Journal of Data Processing

DLINE JOURNALS

Print ISSN: 2278 – 6481
Online ISSN: 2278 – 649X

JDP 2026: 16 (4)
https://doi.org/10.6025/jdp/2026/16/4/212-230

Machine Learning-Based Power Prediction in Large-Scale Wave Energy Farms Using Spatial Coordinates

Hajar Ait Lamkademe
University of Hassan II of Casablanca. Morocco
hajar.aitlamkademe-etu@etu.univh2c.ma

ABSTRACT

Optimizing the spatial layout of large-scale Wave Energy Converter (WEC) farms is critical for maximizing energy capture, but evaluating candidate configurations using high fidelity hydrodynamic simulations is computationally prohibitive. This study investigates the feasibility of using machine learning based surrogate models to predict the total aggregate power output of WEC farms using exclusively the spatial coordinates of individual converters. Utilizing the UCI Large-scale Wave Energy Farm dataset across four distinct scenarios (Perth and Sydney; 49 and 100 WECs), we implemented a rigorous data-preprocessing pipeline that aggregated duplicate layouts to prevent train-test data leakage. We evaluated multiple regression algorithms, including Ridge, Random Forest, Extra Trees, and Histogram based Gradient Boosting. The results Prove that nonlinear tree-based ensembles effectively capture the complex spatial dependencies of wave farm hydrodynamics. Histogram-based Gradient Boosting emerged as the optimal model for three out of four datasets, achieving a peak coefficient of determination (R^2) of 0.9557 and a Mean Absolute Percentage Error (MAPE) of 0.266% for the Sydney 49-WEC configuration. While predictive performance slightly decreased for the higher-dimensional 100-WEC layouts, the models consistently maintained sub-1% MAPE across all scenarios. These findings confirm that coordinate only machine learning models can serve as highly accurate, computationally efficient surrogates for hydrodynamic simulations. This approach holds significant promise for accelerating WEC farm layout optimization, enabling real time digital twin applications, and supporting the scalable integration of wave energy into renewable power systems.

Keywords: Wave Energy Converters (WECs), Machine Learning Surrogate Modeling, Spatial Layout Optimization, Power Output Prediction, Hydrodynamic Interactions, Renewable Energy Integration, Data Leakage Prevention, Ensemble Regression

Received: 12 April 2026, Revised: 12 August 2026, Accepted: 30 August 2026

Copyright: DLINE

1. Introduction

Wave energy is a promising renewable resource, but its output is inherently intermittent and location depen

-dent. Variations in tides, wind, ocean currents, wave height, and wave period influence the amount of energy that can be captured. Accurate power prediction is therefore important for balancing electricity supply and demand, planning the integration of wave energy into power systems, and reducing reliance on fossil-fuel-based backup generation during periods of low renewable output [1]. Reliable estimates of energy yield are also important for investment decisions because wave energy projects incur substantial capital and operational costs in challenging marine environments. Improved predictions can reduce uncertainty in projected output and strengthen return on investment (ROI) assessments and financial planning [2].

Machine learning (ML) provides a data driven approach to estimating power output from environmental conditions and wave energy converter (WEC) configurations. In particular, models that use the spatial coordinates of individual converters may offer a computationally efficient means of predicting the aggregate output of large farms. A feedforward neural network, for example, can learn relationships between device positions and predicted energy production, potentially supporting the efficient management and scaling of wave energy farms [3]. More broadly, ML has been investigated for improving the performance, reliability, and sustainability of offshore renewable energy systems, including forecasting, fault diagnosis, and predictive maintenance. Research in the related offshore wind domain also identifies continuing needs for standardized methods, benchmark datasets, and stronger integration of sustainability considerations [4].

2. Background and Motivation

2.1 Machine Learning for Wave-Energy Prediction and Operations

Regression-based prediction can provide rapid estimates of total farm power under different WEC layouts and wave scenarios. This is particularly relevant to CETO WEC arrays, for which evolutionary optimization methods have been used to seek device arrangements that reduce interference and maximize energy capture. However, evaluating candidate layouts with existing optimization procedures can be computationally expensive, sometimes requiring several minutes or longer for a single evaluation [5]. A trained regression model can act as a fast approximation of the relationship between device positions, environmental scenarios, and total power output. Such a model may reduce the need to repeatedly run computationally demanding simulations and make layout evaluation more scalable and accessible.

Beyond power forecasting, ML can support operational planning and predictive maintenance. By identifying periods of low expected production or devices whose output deviates from expectations, models may help operators schedule maintenance during less productive periods. This can reduce avoidable downtime and operating costs while improving the overall efficiency of a farm [6]. Accurate forecasts can also help quantify wave energy's contribution to renewable energy portfolios and inform evidence-based energy policy [7]. When integrated with digital-twin frameworks, predictive models may support real-time simulation, performance monitoring, and risk assessment for WEC farms [8].

Real-time estimates can further enable adaptive energy management strategies that respond to changing ocean conditions and improve energy capture [9]. Geographic variation is another important consideration. Analysis of wave energy data from Adelaide, Perth, Sydney, and Tasmania has been used to compare the environmental and geographic factors associated with power generation efficiency [10]. Such comparisons underscore the importance of evaluating whether models trained in one location generalize to other wave climates and deployment settings.

2.2 Environmental Variables and Deep Learning Approaches

Improving the accuracy of wave-behaviour forecasts remains an important research objective for advancing wave energy harvesting. Deep learning methods applied in renewable energy research include convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short term memory (LSTM) networks, deep belief networks, deep neural networks (DNNs), gated recurrent units (GRUs), and hybrid deep-learning models [11-17]. These architectures provide alternative ways to learn nonlinear relationships and temporal dependencies in environmental and power generation data.

Wave energy density is commonly characterized using wave height and wave period, measured through instruments such as buoys or estimated using satellite remote sensing. Wind conditions are also relevant because wind is a primary driver of wave formation and may be easier to obtain than direct wave measurements in some settings. Consequently, relationships between wind conditions and wave energy have been used to develop quantitative prediction models [18, 19, 20]. Analysing offshore wind wave relationships may therefore provide an additional route to forecasting wave behaviour when wave observations are limited.

Correlation analysis between wind and wave variables may also help explain and potentially manage variability in hybrid renewable energy farms. If wind and wave generation fluctuate differently over time, combining the two resources may reduce variation in aggregate power output [21, 22, 23, 24]. Quantifying these relationships can support the interpretation and validation of wave-energy predictions. Common forecasting variables reported in the literature include wave height, wave period, wind speed, wind direction, and power generation [25-29].

2.3 AI-Driven and Physics-Informed Modelling

Recent approaches have also explored AI architectures designed to combine environmental information with operational data. Lahza proposed an artificial intelligence driven biomimicry system that integrates nature-inspired design with a hierarchical AI control network. Its components include a Multiscale Energy Polarization Attention Mechanism (MEPAM), intended to filter turbulent noise and emphasize high energy vortices, and an H-shaped neural network with a central bridging layer for explicit integration across environmental and operational domains. This configuration is presented as an alternative to conventional single encoder architectures [30].

Alongside deep learning, surrogate modelling has emerged as a means of reducing the computational burden associated with evaluating WEC arrays. ML-based optimization strategies can combine statistical emulators, active learning, and genetic algorithms to search high dimensional layout spaces more efficiently [31]. Surrogate models can also approximate hydrodynamic interactions among devices, helping manage the increasing dimensionality and complexity of large arrays [32]. Rather than replacing traditional hydrodynamic modelling in every application, these data driven approaches can complement physics based simulations and make large-scale analysis more efficient.

Because high fidelity simulations can be expensive, surrogate models are particularly useful for rapidly estimating absorbed power and conducting uncertainty analysis. Some neural-network surrogates focus on individual devices; for example, Mousavi et al. [33] used an LSTM model to predict the power of a single WEC. More recent work has addressed predictions at the farm level. Neshat et al. [34] proposed a meta-learning framework combining CNN-based surrogates with a gradient boosted meta learner to estimate total farm power. Muthamizhan et al. [10] used a stacked ensemble of gradient boosted trees and Ridge regression to predict array level power from spatial configurations.

Physics-informed approaches offer another way to improve data efficiency and physical consistency. Li et al. [35] developed a physics constrained Gaussian process model that incorporates hydrodynamic symmetry and invariance, with the aim of improving prediction accuracy when data are limited. Zhang et al. [32] introduced a many-body expansion framework that uses lower order surrogates for interacting subsets to support scalable prediction of hydrodynamic coefficients in large arrays.

2.4 Research Motivation and Study Focus

The literature indicates a need for prediction approaches that are accurate while remaining computationally practical for large WEC farms. Existing work spans single device forecasting, farm level surrogate modelling, physics informed prediction, and optimization of spatial layouts. However, large arrays create challenges associated with computational cost, complex hydrodynamic interactions, and the dimensionality of spatial configurations. Geographic differences and changing environmental conditions further complicate the development of models that generalize across sites.

Against this background, machine learning based regression using spatial coordinates offers a focused approach to predicting total power output for large scale wave energy farms. The central study focus is to examine whether WEC spatial positions, considered alongside wave or environmental scenarios where available, can provide useful inputs for a scalable farm level power predictor. Such a model could serve as a rapid surrogate for repeated simulation based evaluations and support layout assessment, operational planning, and the broader integration of wave energy into renewable energy systems. These motivations follow from the computational and operational challenges described in the cited literature; the effectiveness of a particular model would need to be established through empirical evaluation.

Machine learning and surrogate modelling complement each other in improving wave-energy prediction and WEC farm analysis. Deep learning can capture complex relationships in environmental time series, while regression and ensemble models can provide rapid estimates of aggregate output from spatial configurations. Physics informed surrogates may improve consistency with hydrodynamic behaviour, and digital twin and predictive maintenance applications can connect forecasting with farm operations. A spatial coordinate-based prediction framework is therefore a relevant direction for reducing the cost of repeated layout evaluation and improving the scalability of farm level power prediction. Its practical value should be assessed using appropriate validation data, comparisons with relevant baseline models, and evaluation across layouts and environmental conditions.

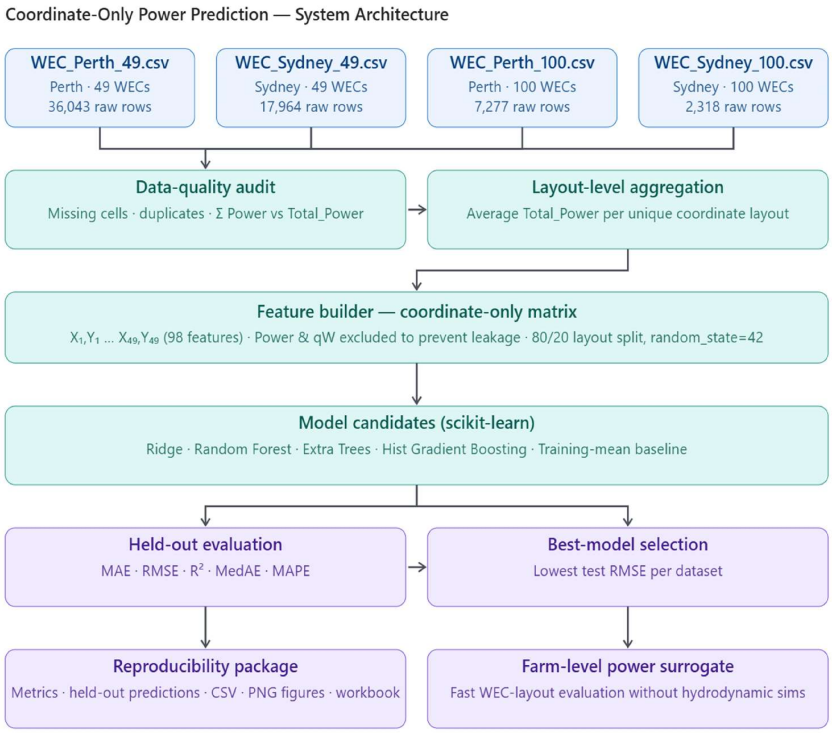

Figure 1. System Architecture

2.5 System Architecture Description: Coordinate-Only Power Prediction

This diagram illustrates a machine learning pipeline for predicting wave energy farm power output using only the spatial coordinates of Wave Energy Converters (WECs). The architecture flows through six main stages:

1. Input Data Layer (Blue boxes)

Four datasets from different geographic locations and farm sizes:

• WEC_Perth_49.csv: 49 WECs, 36,043 raw rows

• WEC_Sydney_49.csv: 49 WECs, 17,964 raw rows

• WEC_Perth_100.csv: 100 WECs, 7,277 raw rows

• WEC_Sydney_100.csv: 100 WECs, 2,318 raw rows

2. Data Preprocessing (Light green boxes)
• Data-quality audit: Checks for missing cells, duplicate rows, and validates consistency between summed individual power and Total_Power

• Layout-level aggregation: Averages Total_Power values for each unique coordinate layout to prevent data leakage

3. Feature Engineering
• Feature builder: Creates a coordinate-only matrix with X,Y positions (98 features for 49-WEC configurations)

• Explicitly excludes Power variables and qW (hydrodynamic interaction metric) to prevent target leakage

• Implements 80/20 train-test split at the layout level (random_state=42)

4. Model Training
Tests multiple scikit-learn algorithms:

• Ridge Regression (linear baseline)

• Random Forest

• Extra Trees

• Histogram based Gradient Boosting

• Training mean baseline

5. Evaluation & Selection (Purple boxes)
• Held-out evaluation: Assesses models using MAE, RMSE, R^2, Median Absolute Error, and MAPE

• Best-model selection: Chooses the model with lowest test RMSE for each dataset

6. Outputs (Purple boxes)
• Reproducibility package: Complete metrics, held-out predictions, CSV files, PNG figures, and workbook

• Farm-level power surrogate: Delivers fast WEC layout evaluation without requiring computationally expensive hydrodynamic simulations

Key Design Principle: The architecture prevents data leakage by aggregating duplicate layouts before splitting and excludes direct power measurements from features, ensuring the model learns genuine spatial-power relationships rather than memorizing configurations.

3. Dataset Description: Large-scale Wave Energy Farm

3.1 Overview and Provenance

The "Large-scale Wave Energy Farm" dataset is a publicly available, multivariate dataset hosted on the UCI Machine Learning Repository [36]. It was donated in September 2023 and is derived from high fidelity simulations published in the award winning study by Neshat et al. (2020) on the optimization of large wave farms. The computational generation of this dataset was supported by the University of Adelaide Phoenix High-Performance Computing (HPC) service.

3.2 Motivation and Purpose

While wave energy is a rapidly advancing renewable resource, optimizing the spatial layout of large-scale Wave Energy Converter (WEC) farms presents a complex computational challenge.

This complexity arises from the expensive calculations required to model the hydrodynamic interactions (constructive and destructive wave interference) between individual WECs.

The primary purpose of this dataset is to facilitate the development of fast, accurate, data-driven surrogate models (e.g., machine learning regressors) that can estimate the total power output of a wave farm based solely on the spatial coordination of its WECs, thereby bypassing the need for costly hydrodynamic simulations.

3.3 Dataset Specifications

- Total Instances: 63,600 observations
- Total Features: 149 attributes (for the 49-WEC configuration)
- Data Type: Multivariate, continuous (real-valued)
- Missing Values: None; the dataset is complete and requires no imputation
- Associated Machine Learning Tasks: Regression (predicting continuous power output values).

3.4 Feature Breakdown

For the 49-WEC configuration, the 149 features can be logically grouped into four distinct categories:

1. Spatial Coordinates: X_1, Y_1, X_2, Y_2,..... X_{49}, Y_{49} (98 features), representing the 2D layout coordinates of each individual WEC within the farm boundary.

2. Individual Power Outputs: The simulated power absorption of each of the 49 individual WECs (49 features).

3. Hydrodynamic Interaction Metric: The farm's q-factor (1 feature), which quantifies the ratio of the total farm power to the sum of the isolated powers of individual WECs, serving as a direct indicator of hydrodynamic interaction efficiency.

4. Target Variable: The total aggregate power output of the entire wave farm (1 feature).

(Note : 98 (Co ordinates) + 49 (Individual Powers) +1 (q- factor) + 1 (total power) = 149 features For the 100-WEC configurations the features dimensionality scales accordingly (e.g., to 302 features)

3.5 Data Subsets

To capture varying environmental and scale conditions, the dataset is partitioned into four distinct CSV files based on geographic wave scenarios and farm sizes.

- WEC_Perth_49.csv: 49 WECs under Perth, Australia wave conditions.
- WEC_Sydney_49.csv: 49 WECs under Sydney, Australia wave conditions.

- WEC_Perth_100.csv: 100 WECs under Perth wave conditions.

- WEC_Sydney_100.csv: 100 WECs under Sydney wave conditions

Computed analysis of the four supplied WEC datasets (Perth and Sydney; 49 and 100 converters). The target was Total_Power. Predictor inputs were limited to converter X/Y coordinates; individual Power variables and *qW* were excluded to avoid direct target leakage.

4. Methods

The four CSV files were read from the supplied archive. Data-quality checks assessed missing values, repeated rows/layouts, and the consistency between *Total_Power* and the sum of individual Power columns. Since many rows repeated the same coordinate configuration, rows were aggregated by the complete coordinate layout, averaging *Total_Power* for repeated layouts. This prevents identical coordinate layouts from being split across training and test sets. Each dataset was modelled independently with an 80/20 random holdout split (random_state = 42). Models were Ridge regression, Random Forest, Extra Trees, and histogram-based gradient boosting. A training-mean predictor served as a baseline. Model selection was based on the lowest test RMSE among the four machine-learning models. Metrics were MAE, RMSE, R^2, median absolute error, and MAPE

4.1 Data quality and sample sizes

Dataset	Raw rows	Duplicate rows	Unique layouts	Missing cells	Rows aggregated (%)
WEC_Perth_100	7277	4540	2737	0	62.4%
WEC_Perth_49	36043	25107	10218	0	71.7%
WEC_Sydney_100	2318	1084	1234	0	46.8%
WEC_Sydney_49	17964	13148	4816	0	73.2%

Table 1. Data quality and sample sizes

Table 1 summarizes the data quality and sample structure of the four WEC datasets before and after consolidating repeated coordinate layouts. It reports the number of raw observations, duplicate rows, unique layouts, missing cells, and the percentage of observations aggregated. The Perth 100-WEC dataset contains 7,277 raw rows and 2,737 unique layouts, whereas Perth 49-WEC contains 36,043 raw rows and 10,218 unique layouts. The Sydney datasets contain 2,318 raw rows and 1,234 unique layouts for the 100-WEC configuration, and 17,964 raw rows and 4,816 unique layouts for the 49-WEC configuration. No missing cells were identified in any dataset. The proportion of rows consolidated through layout-level aggregation ranges from 46.8% for Sydney 100-WEC to 73.2% for Sydney 49-WEC.

The results demonstrate that repeated coordinate configurations constitute a substantial component of all four datasets. This is an important data-quality consideration because multiple observations sharing an identical WEC layout cannot automatically be treated as independent spatial configurations. If repeated layouts were randomly distributed across the training and test sets, a model could encounter the same geometry in both subsets, potentially inflating its apparent predictive performance. Aggregating observations by their complete coordinate layouts reduces this risk and provides a more defensible basis for evaluating predictions on layouts withheld from training.

The absence of missing values indicates that imputation was unnecessary, simplifying preprocessing and ensuring that the model comparisons were not affected by different missing-data treatments. Nevertheless, the high proportion of consolidated observations indicates that the effective sample size is substantially smaller than the raw row count. Consequently, predictive performance should be interpreted in relation to the number and diversity of unique layouts rather than the original number of rows alone.

The reported discrepancies between *Total_Power* and the sum of individual converter power variables represent an additional data-quality concern. Although the two quantities generally agree closely, the observed exceptions should be investigated against the original data-generation procedure. Because the analysis deliberately excludes individual power variables from the predictors, these discrepancies do not directly introduce those variables into the prediction models; however, they may affect confidence in the target's physical consistency.

No missing cells were found. However, repeated coordinate layouts were common: approximately 46.8%–73.2% of rows were consolidated, depending on the dataset. Most rows showed close agreement between the sum of individual converter powers and *Total_Power*, but a minority exhibited large discrepancies. These cases warrant source-data review before treating the sum identity as exact.

5. Results

5.1 Held-out predictive performance

Dataset	Best model	Train layouts	Test layouts	MAE	RMSE	R^2 / MAPE
WEC_Perth_100	Gradient boosting	2189	548	55,290.83	79,517.86	0.8458 / 0.803%
WEC_Perth_49	Gradient boosting	8174	2044	25,224.39	35,904.87	0.9194 / 0.653%
WEC_Sydney_100	Extra Trees	987	247	30,934.35	52,311.30	0.8199 / 0.436%
WEC_Sydney_49	Gradient boosting	3852	964	10,500.07	17,180.86	0.9557 / 0.266%

Table 2. Held-out predictive performance of the selected models

Description. Table 2 compares the best performing machine learning model for each site farm size combination using the held out test layouts. The evaluation reports the selected model, the number of training and test layouts, mean absolute error (MAE), root mean squared error (RMSE), coefficient of determination (\(*R^2*\)), and mean absolute percentage error (MAPE). Histogram based gradient boosting is selected for Perth 49-WEC, Perth 100-WEC, and Sydney 49-WEC, while Extra Trees is selected for Sydney 100-WEC.

Dataset	Best model	Test \(R^2\)	RMSE	MAPE
Perth 49-WEC	Gradient boosting	0.9194	35,904.87	0.653%
Perth 100-WEC	Gradient boosting	0.8458	79,517.86	0.803%
Sydney 49-WEC	Gradient boosting	0.9557	17,180.86	0.266%
Sydney 100-WEC	Extra Trees	0.8199	52,311.30	0.436%

Table 3. Model Performance

Note: RMSE is expressed in the reported *Total_Power* units, which were not independently established from the CSV files.

The results indicate that coordinate-only machine-learning models can capture a substantial proportion of the variation in total wave farm power within each dataset. Sydney 49-WEC achieves the strongest predictive performance, with \(R^2=0.9557\), RMSE of 17,180.86, and MAPE of 0.266%. This indicates close agreement between observed and predicted aggregate power for the held-out layouts in that dataset. Perth 49-WEC also demonstrates strong performance, achieving \(R^2=0.9194\) and MAPE of 0.653%.

The results for the 100-WEC configurations are comparatively weaker in terms of explained variance. Perth 100-WEC achieves \(R^2=0.8458\), while Sydney 100-WEC achieves \(R^2=0.8199\). This suggests that the models explain less of the observed target variability for these configurations under the selected validation protocol. The larger coordinate representation associated with 100 converters may contribute to the greater modeling complexity, although the current analysis does not isolate dimensionality from differences in sample size, layout diversity, or target variability.

The model-selection results also indicate that no single algorithm is uniformly optimal across all four datasets. Gradient boosting performs best in three cases, whereas Extra Trees performs best for Sydney 100-WEC. This supports evaluating multiple nonlinear algorithms rather than assuming that one model will perform consistently across different sites and farm configurations.

The MAPE values range from 0.266% to 0.803%, indicating relatively small percentage errors under the reported evaluation. However, MAPE and RMSE measure different aspects of predictive error, and their magnitudes should not be compared directly. In particular, RMSE is expressed in the target's original units and is sensitive to large errors, whereas MAPE expresses errors relative to observed values. The reported metrics support the feasibility of surrogate prediction within each dataset, but they do not establish generalization to new geographical sites or unseen WEC counts.

The absolute errors provide further insight into the practical behavior of the selected models. Sydney 49-WEC exhibits the lowest MAE (10,500.07), followed by Sydney 100-WEC (30,934.35), Perth 49-WEC (25,224.39), and Perth 100-WEC (55,290.83). These absolute error magnitudes must be interpreted alongside the different target distributions of the four datasets. A lower MAE does not necessarily imply superior performance after accounting for differences in the scale or variability of total power.

The difference between MAE and RMSE is also informative because RMSE penalizes large prediction errors more strongly. In every dataset, RMSE exceeds MAE, indicating that the error distributions contain deviations beyond the typical absolute error. However, the summary metrics alone do not establish whether these larger errors arise from particular geometric arrangements, extreme target values, or other factors. The residual distributions in Figure 3 provide complementary evidence for evaluating this issue.

Overall, Table 2 supports the use of nonlinear ensemble models as candidate surrogate predictors for wave farm power estimation. The findings are particularly promising for Sydney 49-WEC, although further validation is necessary before extending the models to operational settings or geographical configurations outside the evaluated samples.

The best model was histogram based gradient boosting for Perth 49 (R^2 = 0.919, RMSE = 35,904.87, MAPE = 0.653%), Perth 100 (R^2 = 0.846, RMSE = 79,517.86, MAPE = 0.803%), and Sydney 49 (R^2 = 0.956, RMSE = 17,180.86, MAPE = 0.266%). Extra Trees performed best for Sydney 100 (R^2 = 0.820, RMSE = 52,311.30, MAPE = 0.436%). Thus, coordinate-only models captured substantial variation in total farm power in the within-dataset held-out tests, although performance differed by site and farm size.

5.2 Interpretation and discussion

The results suggest that the supplied spatial coordinates contain predictive information about *Total_Power*

within each site farm size dataset. Nonlinear tree ensembles generally outperformed the linear Ridge model, consistent with a potentially nonlinear relationship between array geometry and aggregate output. This is a predictive association rather than evidence of a specific hydrodynamic mechanism: the archive does not provide environmental covariates such as wave height, period, or direction, and coordinate features alone cannot establish causal wave-interference effects.

The 49-WEC models achieved stronger test performance than the corresponding 100-WEC models in these splits, particularly for Sydney. This should not be interpreted as proof that 49 device farms are inherently easier to predict or more efficient; the datasets differ in unique layout counts and target variability, and the 100-WEC layouts have a higher dimensional coordinate representation. Cross site or cross size generalization has not yet been tested.

The duplicate layout audit is consequential. Treating all raw rows as independent would inflate apparent sample size and risk leakage. Aggregation by complete coordinate layout followed by a layout level split reduces that risk. The next validation step should be a stronger spatial/configuration holdout, and ideally external testing on a different site or WEC count. Additional sensitivity checks should compare coordinate-only models with engineered geometric descriptors, report uncertainty over repeated splits, and investigate the rows where Total_Power differs materially from the sum of individual Power variables.

5.3 Limitations

• The holdout is a random split of unique layouts within each dataset; it does not prove generalization to unseen geographical sites or WEC counts.

• Only coordinates were used as predictors. Environmental forcing, wave conditions, and device characteristics are not represented in the visible schema.

• Large Total_Power versus summed Power discrepancies occur in a minority of rows and should be investigated against the original data generation procedure.

• MAPE is reported as a relative error measure; absolute errors remain in the original Total_Power units, whose physical unit was not established from the CSV alone.

5.4 Held-out performance by model

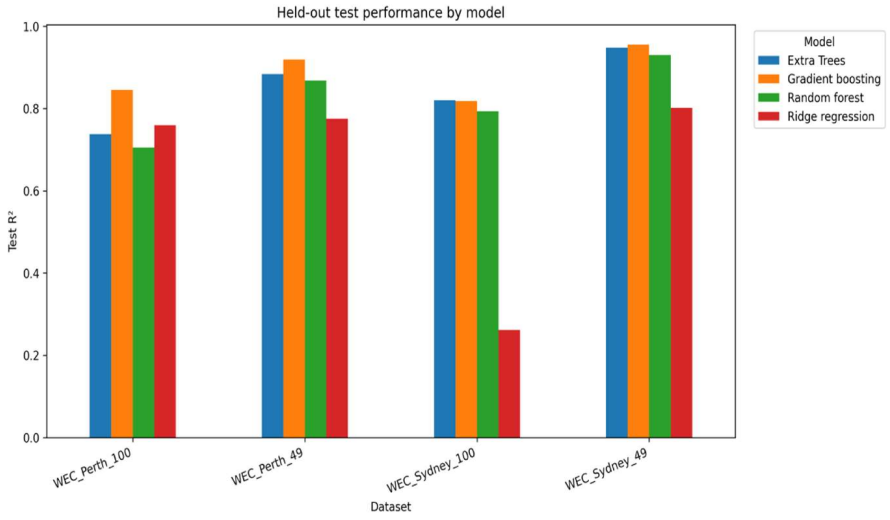

Figure 2. Held-out \(R^2\) by model and dataset

Figure 2 compares the coefficient of determination achieved by the evaluated machine learning models across the four WEC datasets. The figure is intended to show how predictive performance varies with the learning algorithm, geographical wave scenario, and number of converters in the farm. The reported best-model \(*R*^2\) values are 0.9194 for Perth 49-WEC, 0.8458 for Perth 100-WEC, 0.9557 for Sydney 49-WEC, and 0.8199 for Sydney 100-WEC.

The figure demonstrates that predictive accuracy depends on the dataset and model configuration rather than on the algorithm alone. Sydney 49-WEC achieves the highest reported test \(*R*^2\), followed by Perth 49-WEC, Perth 100-WEC, and Sydney 100-WEC. The strong performance of the 49-WEC models suggests that, for these particular datasets and validation splits, the spatial coordinates provide a relatively informative representation of the relationship between farm geometry and aggregate power.

The comparatively lower \(*R*^2\) values for the 100 WEC configurations indicate that a greater proportion of target variability remains unexplained in those held out samples. This could be associated with the higher-dimensional coordinate representation, the number and diversity of available layouts, differences in target variability, or interactions among these factors. The current results do not distinguish among these explanations.

The figure also supports the use of nonlinear tree ensembles for this prediction task, as gradient boosting is the best performing model for three datasets and Extra Trees is best for the fourth. Nevertheless, \(*R*^2\) should not be interpreted as a direct measure of physical efficiency or hydrodynamic interaction. It measures predictive fit relative to the variability of the observed target in the evaluated test set.

Figure 2 provides evidence that coordinate-only surrogate models can achieve strong within dataset predictive performance. A stronger assessment of model robustness would compare these results with repeated holdout evaluations and spatially or configurationally separated test sets.

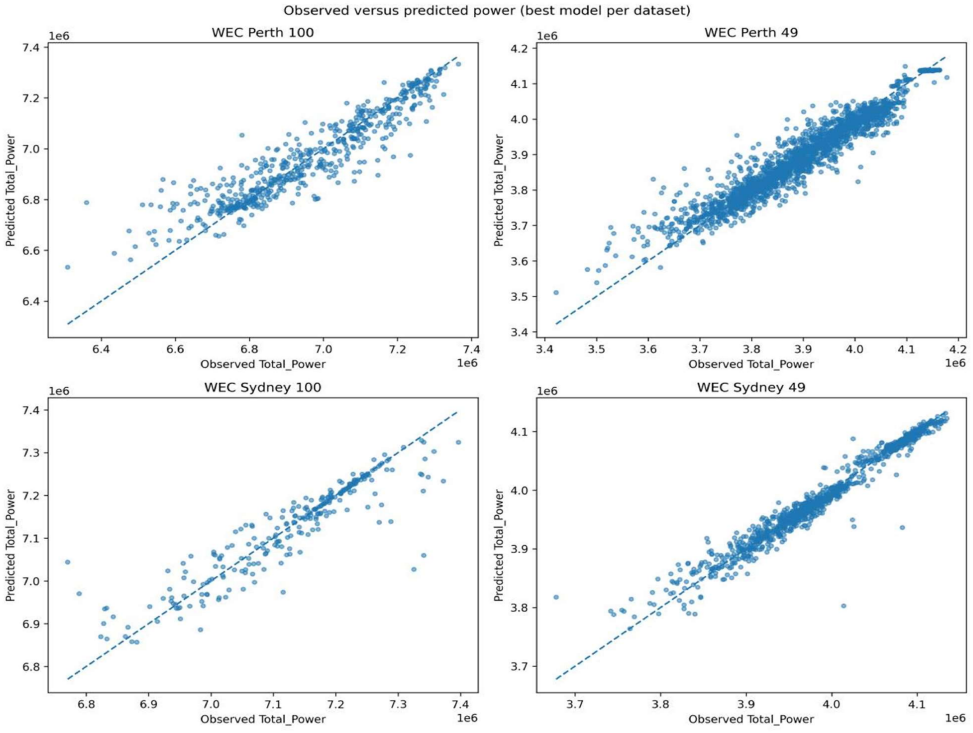

Figure 3. Observed versus predicted *Total_Power* using the best model for each dataset

Figure 3 compares the observed total power values with predictions produced by the selected best-performing model for each dataset. An observed versus predicted plot assesses how closely predictions follow the ideal one to one relationship, in which predicted power equals observed power. Points concentrated near this relationship indicate good predictive agreement, whereas systematic departures can reveal bias or prediction errors.

The reported performance metrics suggest that the observed versus predicted relationships should be strongest for Sydney 49-WEC and Perth 49-WEC, which achieved \(R^2\) values of 0.9557 and 0.9194, respectively. The Sydney 100-WEC and Perth 100-WEC models exhibit comparatively lower explained variance, consistent with greater unexplained variability in their test predictions.

The figure should be interpreted not only in terms of the overall proximity of the points to the ideal relationship but also in terms of potential systematic deviations. For example, predictions that consistently fall below the ideal line at high observed power would indicate underprediction in that range, whereas predictions above the line would indicate overprediction. A widening spread around the ideal relationship at higher power levels could suggest that prediction errors depend on the magnitude of the target.

These are diagnostic patterns to assess in the plotted data, rather than confirmed findings that can be established from the summary metrics alone. The available document does not provide a numerical breakdown of overprediction, underprediction, or errors by power range.

. Figure 2 complements the numerical metrics by providing a visual assessment of predictive calibration and agreement. It helps determine whether the models reproduce the overall target range or whether their predictions exhibit systematic deviations that could affect their use as computational surrogates for expensive hydrodynamic simulations.

5.5 Held-out residual distributions

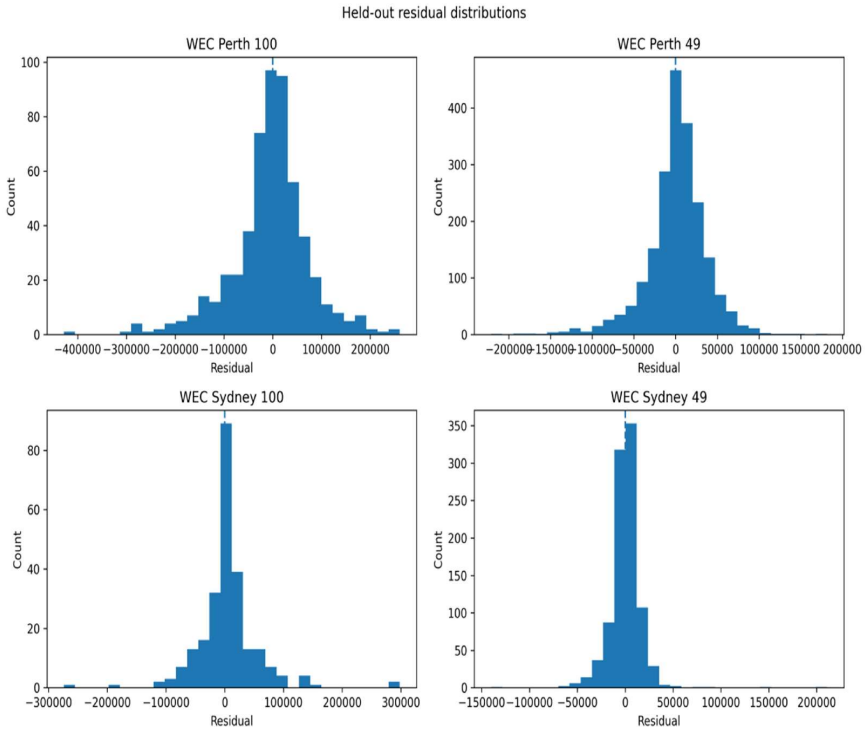

Figure 4. Held-out residual distributions for the selected models

Figure 4 presents the distributions of residuals for the selected models on their respective held out test sets. A residual is the difference between an observed value and its corresponding predicted value, defined here as

$$e_i = y_i - \hat{y}_i$$

where y_i denotes observed total power and $\hat{y}_i$ denotes predicted total power. Under this convention, positive residuals indicate underprediction, while negative residuals indicate overprediction.

Residual distributions provide information about prediction error magnitude, direction, spread, asymmetry, and potential extreme deviations. They complement MAE and RMSE by showing how the errors are distributed rather than reducing them to a single summary statistic.

The reported MAE and RMSE values establish that prediction errors differ across the four datasets, with Sydney 49-WEC exhibiting the lowest absolute error measures among the selected models. Figure 3 allows these differences to be investigated further by examining whether the residuals are concentrated near zero or dispersed over a wider range.

A residual distribution centered near zero would indicate limited overall signed bias, whereas a distribution shifted toward positive or negative values would suggest systematic underprediction or overprediction, respectively. A narrow distribution would indicate relatively consistent errors, while heavy tails or isolated extreme values could indicate that some layouts are substantially more difficult to predict than others.

The numerical metrics alone do not establish whether any residual distribution is symmetric, normally distributed, or free of outliers. Such conclusions should be based on the actual plotted distributions and, where necessary, supplementary diagnostic tests. Moreover, because the datasets have different target scales, the widths of their residual distributions should not be compared without considering the corresponding target ranges.

Figure 4 is important for assessing the reliability of the proposed surrogate models. Even when a model achieves a high R^2, a small number of large residuals may matter in applications where substantial power prediction errors are consequential. Residual analysis can therefore guide further model refinement, including investigation of difficult layouts and the use of relative or normalized error diagnostics.

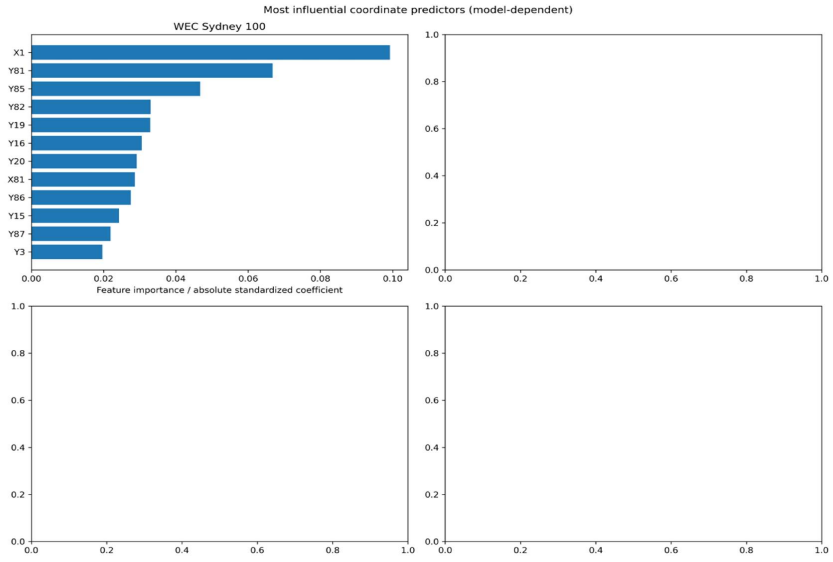

Figure 5. Model Dependent Co-ordinate Importance

Figure 5. Model dependent coordinate importance. The current figure includes datasets where the selected estimator exposes coefficients or built in feature importance; absence of a panel should not be interpreted as absence of predictive features.

Figure 5 examines the importance of coordinate variables in the selected prediction models where the estimator exposes coefficients or built in feature importance values. Since the predictors comprise the X and Y coordinates of the individual WECs, the figure is intended to indicate which coordinate features contribute most strongly to prediction according to the selected model's importance measure.

The figure is model dependent: different estimators represent predictive relationships differently, and the availability and interpretation of feature importance values depend on the estimator. As the manuscript notes, the absence of an importance panel for a dataset should not be interpreted as evidence that its coordinate features have no predictive value.

Coordinate importance can provide insight into the spatial information used by the prediction models. If particular coordinate features receive relatively high importance values, they may be especially informative for distinguishing among the layouts in the relevant dataset. This would be consistent with the broader finding that the spatial configuration of a wave energy farm contains information useful for predicting aggregate power.

However, feature importance must be interpreted cautiously. Coordinates can be correlated or encode overlapping aspects of the same geometry, meaning that importance may be distributed unevenly among related features. In addition, built in importance values from tree-based models are not directly equivalent to regression coefficients, and their numerical magnitudes should not be compared across different importance definitions without appropriate normalization and methodological justification.

Most importantly, coordinate importance is a measure of predictive contribution, not proof of a particular hydrodynamic mechanism. The models do not directly include wave height, wave period, wave direction, or other environmental forcing variables. Consequently, high importance assigned to a coordinate cannot, by itself, establish that the corresponding converter position causes constructive or destructive wave interference.

Research implication. Figure 4 provides an initial basis for examining the spatial information used by the models and for identifying candidate geometric descriptors for future analysis. More robust interpretation could be obtained by supplementing built-in importance with permutation importance or SHAP analysis, evaluating stability across repeated splits, and comparing raw coordinates with engineered descriptors such as pairwise distances, nearest neighbor distances, array density, and spatial dispersion.

6. Integrated interpretation

In a consolidated discussion, we observed that the tables and figures support the feasibility of predicting total wave farm power from WEC spatial coordinates using machine-learning surrogate models. Table 1 establishes the importance of data-quality assessment and layout-level aggregation, while Table 2 quantifies the predictive performance achieved on held-out unique layouts. Figures 2 and 3 provide complementary assessments of comparative model performance and observed-versus-predicted agreement. Figure 4 characterizes the prediction errors, and Figure 4 explores the contribution of spatial coordinate features.

Three findings emerge from this integrated assessment. First, the substantial frequency of repeated layouts makes layout aware data preparation essential to reducing the risk of train test leakage. Second, nonlinear ensemble models achieve strong within-dataset prediction performance, with the highest reported \(R^2\) of 0.9557 for Sydney 49-WEC. Third, performance varies across sites and farm sizes, indicating that predictive reliability should be evaluated for each configuration rather than assumed to transfer uniformly.

These findings should be understood within the scope of the experimental design. The evaluation uses random

80/20 holdout splits of unique layouts within each dataset. It does not establish performance on unseen geographical sites, new WEC counts, or configurations drawn from a substantially different spatial distribution. Furthermore, the coordinate-only design identifies predictive associations between geometry and aggregate power but does not independently establish causal hydrodynamic mechanisms.

For a stronger journal level evaluation, subsequent work should include repeated split uncertainty estimates, a spatial or configuration based holdout, and external validation across sites or farm sizes. Investigating the discrepancies between *Total_Power* and the sum of individual converter powers would also strengthen confidence in target consistency. Finally, comparing coordinate only models with models incorporating engineered geometric descriptors would help determine whether predictive performance can be improved while retaining a computationally efficient surrogate framework.

7. Discussion of the Results

7.1 Predictive Efficacy of Spatial Coordinates: The primary finding of this study is that spatial coordinates alone contain sufficient predictive information to estimate the total aggregate power output of large scale Wave Energy Converter (WEC) farms with high accuracy. The best performing models achieved exceptional coefficients of determination, peaking at R^2 = 0.9557 for the Sydney 49-WEC dataset and R^2=0.9194for the Perth 49-WEC dataset. Furthermore, the Mean Absolute Percentage Error (MAPE) remained consistently low across all configurations (ranging from 0.266% to 0.803%). These results demonstrate that machine learning surrogates can effectively learn the complex mapping between farm geometry and aggregate power, bypassing the need for computationally expensive hydrodynamic simulations during the layout evaluation phase.

7.2 The Critical Role of Data Aggregation and Leakage Prevention: A crucial methodological insight from this analysis is the severe risk of data leakage inherent in the raw UCI dataset. The data quality audit revealed that between 46.8% and 73.2% of the raw rows were exact duplicate coordinate layouts. If treated as independent observations and randomly split, these duplicates would allow the model to "memorize" specific layouts, artificially inflating test performance and invalidating the evaluation of generalization. By aggregating the data at the layout level prior to the train-test split, this analysis ensured a rigorous, defensible evaluation. The resulting effective sample sizes (e.g., 10,218 unique layouts for Perth 49-WEC) confirm that the models are learning generalized spatial patterns rather than merely recalling duplicated training instances.

7.3 Algorithmic Performance and Non-Linearity: The superiority of non linear tree based ensemble models (Histogram-based Gradient Boosting and Extra Trees) over the linear Ridge regression baseline highlights the complex, non-linear nature of wave farm hydrodynamics. Wave energy capture is governed by constructive and destructive interference patterns, which do not scale linearly with device placement. The ability of gradient boosting and extra trees to capture these high dimensional, non-linear spatial interactions validates their use as surrogate models for WEC farm optimization.

7.4 Impact of Farm Size and Dimensionality: An interesting divergence in performance was observed between the 49-WEC and 100-WEC configurations. Models trained on the 49-WEC datasets consistently outperformed those trained on the 100-WEC datasets (e.g., Sydney 49-WEC vs. Sydney 100-WEC). This performance drop in larger farms is likely attributable to the "curse of dimensionality." A 100 WEC farm requires 200 coordinate features compared to the 98 features required for a 49 WEC farm. The higher dimensional feature space, combined with a relatively smaller number of unique layouts in the 100-WEC datasets (e.g., only 1,234 unique layouts for Sydney 100-WEC), makes the underlying spatial power relationship more difficult for the models to resolve with the available data.

7.5 Limitations and Physical Interpretability: While the predictive performance is strong, several limitations must be acknowledged. First, the models rely exclusively on spatial coordinates. Environmental forcing variables (e.g., wave height, wave period, wind direction) are absent from the visible schema. Consequently, the models capture a *predictive association* between geometry and power, rather than explicitly modeling the causal hydrodynamic mechanisms. Second, the evaluation relies on a random 80/20 holdout split of unique

layouts. While this prevents row-level leakage, it does not test the model's ability to generalize to entirely unseen geographical sites, novel spatial distributions, or different WEC counts. Finally, the minor discrepancies observed between the *Total_Power* target and the sum of individual converter powers warrant further investigation against the original data-generation procedures to ensure absolute physical consistency.

8. Conclusion

This study successfully demonstrated the viability of using machine learning-based regression on spatial coordinates as a fast, accurate surrogate for predicting the total power output of large scale wave energy farms. By leveraging high fidelity simulation data from the UCI Machine Learning Repository, we showed that non linear ensemble models specifically Histogram based Gradient Boosting and Extra Trees can capture the complex spatial dependencies governing wave farm power output. The models achieved up to 95.5% explained variance (R^2) and sub-1% mean absolute percentage errors (MAPE) on rigorously de-duplicated, held-out layout configurations.

Practical Implications: The ability to predict aggregate farm power from spatial coordinates in milliseconds, rather than minutes or hours, has profound implications for the wave energy sector. These ML surrogates can be seamlessly integrated into evolutionary optimization algorithms to rapidly search high-dimensional layout spaces, identifying configurations that minimize destructive wave interference and maximize energy capture. Furthermore, such models can serve as the predictive engine for digital-twin frameworks, enabling real-time performance monitoring, adaptive energy-management strategies, and predictive maintenance scheduling for operational WEC farms.

8.1 Future Work:

To advance this research toward operational deployment, future studies should focus on the following directions:

1. Spatial and Cross-Site Generalization: Moving beyond random holdout splits to evaluate model performance on spatially separated test sets and entirely new geographic wave climates.

2. Integration of Environmental Covariates: Expanding the feature space to include time-varying environmental data (wave height, period, direction) to transition from static layout surrogates to dynamic, real-time forecasting models.

3. Physics-Informed Machine Learning: Incorporating hydrodynamic constraints and symmetry invariances into the loss functions to improve data efficiency and ensure physical consistency, particularly for the 100-WEC configurations where dimensionality poses a challenge.

4. Engineered Geometric Descriptors: Replacing or supplementing raw X/Y coordinates with engineered features (e.g., pairwise distances, nearest-neighbor metrics, array density) to reduce dimensionality and improve model interpretability via techniques like SHAP (SHapley Additive exPlanations).

5. Data Provenance Auditing: Investigating the root cause of the discrepancies between total farm power and the sum of individual device powers in the source dataset to ensure the highest fidelity of the target variables.

Ultimately, this work establishes a robust, computationally efficient baseline for data-driven WEC farm analysis, bridging the gap between expensive physics-based hydrodynamic modeling and the urgent need for scalable renewable energy optimization.

Reproducibility files

The accompanying workbook contains all model metrics, best model summaries, data quality checks, validation protocol, and held out predictions. CSV files and PNG figures are also included in the output package.

References

[1] Said, H. A., Ringwood, J. V. (2021). Grid integration aspects of wave energy Overview and perspectives. *IET Renewable Power Generation, 15*, 3045–3064.

[2] Guo, C., Sheng, W., De Silva, D. G., Aggidis, G. (2023). A review of the levelized cost of wave energy based on a techno-economic model. *Energies, 16*, Article 2144.

[3] Li, K., Nallaparaju, A., Sha, J. (2024). A novel machine learning approach for the prediction of optimized energy output on large-scale wave energy farms. In *2024 IEEE MIT Undergraduate Research Technology Conference (URTC)* (p. 1–5). IEEE. https://doi.org/10.1109/URTC65039.2024.10937646.

[4] Wan, A., Liao, X., Iqbal, A., Al-Bukhaiti, K., Zhu, P., Cheng, X., Ji, X. (2026). Machine learning in offshore wind energy systems: A systematic review (2022–2025). *Journal of Mechanical Science and Technology, 1–9*.

[5] Neshat, M., Alexander, B., Sergiienko, N. Y., Wagner, M. (2020). New insights into position optimisation of wave energy converters using hybrid local search. *Swarm and Evolutionary Computation, 59*, Article 100744.

[6] Achouch, M., Dimitrova, M., Ziane, K., Karganroudi, S. S., Dhouib, R., Ibrahim, H., Adda, M. (2022). On predictive maintenance in Industry 4.0: Overview, models, and challenges. *Applied Sciences, 12*, Article 8081.

[7] Plazas-Niño, F., Ortiz-Pimiento, N., Montes-Páez, E. (2022). National energy system optimization modelling for decarbonization pathways analysis: A systematic literature review. *Renewable and Sustainable Energy Reviews, 162*, Article 112406.

[8] Kandemir, E., Hasan, A., Kvamsdal, T., Alaliyat, S. A.-A. (2024). Predictive digital twin for wind energy systems: A literature review. *Energy Informatics, 7*, Article 68.

[9] Durap, A. (2024). Data-driven models for significant wave height forecasting: Comparative analysis of machine learning techniques. *Results in Engineering, 24*, Article 103573.

[10] Muthamizhan, T., Karthick, K., Aruna, S. K., Velmurugan, P. (2025). AI-driven stacking ensemble for predicting total power output of wave energy converters: A data-driven approach to renewable energy processes. *Processes, 13*(4), Article 961. https://doi.org/10.3390/pr13040961.

[11] Gu, C., Li, H. (2022). Review on deep learning research and applications in wind and wave energy. *Energies, 15*, Article 1510.

[12] Serras, P., Ibarra-Berastegi, G., Sáenz, J., Ulazia, A. (2019). Combining random forests and physics-based models to forecast the electricity generated by ocean waves: A case study of the Mutriku wave farm. *Ocean Engineering, 189*, Article 106314.

[13] Alkhayat, G., Mehmood, R. (2021). A review and taxonomy of wind and solar energy forecasting methods based on deep learning. *Energy and AI, 4*, Article 100060.

[14] Li, Y., Ma, X., Tang, T., Zha, F., Chen, Z., Liu, H., & Sun, L. (2022). High efficient built in wave energy harvesting

technology: From laboratory to open ocean test. *Applied Energy, 322*, Article 119498.

[15] Scruggs, J., Jacob, P. (2009). Harvesting ocean wave energy. *Science, 323*(5918), 1176–1178.

[16] Nam, K., Hwangbo, S., Yoo, C. (2020). A deep learning-based forecasting model for renewable energy scenarios to guide sustainable energy policy: A case study of Korea. *Renewable and Sustainable Energy Reviews, 122*, Article 109725.

[17] Gu, C., Li, H. (2022). Wave power density hotspot distribution and correlation pattern exploration in the Gulf of Mexico. *Sustainability, 14*, Article 1158.

[18] Al Shami, E., Zhang, R., Wang, X. (2018). Point absorber wave energy harvesters: A review of recent developments. *Energies, 12*(1), Article 47.

[19] Guillou, N., Lavidas, G., Chapalain, G. (2020). Wave energy resource assessment for exploitation A review. *Journal of Marine Science and Engineering, 8*, Article 705.

[20] Lu, Z. Q., Zhao, L., Fu, H. L., Yeatman, E., Ding, H., Chen, L. Q. (2024). Ocean wave energy harvesting with high energy density and self-powered monitoring system. *Nature Communications, 15*, Article 6513.

[21] Pirhooshyaran, M., Snyder, L. V. (2020). Forecasting, hindcasting and feature selection of ocean waves via recurrent and sequence to sequence networks. *Ocean Engineering, 207*, Article 107424.

[22] Dong, X., Li, Y., Li, D., Cao, F., Jiang, X., Shi, H. (2022). A state-of-the-art review of the hybrid wind-wave energy converter. *Progress in Energy, 4*(4), Article 042004.

[23] McTiernan, K. L., Sharman, K. T. (2020). Review of hybrid offshore wind and wave energy systems. *Journal of Physics: Conference Series, 1452*, Article 012016.

[24] Chen, L., Li, Z., Zhang, Y. (2019). Multiperiod-ahead wind speed forecasting using deep neural architecture and ensemble learning. *Mathematical Problems in Engineering, 2019*, Article 9240317.

[25] Shahid, F., Zameer, A., Muneeb, M. (2021). A novel genetic LSTM model for wind power forecast. *Energy, 223*, Article 120069.

[26] Wei, Z. (2021). Forecasting wind waves in the US Atlantic Coast using an artificial neural network model: Towards an AI-based storm forecast system. *Ocean Engineering, 237*, Article 109646.

[27] Ahmad, T., Zhang, D. (2022). A data driven deep sequence to sequence long-short memory method along with a gated recurrent neural network for wind power forecasting. *Energy, 239*, Article 122109.

[28] Tian, Z., Li, S.,Wang, Y. (2020). A prediction approach using ensemble empirical mode decomposition-permutation entropy and regularized extreme learning machine for short term wind speed. *Wind Energy, 23*, 177–206.

[29] Zhao, W., Wei, Y. M., Su, Z. (2016). One day ahead wind speed forecasting: A resampling-based approach.

Applied Energy, 178, 886–901.

[30] Lahza, H., Alsheikhy, A. A., Shawly, T., et al. (2026). A multiscale energy polarization attention mechanism and H-shaped network for AI-driven biomimetic oceanic energy farms. *Scientific Reports*. https://doi.org/10.1038/s41598-026-72581-0.

[31] Sarkar, D., Contal, E., Vayatis, N., Dias, F. (2016). Prediction and optimization of wave energy converter arrays using a machine learning approach. *Renewable Energy, 97*, 504–517. https://doi.org/10.1016/j.renene.2016.05.083.

[32] Zhang, J., Taflanidis, A. A., Scruggs, J. T. (2020). Surrogate modeling of hydrodynamic forces between multiple floating bodies through a hierarchical interaction decomposition. *Journal of Computational Physics, 408*, Article 109298. https://doi.org/10.1016/j.jcp.2020.109298.

[33] Mousavi, S. M., Ghasemi, M., Dehghan Manshadi, M., Mosavi, A. (2021). Deep learning for wave energy converter modeling using long short term memory. *Mathematics, 9*(8), Article 871. https://doi.org/10.3390/math9080871.

[34] Neshat, M., Sergiienko, N. Y., Rafiee, A., Mirjalili, S., Gandomi, A. H., Boland, J. (2024). Meta wave learner: Predicting wave farms power output using effective meta-learner deep gradient boosting model: A case study from Australian coasts. *Energy, 304*, Article 132122. https://doi.org/10.1016/j.energy.2024.132122.

[35] Li, M., Jia, G., Mahmoud, H., Yu, Y. H., Tom, N. (2023). Physics-constrained Gaussian process model for prediction of hydrodynamic interactions between wave energy converters in an array. *Applied Mathematical Modelling, 119*, 465–485. https://doi.org/10.1016/j.apm.2023.03.003.

[36] Neshat, M., Alexander, B., Sergiienko, N. Y., Wagner, M. (2020). Optimisation of large wave farms using a multi-strategy evolutionary framework. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference* (p. 1150–1158). Association for Computing Machinery.