**Model of Storm Surge Maximum Water Level Increase in a Coastal Area Using Ensemble Machine Learning and Explicable Algorithm**

**Kun Sun1, 2, and Jiayi Pan1, 2, 3\***

1School of Geography and Environment, Jiangxi Normal University, Nanchang 330022, China

2Key Laboratory of Poyang Lake Wetland and Watershed Research, Ministry of Education, Nanchang 330022, China

3Institute of Space and Earth Information Science, The Chinese University of Hong Kong, Shatin, Hong Kong, China

Corresponding author: Jiayi Pan ([panj@cuhk.edu.hk)](mailto:panj@cuhk.edu.hk))

Key Points:

* An ensemble learning model has been developed to predict the peak water level rise during storm surges in Hong Kong's coastal regions.
* The ensemble approach outperforms standalone machine learning methods of Random Forest, Gradient Boosting Decision Tree, and XGBoost.
* An interpretability analysis reveals that gale distance (Gale\_Dis), and the nearest wind speed (N\_WS) are the most influential features for storm surge prediction.

Abstract

This study proposes a novel, new ensemble model (NEM) designed to simulate the maximum water level increases caused by storm surges in a frequently cyclone-affected coastal water of Hong Kong, China. The model relies on storm and water level data spanning 1978 to 2022. The NEM amalgamates three machine learning algorithms: Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and XGBoost (XGB), employing a stacking technique for integration. Six parameters, determined using the Random Forest and Recursive Feature Elimination algorithms (RF-RFE), are used as input features for the NEM. These parameters are the nearest wind speed, gale distance, nearest air pressure, minimum distance, maximum pressure drop within 24 hours, and large wind radius. Model assessment results suggest that the NEM exhibits superior performance over RF, GBDT, and XGB, delivering high stability and precision. It reaches a coefficient of determination (*R*2) up to 0.95 and a mean absolute error (MAE) that fluctuates between 0.08 and 0.20 m for the test dataset. An interpretability analysis conducted using the SHapley Additive exPlanations (SHAP) method shows that gale distance and nearest wind speed are the most significant features for predicting peak water level increases during storm surges. The results of this study could provide practical implications for predictive models concerning storm surges. These findings present essential tools for the mitigation of coastal disasters and the improvement of marine disaster warning systems.

**Plain Language Summary**

This research introduces an innovative ensemble machine learning model that predicts maximum water level rises due to storm surges in Hong Kong's cyclone-prone coastlines. The model blends three machine learning algorithms, namely Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and XGBoost (XGB), using storm and surge data from 1978-2022. It employs six parameters, including nearest wind speed and gale distance, providing superior performance with high stability and precision. Using SHapley Additive exPlanations (SHAP) analysis, gale distance and wind speed emerged as key predictors. The NEM could enhance storm surge models, aiding in coastal disaster mitigation and improving maritime warning systems.

1 Introduction

Storm surge can be defined as an abnormal elevation and recession of seawater primarily triggered by intense winds and sudden atmospheric pressure shifts associated with meteorological disturbances. These disturbances include, but are not limited to, tropical cyclones, extratropical cyclones, explosive cyclones, and various other weather systems. These events can lead to tidal levels greatly exceeding the norm. The Northwest Pacific Ocean (NPO) represents a hotspot for such events, as it experiences nearly a third of the world's cyclonic activities. Hong Kong SAR (HK) of China, situated within influences of the NPO and the Eastern Asian monsoon, frequently encounters active tropical cyclones and related storm surges during the summer and autumn seasons. Between 1954 and 2022, HK witnessed over 370 storm surge incidents precipitated by tropical cyclones, almost 25% of which saw water levels rising more than 1.0 m (Sajjad & Chan, 2020). Given these considerations, the accurate prediction of storm surge water levels becomes of paramount importance for maritime disaster forecasting and the mitigation of economic damage.

Numerical forecasting techniques, grounded in hydrodynamics and air-sea interaction mechanisms, are currently the cornerstone for predicting coastal storm surges. Notable among the commonly employed numerical models are the Sea, Lake, and Overland Surges from Hurricanes (SLOSH) model (Jelesnianski et al. 1992), the Advanced Circulation (ADCIRC) Model (Luettich et al., 1992; Luettich & Westerink, 2004), the Finite-Volume Coastal Ocean Model (FVCOM) (Chen et al., 2003; Chen, 2006), and the Semi-implicit Eulerian-Lagrangian Finite-Element model for cross-scale ocean circulation (SELFE) model (Zhang & Baptista, 2008). The application of these numerical models to storm surge prediction has demonstrated impressive results. Nevertheless, the inherent simplification of oceanic and atmospheric dynamics within these numerical models can induce significant prediction errors. These errors become especially prominent in regions featuring intricate coastlines and complex morphological conditions (Kohno et al., 2018).

With the advent of artificial intelligence, machine (deep) learning methodologies have emerged as viable approaches to storm surge model and predictions, employing storm-related parameters as input factors (Hashemi et al., 2016). For example, Lee (2009) proposed an innovative method for forecasting storm surges and tidal anomalies using the back-propagation neural network (BPN), the results show that the Root Mean Square Error (RMSE ) of water level ranges from 0.05 to 0.43 m. Oliveira (2009) proposed a neural network model (NNM) aimed at predicting variations in coastal sea levels linked to meteorological phenomena. This model underwent testing for water level predictions spanning 6, 12, 18, and 24 hours. Remarkably, the correlation coefficient between the model’s predictions and the observed water levels peaked at an impressive 0.99. Kim (2012) conducted further experimentation using a combination of input data components: sea surface level, sea level pressure, depression rate, wind speed, wind direction, and typhoon position within the neural network, and discovered that enhanced input data, station count, and event frequency improved storm surge prediction accuracy. The accuracy of the model has been further improved, with the highest coefficient of determination reaching ~0.9, and the water level RMSE ranging from 0.10 m to 0.35m. Chen et al. (2012) utilized an artificial neural network (ANN) model, comprising the back-propagation neural network (BPNN) and adaptive neuro-fuzzy inference system (ANFIS) algorithms, to rectify miscalculations of storm surge heights in a two-dimensional hydrodynamic model for typhoon events.

Ayyad et al. (2022a; 2022b) synergistically integrated numerical simulation techniques and artificial neural networks (ANN) to simulate and produce an extensive dataset of synthetic tropical cyclone data, which was subsequently utilized for training, validating, and testing their developed neural network model. Utilizing ANN in conjunction with ADCIRC to forecast storm surge height and conducting comparative analyses of the results, they discovered that the correlation coefficient between the peak storm surge height predictions made by ADCIRC and ANN was impressively high, with a minimum value of 0.975. Additionally, they reported a remarkably low maximum root mean square error (RMSE) of just 0.16 m, underscoring the potential and substantial benefits of employing machine (deep) learning techniques in predicting coastal disasters.

Complementing these findings, other neural network models, such as Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Nonlinear Autoregressive Network with Exogenous Inputs (NARX), have also been applied successfully to the task of storm surge height prediction, as demonstrated by various studies including those by Xie et al. (2023), Wei & Nguyen (2022), and Li et al. (2023), which all reported promising results.

In recent years, more sophisticated machine learning algorithms have been developed and applied to storm surge prediction. Al Kajbaf and Bensi (2020) put forth a comprehensive framework to compare and evaluate the performance of different models in storm surge prediction. This comparison spanned models based on Artificial Neural Networks (ANN), Gaussian Process Regression (GPR), and Support Vector Regression (SVR). The error of ANN, GPR and SVR for the coastal waters is 0.74, 0.87, and 2.49 m, respectively. Lee et al. (2021) introduced a novel alternative modeling approach, namely the one-dimensional Convolutional Neural Network model (C1PKNet), for rapid peak storm surge prediction in vast coastal areas, the coefficient of determination ranges in ~0.9. Utilizing Support Vector Regression and AdaBoost Random Forest in conjunction with numerical simulations, robust storm surge height prediction models were developed, yielding *R2* values between 0.7 and 0.9, and RMSEs from 0.03 to 0.1 meters (Ayyad et al., 2023).

The integration of machine learning models into the field of storm surge prediction has seen a marked increase in popularity, showcasing their ability to deliver considerable predictive accuracy while substantially reducing the time required for computations, in comparison to traditional numerical simulation methods grounded in physical principles. Nonetheless, it is crucial to note that machine learning models tend to underestimate predictions in extreme scenarios. Moreover, the complexity of multi-scale nonlinear interactions associated with storm surges, such as those involving tide-river, tide-wave, and wave-current dynamics, presents substantial challenges for accurate prediction.

The accuracy of storm surge prediction hinges significantly on the chosen machine learning models. Given the constraints of a single machine learning algorithm, numerous studies have validated that an ensemble model is typically more accurate than a singular machine learning model (Alelyani, 2021; Biau et al., 2019; Chen et al., 2020; Cheng et al., 2019; Liang et al., 2020; Mienye & Sun, 2022a). The enhanced accuracy is primarily attributable to the ensemble methods' capacity to constrain the variance and bias error that are typically associated with individual machine learning models. While machine learning boasts exceptional predictive capabilities, the complexity inherent to these models can mask the relationship between input features and output results, resulting in what's often referred to as a "black box" effect (Ribeiro et al., 2016). The "black box" nature of machine learning, wherein the internal workings of the models are not fully transparent or interpretable, remains a significant hurdle to overcome, necessitating further research and development to enhance the reliability and understandability of these models in critical applications such as coastal disaster prediction (Qin et al., 2023).

Thus, this study proposes the application of an ensemble learning algorithm in storm surge modeling. This new ensemble model (NEM) trains multiple base learners and amalgamates their predictions, thereby providing improved performance and generalization compared to a single base learner. In essence, the use of ensemble methods in storm surge prediction enhances model performance and provides a more thorough understanding of the complex nature of storm surges. To address the lack of interpretability of typical machine learning models, this study also establishes an explanation model of the storm surge with the SHapley Additive exPlanations (SHAP) method (Cha et al., 2021; Lundberg & Lee, 2017; Shapley, 1997). These models are trained and evaluated using storm surge and typhoon data sourced from the Hong Kong Observatory (HKO) and the KITAMOTO laboratory in Japan. The SHAP model allows simulating the new ensemble model developed in this study and provides a detailed understanding of the role and contribution of each input feature to the overall prediction. This process demystifies the decision-making mechanism of the NEM, making it more transparent and interpretable.

The remainder of this paper is structured as follows: Section 2 outlines the study area and the data employed in our research. Section 3 delineates the methodology adopted for the study. Our findings are analyzed in Section 4, followed by a discussion of these results in Section 5. Finally, Section 6 offers conclusions drawn from this study.

2 Study Area and Data

2.1 Study area

This study focuses on the Hong Kong Special Administrative Region of China (22°08 ' N-22 °35' N, 113°49 ' E-114 °31' E), situated south of Guangdong Province, west of the Pearl River Estuary (Figure 1). Within this context, four tide gauges located at Quarry Bay (QB), Tai Po Kau (TPK), Tsim Bei Tsui (TBT), and Shek Pik (SP) have been chosen as the study sites for developing our storm surge model (Figure 1). These gauges are strategically distributed across the northeast, southeast, southwest, and northwest regions of Hong Kong.

A map of the world

Description automatically generated

Figure 1 The study area within Hong Kong, with tidal gauge stations highlighted using red dots.

The study assembles a comprehensive dataset of 394 typhoon samples that impacted the study area with the parameters listed in Table 1 including the observed wind speed at the closest point to the tidal gauge station (N\_WS), Observed air pressure at the closest point to the tidal gauge station (N\_PR), the maximum pressure drop within 24 hrs at the observation station (MPD(24h), the largest radius of gale wind (LRGW), etc.

Table 1 The features and their descriptions of the dataset used in this study.

|  |  |
| --- | --- |
| Feature name | Descriptions |
| ID | Typhoon ID |
| Name | Typhoon Name |
| MWS | Maximum wind speed of the typhoon |
| AWS | Average wind speed of the typhoon |
| MCP | Minimum air pressure of the typhoon center |
| ACP | Average at pressure of the typhoon center |
| AS | Average moving speed |
| LRGW | Largest radius of gale wind |
| MPD(24h) | Maximum pressure drop within 24 hours |
| Dis\_Min | Nearest distance of the typhoon centers to the tide gauge station |
| Gale\_Dis | Distance difference between LRGW and Dis\_Min |
| N\_WS | Observed wind speed nearest to the tidal gauge station |
| N\_PR | Observed air pressure nearest to the tidal gauge station |
| WF | Wind momentum flux of the typhoon |
| ACE | Accumulated cyclone energy of the typhoon |
| PDI | Typhoon power dissipation index |
| TA | Typhoon azimuth parameter relative to the observation location |

Figure 2a displays the maximum water level surge distribution, highlighting 133 events with surges under 0.5 m, 193 events from 0.5 to 1 m, 56 events from 1 to 1.5 m, and 12 events over 1.5 m. Figures 2a-d depict histograms for storm parameters including the maximum pressure drop within 24 hours (MPD(24h)), nearest wind speed (N\_WS), and nearest air pressure (N\_PR), showcasing their variability with ranges of 5-70hPa, 20-100 knots, and 930-1010hPa, and averages of 30hPa, 53 knots, and 980hPa, respectively. These data represent a broad range of surge levels and storm conditions, from tropical depressions to super typhoons. More statistical information about these features is listed in Table 2.

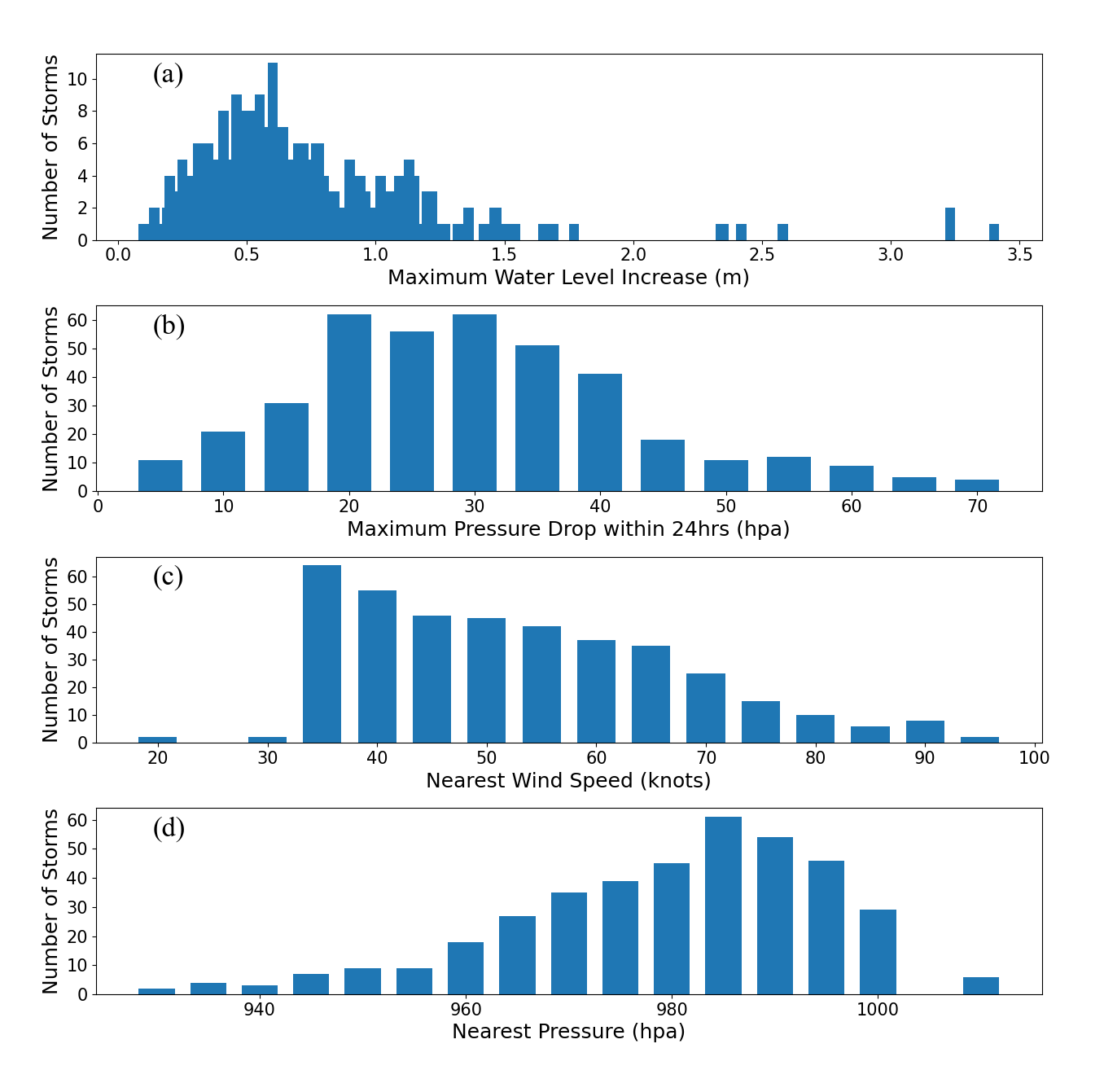


Figure 2 Histograms of some important feature parameters of the maximum water level increase (a) the maximum pressure drop within a 24-hour (b), the nearest wind speed (c), and the nearest air pressure (d).

Table 2 Statistical parameters of maximum water level surge, MPD(24h), N\_WS, and N\_PR

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Features | Mean | Standard Deviation | Maximum | Minimum |
| Maximum water level surge (m) | 0.7 | 0.41 | 3.4 | 0.1 |
| MPD(24h) (hpa) | 30.0 | 13.5 | 70.0 | 5.0 |
| N\_WS (knots) | 52.7 | 14.7 | 95.0 | 20.0 |
| N\_PR (hpa) | 980.0 | 15.3 | 1010.0 | 930.0 |

Table 3 delineates the dataset quantity and temporal coverage for each monitoring station in the study area. In response to this, distinct storm surge models are established and implemented for four separate stations within the study area. Before training the ensemble learning models, the dataset needs to be divided into a training set and a test set. We construct the training set by randomly selecting 80% of the samples at each site (83, 83, 94, 56 cases at Qurry Bay, Tai Po Kau, Tsim Bei Tsui, Shek Pik), while the remaining cases form the test set (19, 19, 24, 14 cases at Qurry Bay, Tai Po Kau, Tsim Bei Tsui, Shek Pik). Between different input factors to eliminate the effects of the dimension of fitting algorithm, the normalized treatment needs to be conducted on the data set, mapping the various factors between [0, 1], given by,

, (1)

where and are the input and normalized values, and are each factor's maximum and minimum values in the data set, respectively.

Table 3 Storm surge cases used in this study

|  |  |  |
| --- | --- | --- |
| Tidal gauge station | Number of cases | Time period |
| Qurry Bay | 102 | 1978 - 2022 |
| Tai Po Kau | 102 | 1978 - 2022 |
| Tsim Bei Tsui | 119 | 1979 - 2022 |
| Shek Pik | 71 | 1998 - 2022 |

3 Methodology

3.1 Ensemble learning model

Ensemble learning leverages the combination of multiple models to enhance performance. By incorporating various regression models, ranging from simple linear regression to more complex machine learning regression models, ensemble learning can produce robust estimation results and superior generalization during the regression process (Schulz et al., 2018; Sagi & Rokach, 2018). In this study, we implement the stacking ensemble method, typically composed of two layers of learners as depicted in Figure 3. The first layer consists of primary learners or base models, and the second layer comprises the meta-learner which combines the base learners' outputs. The fundamental approach involves constructing base learners using the training dataset. Subsequently, the prediction results from all base learners are combined with the true response variables to form a new dataset. A meta-learner is then selected to train and predict based on this new dataset. During the construction of primary learners, the stacking algorithm often employs the bootstrap method for sampling and employs "K-fold cross-validation" for model training. For the amalgamation process, non-linear methods, such as machine learning, are typically used to train the new dataset (Mienye & Sun, 2022b).

A diagram of a company

Description automatically generated

Figure 3 Schematic map of the stacking ensemble method in modelling of maximum increase of storm surge water level.

In this study, three machine learning algorithms, namely Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and XGBoost (XGB), are utilized as the base learners in the ensemble learning model. RF is a bagging ensemble learning algorithm built upon the decision tree algorithm. It employs bootstrap resampling technology and constructs multiple decision tree models from the resampled subsets of the original training set. The final classification or regression result is based on the aggregate results of these decision trees (Breiman, 2001). GBDT combines the gradient boosting algorithm with the decision tree method. It uses boosting technology to construct a powerful learner, with advantages such as non-linearity and no requirement for pre-assumed model structure. Unlike RF, GBDT constructs decision trees iteratively, with each iteration progressing along the direction of the negative gradient of the loss function. Each decision tree learns the residual of the previous tree and accumulates the classification results to provide output. The process halts when the residual is small enough or the number of trees reaches a predetermined limit (Friedman, 2001). XGB is another gradient boosting method, but it incorporates a regularization technique to control overfitting. The primary idea is to grow trees incrementally, performing feature splits to fit the residuals predicted by the previous tree. After training, the scores of all leaf nodes across the trees are summed to obtain the final prediction result (Liang et al., 2020).

The Gaussian Process Regression (GPR) is selected as the meta-learner for this study (Schulz et al., 2018). GPR is a non-parametric regression algorithm that is grounded in the Bayesian framework and utilizes Gaussian process properties. By combining appropriate Gaussian processes, and integrating prior knowledge, the GPR achieves robust predictive and generalization capabilities.

A new ensemble model (NEM) is built upon the selected base learners and the meta-learner. The architecture of this model is presented in Figure 3, with the detailed procedure explained as follows. The initial dataset is split into a training dataset (*D*) and a test dataset (*T*). The training set undergoes 5-fold cross-validation and is segmented into five equal portions. Four of these sections serve as training data, with the remaining part as validation data. For the *xth* fold, base learners are trained on the training data, *Dx* (*x* denotes the fold), and the predictive outputs *Pxt* on the validation data *Dx*, as well as *Qxt* on the test dataset *T*, are obtained. In the training set, the predicted values *Pt = (P1t, P2t, P3t, P4t, P5t)* for validation data after 5-fold cross-validation are compiled, and the corresponding storm surge observation value is *y*t. These results form the training dataset *D't=(Pt, yt)* for the ensemble model, using base learner *ht*. Regarding the test dataset, the predictive outputs of the *T* set after 5-fold cross-validation are *Q1t*, *Q2t*, *Q3t*, *Q4t*, and *Q5t,* and the final prediction result *Qt* of the base learner is their average. If the storm surge observation value is *yt*, then base learner *ht*attains the meta-learning model test data, *T't=(Qt,yt)* via cross-validation. After training and testing the three base learners, we obtain the meta-training data *D'* and test data *T'*. Then, a new ensemble learning model is trained based on the training dataset *D'*, and the final prediction result is produced by predicting the test dataset *T'*.

3.2 SHAP approach

To tackle the issue of explainability in machine learning models employed in this study, we utilize the SHapley Additive exPlanations (SHAP) method. This approach not only visualizes the prediction results but also explains the model's outputs, thereby bolstering the reliability of the results. SHAP, rooted in cooperative game theory, gauges the influence of features on the outcomes by calculating each feature's contribution to the prediction (Lundberg & Lee, 2017). This contribution could either be positive or negative, enhancing or diminishing prediction results, respectively. The absolute value of a feature's contribution mirrors its role in the model, i.e., the larger the absolute contribution value, the greater its importance in the model.

The principle of the SHAP is described as follows. Let be the *i-th* sample, be the *j-th* feature of the *i-th* sample, be the predicted value of the model for this sample, be the baseline (the mean prediction of the model when no input features are given), and be the SHAP value of the *j-th* feature for the *i-th* sample, then SHAP can be written as,

, (2)

where, the is given by,

, (3)

where *F* represents the full set of influencing factors for sample *xi*, *S* denotes any subset of the influencing factors in sample *xi*, *v*(*S*) describes the contribution generated by the joint effect of the influencing factors included in subset *S*, and represents the contribution of influencing factor j to this joint effect. The SHAP value of feature *j*, denoted as *f*(*xj*), is obtained by averaging the contributions of feature *j* for all samples, given by,

, (4)

3.3 Model evaluation metrics

The ensemble learning algorithms' performance is assessed via the coefficient of determination (*R²*), mean squared error (MSE), mean absolute error (MAE), and stability. The coefficient of determination (*R²*) serves as a standard index to gauge the regression model coefficients' goodness of fit; a larger *R²* indicates more accurate model fitting results. The formula for *R²* is as follows,

, (5)

where *y*i represents the storm surge water level observed value, *yi’* is the value fitted by the machine learning algorithm, while and are the average of *y*i and *y*i′, respectively, and *N* represents the volume of data in the dataset, respectively. The root mean squared error (RMSE) is an indicator to evaluate the size of the difference between the fitted result of the regression model and the target value, which is written as,

. (6)

The mean absolute error (MAE) avoids the problem of errors canceling each other out, resulting in an accurate reflection of the magnitude of the actual prediction error, shown as follows,

. (7)

In addition, the mean absolute percentage error (MAPE), also known as mean absolute percentage deviation (MAPD) is derived to reveal prediction relative errors, given by,

. (8)

The model's stability is determined by comparing the accuracy of the model on the training set versus the test set. If the model exhibits high accuracy on the training set but poor accuracy on the test set, it signifies overfitting, indicating poor model stability. Conversely, if the model's accuracy is relatively consistent between the training and test sets, the model is considered stable. However, poor accuracy on both the training and the test set indicates an underfitting phenomenon. A quantitative measure of the model's stability is given by,

， (9)

where *rmsed* and *rmset* represent the RMSE of the model on the training and test sets, respectively.



4 Results, or a descriptive heading about the results

4.1 Feature selection

In this study, due to the variability and interplay between different features, where some positively impact the prediction performance of the model while others, termed as noise, have a negative effect (Alelyani, 2021; Campos et al., 2021), feature selection is conducted on the dataset. This is accomplished using a combination of the Random Forest algorithm (RF) and the Recursive Feature Elimination algorithm (RFE) - the RF-RFE approach based on the Random Forest algorithm. In this process, top significant feature variables are extracted from each variable subset, forming a new feature dataset for model training. By comparing the performance yielded from each subset, the optimal feature variable combination is determined. Applying the RF-RFE process, this study pinpoints six crucial features from the original dataset, constituting the best combination.

Moreover, to enhance the model’s capability in capturing the storm surge dynamics, a feature parameter representing the typhoon azimuth relative to the station location, denoted as TA, was incorporated. This addition aims to address the rightward bias observed in the typhoon’s impact on upper ocean dynamic processes (Price, 1981). The TA value is defined such that a value of 0 indicates the typhoon is situated to the west of the observation station, while a value of 1 signifies the typhoon's position to the east of the station. To ascertain the impact of TA on the machine learning-based prediction of maximum storm surge, we established two distinct feature sets for evaluation. The first set, referred to as F1, comprises the six features outlined in Table 4. The second set, F2, extends F1 by including the TA feature.

Table 4 Feature ranking based on the RF-RFE

|  |  |  |
| --- | --- | --- |
| Rank | Feature | Score |
| 1 | N\_WS (Nearest wind speed) | 0.195 |
| 2 | Gale\_Dis (Distance difference between LRGW and Dis\_Min) | 0.191 |
| 3 | N\_PR (Nearest air pressure) | 0.123 |
| 4 | Dis\_Min (Typhoon nearest distance) | 0.114 |
| 5 | MPD (24h)  (Maximum pressure drop within 24 hrs) | 0.069 |
| 6 | LRGW (Largest radius of gale wind) | 0.056 |

4.2 Performance of the ensemble learning model

Table 5 presents the *R*2 values for the three base learners and the new ensemble model (NEM), with all *R*2 values exceeding 0.96 at all sites for the four algorithms. The NEM outperforms the rest based on the *R*2 values, achieving the highest *R*2 of 0.99 for all the stations. The Random Forest (RF) model exhibits relatively lower *R*2 values. This suggests that the ensemble model may be assimilating the strengths of the three base algorithms to enhance the model's accuracy.

Table 5 *R2* values of the RF, GBDT, XGB, and NEM models based on F1and F2 in the training dataset

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | RF | | GBDT | | XGB | | NEM | |
|  | F1 | F2 | F1 | F2 | F1 | F2 | F1 | F2 |
| Quarry Bay (QB) | 0.96 | 0.97 | 0.98 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| Tai Po Kau (TPK) | 0.98 | 0.98 | 0.98 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| Tsim Bei Tsui (TBT) | 0.98 | 0.98 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| Shek Pik (SP) | 0.97 | 0.98 | 0.98 | 0.98 | 0.97 | 0.98 | 0.99 | 0.99 |

Table 6 presents the performance metrics of the models for the testing dataset. Utilizing the F1 feature set, at Quarry Bay (QB) station, the *R2* values for Random Forest (RF), Gradient Boosting Decision Tree (GBDT), XGBoost (XGB), and New Ensemble Model (NEM) are 0.79, 0.83, 0.87, and 0.89, respectively. Across all gauge stations, which include Quarry Bay, Tai Po Kau, Tsim Bei Tsui, and Shek Pik, the NEM consistently outperforms, achieving the highest *R2* values. The most notable improvement is observed at the TBT station, where RF’s *R2* of 0.80 escalates to 0.93 with NEM. This trend of NEM surpassing the base learners in terms of *R2* values holds true for the testing dataset as well. Switching to the F2 feature combination, which includes TA, all models exhibit varying degrees of enhancement. At the Shek Pik (SP) station, for instance, *R2* values rise from the range of 0.69-0.74 to 0.73-0.77. For NEM specifically, the *R2* values at the four stations increase from 0.89, 0.89, 0.93, and 0.74 to 0.92, 0.93, 0.95, and 0.77, respectively, under the F2 scenario. These results clearly indicate the substantial impact of TA on prediction accuracy, validating its effectiveness in enhancing the model’s performance.

These algorithms register varying *R*2 values across different sites due to the variations in sample sizes. There is a positive correlation between the *R*2 and the size of the sample data; a better model fit is obtained at the TBT station with more training and testing data, whereas the opposite is observed at the SP station.

Table 6 *R*2 values of the RF, GBDT, XGB, and NEM models based on F1and F2 in the test dataset

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | RF | | GBDT | | XGB | | NEM | |
|  | F1 | F2 | F1 | F2 | F1 | F2 | F1 | F2 |
| Quarry Bay (QB) | 0.79 | 0.82 | 0.83 | 0.85 | 0.87 | 0.89 | 0.89 | 0.92 |
| Tai Po Kau (TPK) | 0.86 | 0.88 | 0.86 | 0.88 | 0.87 | 0.90 | 0.89 | 0.93 |
| Tsim Bei Tsui (TBT) | 0.80 | 0.83 | 0.87 | 0.89 | 0.89 | 0.92 | 0.93 | 0.95 |
| Shek Pik (SP) | 0.69 | 0.73 | 0.71 | 0.75 | 0.71 | 0.74 | 0.74 | 0.77 |

Figure 4 showcases the predictive capabilities of the New Ensemble Model (NEM) when employing two different sets of features, F1 and F2. It is evident from the visualization that F2 outperforms F1, despite both feature combinations following generally similar trends. Nonetheless, F1 tends to significantly underestimate extreme high and low values, whereas F2, which compensates for the typhoon's rightward bias in its effects (Price, 1981), provides predictions that are considerably closer to the true values. Consequently, the F2 feature grouping mode is chosen for uniform adoption in the subsequent sections of the study, ensuring more accurate and reliable predictions.

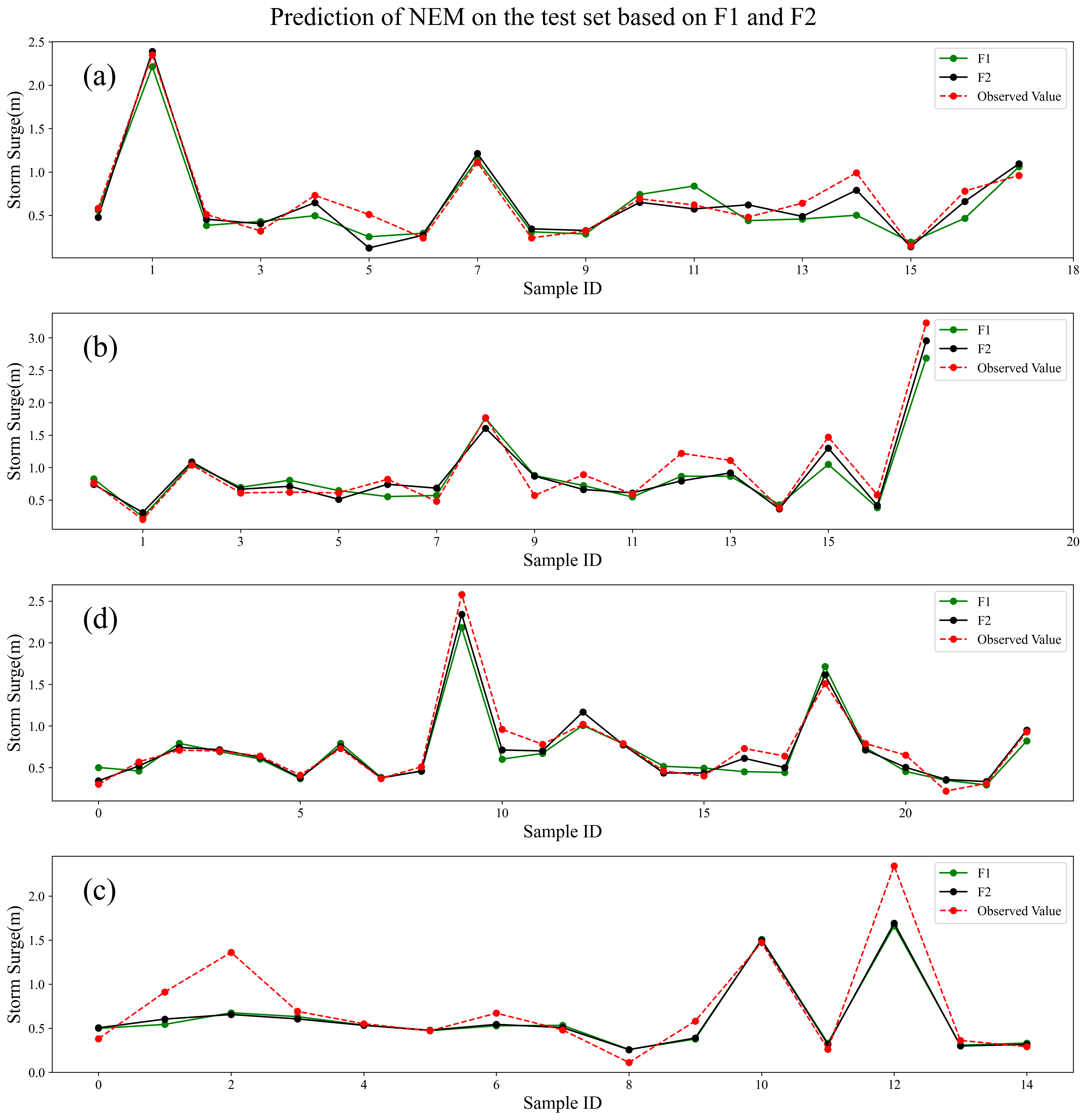


Figure 4: Prediction results from the New Ensemble Model (NEM) at Quarry Bay (QB) (a), Tai Po Kau (TPK) (b), Tsim Bei Tsui (TBT) (c), and Shek Pik (SP) (d) stations using feature combinations F1 (green) and F2 (black), alongside the observed values (red).

To better understand the accuracy of the model predictions, we created scatter plots showcasing the predicted maximum water level increase from the four models versus the observed values at QB, TPK, SP, and TBT stations. These are displayed in Figures 5 to 8 respectively. From these figures, it can be observed that all data points are closely grouped around the 1:1 line. At the QB, TPK, and TBT stations, for water level increases less than 1.5 meters, the discrepancies between the predicted and observed values decrease consistently from the RF model through the GBDT to the XGB model. The XGB model outperforms both the RF and GBDT models in the lower range of water level increases, specifically less than 1.0 meter. However, for water level increases above 1.0 meter, the XGB model does not display a significant advantage over the RF and GBDT models. The incorporation of the three base learner algorithms into the New Ensemble Model (NEM) results in superior accuracy compared to the individual base learner models. This demonstrates the increased efficacy of the NEM over a single machine learning algorithm. Even though there are substantial deviations at the TPK station, Figure 6 illustrates that the NEM is able to effectively reduce prediction errors for the maximum water level increase, particularly in the lower range of less than 1.0 meter.

The performance of each model in terms of the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) on the testing dataset at each station is as follows: For the Random Forest (RF), the RMSE values are 0.20, 0.23, 0.19, and 0.29 m, while the MAE values are 0.17, 0.18, 0.16, and 0.18 m. The Gradient Boosting Decision Tree (GBDT) improves upon the RF with RMSE values of 0.18, 0.23, 0.16, and 0.28 m, and MAE values of 0.14, 0.17, 0.12, and 0.18 m. The XGBoost (XGB), among the three base learners, demonstrates the best overall performance, with RMSE values of 0.16, 0.22, 0.13, and 0.29 m, and MAE values of 0.13, 0.18, 0.10, and 0.20 m. Finally, the New Ensemble Model (NEM), surpasses the individual base learners, providing the lowest error values. The RMSE values for NEM are 0.13, 0.18, 0.10, and 0.27 m at each station, with MAE values of 0.10, 0.15, 0.08, and 0.17 m. This clearly indicates that through ensemble learning, the NEM model achieves the highest accuracy among the four models.

In terms of relative error, the Random Forest (RF) model exhibits the highest Mean Absolute Percentage Error (MAPE) values at QB, TPK, TBT, and SP stations, with corresponding values of 39%, 31%, 27%, and 35%. The Gradient Boosting Decision Tree (GBDT) and XGBoost (XGB) models demonstrate relatively comparable performance. GBDT yields MAPE values of 26%, 23%, 17%, and 28%, while XGB returns values of 25%, 24%, 15%, and 32% at the same stations. Contrastingly, the New Ensemble Model (NEM) achieves comparatively lower MAPE values of 22%, 21%, 12%, and 27%, respectively. Therefore, despite RF recording the highest MAPE values across all four stations and GBDT and XGB showing almost similar performance, the NEM significantly outperforms all these models, boasting the lowest MAPE values across all stations.

To confirm the superior accuracy of the Neural Emulation Model (NEM) over base learners, a new metric, *R2dif* ​, is introduced to quantify the difference in *R2* (coefficient of determination) scores between NEM and base learners (RF, GBDT, and XGBoost) for the test dataset. This metric is defined as,

, (10)

where *R2NEM*​ is the *R2* score of the NEM and *R2BL* is the *R2* score of a base learner. The 95% confidence intervals for *R2dif* ​​ help determine if the improvement in accuracy with NEM is statistically significant. Positive confidence intervals indicate a significant increase in accuracy with the NEM.

The 95% confidence intervals for *R2dif* at the four stations (QB, TPK, TBT, and SP) are presented in Table 7 and range between 0.14-0.23, 0.15-0.25, 0.15-0.21, and 0.06-0.35, respectively, when comparing NEM with RF. For NEM versus other machine learning methods, all confidence intervals are positive. These findings statistically validate that the NEM yields higher *R2* scores than the base learners, signifying a statistically significant improvement in prediction accuracy.

Table 7 The 95% confidence Intervals for *R2dif* ​at Stations QB, TPK, TBT, and SP

|  |  |  |  |
| --- | --- | --- | --- |
| Stations | NEM-RF (m) | NEM-GBDT (m) | NEM-XGB (m) |
| Quarry Bay (QB) | (0.14-0.23) | (0.15-0.21) | (0.17-0.23) |
| Tai Po Kau (TPK) | (0.15-0.25) | (0.16-0.25) | (0.16-0.23) |
| Tsim Bei Tsui (TBT) | (0.15-0.21) | (0.19-0.25) | (0.11-0.17) |
| Shek Pik (SP) | (0.06-0.35) | (0.05-0.38) | (0.1-0.29) |

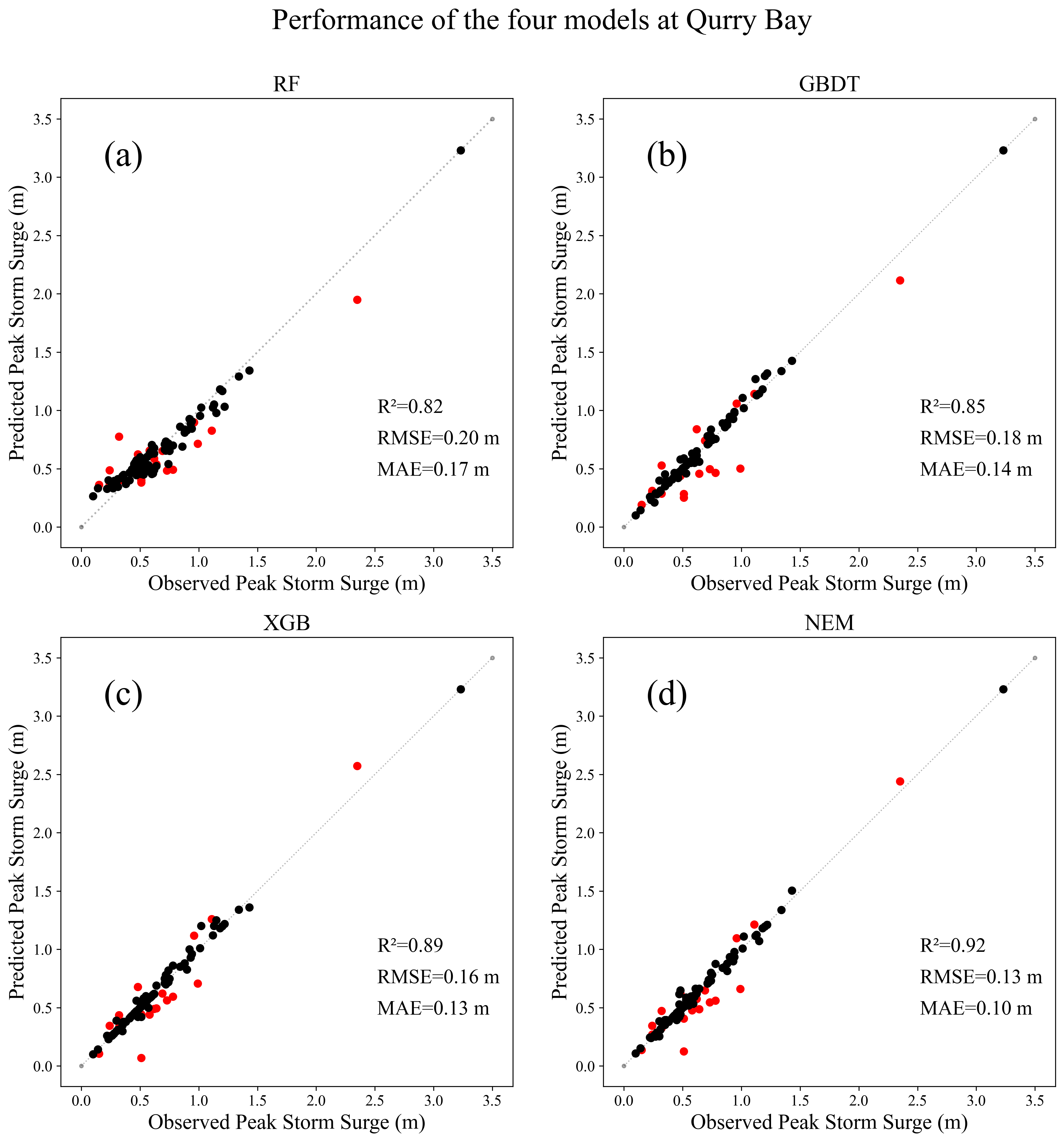


Figure 5 The predicted maximum water level increases versus observed values at the Qurry Bay station for the RF (a), GBDT (b), XGB (c), and NEM (d). The training set and the test set are represented by black and red dots, respectively.

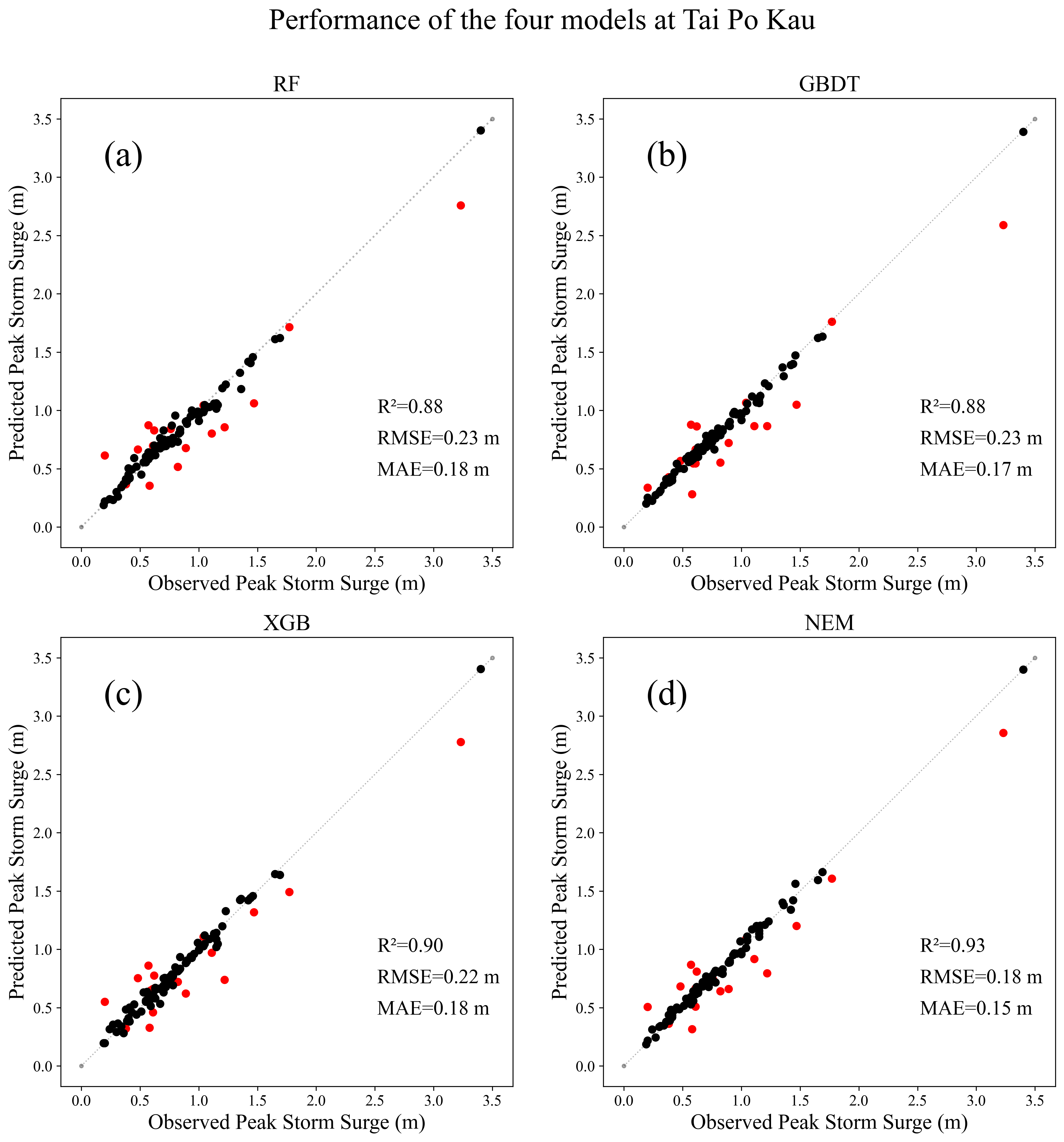


Figure 6 The predicted maximum water level increases versus observed values at the Tao Po Kau station for the RF (a), GBDT (b), XGB (c), and NEM (d). The training set and the test set are represented by black and red dots, respectively.

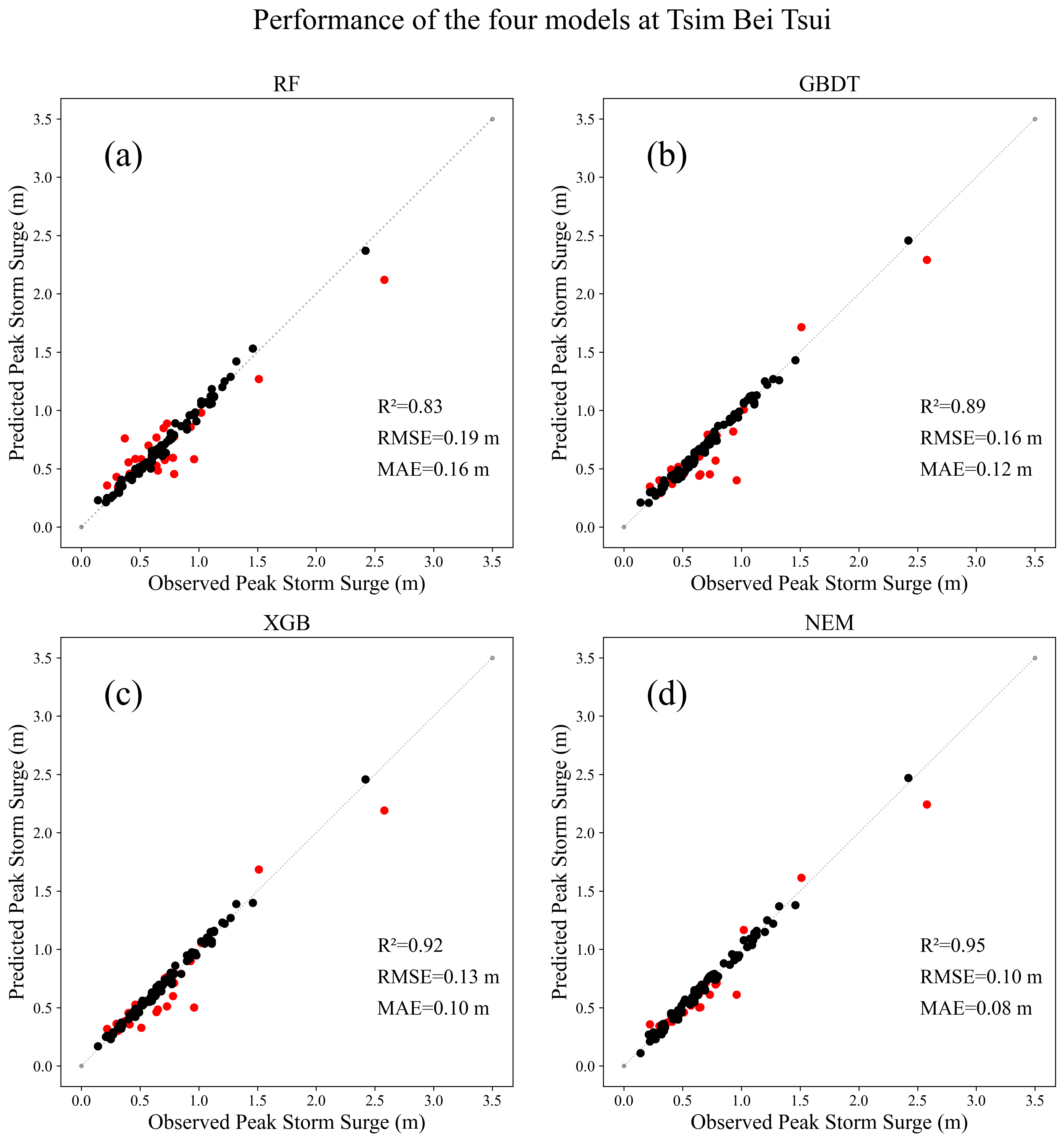


Figure 7 The predicted maximum water level increases versus observed values at the Tsim Bei Tsui station for the RF (a), GBDT (b), XGB (c), and NEM (d). The training set and the test set are represented by black and red dots, respectively.

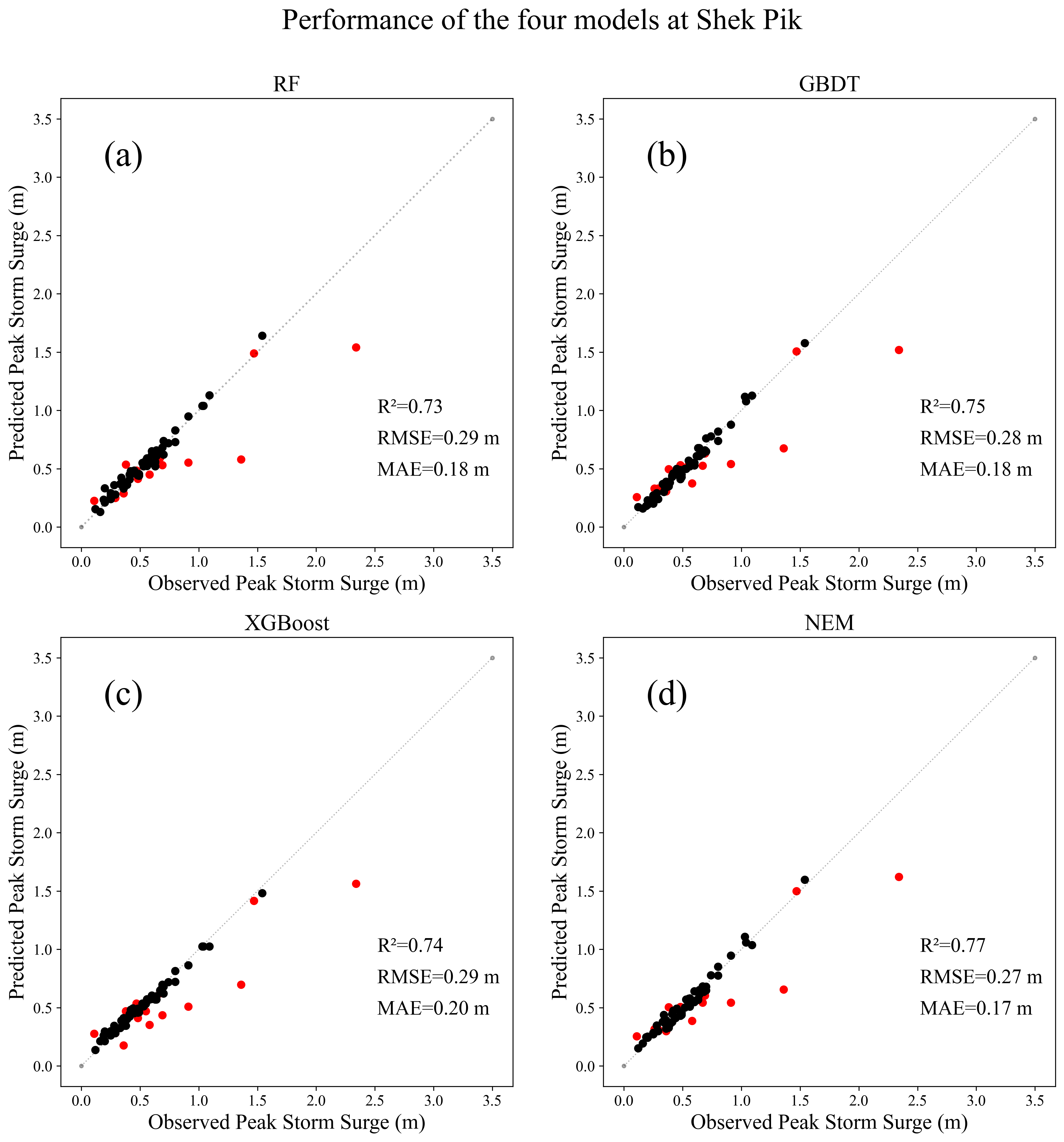


Figure 8 The predicted maximum water level increases versus observed values at the Shek Pik station for the RF (a), GBDT (b), XGB (c), and NEM (d). The training set and the test set are represented by black and red dots, respectively.

Figures 5-8 provides a representation of the Root Mean Square Error (RMSE) generated by each model on both the training and testing datasets at different scales for the four stations. Overall, it is apparent that the RMSE values for the training dataset (*rmsed*) are consistently lower than those for the testing dataset (*rmset*). In the case of the QB station, the stability (*Sta*) values (calculated using Equation 9) for the four models are as follows: 57.14% for the Random Forest (RF), 53.85% for the Gradient Boosting Decision Tree (GBDT), 28.57% for the XGBoost (XGB), and 23.08% for the New Ensemble Model (NEM). This indicates that the XGB and NEM models outperform the RF and GBDT in terms of stability, with NEM showing slightly better performance than XGB. Moreover, GBDT is observed to have better stability than RF. Moving onto the TPK station, the stability values follow a decreasing trend: 56.25% for RF, 47.06% for GBDT, 33.33% for XGB, and 22.22% for NEM. This signifies an improvement in stability as we move from RF to NEM. Similar patterns are observed at the TBT and SP stations. To conclude, when considering performance across all four stations, the NEM outperforms the other models, displaying the highest accuracy and stability.

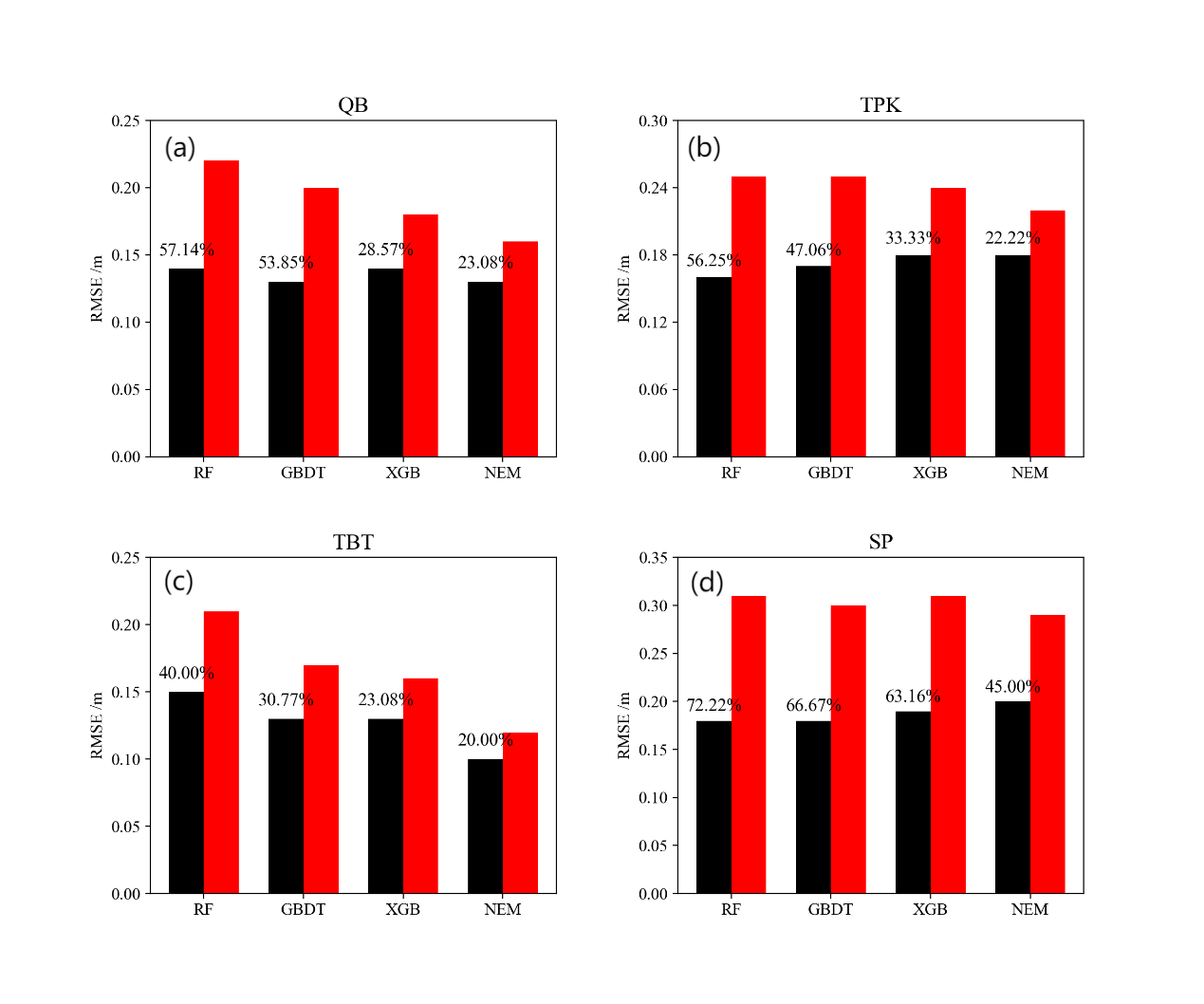


Figure 9 The stability of four models at QB (a), TPK (b), TBT (c), and SP station (d), with the RMSE of training set and the test set represented by black and red bars, respectively.

4.3 Model explicability analysis

In this study, we developed an explanation model to unravel the intricacies of our New Ensemble Model (NEM) using the SHapley Additive exPlanations (SHAP) model. This model employs six distinct features as input parameters, as detailed in Table 4.

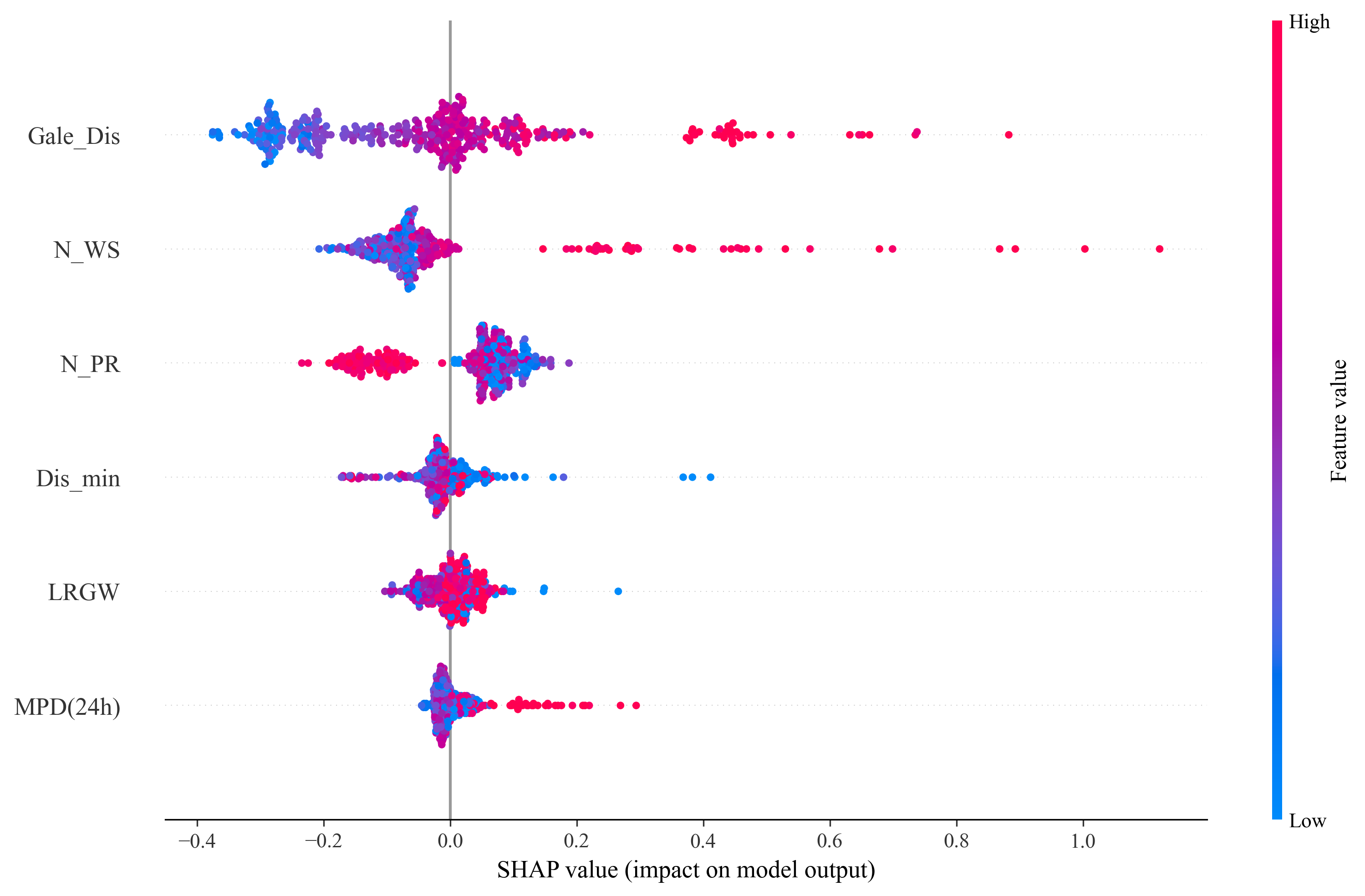


Figure 10 The SHAP values of the NEM model features aggregated from four tide gauge stations.

By summing the SHAP values calculated for the QB, TPK, TBT, and SP stations, we obtained Figure 10. In this figure, the significance of the features diminishes from top to bottom in the order of gale distance (Gale\_Dis), nearest wind speed (N\_WS), nearest air pressure (N\_PR), minimum distance (Dis\_Min), large wind radius (LRGW), and maximum pressure drop within 24 hours (MPD(24h)). The scatter point color-map transitions from blue to red to denote feature importance, with blue being the least and red the most important. Each point represents a SHAP value of a sample, reflecting the contribution of that particular feature to a single prediction. The aggregation of points displays the overall influence - both direction and magnitude - of a given feature on the prediction outcome. Of the six features, the difference between the radius of maximum winds and the nearest distance (Gale\_Dis) stands out as the most significant parameter impacting storm surge elevation predictions. Nearest neighbor wind speed (N\_WS) follows in terms of importance. A positive correlation between their feature values and SHAP values suggests that larger Gale\_Dis and N\_WS values correspond to larger SHAP values. N\_PR, Dis\_Min, LRGW, and MPD(24h) are deemed less critical features. The feature value of N\_PR inversely correlates with its SHAP value. The relationships for Dis\_Min and LRGW are somewhat non-linear. Dis\_Min refers to the smallest distance between the observation station and the nearest recorded typhoon information point. Meanwhile, LRGW denotes the typhoon's large radius of maximum winds, which can somewhat mirror the typhoon's intensity. MPD(24h) signifies the most substantial atmospheric pressure drop within a 24-hour span. Its feature value also correlates positively with the SHAP value.

To better elucidate the influence of each feature on the prediction outcomes and the distribution of their respective SHAP values, we depicted feature dependencies in Figure 11 and showcased the range of SHAP values for each feature along with their next feature. In Figure 11, we present the relationship between the features and their corresponding SHAP values, providing insights into how each feature affects the storm surge prediction. The visualization helps us understand the direction and magnitude of the impact each feature has on the model's predictions.

Figure 11a illustrates the correlation between nearest wind speed (N\_WS) and Gale Distance (Gale\_Dis). It is apparent that blue data points, representing minimal Gale\_Dis values, are densely clustered in areas of high N\_WS, especially when Gale\_Dis is below 0. This suggests that typhoons with smaller wind radii, situated near the observation station, tend to produce lower wind speeds. Yet, as Gale\_Dis exceeds 100, this trend becomes less apparent, highlighting that the factors such as pressure, typhoon trajectory, and alongside proximity influence the observed wind speeds. Conversely,

Figure 11b examines the relationship between Gale\_Dis and N\_WS, revealing no significant linear correlation between these two variables, indicating that the relationship is influenced by other complex factors.

Figure 11c delves into the relationship between N\_PR and N\_WS. Here, we observe that as N\_PR decreases, N\_WS tends to increase. Notably, the SHAP values for N\_PR are evenly distributed, indicating that there is no linear relationship between specific N\_PR values and N\_WS. This, in turn, confirms that NEM's learning is not restricted to particular N\_PR values or ranges. Instead, N\_PR values across the spectrum contribute significantly to the final predictions.

Figure 11d explores the relationship between Dis\_Min and N\_PR. In general, there is no significant linear relationship between the distance from the observation station and atmospheric pressure. However, higher N\_PR values (depicted by the red data points) tend to cluster in the lower Dis\_Min range. This suggests that samples in the immediate vicinity of the study area typically exhibit N\_PR values within the 990-1000hPa range.

Figure 11e investigates the relationship between LRGW and N\_PR, but it does not reveal any significant connection between these two variables. Finally, Figure 11f presents the relationship between MPD (24h) and N\_WS. The color-coding indicates that blue points are concentrated on the left side and transition to red on the right. This pattern suggests that when MPD(24h) has a smaller value, N\_WS tends to be lower, indicating a positive correlation between these variables.

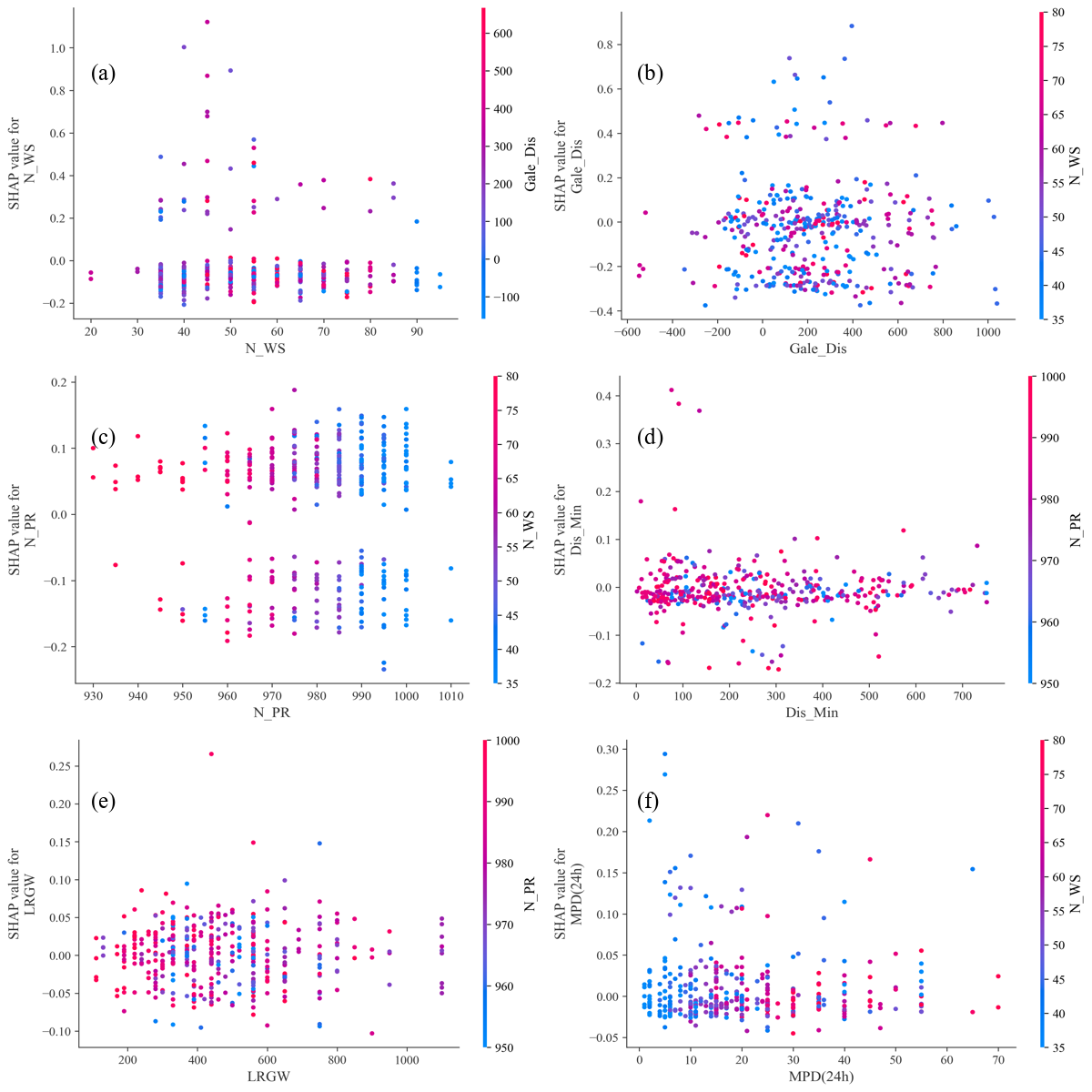


Figure 11. Feature dependencies in the new ensemble model (NEM), generated using SHAP values to show how different features influence the storm surge predictions.

5 Discussions

The application of an ensemble technique in the development of the New Ensemble Model (NEM) for storm surge prediction showcases several advantages over the traditional physical models or numerical simulations. Notably, this method significantly minimizes the time required for forecasting while assuring a desirable level of accuracy.

In our study, three base learners, namely Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and Extreme Gradient Boosting (XGB), were incorporated into NEM. It was observed that NEM outperforms each of these base learners in terms of both accuracy and stability, although we try to tune the hyperparameters leading to the best performance for each of these base learners. This underscores the powerful potential of the ensemble technique in machine learning-based storm surge predictions, and suggests that this approach could be further leveraged in other related studies.

Previous research indicates that wind speed, air pressure, wind direction, and proximity are crucial for predicting storm surges using machine learning techniques (Tiggeloven et al., 2021; Ian et al., 2022). Moreover, to provide a more holistic understanding of storm surges, additional studies have integrated morphological characteristics, oceanic conditions, and temporal patterns into their analyses (Huang et al., 2022; Lockwood et al., 2022; Rajabi-Kiasari et al., 2023, Li et al., 2023). These aspects help to elucidate the dynamics of storm surges, including wind force, oceanic movements, and landform interactions. In this study, we concentrate on forecasting storm surges at individual sites, primarily using typhoon-related variables like nearest wind speed, air pressure, and distance as predictors in our machine learning models to gauge the typhoon's impact on surges. We have also introduced an important aspect, the typhoon azimuth (TA) feature to capture the characteristic rightward bias in a typhoon's trajectory and its implications.

This study reveals a notable influence of TA on the prediction performance of the NEM. We conducted a comparative study between two feature combinations: F1, which does not include TA, and F2, which incorporates the TA feature. The results unequivocally demonstrated enhanced prediction performance when the model was trained with F2 compared to F1, underscoring the significant impact of TA on the model’s predictive capabilities. This indicates that TA indeed has a significant impact on model prediction. For some similar typhoons, for those located on the right-hand-side of observation stations, our model can reveal the rightward bias effects of the typhoons on the upper ocean dynamics.

To demystify the learning process of NEM, we utilized the SHAP method. Our findings indicated that NEM has the capability to extract pertinent information from the given features, subsequently making predictions that are aligned with our general understanding of physical principles. This not only substantiates the effectiveness of NEM but also serves to verify its adherence to the fundamental principles underlying storm surge phenomena.

Contrastingly, many existing studies have either solely employed a single machine learning method, which can result in questionable accuracy, or relied on simulated typhoon data instead of real-life data, potentially leading to less accurate or applicable predictions. Furthermore, our research concentrated exclusively on the peak values of storm surge. However, it's worth noting that the actual storm surge process is a multifaceted physical phenomenon subject to various influencing factors. Therefore, our study, while providing valuable insights, does not encapsulate the complete dynamics of storm surges. Additionally, due to limited data availability, the accuracy of our predictions for certain stations could have been improved.

Despite these limitations, the integration of machine learning techniques through NEM has proven itself as a powerful tool for storm surge prediction. It not only contributes to the advancement of storm surge forecasting but also exemplifies the untapped potential of machine learning in this field.

However, it is imperative that future research aims to address these limitations to further enhance the accuracy and applicability of storm surge predictions. This might involve the expansion of data collection, the incorporation of more factors influencing storm surges, and the continued refinement of machine learning models. As we strive to create more comprehensive, reliable, and faster storm surge prediction models, the lessons learned from this study will be invaluable in guiding future efforts.

5 Conclusions

In this study, we employ three machine learning algorithms - Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and XGBoost - as base learners, paired with a stacking methodology, to establish a novel ensemble model (NEM). This model is designed to simulate the maximum water level increases induced by storm surges, using various storm parameters such as Gale distance (Gale\_Dis), nearest wind speed (N\_WS), nearest air pressure (N\_PR), minimum distance (Dis\_Min), large wind radius (LRGW), and maximum pressure drop within 24 hours (MPD(24h)). Our study undertakes a feature selection analysis, choosing six critical parameters: Gale distance (Gale\_Dis), nearest wind speed (N\_WS), nearest air pressure (N\_PR), minimum distance (Dis\_Min), large wind radius (LRGW), and maximum pressure drop within 24 hours (MPD(24h)). To capture the typhoon's rightward bias on upper ocean dynamics, we incorporated a typhoon azimuth parameter (TA) as a feature, significantly enhancing the model's predictive performance. This led us to employ a comprehensive set of seven parameters as input features for machine learning models such as RF, GBDT, XGBoost, and NEM.

The performance evaluation of the base learner algorithms and the ensemble model suggests that the ensemble mode of the NEM outperforms RF, GBDT, and XGBoost. This observation underscores that the NEM successfully integrates the strengths of the base learners to enhance storm surge water level predictions. For the test dataset, the coefficient of determination (*R2*) for the NEM at the TBT station can reach as high as 0.95, with a mean absolute error (MAE) as low as 0.08 m. These results highlight the ensemble machine learning model's ability to accurately capture the maximum water level increase caused by storm surges.

To provide further insights into the ensemble model's performance, we utilize the SHAP method, an interpretable approach, to analyze the influence of different features in the machine learning black box model. Our results indicate that Gale\_Dis and N\_WS are the most critical features affecting the prediction of extreme storm surge values, with the SHAP value of N\_WS increasing in line with N\_WS. N\_PR, Dis\_Min, and LRGW emerge as secondary features, displaying larger SHAP values as their feature values decrease. MPD(24h) has the least impact on the model prediction but also exhibits a positive correlation with the SHAP value. The SHAP values for the storm surge prediction results elucidate each feature's role in storm surge prediction, further verifying the reliability of the model's prediction outcomes.

**Acknowledgments**

This study is supported by National R&D Program of China grant# 2021YFB3900400.

**Open Research**

The archiving of the data used in this study is underway on the repository: <https://github.com/panj1963/ESS2023/>

**References**

Al Kajbaf, A., & Bensi, M. (2020). Application of surrogate models in estimation of storm surge: A comparative assessment. *Applied Soft Computing*, *91*. <https://doi.org/10.1016/j.asoc.2020.106184>.

Alelyani, S. (2021). Stable bagging feature selection on medical data. *Journal of Big Data*, *8*(1). <https://doi.org/10.1186/s40537-020-00385-8>.

Ayyad, M., Hajj, M. R., & Marsooli, R. (2022a). Artificial intelligence for hurricane storm surge hazard assessment. *Ocean Engineering, 245, 110435*. https://doi.org/10.1016/j.oceaneng.2021.110435.

Ayyad, M., Hajj, M. R., & Marsooli, R. (2022b). Machine learning-based assessment of storm surge in the New York metropolitan area. *Scientific Reports, 12*(1), 19215. <https://doi.org/10.1038/s41598-022-23627-6>.

Ayyad, M., Hajj, M. R., & Marsooli, R. (2023). Climate change impact on hurricane storm surge hazards in New York/New Jersey Coastlines using machine-learning. *Npj Climate and Atmospheric Science, 6*(1), 88. https://doi.org/10.1038/s41612-023-00420-4.

Biau, G., Cadre, B., & Rouviere, L. (2019). Accelerated gradient boosting. *Machine Learning*, *108*(6), 971–992. <https://doi.org/10.1007/s10994-019-05787-1>.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. <https://doi.org/10.1023/A:1010933404324>.

Campos, R. M., Costa, M. O., Almeida, F., & Guedes Soares, C. (2021). Operational Wave Forecast Selection in the Atlantic Ocean Using Random Forests. *Journal of Marine Science and Engineering*, *9*(3). <https://doi.org/10.3390/jmse9030298>.

Cha, Y., Shin, J., Go, B., Lee, D.-S., Kim, Y., Kim, T., & Park, Y.-S. (2021). An interpretable machine learning method for supporting ecosystem management: Application to species distribution models of freshwater macroinvertebrates. *Journal of Environmental Management*, *291*, 112719.

Chen, C., Liu, H., & Beardsley, R. C. (2003). An unstructured, finite-volume, three-dimensional, primitive equation ocean model: application to coastal ocean and estuaries. *Journal of Atmospheric & Oceanic Technology*, *20*, 159-186.

Chen, C. (2006). An unstructured grid, finite-volume coastal ocean model (FVCOM) system. *Oceanography*, *19*, 78–89.

Chen, R., Zhang, W., & Wang, X. (2020). Machine Learning in Tropical Cyclone Forecast Modeling: A Review. *Atmosphere*, *11*(7). <https://doi.org/10.3390/atmos11070676>.

Chen, W.-B., Liu, W.-C., & Hsu, M.-H. (2012a). Predicting typhoon-induced storm surge tide with a two-dimensional hydrodynamic model and artificial neural network model. *Natural Hazards and Earth System Sciences*, *12*(12), 3799–3809. <https://doi.org/10.5194/nhess-12-3799-2012>.

Cheng, J., Li, G., & Chen, X. (2019). Research on Travel Time Prediction Model of Freeway Based on Gradient Boosting Decision Tree. *IEEE ACCESS*, *7*, 7466–7480. <https://doi.org/10.1109/ACCESS.2018.2886549>.

de Oliveira, M. M. F., Ebecken, N. F. F., de Oliveira, J. L. F., & de Azevedo Santos, I. (2009). Neural Network Model to Predict a Storm Surge. *Journal of Applied Meteorology and Climatology, 48*(1), 143–155. <https://doi.org/10.1175/2008JAMC1907.1>.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>.

Hashemi, M. R., Spaulding, M. L., Shaw, A., Farhadi, H., & Lewis, M. (2016). An efficient artificial intelligence model for prediction of tropical storm surge. *Natural Hazards*, *82*(1), 471–491. <https://doi.org/10.1007/s11069-016-2193-4>.

Huang, P.-C. (2022). An effective alternative for predicting coastal floodplain inundation by considering rainfall, storm surge, and downstream topographic characteristics. *Journal of Hydrology, 607,* 127544. <https://doi.org/10.1016/j.jhydrol.2022.127544>

Ian, V. -K., Tse, R., Tang, S. -K., & Pau, G. (2022). Novel Prediction in Storm Surge Using Ensemble Machine Learning Algorithms. *In 2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)* (pp. 1229–1234). <https://doi.org/10.1109/PRAI55851.2022.9904245>

Jelesnianski, C. P., Chen, J., & Shaffer, W. A. (1992). SLOSH: Sea, Lake, and Overland Surges from Hurricanes. *NOAA Technical Report, NWS 48, U.S. Department of Commerce*,71 pp.

Kim, S. Y., Matsumi, Y., Shiozaki, S., & Ota, T. (2012). A study of a real-time storm surge forecast system using a neural network at the Sanin Coast, Japan. *2012 Oceans*, 1–7.

Kitamoto, A. (2005). Digital typhoon: Near real-time aggregation, recombination and delivery of typhoon-related information. *Proceeding of the 4th International Symposium on Digital Earth*.

Lee, J.-W., Irish, J. L., Bensi, M. T., & Marcy, D. C. (2021). Rapid prediction of peak storm surge from tropical cyclone track time series using machine learning. *Coastal Engineering*, *170*. <https://doi.org/10.1016/j.coastaleng.2021.104024>.

Lee, T.-L. (2009). Predictions of typhoon storm surge in Taiwan using artificial neural networks. *Advances in Engineering Software*, *40*(11), 1200–1206. <https://doi.org/10.1016/j.advengsoft.2007.06.005>.

Liang, W., Luo, S., Zhao, G., & Wu, H. (2020). Predicting Hard Rock Pillar Stability Using GBDT, XGBoost, and LightGBM Algorithms. *Mathematics*, *8*(5). <https://doi.org/10.3390/math8050765>.

Li, L., Wu, W., Zhang, W., Zhu, Z., Li, Z., Wang, Y., & Niu, S. (2023). Storm surge level prediction based on improved NARX neural network. *Journal of Computational Electronics, 22*(2), 783–804. <https://doi.org/10.1007/s10825-023-02005-z.>

Lockwood, J. W., Lin, N., Oppenheimer, M., & Lai, C.-Y. (2022). Using Neural Networks to Predict Hurricane Storm Surge and to Assess the Sensitivity of Surge to Storm Characteristics. *Journal of Geophysical Research: Atmospheres, 127*(24), e2022JD037617. https://doi.org/10.1029/2022JD037617

Luettich, R.A., Westerink, J. J., & Scheffner, N.W. (1992). ADCIRC: an advanced three-dimensional circulation model for shelves coasts and estuaries, report 1: theory and methodology of ADCIRC-2DDI and ADCIRC-3DL, Dredging Research Program Technical Report DRP-92-6, U.S. Army Engineers Waterways Experiment Station, Vicksburg, MS, 137pp.

Luettich, R. A., & Westerink, J. J. (2004). *Formulation and numerical implementation of the 2D/3D ADCIRC finite element model version 44. XX* (Vol. 20). R. Luettich Chapel Hill, NC, USA.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Mienye, I. D., & Sun, Y. (2022a). A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE Access*, *10*, 99129–99149. <https://doi.org/10.1109/ACCESS.2022.3207287>.

Mienye, I. D., & Sun, Y. (2022b). A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE Access*, *10*, 99,129–99,149. <https://doi.org/10.1109/ACCESS.2022.3207287>.

Price, J. F. (1981). Upper Ocean Responses to a Hurricane. *Journal of Physical Oceanography, 11*(2), 153-175.

Qin, Y., Su, C., Chu, D., Zhang, J., & Song, J. (2023). A Review of Application of Machine Learning in Storm Surge Problems. *Journal of Marine Science and Engineering*, 11(9). <https://doi.org/10.3390/jmse11091729>.

Rajabi-Kiasari, S., Delpeche-Ellmann, N., & Ellmann, A. (2023). Forecasting of absolute dynamic topography using deep learning algorithm with application to the Baltic Sea. *Computers & Geosciences, 178,* 105406. <https://doi.org/10.1016/j.cageo.2023.105406>

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery*, *8*(4). <https://doi.org/10.1002/widm.1249>.

Sajjad, M., & Chan, J. C. L. (2020). Tropical Cyclone Impacts on Cities: A Case of Hong Kong. *Frontiers in Built Environment*, *6*. <https://doi.org/10.3389/fbuil.2020.575534>.

Schulz, E., Speekenbrink, M., & Krause, A. (2018). A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, *85*, 1–16. <https://doi.org/10.1016/j.jmp.2018.03.001>.

Shapley, L. S. (1997). A value for n-person games. *Classics in Game Theory*, *69*.

Tiggeloven, T., Couasnon, A., van Straaten, C., Muis, S., & Ward, P. J. (2021). Exploring deep learning capabilities for surge predictions in coastal areas. *Scientific Reports*, 11(1), 17224. <https://doi.org/10.1038/s41598-021-96674-0>

Wei, Z., & Nguyen, H. C. (2022). Storm Surge Forecast Using an Encoder–Decoder Recurrent Neural Network Model. *Journal of Marine Science and Engineering, 10*(12). https://doi.org/10.3390/jmse10121980.

Xie, W., Xu, G., Zhang, H., & Dong, C. (2023). Developing a deep learning-based storm surge forecasting model. *Ocean Modelling, 182,* 102179. https://doi.org/10.1016/j.ocemod.2023.102179.

Zhang, Y., & Baptista, A. M. (2008). SELFE: A semi-implicit Eulerian-Lagrangian finite-element model for cross-scale ocean circulation. *Ocean Modelling*, *21*(3–4), 71–96. <https://doi.org/10.1016/j.ocemod.2007.11.005>.