Derivation of a new model for prediction of wave overtopping at rubble mound structures

Ebrahim Jafari¹ and Amir Etemad-Shahidi²*

¹M.Sc, School of Civil Engineering, Iran University of Science and Technology, Narmak, Tehran, Iran, P.O. Box 16765-163, Fax: +9821 77240398, E-mail: ejafari@civileng.iust.ac.ir.

²Professor, School of Civil Engineering, Iran University of Science and Technology, Narmak, Tehran, Iran, P.O. Box 16765-163, Fax: +9821 77240398, E-mail: etemad@iust.ac.ir. (Corresponding Author)

ABSTRACT

Prediction of wave overtopping is a key task in the design and safety assessment of coastal structures. In this study, M5´ model tree as a new soft computing approach was used to develop a model for prediction of wave overtopping rate at rubble mound breakwaters. The main advantages of model trees are that they are easier to deploy and more importantly they produce understandable formulas. The selected data from the CLASH database were used for training of the model and the conventional governing parameters were used as the input parameters. The obtained results were also compared with those of previous models. The accuracy of the model was evaluated by statistical measures, and it was shown that the developed model is more accurate than previous models. Furthermore, the model was validated with the prototype
overtopping measurements in three sites. Results indicated that the developed model outperforms
the previous models in predicting the full scale overtopping rates as well.

**Keywords:** wave overtopping rate, M5’ model tree, rubble mound breakwaters, CLASH
database

**INTRODUCTION**

Coastal structures are designed and constructed to protect coastal regions against storm waves
and high water levels during storm surges. Reliable prediction of wave overtopping plays a
significant role in the design and safety assessment of these structures. Excessive wave
overtopping can harm people, damage vehicles and properties on or close to the breakwaters.
Commonly, the mean wave overtopping rate should be below an acceptable rate under the design
conditions. Construction of a higher structure produces more protection than a lower one, but it
would be costly and greatly impair the harbor aesthetics (Shankar and Jayaratne 2003).

During the recent decades, several approaches have been developed for prediction of overtopping
rate. These approaches can be classified as empirical models, numerical models and soft
computing methods. In fact, the first empirical models were simple regression models
(Verhaeghe 2005). Typically, these models have been developed by fitting dimensionless
parameters to data obtained from physical model tests. The most important empirical formulas
proposed for rubble mound structures are the formulas of Owen (1980; 1982), Van der Meer and
Janssen (1995) and EurOtop (2007). Applications of these formulas are limited to specific
structures and wave conditions that the models were calibrated for them (Neves et al. 2008).

Due to the increase of computing power, the numerical models have become more popular in the
recent years (Neves et al. 2008). Numerical models simulate overtopping flow in numerical
flumes. Several numerical models have been developed such as those of Losada et al. (2008), Neves et al. (2008), and Ingram et al. (2009). The main disadvantage of the numerical models is the large amount of the required time and computing cost for an accurate simulation.

Recently, artificial neural network (ANN) models have been developed to predict mean wave overtopping discharge for a wide range of coastal structures. These studies have been carried out within the European project CLASH (De Rouck et al. 2009). Van Gent et al. (2007) proposed an ANN model for different types of coastal structures. Verhaeghe et al. (2008) developed a 2-phases neural prediction model: a classifier to distinguish negligible from significant overtopping rates and a quantifier for quantification of the overtopping discharges. Nevertheless, ANN is not as transparent and understandable as empirical models and their applications need optimization of the topology of the network (Etemad-Shahidi and Bonakdar 2009).

The aim of this study was that to derive a new accurate and transparent model for estimation of wave overtopping rate at rubble mound breakwaters. For this propose, M5’ model tree (Wang and Witten 1997) as a new approach is used to develop a model. The main advantage of model trees is that, unlike to neural networks, they represent comprehensive and transparent formulas. In the last decade, model trees have been used widely in engineering problems (e.g. Bhattacharya et al. 2007; Fan et al. 2008; Sakhare and Deo 2009; Etemad-Shahidi and Bonakdar 2009; Etemad-Shahidi and Mahjoobi 2009). However, to the authors’ knowledge, this method has not been used in the prediction of overtopping rate yet. The model is trained and tested using small scale data selected from CLASH Database (Van der Meer et al. 2009). The model tree provides simple regression formulas with low computational cost, and the results were compared with those of the previous empirical formulas such as Owen (1980), Van der Meer and Janssen (1995) and EurOtop (2007). In addition, the new model has also been compared with the ANN models proposed by Van Gent et al. (2007) and Verhaeghe et al. (2008).
SELECTION OF DATA FROM THE CLASH DATABASE

The used data set is the database created by the EU-financed CLASH project (De Rouck et al. 2009). Within the CLASH project, an extensive database on wave overtopping consisting of the prototype overtopping measurements and many laboratory tests was created. The database includes many tests, resulting from different experiments performed at several institutions and laboratories with different scales and using various measurement tools with different precisions. For example, analysis of the measurement uncertainties in the test performed in physical modeling of Zeebrugge breakwater has shown that when a test had been repeated, the values of wave height and periods differed in the range of up to 5%. In addition, when weighting of the overtopping water was used for determination of overtopping discharge, overtopping rates varied up to 12% when the test was repeated. More information about the database and experiments has been presented in Verhaeghe (2005) and van der Meer et al. (2009).

The data of wave overtopping at rubble mound structures were selected in this paper. These data contain small scale tests and large scale measurements. Small scale tests include 28 datasets related to more than 1400 tests. A summary of these datasets with their identification in the CLASH database is presented in appendix A. The available large scale measurements for rubble mound breakwaters in the CLASH database are 105 data points, originating from three datasets. The first dataset was the prototype measurements at Ostia and 77 overtopping measurements were recorded at a rubble mound breakwater armored with rocks (Franco et al. 2009). The prototype measurements in Zeebrugge (Belgium) were the second dataset. Overtopping rates were measured at a rubble mound breakwater armored with 25 ton grooved cubes (Geeraerts et al. 2009). The last dataset was the measurements performed in Japan (Goda et al. 1975). 17 prototype measurements were conducted there at a rock armored structure.
Overtopping studies within the CLASH project have shown that the prototype condition cannot be modeled perfectly in the laboratories due to three effects: model effects, scale effects and effects of measurement techniques. For example, effects of sea currents and/or the wind blowing in the field cannot be modeled properly in the laboratory. It was also shown that if large scale tests are used for training of the models (together with the small scale tests), the developed model would possibly be affected by these effects (Verhaeghe 2005). As the majority of overtopping tests in the CLASH database were concerned to small scale tests, large scale tests (including prototype measurements) were excluded during the training process to avoid confusion of the model. Hence, the model was developed based on the small scale tests and then validated using field measurements.

The CLASH database has assigned a complexity factor $CF$ to each structure and a reliability factor $RF$ to each test. The values of $CF$ ranged from $CF=1$ for a very simple structure to $CF=4$ for a very complex structure. In a similar way, $RF=1$ refers to a very reliable test, while $RF=4$ refers to a non-reliable test (Van Gent et al. 2007). In this study, the data with $CF=4$ and $RF=4$ have been left out. In addition, the data of the structures without berms and with single slopes were selected for further processing. Due to the measurement errors, very low overtopping discharges in small scale tests may not be accurate. Hence, the tests with the values of $q < 1 \times 10^{-6}$ m$^3$/s/m were not used for training of the model.

**RESULTS OF PREVIOUS STUDIES FOR SELECTED DATASETS**

A commonly used overtopping model obtained from laboratory tests with irregular waves is the simple regression model proposed by Owen (1980; 1982). Owen presented his formula to predict the dimensionless overtopping rate at straight and bermed impermeable sloping structures (USACE 2006):
\[
q = a \exp \left( -b \frac{R_C}{H_S} \sqrt{\frac{S_{om}}{2 \pi}} \gamma_r \right)
\] (1)

Where \( q \) is the mean overtopping discharge, \( H_S \) is the significant wave height, \( T_m \) is the mean period of the incident waves and \( R_C \) is the crest freeboard. The parameters \( a \) and \( b \) are fitted coefficients (USACE 2006); and \( S_{om} \) is the deepwater wave steepness defined as:

\[
S_{om} = \frac{H_s}{L_{om}}
\] (2)

Where \( L_{om} \) is the deepwater wavelength defined as:

\[
L_{om} = \frac{g \cdot T_m^2}{2\pi}
\] (3)

Scatter diagram of the measured and predicted overtopping rates by the formula of Owen (1982) for the small scale datasets is shown in Fig. 1. The parameter \( q^* \) in this figure is the dimensionless overtopping discharge defined as:

\[
q^* = \frac{q}{\sqrt{g \cdot H_S^3}}
\] (4)

The inclined upper dashed line, solid line and lower dashed line are 10 times over estimated, perfect agreement, and 10 times under estimated lines, respectively. It is clear from Fig. 1 that this formula is inaccurate and somewhat overestimated.

Van der Meer and Janssen (1995) presented the following formulas for estimation of overtopping rate at impermeable sloping structures:

\[
\text{If } \xi_{op} < 2 \text{ then } \frac{q}{\sqrt{g \cdot H_S^3}} \sqrt{\frac{S_{op}}{\tan \alpha}} = 0.06 \exp \left( -5.2 \frac{R_C}{H_S} \sqrt{S_{op}} \frac{1}{\tan \alpha \gamma_r \gamma_b \gamma_h \gamma_\beta} \right)
\] (5)

\[
\text{If } \xi_{op} > 2 \text{ then } \frac{q}{\sqrt{g \cdot H_S^3}} = 0.2 \exp \left( -2.6 \frac{R_C}{H_S} \frac{1}{\gamma_r \gamma_b \gamma_h \gamma_\beta} \right)
\] (6)
Where $\alpha$ is the slope angle and $\xi_{op}$ is the surf similarity parameter (Iribarren number) defined as (USACE 2006):

$$\xi_{op} = \frac{\tan \alpha}{\sqrt{S_{op}}}$$

(7)

The reduction factors include the surface roughness $\gamma_r$ ($\gamma_r = 1$ for smooth slopes), berm effect $\gamma_b$ ($\gamma_b = 1$ for non-bermed profiles), influence of shallow-water conditions $\gamma_h$ ($\gamma_h = 1$ for Rayleigh distributed waves), and effect of angle of incidence $\beta$ of the waves $\gamma_\beta$ ($\gamma_\beta = 1$ for head-on waves).

Fig. 2 shows the comparison between the measured overtopping rates and the predicted ones by the formulas of Van der Meer and Janssen (1995). This figure shows that there is a clear disagreement between the predicted values and measured ones, especially for low discharges. Fig. 2 also demonstrates the high uncertainty of Van der Meer and Janssen's (1995) formula.

Recently, EurOtop (2007) proposed the following formula for the prediction of wave overtopping rate at simple slope breakwaters:

$$\frac{q}{\sqrt{g \cdot H_{m0}^2}} = 0.2 \cdot \exp \left( -2.6 \frac{R_c}{H_{m0} \cdot \gamma_f \cdot \gamma_\beta} \right)$$

(8)

The reduction factor for the effect of oblique waves is as follows:

$$\begin{cases} 
0^\circ \leq |\beta| \leq 80^\circ \rightarrow \gamma_\beta = 1 - 0.0063|\beta| \\
|\beta| > 80^\circ \text{ therelseult } \beta = 80^\circ \text{ canbeapplied }
\end{cases}$$

(9)

EurOtop's formula is simpler than those of Owen (1982) and Van der Meer and Janssen (1995) and does not consider the wave steepness. The scatter diagram of the measured and predicted overtopping discharges by EurOtop (2007) is given in Fig. 3. As can be seen from the figure, there is a large scatter between the measured and predicted overtopping rates.

Van Gent et al. (2007) presented an ANN model with 15 input parameters using the datasets in the CLASH database. Since they used dimensional input parameters from different tests with
different scales for developing their models, they scaled all input parameters to $H_s = 1m$ using Froude's similarity law. Fig. 4 shows the scatter diagram of measured and predicted overtopping rate by ANN model. As seen, the result of ANN model is more accurate than those of other ones. However, the predictions of ANN model are somewhat underestimated.

**MODEL TREE**

Model trees (MTs) are based upon the idea that a complex problem can be solved by dividing into smaller sub problems and solving each sub problem (Bhattacharya et al. 2007). MTs divide the input parameter domain to smaller sub-domains and in each sub-domain, they assign a multiple linear regression model. One of the most commonly used approaches of MTs is M5 algorithm which was first proposed by Quinlan (1992). After that his idea was improved and a more completed algorithm called M5´ was proposed by Wang and Witten (1997) to deal with missing values. As M5´ model tree divide input domain to some sub-domains, the final solution of the problem is formed by considering the combination of some piecewise linear models. Hence, the model shows a nonlinear behavior and can be used to solve nonlinear problems.

M5´ model tree has three main steps: building a tree, pruning the tree and smoothing the tree. First M5´ algorithm forms the basic tree using the standard deviation reduction SDR factor:

$$ SDR = sd(T) - \sum \frac{|T_i|}{|T|} \times sd(T_i) $$

(10)

Where $T$ is the set of data points before splitting, $T_i$ is the data points that result from splitting the domain and fall into one sub-domain according to the chosen splitting parameter and $sd$ is the standard deviation. Standard deviation is used as an error measure for the data points that fall into a sub-domain. M5´ model tree tests different splitting points in the domains for all the input parameters. For each point, M5´ calculates the standard deviation for sub-domains and then
compares it with the standard deviation of data points before dividing the domain to sub-domains. When a point in the input parameters maximizes the standard deviation reduction, it is chosen as the splitting point (node). After that for each sub-domain the splitting process is repeated. The splitting process is stops when the standard deviation reduction is less than 5 percent or a few data points remain in a sub-domain. After the tree has been grown, a linear multiple regression model is calculated for each sub-domain.

The accuracy of the model for training set increases uniformly by growing the tree. However, over-fitting may occur during construction of a model tree based on the training set. Hence, a pruning method is used to avoid this problem. The pruning approach uses the prediction of the expected error at each node for the test data. First, for each of the training examples that reach the node, the average absolute difference between the predicted value and the actual output value is calculated. To prevent underestimating the expected error for new data, the output value is multiplied by \((n+\nu)/(n-\nu)\), where \(n\) is the number of training data points that reach to the node and \(\nu\) is the number of input parameters that represent the output value at that node. The leaf (or sub-domain) can be pruned if the predicted error is lower than the expected one (Witten and Frank 2005).

After pruning, the pruned tree might have sharp discontinuities that possibly occur between adjacent linear models at the leaves; especially for some models built on the small number of training data points (Jung et al. 2010). The smoothing procedure uses the models built in each sub-domain to compute the predicted value. Then, that value is modified along the route back to the root of the tree (first splitting point) by smoothing it at each node. The predicted value by the leaf model is combined with the value predicted by the linear model for each node on the top of the leaf to the root (Quinlan 1992). Generally, smoothing process improves the predictions.
MODELING AND RESULT

In the CLASH database, there are zero and non-zero overtopping rates. Hence, first a model was developed to distinguish between them. For development of this model, different parameters which are more important were selected as the input parameters. These parameters were the dimensionless water depth at the toe of structure $h_t/H_s$, dimensionless crest width of the structure $G_C/H_s$, dimensionless crest freeboard of the structure $R_C/H_s$, tangent of the angle that the structure surface makes with a horizontal plane $\tan(\alpha)$ and deepwater wave steepness $S_{op}$. Both zero and non-zero data points were used for development of this separator model. The M5′ algorithm selected $G_C/H_s$ and $R_C/H_s$ as the input parameters.

Wave overtopping occurs in two forms. Most of the wave overtopping discharges are recorded when the wave run-up is higher than the crest level. The remaining part of wave overtopping is related to fine droplets (usually generated by wave breaking) that are carried over the crest of the structure by their momentum or wind (Verhaeghe 2005). The selection of $G_C/H_s$ and $R_C/H_s$ by M5′ model is physically sound because if the crest of the structure is high and wide, the rising water and fine droplets cannot pass over it. The separator model showed that if $R_C/H_s > 2.08$ and $G_C/H_s > 1.5$, zero overtopping rate occurs. This critical value of about 2 for $R_C/H_s$ is in a good agreement with the proposed value of 1.97 by Van Der Meer and Stam (1992) for maximum value of run-up on rubble mound structures. This is because if $R_C/H_s$ is greater than about 2, the wave run-up cannot reach to the crest of the breakwater and overtopping does not occur. The overtopping is usually measured in the leeside of the breakwater. Hence, the overtopped water has to cross the crest width, to be recorded. When $G_C/H_s$ is large, fewer overtopping water reaches to the leeside of the breakwater crest. The result of the model for $G_C/H_s$ is also supported by findings of Aminti and Franco (1988) showing that by increasing $G_C/H_s$, the overtopping discharge reduces.
After separating non-zero overtopping rates, an estimator model tree was developed for the prediction of wave overtopping discharge. For construction of the model, 67 percent of small scale data were randomly separated as training set. The estimator model was developed based on the training set. The remaining data points (testing set) were used for the validation. Several models were developed using various dimensionless parameters. Table 1 shows various parameters which were tried in developing the new model. As seen in the table, Different combinations of non-dimensional numbers such as the surf similarity parameter, slope angle, dimensionless crest freeboard, wave steepness and etc. have tried and the model which outperformed the other ones was selected. The used input parameters in the selected estimator model tree (EMT) were the dimensionless crest freeboard and tangent of the angle between the structure slope surface and the horizontal plane. The considered dimensionless crest freeboard is defined as:

\[
R^* = \frac{R_c}{H_S \cdot \gamma_f \cdot \gamma_f} \times \frac{\sqrt{S_{op}}}{\tan \alpha}
\]  

(11)

The reduction factors of \(\gamma_f\) and \(\gamma_f\) were assumed to be the same as the proposed values of EurOtop (2007). The output parameter was the same as eq. 4. Some of the previous empirical formula for the prediction of wave overtopping rate such as those of Owen (1982), Van der Meer and Janssen (1995) and EurOtop (2007) have an exponential form. Hence, the exponential form was used for EMT model. M5' model tree can only produce linear relationship between input and output parameters. To remove this restriction, the model was developed using \(\ln(q^*)\) as the output parameter that \(\ln\) is the logarithm in base of e. The data with \(q > 1 \times 10^{-6}\) were used for the development of EMT model. The data points were classified by the separator model and then were used in EMT model. More than 99 percent of them were classified correctly by the
separator model. Table 2 summarizes the ranges of the parameters used in developing of EMT.

The final formulas were:

If \( R^* \leq 0.86 \) then
\[
\frac{q}{\sqrt{g \cdot H_s}} = \exp\left( -6.18 R^* - 3.21 \right)
\] (12)

If \( R^* > 0.86 \) then
\[
\frac{q}{\sqrt{g \cdot H_s}} = \exp\left( -3.1 R^* - 6.05 \tan (\alpha) - 2.63 \right)
\] (13)

It can be seen from eq. 12 and eq. 13 that by increasing dimensionless crest freeboard, the influence of \( R^* \) on wave overtopping rate is reduced. Moreover, the effect of slope angle increases and the parameter of \( \tan (\alpha) \) directly appears in the formula. Fig. 5 shows the predicted value of overtopping discharges by eq. 13 versus relative crest freeboard for different values of slope angles. It is clear from the figure that by decreasing the breakwater slope angle, the amount of overtopping discharge decreases. In addition, the ratio of reducing the overtopping rate is increased by decreasing the slope angle. This can be justified by considering that when the slope angle decreases, the rising water has to travel farther on a rough surface to reach the crest of structure and pass over it. This distance increases by reducing the slope angle. Furthermore, as displayed in Van der Meer and Stam (1992) formula, by decreasing the slope angle, the height of wave run-up (\( R_u / H_s \)) decreases. On the other hand, when the slope angle decreases, the waves break further from the crest of the structure and less fine droplets (generated by wave breaking) can pass the crest.

As discussed before, model tree divides the input parameter domain into a number of subdomains and in each sub-domain it builds a local linear regression model. One of the limitations of the model tree is the discontinuity between two local models. However, M5’ model tree uses "smoothing" process to reduce these discontinuities partly. The value of dimensionless overtopping discharge \( q^* \) using previous equations for \( R^* = 0.86 \) and \( \tan (\alpha) = 0.522 \) (average for
all data) were 0.199 lit/s/m and 0.213 lit/s/m, respectively. Considering these values, the amount of discontinuity is about 7%. Fig. 6 shows dimensionless measured overtopping discharges versus dimensionless crest freeboard. The bold lines are the predictions by the new formulas using \( \tan(\alpha) = 0.522 \). As seen, the mentioned value of discontinuity is negligible considering the uncertainties in the measurements. In addition, using \( \xi_{\text{op}} = 2, R_C/ (H_S \cdot \gamma_f \cdot \gamma_p) = 2.4 \) (average value) and \( \tan(\alpha) = 0.522 \) (average value) in Van der Meer and Janssen (1995) formulas, the amount of discontinuity is 20% which is about three times higher than that of the new model. On the other hand, Fig. 6 shows that by increasing the dimensionless crest freeboard the effect of \( R^* \) decrease. Therefore, the coefficient of \( R^* \) decreases from 6.18 (in eq. 12) to 3.1 (in eq. 13).

As mentioned before, less than 1 percent of non-zero overtopping rates were misclassified by the separator model. These data points mostly had very low values of the \( q_{\text{measured}} \). In addition, for error calculation, they should have numeric value. Hence, the following formula was used for negligible overtopping rates:

\[
\frac{q}{\sqrt{g \cdot H_S}} = \exp\left(-0.6396R^* - 0.7085 \tan \alpha - 11.4897\right)
\]  

(14)

This formula is derived from fitting dimensionless group to wrongly classified data points by the separator model. The power in this equation is always less than \(-11.49\), which indicates that the predicted rates are very low. This formula was then used for data points which were classified as no overtopping by the separator model. The developed formulas were finally combined as follows:
\[
\begin{align*}
\text{if } \frac{R_c}{H} & > 2.08 \text{ and } \frac{G_c}{H} > 1.51 \text{ then} \\
\frac{q}{\sqrt{g \cdot H^3}} &= \exp(-0.6396R^\ast - 0.7085 \tan \alpha - 11.4897) \\
\text{if } R^\ast & \leq 0.86 \text{ then} \\
\frac{q}{\sqrt{g \cdot H^3}} &= \exp(-6.18R^\ast - 3.21) \\
\text{if } R^\ast & > 0.86 \text{ then} \\
\frac{q}{\sqrt{g \cdot H^3}} &= \exp(-3.1R^\ast - 6.05 \tan \alpha - 2.63)
\end{align*}
\] (15)

These formulas are as simple and transparent as the previous empirical. Scatter diagrams of the measured and predicted dimensionless overtopping discharges for test and all data points are shown in Fig. 7. As seen, the predictions of M5 model are more concentrated around the optimal line compared to the previous empirical formulas (Fig. 1, 2 and 3) and close to those of ANN model proposed by Van Gent et al. (2007). M5’ model was trained with 67 percent of the dataset shown on this plot. Therefore, a near perfect agreement between data and M5 model is obtained. In contrast, other models shown on figures 1, 2 and 3 were not trained with any of these data. We have included these models in this comparison to assess their universal applicability, which have performed remarkably well for this dataset.

For evaluating the accuracy of the models, statistical indicators such as geometric mean \(X_G\), geometric standard deviation \(\sigma_{X_G}\) (Goda 2009), \textit{Bias}, the coefficient of determination \(R^2\), root mean square error (\textit{RMSE}) and discrepancy ratio (\textit{DR}) were used. These parameters are defined as:

\[
X_G = \exp\left(\frac{1}{n} \sum_{i=1}^{n} \ln\left(\frac{q_{\text{est}}}{q_{\text{meas}}}\right)\right)
\] (16)

\[
\sigma_{X_G} = \exp\left(\sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\ln\left(\frac{q_{\text{est}}}{q_{\text{meas}}}\right) - \ln\left(x_G\right)\right)^2}\right)
\] (17)
\[ BIAS = \frac{1}{n} \sum_{i=1}^{n} \left( \log q_{\text{est}}^* - \log q_{\text{meas}}^* \right) \]  
\[ R^2 = \frac{\left( \sum_{i=1}^{n} (q_{\text{meas}}^* - \bar{X})(q_{\text{est}}^* - \bar{Y}) \right)^2}{\sum_{i=1}^{n} (q_{\text{meas}}^* - \bar{X})^2 \sum_{i=1}^{n} (q_{\text{est}}^* - \bar{Y})^2} \]  
\[ RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left( \log q_{\text{meas}}^* - \log q_{\text{est}}^* \right)^2} \]  
\[ DR = \frac{q_{\text{est}}^*}{q_{\text{meas}}^*} \]

Where \( q_{\text{meas}}^* \) and \( q_{\text{est}}^* \) are the dimensionless measured and estimated overtopping discharges, \( n \) is the number of the measurements, and \( \bar{X} \) and \( \bar{Y} \) are respectively the average of \( \log(q_{\text{meas}}^*) \) and \( \log(q_{\text{predicted}}^*) \) that \log\ is logarithm in base of 10. In addition, the slope of the regression line for prediction versus measured data was also used for evaluating the model. If a linear regression like “\( y = ax + b + \text{error} \)” is fitted to predicted versus measured scatter charts, the slope of the regression line is equal to \( a \). Table 3 shows the error measures of the previous models and the new model. As seen, almost all the error measures of the new model show improvement in the prediction of overtopping rates. The geometric mean ratio of the new model shows a slight overestimation, while Van der Meer and Janssen and Owen estimations are more overestimated than the model tree. \( X_G \) ratio also shows that both EurOtop formula and ANN model underestimate the overtopping rate. Geometric standard deviation indicates the scatter of predictions. As shown in the Table 3, the predictions of the new model are also less scattered than those of other empirical formulas, and it is close to that of ANN model. \( Bias \) shows the average of absolute error. The negative \( Bias \) values for the EurOtop and ANN models imply that they are underestimated on average. It can be seen from the Table 3 that the average of absolute
error for the new model is much less than those of other models. The coefficient of determination makes sense together with the slope of the regression line. If the slope of the regression line is close to one and the value of $R^2$ is high, then the model is accurate. As seen in the table 3, the new and ANN model performs better than previous empirical formulas and can be used for predictions. The slopes of the most of the models are low. This could be due to simple structure of regression equations. More complex regressions probably result in a better slope; however, the use and the justification of these equations will be more complicated. RMSE error is generally interpreted as another performance indicator. RMSE errors of the different models show that the new model outperforms other ones. The last two rows of the table 3 are related to the discrepancy ratio ($DR$). If the value of log ($DR$) is equal to zero, it means that there is an exact matching between measured and predicted values. The range of log ($DR$) for different models can be seen in Table 3. The last row of table 3 shows the percentage of log ($DR$) values, which are between -1 and 1. It is clear from the table 3 that about 90 percent of model tree predictions are within -1<log ($DR$) <1 range. This percentage for ANN model is about 86.34% and for other formulas this percentage is less than 75 percent.

The histogram of log ($DR$) for different models is shown in Fig. 8. It is clear from this figure that log ($DR$) values for most of the model tree predictions are close to zero and their distribution is nearly symmetrical. However, Van der Meer and Janssen (1995) and Owen (1982) formulas overestimate the overtopping rate. For example, it can be seen from the Fig. 8 that more than 38% of Owen predictions are more than ten times larger than the measured overtopping rates. The values of log ($DR$) for ANN model do not have a symmetrical distribution and most of them are less than zero. Also more than 20 percent of EurOtop predictions are more than 10 times smaller than the measured overtopping rates.
Fig. 9 shows the variation of $DR$ as a function of the relative crest freeboard for different models. A good model is a model that its errors are independent of the input parameters (Sahay and Dutta 2009). Van der Meer and Janssen (1995) and EurOtop (2007) formulas are very sensitive to change of relative crest freeboard. However, according to Fig. 9, model tree had a better performance than other ones in this aspect. In addition, this figure shows that $R_C/H_S$ is included in a better way in the new model. A similar process has also been observed (not shown) for other input parameters such as the surf similarity parameter and slope angle.

In brief, in the development of model trees, minimum effort was required compared to other commonly used soft computing approaches. For example, ANN needs some trial and error for finding optimum neuron and epoch numbers. The main limitation of model trees is that they only can provide linear relationships between input and output parameters (Etemad-Shahidi and Bonakdar 2009). Therefore, input parameters transformations may be required to develop the optimum model. Hence, an appropriate preprocessing of parameters in the M5 model tree can yield a high performance model, i.e. a simple and accurate one.

**APPLICATION EXAMPLES**

In this section M5´ model tree is applied to the prototype measurements in three sites. The measured values were compared with those predicted by the new model, Owen (1982), EurOtop (2007) and the 2-phase neural network model proposed by Verhaeghe et al. (2008). The results of the ANN model proposed by Van Gent et al. (2007) could not be used here; because in the most cases, overtopping rates were estimated negligible and the ANN model did not present any meaningful numeric value for the most of data points. The formulas of Van der Meer and Janssen (1995) were not used as well, because more than 55 percent of 105 data points were out of their range of applicability. Scatter diagrams of the measured and predicted overtopping discharges for
Ostia and Zeebrugge sites are shown in Fig. 10. As seen, the predictions of the new model for Ostia case are more accurate than those of Owen (1982) and EurOtop (2007). For prototype measurements, Verhaeghe et al. (2008) used an enhancement factor to correct the model and scale effects. Table 4 shows the error measures for different models. $X_G$ ratio shows that the models of M5', Owen and EurOtop are all underestimated. This can be justified by model and scale effects which were considered in the ANN model. $\sigma_{XG}$ shows that in Ostia case, the scatter of Owen and EurOtop predictions are more than that of the new model. However, the scatter of ANN model is less than that of the new model. The $RMSE$ rate shows that the new model is more accurate than previous empirical formulas and its accuracy is almost similar to that of the ANN model.

The next application example is the Zeebrugge case (marked in the Fig. 10 by bold black circles). In this case, the armor crest freeboard height $A_C$ is higher than the height of the top of the wave wall or the height of the crest freeboard $R_C$. Since EurOtop manual recommends using $A_C$ instead of $R_C$ in such cases, the parameter of $A_C$ was used to obtain the overtopping rates for Zeebrugge. As seen from the Fig. 10 and Table 4; Owen's formula, the new model and ANN model are all overestimating. However, the new model is still more accurate than others. One reason for these overestimations is that when the armor crest freeboard $A_C$ was considered in the formulas, it was supposed that the impermeable crest height $R_C$ is equal to $A_C$. However, the armor crest height $A_C$ is higher than $R_C$. It means that permeability of the structural crest is more than that of considered, which causes overestimation in overtopping rates. Other error measures also show that the new model outperforms other ones in the Zeebrugge case. In the ANN model, the correction factor for scale and model effects caused the predictions to become much overestimated. It means that this factor still needs to be improved.
The last application example is a breakwater in Japan. This breakwater has an impermeable vertical wave wall on its crest. It means that the \( R_C \) is higher than \( A_C \) in this structure. The scatter diagrams of measured overtopping rates and predicted by different models are shown in Fig. 11. It should be mentioned that the result of ANN model for this case was not available. It is clear from the figure that the previous formulas predicted very low values for overtopping rates. The average of \( R_C/H_s \) and \( G_C/H_s \) for these measurements was 4.12 and 3.91, respectively. Due to the high values of relative crest freeboard \( (R_C/H_s) \) and relative crest width \( (G_C/H_s) \) of the structure, both previous formulas are too underestimated. All of these data points are in the range of negligible overtopping rates of the new model, and the overtopping discharges are predicted by using eq. 14. This example confirms the necessity of separation between negligible and significant overtopping rate before quantifying the overtopping rate.

**SUMMARY AND CONCLUSIONS**

In this study, M5’ model tree was used to predict the overtopping rate at rubble mound structures. The model was developed using small scale data extracted from the CLASH database. First, the data points were separated to zero and non-zero overtopping rates. Dimensionless crest width and dimensionless crest freeboard of the structure were used for this separator model. Then an estimator model was developed for prediction of overtopping rates. Finally, the developed models were combined and a single model was suggested (eq. 15). The proposed model was validated by using the small scale datasets. In addition, performances of the new model and the previous ones were evaluated for field measurements as an application example. It was shown that the proposed simple model can be used successfully for the prediction of wave overtopping rate.
ACKNOWLEDGEMENTS

The authors would like to acknowledge the organizers of EC-project CLASH for their comprehensive and freely available database on the wave overtopping phenomenon. The authors are also grateful to Prof. Marcel Van Gent and Mr. Jan Kramer for providing access to their ANN model. In addition, the authors thank Reza Yasa and Lisham Bonakdar for their help in improving the manuscript. This work was partly supported by the Deputy of Research, Iran University of Science and Technology.
APPENDIX A

Table A
REFERENCES


**Table 1.** Various parameters that used for development of the estimator model.

<table>
<thead>
<tr>
<th>Input parameters</th>
<th>Output parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\zeta_{op}$</td>
<td>$\ln\left(\frac{q}{\sqrt{g \cdot H_S^3}}\right)$</td>
</tr>
<tr>
<td>$\tan\alpha$</td>
<td>$\ln\left(\frac{q}{g \cdot H_S \cdot T_{om}}\right)$</td>
</tr>
<tr>
<td>$\frac{R_C}{H_S \cdot \gamma_f}$</td>
<td>$\frac{R_C}{H_S \cdot \gamma_f} \times \frac{S_{om}}{2\pi}$</td>
</tr>
<tr>
<td>$\frac{R_C}{H_S \cdot \gamma_f} \times \sqrt{S_{op}}$</td>
<td>$\frac{R_C}{H_S \cdot \gamma_f} \times \frac{\sqrt{S_{op}}}{\tan\alpha}$</td>
</tr>
<tr>
<td>Parameter</td>
<td>Training set</td>
</tr>
<tr>
<td>------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>$R_C \ [m]$</td>
<td>0.047 ~ 0.350</td>
</tr>
<tr>
<td>$H_S \ [m]$</td>
<td>0.0385 ~ 0.3103</td>
</tr>
<tr>
<td>$\beta \ [^\circ]$</td>
<td>0 ~ 60</td>
</tr>
<tr>
<td>$\gamma_f \ [-]$</td>
<td>0.38 ~ 0.6</td>
</tr>
<tr>
<td>$T_p \ [S]$</td>
<td>0.8445 ~ 3.6145</td>
</tr>
<tr>
<td>$\tan \ (\alpha) \ [-]$</td>
<td>0.20 ~ 0.75</td>
</tr>
<tr>
<td>$q \ [m^3/s/m]$</td>
<td>1.000E-6 ~ 2.760E-3</td>
</tr>
</tbody>
</table>
Table 3. Statistical measures of different models for small scale experiments.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>all data</td>
<td>all data</td>
<td>all data</td>
<td>all data</td>
<td>all data</td>
</tr>
<tr>
<td>$X_G$</td>
<td>6.014</td>
<td>1.999</td>
<td>0.460</td>
<td>0.389</td>
<td>1.042</td>
</tr>
<tr>
<td>$\sigma_{X_G}$</td>
<td>4.787</td>
<td>10.396</td>
<td>10.613</td>
<td>3.325</td>
<td>3.552</td>
</tr>
<tr>
<td>BIAS</td>
<td>0.779</td>
<td>0.301</td>
<td>-0.337</td>
<td>-0.411</td>
<td>0.018</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.200</td>
<td>0.100</td>
<td>0.235</td>
<td>0.693</td>
<td>0.476</td>
</tr>
<tr>
<td>Slope of regression line</td>
<td>1.738</td>
<td>0.507</td>
<td>0.357</td>
<td>0.541</td>
<td>0.381</td>
</tr>
<tr>
<td>RMSE</td>
<td>1.034</td>
<td>1.060</td>
<td>1.080</td>
<td>0.664</td>
<td>0.551</td>
</tr>
<tr>
<td>Range of Log (DR)</td>
<td>-2.31 to 3.17</td>
<td>-2.78 to 2.53</td>
<td>-3.95 to 1.85</td>
<td>-2.43 to 1.65</td>
<td>-1.55 to</td>
</tr>
<tr>
<td>-1&lt;Log (DR)&lt;1 (%)</td>
<td>60.79</td>
<td>60.81</td>
<td>73.48</td>
<td>86.34</td>
<td>91.95</td>
</tr>
</tbody>
</table>
Table 4. Statistical errors of different models for field measurements.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Ostia</td>
<td>$X_G$</td>
<td>0.049</td>
<td>0.053</td>
<td>2.073</td>
<td>0.853</td>
</tr>
<tr>
<td></td>
<td>$\sigma_{XG}$</td>
<td>6.437</td>
<td>4.254</td>
<td>2.555</td>
<td>3.432</td>
</tr>
<tr>
<td></td>
<td>BIAS</td>
<td>-1.306</td>
<td>-1.278</td>
<td>0.317</td>
<td>-0.069</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>1.536</td>
<td>1.424</td>
<td>0.525</td>
<td>0.540</td>
</tr>
<tr>
<td>Zeebrugge</td>
<td>$X_G$</td>
<td>3.123</td>
<td>0.081</td>
<td>5.168</td>
<td>1.754</td>
</tr>
<tr>
<td></td>
<td>$\sigma_{XG}$</td>
<td>1.661</td>
<td>2.413</td>
<td>2.931</td>
<td>1.989</td>
</tr>
<tr>
<td></td>
<td>BIAS</td>
<td>0.495</td>
<td>-1.089</td>
<td>0.713</td>
<td>0.244</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.541</td>
<td>1.154</td>
<td>0.844</td>
<td>0.386</td>
</tr>
<tr>
<td>Japan</td>
<td>$X_G$</td>
<td>4.00E-07</td>
<td>3.30E-08</td>
<td>-</td>
<td>0.262</td>
</tr>
<tr>
<td></td>
<td>$\sigma_{XG}$</td>
<td>2972.486</td>
<td>6336.609</td>
<td>-</td>
<td>1.840</td>
</tr>
<tr>
<td></td>
<td>BIAS</td>
<td>-6.398</td>
<td>-7.482</td>
<td>-</td>
<td>-0.582</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>7.280</td>
<td>8.393</td>
<td>-</td>
<td>0.639</td>
</tr>
</tbody>
</table>
Table A: The selected small scale datasets from the CLASH, for wave overtopping on the rubble mound structures.

<table>
<thead>
<tr>
<th>Dataset ID</th>
<th>Number of data</th>
<th>$R_c/H_S$</th>
<th>$G_c/H_S$</th>
<th>$\bar{z}_{top}$</th>
<th>$\tan(\alpha)$</th>
<th>$q^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS-3</td>
<td>5</td>
<td>1.07~1.71</td>
<td>0.71~3.12</td>
<td>3.17~3.36</td>
<td>0.67~0.67</td>
<td>0.23E-04~0.11E-02</td>
</tr>
<tr>
<td>DS-4</td>
<td>6</td>
<td>0.50~1.37</td>
<td>0.79~1.54</td>
<td>2.60~2.98</td>
<td>0.40~0.40</td>
<td>0.25E-04~0.11E-01</td>
</tr>
<tr>
<td>DS-15</td>
<td>12</td>
<td>1.31~1.81</td>
<td>0.00~2.04</td>
<td>1.32~2.09</td>
<td>0.20~0.25</td>
<td>0.77E-05~0.12E-03</td>
</tr>
<tr>
<td>DS-21</td>
<td>13</td>
<td>0.59~1.15</td>
<td>1.00~1.54</td>
<td>2.42~2.64</td>
<td>0.50~0.50</td>
<td>0.87E-05~0.18E-02</td>
</tr>
<tr>
<td>DS-27</td>
<td>11</td>
<td>0.82~1.57</td>
<td>1.49~3.88</td>
<td>2.42~9.27</td>
<td>0.50~0.75</td>
<td>0.14E-04~0.38E-02</td>
</tr>
<tr>
<td>DS-32</td>
<td>13</td>
<td>1.00~1.88</td>
<td>0.73~0.92</td>
<td>1.96~3.26</td>
<td>0.50~0.50</td>
<td>0.17E-04~0.21E-02</td>
</tr>
<tr>
<td>DS-35</td>
<td>26</td>
<td>0.69~1.53</td>
<td>0.74~1.23</td>
<td>2.23~3.78</td>
<td>0.50~0.50</td>
<td>0.76E-05~0.47E-02</td>
</tr>
<tr>
<td>DS-45</td>
<td>168</td>
<td>1.13~2.53</td>
<td>0.66~1.47</td>
<td>2.10~5.94</td>
<td>0.50~0.50</td>
<td>0.50E-05~0.33E-02</td>
</tr>
<tr>
<td>DS-211</td>
<td>14</td>
<td>1.21~1.89</td>
<td>1.07~2.25</td>
<td>3.57~3.80</td>
<td>0.67~0.67</td>
<td>0.80E-05~0.53E-03</td>
</tr>
<tr>
<td>DS-221</td>
<td>4</td>
<td>1.12~1.80</td>
<td>0.00~0.00</td>
<td>1.65~3.04</td>
<td>0.33~0.33</td>
<td>0.23E-04~0.14E-03</td>
</tr>
<tr>
<td>DS-304</td>
<td>11</td>
<td>1.25~1.63</td>
<td>1.75~2.28</td>
<td>3.84~4.02</td>
<td>0.67~0.67</td>
<td>0.21E-04~0.10E-02</td>
</tr>
<tr>
<td>DS-326</td>
<td>10</td>
<td>1.03~1.50</td>
<td>1.87~2.35</td>
<td>2.89~3.52</td>
<td>0.50~0.50</td>
<td>0.80E-05~0.10E-03</td>
</tr>
<tr>
<td>DS-328</td>
<td>7</td>
<td>0.77~1.01</td>
<td>1.64~2.15</td>
<td>2.33~3.25</td>
<td>0.50~0.50</td>
<td>0.87E-05~0.18E-03</td>
</tr>
<tr>
<td>DS-330</td>
<td>5</td>
<td>0.78~1.02</td>
<td>0.86~1.65</td>
<td>3.42~3.94</td>
<td>0.67~0.67</td>
<td>0.29E-04~0.11E-02</td>
</tr>
<tr>
<td>DS-331</td>
<td>53</td>
<td>0.63~1.10</td>
<td>1.15~2.72</td>
<td>2.41~3.62</td>
<td>0.50~0.75</td>
<td>0.69E-05~0.68E-02</td>
</tr>
<tr>
<td>DS-352</td>
<td>8</td>
<td>1.46~1.77</td>
<td>1.04~1.20</td>
<td>3.17~4.56</td>
<td>0.50~0.67</td>
<td>0.17E-04~0.59E-03</td>
</tr>
<tr>
<td>DS-356</td>
<td>10</td>
<td>1.54~2.59</td>
<td>0.60~1.01</td>
<td>3.73~7.91</td>
<td>0.75~0.75</td>
<td>0.55E-04~0.51E-03</td>
</tr>
<tr>
<td>DS-379</td>
<td>31</td>
<td>0.81~1.65</td>
<td>1.10~1.36</td>
<td>3.14~5.45</td>
<td>0.67~0.67</td>
<td>0.88E-05~0.54E-02</td>
</tr>
<tr>
<td>DS-510</td>
<td>152</td>
<td>0.62~1.83</td>
<td>0.78~1.78</td>
<td>2.06~5.04</td>
<td>0.50~0.67</td>
<td>0.14E-04~0.85E-02</td>
</tr>
<tr>
<td>DS-701</td>
<td>75</td>
<td>0.92~1.59</td>
<td>1.39~2.41</td>
<td>1.96~4.48</td>
<td>0.44~0.67</td>
<td>0.11E-04~0.13E-02</td>
</tr>
<tr>
<td>DS-702</td>
<td>60</td>
<td>0.61~1.57</td>
<td>0.96~1.71</td>
<td>2.75~5.19</td>
<td>0.67~0.67</td>
<td>0.82E-05~0.54E-02</td>
</tr>
<tr>
<td>DS-705</td>
<td>489</td>
<td>0.53~1.77</td>
<td>0.79~2.19</td>
<td>2.03~4.27</td>
<td>0.50~0.50</td>
<td>0.75E-05~0.16E-02</td>
</tr>
<tr>
<td>DS-801</td>
<td>15</td>
<td>1.11~1.52</td>
<td>2.69~3.70</td>
<td>2.93~3.44</td>
<td>0.50~0.50</td>
<td>0.80E-04~0.16E-02</td>
</tr>
<tr>
<td>DS-913</td>
<td>38</td>
<td>1.06~1.99</td>
<td>0.83~1.25</td>
<td>1.96~3.97</td>
<td>0.33~0.33</td>
<td>0.17E-03~0.68E-02</td>
</tr>
<tr>
<td>DS-915</td>
<td>14</td>
<td>1.24~1.91</td>
<td>0.62~0.95</td>
<td>3.94~6.25</td>
<td>0.75~0.75</td>
<td>0.19E-04~0.38E-02</td>
</tr>
<tr>
<td>DS-916</td>
<td>47</td>
<td>1.09~1.94</td>
<td>0.76~1.11</td>
<td>3.85~7.55</td>
<td>0.75~0.75</td>
<td>0.12E-03~0.92E-02</td>
</tr>
<tr>
<td>DS-917</td>
<td>23</td>
<td>1.19~2.55</td>
<td>0.00~0.00</td>
<td>2.36~7.22</td>
<td>0.29~0.50</td>
<td>0.11E-04~0.17E-02</td>
</tr>
<tr>
<td>DS-954</td>
<td>98</td>
<td>0.65~1.51</td>
<td>1.31~2.41</td>
<td>1.31~2.65</td>
<td>0.25~0.25</td>
<td>0.62E-05~0.19E-02</td>
</tr>
<tr>
<td>Total</td>
<td>1418</td>
<td>0.50~2.59</td>
<td>0.00~3.88</td>
<td>1.31~9.27</td>
<td>0.20~0.75</td>
<td>0.50E-05~0.11E-01</td>
</tr>
</tbody>
</table>