https://dx.doi.org/10.1016/j.apor.2020.102223

Random forest based artificial intelligent model for predicting failure envelopes of caisson foundations in sand

Pin ZHANG[1,2], Yin-Fu JIN[1*], Zhen-Yu YIN[1*] and Yi Yang[3]

1 Department of Civil and Environmental Engineering, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China

2 Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), 1119 Haibin Rd., Nansha District, Guangzhou, China

3 Department of Civil Engineering, Chu Hai College of Higher Education, Hong Kong, P.R. China.

* Corresponding author: Dr. Y-F JIN, E-mail: yinfu.jin9019@gmail.com; Dr Z-Y YIN, Tel: +852 3400 8470, Fax: +852 2334 6389, E-mail: zhenyu.yin@polyu.edu.hk; zhenyu.yin@gmail.com;

Abstract: To reduce the computational cost and improve the accuracy in predicting failure envelopes of caisson foundations, this study proposes an intelligent method using random forest (RF) based on data extended from experiments and calibrated numerical simulations. Two databases are built from the numerical results by coupled Lagrangian finite element method and smoothed particle hydrodynamics with a critical state based simple sand model (CLSPH-SIMSAND). The first database involves the failure envelopes of caisson foundations with various specifications for a given sand, and the second database includes two additional failure envelopes of caisson foundations in other granular soils. The relationship between the characteristic measures of failure envelope and sand properties as well as the specification of caisson foundation is trained by RF using the prepared databases. The results indicate the RF based model is able to accurately learn the failure mechanism of caisson foundation from the raw data. Once a RF based model that can accurately reproduce the failure envelopes of caisson foundations in a given sand is developed, it can be easily modified to predict the failure envelopes of caisson foundations in a random granular soil as long as one numerical result in such soil is added to the database. Therefore, the RF based model is much more convenient than the calibration of parameters used in the conventional analytical solutions and the computational cost is much less than the conventional numerical modelling methods.

Keywords: Caisson foundation; sand; failure envelope; numerical modelling; machine learning; finite element method

1. Introduction

Suction caisson foundation is a form of fixed platform anchor in the form of an open bottomed tube embedded in the seabed, which consists of a thin-walled upturned 'bucket' with cylindrical shape. The caisson penetrates into the seabed by self-weigh and pumping out the water trapped in the caisson cavity. The bearing capacity of foundation can be improved by squeezing the soil in the 'bucket' during undrained loading [1-3]. The caisson foundation has a number of advantages over conventional offshore foundations because of the easy installation and high capacity. Suction caissons are now used extensively worldwide for anchoring large offshore installations, like oil platforms, offshore drillings and accommodation platforms to the seafloor at great depths. In recent years, suction caissons have also seen usage for offshore wind turbines in shallow waters [2, 4, 5]. Caisson foundations of such structures are typically subjected to complex loading force with the combination of vertical (V), horizontal (H) and moment (M) loads [6]. Compared with vertical capacity, horizontal and moment capacities are more critical due to the offshore system responses in combination with the environmental loading [7]. Currently, the state-of-the-art design method for evaluating the bearing capacity of a caisson foundation under combined loading is to use the failure envelope in the V–H–M or H–M space [8, 9].

Generally, the methods for determining the failure envelope can be categorized into two groups: experiment and numerical modelling. The most realistic and intuitive way to obtain the failure envelope is to conduct the experiments, such as field test, centrifuge test and laboratory model test. However, the shape of failure envelope is complex because it is affected by soil properties and specifications of caisson foundations. The cost of experiments to cover all possible paths is large. Moreover, the physical tests tend to be simplified, thereby only limited factors can be considered and it is thus hard to comprehensively capture the failure mechanism of caisson foundation. Compared to the experiment, the numerical

modelling has been extensively used to obtain the failure envelopes [9-12] due to the fast development of computer techniques. However, the computational cost of numerical modelling is also expensive when conducting numerous simulations each time for a given caisson foundation. Furthermore, the calibration of parameters used in constitutive model still requires considerable skills although some parameter identification methods [13-19] and tools [20] have been developed. Furthermore, the yield envelope is a key component in macroelement models, most typical analytical method characterized by computational efficiency with simple formulation and ease of numerical implementation with the integration of finite element code [21-23]. However, parameters of yield envelope need to be calibrated in advance, which brings the challenge to engineers. Therefore, an effective method to predicting the failure envelope, which can comprehensively account for multiple influential factors and guarantee its accuracy, deserves to be developed.

The application of machine learning (ML) algorithms in engineering has recently proliferated such as offshore [24-28], geotechnical [29-32] and structural engineering [33-35], because such algorithms are able to directly learn from raw data and capture the intricate relationship in high-dimensional input and output parameters, meanwhile the accuracy and computational efficiency can be ensured [36-38]. Thus, the ML is a potential method to fast predict the failure envelope of caisson foundation. Various influential factors can be assigned as the input parameters, thereby a well-trained model can make an accurate prediction with less computational cost for cases with various soil types and caisson foundation specifications. The excellent generalization ability of ML algorithms is one of important factors to solve such problem. Numerous research works have comprehensively compared the prediction performance of various ML algorithms such as artificial neural network (ANN), support vector machine (SVM), extreme learning machine (ELM), the evolutionary polynomial regression (EPR) and random forest (RF). RF as an ensemble algorithm has been discovered to exhibit superior performance to other ML algorithms in

74 various domain such as the prediction of tunneling-induced settlement and soil properties [37, 39-42].
75 The overfitting problem of RF is much less than other ML algorithms, and [43] even stated that RF
76 cannot suffer from overfitting problem. Therefore, it is worth trying to apply the RF for predicting failure
77 envelopes of caisson foundations in sand.

78 As one of ML algorithms, RF requires large number of data which is very expensive in engineering
79 practice. One possible solution is to use advanced numerical modeling well calibrated from limited real
80 data (such as coupled Lagrangian finite element method and smoothed particle hydrodynamics with a
81 critical state based simple sand model (CLSPH-SIMSAND) by [9]), through which more simulations can
82 be conducted for different soil and loading conditions to extend database as much as possible. Then, some
83 of data can be used to train the ML based model and the rest for testing the model. This study follows this
84 logic.

85 Hence, this study aims to combine RF with CLSPH-SIMSAND to develop an intelligent model for
86 predicting the failure envelope of caisson foundations in the H–M space. The failure envelopes of various
87 specifications of caisson foundation in a given sand are first generated using CLSPH-SIMSAND. The
88 numerical modelling cases are sufficient enough to ensure that RF can learn the failure mechanism of
89 caisson foundations with various specifications in a given soil. To extend the application scopes of the RF
90 based model, one failure envelope of caisson foundations in sand with different properties are added to
91 the database, which is used to refine RF model so that it can capture the mechanical responses of caisson
92 foundations in sand with novel properties. In this way, the failure envelopes of caisson foundations with
93 various specifications in a random granular soil can be efficiently and accurately predicted.

2. Random forest based methodology

2.1 Brief introduction of random forest

RF is an ensemble algorithm with a collection of unpruned decision trees [43]. Bagging [44] and random feature selection [45] are integrated into the RF to overcome the poor generalization ability of a single decision tree. Fig. 1 illustrates the process of building a RF. A total of *n* subsets is first generated from the training set by bootstrap sampling, thereafter each decision tree is grown independently based on the corresponding bootstrap set. The number of features used for splitting nodes in each decision tree is randomly selected from the feature space of the training set. Such randomization manipulates high diversity among trees in the forest. The output values in each bootstrap set constitute a regression space, as presented in Fig. 1. The training process of RF terminates with the formation of *n* regression spaces. Given a new input dataset **x**, *n* decision trees will generate *n* output value y_i, and the ultimate output of RF is obtained by the aggregation of the results from all decision trees, as shown:

$$y = \frac{1}{n}\sum_{i=1}^{n} y_i(\mathrm{x}) \tag{1}$$

where y_i (**x**) = individual prediction of a tree for an input **x**; *n* = a total of DTs. The training of RF is to determine the values of hyper-parameters. The hyper-parameters of RF are presented in Table 1. The performance of a RF based model is primarily affected by *ntree*, *mtry* and *min_sample_split*. *ntree* controls the number of trees in the RF. The increasing number of *ntree* improves the performance of a RF based model and cannot induce overfitting, but the computational cost is expensive. The diversity of trees in the forest depends on the value of *max_features*. The increasing diversity can improve the generalization ability of RF, but the number of trees needs to increase simultaneously to guarantee its stable prediction performance. The depth of tree is controlled by *min_sample_split*. Setting *min_sample_split* larger causes smaller trees to be grown, thereby may reduce the generalization ability of

116 the RF based model, but small *min_sample_split* easily causes overfitting problem and increases
117 computational cost. Considering only three hyper-parameters need to be determined, grid search method
118 is used to determine their values, because it can traverse all combinations of hyper-parameters and the
119 optimum combination of hyper-parameters can thus be effectively obtained. The prescribed ranges of
120 three hyper-parameters are presented in Table 1, which are sufficiently large to determine the optimum
121 values. The remaining three hyper-parameters can be automatically obtained during the training process
122 after the values of *ntree*, *mtry* and *min_sample_split* are determined.

123 **2.2 Failure envelope description**

124 Physical and numerical modelling results have revealed that the shape of failure envelope is similar to an
125 ellipse in the H–M plane [46-48], thereby analytical solutions focus on using ellipse to fit the failure
126 envelope [9, 49]. [9] have proposed a formulation with three parameters to describe the failure envelope
127 of a caisson foundation in the H–M plane. The definition of parameters used in [9]'s method for
128 describing the failure envelope centered at the origin is presented in Fig. 2, in which the general form of
129 ellipse can be expressed as follows:

130
$$\mathrm{A}_1 X^2 + A_2 XY + A_3 Y^2 + A_4 = 0 \tag{2}$$

131
$$\begin{cases} \mathrm{A}_1 = a^2 \left(\sin\varphi\right)^2 + b^2 \left(\cos\varphi\right)^2 \\ \mathrm{A}_2 = 2\left(b^2 - a^2\right)\sin\varphi\cos\varphi \\ \mathrm{A}_3 = a^2 \left(\cos\varphi\right)^2 + b^2 \left(\sin\varphi\right)^2 \\ \mathrm{A}_4 = -a^2 b^2 \end{cases} \tag{3}$$

132 where a and b = major and minor axis of the failure envelope, respectively; φ = rotation of the failure
133 envelope. The failure enveloped can be obtained if the values of a, b and φ are determined. The RF based
134 model is employed to predict the values of a, b and φ, thereafter the failure envelope can be mapped,
135 which dramatically reduces the computational cost and is also much easier than the direct prediction of
136 failure envelopes.

2.3 Proposed RF-based model for failure envelope prediction

ML is able to take advantage of increases in the amount of available data [50]. The development of ML based model to predict failure envelope thus relies on the collection of experimental or numerical data. The implementation of experimental tests with various soil types or caisson foundation specifications is expensive and time-consuming, thereby it is hard to develop a database based on experiments. Synthetic datasets have been successfully used to develop the ML based model in published research works [19, 51], thereby numerical modelling is a flexible method to generate numerous datasets to train the ML based failure envelope prediction model. However, the combinations of all influential factors are extremely large, it is thus not realistic to simulate all cases. A substitute method is to use numerical modelling to acquire sufficient failure envelopes of caisson foundations in a given soil type, which ensures ML to identify the failure mechanism in such soil type. Next, by adding an additional numerical result in a studied soil type to the original database, ML based model is modified to be adapted to capture the failure mechanism of caisson foundation in the studied soil type.

The flowchart of developing combined RF and CLSPH-SIMSAND model to predict failure envelope of caisson foundation is presented in Fig. 3. The specification of a caisson foundation is represented by skirt length (L) and outer diameter (D). The properties of soils relative density (D_r), friction angle at critical state (ϕ_c), plastic modulus (k_p) and sand grain crushability (Cr) are taken into consideration. Two databases are first generated by using CLSPH-SIMSAND. Herein, the first database includes failure envelopes of various specifications of caisson foundations for a given soil type (sand-O in this study). Such database ensures RF can learn the failure mechanism of a caisson foundation in a given soil type from sufficient datasets. To extend the application scopes of the RF based model, a numerical result of a caisson foundation in the studied soil types in which some properties differ from sand-O is added to the original database, which constitutes the second database. Two studied soil types (sand-I and sand-II in

this study) are used to examine the reliability of the RF based model in this research work. It should be noted that in each studied soil type, only one numerical result of a random caisson foundation specification is appended to the second database. Such additional dataset is used to modify the RF-based model with least computational cost, meanwhile the failure envelopes of caisson foundations in a studied soil type are ensured to be identified. The performance of the RF based model developed based on the first and second databases can be compared to examine the effect of an additional dataset on the improvement in the generalization ability of model. Therefore, the RF based model is trained with six input parameters and three output parameters. Failure envelopes of random specifications of caisson foundation in the studied soil types are predicted using the RF based model, and they are compared with the numerical results of CLSPH-SIMSAND to test the reliability of RF-based model.

2.4 Evaluation indicators

Two indicators mean absolute percentage error (MAPE) and Nash–Sutcliffe model efficiency (NSE) coefficient (see Eqs. [4]–[5]) are used to evaluate model performance. MAPE is an unbiased measure to evaluate the average prediction error of a model, and NSE is used to assess the accuracy of a model [52]. Low value of MAPE and high value of NSE indicate that a model has an excellent performance.

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i^p - y_i^m}{y_i^m}\right| \times 100\% \tag{4}$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}\left(y_i^p - y_i^m\right)^2}{\sum_{i=1}^{n}\left(y_i^m - \overline{y_i^m}\right)^2} \tag{5}$$

where y_i^m and y_i^p = measured and predicted ith output; $\overline{y}^m$ = mean value of measured output; n = a total of datasets.

3. Database of failure envelope

3.1 Data source

[9] proposed a coupled Lagrangian finite element method and smoothed particle hydrodynamics method (CLSPH) combining with an advanced critical state-based sand model (SIMSAND) for the identification of the failure envelope of a caisson foundation in the Baskarp sand with the unit weight of 19.5 kN/m^3 and the specific gravity of 26.5 kN/m^3 [53]. The parameters of the SIMSAND constitutive model were first calibrated using triaxial tests on Baskarp sand. The reliability of numerical results of elaborate CLSPH-SIMSAND based model has been fully validated through one cone penetration test, six model tests and one field test on a reduced scale caisson foundation. Therefore, the numerical results from such research work are used in this study due to the reliability of CLSPH-SIMSAND based model and simulations. The responses of a caisson foundation with L = 1 m and D = 2 m as a typical example are used to introduce the modelling results by CLSPH-SIMSAND, as presented in Fig. 4. Radial displacement tests modelling strategy is used to identify the failure envelope of a caisson foundation, in which the ratio between the applied displacements or the combined rotation-displacement increments maintains constant [54]. Half of applied horizontal displacement and rotation of caisson foundation are presented in Figs. 4(a) and (b), respectively, in which the maximum horizontal displacement increases from 0 to 0.8 m with an interval of 0.04 and the rotation maintains in the range of 0 to 0.2 rad. Another half of applied horizontal displacement and rotation is opposite to the values presented in Figs. 4(a) and (b), but the absolute value keeps unchanged, which is not shown for brevity. Fig. 4(c) presents the loading paths in the H-M plane and the corresponding failure loci. A total of 44 failure loci are determined by the inflexion of loading paths or the ends of loading paths for which the ultimate bearing capacity is hard to reach [5]. Based on the obtained failure loci, the failure envelope can be fitted using Eq. [2], and the corresponding values of a, b and φ can be obtained, as shown in Fig. 4(d). A total of nine parameters are

collected (see Fig. 3) when each simulation is completed.

As mentioned in section 2.3, sand-I and sand-II are used to examine the reliability of the RF based model in this studied. Compared with the sand-O, D_r of sand-I and φ_c of sand-II are changed, respectively. Table 2 summarizes the values of soil parameters of sand-O, sand-I and sand-II. In the first database, the training set includes 56 datasets with different caisson foundation specifications in the sand-O, and the testing set includes 10 datasets with different caisson foundation specifications in the sand-I and sand-II. In second database, 2 datasets with L = 1 m and D = 2 m in the sand-I and sand-II are added to the training set of the database 1, and thus the number of datasets in the testing set decreases to 8. The database used in this study can be freely downloaded for your check and use (see Appendix).

3.2 Global sensitivity analysis

Sensitivity analysis aims to investigate how the model output uncertainty can be apportioned to the uncertainty in each input variable [55], thereby the significance of input parameters to the model can be determined. Variance-based sensitive analysis as a typical global sensitivity analysis method has been successfully used in many domains [36, 56-58], because such method can account for the whole variable space and the coupled effects of parameters on the model can also be considered.

Variance-based sensitivity indices denote the effect of the variation of an input variable on the output. Given a model with k features [$\boldsymbol{X}_1$, $\boldsymbol{X}_2$, …, $\boldsymbol{X}_k$], the model can be expressed by:

$$\mathrm{Y} = f\left(\mathrm{X}_1, \mathrm{X}_2, \ \ldots, \mathrm{X}_i, \ \ldots, \mathrm{X}_k\right) \tag{6}$$

Based on the analysis of variance (ANOVA) functional decomposition [59], the model can be re-expressed by:

$$\mathrm{Y} = f_0 + \sum_{i=1}^{k} f_{\mathrm{i}} + \sum_{i=1}^{k} \sum_{i<j} f_{ij} + \cdots + f_{12 \ldots k} \tag{7}$$

where $f_{\mathrm{i}} = f_{\mathrm{i}}\left(\mathrm{X}_i\right)$, $f_{\mathrm{ij}} = f_{\mathrm{ij}}\left(\mathrm{X}_i, \mathrm{X}_j\right)$ and so on. The total variance can thus be obtained by:

$$V(\mathrm{Y}) = \sum_{i=1}^{k} V_i(\mathrm{Y}) + \sum_{i=1}^{k}\sum_{i<j} V_{ij}(\mathrm{Y}) + \cdots + V_{ij\cdots k}(\mathrm{Y}) \tag{8}$$

where $V_i = V\left(f_i(\mathrm{X}_i)\right)$, $V_{ij} = V\left(f_{ij}(\mathrm{X}_i, \mathrm{X}_j)\right)$ and so on.

The *Sobol*' sensitivity indices are defined as the ratio of partial variance to the total variance [60], thereby the first order S_i and total order S_{Ti} sensitivity indices can be obtained by:

$$S_i = \frac{V_{X_i}\left(E_{X_{\sim i}}(\mathrm{Y} \mid \mathrm{X}_i)\right)}{V(\mathrm{Y})} \tag{9}$$

$$S_{T_i} = 1 - \frac{V\left(E(\mathrm{Y} \mid \mathbf{X}_{\sim i})\right)}{V(\mathrm{Y})} = \frac{E\left(V(\mathrm{Y} \mid \mathbf{X}_{\sim i})\right)}{V(\mathrm{Y})} \tag{10}$$

where $\boldsymbol{X}_i$ = *i*th feature; $\mathbf{X}_{\sim i}$ = matrix of all features without including $\boldsymbol{X}_i$; $\boldsymbol{Y}/\boldsymbol{X}_i = \boldsymbol{Y}$ values under a given $\boldsymbol{X}_i$; $\boldsymbol{Y}/\mathbf{X}_{\sim i} = \boldsymbol{Y}$ values under $\mathbf{X}_{\sim i}$ matrix.

S_i is used to evaluate the effect of feature $\boldsymbol{X}_i$ on the model. S_{Ti} is used to evaluate the total effect of feature, i.e. $\boldsymbol{X}_i$ as well as the interaction between $\boldsymbol{X}_\mathrm{i}$ and remaining features. S_i proposed by [61] and S_{Ti} proposed by [62] are used in this study because of the superiority of such two indices [61], which can be obtained by:

$$V_{X_i}\left(E_{\sim i}(Y \mid X_\mathrm{i})\right) = \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{B})_j \left(f\left(\mathbf{A}_\mathrm{B}^{(i)}\right)_j - f(\mathbf{A})_j\right) \tag{11}$$

$$E\left(V(Y \mid X_{\sim i})\right) = \frac{1}{2n}\sum_{j=1}^{n}\left(f(\mathbf{A})_j - f\left(\mathbf{A}_\mathrm{B}^{(i)}\right)_j\right)^2 \tag{12}$$

$$V(Y) = \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{A})_j^2 - \left(\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{A})_j\right)^2 \tag{13}$$

where n = a total of datasets; $\mathbf{A}$, $\mathbf{B}$ = two independent dataset matrices with $n/2$; $\mathbf{A}_\mathrm{B}^{(i)}$ = a matrix that all columns are from $\mathbf{A}$ except the elements in the *i*th column, which is from $\mathbf{B}$. Therefore, a model with k features means it has k $\mathbf{A}_{\boldsymbol{B}}$ matrices.

As mentioned in the section 2.2, the failure envelope of a caisson foundation can be expressed using $f(a, b, \varphi)$. Given 1000 random datasets, in which a ranges from 1000 to 2000, b ranges from 500 to 1000, and φ ranges from 20 to 60. All of them obey uniform distribution in the corresponding range. Fig. 5

presents the values of sensitive indices of three characteristic measures. It can be observed that the S_i and S_{Ti} values of a are much larger than the b and φ, which indicates the shape of failure envelope of a caisson foundation is primarily affected by a, followed by b and φ. The coupled effects of characteristic measures on the failure envelope cannot be neglected especially for φ, in which the coupled effect is more obviously than itself.

3.2 Data preprocessing

The performance of the model is affected by the scales of input and output parameters. The datasets thus need to be pre-processed. Min-Max normalization method as a linear transformation method has been successfully used in many domains because of its simplicity and effectiveness [36, 63, 64], which is thus used in this study. Such method normalizes all parameters into a same scale using

$$x_{norm} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \left(\bar{x}_{\max} - \bar{x}_{\min} \right) + \bar{x}_{\min} \tag{14}$$

where x = actual value; $x_{\max}$ and $x_{\min}$ = actual maximum and minimum of the parameter x in the database; $\bar{x}_{\max}$ and $\bar{x}_{\min}$ = prescribed lower and upper bounds of normalized x, which are equal to –1 and 1 in this study.

4. Development of RF-based model

4.1 Determination of *min_sample_split*

As mentioned in the section 2.1, the grid search method is used to determine the values of *ntree*, *mtry* and *min_sample_split* of the RF based model. To evaluate the performance of the RF based model, the learning curves on both training and testing sets are investigated. Such method evaluates the model performance by calculating the loss values on both training and testing sets, and it has also been successfully used to examine the underfitting and overfitting problems [65]. Large loss values on both

266 training and testing sets represent that the RF based model exists underfitting problem. Small loss value
267 on the training set and the large loss value on the testing set represent that the RF based model has
268 overfitting problem. Therefore, a model with small loss values on both training and testing sets exists
269 excellent performance. Herein, the loss value is defined as the MAPE as shown in Eq. [4].

270 The loss values on the training and testing sets generated by the RF-based model with different
271 values of *min_sample_split* are presented in Fig. 6. It can be seen from Fig. 6(a) that the performance of
272 the RF based model for the prediction of *a* varies dramatically with the change of *min_sample_split*. The
273 model with *min_sample_split* of 3 obviously loses fidelity on the training set in comparison with other
274 values of *min_sample_split*, although the loss value on the testing set is lowest, thereby such model is not
275 robust and the results are not reliable. Overall, the optimum value of *min_sample_split* is 5, in which the
276 ranges of loss values on both training and testing sets are small, meanwhile the mean value on the testing
277 set is less than the model with other values of *min_sample_split*.

278 Fig. 6(b) presents the effect of *min_sample_split* on the RF based model for predicting *b*. It is clear
279 that the RF based model for predicting *b* is less sensitive to the *min_sample_split* than the prediction of *a*.
280 With the change of *min_sample_split*, the range and mean value of loss values on the training set are
281 roughly identical. On the testing set, the range and mean value of loss values first increase with the value
282 of *min_sample_split* increasing from 1 to 3, thereafter the increasing *min_sample_split* leads to the
283 increase in the range and mean value of loss values, thereby the optimum value of *min_sample_split* is
284 here identical to 6.

285 It can be observed from Fig. 6(c) that the range of loss values on both training and testing sets for the
286 prediction of φ are much less than the predictions of *a* and *b*, which is attributed to the smaller range of φ
287 in the database in comparison with *a* and *b*. The range and mean of loss value on the training set
288 generated by the RF sub-based model for predicting φ with *min_sample_split* of 2 are relatively less than

that generated by the model with other values of *min_sample_split*, meanwhile the range of loss value on the testing set is also small. Therefore, the optimum value of *min_sample_split* is selected as 2.

4.2 Determination of *ntree* and *mtry*

After the value of *min_sample_split* is determined, the coupled effect of *ntree* and *mtry* can be revealed, as presented in Fig. 7. From the perspective of Fig. 7(a), the performance of the RF based model for predicting *a* maintains steadily as the *ntree* exceeds 100, and the corresponding loss values on the training and testing sets decrease with the increasing value of *mtry*. As the value of *mtry* is equal to 1, the RF based model has underfitting problem with high loss values on training and test sets. The performance RF based sub-model with *mtry* of 5 and 6 improves continuously with the increasing *ntree*. Because the trees are diverse in the RF as the *mtry* is identical to 2, 3 and 4, the loss values generated by the corresponding RF based model vary dramatically. The effect of such diversity can be eliminated with the increasing *ntree*, thereby the loss values are roughly unchanged as the *ntree* reaches 100. Herein, the optimum values of *ntree* and *mtry* are identical to 3 and 4, respectively. In Figs. 7(b) and (c), it can be observed that the evolution of loss value on the training set is similar to results generated by the RF based model for predicting *a*. Regarding the RF based model for predicting *b*, an obvious difference on the testing set is that the evolution of loss values is roughly consistent as the *mtry* reaches 4. The optimum values of *ntree* and *mtry* for the RF based model for predicting *b* are 16 and 6, respectively. For the RF based model for φ, it can be observed that the loss value on the testing set generated by RF based model with various *mtry* values continuously decreases with the increasing *ntree*. The model with *ntree* = 82 and *mtry* = 6 shows the best performance. It should be noted that this section only presents the process for determining the optimum hyper-parameters of the RF model based on the second database. The process for developing the optimum RF model based on the first database is similar, which is thus not presented for brevity. Table 3 summarizes the optimum values of *min_sample_split*, *ntree* and *mtry* used in RF

based model developed based on the first and second databases. The optimum model for predicting a, b and φ can be freely downloaded for your check and use (see Appendix).

5. Performance of RF based model

Table 4 presents the values of indicators generated by the RF based model developed based on the first and second databases. The objective of RF based model developed based on the first database is to ensure it can learn the failure mechanism of a caisson foundation in the sand-O. Utilizing such model to directly predict the failure envelopes of caisson foundations with various specifications in the sand-I and sand-II, it can be seen from Table 4 that the prediction error is large, and the largest MAPE reaches 23.7% for predicting a. For the RF based model developed based on the second database, that is, a numerical result in the sand-I and sand-II is added to the database, the errors for predicting a, b and φ obvious reduces especially on the testing set, in which the largest MAPE decreases to 9.5%. NSE values always maintain over 0.9 for the prediction of a, b and φ using the RF based model developed based on both databases. Fig. 8 presents the scatter plots of predicted characteristic measures a, b and φ. It can be observed that the predicted results show excellent agreement with actual results, and all points are close to the line with the slope of 1. It means that the performance of RF based model can be well modified to accurately predict the characteristic measures of failure envelopes of caisson foundations with various specifications in the studied soils, as long as the datasets in a soil type is sufficient enough and one or several numerical results in the studied soil types are included in the database.

Based on the predicted characteristic measures a, b and φ, the failure envelope can be obtained by Eq. [2]–[3]. Fig. 9 and 10 present the predicted failure envelopes of caisson foundations with five specifications in the sand-I and sand-II, compared with numerical results. It is clear that the predicted failure envelopes are roughly consistent with the numerical results. The largest discrepancy between the

predicted and actual failure envelopes exists in the prediction of the caisson foundation with L = 2 and D = 2 m in the sand-II. Global sensitivity analysis indicates the shape of failure envelope is mainly affected by *a*, thereby the prediction error on *a* leads to an obvious difference between the predicted and actual failure envelopes. Overall, the application scopes of RF based model can be extended by adding one or several numerical results in the studied soils to the database. The predicted results are reliable and the main error origins from the prediction of *a* with the value less than 10%.

6. Conclusions

This study combined random forest (RF) with numerical modelling method CLSPH-SIMSAND to develop an intelligent model for predicting failure envelopes of caisson foundations in the H–M space. CLSPH-SIMSAND was employed to generate synthetic datasets of failure envelope. Three characteristics of failure envelope that are the major axis *a*, minor axis *b* and rotation φ of the failure envelope were extracted, thereafter the RF based model was trained to predict *a*, *b* and φ to reproduce failure envelopes. Consequently, the first database was built with the failure envelopes of various specifications of caisson foundations in a given soil, and the second database included additional failure envelopes of caisson foundations in sand with different properties. RF is able to learn the failure mechanism of caisson foundation in a given soil based on the sufficient datasets in the first database. By adding one numerical result in the studied soil types in the database, the RF based model can be well modified to accurately predict the failure envelopes of caisson foundations with various specifications in other soil types, which is much more convenient than the calibration of parameters used in the conventional analytical solutions. Therefore, the predicted results of combined RF and CLSPH-SIMSAND method are reliable and its application scopes can be effectively extended by adding one or several numerical results in a random soil type to the original database with the prediction error less than 10%. Moreover, the prediction of the

356 combined RF and CLSPH-SIMSAND method can be completed within several seconds, thereby the
357 computational cost is much less than the conventional numerical modelling methods.
358

359 ## Appendix

360 The database and the code used in this study can be downloaded at following link:
361 https://www.researchgate.net/publication/339289968_RF_based_failure_envelope_prediction_model
362

363 ## Acknowledges

364 This research was financially supported by Key Special Project for Introduced Talents Team of Southern
365 Marine Science and Engineering Guangdong Laboratory (Guangzhou) (No.: GML2019ZD0503) and FDS
366 project (Grant No. UGC/FDS13/E06/18) from Research Grants Council (RGC) of Hong Kong. The
367 authors also would like to thank Dr. Zhuang JIN for providing the numerical simulations for this study.

368 ## References

369 [1] Mehravar M, Harireche O, Faramarzi A. Evaluation of undrained failure envelopes of caisson
370 foundations under combined loading. Applied Ocean Research. 2016;59:129-37.
371 [2] Hung LC, Kim S-R. Evaluation of undrained bearing capacities of bucket foundations under
372 combined loads. Mar Georesour Geotec. 2013;32:76-92.
373 [3] Vulpe C. Design method for the undrained capacity of skirted circular foundations. Géotechnique.
374 2015;65:669–83.
375 [4] Selmi M, Kormi T, Hentati A, Bel Hadj Ali N. Capacity assessment of offshore skirted foundations
376 under HM combined loading using RFEM. Comput Geotech. 2019;114:103148.
377 [5] Jin Z, Yin Z-Y, Kotronis P, Li Z, Tamagnini C. A hypoplastic macroelement model for a caisson
378 foundation in sand under monotonic and cyclic loadings. Mar Struct. 2019;66:16-26.
379 [6] Fu D, Bienen B, Gaudin C, Cassidy M. Undrained capacity of a hybrid subsea skirted mat with
380 caissons under combined loading. Can Geotech J. 2014;51:934-49.
381 [7] Bienen B, Gaudin C, Cassidy MJ, Rausch L, Purwana OA, Krisdani H. Numerical modelling of a
382 hybrid skirted foundation under combined loading. Comput Geotech. 2012;45:127-39.

[8] Skau KS, Chen Y, Jostad HP. A numerical study of capacity and stiffness of circular skirted foundations in clay subjected to combined static and cyclic general loading. Géotechnique. 2018;68:205-20.
[9] Jin Z, Yin Z-Y, Kotronis P, Li Z. Advanced numerical modelling of caisson foundations in sand to investigate the failure envelope in the H-M-V space. Ocean Eng. 2019;190:106394.
[10] Gerolymos N, Zafeirakos A, Karapiperis K. Generalized failure envelope for caisson foundations in cohesive soil: static and dynamic loading. Soil Dynamics and Earthquake Engineering. 2015;78:154-74.
[11] Yin S, Yi JT. Undrained bearing capacity of deeply embedded skirted spudcan foundations under combined loading in soft clay. Mar Struct. 2019;66:164-77.
[12] Suryasentana SK, Dunne HP, Martin CM, Burd HJ, Byrne BW, Shonberg A. Assessment of numerical procedures for determining shallow foundation failure envelopes. Géotechnique. 2020;70:60-70.
[13] Jin Y-F, Yin Z-Y, Zhou W-H, Horpibulsuk S. Identifying parameters of advanced soil models using an enhanced transitional Markov chain Monte Carlo method. Acta Geotech 2019;14:1925-47
[14] Jin Y-F, Yin Z-Y, Zhou W-H, Huang H-W. Multi-objective optimization-based updating of predictions during excavation. Eng Appl Artif Intell 2019;78:102-23.
[15] Jin Y-F, Yin Z-Y, Zhou W-H, Shao J-F. Bayesian model selection for sand with generalization ability evaluation. Int J Numer Anal Methods Geomech 2019;43:2305-27.
[16] Jin Y-F, Yin Z-Y, Zhou W-H, Yin J-H, Shao J-F. A single-objective EPR based model for creep index of soft clays considering L2 regularization. Eng Geol 2019;248:242-55.
[17] Yin Z-Y, Jin Y-F, Shen JS, Hicher P-Y. Optimization techniques for identifying soil parameters in geotechnical engineering: Comparative study and enhancement. Int J Numer Anal Methods Geomech 2018;42:70-94.
[18] Yin Z-Y, Jin Y-F, Shen S-L, Huang H-W. An efficient optimization method for identifying parameters of soft structured clay by an enhanced genetic algorithm and elastic–viscoplastic model. Acta Geotech 2017;12:849-67.
[19] Zhang P, Yin Z-Y, Jin Y-F, Chan THT. A novel hybrid surrogate intelligent model for creep index prediction based on particle swarm optimization and random forest. Eng Geol 2020;265:105328.
[20] Jin YF, Yin ZY. Enhancement of backtracking search algorithm for identifying soil parameters. Int J Numer Anal Met. 2020.
[21] Li Z, Kotronis P, Escoffier S, Tamagnini C. A hypoplastic macroelement for single vertical piles in sand subject to three-dimensional loading conditions. Acta Geotech. 2015;11:373-90.
[22] Skau KS, Grimstad G, Page AM, Eiksund GR, Jostad HP. A macro-element for integrated time domain analyses representing bucket foundations for offshore wind turbines. Mar Struct. 2018;59:158-78.
[23] Suryasentana SK, Burd HJ, Byrne BW, Shonberg A. A systematic framework for formulating convex failure envelopes in multiple loading dimensions. Géotechnique. 2020;70:343-53.
[24] Ahn Y, Kim Y, Kim S-Y. Database of model-scale sloshing experiment for LNG tank and application of artificial neural network for sloshing load prediction. Mar Struct. 2019;66:66-82.
[25] Zafeirakos A, Gerolymos N. Bearing strength surface for bridge caisson foundations in frictional soil under combined loading. Acta Geotech 2016;11:1189-208.
[26] Pourzangbar A, Brocchini M, Saber A, Mahjoobi J, Mirzaaghasi M, Barzegar M. Prediction of scour depth at breakwaters due to non-breaking waves using machine learning approaches. Applied Ocean Research. 2017;63:120-8.
[27] Yagci O, Kitsikoudis V. Machine learning based mapping of the wave attenuation mechanism of an inclined thin plate. Applied Ocean Research. 2015;53:107-15.

[28] Zhang P, Yin Z-Y, Zheng Y, Gao F-P. A LSTM surrogate modelling approach for caisson foundations. Ocean Eng. 2020;204:107263.
[29] Zhang P, Chen RP, Wu HN. Real-time analysis and regulation of EPB shield steering using Random Forest. Automat Constr. 2019;106:102860.
[30] Chen RP, Zhang P, Kang X, Zhong ZQ, Liu Y, Wu HN. Prediction of maximum surface settlement caused by EPB shield tunneling with ANN methods. Soils Found. 2019;59:284–95.
[31] Zhang P, Li H, Ha QP, Yin Z-Y, Chen R-P. Reinforcement learning based optimizer for improvement of predicting tunneling-induced ground responses. Advanced Engineering Informatics. 2020;45.
[32] Zhang P, Yin ZY, Jin YF, Ye GL. An AI‐based model for describing cyclic characteristics of granular materials. Int J Numer Anal Met. 2020.
[33] Yaseen ZM, Deo RC, Hilal A, Abd AM, Bueno LC, Salcedo-Sanz S, et al. Predicting compressive strength of lightweight foamed concrete using extreme learning machine model. Adv Eng Softw. 2018;115:112-25.
[34] Yun GJ, Ghaboussi J, Elnashai AS. Self-learning simulation method for inverse nonlinear modeling of cyclic behavior of connections. Comput Method Appl M. 2008;197:2836-57.
[35] Sarir P, Shen S-L, Wang Z-F, Chen J, Horpibulsuk S, Pham BT. Optimum model for bearing capacity of concrete-steel columns with AI technology via incorporating the algorithms of IWO and ABC. Eng Comput-Germany. 2019:1-11.
[36] Zhang P. A novel feature selection method based on global sensitivity analysis with application in machine learning-based prediction model. Applied Soft Computing. 2019;85:105859.
[37] Chen RP, Zhang P, Wu HN, Wang ZT, Zhong ZQ. Prediction of shield tunneling-induced ground settlement using machine learning techniques. Front Struct Civ Eng. 2019;13:1363–78.
[38] Atangana Njock PG, Shen S-L, Zhou A, Lyu H-M. Evaluation of soil liquefaction using AI technology incorporating a coupled ENN / t-SNE model. Soil Dyn Earthq Eng. 2020;130:105988.
[39] Ao Y, Li H, Zhu L, Ali S, Yang Z. Identifying channel sand-body from multiple seismic attributes with an improved random forest algorithm. Journal of Petroleum Science and Engineering. 2019;173:781-92.
[40] Matin SS, Farahzadi L, Makaremi S, Chelgani SC, Sattari G. Variable selection and prediction of uniaxial compressive strength and modulus of elasticity by random forest. Appl Soft Comput. 2018;70:980-7.
[41] Zhang P, Yin ZY, Jin YF, Chan T, Gao FP. Intelligent modelling of clay compressibility using hybrid meta-heuristic and machine learning algorithms. Geoscience Frontiers. 2020:in press.
[42] Zhang P, Wu H-N, Chen R-P, Chan THT. Hybrid meta-heuristic and machine learning algorithms for tunneling-induced settlement prediction: A comparative study. Tunnell Undergr Space Technol. 2020;99:103383.
[43] Breiman L. Random Forests. Mach Learn. 2001;45:5–32.
[44] Breiman L. Bagging Predictors. Mach Learn. 1996;24:123-40.
[45] Ho TK. The random subspace method for constructing decision forests. IEEE T Pattern Anal. 1998;20:832-44.
[46] Gourvenec S, Barnett S. Undrained failure envelope for skirted foundations under general loading. Géotechnique. 2011;61:263-70.
[47] Ibsen LB, Barari A, Larsen KA. Adaptive plasticity model for bucket foundations. J Eng Mech-ASCE. 2014;140:361-73.
[48] Liu M, Yang M, Wang H. Bearing behavior of wide-shallow bucket foundation for offshore wind

473 turbines in drained silty sand. Ocean Eng 2014;82:169-79.
474 [49] Villalobos FA, Byrne BW, Houlsby GT. An experimental study of the drained capacity of suction
475 caisson foundations under monotonic loading for offshore applications. Soils Found. 2009;49:477–88.
476 [50] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436-44.
477 [51] Reuter U, Sultan A, Reischl DS. A comparative study of machine learning approaches for modeling
478 concrete failure surfaces. Adv Eng Softw. 2018;116:67-79.
479 [52] Nash JE, Sutcliffe JV. River flow forecasting through conceptual models part I - a discussion of
480 principles. J Hydrol. 1970;10:282-90.
481 [53] Ibsen LB, Liingaard S, Nielsen SA. Bucket foundation, a status. Proceedings of the Copenhagen
482 Offshore Wind. Copenhagen, Denmark2005.
483 [54] Gottardi G, Houlsby GT, Butterfield R. Plastic response of circular footings on sand under general
484 planar. Géotechnique. 1999;49:453-69.
485 [55] Saltelli A, Sobol IM. About the use of rank transformation in sensitivity analysis of model output.
486 Reliab Eng Syst Safe. 1995;50:225-39.
487 [56] Zhao CY, Lavasan AA, Hölter R, Schanz T. Mechanized tunneling induced building settlements and
488 design of optimal monitoring strategies based on sensitivity field. Comput Geotech. 2018;97:246-60.
489 [57] Hamdia KM, Ghasemi H, Zhuang XY, Alajlan N, Rabczuk T. Sensitivity and uncertainty analysis
490 for flexoelectric nanostructures. Comput Method Appl M. 2018;337:95-109.
491 [58] Pérez-Barea JJ, Fernández-Navarro F, Montero-Simó MJ, Araque-Padilla R. A socially responsible
492 consumption index based on non-linear dimensionality reduction and global sensitivity analysis. Appl
493 Soft Comput. 2018;69:599-609.
494 [59] Sobol' IM. Sensitivity estimates for nonlinear mathematical models. Math Model Comput Exp.
495 1993;1:407-14.
496 [60] Sobol' IM. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo
497 estimates. Math Comput Simulat. 2001;55:271-80.
498 [61] Saltelli A, Annoni P, Azzini I, Campolongo F, Ratto M, Tarantola S. Variance based sensitivity
499 analysis of model output. Design and estimator for the total sensitivity index. Comput Phys Commun.
500 2010;181:259-70.
501 [62] Jansen MJW. Analysis of variance designs for model output. Comput Phys Commun. 1999;117:35-
502 43.
503 [63] Patro SGK, Sahu KK. Normalization: a preprocessing stage. arXiv. 2015;1503. 06462.
504 [64] Picasso A, Merello S, Ma Y, Oneto L, Cambria E. Technical analysis and sentiment embeddings for
505 market trend prediction. Expert Syst Appl. 2019;135:60-70.
506 [65] Hassan MM, Gumaei A, Alsanad A, Alrubaian M, Fortino G. A hybrid deep learning model for
507 efficient intrusion detection in big data environment. Information Sciences. 2020;513:386-96.
508
509

Table

Table 1 Hyper-parameters of RF

Parameter	Description	Value
ntree	Number of decision trees	1–500
mtry	Maximum number of features in each decision tree	1–6
max_depth	Maximum depth of decision tree	auto
max_leaf_node	Maximum number of leaf nodes	auto
min_sample_leaf	Minimum number of samples in newly created leaves	auto
min_sample_split	Minimum number of samples required to split an internal node	1–6

Table 2 Values of soil parameters

Soil type	D_r	ϕ_c	k_p	*Cr*
Sand-O	80	35.1	0.0034	72
Sand-I	40	35.1	0.0034	72
Sand-II	80	30	0.0034	72

Table 3 Optimum value of hyper-parameters of RF based models

Hyper-parameter	Database 1			Database 2		
	a	*b*	*φ*	*a*	*b*	*φ*
ntree	2	25	8	3	16	82
mtry	3	5	6	4	6	6
min_sample_split	5	2	5	5	6	2

Table 4 Performance of model trained using the first dataset

RF based model	Database 1				Database 2			
	Training set		Testing set		Training set		Testing set	
	NSE	MAPE	NSE	MAPE	NSE	MAPE	NSE	MAPE
a	0.97	19.4%	0.99	23.7%	0.99	14.6%	0.95	9.5%
b	0.95	16.5%	1.00	10.0%	0.94	19.5%	1.00	4.6%
φ	0.98	4.4%	0.95	1.7%	0.99	2.7%	1.00	1.6%

Figure caption

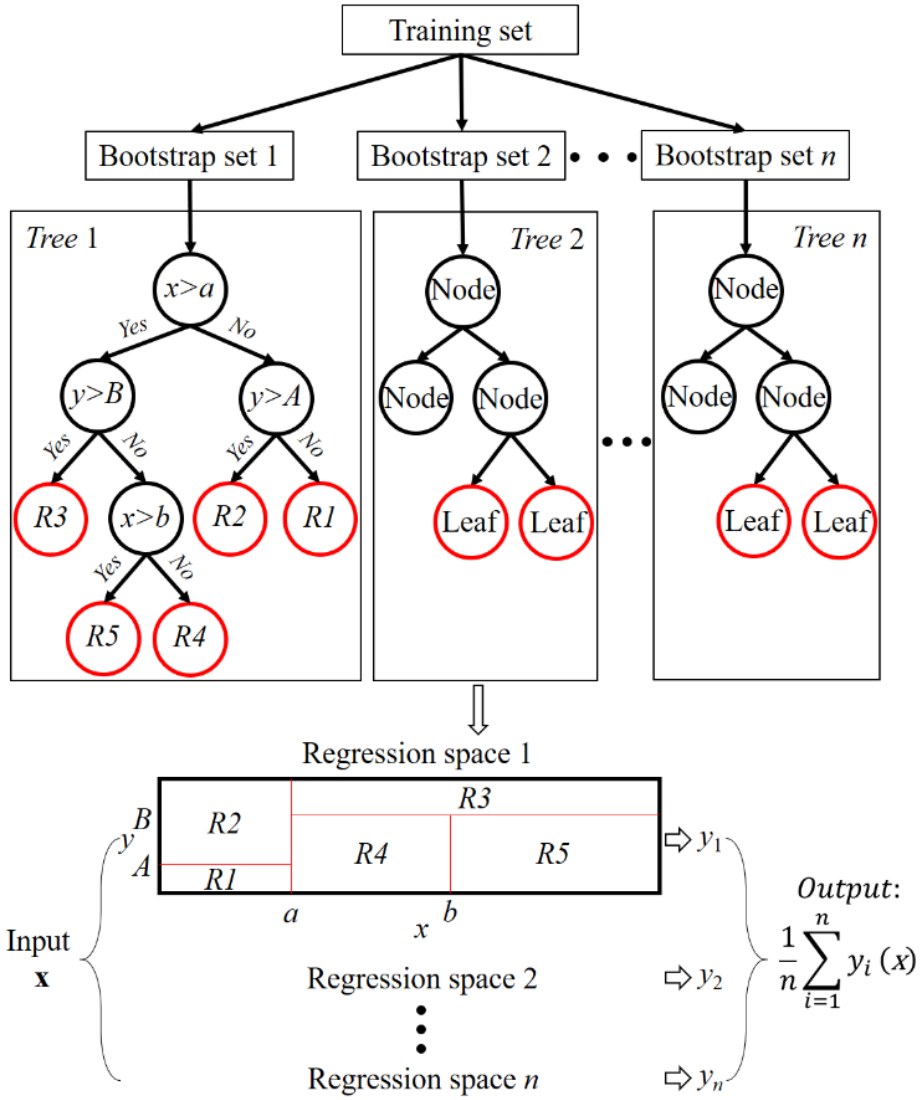

Fig. 1 Schematic view of random forest algorithm

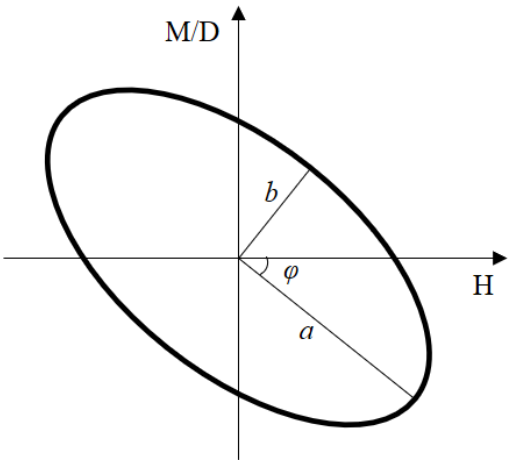

Fig. 2 Definition of parameters describing failure envelope of caisson foundation

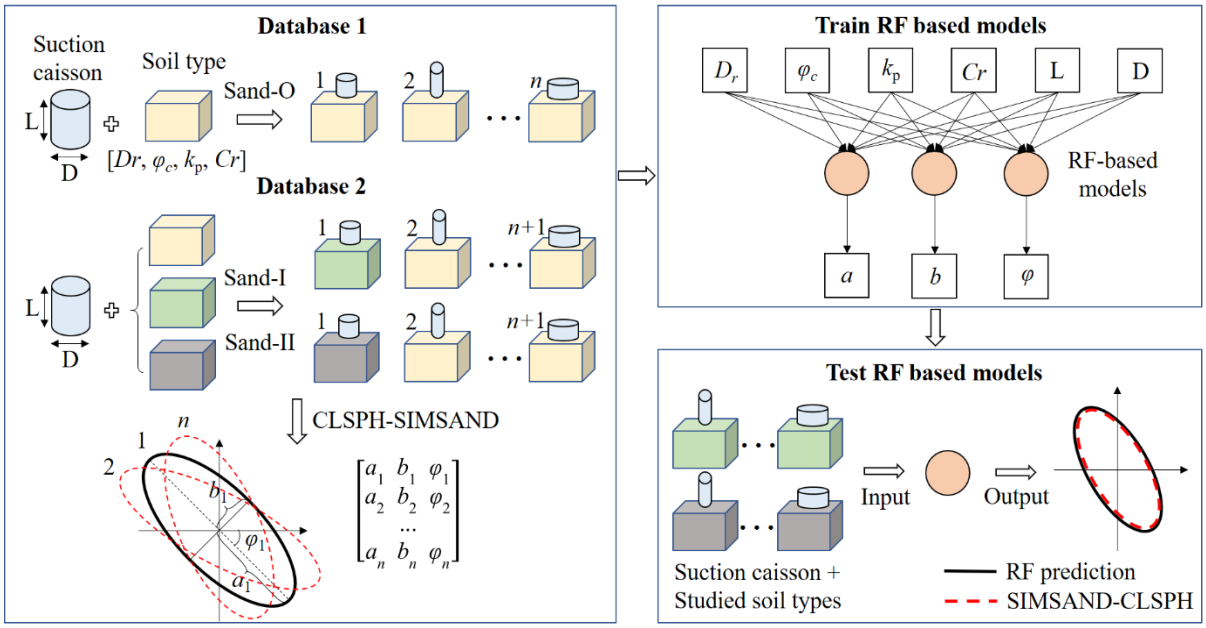

Fig. 3 Flowchart of developing RF based model using CLSPH-SIMSAND simulations

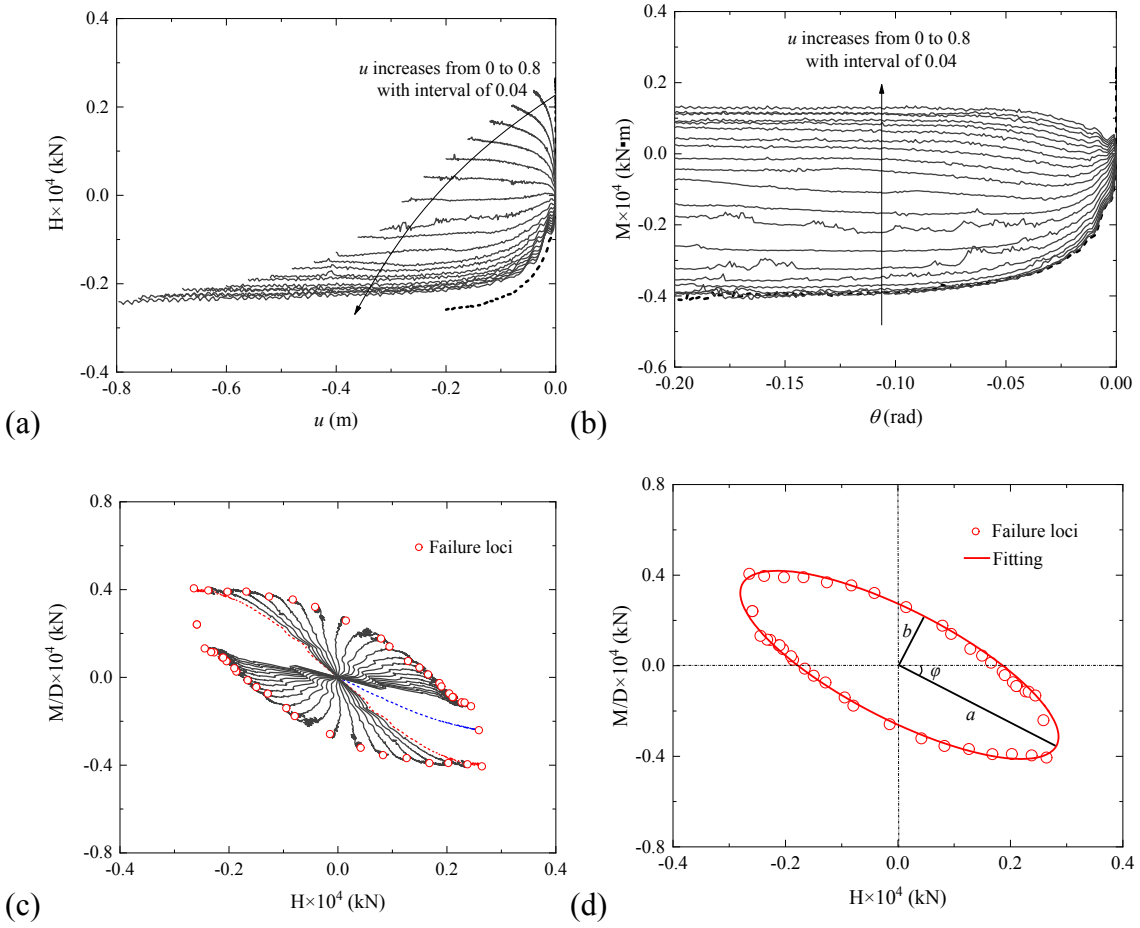

Fig. 4 Results of SPH-SIMSAND numerical modelling for: (a) *u*–H; (b) *θ*–M; (c) H–M/D; (d) failure envelope

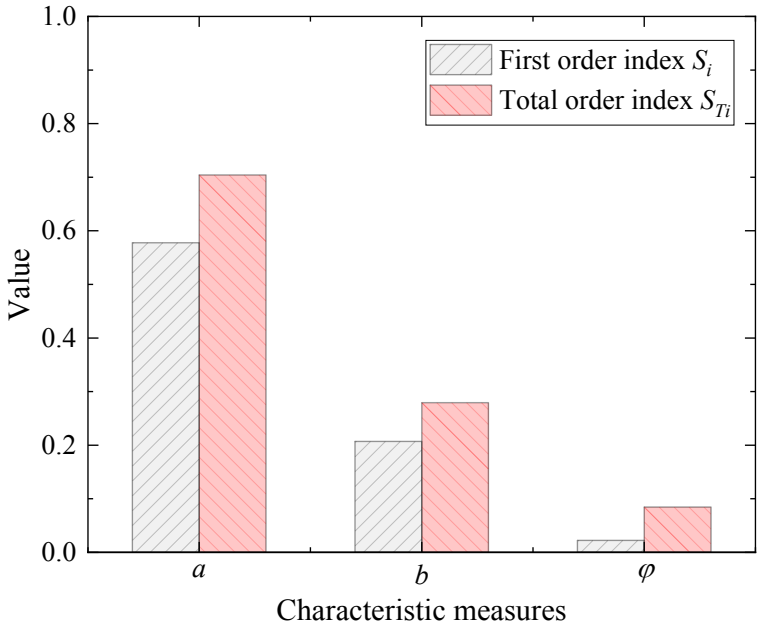

Fig. 5 First and total indices of characteristic measures

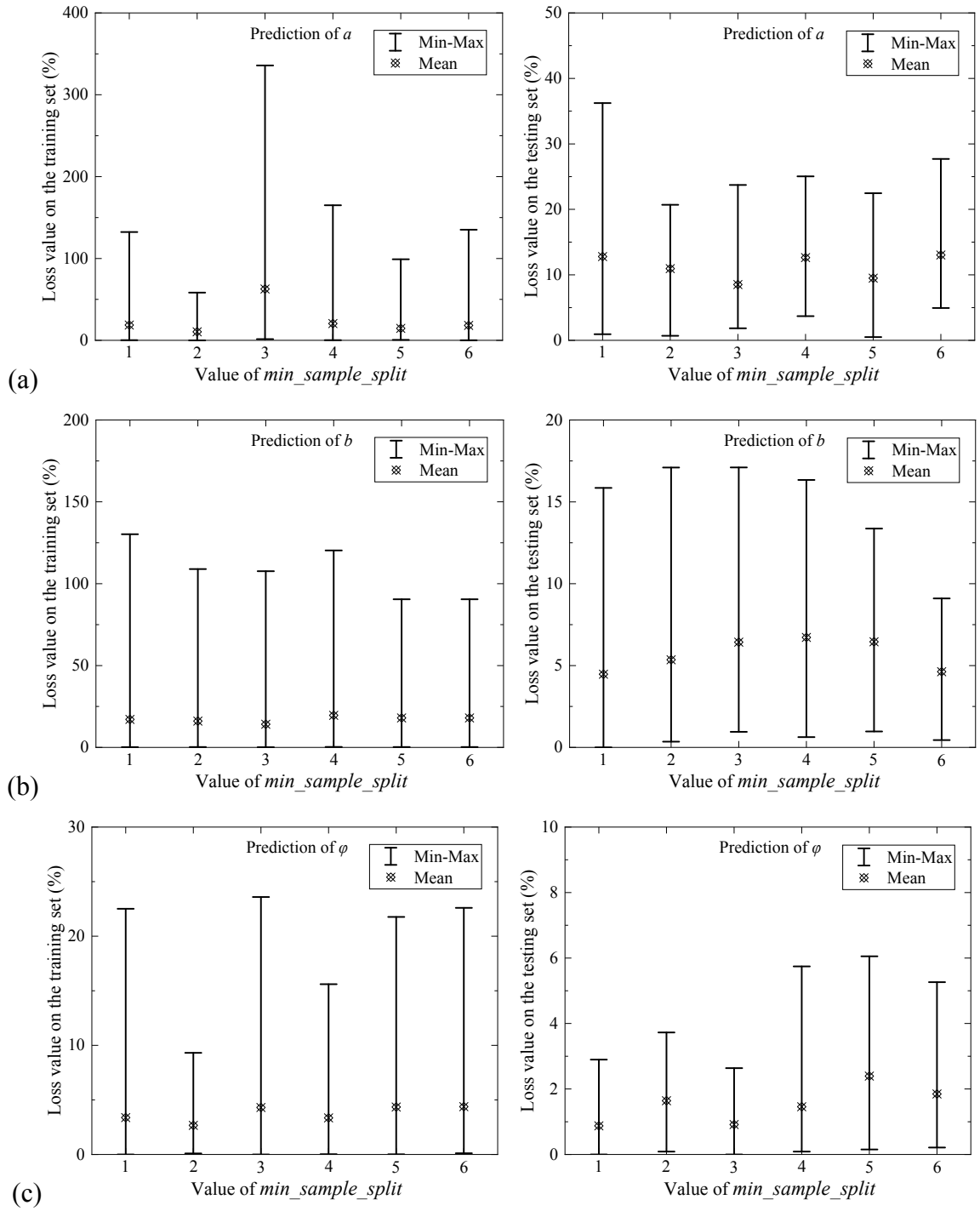

Fig. 6 Effects of *min_sample_split* on the model performance for: (a) *a*; (b) *b*; (c) *φ*

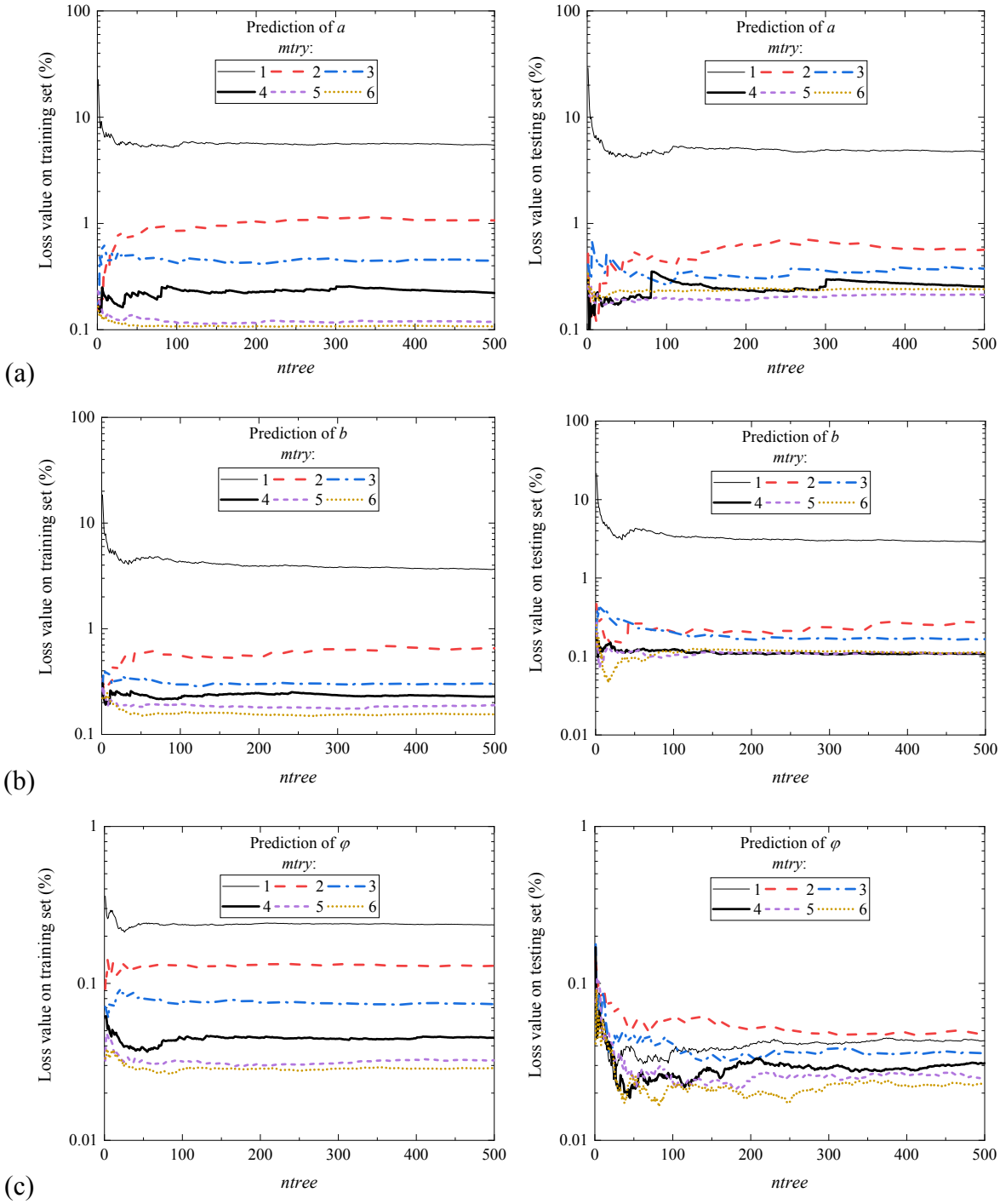

Fig. 7 Effects of *ntree* and *mtry* on the model performance for: (a) *a*; (b) *b*; (c) *φ*

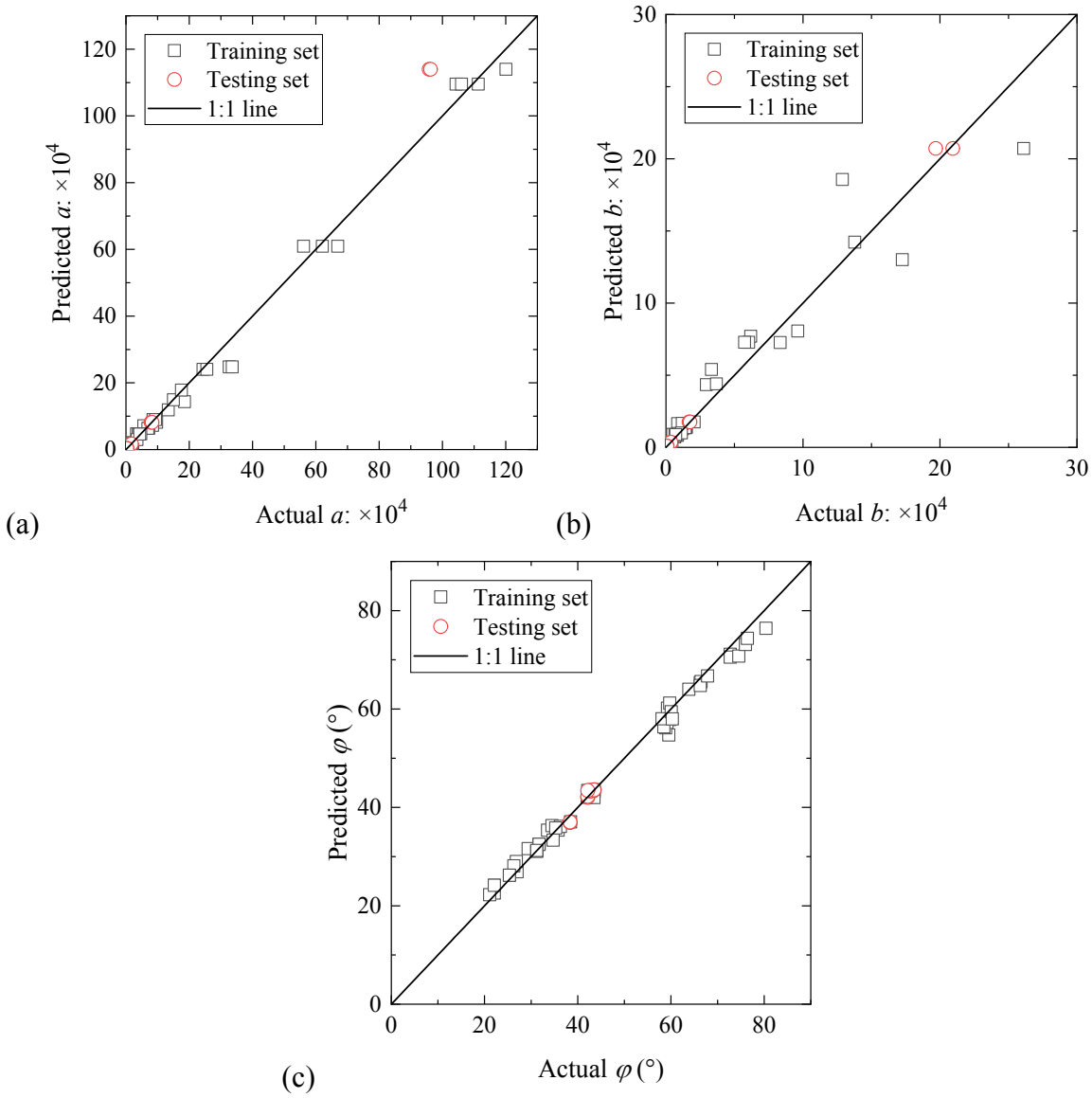

Fig. 8 Comparison between actual and predicted characteristic measures of failure envelope for: (a) a; (b) b; (c) φ

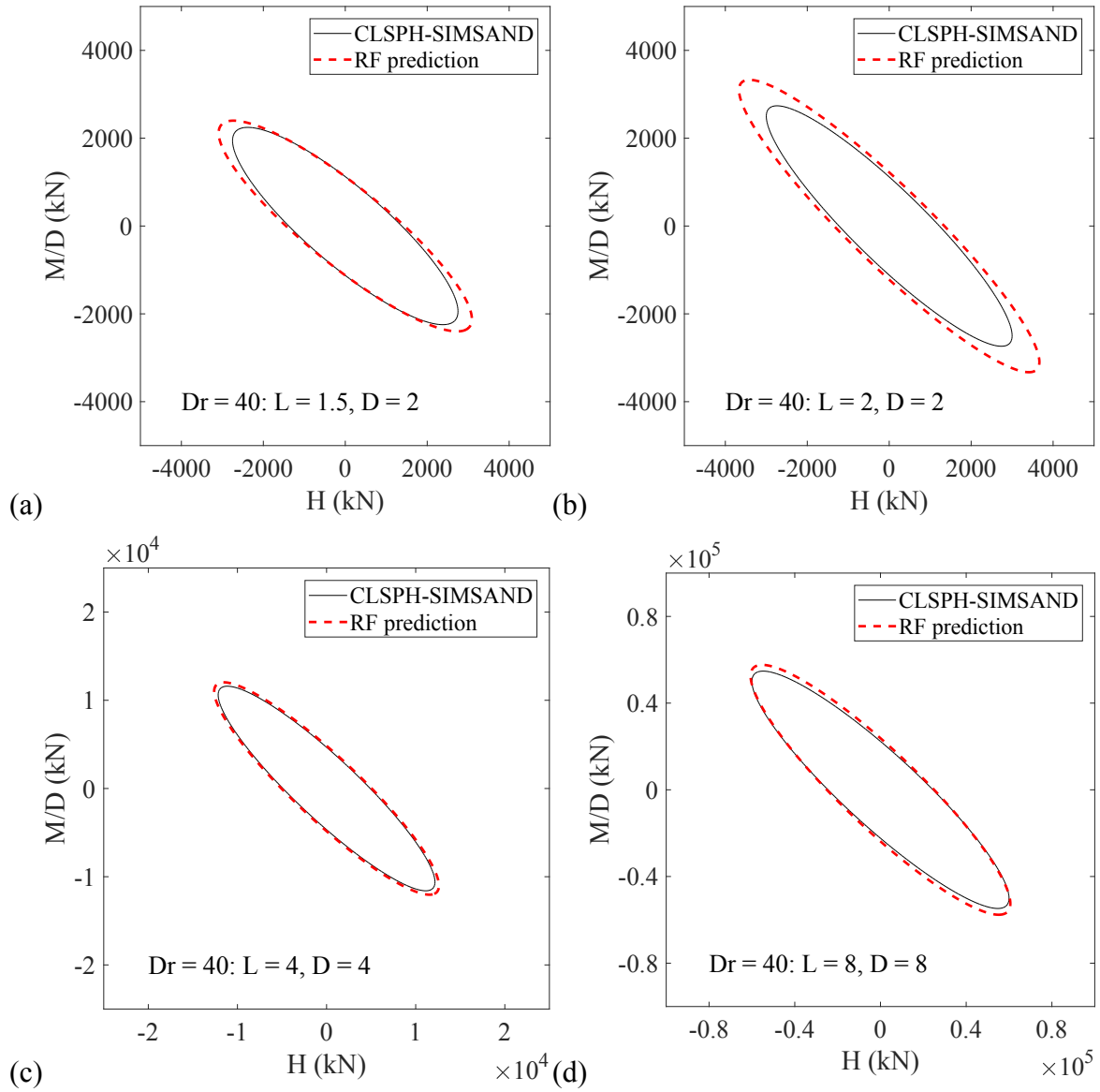

Fig. 9 Comparison between actual and predicted failure envelopes for caisson foundations in sand with D_r = 40: (a) L=1.5 m and D=2 m; (b) L=2 m and D=2 m; (c) L=4 m and D=4 m; (d) L=8 m and D=8 m

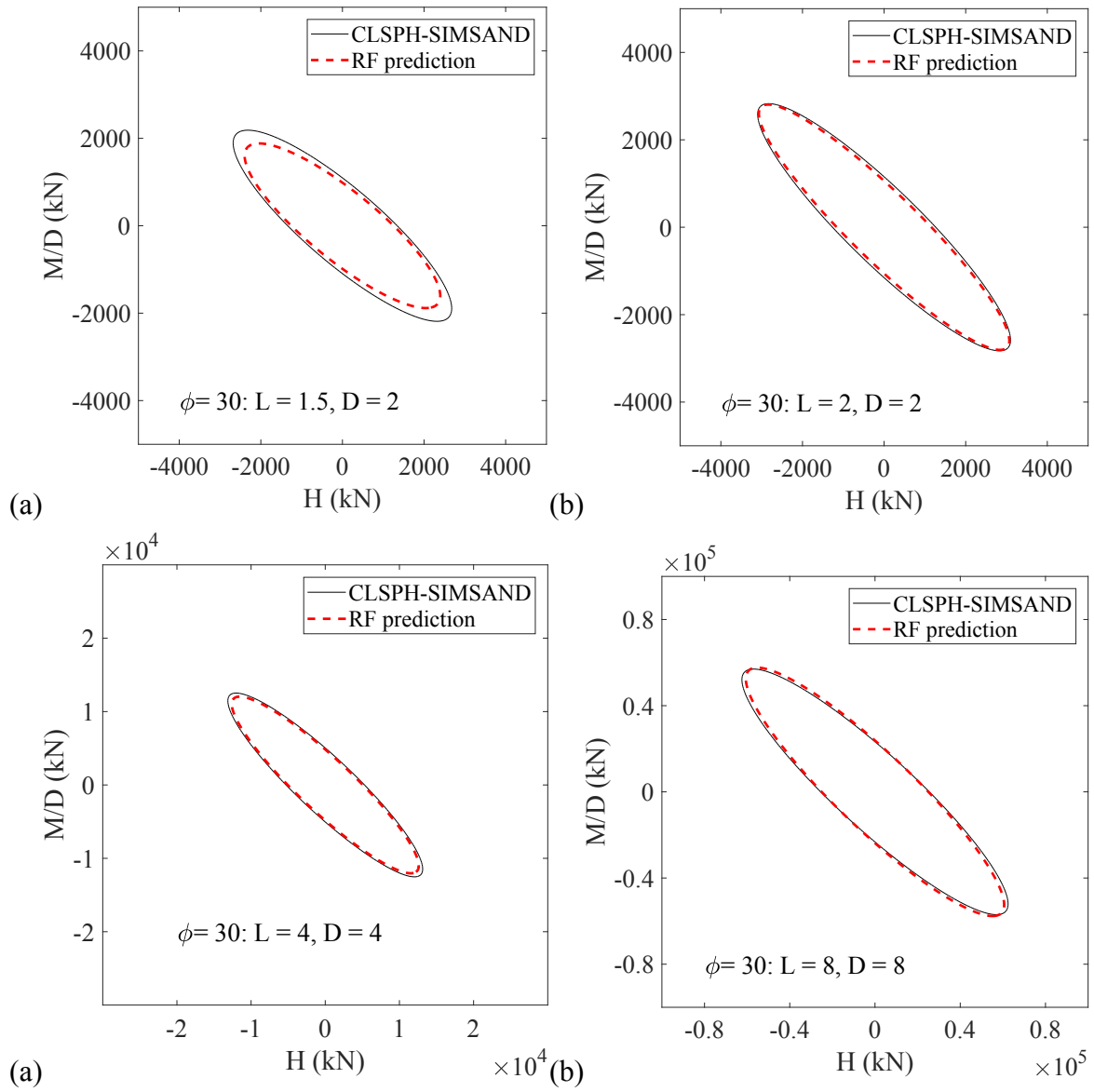

Fig. 10 Comparison between actual and predicted failure envelopes for caisson foundations in sand with $\phi = 30°$: (a) L=1.5 m and D=2 m; (b) L=2 m and D=2 m; (c) L=4 m and D=4 m; (d) L=8 m and D=8 m