

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:<https://orca.cardiff.ac.uk/id/eprint/168803/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Gan, Min, Chen, Yongping, Pan, Shunqi , Lai, Xijun, Pan, Haidong, Wen, Yuncheng and Xia, Mingyan 2024. An improved machine learning-based model to predict estuarine water levels. Ocean Modelling 190 , 102376. 10.1016/j.ocemod.2024.102376

Publishers page: <https://doi.org/10.1016/j.ocemod.2024.102376>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

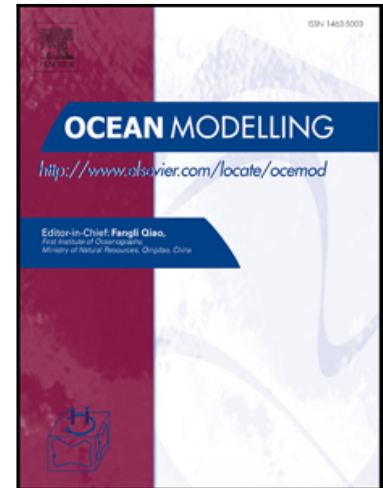
This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



An improved machine learning-based model to predict estuarine water levels

Min Gan , Yongping Chen , Shunqi Pan , Xijun Lai , Haidong Pan ,
Yuncheng Wen , Mingyan Xia

PII: S1463-5003(24)00063-5
DOI: <https://doi.org/10.1016/j.ocemod.2024.102376>
Reference: OCEMOD 102376



To appear in: *Ocean Modelling*

Received date: 5 September 2023
Revised date: 29 January 2024
Accepted date: 5 May 2024

Please cite this article as: Min Gan , Yongping Chen , Shunqi Pan , Xijun Lai , Haidong Pan ,
Yuncheng Wen , Mingyan Xia , An improved machine learning-based model to predict estuarine wa-
ter levels, *Ocean Modelling* (2024), doi: <https://doi.org/10.1016/j.ocemod.2024.102376>

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Highlights:

- Two estuarine water level prediction models are proposed based on the LightGBM
- The two models predict short-term estuarine water level without needing river flow
- Estuarine tidal characteristics demonstrate the weight assignment of the LightGBM

Journal Pre-proof

An improved machine learning-based model to predict estuarine water levels

Min Gan^{1,2,3,4,5}, Yongping Chen^{*3,4}, Shunqi Pan⁵, Xijun Lai^{1,2}, Haidong Pan^{6,7}, Yuncheng Wen^{8,9},
and Mingyan, Xia^{8,9}

¹State Key Laboratory of Lake Science and Environment, Nanjing Institute of Geography and Limnology,
Chinese Academy of Sciences, Nanjing 210008, China

²Poyang Lake Wetland Research Station, Nanjing Institute of Geography and Limnology, Chinese Academy
of Sciences, Jiujiang, China, 332899

³The National Key Laboratory of Water Disaster Prevention, Hohai University, Nanjing 210098, China

⁴College of Harbour, Coastal and Offshore Engineering, Hohai University, Nanjing 210098, China

⁵Hydro-environmental Research Centre, School of Engineering, Cardiff University, Cardiff CF24 3AA,
United Kingdom

⁶First Institute of Oceanography, and Key Laboratory of Marine Science and Numerical Modeling, Ministry
of Natural Resources, Qingdao 266061, China

⁷Laboratory for Regional Oceanography and Numerical Modeling, Pilot National Laboratory for Marine
Science and Technology, Qingdao 266237, China

⁸River and Harbor Engineering Department, Nanjing Hydraulic Research Institute, Nanjing 210029, China

⁹The National Key Laboratory of Water Disaster Prevention, Nanjing Hydraulic Research Institute, Nanjing
210029, China

Corresponding Author: Yongping Chen (ypchen@hhu.edu.cn)

Abstract

The areas around estuaries are typically densely populated and economically developed. Therefore, robust flood risk assessment in these areas is critical. One of the key elements of flood risk assessment is the accurate prediction of estuarine water levels. However, the nonlinear interactions between riverine (i.e., upstream river discharge) and marine (i.e., tides) forces complicate the prediction of estuarine water levels. Traditional physics-based and data-driven models have made significant progress in predicting estuarine water levels, but they require upstream river discharge data as inputs. Considering the lack of such data, the development of new approaches is crucial. This study investigated a machine-learning-based light gradient boosting machine (LightGBM) framework for predicting estuarine water levels using historical water levels as the only inputs. Two prediction models based on the LightGBM framework, denoted as LightGBM1 and LightGBM2, are developed. The LightGBM1 model constructs only a single regression model and uses a recursive approach to generate multidimensional outputs. The LightGBM2 model constructs multiple regression models between the same inputs and outputs in each dimension. The LightGBM1 and LightGBM2 models were applied to the Yangtze estuary as a test case. The results demonstrate that both models are effective at predicting short-term (within 48 hours) estuarine water levels, but the statistical performance of LightGBM2 is better overall. For 24-hour prediction, the root-mean-squared errors of the LightGBM1 and LightGBM2 models are in the ranges of 0.14–0.17 m and 0.12–0.15 m, respectively.

Keywords: Estuarine tide; Yangtze estuary; machine learning; LightGBM

1. Introduction

Because of their favourable environments, estuary zones are densely populated with developed agriculture, transport, communities, and ecosystems (Savenije, 2015). Among the largest 32 cities worldwide, 22 are located near estuary zones (Moftakhari et al., 2019). Estuary zones are highly vulnerable to coastal flooding, particularly flooding associated with global climate change and sea-level rise (Xu et al., 2023). Therefore, the accurate prediction of estuarine water levels is vital for protecting human life and properties.

In oceanic or coastal zones, tide level fluctuations are mainly attributed to astronomic tides, which can be easily analysed and predicted through classical harmonic analysis (Gan et al., 2021b). However, when tides propagate into estuary zones, changes in water depth and estuary width can affect tides (Familkhalili et al., 2020). Furthermore, upstream river discharge has multiple effects on tides. It impedes tidal propagation and causes nonlinear energy transfer between tidal constituents within different tidal species (Gan et al., 2024; Guo et al., 2015), a group of tidal constituents with similar frequencies (Hoitink and Jay, 2016; Zhang et al., 2020). These external forces complicate estuarine tides and make their tidal properties (amplitudes and phases) time-dependent. Numerous studies have reported that estuarine tides worldwide have strong non-stationarity, including the Columbia River Estuary (Gan et al., 2021a; Matte et al., 2013; Pan et al., 2018b), St. Lawrence Fluvial Estuary (Matte et al., 2014), Pearl

Estuary (Cai et al., 2018), and Yangtze Estuary (Chen et al., 2020; Gan et al., 2019). Therefore, considering the non-stationarity of estuarine tides is vital for improving their prediction accuracy.

Researchers have made significant contributions to improving the understanding of estuarine tides using physics-based models. Simple approaches use a series of analytical solutions to describe the variations in tidal waves in estuaries. Kukulka and Jay (2003a) developed a theory representing the variation in tidal properties with upstream river discharge. Kukulka and Jay (2003b) subsequently proposed a framework to describe sub-tidal (low-frequency) water level variations considering marine and riverine forces. Cai et al. (2014) proposed a theoretical framework to describe tidal damping under the influence of upstream river discharge. Based on the analytical solutions presented by Kukulka and Jay (2003a, b) and Jay et al. (2011), Matte et al. (2013) proposed the non-stationary tidal harmonic analysis (NS_TIDE) model, which is an extension of classic harmonic analysis that incorporates the nonlinear influence of upstream river discharge on tides. The NS_TIDE model has become one of the most widely used tools for analysing estuarine tides (Cai et al., 2018; Chen et al., 2020; Gan et al., 2019; Matte et al., 2013, 2014; Pan et al., 2018a, b). Alternatively, full hydrodynamic models require no simplification of nonlinear variables and can directly consider the nonlinear influence of multiple external forces on tides, including upstream river discharge, interventions in hydraulic engineering (Tan et al., 2016), climate change (Liu et al., 2019), and meteorological factors (Dinápoli et al., 2020; Familkhalili

and Talke, 2016).

Artificial intelligence (AI) methods are also excellent choices for predicting estuarine water levels with strong nonlinearity induced by complex interactions between riverine and marine forces. Water level prediction can be mathematically classified as regression, which is a common application area for AI methods. In contrast to physics-based methods, AI methods make no assumptions regarding the functional relationships between inputs and outputs. Their learning abilities enable them to simulate nonlinear or unknown functions (ASCE, 2000a, b). Therefore, AI methods for predicting estuarine water levels have become increasingly common.

One example of an AI-driven estuarine water level prediction is that by Chang and Chen (2003), who applied a radial basis function neural network to conduct one-hour water level predictions in the Danshui River Estuary during the typhoon period. Similarly, Chen et al. (2012) compared the artificial neural network (ANN) and hydrodynamic models to predict water levels in the Danshui River Estuary. Their results indicated that the ANN model achieved accuracy comparable to or even better than that of the hydrodynamic models. More recently, Yoo et al. (2020) tested the performance of a long short-term memory (LSTM) model for predicting the water levels of the Hangang River Estuary in South Korea using different combinations of input parameters such as precipitation, river discharge, and wind speed. Gan et al. (2021b) employed a decision-tree-based light gradient boosting machine (LightGBM) framework to predict the hourly water level of the Lower Columbia River Estuary using

upstream river discharge data as inputs.

Previous studies on estuarine water level prediction have mainly focused on estuaries with observed upstream river discharge data. However, there are thousands of estuaries worldwide (Alder, 2003), some of which do not have observed river discharge data, particularly estuaries with medium or smaller scales. The lack of such observed information may limit the application of physics-based or data-driven models for predicting estuarine water levels. Therefore, new approaches for predicting estuarine water levels when upstream river discharge data is unavailable are needed.

Observed estuarine water levels indirectly reflect the influence of upstream river discharge on tides. Furthermore, estuarine water levels are time series containing periodic fluctuations, particularly on a short-term scale (several days). Therefore, a connection between estuarine water levels may be established (i.e., an auto-correlated relationship). Because machine learning models are adept at learning unknown relationships, we investigated their application to predicting short-term estuarine water levels without prior boundary (upstream and downstream) information. Note that hydrometeorological factors significantly influence estuarine water levels during extreme events. Considering hydrometeorological factors as model inputs for improving the prediction accuracy of machine learning models during severe river floods or typhoons is beyond the scope of this study. The machine learning LightGBM framework was selected for prediction in consideration of its high speed (Ke et al., 2017) and widespread application in finance (Ma et al., 2018; Sun et al., 2020b), medical

science (Gao et al., 2020), and geoscience (Fan et al., 2019). Furthermore, it has already been used successfully to predict estuarine water levels when river discharge information is available (Gan et al., 2021b).

The remainder of this paper is organised as follows. The algorithms utilised in our study are described in Section 2, and study site and field data information are given in Section 3. The results are presented in Section 4, and further discussion is provided in Section 5. The main conclusions are summarised in Section 6.

2. Model description

2.1 LightGBM framework

The LightGBM framework is a decision-tree-based algorithm proposed by Microsoft (Ke et al., 2017). A decision tree algorithm typically consists of root, internal, and leaf nodes. Root nodes contain all input information, whereas internal nodes represent the judgement of input information and are used to conduct further tree node splitting (Wu et al., 2016). Leaf nodes correspond to the outcomes associated with specific inputs. Considering the division of tidal regimes based on tidal form numbers as an example, **Figure 1 (A)** illustrates the concept of the LightGBM framework. When the tidal form number F is in the range no more than 0.25 ($0.25 < F \leq 1.50$; $1.50 < F \leq 3.00$; $F > 3.00$), the tidal form in an area can be classified as a semi-diurnal (mixed but mainly semi-diurnal; mixed but mainly diurnal; diurnal) regime (Frota et al., 2016). All model inputs (tidal form numbers) can be treated as root nodes in the decision tree model. The LightGBM framework can use cut-off values (0.25, 1.50, and 3.00) in a step-by-step fashion to classify all tidal regimes.

In practical applications, a single decision tree is often insufficient to map the relationships between the inputs and outputs accurately. The LightGBM framework uses a gradient boosting decision tree algorithm (Sun et al., 2020a) to construct multiple decision trees. Assuming that the input is X , the output Y is the sum of a series of tree outputs (**Figure 1 (B)**). The mapping relationship between X and Y can be expressed as:

$$Y_I = \sum_{i=1}^I T_i(X, \theta_i), \quad (1)$$

where i is the tree index, I is the total tree number, Y_I is the output from the sum of the I trees, θ_i represents the topological structure, and T_i is the output of the i^{th} tree.

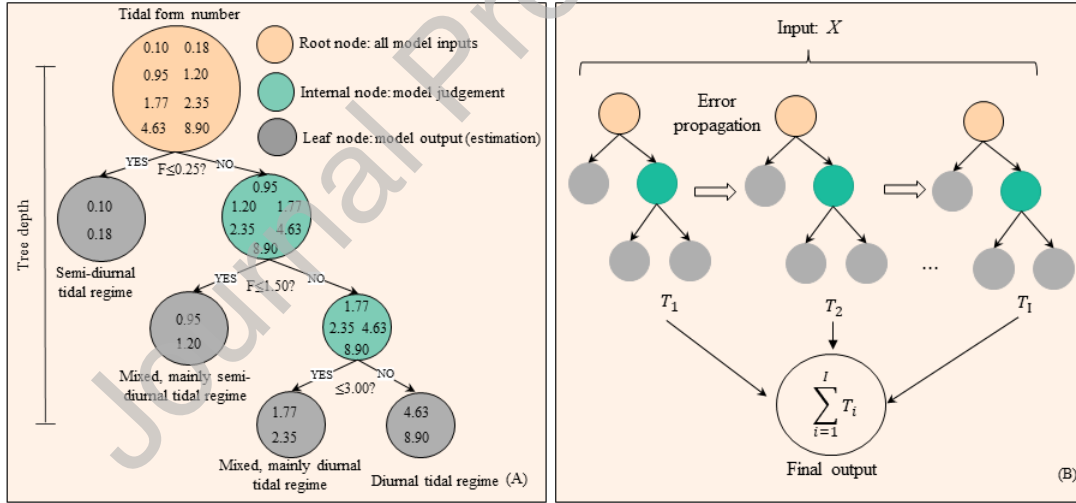


Figure 1. Diagram of the LightGBM framework

During the learning process, the LightGBM framework must be specified using a target function. The target function of the i^{th} tree is expressed as follows:

$$Obj_i = L(\eta - \eta_i) + \Omega_i, \quad (2)$$

where Obj_i denotes the target function of the i^{th} tree, L is the loss function used to

evaluate the errors of the model results relative to the target outputs (i.e., mean squared error in this study), η is the observed estuarine water level, and η_i is the estimated estuarine water level when i trees are used. Ω_i is the following regular function, which is used to improve model robustness and consists of two parts to avoid overfitting:

$$\Omega_i = \alpha \sum_{j=1}^J \omega_j + \beta \sum_{j=1}^J \omega_j^2, \quad (3)$$

where the first and second terms on the right-hand side of Eq. (3) represent lasso regression (L1 regularisation) and ridge regression (L2 regularisation) (Demir-Kavuk et al., 2011), respectively, α and β are corresponding user-defined parameters used to control the learning process, and ω_j is the output of the j^{th} leaf node (grey node in **Figure 1**), which can be expressed as (Gan et al., 2021b):

$$\omega_j = -\frac{sgn(G_j)G_{\alpha,j}}{H_j + \beta}, \quad (4)$$

$$G_{\alpha,j} = \max\{0, (|G_j| - \alpha)\}, \quad (5)$$

where sgn is the sign function, and G_j and H_j are the sums of the first and second derivations of the loss function for the sample points at the j^{th} leaf node, respectively.

2.2 Specification of input and output layers

Figure 2 presents the input and output layers of the LightGBM framework. Because only one-dimensional outputs are supported, only one-step (1 h in this study) predictions can be obtained by the original LightGBM framework when using a time series, which is insufficient for estuarine water level prediction. We propose two indirect methods to derive multidimensional outputs (longer prediction lengths) using the LightGBM framework, as illustrated in **Figure 2**. The past W (window length)

observed water level points are used as inputs to predict the subsequent L (prediction length) water level points (outputs). The first proposed modified LightGBM framework (hereafter referred to as LightGBM1) uses a recursive method to obtain L -dimensional outputs. Each output of the LightGBM1 model is used as one of the inputs for the next step of prediction. Through continuous recursion, the LightGBM1 model can derive L -dimensional outputs. As shown in **Figure 2**, only a single regression model is used in LightGBM1.

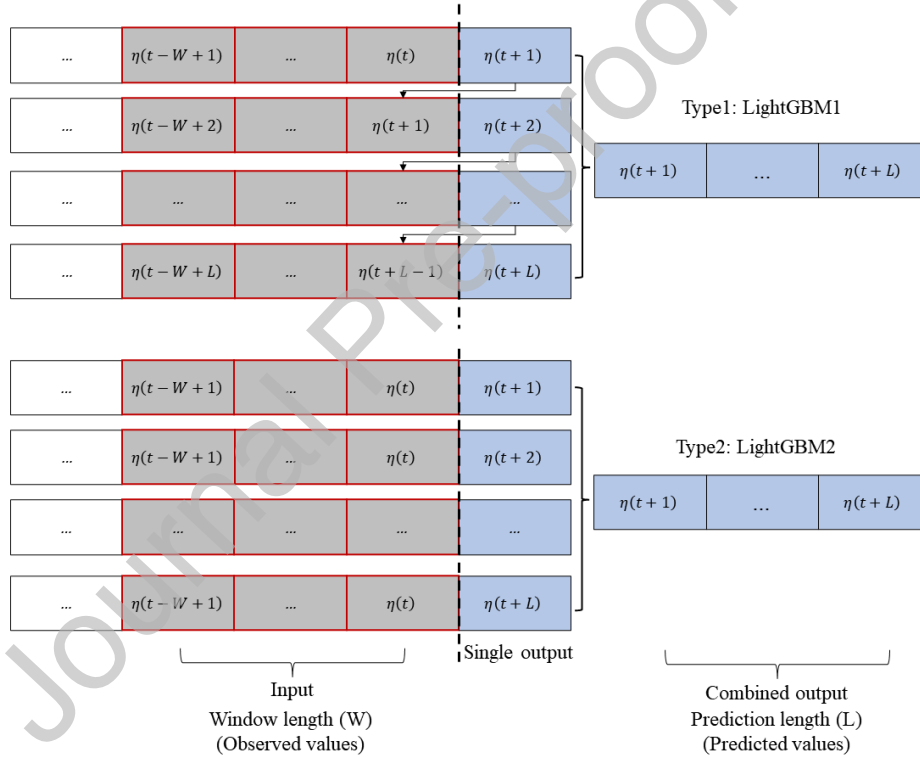


Figure 2. Diagrams of the input and output layers of the LightGBM1 and LightGBM2 models

The second method constructs a multi-regression model between the corresponding inputs and outputs in each step (dimension). Therefore, L regression models are constructed using the second LightGBM framework (hereafter referred to as LightGBM2). If the prediction length is only one step (one hour in this study), then

LightGBM1 and LightGBM2 are equivalent. However, when the prediction length is longer than one step, the LightGBM2 model is expected to be more time-consuming.

The prediction length is specified according to the prediction requirements and was set to 48 hours in this study. With different L values, there may be different optimal W values for maximising prediction performance. For consistency, the same value of W was used to compare variations in the accuracy of LightGBM1 and LightGBM2 with L . Furthermore, the optimal W values of both models were determined according to the minimum errors of single-step prediction (i.e., $L = 1$).

2.3 Parameter tuning

For both LightGBM1 and LightGBM2, two main types of parameters were tuned. The first is the window length parameter (W) for the input structure (**Figure 2**). The second type consists of the internal parameters of the LightGBM framework that control its learning process. Detailed information on the internal parameters of the LightGBM framework can be found in the paper by Ke et al. (2017). An optimisation diagram for all parameters is presented in **Figure 3**, which was modified from the work by Gan et al. (2021b). The data were divided into training and testing datasets. During the training process, a portion of the training data (20% in this study) was used to compare the models trained using different parameter values (i.e., validation period). Once all parameters were optimised, the testing data were used to evaluate the accuracy of the models.

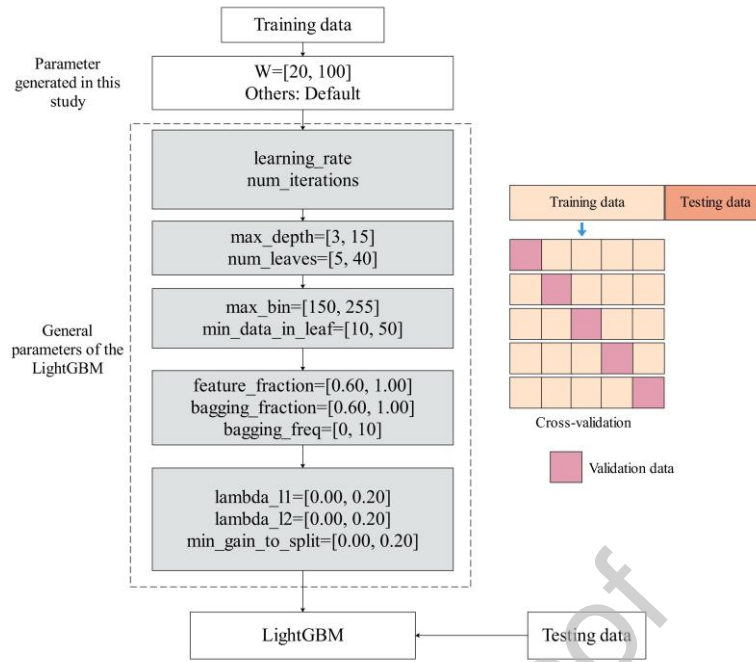


Figure 3. Diagram of parameter optimisation for the LightGBM framework constructed in this study (modified from Gan et al. (2021b))

3. Study site and field data

3.1 Yangtze estuary

The Yangtze River is the longest river in Asia and the third-longest river in the world. It begins in the northern part of the Qinghai-Tibet Plateau (Li et al., 2016). After extending approximately 6,300 km from west to east, the Yangtze River finally flows into the East China Sea. The final stretch of the river is typically referred to as the Yangtze (River) Estuary (**Figure 4**). The Yangtze Estuary includes three-level bifurcations and four outlets (Li et al., 2016). The river discharge is significant and exhibits a distinct seasonal cycle (Gan et al., 2024). The maximum and minimum recorded river discharge values are 92,000 and 4620 m³/s, respectively (Chen et al., 2002). River discharge during the flood season (May to October) accounts for approximately 70% of the yearly river discharge, while river discharge during the dry

season represents the remaining 30% (Song et al., 2013). The monthly mean river discharge values are 45,000 to 60,000 m³/s during the flood season and 10,000 to 20,000 m³/s during the dry season (Guo et al., 2016). Tides in the Yangtze Estuary are dominated by semidiurnal tidal constituents (Gan et al., 2019). M₂ is the dominant tidal constituent, followed by S₂, K₁, and O₁ (Guo et al., 2016). Near the mouth of the Yangtze Estuary, the tides are partially distorted and asymmetric, with the average flood duration (5 h) at the Zhongjun Station near the estuary mouth being shorter than the average ebb duration (7.4 h) (Chen et al., 2016). The mean tidal range near the estuary mouth is 2.0–3.1 m, and the maximum value can reach approximately 4.6 to 6.0 m (Yang et al., 2001). Based on these characteristics, the Yangtze Estuary is classified as a partially mixed mesotidal estuary (Song et al., 2013).

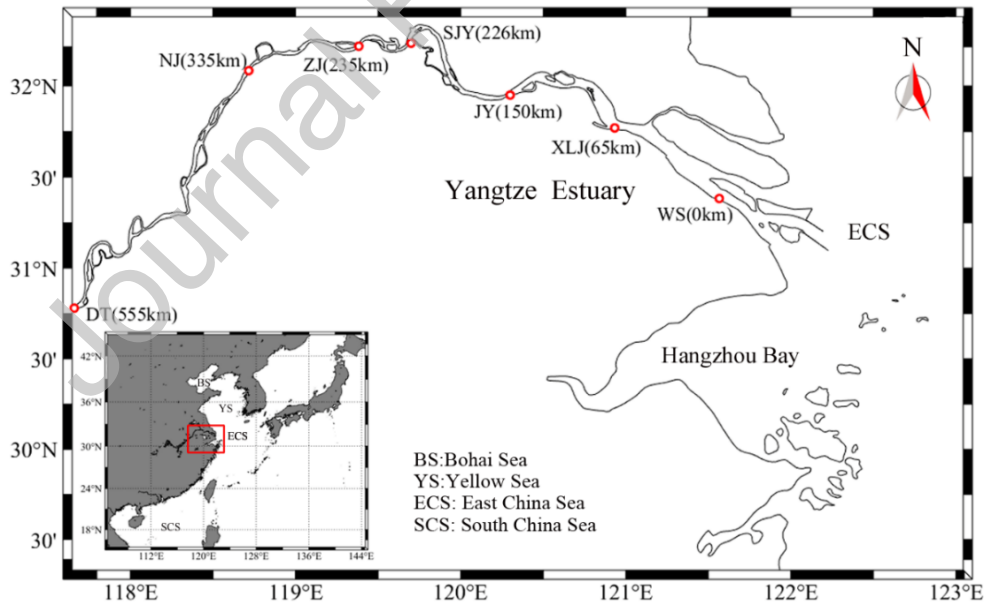


Figure 4. Map of the Yangtze Estuary and the distribution of hydrometric stations considered in this study (taken from Chen et al. (2020))

3.2 Field Data

Figure 4 presents the locations of the tidal stations to collect the water level data used to evaluate the LightGBM models. The water levels at the Xuliujing (XLJ), Jiangyin (JY), Sanjiangying (SJY), Zhenjiang (ZJ), and Nanjing (NJ) stations were used in this study. In the positive landward direction, Wusong (WS) station was used as a reference station to calculate the relative distance of each station. **Figure 5** presents the hourly water level measurements from 2012 to 2020. All water levels were unified to the Wusong datum and times were considered in Chinese standard time. The Yangtze estuarine water levels became more irregular in the upstream direction, indicating the stronger influence of upstream river discharge on tides. In general, JY is considered as a critical point. From the JY station, riverine and marine forces dominate in the upstream and downstream directions, respectively (Guo et al., 2015).

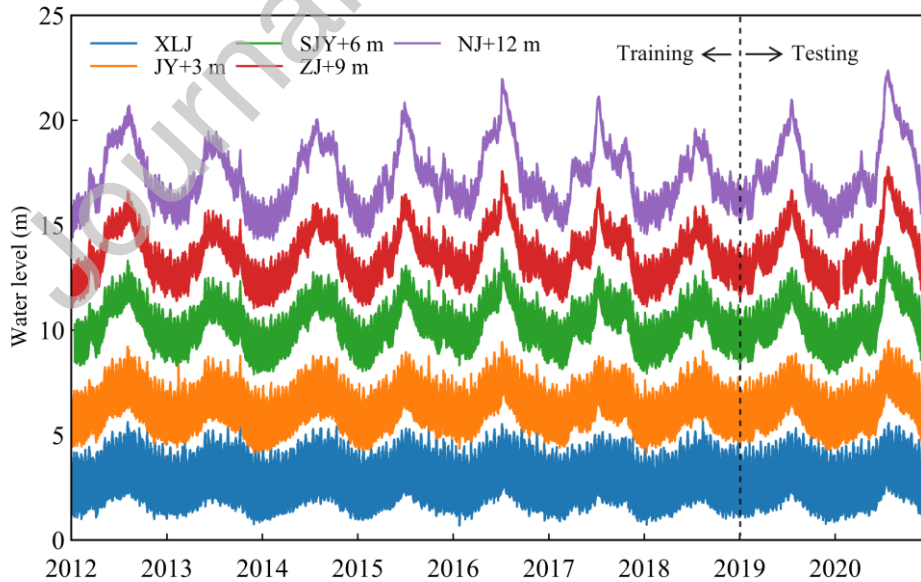


Figure 5. Water level measurements along the Yangtze estuary at XLJ, JY, SJY, ZJ, and NJ

stations

4. Results

4.1 Parameter optimisation

The variations in the root-mean-squared error (RMSE) values of the LightGBM1 and LightGBM2 models for single-step prediction (i.e., $L = 1$) at each station with W are presented in **Figure 6**. As W increase from 20 to 40, all RMSE values at each station decrease significantly. When W is greater than 40, the RMSE values initially exhibit a gradual decreasing trend and eventually reach a stable state with minor fluctuations. The W values corresponding to the minimum RMSE values at stations XLJ, JY, SJY, ZJ, and NJ are 73, 94, 76, 97, and 55, respectively. **Figure 6** reveals that the water levels from the previous two to five days are sufficient for predicting the water levels of the Yangtze Estuary over short timescales.

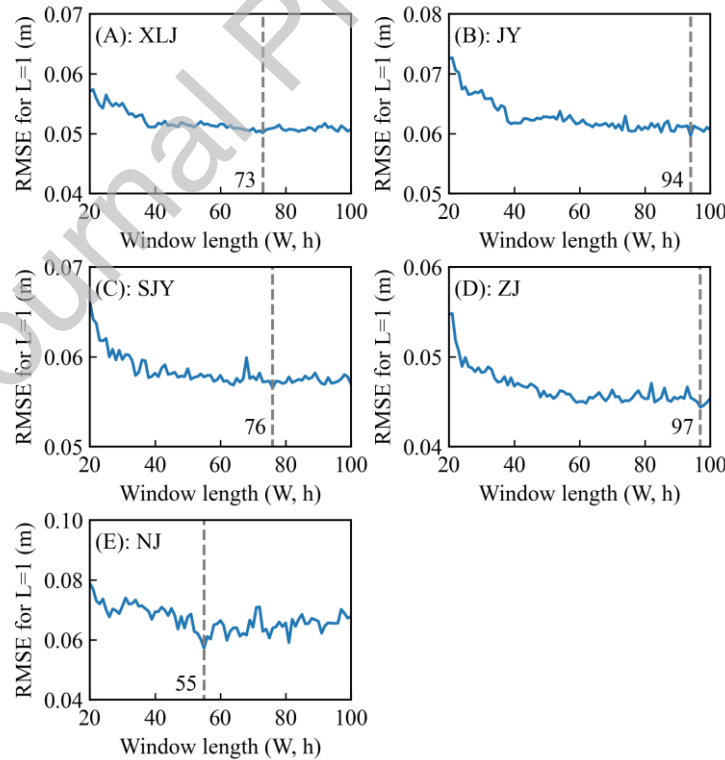


Figure 6. Variation of the RMSE values with different window lengths for the prediction

length of 1 h

Similarly, based on the training process illustrated in **Figure 3**, the remaining parameters of LightGBM1 and LightGBM2 were optimised. The results for the LightGBM1 model are listed in **Table 1**. Because the maximum prediction length (L) was set to 48 hours in this study, there are 48 regression models when using the LightGBM2 model. The parameters of the first regression model in LightGBM2 share the same characteristics as those in LightGBM1 (**Table 1**). The optimised parameters of the other regression models (2nd to 48th) in LightGBM2 were obtained similarly but are not detailed here for the sake of brevity.

Table 1. Optimised parameters for the LightGBM1 model at each station

Parameters	XLJ	JY	SJY	ZJ	NJ
learning_rate	0.20	0.20	0.20	0.20	0.20
num_iterations	2745	1744	3194	2544	2265
max_depth	4	4	6	4	12
num_leaves	9	11	9	20	5
max_bin	245	193	225	255	255
min_data_in_leaf	25	25	23	20	20
feature_fraction	0.690	1.000	1.000	0.760	0.940
bagging_fraction	0.670	0.650	0.650	0.900	0.850
bagging_freq	2	0	0	10	1
lambda_l1	0.170	0.154	0.136	0.132	0.152
lambda_l2	0.060	0.120	0.104	0.000	0.100
min_gain_to_split	0.000	0.000	0.000	0.000	0.000

4.2 Model testing

Based on the previously determined parameters, the LightGBM1 and LightGBM2 prediction models were trained for each station, and the testing data were used to compare their performances. **Figure 7** plots the variation in the RMSE values of

LightGBM1 and LightGBM2 for different prediction lengths. Two general trends can be observed in **Figure 7**. First, the LightGBM2 model performs better than the LightGBM1 model with an increase in L . Second, the RMSE values of both the LightGBM1 and LightGBM2 models exhibit a decreasing trend along the landward direction when L is less than 24. Considering the 24-h prediction as an example, the RMSE values of the LightGBM1 and LightGBM2 models at XLJ are 0.15 m and 0.17 m, respectively. In comparison, the values at the NJ station decrease to 0.12 m and 0.14 m, respectively.

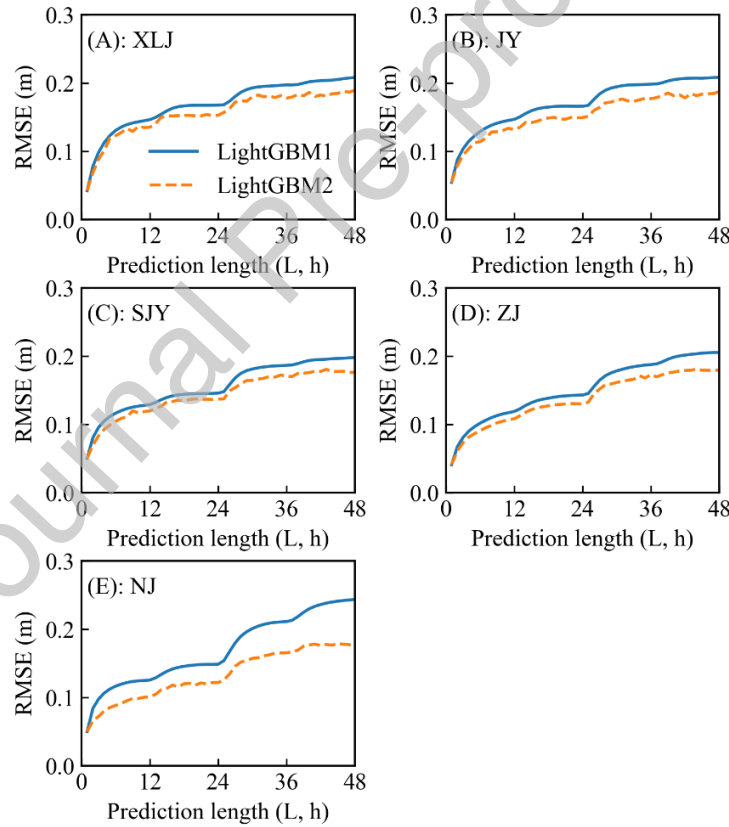


Figure 7. Comparison of the variations in the RMSE values of the LightGBM1 and LightGBM2 models with different prediction lengths

Stations XLJ and NJ are located downstream and upstream of the study area,

respectively (**Figure 4**). Their results were used to evaluate the model characteristics for different tidal reaches. **Figure 8** compares the 24-h predictions of the LightGBM1 and LightGBM2 models to the measurements at the XLJ station. Overall, both the LightGBM1 and LightGBM2 models reasonably predict the water levels at the XLJ station. However, there are two time periods (dashed circles in **Figure 8**) in which more significant errors appear. The details of these two time periods are presented in **Figures 9 and 10**. In **Figure 9**, one can see that significant errors occurred when the Likema typhoon (ID: 201909) influenced the Yangtze Estuary. The maximum absolute error (MAE) values of the LightGBM1 and LightGBM2 models during the Likema typhoon period are 1.09 m and 0.89 m, respectively. Similarly, **Figure 10** illustrates that the second period in **Figure 8** corresponds to the influence of the Tapah (ID: 201917) and Mitag (ID: 201918) typhoons. The MAE values of the LightGBM1 and LightGBM2 models during the Tapah period are 0.71 m and 0.77 m, respectively, which are smaller than their values during the Mitag typhoon (1.36 m and 1.54 m, respectively). Overall, the LightGBM1 and LightGBM2 models both exhibit significant errors at the XLJ station during typhoon periods.

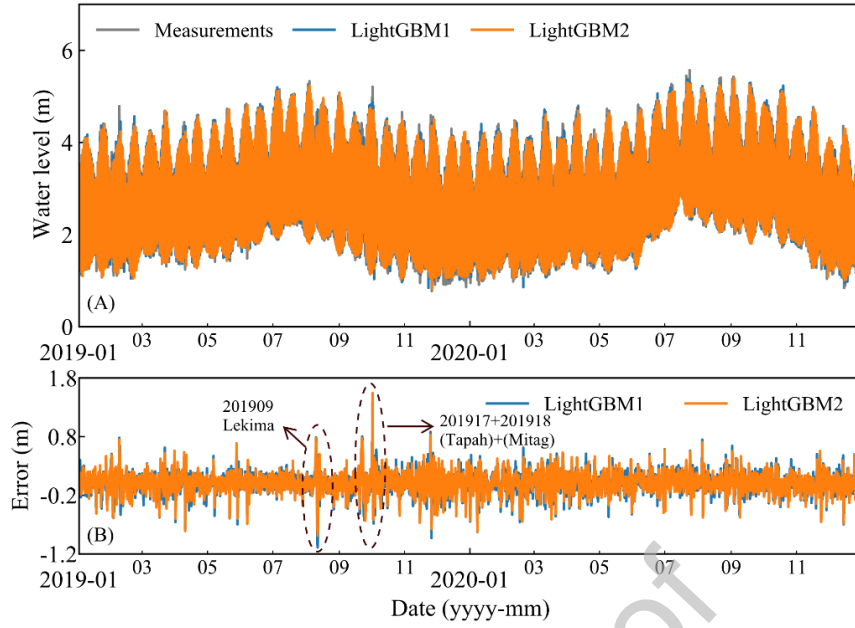


Figure 8. Comparison of the 24-h prediction results of the LightGBM1 and LightGBM2 models with measurements from 2019 to 2020 at the XLJ station (A) and the corresponding model errors (B)

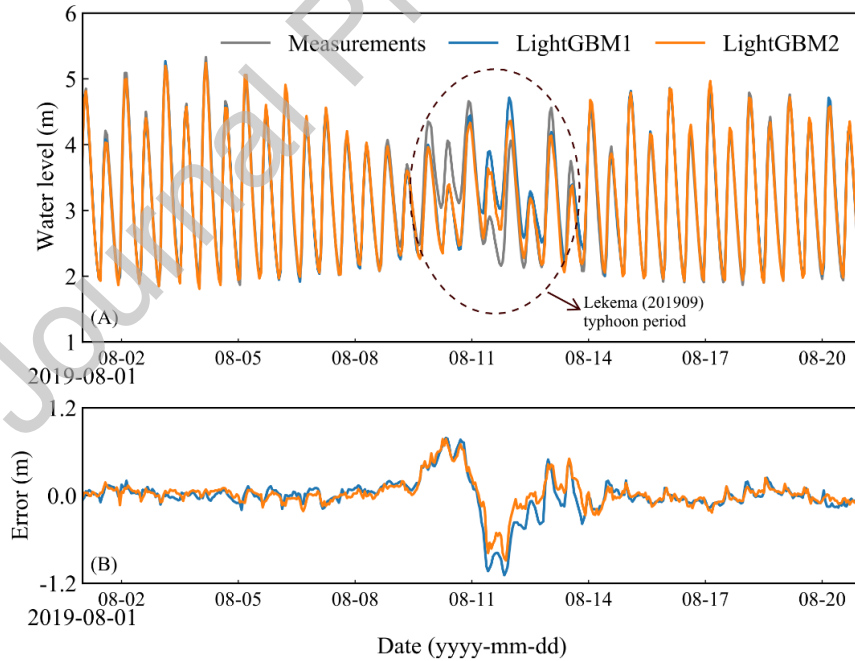


Figure 9. Comparison of the 24-h prediction results of the LightGBM1 and LightGBM2 models with measurements from the Likema (201909) typhoon period at the XLJ station (A) and the corresponding model errors (B)

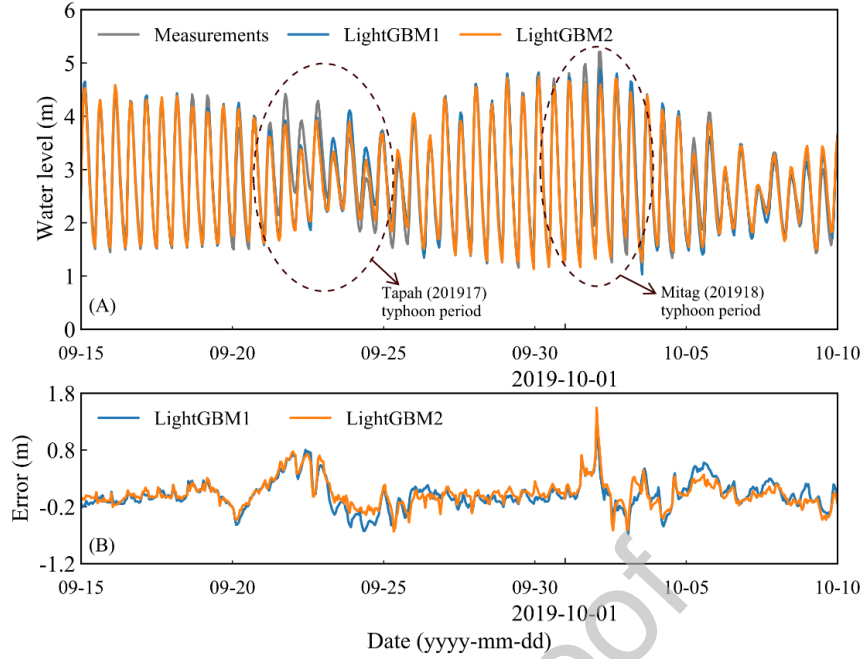


Figure 10. Comparison of the 24-h prediction results of the LightGBM1 and LightGBM2 models with measurements from the Tapah (201917) and Mitag (201918) typhoon periods at the XLJ station (A) and the corresponding model errors (B)

To highlight the relationship between the model errors and typhoon factors, **Figure 11** presents a comparison of the courses and strengths of the Likema, Tapah, and Mitag typhoons. Typhoon information was obtained from the China Meteorological Administration's tropical cyclone database (Feng et al., 2014; Lu et al., 2021). Among the three typhoons, the Likema typhoon is the closest to the Yangtze estuary, followed by Mitag and Tapah. The strength of the Likema typhoon during the time corresponding to the dashed circle in **Figure 9** is classified as a tropical storm, while the Mitag and Tapah typhoons associated with the dashed circle in **Figure 10** belong to severe tropical storms. During the Likema and Tapah typhoon periods, the tide at the XLJ station was close to neap. The peak errors of the two models during the Likema typhoon are greater

than those during the Tapah typhoon. This can be attributed to the closer proximity of the Likema typhoon to the XLJ station compared to the Tapah typhoon. During the Mitag typhoon period (**Figure 10**), the tide at the XLJ station was near spring. Typhoon influences combined with spring tides can enhance the magnitude of storm surges (Stephens et al., 2020). The peak errors of LightGBM1 and LightGBM2 during the Mitag typhoon period are more significant than those during the Likema and Tapah typhoons. Overall, **Figures 8 to 11** demonstrate that both tide strength and typhoon factors (strengths and courses) can affect the errors of the LightGBM1 and LightGBM2 models.

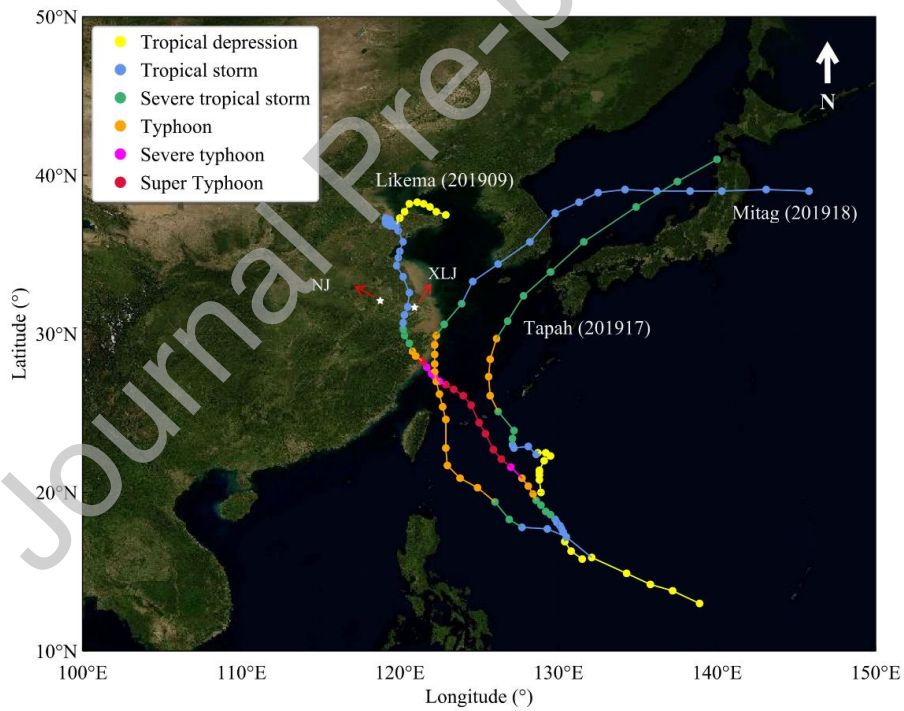


Figure 11. Courses and strengths of the Likema (201909), Tapah (201917), and Mitag (201918) typhoons

For comparison, **Figure 12** plots the 24-h prediction results of the LightGBM1 and LightGBM2 models and the measurements of the NJ station between 2019 and

2020. The overall accuracy is reasonable (RMSE = 0.12 m). Similar to the results for the XLJ station, both models perform better during non-typhoon periods. Even during the three typhoon periods mentioned above, the peak errors of the two models at the NJ station are smaller than those at the XLJ station. The MAE values of the LightGBM1 and LightGBM2 models during the Likema (Tapah, Mitag) typhoon period are 0.43 m and 0.62 m (0.39 m and 0.22 m, 0.49 m and 0.59 m), respectively. The NJ station is much further upstream than the XLJ station (**Figure 4**). Previous studies, such as that by Yin et al. (2021), have indicated that the storm surge at the NJ station is generally smaller than that at the XLJ station, demonstrating that the water level fluctuations of the NJ station are less sensitive to typhoon factors.

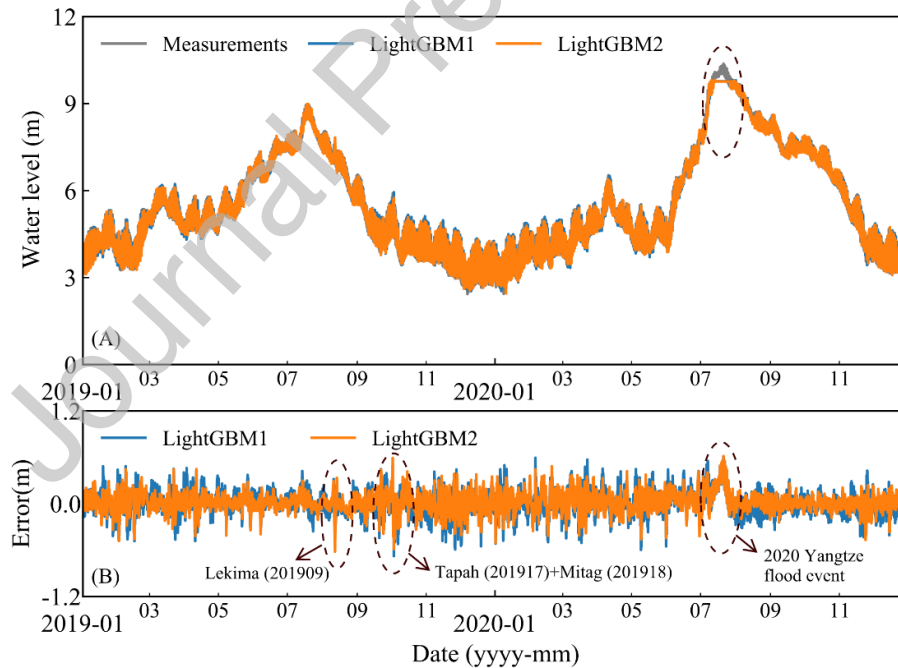


Figure 12. Comparison of the 24-h prediction results of the LightGBM1 and LightGBM2 models with measurements from 2019 to 2020 at the NJ station (A) and the corresponding model errors (B)

Figure 12 also reveals that the two models exhibit more significant errors in July 2020, which corresponds to the flood season of the Yangtze Estuary. To compare model performances during the flood season in detail, **Figures 13** and **14** present the 24-h predictions of the two models, as well as the measurements during July 2019 and 2020, when the yearly maximum water levels appeared. Both models predict the water levels at the NJ station during the flood season of 2019 with reasonable accuracy, although they still exhibit relatively large errors during this period (dashed circle in **Figure 13**). Compared to **Figure 13**, significant model errors are more frequent in **Figure 14** and are continuously distributed. Between 12 July 2020 and 27 July 2020, the results of the LightGBM1 and LightGBM2 models are consistently approximately 0.55 m and 0.56 m lower than the measurements, respectively. During this period, the Yangtze Estuary suffered an extreme flood event that caused the historical maximum water level at the NJ station (10.38 m). The data variation range in **Figure 14** exceeds that of the training data (2012 to 2018), which reduced the performance of the machine learning model.

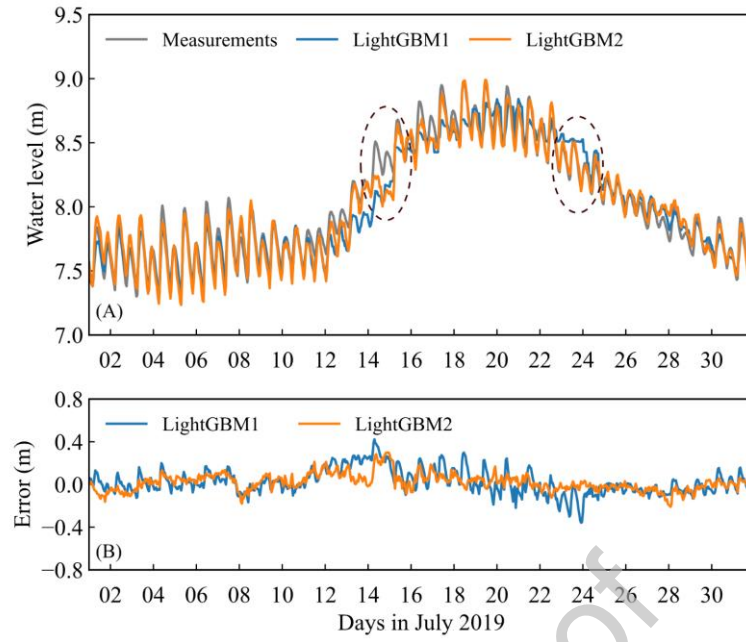


Figure 13. Comparison of the 24-h prediction results of the LightGBM1 and LightGBM2 models with the measurements at NJ station in the flood season of 2019 (A) and the corresponding model errors (B)

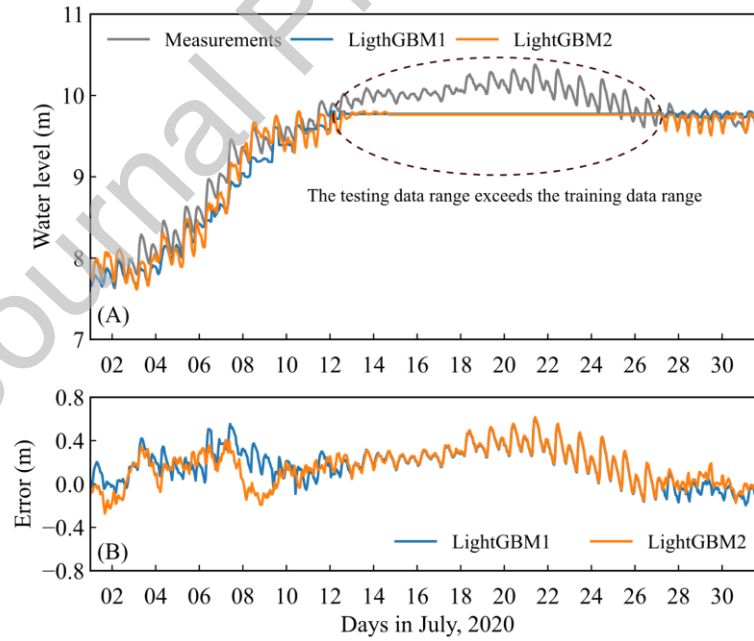


Figure 14. Comparison of the 24-h prediction results of the LightGBM1 and LightGBM2 models with the measurements at NJ station in the flood season of 2020 (A) and the corresponding model errors (B)

5. Discussion

5.1 Differences between training and testing data

Histograms of the training and testing data at the NJ station are presented in **Figure 15** to compare their differences statistically. The maximum water levels of the training and testing data are 9.96 m and 10.38 m, respectively. As shown in **Figure 1 (A)** and **Eq. (4)**, the model output of a leaf node is obtained from the average of all samples at the same leaf node. The minimum number of samples at a leaf node for the NJ station is limited to 20 ($\text{min_data_in_leaf} = 20$ in **Table 1**). Therefore, the leaf node containing the sample with the water level point of 9.96 m in the training dataset also includes other samples whose values are smaller than 9.96 m. When all samples at a leaf node are averaged, the average value is smaller than the maximum value of the sample points. This explains why the maximum predicted water levels of the LightGBM1 and LightGBM2 models in **Figure 14** are near 9.80 m and smaller than 9.96 m. Specifically, the value of 9.80 m is the maximum output of the LightGBM1 and LightGBM2 models learned from the training data.

In previous applications of machine learning models to predict oceanic tide levels, 1 d, 1 week, or 15 d hourly tide level data were sufficient for model training (Lee, 2004; Lee and Jeng, 2002). However, the fluctuations in estuarine water levels are more complex. Influenced by upstream river discharge, estuarine water levels vary with seasonal, annual, or even wet-dry-year variation cycles. To achieve accurate prediction, the training data should contain variations covering the full range of potential conditions. Specifically, the training data should be sufficiently extensive to include

enough extreme events. Otherwise, AI-based models will suffer from degraded performance under extreme conditions, such as the water level prediction of the NJ station during the 2020 extreme flood (**Figure 14**). Therefore, the ideal training data length for estuarine water levels may be much longer than that required to model oceanic tide levels. Overall, **Figures 14** and **15** indicate that the data length can be a factor limiting the performance of machine learning models for predicting estuarine water levels during extreme events.

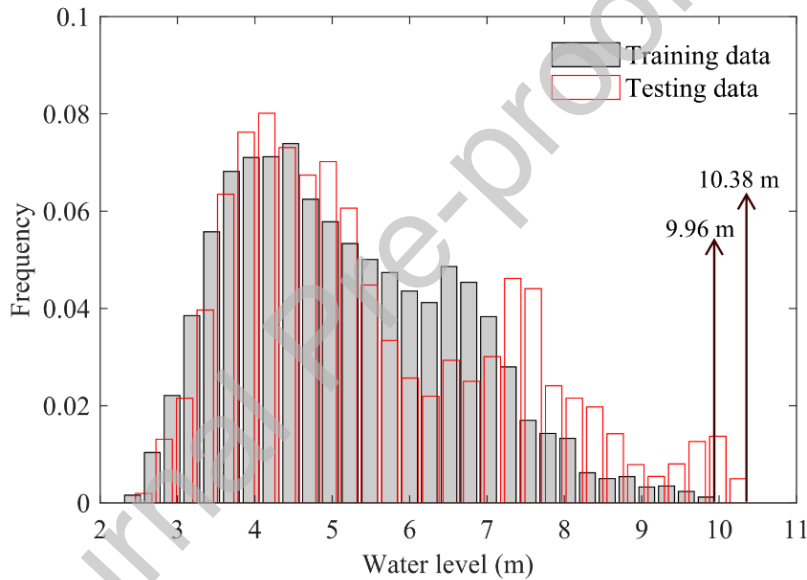


Figure 15. Distributions of the training and testing data at NJ station.

5.2 Connection to a physical background

Although the LightGBM framework is a data-driven algorithm, the weight assignment of the input parameters and model performance can still be correlated with the physical background of the Yangtze estuarine tides. The parameter contributions were evaluated to determine the importance of the parameters in the input layer. The parameter contribution represents the ratio of the number of times the LightGBM

framework selects each parameter to generate a tree node over the total number of times each parameter is chosen. The 10 most important parameters of LightGBM1 (or the first regression model in LightGBM2) for each station are presented in **Figure 16**. The water level from the previous step ($\eta(t)$) contributes the most, indicating that this information most significantly affects the next-step water level ($\eta(t + 1)$). Additionally, the water levels from the past day and half day are also important, with the former being more vital than the latter, as indicated by the diagram in **Figure 16 (F)**. Because tidal variation consists of diurnal and semi-diurnal tidal constituents, the water level points are strongly related to the values of the past day and half day. Furthermore, as a result of the joint influence of diurnal and semi-diurnal tides, a tidal-diurnal inequality appears. Therefore, the magnitude of the water level points is more strongly related to the information from the past day compared to that from the past half day.

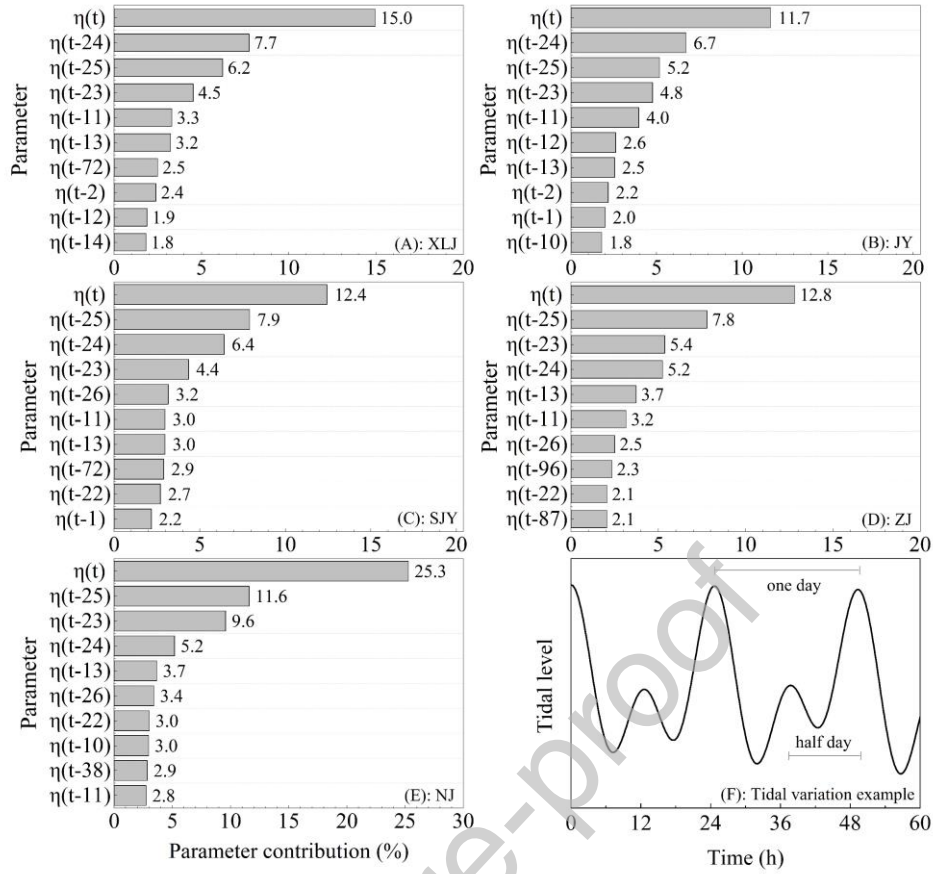


Figure 16. The 10 most important parameters in the LightGBM1 model or the first regression model of the LightGBM2 model for the XLJ (A), JY (B), SJY (C), ZJ (D), and NJ (E) stations, and a diagram illustrating the tidal level variations with diurnal inequality

The Yangtze Estuary tidal characteristics can also illustrate the better performance of the LightGBM models in the upstream direction (**Figure 7**). Historical estuarine water levels were used to predict subsequent estuarine water levels. Therefore, the autocorrelation strength between estuarine water levels can influence model performance. Under the influence of upstream river discharge, high-frequency (i.e., diurnal and semi-diurnal tides) and low-frequency (monthly or annual) tidal variations in the Yangtze Estuary are weakened and enhanced landwards, respectively (Gan et al., 2019, 2024; Guo et al., 2015). Consequently, the water levels on time scales of days are

more autocorrelated landwards owing to their reinforced long-term fluctuation, which enhances their short-term (within 48 hours) predictability, thereby enabling the LightGBM models to perform better landwards.

5.3 Limitations and prospects

Because the two proposed models are specially designed for cases with unavailable boundary information, they cannot consider dramatic meteorological disturbances with their current input settings, resulting in more significant errors during extreme river floods or typhoons. If the LightGBM framework is used to predict estuarine water levels considering dramatic hydrometeorological disturbances, the factors representing external forces should be included as model inputs to improve its performance. Several potential methods for considering such factors are described below.

Machine learning models can predict estuarine water levels during river floods by directly or indirectly considering upstream river discharge. First, the upstream river discharge, downstream tidal range near the estuary mouth, and frequencies of prominent tidal constituents can be used as LightGBM inputs to consider the influence of upstream river discharge and river-tide interactions, which was performed in our previous study (Gan et al., 2021b) and can also be found in other similar studies (Supharatid, 2003). Second, machine learning models can be coupled with physics-based models such as NS_TIDE (Matte et al., 2013) or hydrodynamic models to consider upstream river discharge indirectly to improve the prediction of estuarine

water levels, which is a commonly used modelling concept (Chen et al., 2020; Fei et al., 2023; Zhang et al., 2023).

Regarding the influence of typhoons, numerous studies have reported the application of machine learning models to predict storm surges, but such studies have mainly focused on coastal areas (Hashemi et al., 2016; Hien et al., 2020; Xie et al., 2023). Variables representing storm strength and course are commonly used as model inputs for training. However, estuarine water level prediction during a typhoon period can be more complicated because compound flooding caused by river-tide-surge-rain interactions may occur (Spicer et al., 2019). Additionally, sufficient training data is required for a machine learning model to predict extreme events such as storm surges accurately. Hashemi et al. (2016) pointed out that a limited number of measured typhoon events may be insufficient for developing a robust machine-learning-based prediction model. An effective method to augment typhoon data is to generate synthetic typhoons using statistical models (Nederhoff et al., 2021). Therefore, machine learning models such as the LightGBM model used in this study must be meticulously designed when they are extended to predict estuarine water levels influenced by typhoons, which will be the primary goal of our future research.

6. Conclusions

The nonlinear interactions between river discharge and tides complicate the prediction of estuarine water levels. In particular, when typhoons occur, estuarine water level predictions are more complicated as a result of river-tide-surge-rain interactions.

Traditional physics-based or data-driven models require boundary information on tidal reaches as input data to perform estuarine water level prediction. However, thousands of estuaries worldwide lack boundary observations, such as upstream river discharge, requiring different approaches for predictive modelling. Therefore, we constructed two models (LightGBM1 and LightGBM2) based on the LightGBM framework.

Both the LightGBM1 and LightGBM2 models use historical water levels as the only inputs to predict subsequent water levels. However, they differ significantly in how they derive multidimensional outputs (multistep prediction). LightGBM1 constructs only a single regression model and uses a recursive method for multistep prediction. In contrast, LightGBM2 builds multiple regression models between the same inputs and outputs in each dimension. The LightGBM1 and LightGBM2 models are the same if the prediction length is only one step. However, the training of the LightGBM2 model can be more time-consuming than that of the LightGBM1 model with an increased prediction length.

Statistical results revealed that the accuracy of the LightGBM2 model gradually became higher than that of the LightGBM1 model for longer predictions. For 24-h prediction, the RMSE values of the LightGBM2 model along the Yangtze Estuary are 0.12 to 0.15 m, outperforming the LightGBM1 model (0.14 to 0.17 m). Therefore, the LightGBM2 model is more suitable for the Yangtze Estuary. Generally, the accuracy of the LightGBM2 model increases along the upstream direction, which is related to the autocorrelation enhancement of Yangtze Estuary water levels in the landward direction.

Additionally, the weight assignment of the LightGBM framework to input parameters is associated with the tidal patterns of the Yangtze Estuary.

Although both LightGBM1 and LightGBM2 statistically predict the water levels along the Yangtze Estuary with reasonable accuracy, they still exhibit more significant errors at specific times, such as during typhoon periods. The MAE values during the Likema, Tapah, and Mitag typhoons range from 0.22 m to 1.54 m. These error magnitudes are related to typhoon factors (strength and course), tide strength, and the location of the tide reach. The results at the NJ station reveal that a shortage of training data (not covering enough extreme flood events) could also limit the accuracy of both models. Their predictions at the NJ station during the 2020 extreme flood event are consistently lower than the measured values, and their MAE values reach 0.55 m and 0.56 m, respectively. Therefore, further investigation is required to improve the performance of LightGBM1 and LightGBM2 during typhoon periods and extreme flood events by considering hydrometeorological data as additional inputs.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability Statement:

The data are not publicly available due to confidentiality. The framework of the

LightGBM can be obtained from <https://lightgbm.readthedocs.io/en/v3.3.2/Python-Intro.html>.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (NO. 2023YFC3008100 2021YFC30001000), the National Natural Science Foundation of China (No. 42171012 42206022), the Strategic Priority Research Program of Chinese Academy of Sciences (No. XDA230402), the Key Research Program of Department of Water Resources of Zhejiang Province, China (RA2202), the Science and Technology Planning Project of NIGLAS (No. NIGLAS2022GS07), and the Central Public-Interest Scientific Institution Basal Research Fund (No. Y222002). The first author would like to acknowledge the financial support from the China Scholarship Council (CSC) under the PhD exchange program (No. 201906710022) with Cardiff University and the financial support of the Jiangsu Funding Program for Excellent Postdoctoral Talent (No. 2023ZB099).

References

- Alder, J., 2003. Putting the coast in the “Sea Around Us”. <http://data.unep-wcmc.org/datasets/23>
- ASCE, 2000a. Artificial neural networks in hydrology. I: Preliminary concepts. Journal of Hydrologic Engineering, 5(2): 115-123. DOI:10.1061/(asce)1084-0699(2000)5:2(115)
- ASCE, 2000b. Artificial neural networks in hydrology. II: Hydrologic applications.

- Journal of Hydrologic Engineering, 5(2): 124-137. DOI:10.1061/(asce)1084-0699(2000)5:2(124)
- Cai, H., Savenije, H., Toffolon, M., 2014. Linking the river to the estuary: influence of river discharge on tidal damping. *Hydrology and Earth System Sciences*, 18(1): 287-304. DOI:10.5194/hess-18-287-2014
- Cai, H., Yang, Q., Zhang, Z., Guo, X., Liu, F., Ou, S., 2018. Impact of river-tide dynamics on the temporal-spatial distribution of residual water level in the Pearl river channel networks. *Estuaries and Coasts*, 41(7): 1885-1903. DOI:10.1007/s12237-018-0399-2
- Chang, F., Chen, Y., 2003. Estuary water-stage forecasting by using radial basis function neural network. *Journal of Hydrology*, 270(1-2): 158-166. DOI:10.1016/s0022-1694(02)00289-5
- Chen, W., Chen, K., Kuang, C., Zhu, D.Z., He, L., Mao, X., Liang, H., Song, H., 2016. Influence of sea level rise on saline water intrusion in the Yangtze River Estuary, China. *Applied Ocean Research*, 54: 12-25. DOI:10.1016/j.apor.2015.11.002
- Chen, W., Liu, W., Hsu, M., 2012. Comparison of ANN approach with 2D and 3D hydrodynamic models for simulating estuary water stage. *Advances in Engineering Software*, 45(1): 69-79. DOI:10.1016/j.advengsoft.2011.09.018
- Chen, X., Zhang, D., Zhang, E., 2002. The South to North Water Diversions in China: Review and Comments. *Journal of Environmental Planning and Management*, 45(6): 927-932. DOI:10.1080/0964056022000024415

- Chen, Y., Gan, M., Pan, S., Pan, H., Zhu, X., Tao, Z., 2020. Application of autoregressive (AR) analysis to improve short-term prediction of water levels in the Yangtze estuary. *Journal of Hydrology*, 590: 125386. DOI:10.1016/j.jhydrol.2020.125386
- Demir-Kavuk, O., Kamada, M., Akutsu, T., Knapp, E.-W., 2011. Prediction using stepwise L1, L2 regularization and feature selection for small data sets with large number of features. *BMC Bioinformatics*, 12(1): 1-10. DOI:10.1186/1471-2105-12-412
- Dinápoli, M.G., Simionato, C.G., Moreira, D., 2020. Nonlinear tide-surge interactions in the Río de la Plata Estuary. *Estuarine, Coastal and Shelf Science*, 241. DOI:10.1016/j.ecss.2020.106834
- Familkhalili, R., Talke, S.A., 2016. The effect of channel deepening on tides and storm surge: A case study of Wilmington, NC. *Geophysical Research Letters*, 43(17): 9138-9147. DOI:10.1002/2016gl069494
- Familkhalili, R., Talke, S.A., Jay, D.A., 2020. Tide - Storm Surge Interactions in Highly Altered Estuaries: How Channel Deepening Increases Surge Vulnerability. *Journal of Geophysical Research: Oceans*, 125(4). DOI:10.1029/2019jc015286
- Fan, J., Ma, X., Wu, L., Zhang, F., Yu, X., Zeng, W., 2019. Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. *Agricultural Water Management*, 225: 105758. DOI:10.1016/j.agwat.2019.105758

- Fei, K., Du, H., Gao, L., 2023. Accurate water level predictions in a tidal reach: Integration of Physics-based and Machine learning approaches. *Journal of Hydrology*, 622. DOI:10.1016/j.jhydrol.2023.129705
- Feng, J., Lu, X., Yu, H., Zhang, W., Ying, M., Fan, Y., Zhu, Y., Chen, D., 2014. An Overview of the China Meteorological Administration Tropical Cyclone Database. *Journal of Atmospheric and Oceanic Technology*, 31(2): 287-301. DOI:10.1175/jtech-d-12-00119.1
- Frota, F.F., Truccolo, E.C., Schettini, C.A., 2016. Tidal and sub-tidal sea level variability at the northern shelf of the Brazilian Northeast Region. *An Acad Bras Cienc*, 88(3): 1371-1386. DOI:10.1590/0001-3765201620150162
- Gan, M., Chen, Y., Pan, H., Pan, S., Lai, X., Wen, Y., Xia, M., Zhang, Y., 2024. Study on the spatiotemporal variation of the Yangtze estuarine tidal species. *Estuarine, Coastal and Shelf Science*. DOI:10.1016/j.ecss.2024.108637
- Gan, M., Chen, Y., Pan, S., Li, J., Zhou, Z., 2019. A modified nonstationary tidal harmonic analysis model for the Yangtze estuarine tides. *Journal of Atmospheric and Oceanic Technology*, 36(4): 513-525. DOI:10.1175/jtech-d-18-0199.1
- Gan, M., Pan, H., Chen, Y., Pan, S., 2021a. Application of the Variational Mode Decomposition (VMD) method to river tides. *Estuarine, Coastal and Shelf Science*, 261. DOI:10.1016/j.ecss.2021.107570
- Gan, M., Pan, S., Chen, Y., Cheng, C., Pan, H., Zhu, X., 2021b. Application of the machine learning LightGBM model to the prediction of the water levels of the lower

- Columbia River. *Journal of Marine Science and Engineering*, 9(5): 496.
DOI:10.3390/jmse9050496
- Gao, H., Ye, Z., Dong, J., Gao, H., Yu, H., Li, H., Ouyang, D., 2020. Predicting drug/phospholipid complexation by the lightGBM method. *Chemical Physics Letters*, 747. DOI:10.1016/j.cplett.2020.137354
- Guo, L., van der Wegen, M., Jay, D.A., Matte, P., Wang, Z.B., Roelvink, D., He, Q., 2015. River-tide dynamics: Exploration of nonstationary and nonlinear tidal behavior in the Yangtze River estuary. *Journal of Geophysical Research: Oceans*, 120(5): 3499-3521. DOI:10.1002/2014jc010491
- Guo, L., Wegen, M., Wang, Z.B., Roelvink, D., He, Q., 2016. Exploring the impacts of multiple tidal constituents and varying river flow on long - term, large - scale estuarine morphodynamics by means of a 1 - D model. *Journal of Geophysical Research: Earth Surface*, 121(5): 1000-1022. DOI:10.1002/2016jf003821
- 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. DOI:10.1007/s11069-016-2193-4
- Hien, N.T., Tran, C.T., Nguyen, X.H., Kim, S., Phai, V.D., Thuy, N.B., Van Manh, N., 2020. Genetic Programming for storm surge forecasting. *Ocean Engineering*, 215. DOI:10.1016/j.oceaneng.2020.107812
- Hoitink, A., Jay, D.A., 2016. Tidal river dynamics: Implications for deltas. *Reviews of Geophysics*, 54(1): 240-272. DOI:10.1002/2015RG000507

Jay, D.A., Leffler, K., Degens, S., 2011. Long-Term Evolution of Columbia River Tides.

Journal of Waterway, Port, Coastal, and Ocean Engineering, 137(4): 182-191.

DOI:10.1061/(asce)ww.1943-5460.0000082

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017.

Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural

information processing systems, 30: 3146-3154. DOI:10.5555/3294996.3295074

Kukulka, T., Jay, D.A., 2003a. Impacts of Columbia River discharge on salmonid

habitat: 1. A nonstationary fluvial tide model. Journal of Geophysical Research-

Oceans, 108(C9): 3293. DOI:10.1029/2002jc001382

Kukulka, T., Jay, D.A., 2003b. Impacts of Columbia River discharge on salmonid

habitat: 2. Changes in shallow-water habitat. Journal of Geophysical Research-

Oceans, 108(C9): 3294. DOI:10.1029/2003jc001829

Lee, T.-L., 2004. Back-propagation neural network for long-term tidal predictions.

Ocean Engineering, 31(2): 225-238. DOI:10.1016/s0029-8018(03)00115-x

Lee, T., Jeng, D., 2002. Application of artificial neural networks in tide-forecasting.

Ocean Engineering, 29(9): 1003-1022. DOI:10.1016/s0029-8018(01)00068-3

Li, G., Xiang, X., Guo, C., 2016. Analysis of nonstationary change of annual maximum

level records in the Yangtze River estuary. Advances in Meteorology, 2016: 1-14.

DOI:10.1155/2016/7205723

Liu, Z.C., de Lange, W.P., Bryan, K.R., 2019. Estuary rejuvenation in response to sea

level rise: an example from Tairua Estuary, New Zealand. Geo-Marine Letters,

40(2): 269-280. DOI:10.1007/s00367-019-00603-0

Lu, X., Yu, H., Ying, M., Zhao, B., Zhang, S., Lin, L., Bai, L., Wan, R., 2021. Western North Pacific Tropical Cyclone Database Created by the China Meteorological Administration. *Advances in Atmospheric Sciences*, 38(4): 690-699. DOI:10.1007/s00376-020-0211-7

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., Niu, X., 2018. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications*, 31: 24-39. DOI:10.1016/j.elerap.2018.08.002

Matte, P., Jay, D.A., Zaron, E.D., 2013. Adaptation of classical tidal harmonic analysis to nonstationary tides, with application to river tides. *Journal of Atmospheric and Oceanic Technology*, 30(3): 569-589. DOI:10.1175/Jtech-D-12-00016.1

Matte, P., Secretan, Y., Morin, J., 2014. Temporal and spatial variability of tidal-fluvial dynamics in the St. Lawrence fluvial estuary: An application of nonstationary tidal harmonic analysis. *Journal of Geophysical Research: Oceans*, 119(9): 5724-5744. DOI:10.1002/2014jc009791

Moftakhari, H., Schubert, J.E., AghaKouchak, A., Matthew, R.A., Sanders, B.F., 2019. Linking statistical and hydrodynamic modeling for compound flood hazard assessment in tidal channels and estuaries. *Advances in Water Resources*, 128: 28-38. DOI:10.1016/j.advwatres.2019.04.009

Nederhoff, K., Hoek, J., Leijnse, T., van Ormondt, M., Caires, S., Giardino, A., 2021.

Simulating synthetic tropical cyclone tracks for statistically reliable wind and pressure estimations. *Natural Hazards and Earth System Sciences*, 21(3): 861-878.

DOI:10.5194/nhess-21-861-2021

Pan, H., Guo, Z., Wang, Y., Lv, X., 2018a. Application of the EMD Method to River

Tides. *Journal of Atmospheric and Oceanic Technology*, 35(4): 809-819.

DOI:10.1175/jtech-d-17-0185.1

Pan, H., Lv, X., Wang, Y., Matte, P., Chen, H., Jin, G., 2018b. Exploration of Tidal-

Fluvial Interaction in the Columbia River Estuary Using S_TIDE. *Journal of*

Geophysical Research: Oceans, 123(9): 6598-6619. DOI:10.1029/2018jc014146

Savenije, H.H.G., 2015. Prediction in ungauged estuaries: An integrated theory. *Water*

Resources Research, 51(4): 2464-2476. DOI:10.1002/2015wr016936

Song, D., Wang, X.H., Cao, Z., Guan, W., 2013. Suspended sediment transport in the

Deepwater Navigation Channel, Yangtze River Estuary, China, in the dry season

2009: 1. Observations over spring and neap tidal cycles. *Journal of Geophysical*

Research: Oceans, 118(10): 5555-5567. DOI:10.1002/jgrc.20410

Spicer, P., Huguenard, K., Ross, L., Rickard, L.N., 2019. High - Frequency Tide -

Surge - River Interaction in Estuaries: Causes and Implications for Coastal

Flooding. *Journal of Geophysical Research: Oceans*, 124(12): 9517-9530.

DOI:10.1029/2019jc015466

Stephens, S.A., Bell, R.G., Haigh, I.D., 2020. Spatial and temporal analysis of extreme

- storm-tide and skew-surge events around the coastline of New Zealand. *Natural Hazards and Earth System Sciences*, 20(3): 783-796. DOI:10.5194/nhess-20-783-2020
- Sun, R., Wang, G., Zhang, W., Hsu, L.-T., Ochieng, W.Y., 2020a. A gradient boosting decision tree based GPS signal reception classification algorithm. *Applied Soft Computing*, 86. DOI:10.1016/j.asoc.2019.105942
- Sun, X., Liu, M., Sima, Z., 2020b. A novel cryptocurrency price trend forecasting model based on LightGBM. *Finance Research Letters*, 32. DOI:10.1016/j.frl.2018.12.032
- Supharatid, S., 2003. Application of a neural network model in establishing a stage–discharge relationship for a tidal river. *Hydrological processes*, 17(15): 3085-3099. DOI:10.1002/hyp.1278
- Tan, Y., Yang, F., Xie, D., 2016. The Change of Tidal Characteristics under the Influence of Human Activities in the Yangtze River Estuary. *Journal of Coastal Research*, 75(sp1): 163-167. DOI:10.2112/si75-033.1
- Wu, C.-C., Chen, Y.-L., Liu, Y.-H., Yang, X.-Y., 2016. Decision tree induction with a constrained number of leaf nodes. *Applied Intelligence*, 45(3): 673-685. DOI:10.1007/s10489-016-0785-z
- Xie, W., Xu, G., Zhang, H., Dong, C., 2023. Developing a deep learning-based storm surge forecasting model. *Ocean Modelling*, 182. DOI:10.1016/j.ocemod.2023.102179
- Xu, K., Zhuang, Y., Bin, L., Wang, C., Tian, F., 2023. Impact assessment of climate

change on compound flooding in a coastal city. *Journal of Hydrology*, 617: 129166.

DOI:10.1016/j.jhydrol.2023.129166

Yang, S.L., Ding, P.X., Chen, S.L., 2001. Changes in progradation rate of the tidal flats at the mouth of the Changjiang (Yangtze) River, China. *Geomorphology*, 38(1-2): 167-180. DOI:Doi 10.1016/S0169-555x(00)00079-9

Yin, C., Zhang, W., Xiong, M., Wang, J., Zhou, C., Dou, X., Zhang, J., 2021. Storm surge responses to the representative tracks and storm timing in the Yangtze Estuary, China. *Ocean Engineering*, 233. DOI:10.1016/j.oceaneng.2021.109020

Yoo, H.J., Kim, D.H., Kwon, H.-H., Lee, S.O., 2020. Data Driven Water Surface Elevation Forecasting Model with Hybrid Activation Function—A Case Study for Hangang River, South Korea. *Applied Sciences*, 10(4). DOI:10.3390/app10041424

Zhang, M., Yang, H., Tang, Q., Cai, H., Zhu, Z., Feng, A., Luo, M., Gao, H., Tian, X., 2020. Impacts of secondary and quarter-diurnal tidal species on backwater hydrodynamics in tidal rivers. *Advances in Water Resources*, 143. DOI:10.1016/j.advwatres.2020.103660

Zhang, Z., Zhang, L., Yue, S., Wu, J., Guo, F., 2023. Correction of nonstationary tidal prediction using deep-learning neural network models in tidal estuaries and rivers. *Journal of Hydrology*, 622. DOI:10.1016/j.jhydrol.2023.129686

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

--