NATIONAL AUTONOMOUS UNIVERSITY OF MEXICO

POSTGRADUATE DEGREE IN MARINE SCIENCES AND LIMNOLOGY

FROM DESCRIPTIVE STATISTICS TO MACHINE LEARNING: A COMPREHENSIVE APPROACH TO WAVE ANALYSIS WITH RESISTIVE SENSORS

Research work

Statistical Analysis and Data Science Applied to Environmental Data Using Python

PRESENTS:
JOSÉ ROLANDO BOFFILL VÁZQUEZ

MEXICO, SISAL, SEPTEMBER,

(2025)

1. INTRODUCTION 1

1.2 Research questions and hypotheses 2

1.3 Objectives 2
1.3.1 General objectives 2
1.3.2 Specific objectives 2

2. METHODOLOGY 3

2.1. Channel Description 3

2.2. Experimental design 3

2.3. Test Configuration 4

2.4. Instrumentation 4

2.4 Wave analysis 5

2.5. Statistical analysis 5

2.6. Modeling with Machine Learning 6

3. RESULTS 6

3.1. Wave characterisation using free surface time series 7

3.2. Descriptive statistics 7
3.2.1 Descriptive analysis of mean wave height 8
3.2.2 Descriptive analysis of mean water level 9

3.3. Non-statistical tests 11
3.3.1. Analysis of mean wave height 11
3.3.2 Analysis of mean water level 12

3.4. Machine-learning models 12

4. DISCUSSION 14

5. CONCLUSIONS 14

6. ACKNOWLEDGEMENTS 15

7. REFERENCES 15

1. Introduction

Wave breaking in the ocean is a highly important phenomenon, since breaking waves play a significant role in all aspects of air–sea exchange processes, including momentum, heat and mass transfer (Xu et al., 1986).

When waves break, they induce a change in the radiation stress tensor (the excess momentum flux associated with the presence of waves (Bowen et al., 1968)) and this change produces variations in the mean free-surface elevation (Longuet-Higgins & Stewart, 1964) when waves interact with a sloping beach. The set-up is confined to the breaking or surf zone, shoreward of the initial breaking point; conversely, the set-down is a depression in the mean level that occurs shoreward of the breaking point, where it attains its maximum value (see **Figure 1**).

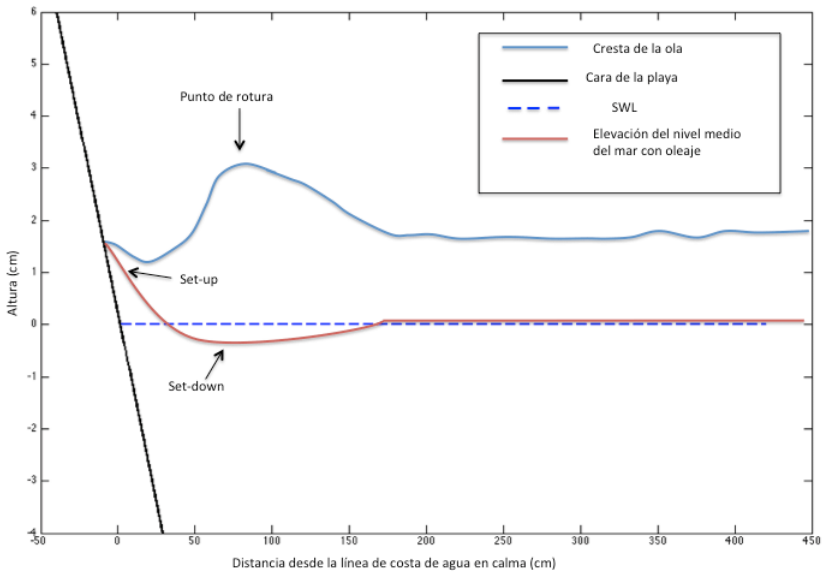

Figure 1. Set-up and set down induced by waves in the breaking zone (From García, 2016).

Laboratory experimentation makes it possible to recreate simplified, scaled scenarios of reality, enabling a deeper understanding of the processes involved while retaining the fundamental physics of the problem (Hughes, 1993).

Measurement of the free surface is a fundamental element in the study of hydrodynamic processes because it allows characterization of wave behaviour and its interaction with coastal structures or environments. Analysis of these signals is essential both in laboratory experimental studies and in coastal and hydraulic engineering applications, where

understanding the variability of wave and water-level parameters is critical for design, management and prediction.

In this context, one of the main challenges is to characterize the spatial and temporal variability recorded by multiple sensors, which requires data manipulation, cleaning and analysis techniques that ensure the quality of the information. Likewise, the growing availability of statistical and machine-learning methods offers new tools to detect patterns, contrast differences between scenarios, and predict relevant behaviours under controlled conditions.

1.2 Research questions and hypotheses

Are there significant differences in wave height and mean water-level variation between the tests, and is it possible to predict these variables from machine-learning techniques applied to the readings of resistive sensors?

There are statistically significant differences in mean wave height and mean water-level variation between the tests; these variables possess predictive potential that can be exploited using machine-learning models.

1.3 Objectives

This practice was developed with the following general and specific objectives.

1.3.1 General objectives

Calculate characteristic wave parameters such as mean wave height and mean water-level variation, and apply descriptive and inferential statistical analysis as well as classification and regression techniques using machine-learning algorithms.

1.3.2 Specific objectives

Clean and prepare time series from multiple sensors; construct a homogeneous database.

Carry out exploratory analysis and visualizations that characterize spatial and temporal differences.

Calculate descriptive statistics and study the distributions.

Apply parametric and nonparametric tests to contrast hypotheses about differences between scenarios.

Implement classification models (k-NN, decision trees) and regression models (linear regression and regression trees) to evaluate the predictive potential of the data.

2. Methodology

The laboratory tests, conducted on an idealized beach model, were performed in the Coastal Engineering and Processes Laboratory (LIPC) of the Institute of Engineering at the National Autonomous University of Mexico (UNAM).

2.1. Channel Description

The wave channel is 40 m long, 0.8 m wide and 1.27 m high (**Figure 2**). It is equipped with a unidirectional wavemaker composed of a piston-type paddle with a power of 7.5 kW and 1.2 m stroke. This system can generate regular waves, second-order irregular waves and solitary waves (VTI, 2015). The generation system uses the AwaSys6 software, developed by Aalborg University, to control the wavemaker. It includes an active wave-absorption system that suppresses reflected waves incident on the paddle (Rodríguez et al., 2023).

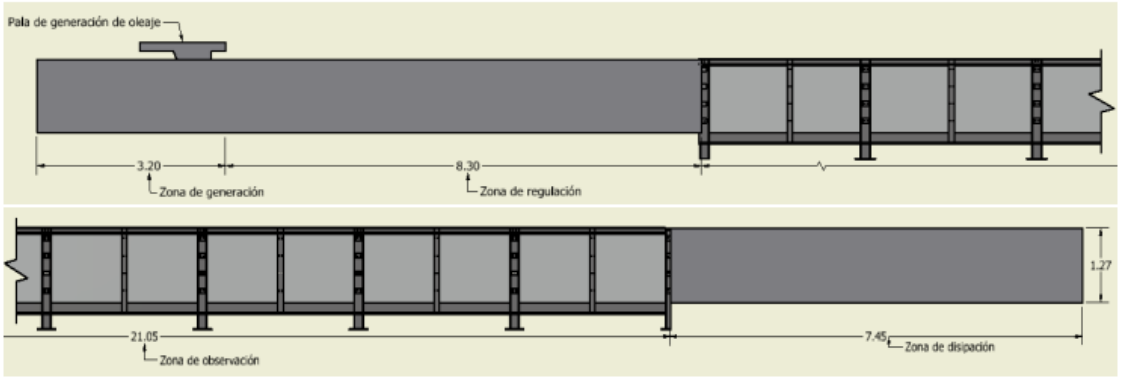

Figure 2. Schematic of the wave channel (From Rodríguez et al., 2023).

2.2. Experimental design

A two-dimensional (2D) idealized model was constructed to examine the hydrodynamic processes. In the construction of the bottom profile (**Figure 3**) a fixed bed model representing a beach with a 1:10 slope was considered; the ramp dimensions are 10 m in length, 0.8 m in width and 1.0 m in height, and the still-water level (MSL) was set at

0.58 m.

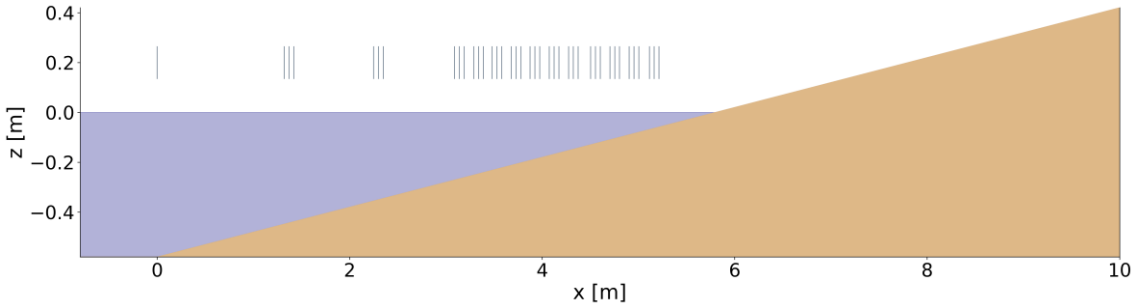

Figure 3. Schematic representation of the configuration of the resistive sensors for the study cases (the horizontal scale has been exaggerated for clarity).

2.3. Test Configuration

Three tests were carried out on fixed beds, considering a smooth surface. Experiments were run under regular-wave conditions with various periods, as detailed in **Table 1**. Each test lasted 480 seconds. In all cases, data acquisition began at the exact moment the wavemaker was started, with the water level at rest, and measurements were recorded at 50 Hz at each sensor.

Table 1. Study cases.

TRIALS	H[m]	T[s]	h0 [m]
H10T12	0.1	3.0	0.58
H10T25	0.1	2.5	0.58
H10T30	0.1	1.2	0.58

2.4. Instrumentation

To obtain time records of the free surface elevation, level sensors operating on the principle of electrical resistivity were used. These instruments determine the water-surface level via closure of an electrical circuit between two electrodes. The resulting potential difference is recorded as a voltage and is directly proportional to the water level (Rodríguez et al., 2023). From the voltage data measured by these sensors, elevations can be obtained using a calibration curve. The level sensors were calibrated simultaneously for 20 seconds, performing measurements at six different still-water levels (0.58 m, 0.62 m, 0.64 m and 0.66 m) and fitting a first-order polynomial to the measured voltages. This yields conversion of the measured voltage signals to centimetres. The elevation time series are then used to determine wave heights and periods at each sensor.

2.4 Wave analysis

After obtaining the elevation time series, data processing was performed using the Python 3 programming language (version 3.11). This process included filtering, statistical analysis and calculation of characteristic wave parameters such as mean wave height and mean water-level variation. Python, through specialized libraries such as NumPy, SciPy and Pandas, facilitates efficient handling of large datasets and development of algorithms specific to wave analysis.

A resampling procedure was applied, reducing the sampling frequency from 50 Hz to 20 Hz. To preserve the main characteristics of the signal while attenuating high-frequency noise, a Savitzky–Golay filter was used. Subsequently, outlier detection and correction were carried out, followed by homogenization of the time series in order to form a consistent database suitable for subsequent statistical analysis and modelling.

Using the zero down-crossing method on the free-surface series from each sensor, the parameters of each wave (period (T) and mean height ($\bar{H}$)) were calculated.

$$\bar{H} = \frac{1}{n}\sum_{i=1}^{n} H_i$$

The experimental mean level ($\bar{\eta}$) for each sensor and study case was determined from the still-water mean ($\bar{\eta}_{rep}$) and the sample mean ($\bar{\eta}_{med}$) calculated from a subset of the recorded free-surface series.

$$\bar{\eta} = \bar{\eta}_{med} - \bar{\eta}_{rep}$$

The still-water mean was computed from the first 20 seconds of the time series, while the mean was computed using the remaining 460 seconds of recorded data.

2.5. Statistical analysis

Statistical analysis was carried out at three complementary levels. First, a descriptive analysis was performed that included measures of central tendency (mean and median) and measures of dispersion (variance, standard deviation, interquartile range and coefficient of variation). This analysis was complemented with graphical representations that characterize the variability of the series.

Second, boxplots were used to identify data dispersion and detect outliers. Kernel-smoothed histograms for each test were used to explore the shape of the distributions and possible skewness. Quantile–quantile (QQ) plots were also produced to compare observed values with a theoretical normal distribution, facilitating identification of departures from normality.

Finally, inferential statistical tests were applied: parametric tests (Student's t-test, ANOVA and Chi-square) and nonparametric tests (Mann–Whitney, Wilcoxon, Kruskal–Wallis and Friedman), chosen according to the nature of the data and the objectives of the analysis. This comprehensive approach provides a deeper understanding of the characteristics and relationships in the analysed data.

2.6. Modeling with Machine Learning

Two complementary machine-learning approaches were considered. For classification, wave categories (low, medium and high) were defined and supervised algorithms (k-nearest neighbours (k-NN) and decision trees) were implemented to identify these categories. For regression, models aimed at predicting continuous variables of interest (specifically mean wave heigh ($\bar{H}$)) were applied, using both linear regression and regression trees. Model quality and predictive capability were evaluated with performance metrics including accuracy and F1-score for classification, and RMSE and the coefficient of determination (R^2) for regression, incorporating cross-validation to ensure robustness and generalizability.

3. Results

This chapter presents the results of the experimental study. The main objective is to evaluate the impact of the tested configurations on wave transformation under different energy and period conditions. The research was carried out using data generated from physical modelling in a wave channel. In addition, a process of validation and calibration of a numerical model was applied, which allowed obtaining results with high spatial resolution.

3.1. Wave characterisation using free surface time series

The **Figure 4** how's the variation of wave height and the mean water-level variation along the regular-wave periods studied. In subfigures (a), (c) and (e) the evolution of wave height is compared, where an initial increase followed by progressive dissipation shoreward is observed.

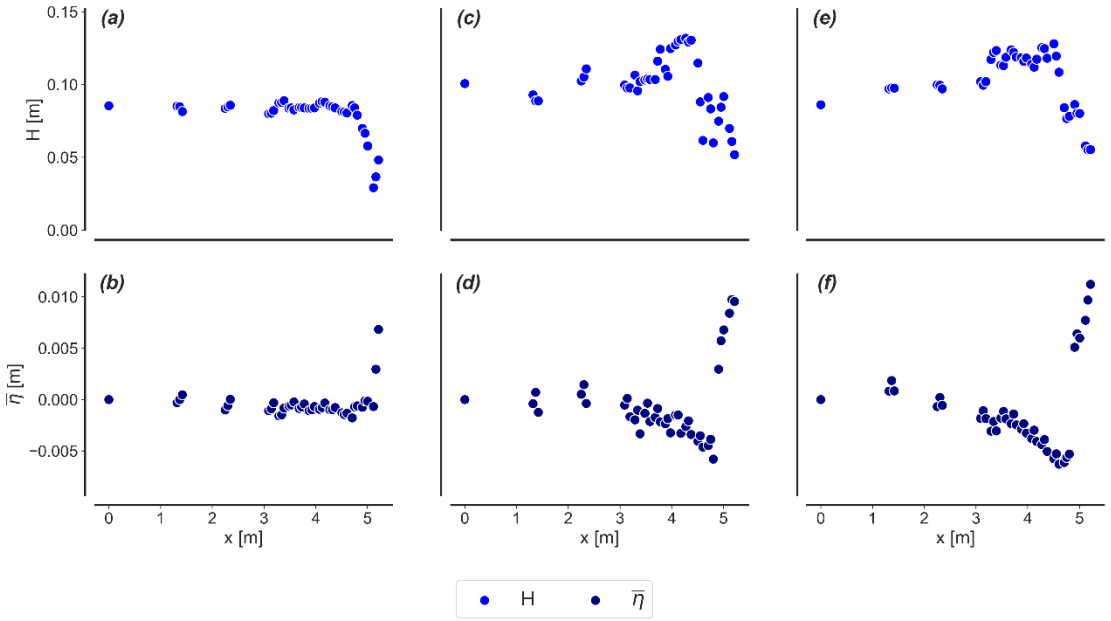

Figure 4. Results of the regular-wave tests for the cases corresponding to T = 1.2 s ((a)–(b)), T = 2.5 s ((c)–(d)) and T = 3.0 s ((e)–(f)).

The mean water-level variation, shown in subfigures (b), (d) and (f), reveals a maximum depression (set down) at the breaking point and a gradual increase (set up) in the region landward of it. The mean level remains practically constant until the breaking zone, where the set-down reaches its maximum value, followed by a gradual growth of the set-up up to the beach toe.

3.2. Descriptive statistics

The following figures and tables present the descriptive analysis of mean wave height ($\bar{H}$) and mean water level ($\bar{\eta}$) for the three tests studied (H10T12, H10T25 and H10T30).

3.2.1 Descriptive analysis of mean wave height

The boxplots (**Figure 5**) reveal differences in dispersion and presence of outliers across the three cases. Greater variability is observed in H10T25 and H10T30, while H10T12 shows a more concentrated distribution. These results suggest that under longer-period conditions the wave response becomes less homogeneous and more prone to extreme fluctuations.

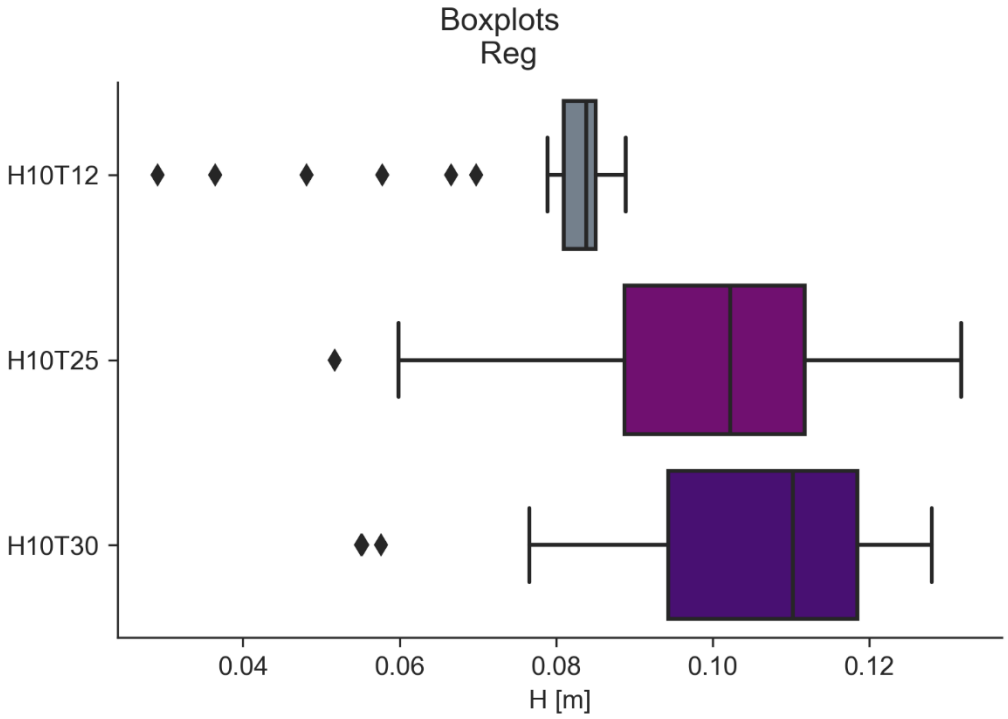

Figure 5. Boxplots of mean wave height ($\bar{H}$) for three tests: H10T12, H10T25 and H10T30.

Mean wave height increased progressively across the three tests (**Table 2**), rising from average values of 0.079 m in H10T12 to 0.099 m in H10T25 and reaching 0.103 m in H10T30. This increase was accompanied by higher relative dispersion, reflected in coefficients of variation between 0.17 and 0.21, indicating that more energetic wave conditions were associated with greater variability.

Table 2. Descriptive statistics for mean wave height ($\bar{H}$)

Practice	stocking	median	variance	Std	IQR	CV
H10T12	0.079183	0.083880	0.000180	0.013419	0.004115	0.169469
H10T25	0.099850	0.102216	0.000437	0.020912	0.023056	0.209433
H10T30	0.103088	0.110229	0.000404	0.020091	0.024214	0.194895

Regarding distribution shape, the kernel-smoothed histograms (**Figure 6**) showed that in H10T12 values are more concentrated around the mean, with marked skewness, while in H10T25 and H10T30 the distributions are more spread and closer to a unimodal shape. However, the QQ-plots (**Figure 7**) revealed that the normality hypothesis is not fully satisfied in any test: H10T12 exhibits notable deviations in the tails, whereas H10T25 and H10T30 align more closely with the theoretical line, although discrepancies persist in the extremes.

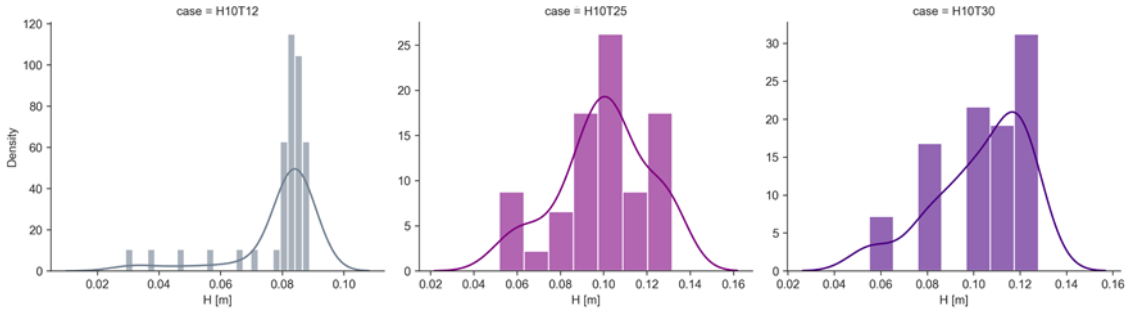

Figure 6. Distribution analysis via kernel-smoothed histograms for mean wave height ($\overline{H}$) the three tests: H10T12, H10T25 and H10T30.

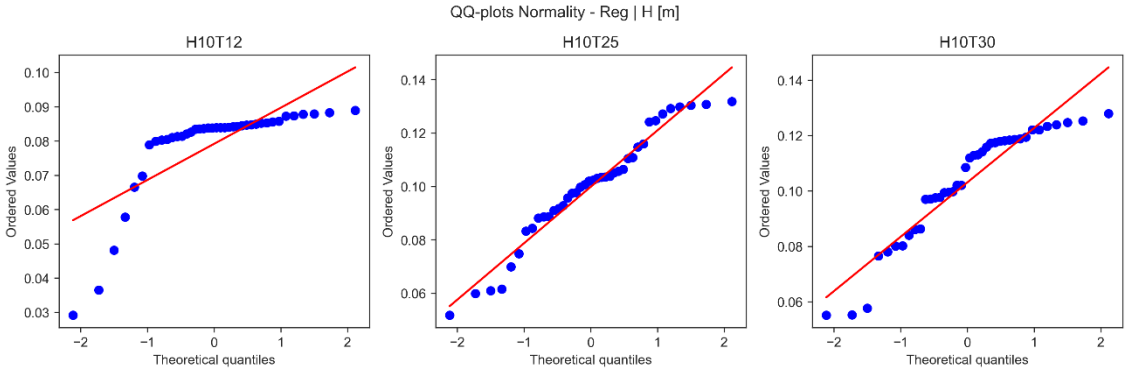

Figure 7.: QQ-plots for mean wave height ($\overline{H}$) in the three tests: H10T12, H10T25 and H10T30.

3.2.2 Descriptive analysis of mean water level

The boxplots (**Figure 8**) show differences in dispersion and outlier presence between cases: while H10T12 exhibits reduced variability and few outliers, H10T25 and especially H10T30 display increased dispersion and a higher frequency of extreme values, suggesting a more unstable free-surface behaviour under stronger wave conditions.

The table of descriptive statistics (**Table 3**) complements this observation by showing that both the mean and median are negative in all tests, with larger magnitudes in H10T25 and H10T30. This reflects a progressive mean lowering of the water level as the test

energy increases. Variance and standard deviation also increase from H10T12 to H10T30, confirming the trend toward greater dispersion in the more energetic experiments.

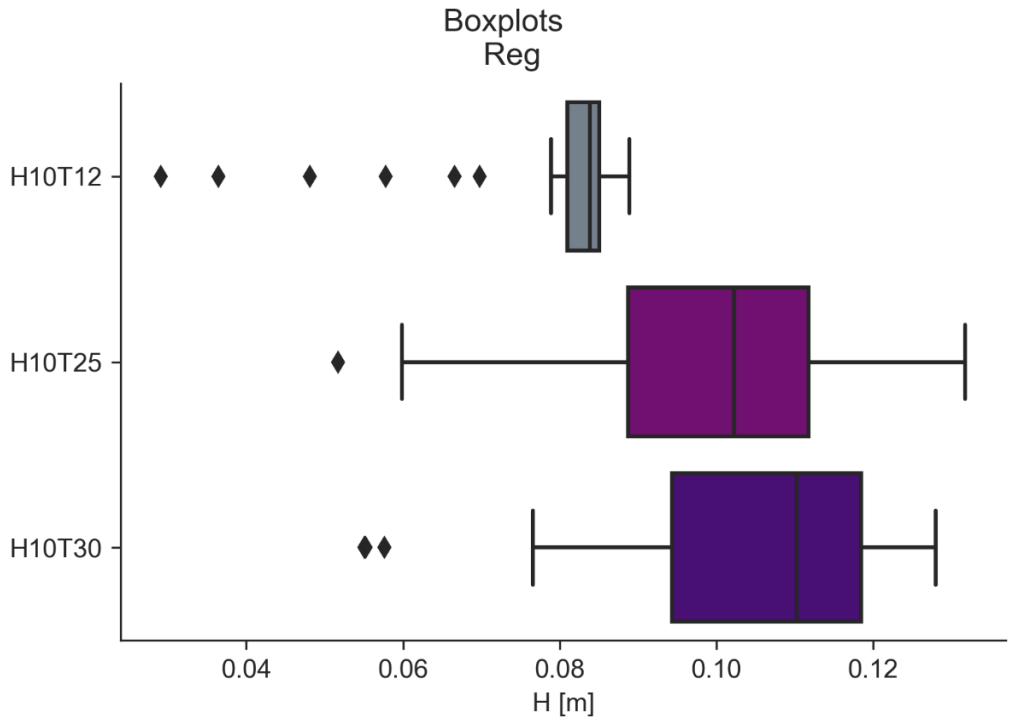

Figure 8. Boxplots of mean water-level variation ($\bar{\eta}$) or three tests: H10T12, H10T25 and H10T30.

Table 3. Descriptive statistics for mean water-level variation ($\bar{\eta}$)

Practice	media	median	variance	Std	IQR	CV
H10T12	-0.002685	-0.002955	0.000002	0.001398	0.000679	0.52054
H10T25	-0.004002	-0.004967	0.000014	0.00375	0.002804	0.937102
H10T30	-0.004724	-0.005667	0.000018	0.004292	0.003882	0.908626

Distribution analysis via kernel-smoothed histograms (**Figure 9**) reveals differentiated behavior's: in H10T12 the data concentrate around values close to zero, whereas in H10T25 and H10T30 the distributions are wider and skewed, with heavier tails toward both negative and positive values. Finally, the QQ-plots (**Figure 10**) indicate that empirical distributions deviate from the theoretical normal in all three cases, particularly in the tails of H10T25 and H10T30, confirming the presence of heavy tails and departure from normality.

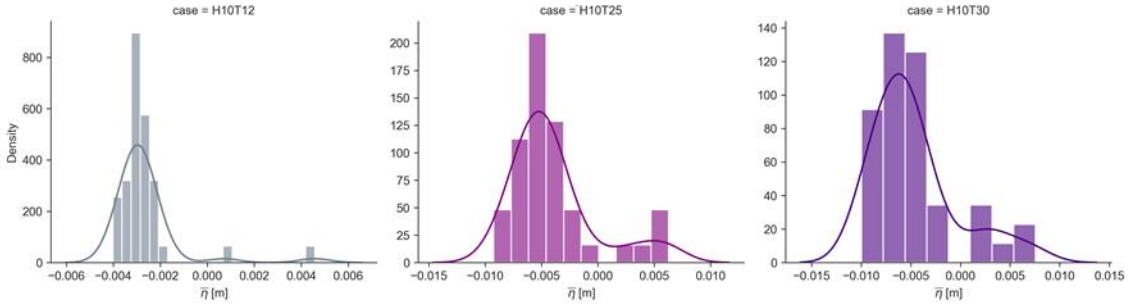

Figure 9. Distribution analysis via kernel-smoothed histograms for mean water-level variation ($\bar{\eta}$) in the three tests: H10T12, H10T25 and H10T30.

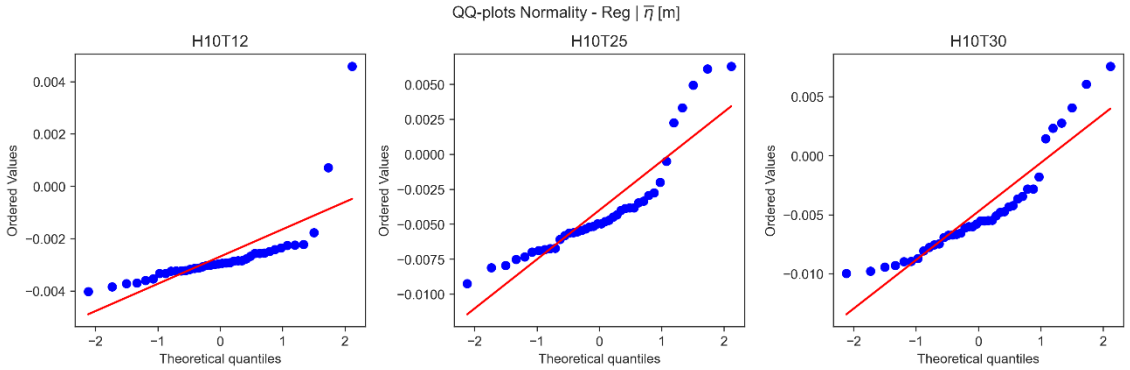

Figure 10. QQ-plots for mean water-level variation ($\bar{\eta}$) in the three tests: H10T12, H10T25 and H10T30.

3.3. Non-statistical tests

Nonparametric methods were adopted because the series exhibited skewness, outliers and departures from normality, which limit the validity of parametric tests. These procedures allow robust contrasts of differences between scenarios without strict distributional assumptions.

3.3.1. Analysis of mean wave height

The nonparametric analysis (see **Table 4**) using the Friedman test confirms the existence of significant differences in mean wave height among the three scenarios. Wilcoxon post-hoc contrasts indicate that differences are concentrated between H10T12 and the tests H10T25 and H10T30, with highly significant p-values. No difference was detected between H10T25 and H10T30, suggesting that beyond a certain period threshold ($\geq$ 2.5 s) the mean wave height tends to stabilise.

Table 4. Variation values for mean wave height ($\bar{H}$)

Comparison	Statistical	p-value	P-Adjusted	Significant
Friedman (global)	43.55	3.49E-10	-	Yes
H10T12 vs H10T25	34	6.78E-09	2.03E-08	Yes
H10T12 vs H10T30	10	7.82E-11	2.35E-10	Yes
H10T25 vs H10T30	338	0.3403	1	No

3.3.2 Analysis of mean water level

The Friedman test revealed significant differences in mean water-level variation between the three scenarios (see **Table 5**), confirming a systematic effect of period on the mean level displacement. Wilcoxon post-hoc contrasts showed that all pairs of tests differ significantly, indicating that both the transition from H10T12 to H10T25 and from H10T25 to H10T30 produce statistically relevant changes. This pattern suggests a progressive and consistent trend of mean-level variation with increasing period, without evidence of stabilisation between the tested scenarios.

Table 5. Values of mean water-level variation ($\bar{\eta}$)

Comparison	Statistical	P-Value	P-Adjusted	Significant
Friedman (global)	30.05	2.98E-07	-	Yes
H10T12 vs H10T25	190	0.00251	0.0075	Yes
H10T12 vs H10T30	178	0.00137	0.0041	Yes
H10T25 vs H10T30	153	0.000336	0.001	Yes

3.4. Machine-learning models

For regular waves (H10T12, H10T25 and H10T30), feature vectors were constructed on sliding windows from basic statistical parameters of the signal: mean wave height, mean absolute variation, standard deviation, skewness and kurtosis. These variables were used to train classifiers to identify the corresponding wave case.

Results show that the k-NN model (**Figure 11**(a)) reached an accuracy of 0.833 and an F1 score of 0.833, with some classification errors between the low and medium levels. The decision tree (**Figure 11**(b)) performed better, obtaining an accuracy of 0.917 and an F1 of 0.917, correctly classifying most cases and showing less confusion between classes.

This indicates that decision trees are more effective than k-NN at discriminating among regular-wave conditions.

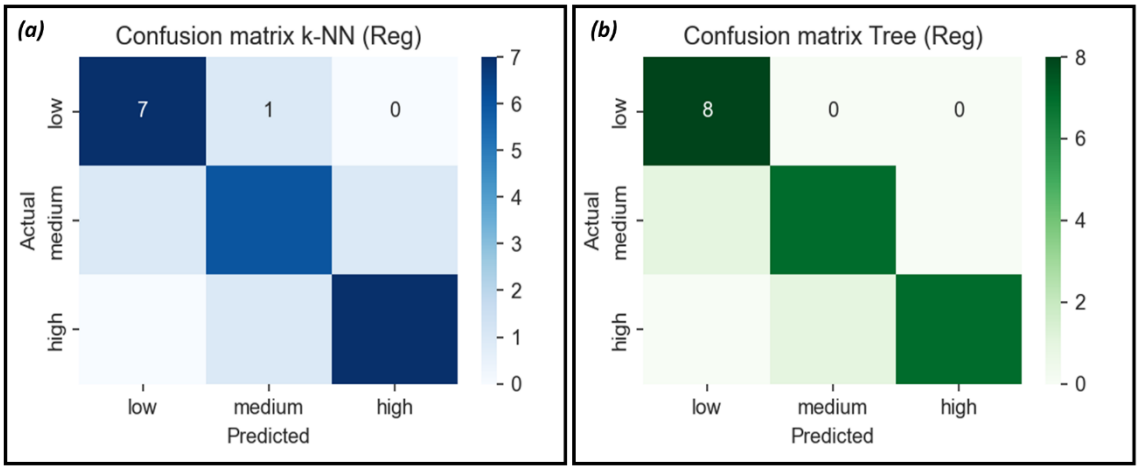

Figure 11.: k-NN model (a) and decision tree (b) for regular waves.

For regression (**Figure 12**), prediction of wave height at a target sensor was evaluated using as predictors the features from neighbouring sensors or previous conditions. Linear regression showed an almost perfect fit ($R^2 = 1.000$, RMSE ≈ 0.0008), with predicted values lying practically on the diagonal in the Predicted vs. Observed plot. In contrast, the regression tree obtained $R^2 = 0.969$ and RMSE = 0.0413, reflecting adequate performance but with greater scatter around the diagonal, especially for intermediate values.

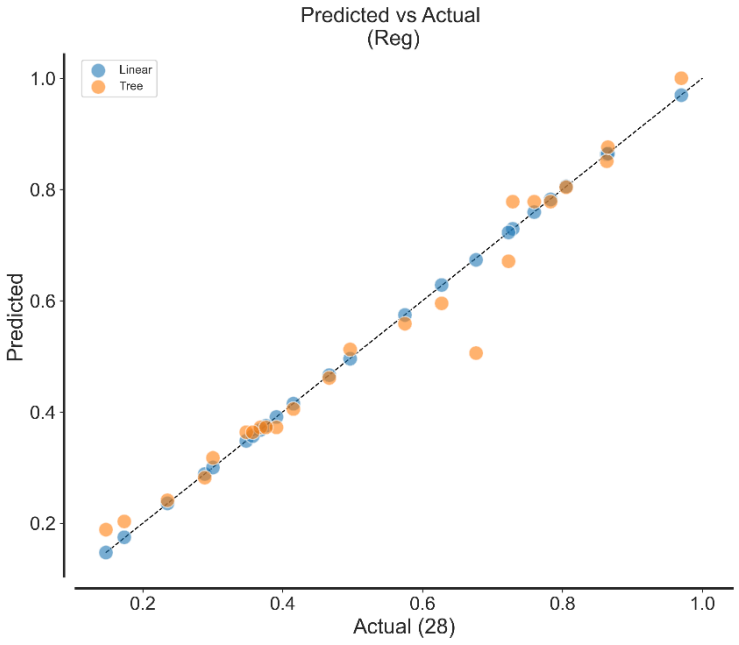

Figure 12. Linear regression for regular waves

4. Discussion

Statistical analyses confirmed that both mean wave height and mean water-level variation differ significantly between scenarios, reflecting a direct effect of period on free-surface dynamics. In particular, nonparametric results showed that for wave height the differences are concentrated in the transition from H10T12 to H10T25, while mean water level exhibited significant differences across all pairs, suggesting a progressive trend without stabilisation.

Complementing these findings with a machine-learning approach showed that classifiers can discriminate between regular-wave conditions using simple statistical variables. Although k-NN achieved acceptable performance, the decision tree demonstrated greater classification capability, reducing confusion between wave levels. In regression, linear regression achieved an almost perfect fit, whereas the regression tree, although precise, showed greater dispersion. Together, these results confirm that both traditional statistical methods and machine-learning approaches provide complementary and robust information to characterise and predict wave behaviour under controlled conditions.

5. Conclusions

- Under regular-wave conditions, wave height increases up to the breaking zone and then dissipates progressively shoreward, while the mean water level exhibits a maximum set-down at the breaking point followed by a gradual set-up toward the beach toe.
- Mean wave height increased progressively across the tests, accompanied by greater dispersion and the presence of extreme values.
- The distributions showed skewness for wave-height values and QQ-plots confirmed deviations from normality, which highlights the advisability of applying nonparametric or robust statistical methods when comparing cases.
- Mean water-level variation exhibited a progressive decrease from H10T12 to H10T30, together with asymmetric distributions and the presence of outliers.
- Results indicate that increasing period induces systematic displacements of the mean level, with deviations from normality that justify using nonparametric tests in comparative analysis.

- Mean wave height differs significantly only between H10T12 and the other cases, while mean water level differs among all scenarios, evidencing its greater sensitivity to wave period.
- Decision trees proved more effective for classifying regular-wave scenarios, whereas linear regression provided the best accuracy for predicting wave heights at a target sensor.

6. Acknowledgements

We thank the staff of the Coastal Engineering and Processes Laboratory (LIPC) at the Institute of Engineering, UNAM, in particular Oceanographer Camilo Sergio Rendón Valdez, for the support provided during the experiments and data acquisition. We also acknowledge Dr. Alec Torres Freyermuth for his supervision of the tests, and Dr. Gabriela Medellín Mayoral and Dr. José Carlos Pintado Patiño, members of the committee, for their valuable guidance and recommendations during the experimental development.

7. References

1. Bowen, A. J., Linman, D., and Simmons, V. P. (1968) *Wave Set - Down and Set-Up*. Journal of Geophysical Research, 73, 8.
2. García, A.D. (2016). *Variaciones de la línea de costa debido a la marea y la obtención de modelos digitales de elevación* [Título profesional, Universidad Nacional Autónoma de México, Facultad de Ciencias, Unidad Multidisciplinaria de Docencia e Investigación Sisal]. https://ru.dgb.unam.mx/bitstream/20.500.14330/TES01000769011/3/0769011.pdf
3. Hughes, S. A. (1993). *Physical models and laboratory techniques in coastal engineering*, 7. World Scientific.
4. Longuet-Higgins, M. S. & Stewart, R. W. (1962). *Radiation stresses and mass transport in gravity waves with applications to surf beats*. J. Fluid Mech, 13, 481-504.
5. Longuet-Higgins, M. S. & Stewart, R. W. (1964). *Radiation stresses in water waves; a physical discussion, with applications.* Deep-Sea Research, 11, 529-562.
6. Rodríguez, J. B.; Rendón, C.S. V. & Torres-Freyermuth A. (2023) Modelado físico de la transformación del oleaje en arrecifes coralinos durante eventos extremos, Autor.
7. Xu, D., Hwamg, P.A., and Wu, J. (1986) *Breaking of wind-generated waves*. J. Phys.Oceanogr., 16, 2172-2178