

Article

# Resilient Coastal Protection Infrastructures: Probabilistic Sensitivity Analysis of Wave Overtopping Using Gaussian Process Surrogate Models

Paul Kent <sup>1</sup>, Soroush Abolfathi <sup>2</sup>, Hannah Al Ali <sup>3</sup>, Tabassom Sedighi <sup>4</sup>, Omid Chatrabgoun <sup>5</sup>  
and Alireza Daneshkhah <sup>3,\*</sup>

<sup>1</sup> Department of Computer Science, University of Exeter, Exeter EX1 2LU, UK; p.u.kent@exeter.ac.uk

<sup>2</sup> School of Engineering, The University of Warwick, Coventry CV4 7AL, UK; soroush.abolfathi@warwick.ac.uk

<sup>3</sup> Faculty of Mathematics and Data Science, Emirates Aviation University, Dubai P.O. Box 53044, United Arab Emirates; hannah.alali@emirates.com

<sup>4</sup> International Policing and Public Protection Research Institute (IPPPRI), Anglia Ruskin University, Cambridge CB1 1PT, UK; tabassom.sedighi@aru.ac.uk

<sup>5</sup> School of Computing, Mathematics and Data Science, Coventry University, Coventry CV1 5FB, UK; omid.chatrabgoun@coventry.ac.uk

\* Correspondence: alireza.daneshkhah@emirates.com

**Abstract:** This paper presents a novel mathematical framework for assessing and predicting the resilience of critical coastal infrastructures against wave overtopping hazards and extreme climatic events. A probabilistic sensitivity analysis model is developed to evaluate the relative influence of hydrodynamic, geomorphological, and structural factors contributing to wave overtopping dynamics. Additionally, a stochastic Gaussian process (GP) model is introduced to predict the mean overtopping discharge from coastal defences. Both the sensitivity analysis and the predictive models are validated using a large homogeneous dataset comprising 163 laboratory and field-scale tests. Statistical evaluations demonstrate the superior performance of the GPs in identifying key parameters driving wave overtopping and predicting mean discharge rates, outperforming existing regression-based formulae. The proposed model offers a robust predictive tool for assessing the performance of critical coastal protection infrastructures under various climate scenarios.

**Keywords:** climate resilience; coastal flooding; Gaussian processes; probabilistic sensitivity analysis; wave overtopping; coastal defence



**Citation:** Kent, P.; Abolfathi, S.; Al Ali, H.; Sedighi, T.; Chatrabgoun, O.; Daneshkhah, A. Resilient Coastal Protection Infrastructures: Probabilistic Sensitivity Analysis of Wave Overtopping Using Gaussian Process Surrogate Models. *Sustainability* **2024**, *16*, 9110. <https://doi.org/10.3390/su16209110>

Academic Editor: Miklas Scholz

Received: 17 May 2024

Revised: 12 September 2024

Accepted: 9 October 2024

Published: 21 October 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Coastal regions are densely populated, with over 2.4 billion people (40% of the world's population) living within 100 km of the coast [1]. More than 600 million people (10% of the world's population) live in coastal areas less than 10 meters above sea level, making them highly vulnerable to flooding and inundation during extreme climatic events. Over the past century, global sea levels have risen at an accelerating rate, with projections suggesting a rise of over a meter by 2100. Such increases contribute to a host of adverse outcomes, including intensified erosion, storm surge flooding, inundation, contamination of freshwater, and the loss of coastal lowlands and wetlands [2,3]. The rise in sea level and associated risks affect two-thirds of the world's 23 most densely populated cities [4]. In recent years, coastal flooding triggered by extreme events has had devastating socioeconomic impacts. The scientific prediction for a high-emission climate change scenario (+2 °C temperature rise) estimates an annual \$14 trillion cost of flooding worldwide [5], with significant damage to critical infrastructures.

Coastal defences play a vital role in protecting coasts from flooding and erosion. The effective and optimal design of coastal protection structures are important for safeguarding coastal communities from storm surges and extreme climatic events [6]. In the past decades,

several types of flood protection schemes, including hard-engineered, soft nature-based, and hybrid solutions, have been developed and tested [7]. Hard defences, in particular, vary widely in terms of shape, structural design, and materials. Extensive research has been conducted to evaluate the performance of these defences under changing hydrodynamic, geomorphological, and structural conditions. Wave overtopping is a crucial design parameter for coastal defences as it directly affects both the structural integrity of the defence and the level of protection it provides. Given the direct link between wave overtopping and the vulnerability of coastal areas to flooding, it is vital to develop robust methods for evaluating and predicting wave overtopping under different hydrodynamic and structural conditions. However, accurately predicting the performance of critical coastal infrastructure remains a significant challenge. This difficulty stems from the complexity of nearshore processes, which operate at varying spatial and temporal scales. It is further compounded by the non-linear interactions between hydrodynamics and geomorphology, the complex geometry of nearshore areas and coastal defences, and the uncertainties associated with the incident wave climate. Furthermore, a lack of comprehensive data and field observations exacerbates these challenges [8,9].

The interaction between nearshore wave dynamics and defence structures leads to complex hydrodynamic responses, which are influenced by factors such as water depth  $h$ , incident wave height  $H$ , and period  $T$ . Previous studies undertook extensive laboratory physical modelling and field measurements to investigate the effects of wave structure interactions on the structural response to wave overtopping. The majority of these investigations led to empirical formulae for overtopping; however, these formulae are typically valid only within specific hydrodynamic ranges or for particular structural configurations. In recent decades, efforts have been made to develop integrated tide-surge and wave flood models, often using RANS-VOF models and shallow water equations [10,11]. Despite their utility, these models are limited by their depth- and time-averaged nature of flow solvers and the inherent simplifications in turbulence closure models. More recently, Lagrangian particle-based models have been employed to simulate wave interactions with coastal infrastructures and overtopping processes ([12,13]). These models are advantageous because they account for depth-varying hydrodynamics in wave–structure interactions. However, they are often case-specific, computationally expensive, and time-consuming. To address the challenges associated with understanding wave overtopping processes from laboratory and numerical models, machine learning approaches have gained traction as a reliable method for predicting wave run-up and overtopping at coastal defences [6,14]. Machine learning techniques have shown superior predictive capabilities compared with traditional overtopping formulae based on a regression analysis of physical model data. Nonetheless, existing predictive tools for wave overtopping are generally limited by their focus on specific geometries and hydrodynamic conditions. There remains a critical need to develop a more comprehensive predictive framework capable of providing robust overtopping estimates across a wide range of defence structures and hydro-meteorological conditions.

We now explore the complex processes driving overtopping. A thorough understanding of these processes will provide insights for selecting input parameters to develop an efficient model and improving predictive capabilities.

## 2. Wave Overtopping Processes

Overtopping is a complex, multi-faceted phenomenon influenced by the interaction between nearshore wave processes, seabed topography, and the configuration of coastal defence structures. The complicated interactions between the incident wave and coastal defences can generate overtopping flows that are sudden and do not follow the typical wave up-rush (run-up) processes. To fully understand wave–structure interactions, it is necessary to study the nearshore wave processes. Waves approaching coastal protection structures can be classified into ‘breaking’ or ‘non-breaking’ waves. Extensive research has been conducted to explore the distinct interactions between these wave types and coastal defences (e.g., see [15,16]). Although these two terms may not be entirely precise, they are

widely used in the literature to characterise wave overtopping behaviours. This section provides an overview of the key concepts related to wave overtopping.

Wave breaking on beaches and gently sloping structures is commonly characterised by the Iribarren number, also known as the surf similarity parameter (shown in Equation (1)). The Iribarren number defines four distinct breaking regimes, including spilling ( $\xi_{op} < 0.4$ ), plunging ( $0.4 < \xi_{op} < 2.3$ ), collapsing ( $2.3 < \xi_{op} < 3.2$ ), and surging ( $\xi_{op} > 3.2$ ). These regimes are frequently used in design manuals [17] to assess armour stability for sloped structures.

$$\xi_{op} = \frac{\tan(\alpha)}{(S_{op})^{0.5}} \quad (1)$$

where  $\alpha$  is the beach slope and  $S_{op}$  is the wave steepness.

For vertical walls and steep structures, the impulsive and non-impulsive breaking is defined by the slope and the incident wave's wavelength. Besley et al. [18] introduced the wave breaking parameter,  $h_*$ , which is based on the depth at the toe of the wall ( $h_s$ ) and nearshore incident wave conditions (Equation (2)).

$$h_* = \frac{h_s}{H_{si}} \left( \frac{2\pi h_s}{g T_m^2} \right) \quad (2)$$

where  $H_{si}$  is the inshore significant wave height and  $T_m$  is the averaged wave period determined from spectral moments or a zero-crossing analysis. Besley et al. [18] suggested that impulsive wave conditions occur at the wall when  $h_* \leq 0.3$  and that pulsating conditions occur when  $h_* > 0.3$ .

Wave overtopping rates must remain below tolerable thresholds under design and operational conditions to ensure the safety of people and property on or behind coastal defence structures [19–21]. The volume of overtopping is significantly influenced by the nature of wave interactions with the defence structure. When waves break onto or over the structure, overtopping tends to produce relatively continuous volumes of water, known as 'green water'. In contrast, when waves break seaward of the structure, overtopping occurs in the form of fine droplets—a phenomenon referred to as 'splash overtopping'. This type of overtopping is typically carried by the wave's momentum or enhanced by onshore winds [22]. Onshore winds can play a critical role in intensifying overtopping, especially when reflected waves (e.g., from steep walls) interact with incoming waves, creating local clapotis effects. Although onshore wind significantly influences spray overtopping, particularly from vertical walls, limited data are quantifying its impact. Research indicates that while onshore wind has minimal effect on 'green overtopping', it can amplify spray overtopping from vertical walls by a factor of up to three [23].

### 3. Database

The overtopping data utilised in this study are derived from both two-dimensional and three-dimensional physical modelling tests, as well as full-scale field-based (prototype) measurements reported in the CLASH dataset [24]. This comprehensive dataset comprises 163 test series with a total of 10,532 data points and 31 parameters. The dataset includes overtopping records for a range of hydrodynamics and structural configurations, covering both small-scale and large-scale laboratory studies, as well as field measurements. Table 1 summarises the parameters recorded in the dataset. The homogeneous database includes measurements from various types of hard-engineered coastal protection structures, including vertical structures, rubble mound breakwaters (with rock or concrete armour), dikes, berm breakwaters, and composite structures. Notably, the dataset also features tests with zero overtopping, making it particularly valuable for developing predictive models. De Rouck et al. [25] compared the dataset to empirical overtopping prediction formulae for vertical structures [22], sloping structures ( $\gamma_f = 0.5$ , as suggested in [26]), and dikes [26]. This comparison aimed to identify outliers within the dataset. A general form of the overtopping formulae proposed in the EurOtop manual is presented in Equation (3) [27].

$$\frac{q}{(gH_{m0}^3)^{0.5}} = A \exp\left(-\frac{BR_c}{\gamma H_{m0}}\right) \quad (3)$$

where  $H_{m0}$  denotes the significant wave height based on spectral analysis,  $R_c$  is the structure crest freeboard relative to still water line,  $\gamma$  is a correction factor for the roughness and angle of wave approach or structure geometry, and  $A$  and  $B$  are constants (fitting coefficients) that are determined for different types of structures based on a regression analysis.

**Table 1.** Summary of dataset parameters used in this study and the experimental range.

Parameter	Range	Parameter	Range	Description
Structural parameters		Hydrodynamic parameters		
$h_{deep}$ [m]	(0, 100)	$H_{m0\ deep}$ [m]	(0.003, 5.920)	$= 4\sqrt{m_0}$
$m$ [-]	(6, 1000)	$T_p\ deep$ [s]	[0.545, 15]	
$h$ [m]	(0.029, 9.32)	$T_m\ deep$ [s]	(0.454, 12.5)	$= m_2/m_0$
$h_t$ [m]	(0.025, 7.78)	$T_{m-1,0\ deep}$ [s]	(0.495, 13.636)	$= m_{-1}/m_0$
$B_t$ [m]	(0, 10)	$\beta^{[0]}$	(0, 80)	
$\gamma_f$ [-]	(0.35, 1)	$H_{m0\ toe}$ [m]	(0.003, 3.8)	$= 4\sqrt{m_0}$
$\cot\alpha_d$ [-]	(0, 7)	$T_p\ toe$ [s]	(0.545, 16.4)	
$\cot\alpha_d$ [-]	(-5, 9.706)	$T_m\ toe$ [s]	(0.454, 11.881)	$= m_2/m_0$
$\cot\alpha_{excl}$ [-]	(-1.533, 8.144)	$T_{m-1,0\ toe}$ [s]	(0.495, 10.64)	$= m_{-1}/m_0$
$\cot\alpha_{incl}$ [-]	(-1.533, 12.821)	$q$ [m <sup>3</sup> /s.m]	(0, $1.65 \times 10^{-1}$ )	
$R_c$ [m]	(0, 8.345)	$P_{ow}$ [-]	(0, 81)	
$B$ [m]	(0, 8)	General parameters		
$h_b$ [m]	(-0.208, 1.175)	RF [-]	(1, 4)	
$\tan\alpha_B$ [-]	(0, 0.125)	CF [-]	(1, 4)	
$B_h$ [m]	(0, 8)			
$A_c$ [m]	(0, 7.87)			
$G_c$ [m]	(0, 5.6)			

All data used in this study are classified with a reliability factor (RF) and complexity factor (CF), based on the screening study conducted in [28]. Table 1 presents the range of RF and CF for the CLASH dataset. Tests with an RF/CF of four are deemed unreliable and are excluded from the modelling process in this study. Figure 1 shows a schematic of the hydrodynamic and structural parameters considered in this study for the probabilistic sensitivity analysis and for developing a predictive model. To present a holistic overview of the data used for this study and the range they cover, Figure 2 plots dimensionless wave overtopping rates measured against a relative crest freeboard. The dataset effectively covers the range of  $10^{-6} \leq q \leq 10^{-1}$  and  $0.3 \leq R_c/H_{m0} \leq 3.5$ . Notably, outliers are present, manifested as large, unexpected overtopping values. These outliers are primarily attributed to experimental conditions, such as very small wave steepness and shallow foreshores. Additionally, the unrealistically low wave overtopping discharges observed for cases with small relative freeboards are associated with tests involving high and wide-crested rubble mound armour.

The range of wave steepness as a function of the wave height was determined for all the data and is presented in Figure 3. Wave steepnesses ( $S_{op}$ ) exceeding 0.07 are physically unrealistic, while values lower than 0.005 are challenging to generate reliably in controlled conditions. Therefore, tests with  $S_{op} > 0.07$  or  $S_{op} < 0.005$  are considered unreliable and are excluded from further analysis in this study.

Figure 4 illustrates the combination of upper and down slopes for structures with and without berms. The majority of the tests involve uniform sloping structures. A negative upper slope corresponds to structures with a large return wall, while data with  $\cot(\alpha_{up}) = 0$  represent structures with a vertical upper section and a sloping lower section. Data points along the vertical axis correspond to structures with a vertical down slope and sloping upper section. Further analysis of the dataset reveals that most data points fall within the range of  $0 < \text{width}/H_{m0,toe} < 10$  and  $-5 < \text{level}/H_{m0,toe} < 5$ .

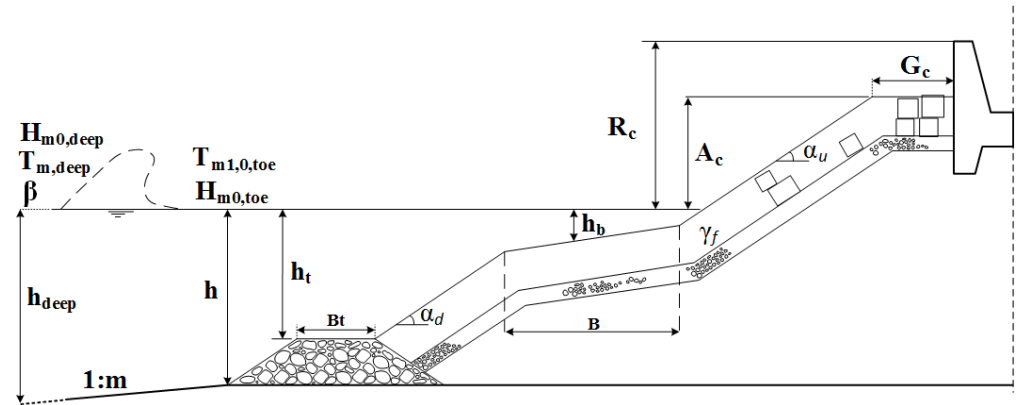


Figure 1. Schematic description of parameters recorded in the CLASH overtopping database (adopted from [28]).

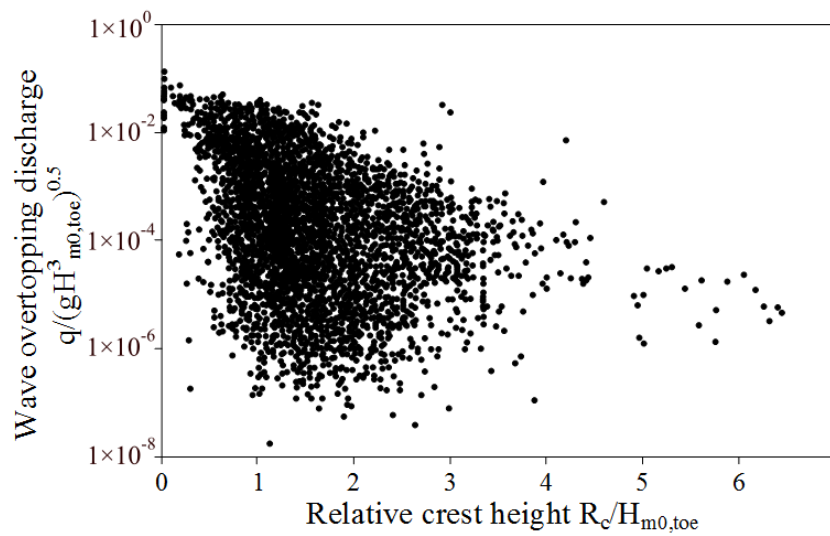


Figure 2. Dimensionless overtopping records for a range of relative crest heights tested within the CLASH dataset (adopted from [28]).

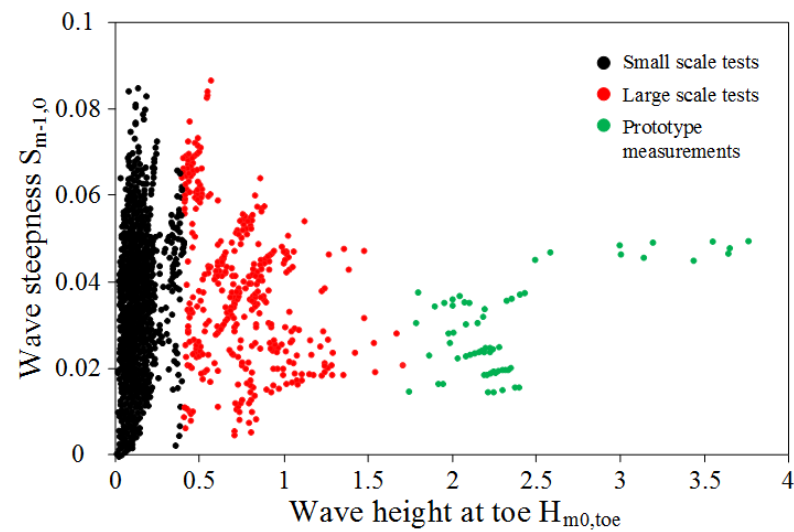


Figure 3. Range of wave steepness as a function of the wave height for all data (adopted from [28]).

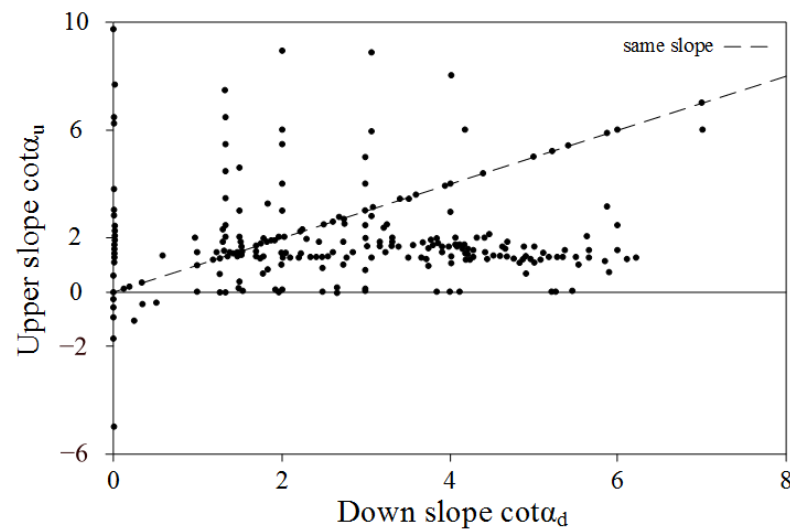


Figure 4. Upper slope  $\cot \alpha_u$  versus down slope  $\cot \alpha_d$  (adopted from [28]).

#### 4. Method

This section outlines the use of Gaussian process regression (GPR) to probabilistically model the complex relationships between wave overtopping ( $q$ ) and the input parameters listed in Table 1. Before introducing the GPR approach, the variance-based and emulator-based methods are described as efficient probabilistic sensitivity analysis (SA) techniques. These methods are used to determine the relative importance and influence of each input parameter on wave overtopping. By building a model that focuses on the most influential input parameters, we aim to reduce model complexity, prevent overfitting, and improve predictive accuracy.

##### 4.1. Probabilistic Sensitivity Analysis

In this paper, a global SA of the model output is performed, which evaluates the relative importance of input parameters when they are varied extensively, accounting for their uncertainties over a broad range. One approach to global SA is the analysis of variance of the model response originally proposed by [29]. This method identifies the contribution of individual inputs or groups of inputs to the overall variance in the model's output. It also assesses the total effect of each input on output variance, including both its marginal influence and its interaction with other inputs. Several computational techniques can be applied to carry out this SA, as detailed in [29–31]. This study adopts the emulator-based method outlined by [32] to determine the sensitivity measures.

To conduct the SA, we consider how a function  $f(\mathbf{x})$  depends on its input variables. In this study,  $f$  typically represents the function that computes wave overtopping as a function of a vector of input parameters illustrated in Table 1. Some key notations are introduced below. We define a  $d$ -dimensional random vector as  $\mathbf{X} = (X_1, \dots, X_d)$ , where  $X_i$  denotes the  $i$ th element of  $\mathbf{X}$ . The sub-vector  $(X_i, X_j)$  is represented as  $\mathbf{X}_{i,j}$ . More generally, if  $p$  represents a set of indices, then  $\mathbf{X}_p$  indicates the sub-vector of  $\mathbf{X}$  comprising elements with those indices.  $\mathbf{X}_{-i}$  is defined as the sub-vector of  $\mathbf{X}$  containing all elements except  $x_i$ . Similarly,  $\mathbf{x} = (x_1, \dots, x_d)$  represents the corresponding observed random vector  $\mathbf{X}$ . In this context,  $\mathbf{X}$  serves as an input vector consisting of all input parameters outlined in Table 1, while  $q$  (wave overtopping) is regarded as the output variable and is denoted by  $Y$ .

##### 4.1.1. Function Decomposition for Main Effects and Interactions

Sobol [29] demonstrates that any function  $f(\cdot)$  with quadratic integrability can be expressed through its main effects (MEs) and interactions, as follows:

$$y = f(\mathbf{x}) = z_0 + \sum_{i=1}^d z_i(x_i) + \sum_{i < j} z_{i,j}(\mathbf{x}_{i,j}) + \dots + z_{1,2,\dots,d}(\mathbf{x}). \quad (4)$$

In the context of a relationship between  $y$  and  $\mathbf{x}$ , expressed by  $y = f(\mathbf{x})$  (where  $f(\cdot)$  represents a function involving uncertain quantities  $\mathbf{x}$ ), the expected value of which is represented by  $z_0 = E[f(\mathbf{X})]$ . The function  $z_i(x_i)$ , as seen in Equation (5), stands for the “main effect” of the  $i^{\text{th}}$  variable,  $x_i$ , and is expressed as:

$$z_i(x_i) = E[f(\mathbf{X}) | x_i] - E[f(\mathbf{X})]. \quad (5)$$

The ME,  $z_i(x_i)$ , is the function solely dependent on  $x_i$  that provides the optimal approximation of  $f(\cdot)$  by minimising the variance when averaged over all other variables [32,33].

The function  $z_{i,j}(x_{i,j})$ , as presented in Equation (6), characterises the first-order interaction between variables  $x_i$  and  $x_j$ . Likewise,  $z_{i,j,k}(x_{i,j,k})$  represents the second-order interaction among  $x_i$ ,  $x_j$ , and  $x_k$ , with this pattern extending to higher-order interactions.

$$z_{i,j}(x_{i,j}) = E[f(\mathbf{X}) | \mathbf{x}_{i,j}] - z_i(x_i) - z_j(x_j) - E[f(\mathbf{X})]. \quad (6)$$

In the context of this study, the main effect (ME) represents the expected change in wave overtopping when input parameter  $i$  takes a specific value  $x_i$ , while the uncertainty in the remaining parameters is still accounted for.

It is common for the functions in Sobol’s decomposition to be pairwise orthogonal. The definitions of the main effects and interaction terms in Sobol’s decomposition, as given in Equation (4), depend on the distribution of the input parameters,  $\mathbf{X}$  denoted by  $G$  (further details are discussed in Section 4.2.2).

The sensitivity metrics evaluated in this study help identify which input parameters in  $\mathbf{X}$  most significantly contribute to the uncertainty in  $f(\cdot)$ . Examining the MEs and, where applicable, first-order interaction terms provides valuable insights. Their corresponding plots offer a useful visual tool for understanding how individual inputs affect the model’s output and how these inputs interact to influence the overall behaviour of the model.

#### 4.1.2. Variance-Based Methods

Variance-based methods assess the sensitivity of the output, in this case, the overtopping volume,  $Y = f(\mathbf{X})$ , by analysing how changes in model input parameters affect the variance of  $Y$ . A comprehensive review of this approach can be found in [30]. Two key sensitivity measures for the model output  $Y$  with respect to an individual input  $x_i$  are introduced. The variance of the ME indicates the potential reduction in the overall variance of  $f(\cdot)$  if  $x_i$ ’s were known. The first of these measures is defined in Equation (5).

$$V_i = \text{var}\{E(Y | X_i)\}. \quad (7)$$

The second variance-based SA measure, proposed by [34], can be written as:

$$V_{T_i} = \text{var}(Y) - \text{var}\{E(Y | \mathbf{X}_{-i})\} \quad (8)$$

which is the remaining uncertainty in  $Y$  that is unexplained after everything has been learnt except  $x_i$ .

These two measures (Equations (6) and (7)) can be converted into scale-invariant measures by dividing by  $\text{var}(Y)$ , as follows:

$$S_i = \frac{V_i}{\text{var}(Y)}, \quad S_{T_i} = \frac{V_{T_i}}{\text{var}(Y)} = 1 - S_{-i} \quad (9)$$

where  $S_i$  is the ME index of  $x_i$  and  $S_{T_i}$  is the total effect index of  $x_i$ .

The variance measures are linked to the Sobol decomposition when the parameters are independent. The total variance of  $f$  can be represented as the sum of the variances for each term as given in Equation (7) (see [32,33] for further details on the Sobol decomposition).

#### 4.2. Emulator-Based Sensitivity Analysis

Theoretically, if the function  $f(\mathbf{x})$  was not extremely complex, the sensitivity measures discussed earlier could be computed analytically. However, given the complexity of the models in this study, these measures cannot be evaluated analytically. Instead, a computationally efficient and robust model is required to compute the sensitivity measures discussed in Section 4.1.

When  $f(\mathbf{x})$  is computationally inexpensive and can be quickly computed and easily evaluated for many different inputs, standard Monte Carlo (MC) methods are sufficient for estimating  $var(Y)$  and other sensitivity measures introduced in Section 4.1. However, the computation techniques proposed by [29,30] require many thousands of function evaluations, making them impractical for more expensive functions. To address this computational complexity, the promising methodology outlined in [32], based on the Bayesian paradigm, enables the estimation of all necessary quantities for sensitivity analysis in modeling and predicting wave overtopping for coastal defenses.

Because the functional relationship  $f(\cdot)$  between wave overtopping and the input parameters given in Table 1 is unknown for any specific input configuration  $\mathbf{x}$  until the model is actually run for those inputs, we need to make some assumptions. Within a Bayesian framework, it is appropriate to define a prior distribution for the values of  $f(\mathbf{x})$  at various  $\mathbf{x}$  points. This prior distribution is then updated using Bayesian methods, with data  $\mathcal{D} = \{(\mathbf{x}_i, y_i) : y_i = f(\mathbf{x}_i), i = 1, \dots, n\}$  being generated from a sequence of model simulations. The outcome is a posterior distribution for  $f(\cdot)$ , enabling formal Bayesian inferences regarding the sensitivity analysis measures described earlier.

While uncertainty persists regarding the function  $f(\cdot)$  at input or parameter values where it has not been evaluated, considering the correlation between function values at different points can help reduce this uncertainty. Typically, the expected value of the posterior distribution serves as a point estimate for  $f(\cdot)$ . In SA, two distinct distributions are utilised: first, the distribution  $G$ , representing uncertainty in model inputs/parameters  $\mathbf{x}$ , which is propagated to output values via  $f(\cdot)$ ; and second, the posterior distribution on  $f(\cdot)$ , serving a purely computational role. This latter can be refined by increasing the number of training points  $\mathbf{x}$ , but it has no operational interpretation.

##### 4.2.1. Gaussian Process Emulators

Consider the function or complex model under investigation as a deterministic code that yields an output  $y = f(\mathbf{x})$  for a given input vector  $\mathbf{x}$ . A GP emulator offers a prior representation of our uncertain understanding of this function's values before the deterministic code is executed. Evaluating the code at various input configurations generates the data required to construct a posterior distribution. The process for defining the prior distribution for  $f(\cdot)$  involves two stages: initially, a GP is chosen as the prior distribution based on specific hyperparameters. Subsequently, these hyperparameters are estimated to define the mean vector and covariance matrix of the GP model. The key requirement for employing the GP is that  $f(\cdot)$  should be a smooth function. Thus, knowing the value of  $f(\mathbf{x})$  should provide insight into the value of  $f(\mathbf{x}')$  for  $\mathbf{x}$  near  $\mathbf{x}'$ . This smoothness assumption gives GPs a significant computational advantage over Monte Carlo (MC) methods, which often ignore the expected similarity of function values at proximate points. In this study, we select high-reliability points (see Section 5.1) to minimise noise in our model and select a subset of data that only includes overtopping measurements in the presence of a berm, thereby eliminating a distinct class of events that could violate the continuity assumption inherent in GPs. One of the key strengths of GPs is their ability to provide uncertainty quantification in predictions, even when the smoothness assumption is partially violated. While incorporating non-stationary kernels could allow us to handle potential discontinuities more explicitly, such an approach lies beyond the scope of this study. Instead, we rely on careful data selection to maintain model validity and enhance prediction robustness.

Using a GP prior for  $f(\cdot)$  means that the uncertainty about  $f(\mathbf{x}_1), \dots, f(\mathbf{x}_n)$ , given any set of points  $\mathbf{x}_1, \dots, \mathbf{x}_n$ , can be expressed as a multivariate normal distribution. Con-



sequently, we need to make feasible prior assumptions about the mean and covariance. The mean of  $f(\mathbf{x})$ , given the hyperparameters  $\beta$ , is modelled as

$$E[f(\mathbf{x})|\beta] = \mathbf{h}(\mathbf{x})^T \beta \quad (10)$$

where  $\mathbf{h}(\cdot)$  is a vector of  $q$  known functions of  $\mathbf{x}$  and  $\beta$  is a vector of coefficients. The selection of  $\mathbf{h}(\cdot)$  is flexible, but it should be made to incorporate any prior beliefs we have about the form of  $f(\cdot)$ . The covariance between  $f(\mathbf{x})$  and  $f(\mathbf{x}')$  is given by:

$$\text{cov}(f(\mathbf{x}), f(\mathbf{x}')|\sigma^2) = \sigma^2 c(\mathbf{x}, \mathbf{x}') \quad (11)$$

where  $c(\cdot, \cdot)$  is a monotone correlation function on  $\mathbb{R}^+$  with  $c(\mathbf{x}, \mathbf{x}) = 1$ , and it decreases as  $|\mathbf{x} - \mathbf{x}'|$  increases. Furthermore, the function  $c(\cdot, \cdot)$  must ensure that the covariance matrix of any set of outputs  $\{y_1 = f(\mathbf{x}_1), \dots, y_n = f(\mathbf{x}_n)\}$  is positive semi-definite. Throughout this paper, we use the following correlation function, which satisfies all the conditions mentioned above and is widely used for its computational convenience:

$$c(\mathbf{x}, \mathbf{x}') = \exp\{-(\mathbf{x} - \mathbf{x}')^T \mathbf{B}(\mathbf{x} - \mathbf{x}')\}, \quad (12)$$

where  $\mathbf{B}$  is a diagonal matrix composed of positive smoothness parameters  $\{(\sqrt{2}b_i)^{-2}\}_{i=1}^d$ , where  $d$  represents the dimension of  $\mathbf{x}$ . It should be noted that  $\mathbf{B}$  functions to re-scale the distance between  $\mathbf{x}$  and  $\mathbf{x}'$ , thereby determining the proximity required between two inputs  $\mathbf{x}$  and  $\mathbf{x}'$  for the correlation between  $f(\mathbf{x})$  and  $f(\mathbf{x}')$  to achieve a specific value.

The normal inverse gamma distribution for  $(\beta, \sigma^2)$  was proposed in [32] for fixed hyperparameters  $\mathbf{z}$ ,  $V$ ,  $a$  and  $d$ , as represented by

$$p(\beta, \sigma^2) \propto (\sigma^2)^{-\frac{1}{2}(d+q+2)} \exp\{-\{(\beta - \mathbf{z})^T V^{-1}(\beta - \mathbf{z}) + a\} / (2\sigma^2)\}$$

The function  $f(\cdot)$  yields outputs at  $n$  predetermined design points  $\mathbf{x}_1, \dots, \mathbf{x}_n$ , generating the dataset as  $\mathbf{y} = \{f(\mathbf{x}_1), \dots, f(\mathbf{x}_n)\}$ . Unlike MC methods, these points are intentionally chosen to provide informative insights about  $f(\cdot)$ . Typically, these design points are strategically distributed across the input space  $\mathcal{X}$  of  $\mathbf{X}$ , the unknown variables, guided by the probability distribution  $G(\mathbf{X})$ . Consequently, the selection of design points is influenced by  $G(\cdot)$ , as detailed in [35]. The standardised posterior distribution of  $f(\cdot)$ , conditioned on  $\mathbf{y} = \{f(\mathbf{x}_1), \dots, f(\mathbf{x}_n)\}$ , is then determined as:

$$\frac{f(\mathbf{x}) - m^*(\mathbf{x})}{\hat{\sigma} \sqrt{c^*(\mathbf{x}, \mathbf{x}')}} \mid \mathbf{y} \sim t_{d+n}, \quad (13)$$

where  $t_{d+n}$  stands for a Student's  $t$ -distribution with  $n + d$  degrees of freedom.

The mean of the resulting posterior is then given by

$$m^*(\mathbf{x}) = \mathbf{h}(\mathbf{x})^T \hat{\beta} + \mathbf{t}(\mathbf{x})^T A^{-1}(\mathbf{y} - H\hat{\beta}), \quad (14)$$

the updated correlation function described in Equation (12) given the observed data can be written as:

$$c^*(\mathbf{x}, \mathbf{x}') = c(\mathbf{x}, \mathbf{x}') - \mathbf{t}(\mathbf{x})^T A^{-1} \mathbf{t}(\mathbf{x}') + (\mathbf{h}(\mathbf{x}))^T \\ - \mathbf{t}(\mathbf{x})^T A^{-1} H)(H^T A^{-1} H)^{-1} (h(\mathbf{x}')^T - \mathbf{t}(\mathbf{x}')^T A^{-1} H)^T \quad (15)$$

and

$$\mathbf{t}(\mathbf{x})^T = (c(\mathbf{x}, \mathbf{x}_1), \dots, c(\mathbf{x}, \mathbf{x}_n)), \quad (16) \\ H^T = (\mathbf{h}^T(\mathbf{x}_1)^T, \dots, \mathbf{h}^T(\mathbf{x}_n)^T),$$

$$A = \begin{pmatrix} 1 & c(\mathbf{x}_1, \mathbf{x}_2) & \dots & c(\mathbf{x}_1, \mathbf{x}_n) \\ c(\mathbf{x}_2, \mathbf{x}_1) & 1 & & \vdots \\ \vdots & & \ddots & \\ c(\mathbf{x}_n, \mathbf{x}_1) & \dots & & 1 \end{pmatrix} \quad (17)$$

$$\begin{aligned} \boldsymbol{\beta} &= V^*(V^{-1}\mathbf{z} + H^T A^{-1}\mathbf{y}), \\ \hat{\sigma}^2 &= \frac{\{a + \mathbf{z}^T V^{-1}\mathbf{z} + \mathbf{y}^T A^{-1}\mathbf{y} - \hat{\boldsymbol{\beta}}^T (V^*)^{-1} \hat{\boldsymbol{\beta}}\}}{(n + d - 2)} \\ V^* &= (V^{-1} + H^T A^{-1} H)^{-1}. \end{aligned}$$

The outputs for any set of inputs will follow a multivariate  $t$ -distribution, with the covariance between any two outputs being defined by Equation (13). The  $t$ -distribution arises as the marginal distribution for  $f(\cdot)$  after integrating the hyperparameters  $\boldsymbol{\beta}$  and  $\sigma^2$ . In practice, additional hyperparameters, known as smoothness parameters  $\mathbf{B}$ , are involved in modelling the correlation function,  $c(\cdot, \cdot)$ . It is often impractical to give  $\mathbf{B}$  a fully analytical Bayesian treatment as integrating the posterior distribution analytically with respect to these parameters is generally impossible. One straightforward approach is to keep  $\mathbf{B}$  fixed. Alternatively, numerical methods, such as Markov chain Monte Carlo (MCMC) sampling, can be used to integrate the posterior distribution, though this is computationally intensive. A practical and robust approach is to estimate the hyperparameters of  $c(\cdot, \cdot)$  from the posterior distribution and substitute these estimates into  $c(\cdot, \cdot)$  in the relevant formulae [36]. These estimates can be obtained using the posterior mode combined with a cross-validation approach [37]. The GEM-SA tool can estimate the smoothness parameters using either method.

#### 4.2.2. Analysis of Main Effects and Interactions

This section explains how sensitivity metrics discussed earlier can be estimated using the GP posterior distribution obtained in Section 4.2.1. An important insight from [32] is that inferences about  $f(\cdot)$  can be used to derive information about the main and interaction effects of  $f(\cdot)$ . This is because these effects are linear functions of  $f(\cdot)$  and  $t_{d+n}$  after standardisation, as illustrated in (13). As a result, the derived posterior for the main and interaction effects will also be  $t_{d+n}$ . In particular, if the posterior mean of  $f(\cdot)$  is expressed as shown in Equation (14), subsequently for

$$E(Y | \mathbf{x}_p) = \int \chi_{-p} f(\mathbf{x}) dG_{-p|p}(\mathbf{x}_{-p} | \mathbf{x}_p) \quad (18)$$

(where  $\chi_{-p}$  refers to the input space corresponding to  $\mathbf{x}_{-p}$ , while  $G_{-p|p}(\mathbf{x}_{-p} | \mathbf{x}_p)$  represents the conditional distribution of  $\mathbf{x}_{-p}$  given  $\mathbf{x}_p$  under  $G$ ), the posterior mean of this quantity can be written as:

$$E_{\text{post}} E(Y | \mathbf{x}_p) = R_p(\mathbf{x}_p) \hat{\boldsymbol{\beta}} + T_p(\mathbf{x}_p) \mathbf{e} \quad (19)$$

where

$$R_p(\mathbf{x}_p) = \int_{\chi_{-p}} h(\mathbf{x})^T dG_{-p|p}(\mathbf{x}_{-p} | \mathbf{x}_p), \quad (20)$$

$$T_p(\mathbf{x}_p) = \int_{\chi_{-p}} t(\mathbf{x})^T dG_{-p|p}(\mathbf{x}_{-p} | \mathbf{x}_p) \quad (21)$$

and  $\mathbf{e} = A^{-1}(\mathbf{y} - H\hat{\boldsymbol{\beta}})$ .

Similarly, the mean of posterior for ME or interaction can be derived in the following way:

$$E_{\text{post}} \{z_i(x_i)\} = \{R_i(x_i) - R\} \hat{\boldsymbol{\beta}} + \{T_i(x_i) - T\} \mathbf{e}. \quad (22)$$

In a similar manner, the standard deviations of the MEs and interactions can be derived; see [32] for more details on the computational aspects.

The posterior mean of the ME  $E_{pos}(z_i(x_i))$  can be plotted against  $x_i$ , with bounds representing, for example, plus and minus two posterior standard deviations. By standardising the input variables, we can visualise  $E_{pos}(z_i(x_i))$  for  $i = 1, \dots, d$  on a single plot. This provides a concise graphical summary of the influence that each input variable exerts on the model's output. Section 5 will present this plot using the dataset introduced in Section 3, offering insights into the impact of different parameters on wave overtopping predictions.

Direct posterior inference for the variance-based measures, specifically  $V_{T_i}$  discussed in Section 4.1.2, presents greater challenges due to their nature as quadratic functionals of the underlying function  $f(\cdot)$ . These measures are more complex than simple linear functionals like the main effects. To handle this computational complexity and derive these sensitivity measures within a Bayesian framework, advanced techniques are required. An in-depth exploration of these methods, including how Gaussian process (GP) emulators can be utilised to compute such measures, is detailed in [32].

## 5. Results

### 5.1. Data Preparation and Initial Examination of the CLASH Dataset

This section describes the pre-processing steps taken prior to implementing a novel probabilistic sensitivity analysis approach and predictive modelling for the wave overtopping dataset (Section 3). Initial examination and visualisation of the CLASH dataset in Section 3 revealed a highly non-linear and complex relationship between variables. As a result, standard linear regression models are not suitable for modelling wave overtopping data with such a complex dependency structure. Previous studies have attempted to model these intricate relationships between variables in wave run-up and overtopping using artificial neural networks (ANNs) [38,39].

The emulator-based sensitivity analysis, outlined in Section 4.2, will be used to perform an SA of the parameters influencing wave overtopping. This probabilistic method is developed based on the GP regression as a computationally efficient non-parametric Bayesian machine learning technique. To train the GP model required for computing the SA measures described in Sections 4.1.2 and 4.2.2, and to address the other computational objectives discussed in Section 4, the following steps were implemented:

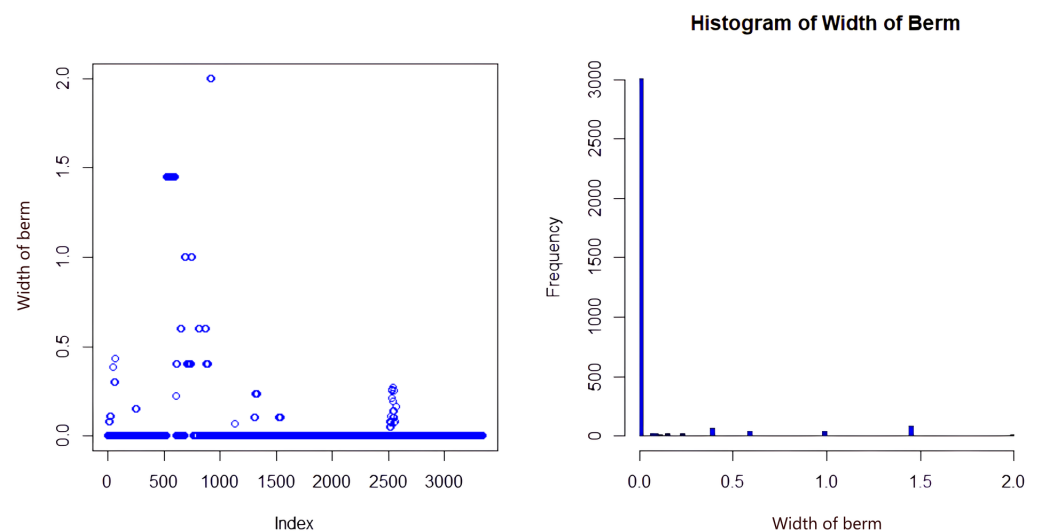
1. Database cleaning and the selection of a highly reliable subset.
2. Perform an exploratory analysis of the subset.
3. Fit a Gaussian process regression model for the selected subset of the dataset.
4. Compute the SA measures, including the variance-based indexes and the main effects.
5. Illustrate the corresponding SA plots.
6. Interpret the SA results, perform an uncertainty analysis, and draw conclusions about the most influencing input parameters affecting the wave overtopping.

The CLASH wave overtopping database contains observations and measurements collected from several sources with varying levels of reliability. Numerical modelling has been implemented to fill the gaps in the dataset, and in order to discern between observed and modelled data, each row was given a reliability rating ranging between 1 and 4, with 1 being for "highly reliable" data points and 4 for "highly unreliable" data.

Selecting only the highest reliability entries in the CLASH dataset offers two significant benefits: it reduces the training time for the Gaussian process regression (GPR) model and increases accuracy due to the improved quality of the training points. This initial screening of the dataset, focusing on high-reliability data, results in a dataset of size  $3385 \times 29$ . After removing any entries with missing values, the final dataset is further reduced to 3208 observations, consisting of 27 feature variables and 1 output ( $q[m^3/s.m]$ ), which represents the wave overtopping discharge per unit width. The choice of using GPR for the model development is well justified because GPR is known for its efficiency in constructing reliable models with far fewer training points compared with methods like artificial neural networks (ANNs). ANNs, as employed in previous studies, typically require large training

datasets to generalise effectively. In contrast, GPR’s ability to incorporate uncertainty and work with smaller, higher-quality datasets makes it a more appropriate choice for this study. This further supports the adoption of the GPR approach for modelling wave overtopping, especially given the non-linear and complex dependency structure present in the data.

The dataset screening process was further refined by eliminating non-informative feature variables. This was accomplished through an initial data analysis involving the regression of feature variables against the output variable. During this analysis, the scatter plot (Figure 5 left) and histogram (Figure 5 right) of ‘Width of Berm (B)’ revealed that the recorded values for this variable were predominantly zeros. This presented a potential issue for the modelling process. Not only does a variable with mostly zero values contribute little information to the model, but it can also cause numerical challenges during the training phase, particularly for data-driven models like the GPR approach employed here. To mitigate these issues, this study opted to focus on analysing a smaller subset of the dataset, consisting of 330 high-reliability data points. By selecting this high-quality subset, the aim was to probabilistically identify the most suitable model linking wave overtopping to the remaining 27 input parameters. This step ensures that the model is developed using informative and reliable data, thus improving the overall accuracy and robustness of the predictive modelling process.



**Figure 5.** Scatter plot (left) and histogram (right) of ‘width of berm’ in the dataset with the highest level of reliability.

### 5.2. GP-Based Sensitivity Analysis for the Wave Overtopping Dataset

The SA of the overtopping parameter,  $q$ , with respect to the changes in the input parameters described in Section 5.1, has been conducted. The results, expressed in terms of variance-based measures and total effects, are presented in Table 2.

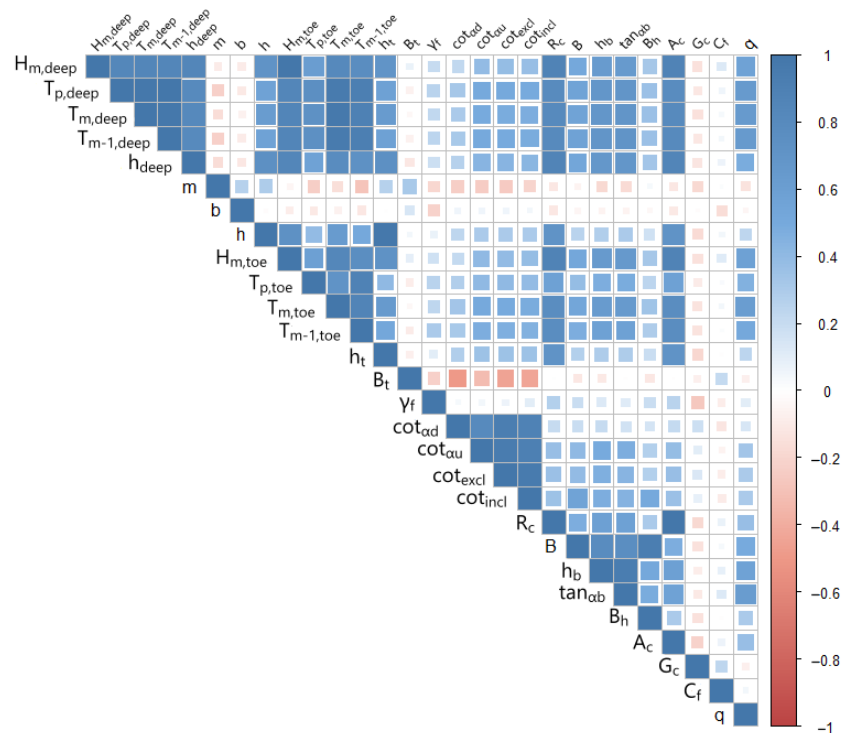
To identify the variables that are most influential on the overtopping parameter, the correlation coefficients between the input variables and the wave overtopping discharge parameter,  $q$ , were examined. Figure 6 illustrates the correlation matrix of these variables. Several strong positive correlations are apparent. For example, significant correlations are observed between  $(H_{(m, deep)}, q)$ ,  $(T_{(p, deep)}, q)$ , and  $(T_{(m, deep)}, q)$ . In contrast, the correlations between  $(m, q)$  and  $(b, q)$  are very weak. Despite Figure 6 being useful for assessing the strength and direction of correlations between the input variables and  $q$ , it does not provide insight into the underlying functional relationships between the input parameters and  $q$ .

**Table 2.** The emulator-based SA of the overtopping parameter with respect to the changes in other input parameters. The seven parameters contributing the greatest variance are in bold.

Parameters	Variance (%)	Total Effect
<b>Signif wave height (<math>H_{m,deep}</math>)</b>	<b>10.93</b>	<b>11.22</b>
<b>Peak period in the deep (<math>T_{p,deep}</math>)</b>	<b>5.53</b>	<b>5.83</b>
Mean period $m2/m0$ , deep ( $T_{m,deep}$ )	0.79	0.89
<b>Mean period, deep (<math>T_{m-1,deep}</math>)</b>	<b>7.63</b>	<b>7.77</b>
Off-shore Water depth, ( $H_{deep}$ )	1.93	2.04
Slope of foreshore ( $m$ )	1.28	1.38
Angle of wave attack ( $\beta$ )	0.85	0.96
Water depth at toe ( $h$ )	1.59	1.69
<b>Signif wave height at toe (<math>H_{m,toe}</math>)</b>	<b>8.11</b>	<b>8.22</b>
<b>Peak period, toe (<math>T_{p,toe}</math>)</b>	<b>15.75</b>	<b>15.86</b>
<b>Mean wave period, toe (<math>T_{m,toe}</math>)</b>	<b>8.48</b>	<b>8.59</b>
Spectral wave period at toe ( $T_{m-1,toe}$ )	4.06	4.17
Water depth on toe ( $h_t$ )	1.27	1.40
Toe width ( $B_t$ )	0.95	1.06
Roughness/perm factor ( $\gamma_f$ )	0.06	0.17
Cot downward slope, berm ( $\cot\alpha_d$ )	4.01	4.30
Cot upward slope, berm ( $\cot\alpha_u$ )	0.65	0.68
<b>Cot slope, excl berm (<math>\cot\alpha_{excl}</math>)</b>	<b>16.91</b>	<b>17.02</b>
Cot slope, incl berm ( $\cot\alpha_{incl}$ )	1.07	1.37
Crest freeboard ( $R_c$ )	0.74	1.04
Width of berm ( $B$ )	2.15	2.22
Water depth on berm ( $h_b$ )	2.51	2.65
tan of slope of berm ( $\tan ab$ )	0.32	0.43
Width of horizontally schematised berm ( $B_h$ )	1.50	1.62
Width of crest ( $G_c$ )	0.39	0.50
Armour crest freeboard ( $A_c$ )	0.20	0.31
Total variance (%)	99.64	
Estimated mean output	0.00779018	
Estimated variance output	$1.35091 \times 10^{-6}$	

Figure 7 illustrates the scatter plots between several input parameters (mainly structural and geomorphological parameters, as listed in Table 1) and the wave overtopping discharge parameter,  $q$ . Due to space constraints, not all input parameters could be included in this figure. From the scatter plots and the computed correlations between each input variable and  $q$ , it becomes clear that the relationships between these variables and the wave overtopping discharge are highly non-linear. Moreover, some input parameters, such as  $\cot\alpha_{excl}$  or  $\cot\alpha_{incl}$ , exhibit distributions with multiple modes, suggesting that their relationships with  $q$  and even other input parameters defy interpretation via any known mathematical functional forms. Consequently, it would be misleading to rely solely on correlation coefficients for the SA of this complex, non-linear system, as highlighted by [40].

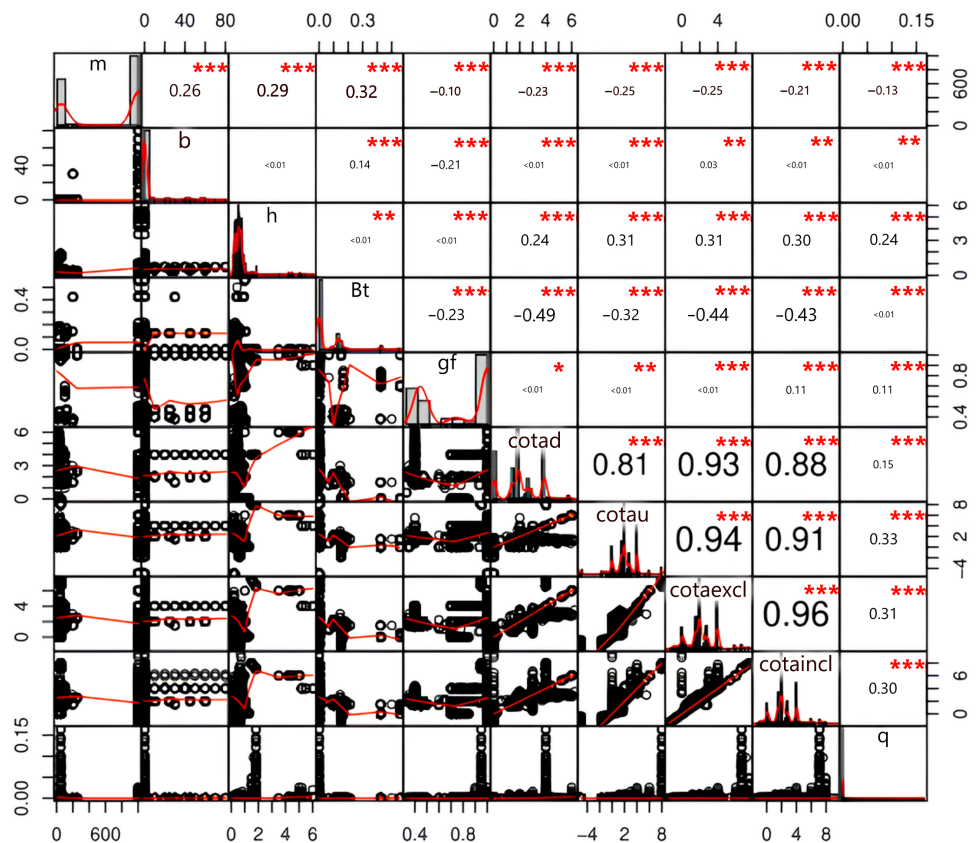
Using correlation coefficients in such a scenario could obscure the true effects of input variables on the prediction of the wave overtopping discharge parameter. Additionally, conventional SA methods, including Markov chain Monte Carlo (MCMC)-based approaches [41], may struggle to capture the complex, non-linear dependencies and interactions inherent in this model.



**Figure 6.** Correlation plot of all variables in the subset of the CLASH dataset.

To examine the variation of model output with respect to the uncertainty in the input variables, we further develop the probabilistic SA approaches described in Section 4.1. The GP emulator helps to overcome the computational challenges involved in calculating the SA indices for the complex, non-linear system considered in this study. These SA indices allow us to evaluate the contribution of each input variable to the overall variability of the system. Specifically, by leveraging the GP emulator, we efficiently compute these indices and quantify the influence of each variable on the wave overtopping model's output. The probabilistic SA enables us to assess how uncertainty in the input variables propagates through the model and affects the overtopping discharge parameter. This approach is crucial for understanding which variables are the most influential in predicting wave overtopping, allowing for a more focused analysis and potential simplifications in the model complexity without compromising accuracy. The SA indices thus provide a powerful tool for identifying key drivers of uncertainty and guiding decision making in model refinement and resource allocation.

Table 2 provides the results of the emulator-based SA for the wave overtopping parameter with respect to changes in the input parameters discussed in Table 1. The results highlight that a significant portion of the variability in wave overtopping discharge, approximately 20%, can be attributed to 'significant wave height' measurements. Furthermore, the results indicate that  $H_{m,deep}$  accounts for 10.93% and  $H_{m,toe}$  accounts for 8.11% of the total variance. Further investigation into the relationship between these hydrodynamic parameters and wave overtopping discharge reveals strong correlations. Figure 8 illustrates that significant wave height parameters ( $H_{m,deep}$ ,  $H_{m,toe}$ ) are also highly correlated. This correlation indicates that significant wave height is a key factor influencing wave overtopping and suggests it could be pivotal in simplifying the final model or reducing model dimensionality. Given these findings, a significant wave height should be included as one of the primary features in any predictive model for wave overtopping. This inclusion will enhance model accuracy and focus, leveraging the substantial variance contribution from these hydrodynamic parameters.

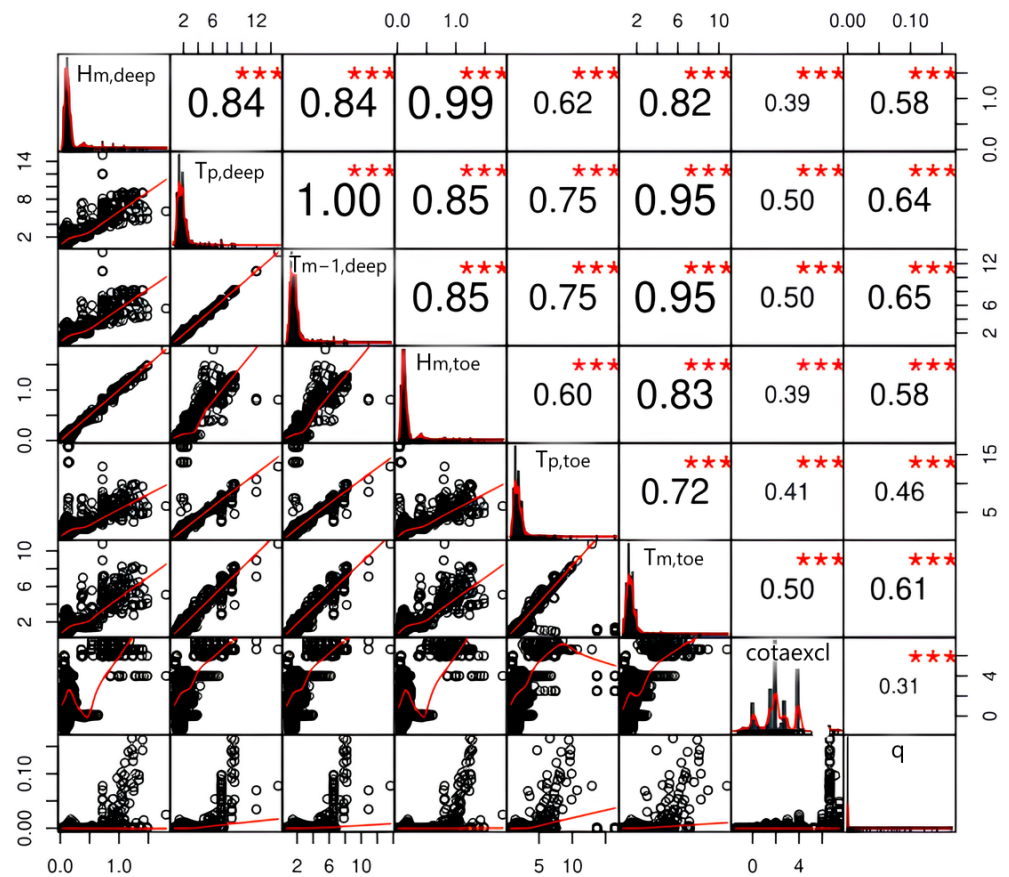


**Figure 7.** Correlation plot between some input parameters (randomly selected) and the overtopping parameter. Red stars indicate the level of statistical significance: \* for  $p \leq 0.05$ , \*\* for  $p \leq 0.01$ , and \*\*\* for  $p \leq 0.001$ .

The most influential input parameter identified by the models is the ‘mean cotangent of structure slope without contribution to the berm’, denoted by  $\cot \alpha_{excl}$ , which contributes approximately 17% to the total variance of the wave overtopping output. This finding of the model aligns with fundamental underlying theories, which emphasise the importance of the slope in determining the wave breaker index (Iribarren number). The Iribarren number provides an important role in the wave breaking process, affecting turbulent kinetic energy and momentum transfer, both of which are crucial factors in wave overtopping at coastal defence structures.

In addition to slope, several wave period-related parameters also significantly influence wave overtopping. The ‘Peak wave period at toe’,  $T_{p,toe}$ , contributes 15.75%, while the ‘off-shore peak wave period in the deep water’,  $T_{p,deep}$ , and ‘off-shore spectral wave period’,  $T_{m-1,deep}$ , contribute 5.53% and 7.63% of total variance contribution toward the output’s variance, respectively. These findings are consistent with existing theories as wave periods directly impact the frequency and intensity of wave impacts on coastal structures.

Furthermore, parameters such as ‘Roughness/permeability factor for the structure ( $\gamma_f$ )’, ‘Width of the structure crest ( $G_s$ )’, and ‘Armour crest freeboard ( $G_s$ )’ contribute relatively little to the overall variance in the wave overtopping model (see Table 2). These parameters, with their relatively insignificant contributions to variance, suggest that they are not critical for predicting the wave overtopping discharge. Notably, 99% of the variability in the output is attributed to MEs, which further underscores the suitability of using GPR in this analysis.



**Figure 8.** Correlation plot between the most sensitive input parameters and the wave overtopping discharge. Red Stars indicate significance lvl.

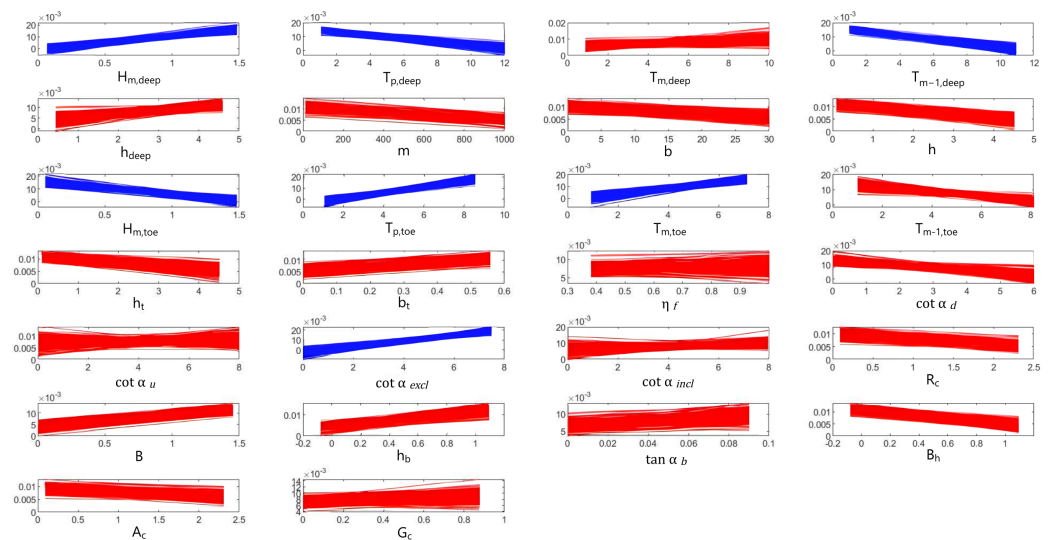
Figure 9 illustrates the estimated MEs,  $E(q|X_i)$ , where  $X_i$ , defined in the general case in Equation (18), is an input parameter (detailed in Table 1). These MEs are approximated using the Gaussian process (GP) model as described in Section 4.2.2. This approach allows us to evaluate the sensitivity of the wave overtopping parameter with respect to individual input variables while maintaining computational efficiency and flexibility. To compute the SA measures reported in Table 2 and the MEs for each input parameter, the GP model was trained using only 330 high-reliability data points after cleaning the dataset, as described in Section 5.1.

The SA measures proposed in this study are computed by developing a custom code in R. To facilitate computations, the input variables and the output (wave overtopping discharge) were standardised prior to fitting the GP emulator. Standardising the data not only helps to avoid numerical issues but also makes the computations more consistent across different models. Once the GP model was trained, the results were then transformed back to their original scale to make them interpretable and applicable to real-world scenarios. This process, as supported by [42], showcases the advantages of using standardised data, particularly when working with complex models like the GP emulator. This ensures numerical stability while maintaining flexibility across a wide range of computational implementations. The final model results can then be transformed back to the original scale.

Figure 9 shows the estimated MEs of the overtopping parameter in response to changes in various input parameters. The figure highlights that the overtopping parameter is most sensitive to uncertainties in terms of  $\cot \alpha_{excl}$ ,  $T_{p,toe}$ ,  $H_{m,deep}$ ,  $T_{m,toe}$ ,  $H_{m,toe}$ ,  $T_{m-1,deep}$ , and  $T_{p,deep}$ , respectively. The width of the uncertainty bands depicted in Figure 9 represents the uncertainty of the GP emulator linked to each input. This uncertainty quantification provides insight into the reliability of predictions for each parameter. For instance, when  $H_{m,deep}$  is fixed at a value of 0.5, the corresponding point on the graph represents the



expected value of the overtopping discharge parameter,  $q$ , derived by averaging across the remaining input parameters.

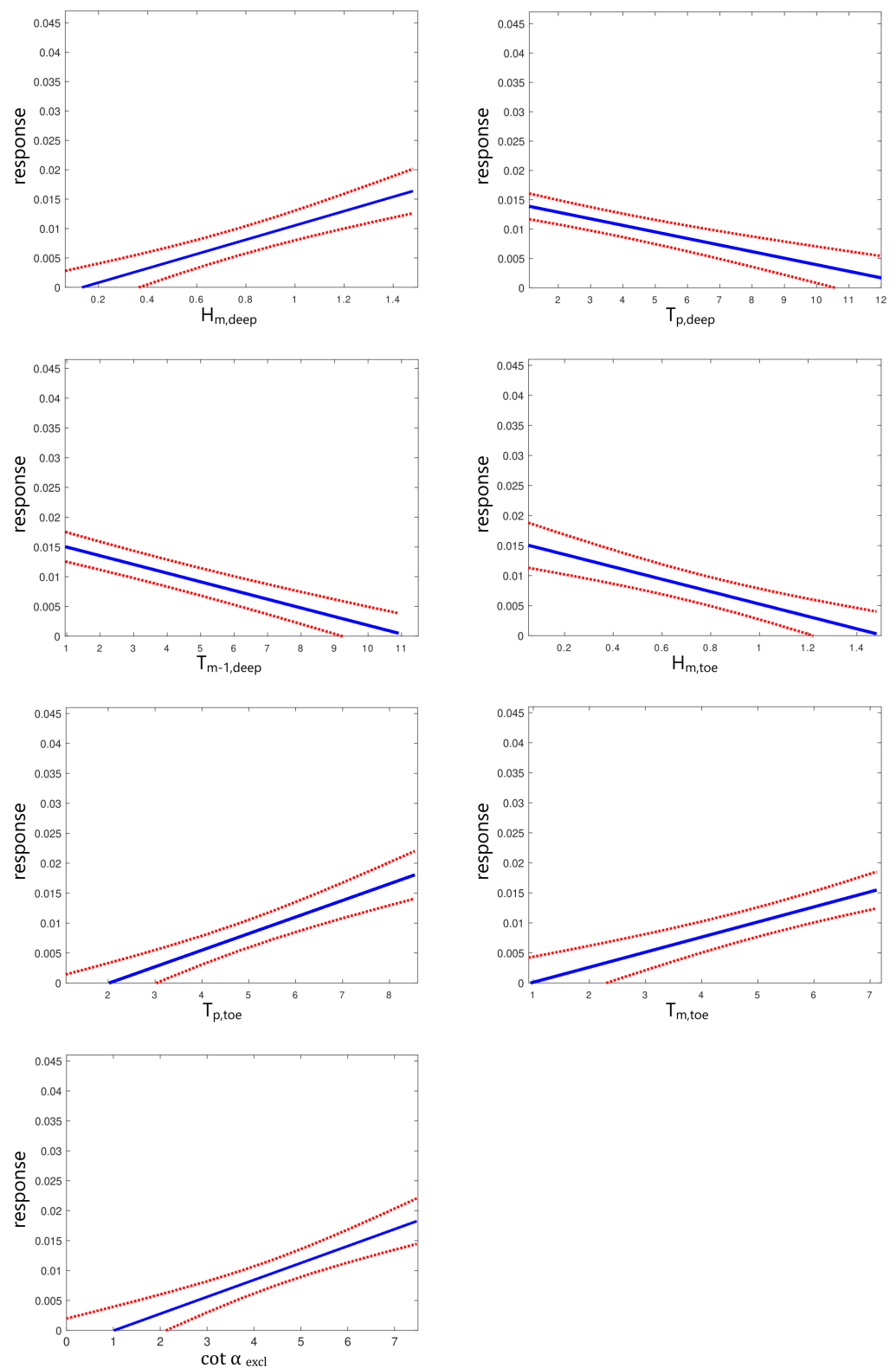


**Figure 9.** The estimated main effects of the overtopping parameter with respect to the changes in other input parameters using the emulator-based SA method. The seven parameters contributing the greatest variance are in blue.

The emulator-based SA method allows for systematic exploration by varying small groups of input variables, while others are held fixed at their default values. This approach enables the examination of the sensitivity of the output (overtopping parameter,  $q$ ) to particular input variables while controlling for other factors. When comparing the thickness of the ME plots, it becomes clear that there is less uncertainty associated with certain key variables, such as  $\cot \alpha_{excl}$ ,  $T_{p,toe}$ ,  $H_{m,deep}$ ,  $T_{m,toe}$ ,  $H_{m,toe}$ ,  $T_{m-1,deep}$ , and  $T_{p,deep}$ , which are illustrated in blue, than the rest of the input parameters. The thickness of the ME plots is the result of simulating multiple (in this case, 200) realisations from the posterior distribution of the output (overtopping parameter  $q$ ). These realisations are computed at a regularly spaced grid of input points, as described in Section 4.2.1. The thicker ME bands indicate more uncertainty, while thinner bands suggest a higher degree of confidence in the model's predictions for those variables. The variability in the spread of ME lines illustrates the combined uncertainty present in both the model and input parameters. Thus, these sensitivity measures themselves are subject to uncertainty, and generating multiple realisations helps to reflect that uncertainty in a probabilistic framework.

To enhance the clarity of the visualisations of the ME plots illustrated in Figure 9, we now focus on the ME plots of the most sensitive input parameters, namely  $\cot \alpha_{excl}$ ,  $T_{p,toe}$ ,  $H_{m,deep}$ ,  $T_{m,toe}$ ,  $H_{m,toe}$ ,  $T_{m-1,deep}$ , and  $T_{p,deep}$ . The improved visualisations are presented in Figure 10, which provides a more detailed focus on the most influential parameters. In each ME plot, the blue solid line represents the posterior mean of the ME, calculated as described in Section 4.2.2. This line indicates the central tendency of the model's prediction for the effect of each input parameter on the output given the available data. The red dotted lines show the 95% confidence interval around the ME (see Section 4.2.2 for details).

This confidence interval is derived from the posterior distribution of the emulator and illustrates the uncertainty in the estimated ME. It is worth noting that this confidence interval offers an alternative method for assessing uncertainty compared with the approach of using multiple realisations simulated from the posterior distribution of the overtopping parameter,  $q$ . Both methods provide insight into the uncertainty surrounding the estimated ME, though they approach the analysis from slightly different perspectives.



**Figure 10.** The main effects (blue lines) and their 95% confidence intervals (red lines) for the most sensitive input parameters, as illustrated by the blue images in Figure 9.

One final point to note about the proposed method in this study is that the machine learning Gaussian process (GP) sensitivity analysis results for the wave overtopping discharge parameter are consistent with the findings derived from the variance-based sensitivity analysis (SA) measures, as reported in Table 2. This agreement reinforces the

robustness of the GP-based approach, indicating that it effectively captures the influential input parameters and their contributions to the variability of the wave overtopping discharge parameter. The concordance between the two approaches—one probabilistic (GP-based) and the other variance-based—validates the applicability of the GP method for complex, non-linear systems like the one studied here. It suggests that the GP emulator can be reliably used as a computationally efficient alternative to traditional variance-based methods for sensitivity analysis, especially in cases where the underlying functional relationships between inputs and outputs are highly non-linear or difficult to model with simpler techniques.

## 6. Conclusions

This study outlines a novel probabilistic machine learning approach for sensitivity analysis of complex, non-linear, multi-variable datasets on wave overtopping from coastal defences. The overtopping data used in this study are based on 163 two- and three-dimensional physical modelling tests as well as full-scale field-based (prototype) measurements reported in the CLASH dataset. The homogeneous database of this study includes measurements for all common types of hard-engineered coastal protection structures and captures data on the hydrodynamics, geomorphological and structural parameters influencing wave overtopping discharge.

The analysis revealed that the overtopping process is highly complex, driven by non-linear interactions between wave kinematics, geomorphological features, and structural parameters. For the first time, this paper proposes a mathematically robust and computationally efficient framework based on the GP emulator to perform a sensitivity analysis of the multi-faceted wave overtopping problem. This framework analyses how variations in key input variables influence the wave overtopping from coastal defences.

The GP model developed in this study was successfully tested and validated using data from the CLASH database. The proposed method allows for an effective sensitivity analysis with significantly fewer model runs compared with conventional SA approaches, including MCMC-based methods [33].

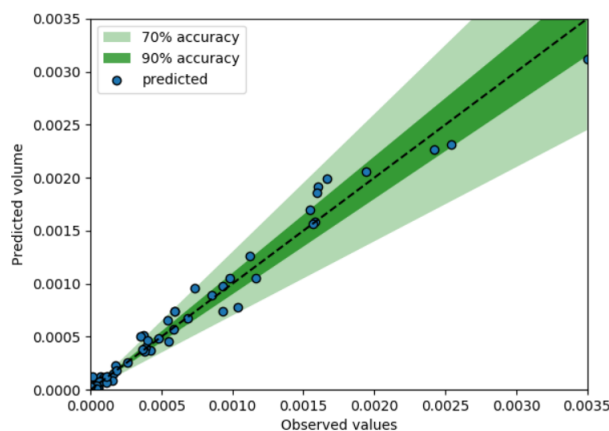
The machine learning-based sensitivity analysis model presented in this paper is crucial for predicting wave overtopping discharge from critical coastal infrastructures, a key factor in forecasting coastal flooding. The results of the sensitivity analysis can be directly applied to improve predictive models by highlighting the parameters with the most significant impact on wave overtopping, thereby enhancing the accuracy of assessments.

Using the GP emulator, computationally expensive sensitivity measures, such as variance-based analyses and main effects (MEs) of the overtopping parameters, were efficiently computed. The study found that significant wave height features ( $H_{m,deep}$ ,  $H_{m,toe}$ ), slope without berm contribution ( $\cot \alpha_{excl}$ ), peak wave period at the toe ( $T_{p,toe}$ ), offshore peak wave period ( $T_{p,deep}$ ), and offshore spectral wave period ( $T_{m-1,deep}$ ) are the most influential factors determining the intensity of wave overtopping discharge.

The results from the GP-based SA methods were leveraged to simplify and develop a robust predictive model for wave overtopping and coastal flooding based on the CLASH database. Figure 11 illustrates the initial results of predictive modelling for wave overtopping discharges from vertical defence structures (e.g., seawalls). The GP-based model demonstrates an efficient and reliable prediction of wave overtopping volumes, showing potential to outperform existing methods, such as those based on artificial neural networks, as reported in [38,39]. Additionally, due to the capability of GPs to make accurate predictions with far fewer data points, compared with ANNs, the proposed method allows for SA and modelling on smaller data subsets, enabling rigorous analyses, even with limited data availability.

The mathematical framework presented in this paper, along with the machine learning Gaussian process (GP)-based model developed, can significantly contribute to creating a reliable decision support tool. This tool would empower coastal scientists and engineers

to better evaluate and predict the performance of key coastal infrastructures under the growing threats posed by climate change.



**Figure 11.** Overtopping predictions from the GP model versus observation data for a vertical seawall with no berm.

This study provides a foundation for future work. While our method exhibits impressive predictive power on the CLASH dataset, many real-world coastlines are under-represented in these data. Future research should investigate the generalisability of this approach, which would require an expanded data collection. This paper primarily focused on topological and hydrological parameters, yet other sources of variability—such as wind direction and speed—are known to affect overtopping. The proposed methodology could be extended to explore these additional parameters, although collecting such data may be expensive. This enhances the value of our data-efficient machine learning approach.

With the availability of more data, future work should explore methods to enhance the computational efficiency of the GP model, particularly through approaches such as sparse GPs [43].

Although we introduced a novel method and demonstrated its promising performance relative to the current state of the art, future research should undertake a thorough comparison with other existing methods. Such work could help clarify the specific use cases where our approach excels, contributing to a deeper understanding of when and where to apply various predictive techniques.

One particularly impactful area for future exploration is spatio-temporal forecasting. Given sufficient data, GPs could be employed to predict the effects of climate change on coastal regions, forecasting the likelihood of wave overtopping in the coming decades. These predictive insights could help identify critical risks to vulnerable coastal areas well in advance, enabling timely and effective preventive measures.

**Author Contributions:** Conceptualization, P.K., S.A. and A.D.; methodology, P.K., S.A., H.A.A., T.S., O.C. and A.D.; software, P.K., T.S., O.C. and A.D.; validation, P.K., S.A., H.A.A., T.S., O.C. and A.D.; formal analysis, P.K., S.A., H.A.A., T.S., O.C. and A.D.; investigation, P.K., S.A., H.A.A., T.S., O.C. and A.D.; resources, P.K. and S.A.; data curation, P.K. and S.A.; writing—original draft preparation, P.K., S.A., H.A.A., T.S., O.C. and A.D.; writing—review and editing, P.K., S.A., H.A.A., T.S., O.C. and A.D.; visualization, P.K., O.C. and A.D.; supervision, S.A. and A.D.; project administration, S.A. and A.D.; funding acquisition, A.D. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The CLASH dataset used in this article is publicly available from the EurOtop website, accessible via the following link: <http://www.overtopping-manual.com/eurotop/>

[neural-networks-and-databases/](#) (accessed on 12 April 2024). The authors can provide the codes and any other relevant details upon request.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. United Nations. Fact Sheet: People and Oceans, 2007. Available online: [https://sustainabledevelopment.un.org/content/documents/Ocean\\_Factsheet\\_People.pdf](https://sustainabledevelopment.un.org/content/documents/Ocean_Factsheet_People.pdf) (accessed on 4 September 2024).
2. Lee, J.Y.; Marotzke, J.; Bala, G.; Cao, L.; Corti, S.; Dunne, J.P.; Engelbrecht, F.; Fischer, E.; Fyfe, J.C.; Jones, C.; et al. Future global climate: Scenario-based projections and near-term information. In *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*; Cambridge University Press: Cambridge, UK, 2021; pp. 553–672.
3. Fanous, M.; Eden, J.M.; Remesan, R.; Daneshkhah, A. Challenges and prospects of climate change impact assessment on mangrove environments through mathematical models. *Environ. Model. Softw.* **2023**, *162*, 105658. [[CrossRef](#)]
4. United Nations. The Climate Crisis—A Race We Can Win. 2020. Available online: <https://www.un.org/en/un75/climate-crisis-race-we-can-win> (accessed on 4 September 2024).
5. Jevrejeva, S.; Jackson, L.; Grinsted, A.; Lincke, D.; Marzeion, B. Flood damage costs under the sea level rise with warming of 1.5 °C and 2 °C. *Environ. Res. Lett.* **2018**, *13*, 074014. [[CrossRef](#)]
6. Donnelly, J.; Abolfathi, S.; Daneshkhah, A. A physics-informed neural network surrogate model for tidal simulations. In Proceedings of the 5th ECCOMAS Thematic Conference on Uncertainty Quantification in Computational Science and Engineering, Athens, Greece, 12–14 June 2023; pp. 836–844.
7. Liu, N.; Salauddin, M.; Yeganeh-Bakhtiari, A.; Pearson, J.; Abolfathi, S. The impact of eco-retrofitting on coastal resilience enhancement—A physical modelling study. In Proceedings of the IOP Conference Series: Earth and Environmental Science, Mohali, India, 23–24 June 2022; IOP Publishing: Bristol, UK, 2022; Volume 1072, p. 012005.
8. Gallien, T.; Sanders, B.; Flick, R. Urban coastal flood prediction: Integrating wave overtopping, flood defenses and drainage. *Coast. Eng.* **2014**, *91*, 18–28. [[CrossRef](#)]
9. Xie, D.; Zou, Q.P.; Mignone, A.; MacRae, J.D. Coastal flooding from wave overtopping and sea level rise adaptation in the northeastern USA. *Coast. Eng.* **2019**, *150*, 39–58. [[CrossRef](#)]
10. Lynett, P.J.; Melby, J.A.; Kim, D.H. An application of Boussinesq modelling to hurricane wave overtopping and inundation. *Ocean Eng.* **2010**, *37*, 135–153. [[CrossRef](#)]
11. Gallien, T. Validated coastal flood modelling at Imperial Beach, California: Comparing total water level, empirical and numerical overtopping methodologies. *Coast. Eng.* **2016**, *111*, 95–104. [[CrossRef](#)]
12. Abolfathi, S.; Pearson, J. Numerical Modelling of Wave Runup & Overtopping under Influence of Complex Geometries. *Coast. Eng. Proc.* **2018**, *1*, 44.
13. Torabbeigi, M.; Akbari, H.; Adibzade, M.; Abolfathi, S. modelling wave dynamics with coastal vegetation using a smoothed particle hydrodynamics porous flow model. *Ocean Eng.* **2024**, *311*, 118756. [[CrossRef](#)]
14. Pillai, K.; Etemad-Shahidi, A.; Lemckert, C. Wave overtopping at berm breakwaters: Experimental study and development of prediction formula. *Coast. Eng.* **2017**, *130*, 85–102. [[CrossRef](#)]
15. De Chowdhury, S.; Anand, K.; Sannasiraj, S.; Sundar, V. Nonlinear wave interaction with curved front seawalls. *Ocean Eng.* **2017**, *140*, 84–96. [[CrossRef](#)]
16. Salauddin, M.; Pearson, J. Wave overtopping and toe scouring at a plain vertical seawall with shingle foreshore: A physical model study. *Ocean Eng.* **2019**, *171*, 286–299. [[CrossRef](#)]
17. Battjes, J.A. Surf similarity. In *Coastal Engineering 1974*; American Society of Civil Engineers: Reston, VA, USA, 1974; pp. 466–480.
18. Besley, P.; Stewart, T.; Allsop, N. Overtopping of vertical structures: New prediction methods to account for shallow water conditions. In Proceedings of the Coastlines, Structures and Breakwaters, London, UK, 19–20 March 1998; pp. 46–57.
19. Goda, Y. Derivation of unified wave overtopping formulas for seawalls with smooth, impermeable surfaces based on selected CLASH datasets. *Coast. Eng.* **2009**, *56*, 385–399. [[CrossRef](#)]
20. Van der Meer, J.; Sigurdarson, S. Geometrical design of berm breakwaters. *Coast. Eng. Proc.* **2014**, *1*, 25. [[CrossRef](#)]
21. Sigurdarson, S.; Van der Meer, J. *Design and Construction of Berm Breakwaters*; World Scientific: Singapore, 2016; Volume 40.
22. Allsop, W.; Bruce, T.; Pearson, J.; Besley, P. Wave overtopping at vertical and steep seawalls. In *Proceedings of the Institution of Civil Engineers-Maritime Engineering*; Thomas Telford Ltd.: London, UK, 2005; Volume 158, pp. 103–114.
23. Pullen, T.; Allsop, N.; Pearson, J.; Bruce, T. Violent wave overtopping discharges and the safe use of seawalls. In Proceedings of the Defra Flood & Coastal Management Conference, York, UK, 29 June–1 July 2004; Flood Management Division, Department for Environment Food and Rural Affairs: London, UK, 2004.
24. van der Meer, J.W.; Verhaeghe, H.; Steendam, G.J. The new wave overtopping database for coastal structures. *Coast. Eng.* **2009**, *56*, 108–120. [[CrossRef](#)]
25. De Rouck, J.; Van der Meer, J.; Allsop, N.; Franco, L.; Verhaeghe, H. Wave overtopping at coastal structures: Development of a database towards up-graded prediction methods. In *Coastal Engineering 2002: Solving Coastal Conundrums*; World Scientific: Singapore, 2003; pp. 2140–2152.

26. Van der Meer, J. Technical Report Wave Run-Up and Wave Overtopping at Dikes. TAW Report (Incorporated in the EurOtop Manual), 2002. Available online: <http://www.overtopping-manual.com/assets/downloads/TRRunupOvertopping.pdf> (accessed on 31 May 2002).
27. Van der Meer, J.; Allsop, N.; Bruce, T.; De Rouck, J.; Kortenhaus, A.; Pullen, T.; Schüttrumpf, H.; Troch, P.; Zanuttigh, B. *Manual on Wave Overtopping of Sea Defences and Related Structures: An Overtopping Manual Largely Based on European Research, But for Worldwide Application*; EurOtop: London, UK, 2016; p. 264.
28. Steendam, G.J.; Van Der Meer, J.W.; Verhaeghe, H.; Besley, P.; Franco, L.; Van Gent, M.R. The international database on wave overtopping. In *Coastal Engineering 2004: (In 4 Volumes)*; World Scientific: Singapore, 2005; pp. 4301–4313.
29. Sobol, I.M. Sensitivity estimates for nonlinear mathematical models. *Math. Model. Comput. Exp.* **1993**, *1*, 407–414.
30. Saltelli, A.; Tarantola, S.; Chan, K.S. A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics* **1999**, *41*, 39–56. [[CrossRef](#)]
31. Sudret, B. Global sensitivity analysis using polynomial chaos expansions. *Reliab. Eng. Syst. Saf.* **2008**, *93*, 964–979. [[CrossRef](#)]
32. Oakley, J.E.; O'Hagan, A. Probabilistic sensitivity analysis of complex models: A Bayesian approach. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2004**, *66*, 751–769. [[CrossRef](#)]
33. Daneshkhah, A.; Bedford, T. Probabilistic sensitivity analysis of system availability using Gaussian processes. *Reliab. Eng. Syst. Saf.* **2013**, *112*, 82–93. [[CrossRef](#)]
34. Homma, T.; Saltelli, A. Importance measures in global sensitivity analysis of nonlinear models. *Reliab. Eng. Syst. Saf.* **1996**, *52*, 1–17. [[CrossRef](#)]
35. Sacks, J.; Welch, W.J.; Mitchell, T.J.; Wynn, H.P. Design and analysis of computer experiments. *Stat. Sci.* **1989**, *4*, 409–435. [[CrossRef](#)]
36. Kennedy, M.C.; O'Hagan, A. Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2001**, *63*, 425–464. [[CrossRef](#)]
37. O'Hagan, A.; Bernardo, J.M.; Berger, J.O.; Dawid, A.P. *Uncertainty Analysis and other Inference Tools for Complex Computer Codes*; Smith, A.F.M., Dyy, M.C., Oakley, J.E., Eds.; Oxford University Press: Oxford, UK, 1998.
38. Erdik, T.; Savci, M.; Şen, Z. Artificial neural networks for predicting maximum wave runup on rubble mound structures. *Expert Syst. Appl.* **2009**, *36*, 6403–6408. [[CrossRef](#)]
39. van Gent, M.R.; van den Boogaard, H.F.; Pozueta, B.; Medina, J.R. Neural network modelling of wave overtopping at coastal structures. *Coast. Eng.* **2007**, *54*, 586–593. [[CrossRef](#)]
40. Taddy, M.A.; Lee, H.K.; Gray, G.A.; Griffin, J.D. Bayesian guided pattern search for robust local optimization. *Technometrics* **2009**, *51*, 389–401. [[CrossRef](#)]
41. Pianosi, F.; Beven, K.; Freer, J.; Hall, J.W.; Rougier, J.; Stephenson, D.B.; Wagener, T. Sensitivity analysis of environmental models: A systematic review with practical workflow. *Environ. Model. Softw.* **2016**, *79*, 214–232. [[CrossRef](#)]
42. Kennedy, M.; Petropoulos, G. GEM-SA: The Gaussian Emulation Machine for Sensitivity Analysis. In *Sensitivity Analysis in Earth Observation Modelling*; Elsevier: Amsterdam, The Netherlands, 2017; pp. 341–361.
43. Quinonero-Candela, J.; Rasmussen, C.E. A unifying view of sparse approximate Gaussian process regression. *J. Mach. Learn. Res.* **2005**, *6*, 1939–1959.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.