

2021

Prediction of mean wave overtopping at simple sloped breakwaters using kernel-based methods

Shabnam Hosseinzadeh

Amir Etemad-Shahidi
Edith Cowan University

Ali Koosheh

Follow this and additional works at: <https://ro.ecu.edu.au/ecuworkspost2013>




Part of the [Civil and Environmental Engineering Commons](#), and the [Environmental Sciences Commons](#)

[10.2166/hydro.2021.046](https://doi.org/10.2166/hydro.2021.046)

Hosseinzadeh, S., Etemad-Shahidi, A., & Koosheh, A. (2021). Prediction of mean wave overtopping at simple sloped breakwaters using kernel-based methods. *Journal of Hydroinformatics*, 23(5), 1030-1049. <https://doi.org/10.2166/hydro.2021.046>

This Journal Article is posted at Research Online.
<https://ro.ecu.edu.au/ecuworkspost2013/11402>

Prediction of mean wave overtopping at simple sloped breakwaters using kernel-based methods

Shabnam Hosseinzadeh^a, Amir Etemad-Shahidi ^{b,c,*} and Ali Koosheh^b

^a Faculty of Civil Engineering, University of Tabriz, Tabriz, East Azerbaijan Province, Iran

^b School of Engineering and Built Environment, Griffith University, Southport, QLD 4222, Australia

^c School of Engineering, Edith Cowan University, Joondalup, WA 6027, Australia

*Corresponding author. E-mail: a.etemadshahidi@griffith.edu.au

 AE, 0000-0002-8489-7526

ABSTRACT

The accurate prediction of the mean wave overtopping rate at breakwaters is vital for a safe design. Hence, providing a robust tool as a preliminary estimator can be useful for practitioners. Recently, soft computing tools such as artificial neural networks (ANN) have been developed as alternatives to traditional overtopping formulae. The goal of this paper is to assess the capabilities of two kernel-based methods, namely Gaussian process regression (GPR) and support vector regression for the prediction of mean wave overtopping rate at sloped breakwaters. An extensive dataset taken from the EurOtop database, including rubble mound structures with permeable core, straight slopes, without berm, and crown wall, was employed to develop the models. Different combinations of the important dimensionless parameters representing structural features and wave conditions were tested based on the sensitivity analysis for developing the models. The obtained results were compared with those of the ANN model and the existing empirical formulae. The modified Taylor diagram was used to compare the models graphically. The results showed the superiority of kernel-based models, especially the GPR model over the ANN model and empirical formulae. In addition, the optimal input combination was introduced based on accuracy and the number of input parameters criteria. Finally, the physical consistencies of developed models were investigated, the results of which demonstrated the reliability of kernel-based models in terms of delivering physics of overtopping phenomenon.

Key words: ARD-Mattern5/2-Gaussian process regression (GPR), FFBP-artificial neural network (ANN), kernel-based models, mean wave overtopping, RBF-support vector regression (SVR), simple sloped breakwaters

HIGHLIGHTS

- Gaussian process regression (GPR) and support vector regression (SVR) methods were employed to predict the mean wave overtopping rate at simple sloped breakwaters.
- The performances of GPR and SVR models were compared with those of ANN model and existing empirical formulae.
- GPR and SVR models showed better performances compared to those of the ANN model and empirical formulae.
- The optimal input combination with fewer number of input parameters, extracted from sensitivity analysis, and high accuracy was introduced.
- Physical consistency of developed GPR and SVR models were investigated based on the observed trend between the most effective input parameter and mean wave overtopping rate.

INTRODUCTION

Breakwaters are designed to protect harbours and infrastructures against wave attacks. Recently, due to the potential impact of climate change and sea-level rise, the safety and performance of breakwaters have become more important for coastal engineers. Excessive overtopping can also greatly threaten the stability of a breakwater or cause damage to nearby equipment or properties. Conventionally, the mean wave overtopping rate (q) as one of the important hydraulic responses needs to be limited.

During recent decades, several methods have been applied to predict wave overtopping phenomena at coastal structures including numerical, empirical, and soft computing methods. Numerical models (e.g. Losada *et al.* 2008; Neves *et al.*

This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (CC BY 4.0), which permits copying, adaptation and redistribution, provided the original work is properly cited (<http://creativecommons.org/licenses/by/4.0/>).

2008; Ingram *et al.* 2009; Zhang *et al.* 2020) have been used for situations in which empirical test data are limited, or reliable results may not be obtained (van der Meer *et al.* 2017). Nevertheless, the application of numerical models is time-consuming and computationally expensive, especially when high accuracy is required.

The existing empirical formulae to predict mean overtopping rate (e.g. Owen 1980; van der Meer & Janssen 1995; EurOtop 2007; EurOtop 2018; Shaeri & Etemad-Shahidi 2021) have mostly been derived by regression analysis of small-scale experiments. The mentioned formulae correlate the dimensionless mean overtopping rate to dimensionless wave and structural parameters through physical arguments. However, poor predictions of mean overtopping rate at armoured structures using empirical formulae have been reported in the literature (e.g. Koosheh *et al.* 2020). Figure 1 displays the performances of Jafari & Etemad-Shahidi (2011) (hereafter JE), and EurOtop (2018) (mean approach: Equation (6.5)) (hereafter ET), formulae for simple sloped breakwaters. The dimensionless measured and predicted mean overtopping rates defined as $q^* = q/(g \cdot H_{m0,t}^3)^{1/2}$ are shown in this figure. Here, q ($m^3/s/m$) is the dimensional mean overtopping rate per unit width, g (m/s^2) represents the gravitational acceleration, and $H_{m0,t}$ (m) refers to the significant wave height at the toe of the structure. In this figure, the data of rubble mound structures with permeable core and simple slope without crown wall, including both head-on and oblique waves, have been selected from the EurOtop (2018) database. More details of the dataset used are given in the section of the used dataset. As seen, some predictions lie out of 10 times over/under estimation lines (dashed). The ET formula remarkably underestimates overtopping rates, which could be misleading for the design procedure.

In recent decades, several applications of soft computing techniques (e.g. artificial neural network (ANN)) for water engineering problems can be found (e.g. Ayoubloo *et al.* 2010; Kazeminezhad *et al.* 2010; Cini & Deo 2013; Ghaemi *et al.* 2013; Moghaddas *et al.* 2021). These techniques provide a quick and cost-effective solution that can be useful for complicated problems. Due to the complex nature of the overtopping process and the existing limitations of empirical formulae, some soft computing approaches, as alternative tools, have been implemented to predict the mean overtopping rate for a broad range of coastal structures. Among them all, initially developed within the CLASH (De Rouck & Geeraerts 2005) project and presented by EurOtop (2007), the ANN model is the most well-known soft computing tool applicable for a wide range of structures. In the training of this model, dimensional input parameters have been used which may not be appropriate for all cases with different scales. Recently, Zanuttigh *et al.* (2016) developed an improved neural network for a broad range of coastal structures, released in EurOtop (2018) and EurOtop (2018)-ANN, using dimensionless input parameters based on

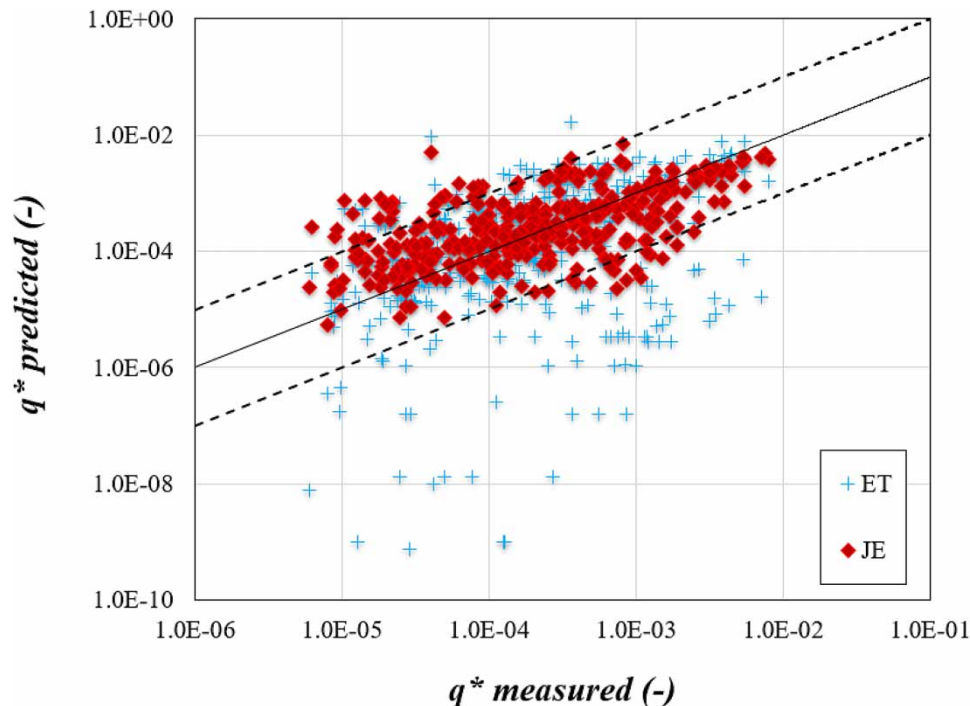


Figure 1 | Comparison of dimensionless measured and predicted overtopping rates $q^* = q/(g \cdot H_{m0,t}^3)^{1/2}$ by Jafari & Etemad-Shahidi (2011) and EurOtop (2018) formulae (the solid line displays perfect agreement and the others demonstrate 10 times over-/underestimations).

an extended database (Zanuttigh *et al.* 2014) mainly derived from the CLASH database. However, in terms of the accuracy of wave overtopping prediction, the recent version does not show a significantly better performance in comparison to CLASH-ANN (Formentin *et al.* 2017; Pillai *et al.* 2017). Besides the mentioned ANN applicable for the wide range of coastal structures, some other studies focused on specific types of structures using soft computing approaches. For example, Molines & Medina (2016) applied ANN to derive an explicit wave overtopping formula for breakwaters with crown wall. However, they achieved the same prediction accuracy compared to CLASH-ANN. The group method of data handling (GMDH) algorithm was used by Lee & Suh (2019) to develop wave overtopping formulae for inclined seawalls. It was shown that GMDH has a better performance compared to the empirical formulae, while its accuracy is similar to that of the EurOtop-ANN model.

This study aims to provide an overview of kernel-based methods, as soft computing tools, to investigate their capabilities for the prediction of mean wave overtopping rate at simple sloped breakwaters. Gaussian process regression (GPR) (Rasmussen & Williams 2006) and support vector regression (SVR) (Vapnik 1995) as kernel-based methods are flexible, as they can handle nonlinear problems. These methods have been recently used in different fields of engineering problems representing promising performance compared to the other soft computing methods (e.g. Ghazanfari-Hashemi *et al.* 2011; Grbić *et al.* 2013; Sun *et al.* 2014; Roushangar & Koosheh 2015; Roushangar *et al.* 2016; Najafzadeh & Oliveto 2020). To the best of the authors' knowledge, SVR and GPR methods have not been implemented for the prediction of the mean wave overtopping rate so far. To develop the models, an extensive dataset selected from the EurOtop (2018) database was used. Also, to evaluate the performances of the kernel-based methods, the results of the analysis were compared with those of ANN, as a benchmark tool for overtopping problems, as well as recently proposed empirical formulae. Moreover, the key variables of overtopping, representing structural and wave conditions features at rubble mound breakwaters, were determined based on sensitivity analysis. To evaluate the reliability of used kernel-based models, a physical consistency test between the key input parameter and mean overtopping rate was investigated. This evaluation was based on a parametric analysis between the most effective input parameter and output one to recognize the existing trend and compare it with the identified physical pattern of overtopping.

MATERIALS AND METHODS

Support vector regression

Initially proposed by Vapnik (1995), the support vector machine (SVM) is commonly implemented for classification purposes in statistical learning problems. In contrast to the conventional neural networks in which empirical risk is minimized, the SVM approach minimizes an upper bound on the expected risk. This equips SVM with a greater ability to generalize, which is the goal of statistical learning (Gunn 1998). In the SVM formulation, the original training data are transformed into a higher dimensional space using nonlinear mapping functions to make data easily separable. Here, the purpose is to find an optimal hyperplane that separates the samples of two classes by considering the widest margin between them within the new space. SVR is an adaption of SVM which can be used as a predictive tool for regression problems. SVR tries to find as flat an optimal function as possible that has the most deviation from the training data while balancing model complexity and prediction error. Since SVR uses a symmetrical loss function with equal penalties for the high and low misestimation, a tube with the radius of ε is formed around the estimation function. In this manner, the points outside of the tube are proportionally penalized to their distance regarding the function. The significant advantage of SVR is that its computational complexity is independent of the dimensionality of input spaces (Awad & Khanna 2015).

The general SVR formulation can be written as follows:

$$f(x) = w\varphi(x) + b \quad (1)$$

where w represents the weight factor, $\varphi(x)$ is known as a nonlinear function in the feature of input x , and b is called the bias. By minimizing regularized risk function, these factors can be obtained as follows:

$$\text{Min } R = C \frac{1}{N} \sum_{i=1}^N L_{\varepsilon}(t_i, y_i) + \frac{1}{2} \|w\|^2 \quad (2)$$

The constant C is the cost factor for performing the trade-off between the weight factor and approximation error. The term $\|w\|^2$ represents the norm of the inner product of the w vector and its transposed form ($w^T \cdot w$). $L_\varepsilon(t_i, y_i)$ is the loss function in which y_i is the predicted value and t_i is the observed value in period i . To ensure the convergence of the optimization process within the finite number of steps, the loss function needs to be symmetric and convex. The simplest loss function is provided in Equation (3) where, for the data out of the tube, the loss will increase linearly:

$$L_\varepsilon(t_i, y_i) = \begin{cases} |t_i - y_i| - \varepsilon: & |t_i - y_i| \geq \varepsilon \\ 0: & \text{otherwise} \end{cases} \quad (3)$$

The slack variables (ξ, ξ^*) are defined to specify the upper and lower training errors subject to an error tolerance ε . Hence, Equation (2) can be re-written in the below form:

$$\text{Min } R = C \sum_{i=1}^N (\xi + \xi^*) + \frac{1}{2} \|w\|^2 \quad (4)$$

$$\text{Subject to: } \begin{aligned} t_i - w_i \varphi(x_i) - b &\leq \varepsilon + \xi_i \\ w_i \varphi(x_i) + b - t_i &\leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* &\geq 0 \end{aligned}$$

The dual Lagrangian form can then be obtained by applying Lagrangian multipliers (α_i and α_i^*) and Karush–Kuhn–Tucher condition:

$$\max L(\alpha_i, \alpha_i^*) = -\varepsilon \sum_{i=1}^N (\alpha_i + \alpha_i^*) + \sum_{i=1}^N t_i (\alpha_i - \alpha_i^*) - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) K(x_i, x_j) \quad (5)$$

$$\text{subject to: } \begin{aligned} \sum_{i=1}^N (\alpha_i - \alpha_i^*) &= 0 \\ 0 \leq \alpha_i \leq C & \quad i = 1, 2, \dots, N \\ 0 \leq \alpha_i^* \leq C & \quad i = 1, 2, \dots, N \end{aligned}$$

As the inner product of two vectors, x_i and x_j in the feature space $\varphi(x_i)$ and $\varphi(x_j)$, kernel function $K(x_i, x_j)$ transforms data into the new space. Several kernel functions, which have their own variable parameters to adjust the flexibility of the regression function, have been implemented in the literature. Obviously, the selection of kernel functions depends on the nature of the data and the problem. However, radial basis function (RBF) was selected for the present study which has been publicly accepted as a good kernel especially for cases without prior knowledge of the data characteristics (Roushangar & Koosheh 2015):

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (6)$$

where σ stands for the kernel parameter. By calculating α_i and α_i^* , the regression function is obtained as follows:

$$f(x) = \sum_{i=1}^N (\alpha_i - \alpha_i^*) K(x_i, x_j) + b \quad (7)$$

The implementation of SVR entails the allocation of an optimization where several parameters such as ε and kernel variables need to be adjusted. Hence, the SVR developed in the MATLAB software was used, and has been equipped with an automated optimization tool.

Gaussian process regression

As a generalization of Gaussian probability (GP) distribution, GPR is a nonparametric and probabilistic approach that can be applied for a variety of nonlinear problems (Rasmussen & Williams 2006). Based on the assumption that the learning sample follows the prior probabilities of GP, the corresponding posterior probability is calculated. GPR uses a kernel to define the covariance of a prior distribution over the target functions. Here, the covariance function plays an important role, as it encodes the prior assumptions about the underlying process that generated the data (Hu & Wang 2015). Assuming $\mathbf{X} \times \mathbf{Y}$ represents the input and output domains from which n pairs (x_i, y_i) are drawn independently and identically distributed. Let f represent an unknown function which maps $\mathbf{X} \rightarrow \mathbf{Y}$. Hence, the regression functional form can be described as follows:

$$y_i = f(x_i) + \varepsilon_i \quad (8)$$

where ε is the Gaussian noise with variance σ_n^2 . The function f can be expressed as a Gaussian process (GP):

$$f(x) \sim \mathcal{GP}(M(x), K(x, x')) \quad (9)$$

GP is a distribution over functions defined by a mean and a covariance function. Conventionally, for the basic GPR, $M(x) = 0$ is assumed to avoid expensive posterior computations (Aye & Heyns 2017). $K(x, x')$ is the covariance (kernel) function by which the dependence between the function values at input points x and x' can be modelled. The expected smoothness and likely patterns in the used data should be considered for the selection of an appropriate kernel (Schulz et al. 2018). After testing different kernels to find a suitable one leading to accurate results, 'ARD (automatic relevance determination) Matern 5/2' kernel was selected for the GPR modelling, the expression of which is as follows:

$$K(x, x') = \sigma_f^2 \left(1 + \sqrt{5}r + \frac{5}{3}r^2 \right) \exp(-\sqrt{5}r) \quad (10)$$

where

$$r = \sqrt{\sum_{m=1}^d \frac{(x_m - x'_m)^2}{\sigma_m^2}} \quad (11)$$

where σ_f represents the single standard deviation, σ_m is the length scale for each predictor m ($m = 1, 2, \dots, d$).

By knowing the observation data $\mathcal{D} = \{\mathbf{X}, \mathbf{Y}\}$, the predictions for the new input \mathbf{X}_* should be obtained by drawing f_* from the posterior distribution $P(f|\mathcal{D})$. The distributions of f_* and \mathbf{Y} , which follow a normal distribution, can be written as follows:

$$\begin{bmatrix} \mathbf{Y} \\ f_* \end{bmatrix} \sim \mathcal{N} \left(0, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) + \sigma_\varepsilon^2 \mathbf{I} & K(\mathbf{X}, \mathbf{X}_*) \\ K(\mathbf{X}_*, \mathbf{X}) & K(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix} \right) \quad (12)$$

where \mathbf{I} is an identity matrix and σ_ε^2 stands for the noise level of observations. By imposing restrictions on the joint prior distribution, the posterior distribution over f_* can be derived.

$$f_* | \mathbf{X}, \mathbf{Y}, \mathbf{X}_* \sim \mathcal{N}(\bar{f}_*, \text{cov}(f_*)) \quad (13)$$

where

$$\bar{f}_* = E[f_* | \mathbf{X}, \mathbf{Y}, \mathbf{X}_*] = M(\mathbf{X}_*) + K(\mathbf{X}_*, \mathbf{X}) [K(\mathbf{X}, \mathbf{X}) + \sigma_\varepsilon^2 \mathbf{I}]^{-1} (\mathbf{Y} - M(\mathbf{X})) \quad (14)$$

$$\text{cov}(f_*) = K(\mathbf{X}_*, \mathbf{X}_*) - K(\mathbf{X}_*, \mathbf{X}) [K(\mathbf{X}, \mathbf{X}) + \sigma_\varepsilon^2 \mathbf{I}]^{-1} K(\mathbf{X}, \mathbf{X}_*) \quad (15)$$

After determining mean and covariance functions, the corresponding hyper-parameters (θ) are still unknown in the GPR formulation and need to be obtained from the training dataset. To estimate the parameters of the GPR model, maximum

likelihood estimation is commonly used (Melo 2012). Based on Bayes' rule, the marginal likelihood can be written as follows:

$$P(\mathbf{Y}|\mathbf{X}) = \int P(\mathbf{Y}|f, \mathbf{X})P(f, \mathbf{X})df \quad (16)$$

Maximizing the log-marginal likelihood gives:

$$\log(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) = \frac{-1}{2}(\mathbf{Y}-M)^T [K(\mathbf{X}, \mathbf{X}) + \sigma_\epsilon^2 \mathbf{I}]^{-1}(\mathbf{Y}-M) - \frac{1}{2} \log|K(\mathbf{X}, \mathbf{X}) + \sigma_\epsilon^2 \mathbf{I}| - \frac{N}{2} \log 2\pi \quad (17)$$

Eventually, the optimal hyper-parameters ($\boldsymbol{\theta}$) can be calculated using the conjugate gradient algorithm (Rasmussen & Williams 2006):

$$\boldsymbol{\theta}' = \arg \max_{\boldsymbol{\theta}} \log P(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) \quad (18)$$

It should be mentioned that the automatic optimization of GPR was developed in the MATLAB software for this study.

Artificial neural network

ANN is a well-known artificial intelligence (AI) model which is based on the framework of the biological human nervous system. It can be used for finding a relationship between the inputs and output called regression. The regression is performed through configuring a flexible architecture of ANN, which consists of input, hidden, and output layers. These layers are connected by neurons where the number of neurons in the input and output layers corresponds to the number of used parameters as inputs and output, while the number of neurons in the hidden layer can be varied to find the best architecture for the problem. In the present study, the criteria used by Pourzangbar *et al.* (2017) for selecting the optimum number of neurons were applied. A three-layer feed-forward (FF) network with the Levenberg–Marquardt back-propagation (BP) training algorithm was utilized for the modelling process. In the feed-forward back-propagation (FFBP) neural network, the term FF illustrates how the neural network process works when the neurons are connected forward, while the BP term points out how the weights of different layers are adjusted in the training procedure using the output estimated by model (Zanganeh *et al.* 2016). The mathematical expression of this network is as follows:

$$y_o = f\left(\sum_{i=1}^n (w_{io}p_i - b_o)\right) \quad (19)$$

where y_o is the output of neuron o , w_{io} represents the weight vector, p_i is the input vector for neuron i ($i = 1, \dots, n$), b_o represents the bias for neuron o , and f is the network transfer function. The tangent sigmoid function is selected, which can be defined as follows (Haykin 2009):

$$f(x) = \frac{2}{(1 + e^{-2x})} - 1 \quad (20)$$

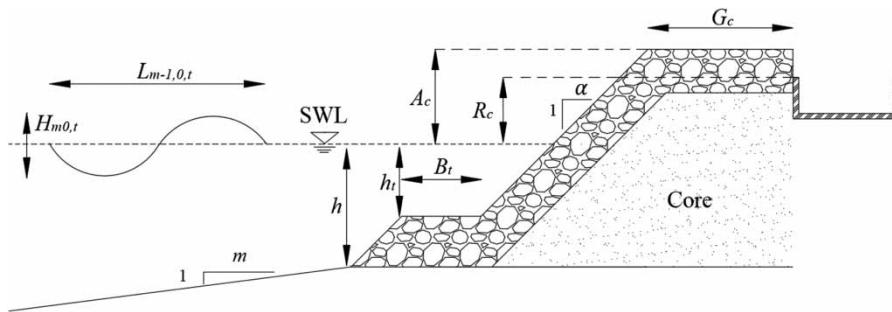
Used dataset

An extensive collected dataset of CLASH (De Rouck & Geeraerts 2005), updated and reorganized later within EurOtop (2018), was employed in the present study. The details of used dimensionless input parameters are given in Table 1.

It should be mentioned that these input parameters were selected based on the EurOtop (2018)-ANN (hereafter ET-ANN) inputs without the berm indicators. In Table 1, $H_{m0,t}$ and $L_{m-1,0,t}$ represent significant wave height and wavelength at the toe of the structure, respectively. Here, $L_{m-1,0,t}$ is $1.56T_{m-1,0,t}^2$ where $T_{m-1,0,t}$ is the spectral wave period at the toe of the structure. h is the water depth at the toe of structure, h_i is the toe submergence depth, and B_t is the toe width. The parameters R_c , A_c , and G_c are the crest freeboard, crest height, and crest width, respectively. In addition, D stands for the average size of the structural elements in the run-up/down area. The symbols relevant to the overtopping are shown in Figure 2.

Table 1 | The range and definitions of dimensionless used input parameters

Input	Type	Representation of	Range
$S_{m-1,0,t} = H_{m0,t}/L_{m-1,0,t}$	Wave attack	Wave steepness (breaking)	0.002–0.071
β (°)	Wave attack	Wave obliquity	0–60
$h/L_{m-1,0,t}$	Wave attack	Shoaling parameter	0.009–0.731
$h_t/H_{m0,t}$	Geometry	Effect of the toe submergence	0.949–14.404
$B_t/L_{m-1,0,t}$	Geometry	Effect of the toe width	0–0.14
$R_c/H_{m0,t}$	Geometry	Relative crest freeboard	0.096–2.617
$A_c/H_{m0,t}$	Geometry	Relative armoured freeboard	0.096–2.617
$G_c/L_{m-1,0,t}$	Geometry	Relative crest width	0–0.425
m	Geometry	Foreshore slope	12–1,000
$\cot \alpha$	Geometry	Slope of the structure	1.33–2
γ_f	Structural features	Roughness factor	0.38–0.5
$D/H_{m0,t}$	Structural features	Indication of structure stability	0.131–2.682

**Figure 2** | Schematic diagram of simple sloped breakwater.

To refine the permeable simple sloped breakwaters, small-scale records ($H_{m0,t} \leq 0.5$) with the roughness factors $\gamma_f < 0.6$ (except $\gamma_f = 0.55$) and mild slope ($1.33 \leq \cot \alpha \leq 2$) were chosen. The factors RF and CF, varying from 1 to 4, represent the reliability and complexity level of given data. These factors somehow describe the measurements accuracy of each test or how well the geometry of a structure could be described by the geometrical parameters. Records with the lowest reliability (RF = 4) and the highest level of complexity (CF = 4) were ignored to consider only good quality data in the analysis (see also Etemad-Shahidi & Jafari 2014; Shaeri & Etemad-Shahidi 2021). Low overtopping rates ($q \leq 1 \times 10^{-6} \text{ m}^3/\text{s}/\text{m}$) were also removed, as they may be affected by measurement errors (e.g. Verhaeghe *et al.* 2008; Etemad-Shahidi *et al.* 2016). The records with the emerged crest ($R_c > 0$) and simple slopes ($\cot \alpha_u = \cot \alpha_d$), without berm ($B = 0$), and the crown wall ($R_c \leq A_c$) were selected. In this way, a total number of 1,220 small-scale records remained for further analysis: 70% of these selected for the training, and the rest were used for the testing. It should be mentioned that the selected data for permeable simple sloped rubble mound structures include rock permeable straight slopes denoted by the label 'A', armour units straight slopes represented by the label 'C', and oblique wave attack with the label 'G' (see Zanuttigh *et al.* (2016) for details). Table 2 provides the details of used data.

As the perfect reproduction of wave and structure interaction is not possible in the small-scale physical models in a laboratory, the existence of scale effects is unavoidable. This is because the simultaneous fulfilment of scaling laws or similarity principle (i.e. Froude and Reynolds) is unachievable in the physical modelling. Therefore, a significant difference between field and model measurements of overtopping rate on rubble mound structures, especially for low rates, has been reported in EurOtop (2018). This difference is more considerable for the longer and flatter slopes where the zero overtopping is predicted in the laboratory for an overtopped prototype situation (Koosheh *et al.* 2021). Thus, using field measurements along

Table 2 | Details of used data extracted from the EurOtop (2018) database

Data label	Number of data	Data label	Number of data
A-2	32	C-25	5
A-3	84	C-26	6
A-5	4	C-27	6
A-14	3	C-28	6
A-33	13	C-29	6
A-35	146	C-30	6
A-36	2	C-31	5
A-38	62	C-40	14
A-39	60	C-41	47
A-42	18	C-44	10
A-43	8	C-45	5
C-1	12	C-46	6
C-2	12	C-47	1
C-3	10	C-51	107
C-5	12	C-58	25
C-6	13	G-2	192
C-8	25	G-3	157
C-9	11	G-4	56
C-10	13	G-10	3

with small-scale data to develop the models can lead to model confusion due to scale effects (e.g. Jafari & Etemad-Shahidi 2011; see Supplementary Appendix A for details). For this reason, the large-scale and field measurements were excluded, and only small-scale laboratory tests were selected to develop the models. To generalize the developed models for the large-scale and field measurements, the scale effect correction proposed by EurOtop (2018) for rubble mound structures can be applied. This correction is derived from the prototype and laboratory observations and suggests an increasing ratio ($f_q \geq 1$) for upscaled wave overtopping only when the discharge (q_{up}) is smaller than about 1×10^{-3} (m³/s/m). This adjustment factor also depends on the slope of the structure (see EurOtop (2018) for details).

Modelling overview

When sea waves run-up above the coastal structures and water overflows, wave overtopping occurs. The mean overtopping rate is defined as the average discharge per metre width of the structure and commonly expressed in m³/s/m. This parameter as the response of structure against incident wave depends on its geometrical features such as crest freeboard and seaward slope but also on local wave conditions such as wave height, wave period, and water depth. The dimensionless mean overtopping rate ($q^* = q / (g \cdot H_{m0,t}^3)^{1/2}$) is usually correlated to the dimensionless form of the shown parameters to generalize the results. The mentioned dimensionless overtopping rate is employed based on the assumption of critical flow conditions on the crest of the structure (Altomare *et al.* 2020). The selection of input parameters is an important step for modelling. The appropriate combinations of the most effective parameters can enhance the performance of models. Table 3 shows the used input combinations to feed GPR, SVR, and ANN models. These input combinations are taken from ET-ANN (combinations a and b) and existing empirical formulae such as JE and ET (combinations c and d). Regarding the specific studied structure, the berm indicators were excluded for configuring the input combination a. The key point for the configuration of the used dimensionless input parameters in the combination a, as the most comprehensive one, is using the significant wave height ($H_{m0,t}$) to scale the structure heights (A_c , h_t , and R_c) as well as using wave length ($L_{m-1,0,t}$) to scale structure widths (B_t and G_c). The wave dissipation caused by breaking wave on the toe of structure (h_t) and possible wave overtopping (R_c and A_c) can be considered using $H_{m0,t}$ as a height-scaling parameter. In addition, two key procedures such as breaking by steepness and shoaling described by $S_{m-1,0,t}$ and $h/L_{m-1,0,t}$, respectively, were considered in this input combination. To achieve the modelling with

Table 3 | Used input combinations for developing the GPR, SVR, and ANN models

Input combinations	
(a)	$\frac{D}{H_{m0,t}}, m, \frac{A_c}{H_{m0,t}}, \frac{B_t}{L_{m-1,0,t}}, \frac{h_t}{H_{m0,t}}, \frac{h}{L_{m-1,0,t}}, \frac{R_c}{H_{m0,t}}, \frac{G_c}{L_{m-1,0,t}}, S_{m-1,0,t}, \tan \alpha, \cos \beta, \gamma_f$
(b)	$\frac{h}{L_{m-1,0,t}}, \frac{R_c}{H_{m0,t}}, \frac{G_c}{L_{m-1,0,t}}, S_{m-1,0,t}, \tan \alpha, \cos \beta, \gamma_f$
(c)	$\frac{R_c}{H_{m0,t}}, \frac{G_c}{H_{m0,t}}, \tan \alpha, R^*$
d)	$\frac{R_c}{H_{m0,t} \cdot \gamma_\beta \cdot \gamma_f}$

as few as possible input parameters, the most influential parameters reported in the literature (Pillai *et al.* 2017) and extracted from sensitivity analysis were considered to configure input combination b.

The results of sensitivity analysis to identify the key governing parameters are shown in Table 4. According to this table, the prediction errors were estimated by eliminating each parameter one by one. If the elimination of a parameter does affect the results marginally, that parameter could be neglected for further modelling; otherwise, the parameter needs to be included. For example, by elimination of some input parameters such as $A_c/H_{m0,t}$, $D/H_{m0,t}$, m , $B_t/L_{m-1,0,t}$, $h_t/H_{m0,t}$, no significant change was observed in the accuracy metrics. On the other hand, by the elimination of parameters such as wave steepness ($S_{m-1,0,t}$) or oblique wave factor (γ_β), centred pattern-root-mean-square error (c-RMSE) increases by 23 and 50%, respectively. This implies that the mentioned parameters should be used in the modelling as key parameters. Among all input parameters, the relative crest freeboard ($R_c/H_{m0,t}$) is the most effective one as modelling excluding this parameter can lead to large prediction errors (Model 13: c-RMSE = 0.54 and BIAS = -0.06).

Input combinations c and d were selected to fairly compare soft computing models with empirical formulae (JE and ET, respectively) using the same input parameters. Here, the mathematical expressions of used formulae are given. For rubble mound structures, EurOtop (2018) proposed a simple exponential formula as follows:

$$\frac{q}{\sqrt{g} \cdot H_{m0}^3} = 0.09 \cdot \exp \left[- \left(1.5 \frac{R_c}{H_{m0} \cdot \gamma_f \cdot \gamma_\beta} \right)^{1.5} \right] \tag{21}$$

Table 4 | Results of sensitivity analysis

		Models												
		1	2	3	4	5	6	7	8	9	10	11	12	13
Inputs	$A_c/H_{m0,t}$	✓	–	–	–	–	–	–	–	–	–	–	–	–
	$D/H_{m0,t}$	✓	✓	–	–	–	–	–	–	–	–	–	–	–
	m	✓	✓	✓	–	–	–	–	–	–	–	–	–	–
	$B_t/L_{m-1,0,t}$	✓	✓	✓	✓	–	–	–	–	–	–	–	–	–
	$h_t/H_{m0,t}$	✓	✓	✓	✓	✓	–	–	–	–	–	–	–	–
	$h/L_{m-1,0,t}$	✓	✓	✓	✓	✓	✓	–	✓	✓	✓	✓	✓	✓
	$G_c/L_{m-1,0,t}$	✓	✓	✓	✓	✓	✓	✓	–	✓	✓	✓	✓	✓
	$S_{m-1,0,t}$	✓	✓	✓	✓	✓	✓	✓	✓	–	✓	✓	✓	✓
	$\tan \alpha$	✓	✓	✓	✓	✓	✓	✓	✓	✓	–	✓	✓	✓
	γ_β	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	–	✓	✓
	γ_f	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	–	✓
	$R_c/H_{m0,t}$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	–
	Error metrics	BIAS	0.01	0	0	0	0.01	0	0	-0.04	-0.02	0	-0.02	0
c-RMSE		0.26	0.26	0.26	0.28	0.28	0.28	0.32	0.35	0.32	0.30	0.39	0.31	0.54

where γ_β and γ_f are the reduction factors of wave obliquity and structure's surface roughness, respectively. For head-on waves, γ_β is assumed equal to one, and γ_f can be calculated as follows:

$$\gamma_\beta = \begin{cases} \cos^2(|\beta| - 10^\circ) & \text{with a minimum of } \gamma_\beta = 0.6 \text{ (long-crested)} \\ 1 & \text{for } |\beta| = 0-10^\circ \end{cases} \quad (22)$$

Likewise, for smooth structures, the roughness factor (γ_f) is equal to one and for other types can be determined based on used materials and the structure's permeability (EurOtop 2018). The roughness factor also needs to be modified when $Ir_{m-1,0} > 5.0$ which increases linearly up to 1 at $Ir_{m-1,0} = 10$ as below:

$$\gamma_{f\text{mod}} = \gamma_f + \frac{(Ir_{m-1,0} - 5)(1 - \gamma_f)}{5} \quad (23)$$

where $Ir_{m-1,0}$ demonstrates the Iribarren number based on $T_{m-1,0,t}$ defined as $\tan \alpha / \sqrt{S_{m-1,0,t}}$ where $S_{m-1,0,t}$ is the wave steepness defined by $H_{m0,t}/L_{m-1,0,t}$. It should be mentioned that the Iribarren number represents the wave breaking condition ($Ir_{m-1,0} < 1.8$) and non-breaking condition ($Ir_{m-1,0} > 1.8$) (EurOtop 2018). Based on Equation (23), the roughness factor for the data with $Ir_{m-1,0,t} > 5$ (around 6% of all data) was modified for all input combinations applied in this study.

Jafari & Etemad-Shahidi (2011) suggested multi-conditional formulae for rubble mound structures using the CLASH database as follows:

$$\begin{cases} \text{if } \frac{R_c}{H_{m0}} > 2.08 \text{ and } \frac{G_c}{H_{m0}} > 1.51; & \frac{q}{\sqrt{g \cdot H_{m0}^3}} = \exp(-0.6396 R^* - 0.7085 \tan \alpha - 11.4897) \\ \text{if } R^* \leq 0.86; & \frac{q}{\sqrt{g \cdot H_{m0}^3}} = \exp(-6.18 R^* - 3.21) \\ \text{if } R^* > 0.86; & \frac{q}{\sqrt{g \cdot H_{m0}^3}} = \exp(-3.1 R^* - 6.05 \tan \alpha - 2.63) \end{cases} \quad (24)$$

where H_{m0} is the significant wave height at the toe of structure and R^* is defined as follows:

$$R^* = \frac{R_c}{H_{m0} \cdot \gamma_\beta \cdot \gamma_f} \times \frac{\sqrt{S_{\text{op}}}}{\tan \alpha} \quad (25)$$

As seen, all used input parameters are dimensionless to unify the whole dataset regardless of different model scales of tests. Moreover, employing dimensionless parameters improves the analysis, fitting, and interpretation of results as well as the generalization capacity of the developed model. It should be mentioned that the dimensionless wave overtopping rate in the form of $q^* = q/(g \cdot H_{m0,t}^3)^{1/2}$ was selected as the output of the models.

Performance measures

The capabilities of the models were evaluated using the discrepancy ratio (DR) and accuracy metrics of modified Taylor's diagram (Elvidge *et al.* 2014), i.e. standard deviation (σ), Pearson's correlation coefficient (R), c-RMSE, and BIAS. Taylor's diagram, initially proposed by Taylor (2001), is a mathematical diagram which graphically illustrates how realistic the models are and simplifies the comparison process. This is obtained by finding a geometric relation between standard deviation, c-RMSE, and Pearson's correlation coefficient. c-RMSE defined as mean-removed RMSE and represented by E can be calculated as follows:

$$E^2 = \frac{1}{n} \sum_{i=1}^n [(\log q_{pi}^* - \bar{Y}) - (\log q_{mi}^* - \bar{X})]^2 \quad (26)$$

Indeed, c-RMSE can be equated with the standard deviation of the model error based on the mathematical operations applied to the above equation as follows:

$$E^2 = \frac{1}{n} \sum_{i=1}^n [(\log q_{pi}^* - \log q_{mi}^*) - (\bar{Y} - \bar{X})]^2 \quad (27)$$

$$E^2 = \frac{1}{n} \sum_{i=1}^n [(\log q_{pi}^* - \log q_{mi}^*) - (\bar{Y} - \bar{X})]^2 \quad (28)$$

where q_m^* and q_p^* are dimensionless measured and predicted overtopping rates, n is the number of records, and \bar{X} and \bar{Y} are the average values of $\log q_m^*$ and $\log q_p^*$, respectively. c-RMSE is always non-negative where a value of zero represents the perfect fit of prediction to the measured data. In addition, given that c-RMSE is a scale-dependent parameter and the target parameter in the present study is dimensionless (q^*), the c-RMSE value will be dimensionless.

The used metrics in Taylor's diagram are related by the following equation:

$$E^2 = \sigma_p^2 + \sigma_m^2 - 2\sigma_p\sigma_mR \quad (29)$$

where σ_p and σ_m represent the standard deviation of predicted and measured values, respectively, and R is the Pearson correlation coefficient. These parameters are expressed as follows:

$$\sigma_p = \sqrt{\frac{1}{n} \sum_{i=1}^n (\log q_{pi}^* - \bar{Y})^2} \quad (30)$$

$$\sigma_m = \sqrt{\frac{1}{n} \sum_{i=1}^n (\log q_{mi}^* - \bar{X})^2} \quad (31)$$

$$R = \frac{\sum_{i=1}^n (\log q_{mi}^* - \bar{X})(\log q_{pi}^* - \bar{Y})}{\sqrt{\sum_{i=1}^n (\log q_{mi}^* - \bar{X})^2} \sqrt{\sum_{i=1}^n (\log q_{pi}^* - \bar{Y})^2}} \quad (32)$$

Standard deviation represents the amount of dispersion of a set of values in comparison to the average expected value. A low value of standard deviation points out the closeness of values to the mean of the set, while a high standard deviation illustrates that there are widespread values around the average value. R is a measure of linear correlation between two sets of data. The value of the correlation coefficient shows the strength of the relationship between the measured values and predicted ones.

The law of cosines (Pickover 2009) defined by $c^2 = a^2 + b^2 - 2ab \cos \varphi$ (where a , b , and c represent triangle sides and φ is the angle between the sides a and b) is a key to forming the geometrical connection between four quantities (σ_p , σ_m , R , and E) which underlie the Taylor diagram (Figure 3).

Elvidge *et al.* (2014) proposed a modified Taylor's diagram in which BIAS as a complementary accuracy metric was added in contours. The BIAS can be calculated as follows:

$$\text{BIAS} = \frac{1}{n} \sum_{i=1}^n (\log q_{pi}^* - \log q_{mi}^*) \quad (33)$$

BIAS is a systematic error that is achieved from an estimation process not giving accurate results on average. The positive and negative BIAS values show the overestimation or underestimation of the modelling where for the best fit, the BIAS value

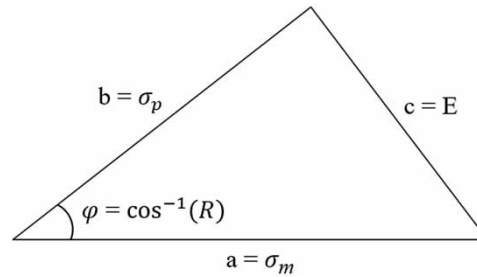


Figure 3 | Geometrical relationship between metrics plotted on Taylor's diagram based on the cosines law.

will be equal to zero. Also, the mathematical formulation of another used metric, namely DR, can be expressed as follows:

$$DR_i = \frac{q_{pi}^*}{q_{mi}^*} \quad (34)$$

where $DR_i = 1$ means that there is an exact match between the measured and predicted values.

RESULTS AND DISCUSSION

The scatterplots of measured values against the predicted ones for all input combinations and the test data using all applied models (GPR, SVR, and ANN) and empirical formulae (JE and ET) are given in [Figure 4](#). Here, the solid line shows the perfect agreement, and the dashed lines represent 10 and 0.1 times over-/underestimations, respectively. As can be seen, for all input combinations, predicted values by soft computing methods are less scattered than those by empirical formulae. The plot given for input combination b is slightly more scattered than that for input combination a. In the case of input combinations of a and b, all predicted values by the GPR model lie between over-/underestimation lines which show its higher accuracy. [Figures 4\(c\)](#) and [4\(d\)](#) compare the prediction performance of soft computing models against empirical formulae using the same input parameters. As can be seen, the ET formula underestimates the low overtopping rates significantly, while slight underestimations and some overestimations were observed for the JE formula.

The accuracy metrics of GPR, SVR, ANN models, and empirical formulae for both all and test datasets are presented in [Table 5](#). Here, lowercase letters a, b, c, and d correspond to the input combinations given in [Table 3](#). The capital letters G, S, and N also denote GPR, SVR, and ANN methods, respectively.

According to [Table 5](#), it could be inferred that models using input combinations of a and b (taken from [EurOtop \(2018\)](#)-ANN) provide better results in comparison to other models (using input of empirical formulae). As an example, both G(a) and G(b) models with $c\text{-RMSE} = 0.24$ and 0.25 are more accurate compared with G(c) and G(d) models with $c\text{-RMSE} = 0.31$ and 0.62 , respectively. Models with input combination b, as the reduced form of input combination a, show an acceptable accuracy. For G models, almost similar accuracy metrics were obtained for input combinations a and b with a slight difference in the $c\text{-RMSE}$. Since the eliminated parameters in the input combination b are mostly the less influential ones in the overtopping process (at least for study case), the results could be expected. The input combination b consists of all conventionally known effective parameters namely $R_c/H_{m0,t}$, $S_{m-1,0,t}$, $\tan \alpha$, $\cos \beta$, and γ_f . Besides the mentioned parameters, the parameter $(h/L_{m-1,0,t})$ has an effective contribution in the modelling which is supported by the findings of [Cheon & Suh \(2016\)](#) and [Pillai et al. \(2017\)](#). In addition, since breakwaters have a permeable crest, the position of the box collecting overtopping on it can be a significant factor for the measurement ([Jafari & Etemad-Shahidi 2011](#)). This issue can be considered by using the relative crest width $(G_c/L_{m-1,0,t})$ by the increase of which the overtopping is expected to decrease as water percolates into the permeable surface ([Pillai et al. 2017](#)). The comparison of models with input combination b demonstrates the highest accuracy for model G ($c\text{-RMSE} = 0.25$). However, the good performance of model S with $\text{BIAS} = 0$ should not be overlooked. The improvements of the model G(b) compared to N(b), JE, and ET formulae can be accounted for 22, 60, and 79% in terms of $c\text{-RMSE}$ value.

In general, soft computing models show better performance than empirical formulae. This superiority can be seen even in the cases where the same input parameters are used. For example, the models G(c), S(c), and N(c) have the $c\text{-RMSE}$ values of 0.31, 0.37, and 0.4, while JE gives $c\text{-RMSE} = 0.63$. Comparing the accuracy metrics of the most accurate soft computing

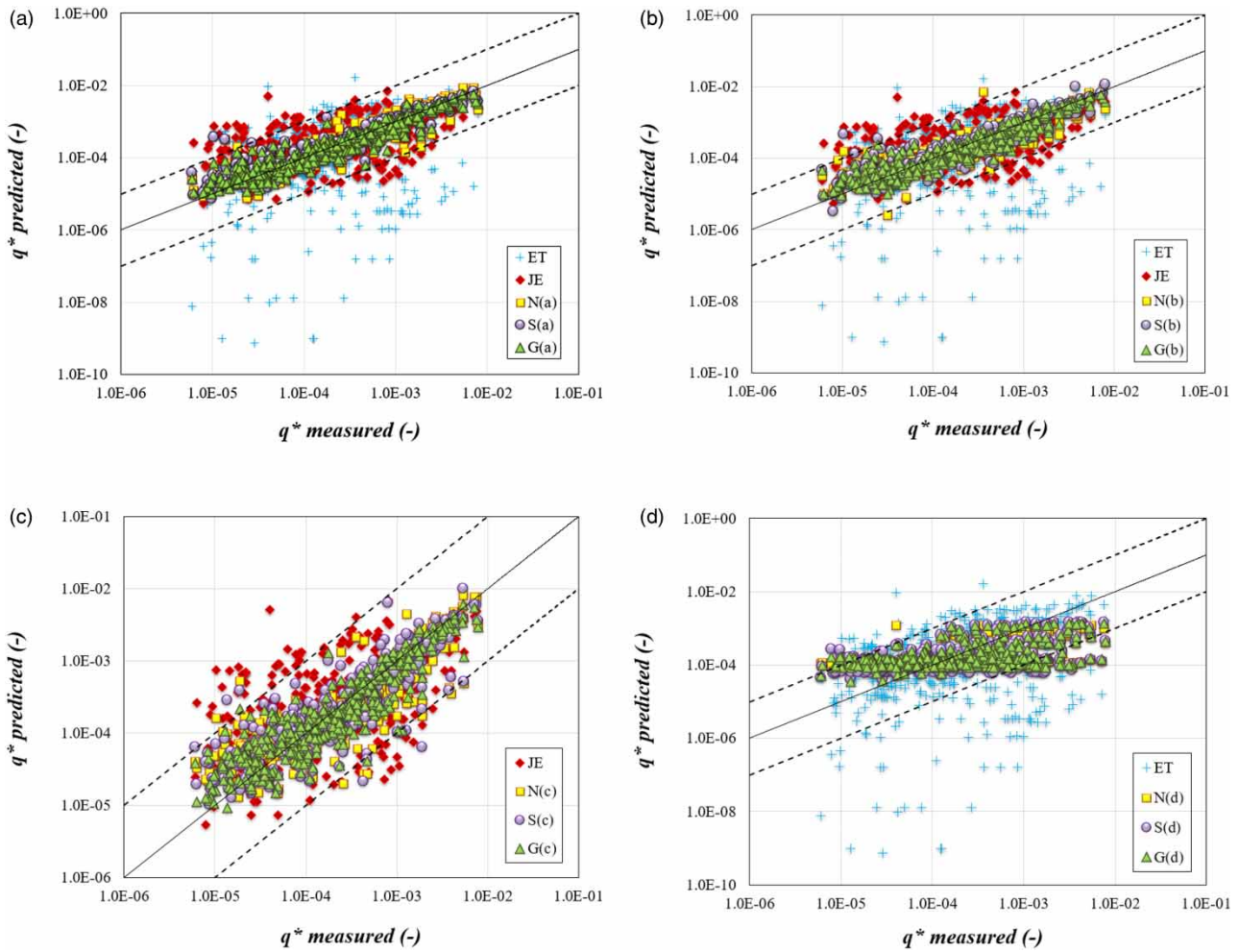


Figure 4 | Measured versus predicted overtopping rates using JE and ET formulae as well as ANN (N), SVR (S), and GPR (G) models for the input combinations a (a), b (b), c (c), and d (d); test data.

Table 5 | Accuracy metrics of different developed models (GPR, SVR, and ANN) and empirical formulae for test (all) data

Models	Error metrics	
	c-RMSE	BIAS
G(a)	0.24 (0.19)	-0.02 (-0.01)
G(b)	0.25 (0.22)	-0.02 (-0.01)
G(c)	0.31 (0.27)	-0.02 (-0.01)
G(d)	0.62 (0.63)	-0.04 (-0.01)
S(a)	0.26 (0.22)	0.01 (0.01)
S(b)	0.28 (0.24)	0 (0.01)
S(c)	0.37 (0.33)	-0.01 (0)
S(d)	0.63 (0.63)	-0.07 (-0.04)
N(a)	0.28 (0.24)	0.01 (0.01)
N(b)	0.32 (0.30)	-0.01 (0)
N(c)	0.40 (0.40)	-0.02 (0)
N(d)	0.63 (0.64)	-0.04 (-0.01)
JE	0.63	0.13
ET	1.20	-0.29

method for the input combination c, G(c) with the JE formula representing the improvements of about 51 and 85% in terms of c-RMSE and BIAS, respectively. Developed models using input combination d (taken from the ET formula) result in a better accuracy metrics compared to ET. For example, c-RMSE has been reduced from 1.20 (ET) to 0.62 (G(d)). However, the comparison of Figure 4(d) and other panels demonstrates the unsuitability of this input combination, as it lacks some key parameters compared to other input combinations. It should also be mentioned that for input combinations c and d, GPR models outperform SVR and ANN models. Besides the better performance of JE than ET formula, soft computing models fed by input combination of c, taken from the JE formula, perform better compared with the input combination in which ET parameters have been used (input combination d). This can be explained by overlooking some parameters in the input combination d such as $S_{m-1,0,t}$, $\tan \alpha$, and $G_c/H_{m-1,0,t}$.

Overall, considering both accuracy and the number of input parameters, input combination b can be introduced as the optimal one. In addition, comparing the results of the analysis of all applied models for the test dataset represents the good capability of kernel-based models compared to ANN and empirical formulae. It can also be seen that the GPR model performs slightly better than the SVR model.

Figure 5 shows the used modified Taylor diagram which graphically summarizes the results of the analysis. The advantage of using modified Taylor's diagram is plotting all meaningful accuracy metrics in one diagram, which can be more helpful for the comparison of the models. In this diagram, (1) the azimuthal angle shows the correlation coefficient, (2) the radial distance represents the standard deviation of models, (3) the blue-coloured dashed line shows the standard deviation of measured values, (4) the red-coloured circular dashed lines with the centre of measured standard deviation inside the diagram display the c-RMSE, (5) BIAS is presented by contours. Predicted patterns, which are in good agreement with the observations, will lie nearest the point marked 'Measured' on the horizontal axis. This point is the representation of the highest correlation ($R = 1$) and lowest c-RMSE ($=0$) and a similar dispersion pattern of predicted values compared to the measured ones ($\sigma_p = \sigma_m = 0.72$).

As can be seen, each diagram (a, b, c, and d) refers to the used input combinations in this study. According to diagram (a), it is evident that the GPR and SVR models agree best with observations by the lowest c-RMSE and highest correlation coefficient (R), respectively. However, the spatial variability of the ANN model is lower compared to the others, as it is close to the blue-dashed line. Also, the ANN and SVR models in yellow show few overestimations, while the GPR model in green indicates the underestimation. Attending to diagram (b), the relative merit of kernel-based models, especially GPR, can be inferred from the location of the models. As seen, the GPR model is in the nearest spot to the measured point on the horizontal axis in the case of either radial distance (close standard deviation to the measured values) or azimuthal angle (lowest c-RMSE and highest correlation coefficient). However, according to the BIAS contour, SVR and ANN models show slight overestimations, while GPR is underestimated. Given that the differences of BIAS values for the models regardless of their over-/underestimations are negligible, the GPR model can be considered as the most accurate one for the input combination b. In diagrams (c) and (d), the applied soft computing models are compared with the empirical formulae. Based on diagram (c), the location of the JE formula, standing further than those of the applied soft computing models with respect to the optimal point on the horizontal axis, confirms its unreliable estimation. Also, the superiority of used soft computing models can be obviously seen in diagram (d) where the point representing the ET formula is in the furthest location from the measured point on the horizontal axis.

Overall, Figure 5 demonstrates the higher capability of kernel-based models in comparison to ANN and empirical formulae in the prediction of wave overtopping at simple sloped breakwaters. G(a) and G(b), as the most accurate ones, are the nearest to the measured point on the horizontal axis.

The distribution of $\log(\text{DR})$ for the predicted values using the input combination b was further analysed (Figure 6). The narrower distribution of predictions using all soft computing methods especially GPR ($-1 < \log \text{DR} < 1$) in comparison with JE ($-2 < \log \text{DR} < 2$) and ET (less than $-2 < \log \text{DR} < 2$) formulae can be observed. Also, as seen all models (except the JE formula) have negative skewness indicating underestimation, while the JE formula with positive skewness illustrates overestimation.

A good model is a model with errors that are independent of the input parameters (Sahay & Dutta 2009) and has no systematic error. Hence, the variation of DR as a function of the relative crest freeboard ($R_c/H_{m0,t}$) is shown in Figure 7 for the input combination b as optimal one using different models (JE, ET, ANN, SVR, and GPR). As shown, it can be concluded that the relative crest freeboard ($R_c/H_{m0,t}$) has been used appropriately in all models except the ET formula where a systematic error is observed. DR values of applied models (GPR, SVR, and ANN) are less sensitive to the change of relative crest

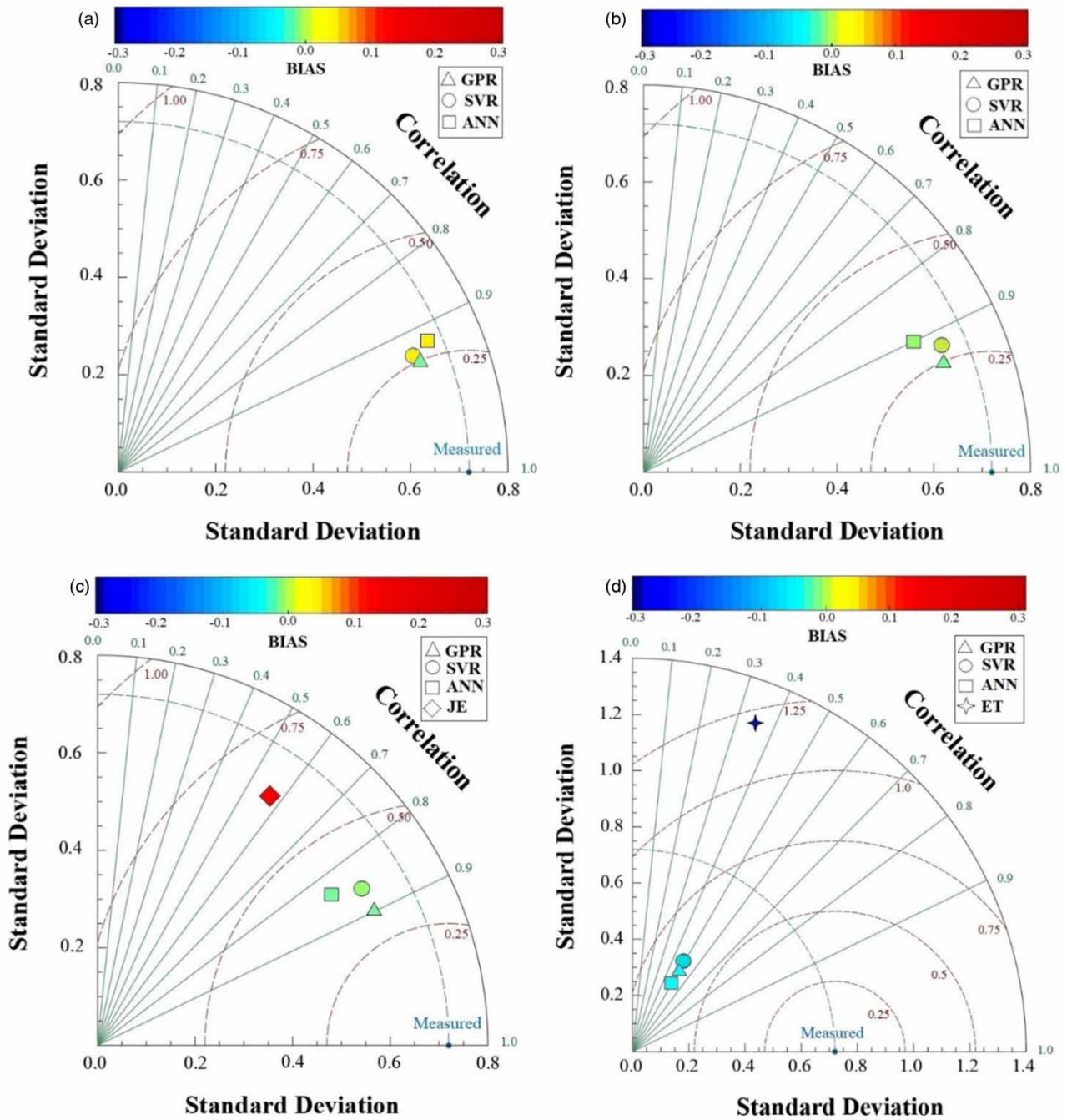


Figure 5 | Modified Taylor's diagrams for the input combinations a (a), b (b), c (c), and d (d); test data. Please refer to the online version of this paper to see this figure in colour: <http://dx.doi.org/10.2166/hydro.2021.046>.

freeboard as well as more symmetric than those of the ET formula. In addition, comparing the dispersion of the data points around $DR = 1$ for all models indicates the good capability of soft computing methods especially the GPR for the prediction of overtopping rate.

To investigate the physical consistency of developed GPR and SVR models, a parametric analysis showing the relationship between the most important input parameter ($R_c/H_{m0, t}$), mentioned in the literature and extracted from sensitivity analysis, and overtopping rate was conducted. Figure 8 represents the predicted values of dimensionless mean overtopping rate against relative crest freeboard ($R_c/H_{m0, t}$) for GPR and SVR using the input combination b as the optimal one. As seen, a decreasing

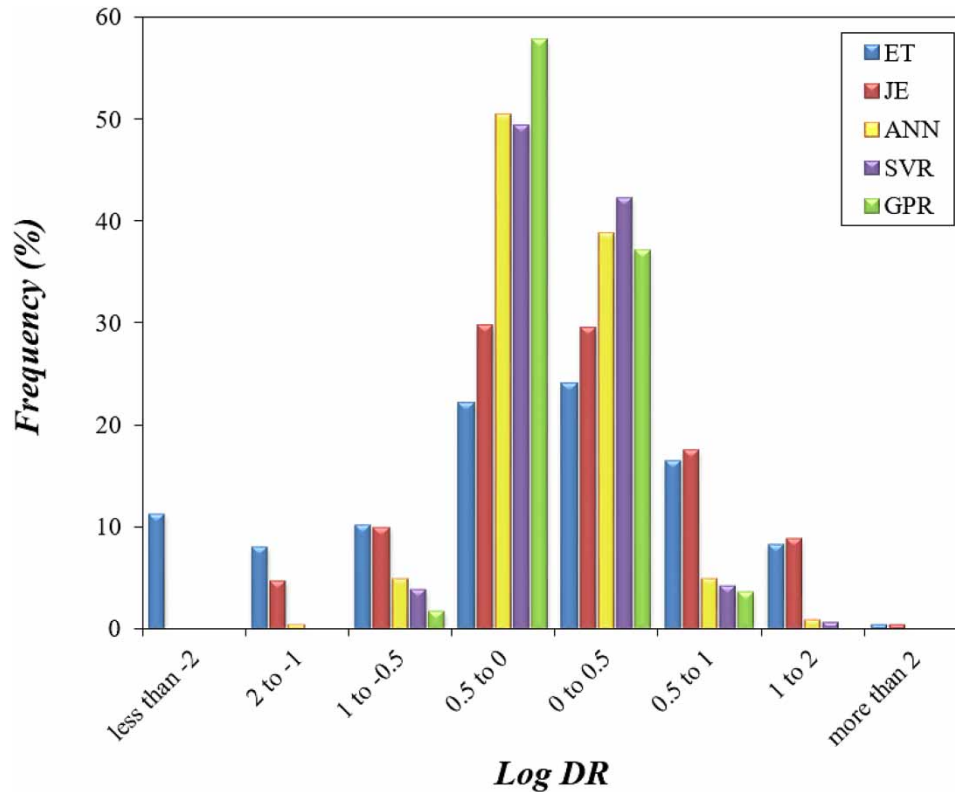


Figure 6 | Histogram of log DR for input combination b, different soft computing methods (ANN, SVR, and GPR) and empirical formulae (JE and ET); test data.

trend between the relative crest freeboard and mean overtopping rate is observed for both developed models demonstrating good agreement with the existing physical pattern.

To introduce an appropriate model, different criteria such as accuracy, simplicity, and computational cost should be considered. For developing kernel-based models, the manual adjustment of structure is not required where the optimal structure is obtained through an automatic process. This feature makes them more user-friendly, especially for those who are not quite familiar with the optimization process. In addition, regarding the good prediction accuracies of the kernel-based models compared to ANN, these models can be applied as efficient soft computing tools for the estimation of wave overtopping at coastal structures. Moreover, similar to most of the other soft computing models (e.g. Zanuttigh *et al.* 2016), the kernel-based models do not provide formulas but can be used in practice considering dimensionless parameters as the input. The m. file (MATLAB) of the developed models is provided as the supplementary file to be used by practitioners.

SUMMARY AND CONCLUSION

In this study, kernel-based methods (GPR and SVR) were employed to estimate the mean wave overtopping rate at simple sloped breakwaters. To investigate the capability of kernel-based models, the ANN method as a well-known soft computing tool as well as recently proposed empirical formulae (JE and ET) were applied as benchmarks. The existing laboratory tests from the EurOtop (2018) database were used for the modelling process. Conventionally used wave and structural parameters in the existing models were selected to define different input combinations. A sensitivity analysis was performed to recognize the most important parameters to configure the optimal input combination. To evaluate the reliability of kernel-based models in terms of physical consistency, a parametric analysis representing the simulated trend between the relative crest freeboard ($R_c/H_{m0,t}$), as the most important parameter, and overtopping rate using GPR and SVR was carried out.

According to the obtained results, the main findings of this study can be summarized as below:

- Input combinations taken from EurOtop (2018)-ANN can lead to more accurate predictions in comparison to those obtained from empirical formulae.

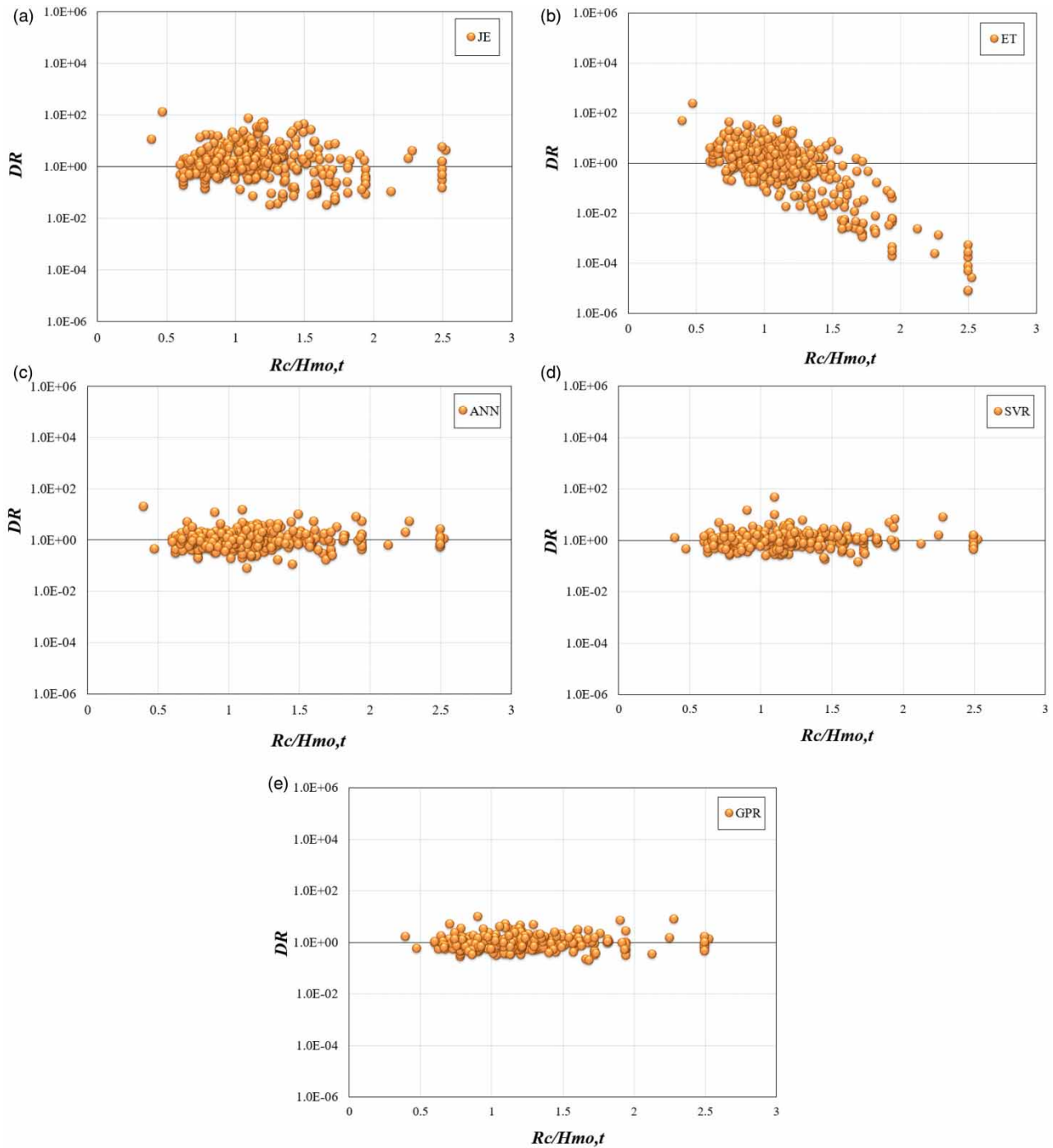


Figure 7 | Variation of DR as a function of the relative crest freeboard for the input combination b using (a) JE, (b) ET, (c) ANN, (d) SVR, and (e) GPR models; test data.

- The kernel-based models, especially GPR, perform better than the ANN and empirical formulae (JE and ET).
- The input combination b, with acceptable accuracy and as few as possible parameters obtained from sensitivity analysis, was introduced as the optimal one for modelling.
- In addition to commonly known effective parameters such as the relative crest freeboard ($R_c/H_{m0,t}$), relative crest width ($G_c/L_{m-1,0,t}$) and relative water depth at the toe of structure ($h/L_{m-1,0,t}$) were recommended to be considered in the prediction of overtopping.

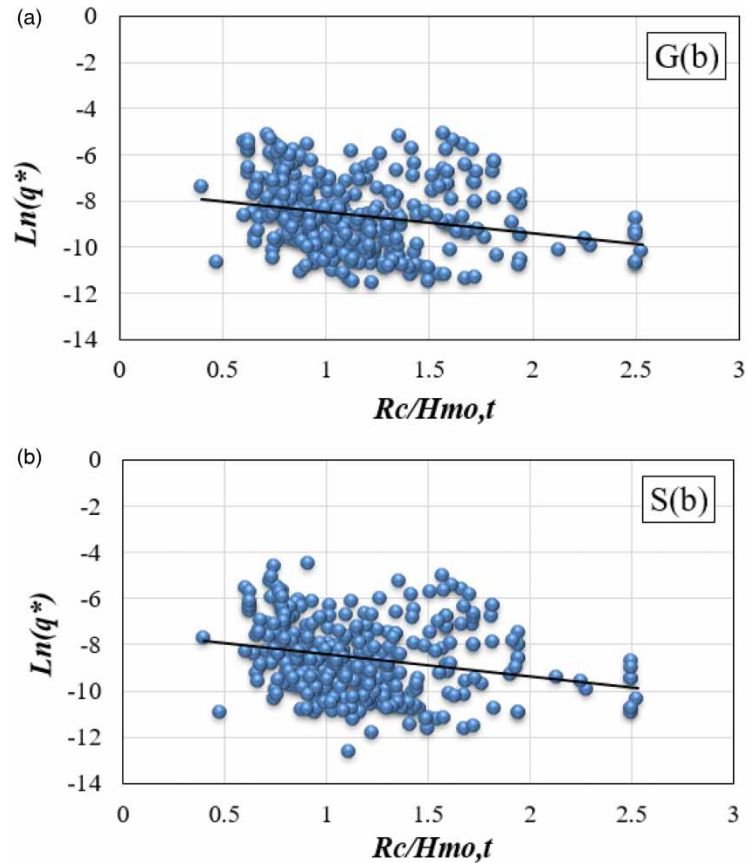


Figure 8 | Variation of $\ln(q^*)$ versus $R_c/H_{m0,t}$ using (a) GPR and (b) SVR models for input combination b.

- The GPR and SVR models can be used as reliable models, as the physics of overtopping phenomenon is preserved in modelling.
- The implementations of both GPR and SVR models are simple, as the structural parameters are optimized automatically. Hence, they are recommended for other similar studies.
- Among all models, considering both the simplicity of application and accuracy criteria the GPR model can be applied as an alternative tool for the prediction of wave overtopping rate at simple sloped breakwaters.

This study was conducted for simple sloped breakwaters. Investigating the capability of the kernel-based models for a larger database covering a variety of coastal structures can be the aim of future studies.

ACKNOWLEDGEMENTS

The authors appreciate A/Prof. Barbara Zanuttigh, Dr Sara Mizar Formentin, Prof. Jentsje W. Van der Meer, and Dr Tonino Liserra for kindly providing the [EurOtop \(2018\)](#) database.

DATA AVAILABILITY STATEMENT

All relevant data are available from an online repository or repositories (<http://www.overtopping-manual.com/eurotop/neural-networks-and-databases/>).

REFERENCES

- Altomare, C., Laucelli, D. B., Mase, H. & Gironella, X. 2020 Determination of semi-empirical models for mean wave overtopping using an evolutionary polynomial paradigm. *Journal of Marine Science and Engineering* 8 (8), 6–8. <https://doi.org/10.3390/JMSE8080570>.

- Awad, M. & Khanna, R. 2015 Support vector regression. In: *Efficient Learning Machines* (J. Pepper, S. Weiss, P. Hauke, R. Hutchinson & D. Pundick, eds). Apress, Berkeley, CA, pp. 67–80.
- Aye, S. A. & Heyns, P. S. 2017 An integrated Gaussian process regression for prediction of remaining useful life of slow speed bearings based on acoustic emission. *Mechanical Systems and Signal Processing* **84**, 485–498.
- Ayoubloo, M. K., Etemad-Shahidi, A. & Mahjoob, J. 2010 Evaluation of regular wave scour around a circular pile using data mining approaches. *Applied Ocean Research* **32**, 34–39.
- Cheon, S. H. & Suh, K. D. 2016 Effect of sea level rise on nearshore significant waves and coastal structures. *Ocean Engineering* **114**, 280–289. <https://doi.org/10.1016/j.oceaneng.2016.01.026>.
- Cini, S. & Deo, M. C. 2013 Real time current prediction with recurrent neural networks and model tree. *International Journal of Ocean System Engineering* **3** (3), 116–130. <https://doi.org/10.5574/ijose.2013.3.3.116>.
- De Rouck, J. & Geeraerts, J. 2005 *CLASH-Crest Level Assessment of Coastal Structures by Full Scale Monitoring, Neural Network Prediction and Hazard Analysis on Permissible Wave Overtopping*. Final Report: Full Scientific and Technical Report.
- Elvidge, S., Angling, M. J. & Nava, B. 2014 On the use of modified Taylor diagrams to compare ionospheric assimilation models. In *31th URSI General Assembly and Scientific Symposium, URSI GASS 2014, Table 2*, pp. 737–745. <https://doi.org/10.1109/URSIGASS.2014.6929835>.
- Etemad-Shahidi, A. & Jafari, E. 2014 New formulae for prediction of wave overtopping at inclined structures with smooth impermeable surface. *Ocean Engineering* **84**, 124–132.
- Etemad-Shahidi, A., Shaeri, S. & Jafari, E. 2016 Prediction of wave overtopping at vertical structures. *Coastal Engineering* **109**, 42–52.
- EurOtop 2007 *Wave Overtopping of Sea Defences and Related Structures–Assessment Manual* (N. W. H. Allsop, T. Pullen, T. Bruce, J. W. Van der Meer, H. Schüttrumpf & A. Kortenhaus). Available from: www.overtopping-manual.com.
- EurOtop 2018 *Manual on Wave Overtopping of Sea Defences and Related Structures. An Overtopping Manual Largely Based on European Research, But for Worldwide Application* (J. W. Van der Meer, N. W. H. Allsop, T. Bruce, J. De Rouck, A. Kortenhaus, T. Pullen, H. Schüttrumpf, P. Troch & B. Zanuttigh). Available from: www.overtopping-manual.com.
- Formentin, S. M., Zanuttigh, B. & van der Meer, J. W. 2017 A neural network tool for predicting wave reflection, overtopping and transmission. *Coastal Engineering Journal* **59** (1), 1750006.
- Ghaemi, N., Etemad-Shahidi, A. & Ataie-Ashtiani, B. 2013 Estimation of current-induced pile groups scour using a rule-based methods. *Journal of Hydroinformatics* **15** (2), 516–528.
- Ghazanfari-Hashemi, S., Etemad-Shahidi, A., Kazeminezhad, M. H. & Mansoori, A. R. 2011 Prediction of pile group scour in waves using support vector machines and ANN. *Journal of Hydroinformatics* **13** (4), 609–620. <https://doi.org/10.2166/hydro.2010.107>.
- Grbic, R., Kurtagić, D. & Sliškočić, D. 2013 Stream water temperature prediction based on Gaussian process regression. *Expert Systems with Applications* **40** (18), 7407–7414. <https://doi.org/10.1016/j.eswa.2013.06.077>.
- Gunn, S. R. 1998 Support vector machines for classification and regression. *ISIS Technical Report* **14** (1), 5–16.
- Haykin, S. S. 2009 *Neural Networks and Learning Machines*. Prentice Hall, New York.
- Hu, J. & Wang, J. 2015 Short-term wind speed prediction using empirical wavelet transform and Gaussian process regression. *Energy* **93**, 1456–1466.
- Ingram, D. M., Gao, F., Causon, D. M., Mingham, C. G. & Troch, P. 2009 Numerical investigations of wave overtopping at coastal structures. *Coastal Engineering* **56** (2), 190–202.
- Jafari, E. & Etemad-Shahidi, A. 2011 Derivation of a new model for prediction of wave overtopping at rubble mound structures. *Journal of Waterway, Port, Coastal and Ocean Engineering* **138** (1), 42–52. [https://doi.org/10.1061/\(ASCE\)WW.1943-5460.0000099](https://doi.org/10.1061/(ASCE)WW.1943-5460.0000099).
- Kazeminezhad, M. H., Etemad-Shahidi, A. & Bakhtiari, Y. 2010 An alternative approach for investigating of the wave-induced scour around pipelines. *Journal of Hydroinformatics* **12** (1), 51–65.
- Koosheh, A., Etemad-Shahidi, A., Cartwright, N., Tomlinson, R. & Hosseinzadeh, S. 2020 The comparison of empirical formulae for the prediction of mean wave overtopping rate at armoured sloped structures. *Coastal Engineering Proceedings* **36**, 22–22.
- Koosheh, A., Etemad-Shahidi, A., Cartwright, N., Tomlinson, R. & van Gent, M. R. 2021 Individual wave overtopping at coastal structures: a critical review and the existing challenges. *Applied Ocean Research* **106**, 102476.
- Lee, S. B. & Suh, K. D. 2019 Development of wave overtopping formulas for inclined seawalls using GMDH algorithm. *KSCE Journal of Civil Engineering* **23**, 1899–1910. <https://doi.org/10.1007/s12205-019-1298-1>.
- Losada, I. J., Lara, J. L., Guaniche, R. & Gonzalez-Ondina, J. M. 2008 Numerical analysis of wave overtopping of rubble mound breakwaters. *Coastal Engineering* **55** (1), 47–62.
- Melo, J. 2012 *Gaussian Processes for Regression: A Tutorial*. Technical Report. University of Porto.
- Moghaddas, F., Kabiri-Samani, A., Zekri, M. & Azamathulla, H. M. 2021 Combined APSO-ANN and APSO-ANFIS models for prediction of pressure loss in air-water two-phase slug flow in a horizontal pipeline. *Journal of Hydroinformatics* **23** (1), 88–102. <https://doi.org/10.2166/hydro.2020.300>.
- Molines, J. & Medina, J. R. 2016 Explicit wave-overtopping formula for mound breakwaters with crown walls using CLASH neural network-derived data. *Journal of Waterway, Port, Coastal and Ocean Engineering* **142** (3), 1–13. [https://doi.org/10.1061/\(ASCE\)WW.1943-5460.0000322](https://doi.org/10.1061/(ASCE)WW.1943-5460.0000322).
- Najafzadeh, M. & Oliveto, G. 2020 Riprap incipient motion for overtopping flows with machine learning models. *Journal of Hydroinformatics* **22** (4), 749–767. <https://doi.org/10.2166/hydro.2020.129>.

- Neves, M. G., Reis, M. T., Losada, I. J. & Hu, K. 2008 Wave overtopping of Póvoa de Varzim breakwater: physical and numerical simulations. *Journal of Waterway, Port, Coastal, and Ocean Engineering* **134** (4), 226–236.
- Owen, M. W. 1980 *Design of Sea Walls Allowing for Wave Overtopping*. Report EX. 924. Hydraulics Research Station, Wallingford, UK.
- Pickover, C. A. 2009 *The Math Book: From Pythagoras to the 57th Dimension, 250 Milestones in the History of Mathematics*. Sterling Publishing Company, New York.
- Pillai, K., Etemad-Shahidi, A. & Lemckert, C. 2017 Wave overtopping at berm breakwaters: experimental study and development of prediction formula. *Coastal Engineering* **130** (October), 85–102. <https://doi.org/10.1016/j.coastaleng.2017.10.004>.
- Pourzangbar, A., Saber, A., Yeganeh-Bakhtiary, A. & Ahari, L. R. 2017 Predicting scour depth at seawalls using GP and ANNs. *Journal of Hydroinformatics* **19** (3), 349–363. <https://doi.org/10.2166/hydro.2017.125>.
- Rasmussen, C. E. & Williams, C. K. I. 2006 *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA.
- Roushangar, K. & Koosheh, A. 2015 Evaluation of GA-SVR method for modeling bed load transport in gravel-bed rivers. *Journal of Hydrology* **527**, 1142–1152. <https://doi.org/10.1016/j.jhydrol.2015.06.006>.
- Roushangar, K., Hosseinzadeh, S. & Shiri, J. 2016 Local vs. cross station simulation of suspended sediment load in successive hydrometric stations: heuristic modeling approach. *Journal of Mountain Science* **13** (2), 1773–1788.
- Sahay, R. R. & Dutta, S. 2009 Prediction of longitudinal dispersion coefficients in natural rivers using genetic algorithm. *Hydrology Research* **40** (6), 544–552.
- Schulz, E., Speekenbrink, M. & Krause, A. 2018 A tutorial on Gaussian process regression: modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology* **85**, 1–16.
- Shaeri, S. & Etemad-Shahidi, A. 2021 Wave overtopping at vertical and battered smooth impermeable structures. *Coastal Engineering* **166**, 103889. <https://doi.org/10.1016/j.coastaleng.2021.103889>.
- Sun, A. Y., Wang, D. & Xu, X. 2014 Monthly streamflow forecasting using Gaussian process regression. *Journal of Hydrology* **511**, 72–81. <https://doi.org/10.1016/j.jhydrol.2014.01.023>.
- Taylor, K. E. 2001 Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres* **106** (D7), 7183–7192.
- Van der Meer, J. W. & Janssen, J. P. F. 1995 *Wave Run-up and Wave Overtopping at Dikes. Wave Forces on Inclined and Vertical Structures*. ASCE – TaskCommittee Reports, pp. 1–27.
- Van der Meer, J., Pullen, T., Allsop, W., Bruce, T., Schüttrumpf, H. & Kortenhaus, A. 2017 Prediction of overtopping. In: *Handbook of Coastal and Ocean Engineering: Expanded Edition*, Vol. 1–2, pp. 563–604. https://doi.org/10.1142/9789813204027_0021.
- Vapnik, V. N. 1995 *The Nature of Statistical Learning Theory*. Data Mining and Knowledge Discovery, Springer, New York, pp. 1–47.
- Verhaeghe, H., De Rouck, J. & Van Der Meer, J. 2008 Combined classifier–quantifier model: a 2-phases neural model for prediction of wave overtopping at coastal structures. *Coastal Engineering* **55**, 357–374.
- Zanganeh, M., Yeganeh-Bakhtiary, A. & Yamashita, T. 2016 ANFIS and ANN models for the estimation of wind and wave-induced current velocities at Joetsu-Ogata coast. *Journal of Hydroinformatics* **18** (2), 371–391. <https://doi.org/10.2166/hydro.2015.099>.
- Zanuttigh, B., Formentin, S. M. & van der Meer, J. W. 2014 Advances in modelling wave-structure interaction through artificial neural networks. *Coastal Engineering Proceedings* **1** (34), 69.
- Zanuttigh, B., Formentin, S. M. & van der Meer, J. W. 2016 Prediction of extreme and tolerable wave overtopping discharges through an advanced neural network. *Ocean Engineering* **127**, 7–22. <https://doi.org/10.1016/j.oceaneng.2016.09.032>.
- Zhang, N., Zhang, Q., Wang, K. H., Zou, G., Jiang, X., Yang, A. & Li, Y. 2020 Numerical simulation of wave overtopping on breakwater with an armor layer of accropode using SWASH model. *Water* **12** (2). <https://doi.org/10.3390/w12020386>.

First received 31 March 2021; accepted in revised form 2 August 2021. Available online 16 August 2021