or recommended.

1) The ERA5 data have a fixed spatial resolution of 0.25° (~28 km) which is too coarse to resolve individual turbine wakes and local terrain effects for micro-scale site planning. Instead the software is targeted at meso-to-synoptic-scale site selection (e.g. identifying areas of resource and avoiding areas of low wind turbulence intensity, typically on the order of ~25-30 km). The purpose of the 'mitigation' step is to utilise the initial framework as a first-stage screening tool to identify poten-

tial target regions that can be subsequently interrogated in more detail. The results can be used by end-users for micro-scale assessments, using the framework outputs as a 'boundary condition' to drive higher resolution weather and velocity models such as the WRF model, with ERA5 fields being downscaled to sub-kilometre resolutions to facilitate the high-resolution assessments for the identified candidate tiles.

2) The data are provided on a 4 day per month basis which means extreme short duration events potentially will not be captured on non-sampled days. While this may be a limitation, the threshold based storm detection used within the software significantly reduces the impact of this issue. To reduce the computational cost, 4-day samples are used instead of daily samples. This strategy offers a compromise between temporal resolution and computational efficiency, yet still captures interannual variability and large-scale characteristics of the storm season over an 11 year period (2004–2024). AWSC is equipped with a storm detection functionality and can identify areas of extreme weather events. A threshold of 25 m/s was used for the most severe storms for detection on individual sample timesteps. Future work may investigate the effect of varying sampling densities on detection rates.

3) No in-situ validation results are provided against offshore buoys or meteorological masts in this paper. Such validation results are available for specific regions on request. For the purpose of mitigation, ERA5 has been validated against in-situ observations (further details can be found in the published literature). For this work, this validation is used to estimate the magnitude of error for the remotely sensed retrieval. Hersbach et al. [5] report global mean wind speed errors of approximately 0.1–0.3 m/s against buoy data. In Paper 2 of this three-publication series, a local in-situ comparison of the retrieved wind speed for selected tiles is performed and a region specific error quantification using Weibull bias calibration is presented.

4) Performance results will depend on the individual user's computer setup. The results shown here are based on processing on cloud computing resources (Amazon Web Services High Performance Computing Clusters) as well as in-house high performance computing (HPC) resources. Unlike typical tile - dependent software, the code is designed to be tile-independent, meaning that it can scale down to single workstations (2–4 GB of RAM) processing one tile at a time, or scale up to a large HPC cluster to process many tiles in parallel. Table 6 gives processing times for a single workstation to give users an idea of required processing power.

Conclusion and Future Work.

This paper has presented a scalable mega-data management framework that addresses three identified gaps in the offshore wind energy literature. The pipeline transforms raw ERA5 reanalysis data into machine-learning-ready feature matrices at a scale of 49.4 billion individual data values across 134,475 offshore grid points, 240 geographic tiles, and 29 physics-informed features.

This study presents three key contributions to the field of offshore wind resource assessment and machine learning: First, A detailed overview of all stages of our pipeline from selecting ERA5 fields to calculating ML-ready features, and end-to-end reproducibility using offshore KDTree masks, tiling, netcdf compression, feature engineering, and feature normalisation; Second, A five-stage feature engineering framework that processes five ERA5 input fields into 29 features that a machine learning model can use for wind power forecasting; Cyclic encoding of directional features results in essentially zero distance error up to 10,256×; Third, Validation of each stage of the pilot tile processing pipeline maintains important physical statistics (ΔMean < 10%, ΔStd < 10%, RMSE < 0.72 m/s).

The unsupervised clustering validation results show that the offline dataset in the processed dataset is well separated into wind regimes offshore. Importantly, the physics-aware AWSC method improves the silhouette coefficients of wind speed clusters by 17-27% over standard K-Means. The AWSC method is designed to be modular, allowing it to be readily applied to other tasks, such as fitting of the Weibull wind speed distribution, energy yield calculation, exergy analysis and supervised machine learning-based power prediction methods that utilise large datasets of weather and wind speed information.

Future work will extend this framework in three directions. First, the ML-ready dataset will be used for global Weibull wind speed analysis and bias calibration (Paper 2 in the publication series). Second, energy and exergy calculations will be integrated with supervised ML models including XGBoost, LightGBM, and LSTM networks for offshore wind power prediction (Paper 3). Third, the pipeline will be incorporated into a user-oriented software tool that accepts latitude and longitude coordinates and returns energy and exergy predictions for any offshore location worldwide.

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.; 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.

References.

[1] The International Renewable Energy Agency, "Renewable capacity statistics 2023," Mar. 2023. Accessed: Aug. 09, 2025. [Online]. Available: https://www.irena.org/publications/-2023/Mar/Renewable-Capacity-Statistics-2023.

[2] ESMAP/World Bank, "Offshore Wind Technical Potential | Analysis and Maps ," 2020. Accessed: Aug. 09, 2025. [Online]. Available: https://esmap.org/esmap_offshorewind_-techpotential_analysis_maps.

[3] International Energy Agency, "Net Zero Roadmap: A Global Pathway to Keep the 1.5 °C Goal in Reach," 2023. Accessed: Aug. 09, 2025. [Online]. Available: https://www.iea.-org/reports/net-zero-roadmap-a-global-pathway-to-keep-the-15-0c-goal-in-reach.

[4] World Forum Offshore Wind, "Global Offshore Wind Report ," 2023. Accessed: Aug. 09, 2025. [Online]. Available: https://wfo-global.org/wp-content/uploads/2024/04/WFO- Report-2024Q1.pdf.

[5] 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.

[6] G. L. Geerts and D. E. O'Leary, "V-Matrix: A wave theory of value creation for big data," *International Journal of Accounting Information Systems*, vol. 47, p. 100575, Dec. 2022, doi: 10.1016/j.accinf.2022.100575.

[7] M. Govett *et al.*, "Exascale Computing and Data Handling: Challenges and Opportunities for Weather and Climate Prediction," *Bull. Am. Meteorol. Soc.*, vol. 105, no. 12, pp. E2385–E2404, Dec. 2024, doi: 10.1175/BAMS-D-23-0220.1.

[8] A. Badshah, A. Daud, R. Alharbey, A. Banjar, A. Bukhari, and B. Alshemaimri, "Big data applications: overview, challenges and future," *Artif. Intell. Rev.*, vol. 57, no. 11, p. 290, Sep. 2024, doi: 10.1007/s10462-024-10938-5.

[9] F. Zhao, C. Zhang, and B. Geng, "Deep Multimodal Data Fusion," *ACM Comput. Surv.*, vol. 56, no. 9, pp. 1–36, Oct. 2024, doi: 10.1145/3649447.

[10] Y. Zekeik, M. J. OrtizBevia, F. J. Alvarez-Garcia, A. Haddi, Y. El Mourabit, and A. Alrubaye, "Improved wind power assessments by bias adjusted reanalysed data with applications near Morocco's coast," *Sci. Rep.*, vol. 14, no. 1, p. 27711, Nov. 2024, doi: 10.1038/s41598-024-77765-0.

[11] M. Shahnawaz and M. Kumar, "A Comprehensive Survey on Big Data Analytics: Characteristics, Tools and Techniques," *ACM Comput. Surv.*, vol. 57, no. 8, pp. 1–33, Aug. 2025, doi: 10.1145/3718364.

[12] R. Adams *et al.*, "Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis," *Nat. Med.*, vol. 28, no. 7, pp. 1455–1460, Jul. 2022, doi: 10.1038/s41591-022-01894-0.

[13] E. Cimadevilla, B. N. Lawrence, and A. S. Cofiño, "The Earth System Grid Federation (ESGF) Virtual Aggregation (CMIP6 v20240125)," *Geosci. Model Dev.*, vol. 18, no. 8, pp. 2461–2478, Apr. 2025, doi: 10.5194/gmd-18-2461-2025.

[14] J. L. Schnase, "Climate Analytics as a Service," in *Cloud Computing in Ocean and Atmospheric Sciences*, Elsevier, 2016, pp. 187–219. doi: 10.1016/B978-0-12-803192-6.00011-6.

[15] ORNL, "Adaptable IO System delivers data for high-performance computing science at ORNL," ORNL. Accessed: Aug. 09, 2025. [Online]. Available: https://www.ornl.gov/news/adaptable-io-system-delivers-data-high-performance - computing -science-ornl?utm_source=chatgpt.com.

[16] Z. Li, Q. Huang, G. J. Carbone, and F. Hu, "A high performance query analytical framework for supporting data-intensive climate studies," *Comput. Environ. Urban Syst.*, vol. 62, pp. 210–221, Mar. 2017, doi: 10.1016/j.compenvurbsys.20-16.12.003.

[17] J. K. Afriyie *et al.*, "A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions," *Decision Analytics Journal*, vol. 6, p. 100163, Mar. 2023, doi: 10.1016/j.dajour.2023.100163.

[18] T. Pelser, J. M. Weinand, P. Kuckertz, R. McKenna, J. Linssen, and D. Stolten, "Reviewing accuracy & reproducibility of large-scale wind resource assessments," *Advances in Applied Energy*, vol. 13, p. 100158, Feb. 2024, doi: 10.1016/j.adapen.2023.100158.

[19] R. Du *et al.*, "3DTCN-CBAM-LSTM short-term power multi-step prediction model for offshore wind power based on data space and multi-field cluster spatio-temporal correlation," *Appl. Energy*, vol. 376, p. 124169, Dec. 2024, doi: 10.1016/j.-apenergy.2024.124169.

[20] S. Hanifi, S. Lotfian, H. Zare-Behtash, and A. Cammarano, "Offshore Wind Power Forecasting—A New Hyperparameter Optimisation Algorithm for Deep Learning Models," *Energies (Basel).*, vol. 15, no. 19, p. 6919, Sep. 2022, doi: 10.3390/en15196919.

[21] J.-W. Ding, M.-J. Chuang, J.-S. Tseng, and I.-Y. L. Hsieh, "Reanalysis and Ground Station data: Advanced data preprocessing in deep learning for wind power prediction," *Appl. Energy*, vol. 375, p. 124129, Dec. 2024, doi: 10.1016/j.apenergy.2024.124129.

[22] N. N. Davis *et al.*, "The Global Wind Atlas: A High-Resolution Dataset of Climatologies and Associated Web-Based Application," *Bull. Am. Meteorol. Soc.*, vol. 104, no. 8, pp. E1507–E1525, Aug. 2023, doi: 10.1175/BAMS-D-21-0075.1.

[23] I. Staffell and S. Pfenninger, "Using bias-corrected reanalysis to simulate current and future wind power output," *Energy*, vol. 114, pp. 1224–1239, Nov. 2016, doi: 10.1016/j.energy.2016.08.068.

[24] A. Mitra, "Regime Characterization of Offshore Wind Resource Using Unsupervised Learning," Sep. 2025. doi: 10.2172/2589309.

[25] S. Yang, H. Yuan, and L. Dong, "Offshore wind resource assessment by characterizing weather regimes based on self-organizing map," *Environmental Research Letters*, vol. 17, no. 12, p. 124009, Dec. 2022, doi: 10.1088/1748-9326/aca2c2.

[26] Y. Yang, H. Lou, J. Wu, S. Zhang, and S. Gao, "A survey on wind power forecasting with machine learning approaches," *Neural Comput. Appl.*, vol. 36, no. 21, pp. 12753–12773, Jul. 2024, doi: 10.1007/s00521-024-09923-4.

[27] NetCDF 4.9.3, "Introduction and Overview-What is NetCDF?" Accessed: Aug. 10, 2025. [Online]. Available: https://docs.unidata.ucar.edu/netcdf-c/current/index.html.

[28] G. E. Loeb N., "Ask a Climate Expert: How do I use netCDF files?," ClimateData.ca. Accessed: Nov. 10, 2024. [Online]. Available: https://climatedata.ca/ask-a-climate-expert-how-do-i-use-netcdf-files/.

[29] Natural Earth, "Coastline - Free vector and raster map data at 1:10m, 1:50m, and 1:110m scales." Accessed: Aug. 09, 2025. [Online]. Available: https://www.naturalearthdata.com/downloads/10m-physical-vectors/10m-coastline/.

[30] SciPY, "CKDTree — SciPY V1.16.1 Manual." Accessed: Aug. 09, 2025. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.cKDTree.html.

[31] W. J. Leong, "The Pangeo Machine Learning Ecosystem," Xarray Blog. Accessed: Apr. 04, 2026. [Online]. Available: https://xarray.dev/blog/pangeo-ml-ecosystem-2023.

[32] C. Stern *et al.*, "Pangeo Forge: Crowdsourcing Analysis-Ready, Cloud Optimized Data Production," *Frontiers in Climate*, vol. 3, Feb. 2022, doi: 10.3389/fclim.2021.782909.

[33] Xarray, "Parallel computing with Dask." Accessed: Aug. 09, 2025. [Online]. Available: https://docs.xarray.dev/en/v2023.04.1/user-guide/dask.html#reading-and-writing-data.

[34] dask, "Best Practices — Dask documentation." Accessed: Aug. 09, 2025. [Online]. Available: https://docs.dask.org/en/latest/array-best-practices.html?.

[35] The HDF Group, "Chunking in HDF5." Accessed: Aug. 09, 2025. [Online]. Available: https://support.hdfgroup.org/documentation/hdf5-docs/advanced_topics/chunking_in_hdf5.html?.

[36] 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.

[37] Scikit-learn, "KMEANS." Accessed: Aug. 09, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

[38] Scikit-learn, "DBSCAN." Accessed: Aug. 09, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html?.

[39] Scikit-learn, "OPTICS." Accessed: Aug. 09, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.OPTICS.html?.

[40] Scikit-learn, "GaussianMixture." Accessed: Aug. 09, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html?.

[41] H. R. Soltani Motlagh, S. B. Issa-Zadeh, A. H. Kalifullah, and M. R. Zoolfakar, "Physics-Aware and Storm-Adaptive Machine Learning Clustering of Offshore Wind Regimes," *JMR (Journal of Maritime Research)*, no. submitted for publication, 2026.

[42] Scikit-learn, "Selecting the number of clusters with silhouette analysis on KMeans clustering." Accessed: Aug. 09, 2025. [Online]. Available: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html?.

[43] M. Hajihosseinlou, A. Maghsoudi, and R. Ghezelbash, "A comprehensive evaluation of OPTICS, GMM and K-means clustering methodologies for geochemical anomaly detection connected with sample catchment basins," *Geochemistry*, vol. 84, no. 2, p. 126094, May 2024, doi: 10.1016/j.chemer.2024.126094.

[44] A. A. Mas'ud, A. Sundaram, J. A. Ardila - Rey, R. Schurch, F. Muhammad-Sukki, and N. A. Bani, "Application of the Gaussian Mixture Model to Classify Stages of Electrical Tree Growth in Epoxy Resin," *Sensors*, vol. 21, no. 7, p. 2562, Apr. 2021, doi: 10.3390/s21072562.