

**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](#_Toc208744246)

[1.2 Research questions and hypotheses 2](#_Toc208744247)

[1.3 Objectives 2](#_Toc208744248)

[1.3.1 General objectives 2](#_Toc208744249)

[1.3.2 Specific objectives 2](#_Toc208744250)

[2. Methodology 3](#_Toc208744251)

[2.1 Channel Description 3](#_Toc208744252)

[2.2 Experimental design 3](#_Toc208744253)

[2.3 Test Configuration 4](#_Toc208744254)

[2.4 Instrumentation 4](#_Toc208744255)

[2.4 Wave analysis 5](#_Toc208744256)

[2.5 Statistical analysis 6](#_Toc208744257)

[2.6. Modeling with Machine Learning 6](#_Toc208744258)

[3. Results 7](#_Toc208744259)

[3.1. Wave characterization using free surface time series 7](#_Toc208744260)

[3.2. Descriptive statistics 8](#_Toc208744261)

[3.2.1 Descriptive analysis of mean wave height 8](#_Toc208744262)

[3.2.2 Descriptive analysis of mean water level 9](#_Toc208744263)

[3.3. Non-parametric tests 11](#_Toc208744264)

[3.3.1. Average wave height analysis 11](#_Toc208744265)

[3.3.2 Average water level analysis 12](#_Toc208744266)

[3.4. ML Models 12](#_Toc208744267)

[4. Discussion 14](#_Toc208744268)

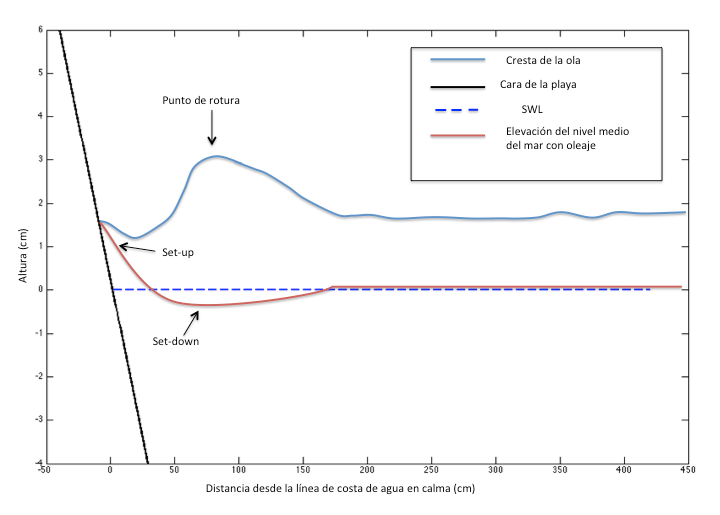
[5. Conclusions 15](#_Toc208744269)

[6. Acknowledgements 15](#_Toc208744270)

[7. References 16](#_Toc208744271)

# 1. Introduction

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

When waves break, they cause a change in the radiation stress, which represents the excess flux of momentum due to the presence of waves (Bowen et al., 1968). This change produces variations in the mean sea surface elevation (Longuet-Higgins & Stewart, 1964) when water waves encounter a sloping beach. The set-up is confined to the surf zone, landward from the initial breaking point; conversely, the set-down is a depression in the mean level that occurs before breaking, where it reaches its maximum value (see **Figure 1**).

**Figure 1.** Set up and set down caused by the waves in the surf area (Taken from García, 2016)

Laboratory experimentation allows the recreation of certain simplified real-world scenarios at reduced scale, providing a better understanding of the processes involved in the phenomenon under study, while preserving the fundamental physics of the problem (Hughes, 1993).

The measurement of the free surface is a fundamental element in the study of hydrodynamic processes, as it allows the characterization of wave behavior and its interaction with coastal structures or environments. The analysis of these signals is essential both in laboratory experimental studies and in coastal and hydraulic engineering applications, where it is critical to understand the variability of wave and water-level parameters for purposes of design, management, and prediction.

In this context, one of the main challenges lies in characterizing the spatial and temporal variability recorded by multiple sensors, which requires data handling, cleaning, and analysis techniques that ensure information quality. Likewise, the increasing availability of statistical methods and machine learning techniques offers new tools to detect patterns, test differences between scenarios, and predict relevant behaviors under controlled conditions.

## 1.2 Research questions and hypotheses

Are there significant differences in wave height values and mean water level variation between tests, and is it possible to predict these variables from Machine Learning techniques applied to resistive sensor values?

There are statistically significant differences in the values of average wave height and in the average variation of the water level between the trials, these variables have a predictive potential that can be exploited by Machine Learning models.

## 1.3 Objectives

In the development of this practice, the following general objectives and specific objectives are formulated.

### 1.3.1 General objectives

Calculate characteristic parameters of the waves such as the average height and the average variation of the water level, 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; build a homogeneous database.

Perform exploratory analyses and visualizations that characterize spatial and temporal differences.

Calculate descriptive statistics and study distributions.

Apply parametric and non-parametric tests to test hypotheses about differences between scenarios.

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

# 2. ****Methodology****

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

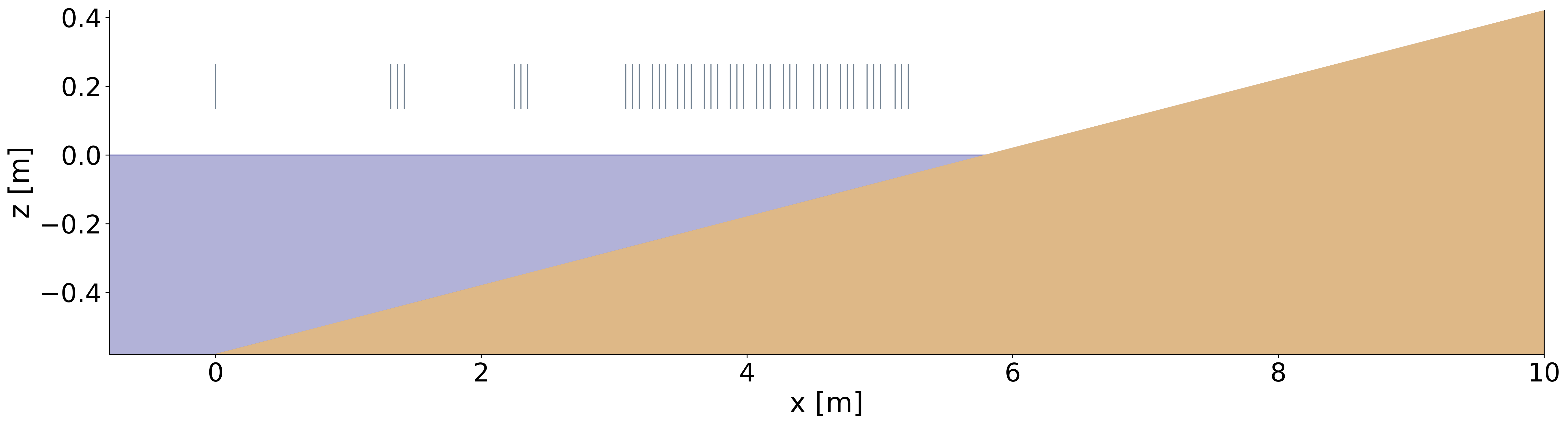
### 2.1 Channel Description

The wave channel has a length of 40 m, width of 0.8 m, and 1.27 m in height (**Figure 2**). It is equipped with a unidirectional wave generator (GOU), made up of a piston-type blade with a power of 7.5 kW and 1.2 m of stroke. This system is capable of generating regular waves, irregular second-order waves, and solitary waves (VTI, 2015). The generation system uses the AwaSys6 software, developed by Aalborg University, which is responsible for controlling the GOU. It has an active wave absorption system that suppresses the reflected waves that affect the blade (Rodríguez et al., 2023).

**Figure 2.** Wave channel diagram. (Taken from Rodríguez et al., 2023).

### 2.2 Experimental design

An idealized two-dimensional (2D) model was constructed to examine hydrodynamic processes. In the construction of the profile (**Figure 3**) a fixed bottom model was considered that represents a beach with a slope of 1:10, the dimensions of the ramp are 10 m in length, 0.8 m in width and 1.0 m in height, the water level at rest (MSL) was fixed at 0.58 m.

**Figure 3.** Schematic representation of the resistive sensor configuration for the case studies (the x-axis scale has been exaggerated for easy visualization).

### 2.3 Test Configuration

3 tests were carried out on fixed bottoms, considering a smooth surface. The experiments were carried out under regular wave conditions with different periods, as detailed in **Tables 1**. Each test lasted 480 seconds, in all cases, data acquisition began at the exact moment the wave generator was put into operation, with the water level at rest, and measurements at a frequency of 50 Hz were recorded on each sensor.

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |

**Table 1.** Case studies.

### 2.4 Instrumentation

To obtain the temporal records of the free surface of the water, level sensors were used, which act under the principle of electrical resistivity. By means of these instruments, the level of the surface of the water for the case studies was determined from the closure of the electrical circuit between two electrodes. The potential difference is recorded in voltage and is directly proportional to the water level. (Rodriguez et al., 2023)

From the voltage data measured by these sensors, elevations can be obtained by means of a calibration curve. The level sensors were calibrated at the same time for 20 seconds, making measurements at rest at six different 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 results in the conversion of the measured signals from voltage to centimeters. The elevation time series is used to determine the wave heights and periods on each sensor.

### 2.4 Wave analysis

After obtaining the time series of the elevation of the free surface, the data is processed using the Python3 programming language, in its version 3.11. This process includes filtration, statistical analysis and calculation of characteristic parameters of the waves, such as the average height and the average variation in the water level. Python, thanks to its specialized libraries such as NumPy, SciPy and Pandas, facilitates the efficient handling of large volumes of data and the development of specific algorithms for wave analysis.

**A resampling procedure was applied, reducing the sampling rate from 50 Hz to 20 Hz. To preserve the main characteristics of the signal and at the same time attenuate high-frequency noise, a Savitzky–Golay filter was used. Subsequently, the detection and correction of outliers was carried out, followed by the homogenization of the time series, in order to form a consistent and adequate database for subsequent statistical and modeling analyses.**

Applying the method of downward crossings by zeros, to the free surface series of each sensor, the parameters of each wave (period (T) and mean height ()) are calculated.

The experimental mean level () for each sensor in the case studies was determined from the values of the mean resting level () and the mean () calculated for a group of data from the recorded free surface series.

The mean resting level was calculated by means of the first 20 mean seconds in the time series, while the mean was calculated with the data recorded for the remaining 460 seconds.

### 2.5 Statistical analysis

The statistical analysis was carried out at three complementary levels. First, a descriptive analysis was performed that included the calculation of measures of central tendency, such as mean and median, as well as measures of dispersion, which encompass variance, standard deviation, interquartile range, and index of variation. This analysis was complemented with graphical representations that allow characterizing the variability of the data series.

Second, boxplots were used to identify data sprawl and detect outliers. In addition, histograms with density adjustments were used for each assay, which is useful for exploring the shape of distributions and possible asymmetries. Theoretical quantile plots (QQ-plots) were also made to contrast the observed values with a theoretical normal distribution, thus facilitating the identification of deviations from normal.

Finally, inferential statistical tests were applied, both parametric (Student's t, ANOVA and Chi-square) and non-parametric (Mann–Whitney, Wilcoxon, Kruskal–Wallis and Friedman), depending on the nature of the data and the objectives of the analysis. This comprehensive approach allows for a deeper understanding of the characteristics and relationships in the analyzed data.

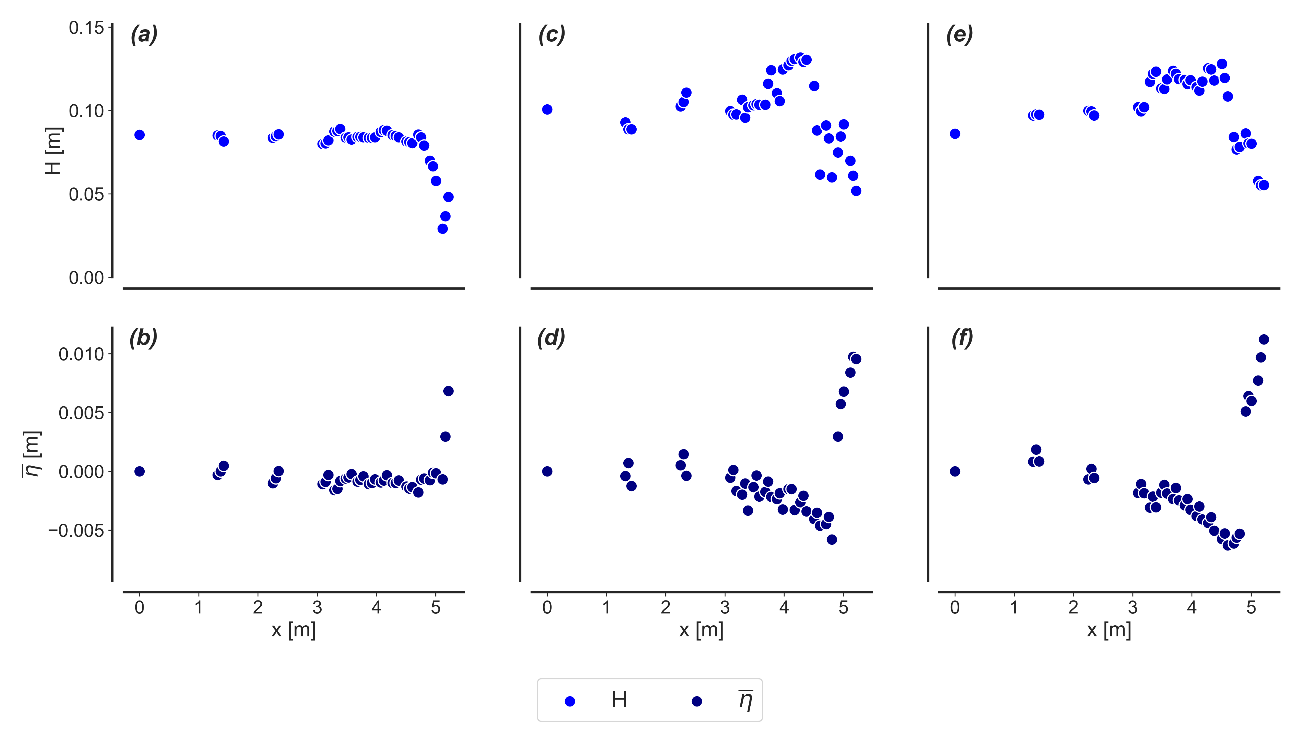
### 2.6. Modeling with Machine Learning

In the field of machine learning, two complementary approaches were considered. On the one hand, the classification consisted of the definition of wave categories (low, medium and high), for whose identification supervised algorithms of k-nearest neighbors (k-NN) and decision trees were implemented. On the other hand, regression models were applied aimed at predicting continuous variables of interest, specifically mean wave height () and mean water level (), using both linear regression and regression trees. The quality and predictive capacity of the models was evaluated using performance metrics, including accuracy and F1-score in classification, as well as RMSE and coefficient of determination (R²) in regression, also incorporating cross-validation to guarantee the robustness and generalizability of the results.

# 3. ****Results****

This chapter presents the results of an experimental study. The main objective is to evaluate the impact of these configurations on the transformation of the waves under various energy conditions. To carry out this research, tests were carried out in a wave channel, using databases generated from physical modeling. In addition, a validation and calibration process were applied to the numerical model, which allowed achieving a high spatial resolution in the results obtained.

### 3.1. Wave characterization using free surface time series

The **Figure 4** presents the variation in wave height and the average variations in mean water level over the periods of regular waves studied. Subfigures (a), (c) and (e) compare the evolution of wave height, where an initial increase followed by a progressive dissipation towards the coast can be seen.

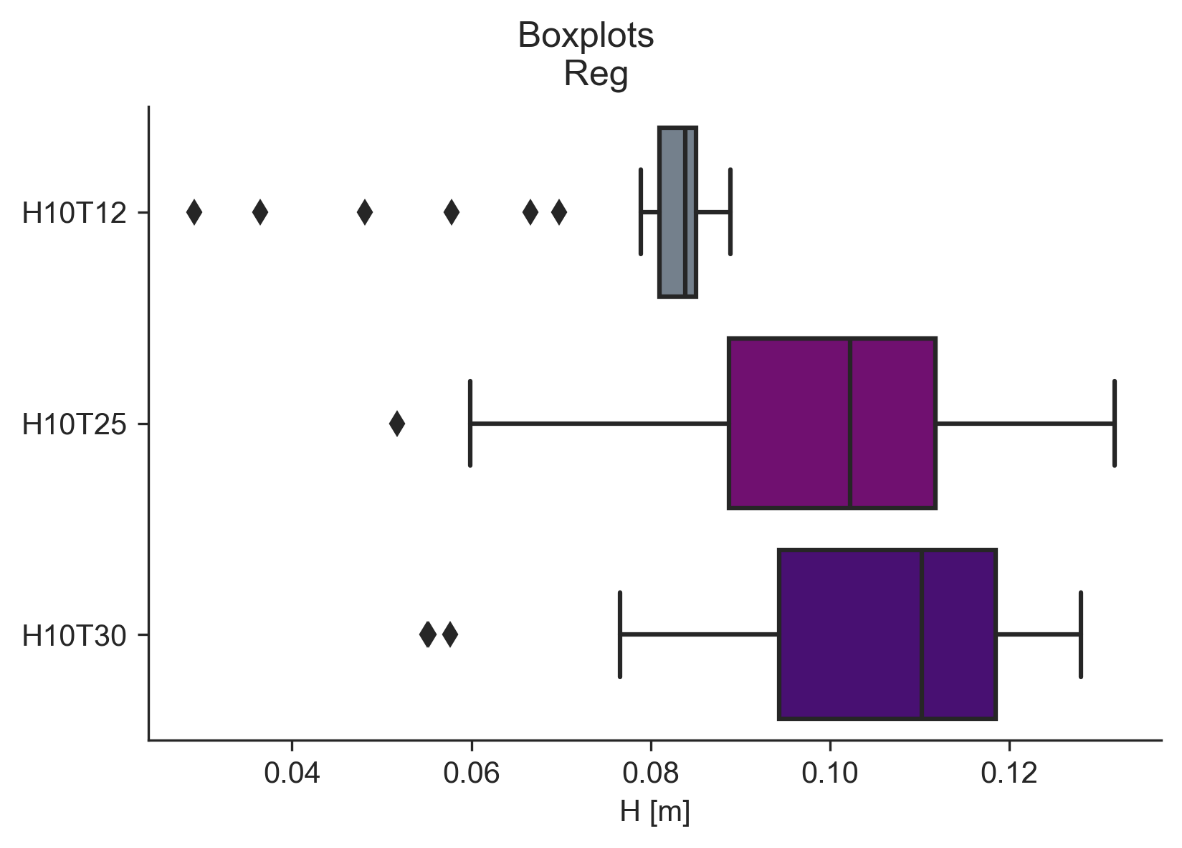
**Figure 4.** Results of the tests with regular waves for the cases corresponding to T = 1.2 s ((a) – (b)), T = 2.5 s ((c) – (d)) and T = 3.0 s ((e) and (f)).

The variation in the mean water level, represented in subfigures (b), (d) and (f), reveals a maximum depression (set down) at the breaking point and a gradual increase (set up) in the posterior region. The average 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 to the foot of the beach.

### 3.2. Descriptive statistics

The **following figures and tables** present the descriptive analysis of mean wave height () and mean water level () for the three trials studied (H10T12, H10T25 and H10T30).

#### 3.2.1 Descriptive analysis of mean wave height

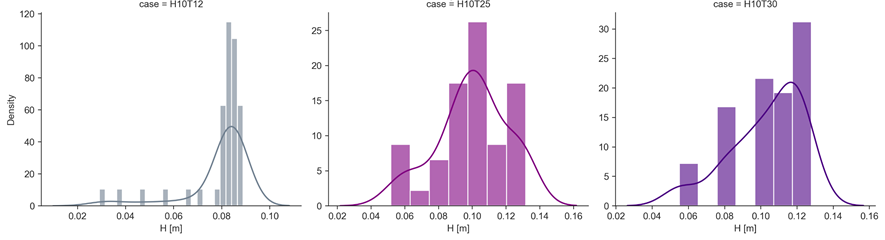
The **boxplots (Figure 5)** allow us to identify differences in dispersion and in the presence of outliers in all three cases. Greater variability is observed in H10T25 and H10T30 cases, while H10T12 has a more concentrated distribution. These results suggest that, under longer-period conditions, the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

**Figure 5.** Boxplots corresponding to the mean wave height () in three tests: H10T12, H10T25 and H10T30.

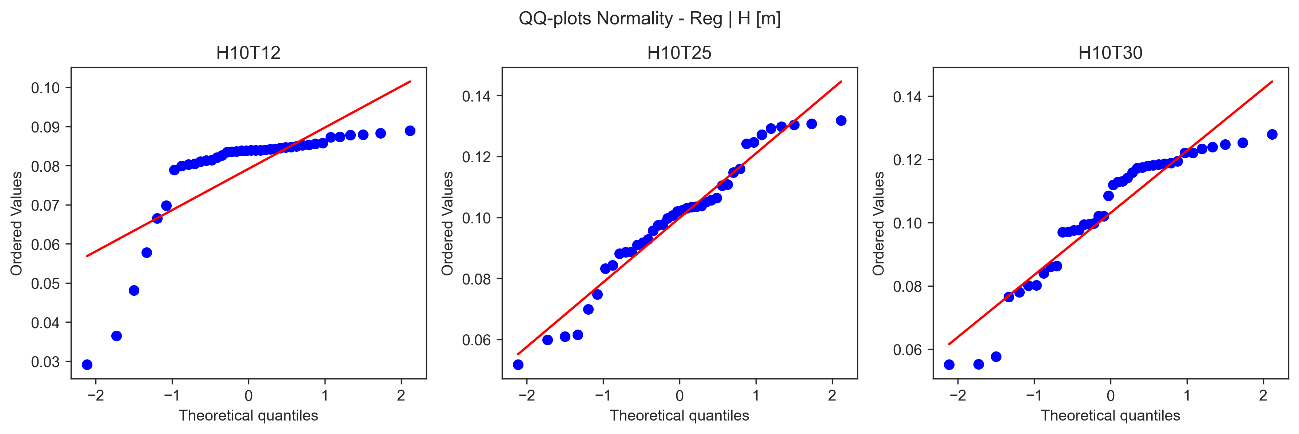
The mean wave height showed a progressive increase between the three trials considered (**Table 2**), going from average values of 0.079 m in H10Q12 to 0.099 m in H10T25 and reaching 0.103 m in H10T30. This increase was accompanied by a greater relative dispersion, reflected in coefficients of variation 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 values ().

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 |

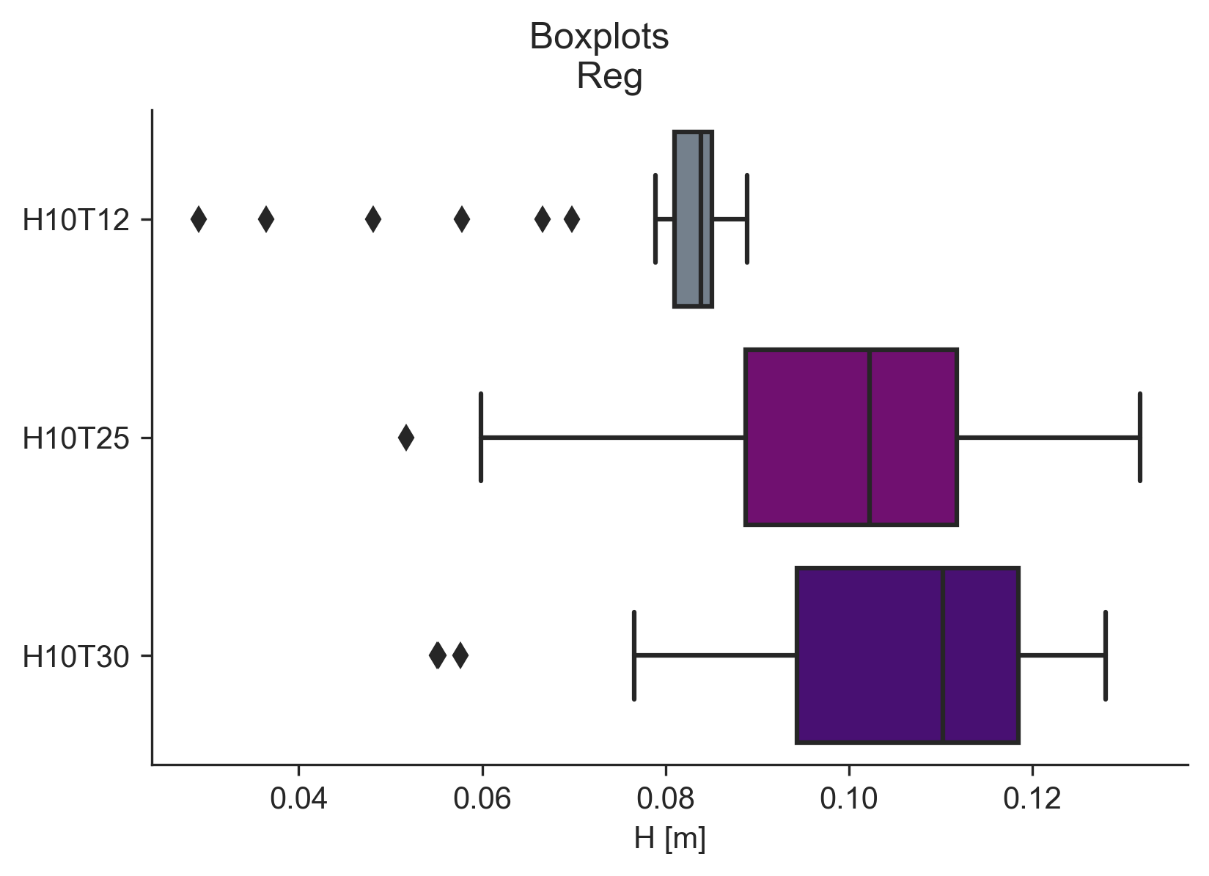
Regarding the shape of the distributions, the **histograms with density adjustment (Figure 6)** showed that in H10T12 the concentration of values is higher around the mean, with marked asymmetry, while in H10T25 and H10T30 the distributions are more widespread and closer to a unimodal shape. However, the **QQ-plots (Figure 7)** revealed that the normality hypothesis is not fully fulfilled in any of the trials: in H10T12 the deviations are noticeable at the extremes, while in H10T25 and H10T30 the alignment with the theoretical line is closer, although discrepancies persist in the tails.

**Figure 6.** Analysis of distributions using density-adjusted histograms corresponding to mean wave height () in three trials: H10T12, H10T25 and H10T30.

**Figure 7.**: **QQ-plots** corresponding to the mean wave height () in three tests: H10T12, H10T25 and H10T30.

#### 3.2.2 Descriptive analysis of mean water level

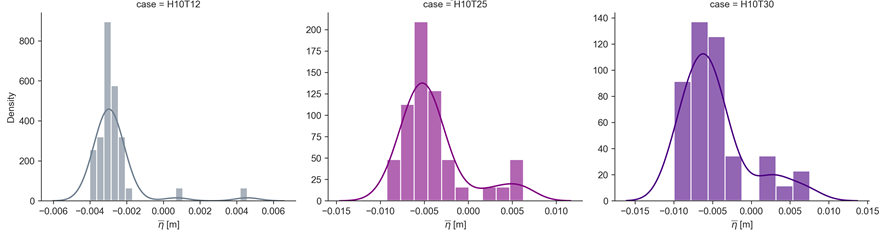
The **boxplots** (**Figure 8**) showed differences in dispersion and in the presence of outliers between cases: while in H10T12 the variability is reduced and outliers are scarce, in H10T25 and with greater emphasis on H10T30 an increase in dispersion and a greater frequency of extreme values are observed, which suggests a more unstable behavior of the free surface under conditions of greater wave intensity.

****The **table of descriptive statistics (Table 3)** complements this observation, showing that both the mean and the median have negative values in all trials, with more accentuated magnitudes in H10T25 and H10T30. This reflects a progressive average decrease in the water level as the intensity of the test increases. Likewise, the variance and standard deviation exhibit an increasing pattern from H10T12 to H10T30, confirming the trend towards greater dispersion in the most energetic experiments.

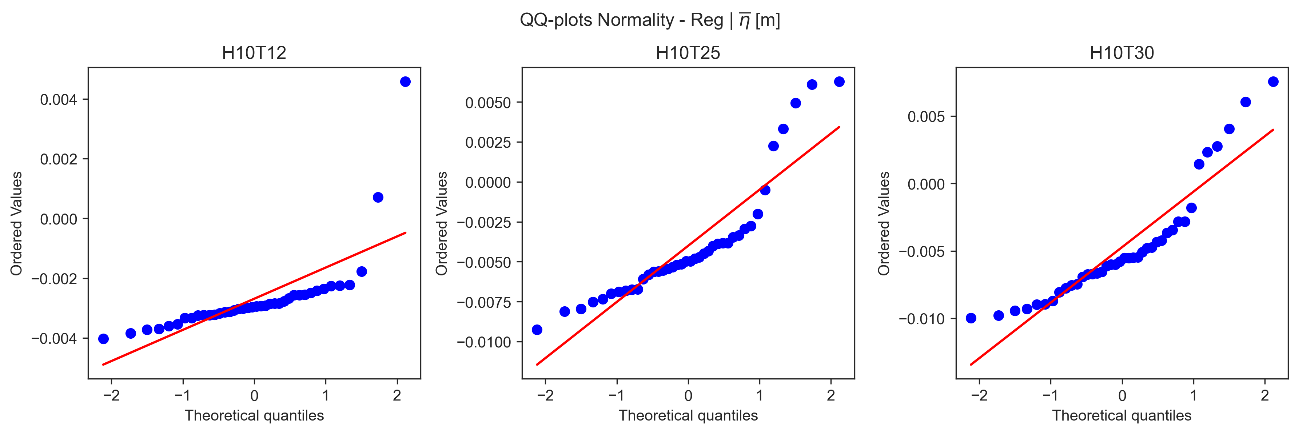
**Figure 8.** Boxplots corresponding to the mean variation of the water level () in three tests: H10T12, H10T25 and H10T30.

**Table 3.** Descriptive statistics for mean water level variation values ().

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 |

The analysis of the **distributions** using histograms with density adjustment (**Figure 9**) reveals differentiated behaviors: in H10T12 the data are concentrated around values close to zero, while in H10T25 and H10T30 a wider and asymmetrical distribution is observed, with more pronounced tails towards negative and positive values. Finally, the **QQ-plots (Figure 10)** indicate that in all three cases the empirical distribution deviates from the theoretical normal, especially at the extremes of H10T25 and H10T30, which confirms the presence of heavy tails and a departure from the assumption of normality.

**Figure 9.** Analysis of distributions using density-adjusted histograms corresponding to mean water level variation () in three assays: H10T12, H10T25 and H10T30.

**Figure 10. QQ-plots** corresponding to the mean variation of the water level () in three tests: H10T12, H10T25 and H10T30.

### 3.3. Non-parametric tests

Non-parametric methods were adopted because the series presented asymmetries, outliers and deviations from normal, which limits the validity of the parametric tests. These tests allow us to contrast differences between scenarios in a robust way, without requiring strict distribution assumptions.

#### 3.3.1. Average wave height analysis

The non-parametric analysis (See **Table 4**) using the Friedman test confirms the existence of significant differences in the mean wave height between the three scenarios evaluated. Wilcoxon's post-hoc contrasts indicate that the differences are mainly concentrated between the H10T12 assay and the H10T25 and H10T30 assays, with highly significant p-values. On the other hand, no differences were detected between H10T25 and H10T30, suggesting that from a certain period threshold (≥ 2.5 s) the average wave height tends to stabilize.

**Table 4.** Mean wave height variation values ().

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 Average water level analysis

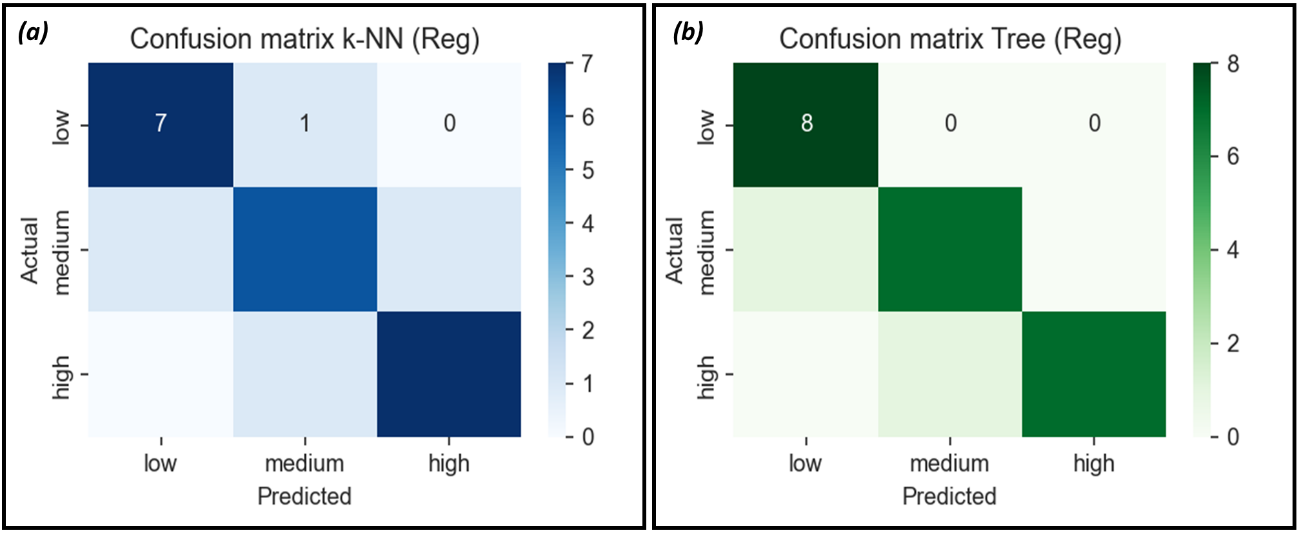
The Friedman test showed significant differences in mean water level variation between the three scenarios, confirming a systematic effect of the period on level shift. Wilcoxon's post-hoc contrasts revealed that all test pairs differ significantly, indicating that both the transition from H10T12 to H10T25 and from H10T25 to H10T30 generate statistically relevant changes. This pattern suggests a progressive and consistent trend of variation in the level with increasing periods, with no evidence of stabilization between the scenarios evaluated.

**Table 5.** Values of average change in the water level ().

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 |

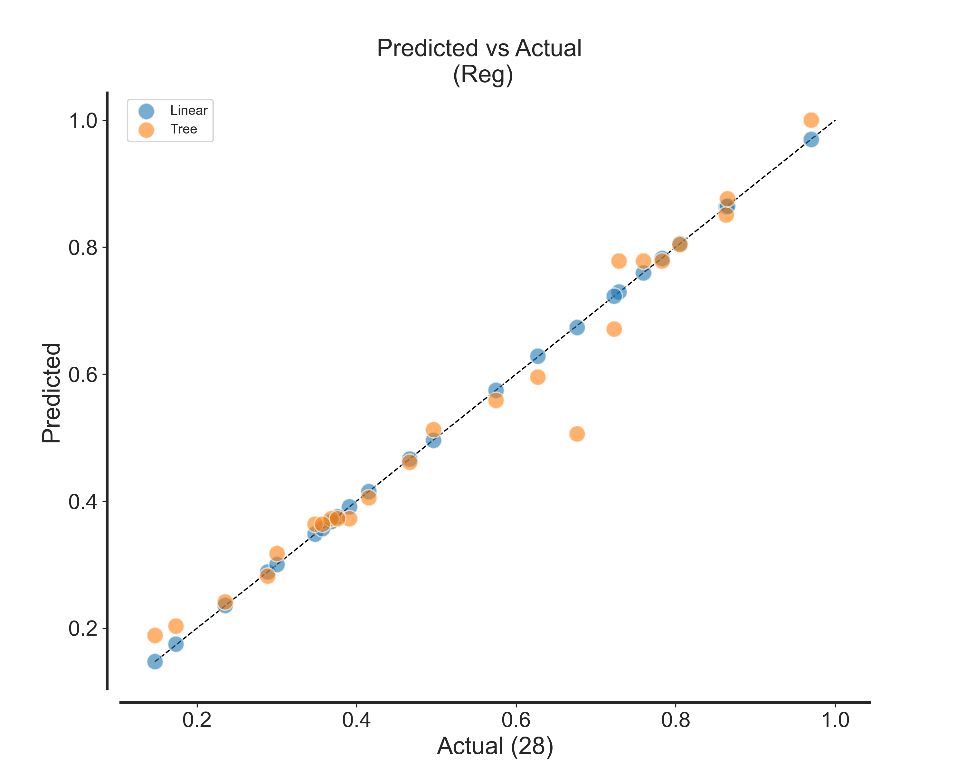
### 3.4. ML Models

In the case of regular waves (H10T12, H10T20 and H10T30), feature vectors were constructed per window from basic statistical parameters of the signal: mean wave height, mean absolute variation, standard deviation, skewness and kurtosis. With these variables, classifiers were trained to identify the corresponding case of waves.

The results show that the k-NN model (**Figure 11** (a)) reached an accuracy of 0.833 and an f1 value of 0.833, presenting some classification errors between medium and low levels. On the other hand, the decision tree (**Figure 11** (b)) obtained a superior performance, with a precision of 0.917 and an f1 of 0.917, correctly classifying most cases and with less confusion between classes. This indicates that decision trees are more effective than k-NN at discriminating between regular wave conditions.

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

Regarding regression (**Figure 12**), wave height prediction in a target sensor was evaluated using the characteristics of neighboring sensors or previous conditions as predictors. Linear regression showed a near-perfect fit (R² = 1.000, RMSE ≈ 0.0008), with predicted values practically on the diagonal in the Predicted vs Actual graph. In contrast, the regression tree obtained an R² of 0.969 and an RMSE of 0.0413, which reflects an adequate performance, but with greater dispersion around the diagonal, especially in intermediate values.

**Figure 12.** Linear Regression for regular waves.

# 4. ****Discussion****

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

By complementing these findings with the Machine Learning approach, it was observed that classifiers are able to discriminate between regular wave conditions based on simple statistical variables. Although the k-NN model achieved acceptable performance, the decision tree showed greater classification capacity, reducing confusion between wave levels. In the case of regression, the linear regression achieved a practically perfect fit, while the regression tree, although precise, showed greater dispersion. Taken together, the results confirm that both traditional statistical methods and machine learning approaches provide complementary and robust information to characterize and predict wave behavior under controlled conditions.

# 5. ****Conclusions****

* The results show that, under regular waves, the wave height increases to the break zone and then progressively dissipates towards the coast, while the average water level presents a maximum set down at the breaking point followed by a gradual set up to the foot of the beach.
* The average wave height showed a progressive increase between the trials, accompanied by greater dispersion and the presence of extreme values.
* The distributions showed asymmetries between the mean height values and the QQ-plots confirmed deviations from normality, which highlights the convenience of applying non-parametric or robust statistical methods in the comparison between cases.
* The mean variation in water level showed a progressive decrease from T12 to T30, accompanied by asymmetric distributions and the presence of outliers.
* The results show that the increase in the period induces systematic displacements from the mean level, with deviations from normality that justify the use of non-parametric tests in the comparative analysis.
* The average wave height only differs significantly between H10T12 and the other cases, while the average water level shows differences in all scenarios, which shows its greater sensitivity to the wave period.
* The results show that decision trees are more effective in classifying regular wave scenarios, while linear regression offers the best accuracy in predicting wave heights on a target sensor.

# 6. Acknowledgements

The staff of the Coastal Engineering and Processes Laboratory (LIPC) of the Institute of Engineering of the National Autonomous University of Mexico (UNAM) is thanked, in particular the Ocean. Camilo Sergio Rendón Valdez, for the support provided during the performance of the experiments and in the acquisition of the data. Likewise, Dr. Alec Torres Freyermuth is recognized for his leadership in the execution of the trials, as well as Dr. Gabriela Medellín Mayoral and Dr. José Carlos Pintado Patiño, members of the committee, for their valuable guidance and recommendations throughout 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