TẠP CHÍ
KHÍ TƯỢNG THỦY VĂN

Bài báo khoa học

Kết quả bước đầu thử nghiệm thuật toán XGBoost dự báo nước dâng do bão tại trạm Hòn Dáu

Bùi Mạnh Hà[1], Nguyễn Bá Thủy[1]*, Phạm Khánh Ngọc[1], Phạm Văn Tiến[2]

[1] Trung tâm Dự báo khí tượng thủy văn quốc gia; manhhamhc@gmail.com; thuybanguyen@gmail.com; ngocpkchibo@gmail.com
[2] Viện Khoa học Khí tượng Thủy văn và Biến đổi khí hậu; phamvantien@gmail.com

*Tác giả liên hệ: thuybanguyen@gmail.com; Tel.: +84–975853471

Ban Biên tập nhận bài: 8/10/2022; Ngày phản biện xong: 1/11/2023; Ngày đăng bài: 25/12/2023

Tóm tắt: Trong nghiên cứu này, thuật toán tăng cường độ dốc cấp cao XGBoost (Extreme Gradient Boosting, sau đây gọi là mô hình XGBoost) được ứng dụng để xây dựng công cụ dự báo nước dâng do bão tại Hòn Dáu. Mô hình XGBoost được xây dựng với 4 phương án sử dụng dữ liệu khác nhau (04 mô hình): mô hình XGBoost đơn biến, mô hình XGBoost đa biến I, mô hình XGBoost đa biến II và mô hình XGBoost sử dụng dữ liệu chéo. Bộ dữ liệu trong 28 cơn bão ảnh hưởng tới trạm Hòn Dáu giai đoạn 2002-2021 được thu thập để xây dựng các mô hình và kiểm định kết quả dự báo. Kết quả thử nghiệm mô hình XGBoost dự báo nước dâng do bão cho thấy, mô hình XGBoos đơn biến cho độ tin cậy thấp ở tất cả các thời hạn dự báo. Trong khi đó, hai mô hình XGBoos đa biến và mô hình sử dụng dữ liệu chéo đều cho kết quả tin cậy cao, với phần lớn hệ số tương quan giữa dự báo và quan trắc đều trên 80%. Kết quả của nghiên cứu làm cơ sở lựa chọn công cụ dự báo nước dâng do bão tại Hòn Dáu tuỳ thuộc vào hiện trạng số liệu quan trắc khí tượng, hải văn.

Từ khóa: Dự báo nước dâng do bão; XGBoost; Machine Learning; AI.

1. Mở đầu

Thông thường, các mô hình số như POM [1–2], ROMS [3–4], ADCIRC [5–6], SuWAT [7–10], Delft3D [11–12] được sử dụng để dự báo nước dâng do bão có độ tin cậy cao do sử dụng các phương trình vật lý chính xác, nhưng thường đi kèm với đó là cần tài nguyên tính toán lớn, thời gian vận hành khá đáng kể. Với ưu điểm là linh hoạt và mạnh mẽ, có khả năng xác định các mối quan hệ phức tạp trong dung lượng lớn dữ liệu đầu vào nên trí tuệ nhân tạo (*AI:* Artificial Intelligence) sẽ phù hợp khi áp dụng để triển khai học máy chuyên sâu, phân tích dữ liệu đa yếu tố nhanh và chính xác. Dự báo nước dâng do bão theo hướng sử dụng phương pháp học máy đã được nhiều nhà khoa học trên thế giới nghiên cứu và phát triển mạnh mẽ trong thời gian gần đây. Mạng nơ-ron nhân tạo (*ANN: Artificial Neural Networks*) đã được sử dụng phổ biến trong dự báo độ cao nước dâng do bão [13–19].Với nguyên lý kết hợp các mô hình học tập có sai số cao thành một cây học tập mạnh hơn theo kiểu tuần tự nhằm mục đích xử lý bài toán học máy có giám sát với độ tin cậy cao mà mô hình XGBoost được ứng dụng nhiều trong dự báo liên quan đến lĩnh vực khí tượng thủy văn, quản lý rủi ro thiên tai, trong đó có dự báo nước dâng do bão.

Trong nghiên cứu [20–21] đã ứng dụng mô hình học máy XGBoost để dự báo độ cao nước dâng lớn nhất cho một số khu vực ven biển Hoa Kỳ, kết quả nghiên cứu cho thấy mô hình XGBoost dự báo nước dâng bão có độ tin cậy tương đương mô hình ADCIRC. Nghiên cứu [22] sử dụng mô hình XGBoost dự báo nước dâng do bão khu vực Phúc Kiến và Quảng

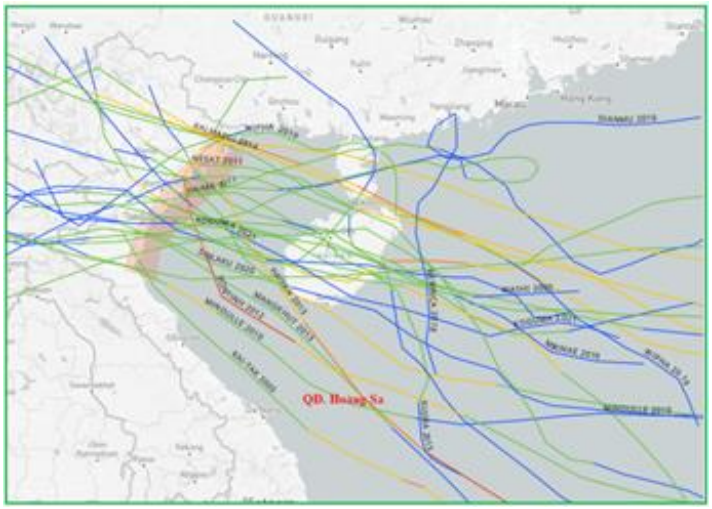
Đông, Trung Quốc cho thấy mô hình XGBoost dự báo tốt hơn so với mô hình BPNN (*Backpropagation Neural Network*) và SVR (*Support Vector Regression*). Nghiên cứu [23] sử dụng bốn mô hình XGBoost, MLR (*Multiple Linear Regression*), SVR, RF (*Rondom Forest*) để dự tính sóng leo trên bãi biển dốc cho thấy mô hình XGBoost có hiệu suất dự báo vượt trội so với ba mô hình còn lại là MLR, SVR và RF. Nhóm nghiên cứu [24] đã sử dụng mô hình GN-VSIDM là sự kết hợp của mô hình GN (*Gaussian Noise*) với mô hình VSIDM (*Diffusion Model based on the Vibrating String*) và mô hình XGBoost để ước tính thiệt hại do nước dâng bão cho thấy mô hình tổ hợp GN-VSIDM-XGBoost là mô hình ước tính thiệt hại do nước dâng bão tối ưu. Chỉ số RMSE và R^2 đánh giá mô hình GN -VSIDM- XGBoost lần lượt là 0,1089 và 0,8292. Nghiên cứu [25] đã sử dụng mô hình XGBoost để dự báo mực nước ngầm tại Selangor, Malaysia cho thấy rằng mô hình XGBoost có kết quả dự báo tốt hơn so với mô hình ANN và SVR.

Việt Nam là khu vực thường xuyên chịu ảnh hưởng của ATNĐ, bão và tác động kèm theo là nước dang do bão. Do vậy, dự báo chính xác độ cao và thời điểm xuất hiện nước dâng do bão là rất quan trọng nhằm giảm thiểu rủi ro do nước dâng bão đến cơ sở hạ tầng và thiệt hại về người. Công nghệ dự báo nước dâng do bão bằng mô hình số hiện tại đang phát triển rất mạnh, có thể nói chung là gần đến giới hạn về thuật toán mô hình. Ưu điểm của mô hình số dự báo nước dâng do bão đã được khẳng định nhưng để dự báo chi tiết thì mô hình số đòi hỏi nguồn tài nguyên tính toán lớn, điều này làm cho việc nghiên cứu áp dụng mô hình số bị giới hạn khi cần dự báo chi tiết trên phạm vi không gian rộng. Để khắc phục các vấn đề về tài nguyên tính toán khi dự báo bằng mô hình số cũng như sự hạn chế về số liệu quan trắc mực nước tại các trạm khí tượng hải văn ven biển hiện nay tại Việt Nam thì nghiên cứu và xây dựng mô hình học máy dự báo nước dâng do bão là hướng tiếp cận hợp lý. Trong nghiên cứu này, nghiên sử dụng mô hình XGBoost với các dữ liệu quan trắc khác nhau để dự báo độ cao nước dâng do bão tại Hòn Dáu, đây là trạm có số liệu quan trắc các yếu tố khí tượng, hải văn đầy đủ và dài nhất so với các trạm khí tượng hải văn trên khắc cả nước. Các giá trị dự báo của mô hình sau đó sẽ được so sánh với các giá trị thực đo để đánh giá kỹ năng của mô hình XGBoost trong nghiệp vụ dự báo nước dâng do bão. Tính mới của nghiên cứu này

2. Số liệu và phương pháp nghiên cứu

2.1. Số liệu phục vụ xây dựng mô hình XGBoost

Bộ dữ liệu đầu vào để xây dựng mô hình XGBoost dự báo nước dâng tại Hòn Dáu gồm gió, khí áp, mực nước quan trắc tại trạm Hòn Dáu, Hòn Ngư, Sơn Trà, các tham số bão trong các cơn bão ảnh hưởng trực tiếp đến trạm Hòn Dáu gồm vị trí tâm bão, khí áp tại tâm bão, vận tốc gió cực đại, và hướng di chuyển tại các bước thời gian 06 giờ. Bộ dữ liệu quan trắc và các tham số bão trong 20 năm gần đây (2002-2021) được thu thập, bao gồm 28 cơn bão có khả năng gây nước dâng tại Hòn Dáu. Quỹ đạo các cơn bão được thu thập thể hiện trên hình 1. Trong đó, 80% dữ liệu dành cho huấn luyện mô hình (training) và 20% dữ liệu dùng để kiểm định mô hình (testing). Dữ liệu trong 02 cơn bão



Hình 1. Quỹ đạo các cơn bão trong 20 năm (2002-2021) với số liệu được sử dụng xây dựng mô hình dự báo nước dâng tại trạm Hòn Dáu.

Wutip (năm 2013) và Doksuri (2017) được sử dụng để kiểm định mô hình, đây là 2 cơn bão mạnh, gây nước dâng lớn đáng kể (lớn hơn 0,5m) tại Hòn Dáu. Bộ dữ liệu trong thời gian 02 cơn bão này ảnh hưởng sẽ không tham gia vào huấn luyện mô hình để đảm bảo tính khách quan.

Mô hình XGBoost được thiết kế dự báo nước dâng do bão tại Hòn Dáu với 04 phương án như trên bảng 1. Các siêu tham số chính để thực hiện hiệu chỉnh mô hình XGBoost trong quá trình huấn luyện gồm: learning_rate, max_depth, l1_reg, l2_reg, subsample, gamma, min_child_weight, n_estimators và early_stopping. Phạm vi điều chỉnh bộ siêu tham số này được thể hiện trong bảng 2. Để điều chỉnh bộ siêu tham số này, nghiên cứu lựa chọn thư viện Optunn nhằm hỗ trợ tự động điều chỉnh tham số mô hình để mô hình có thể đạt được hiệu năng tốt nhất, kết quả thể hiện trong bảng 3.

Bảng 1. Thiết lập mô hình XGBoost cho trạm Hòn Dáu.

TT	Mô hình	Tham số đầu vào	Tham số dự báo
1	Mô hình XGBoost đơn biến	SL tại Hòn Dáu	SS
2	Mô hình XGBoost đa biến I (06 biến)	LG, LT, CAP, MWS, MV, HWS tại Hòn Dáu	SS
3	Mô hình XGBoost đa biến II (09 biến)	SL, OWS, OP, LG, LT, CAP, MWS, MV, HWS tại Hòn Dáu	SS
4	Mô hình XGBoost dữ liệu chéo (13 biến)	SL (Hòn Dáu) OWS, OP (Hòn Ngư và Sơn Trà), LG, LT, CAP, MWS, MV, HWS	SS

Trong đó SS là độ cao nước dâng dự báo tại trạm; SL là mực nước quan trắc tại trạm; OWS là vận tốc gió quan trắc tại trạm; OP là áp suất quan trắc tại trạm; CAP là khí áp tại tâm bão; MWS là vận tốc gió cực đại trong bão; LG là kinh độ (tâm bão); LT là vĩ độ (tâm bão); MV là tốc độ di chuyển của bão; HWS là hướng di chuyển của bão.

Bảng 2. Phạm vi điều chỉnh các siêu tham số của mô hình XGBoost.

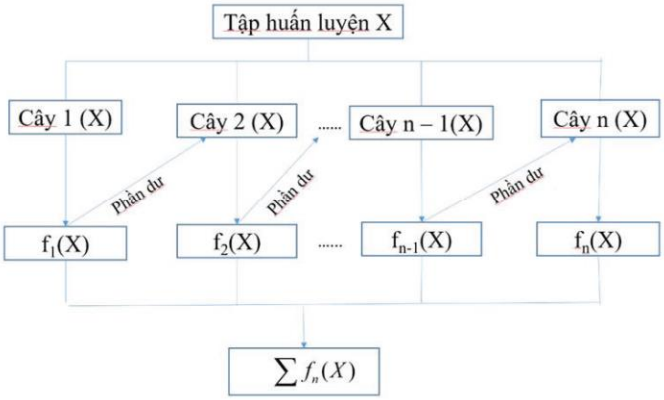
Siêu tham số	Phạm vi điều chỉnh	Siêu tham số	Phạm vi điều chỉnh
learning_rate	[0.00001 ÷ 0.50]	gamma	[1e-8 ÷ 1.0]
max_depth	[1 ÷ 20]	min_child_weight	[1e-8 ÷ 1.0]
l1_reg	[0.00001 ÷ 10.0]	n_estimators	[10 ÷ 1000]
l2_reg	[0.00001 ÷ 10.0]	early_stopping	20
subsample	[0.0 ÷ 1.0]		

Bảng 3. Bộ các siêu tham số tối ưu của mô hình XGBoost tại Hòn Dáu.

Siêu tham số	Mô hình đơn biến	Mô hình đa biến I	Mô hình đa biến II	Sử dụng chéo dữ liệu
learning_rate	0,5	0,4	0,43	0,49
max_depth	18	19	14	16
l1_reg	2,2e-05	4,6e-05	0,4	0,53
l2_reg	4,15	0,32	0,37	2,2e-03
subsample	0,94	0,88	0,91	0,98
gamma	0,00015	0,05	3,99e-05	3,2e-04
min_child_weight	0,00048	0,3e-06	2,36e-06	2,3e-04
n_estimators	644	756	666	558
early_stopping	20	20	20	20

2.2. Mô hình XGBoost

Mô hình XGBoost sử dụng thuật toán tăng cường độ dốc cấp cao, là sự mở rộng của thuật toán Gradient Tree Boosting (GTB) được đề xuất bởi Friedman [26]. Nguyên lý cơ bản được sử dụng trong mô hình XGBoost là việc kết hợp các cây mô hình học tập có độ sai số cao thành một cây mô hình học tập mạnh hơn theo kiểu tuần tự, nghĩa là đào tạo các mô hình mới tốt hơn từ việc kết hợp các mô hình yếu trước đó để bù đắp các thiếu sót trong các mô hình trước, với sơ đồ trên hình 2.



Hình 2. Sơ đồ thuật toán của mô hình XGBoost.

Mô hình XGBoost nhằm mục đích xử lý bài toán học máy có giám sát cho độ tin cậy cao. Với phương pháp học máy chuyên sâu thông thường chỉ nhận nguồn vào là dạng thô, khi đó phải quy đổi sang n-vector trong khoảng trống số thực thì XGBoost nhận nguồn dữ liệu đầu vào là dạng bảng với mọi kích cỡ dữ liệu và dạng tài liệu gồm có cả phân loại. XGBoost có tốc độ huấn luyện nhanh do có thể tính toán song song khi sử dụng tất cả các lõi CPU trong quá trình đào tạo. XGBoost là sự mở rộng của thuật toán GTB tuy nhiên kèm theo đó là những cải tiến để tối ưu thuật toán, bộ nhớ đệm của cấu trúc dữ liệu, sự kết hợp tối ưu giữa phần mềm và phần cứng nên có khả năng ứng dụng với bộ dữ liệu lớn. Thư viện của XGBoost triển khai thuật toán cây quyết định tăng cường độ dốc. Tăng cường độ dốc là một cách tiếp cận trong đó các mô hình mới được tạo ra để dự đoán phần dư hoặc sai sót của các mô hình trước đó rồi cộng lại với nhau để đưa ra dự đoán cuối cùng. XGBoost được gọi là độ dốc tăng cường vì nó sử dụng thuật toán giảm độ dốc để giảm thiểu tổn thất khi thêm mới các mô hình. Cách tiếp cận này hỗ trợ cả các vấn đề về mô hình dự đoán hồi quy và phân loại.

Lý thuyết của thuật toán như sau, giả sử rằng chúng ta có một tập huấn luyện có N mẫu:

$$X = \{X_1, X_2, X_3, \ldots X_n,\} \quad (1)$$

Với thông số đầu ra xác định là:

$$Y = \{Y_1, Y_2, Y_3, \ldots Y_n,\} \quad (2)$$

Như vậy trong thuật toán XGBoost, tại vòng lặp đầu tiên một cây học tập bất kỳ được tạo ra và ước lượng các giá trị đầu ra $f_1(X)$. Các thông số ước lượng này sẽ có sai khác với giá trị chính xác y một lượng giá trị được gọi là phần dư. Phần dư có thể hiểu là biểu thị cho sự sai số của mô hình. Muốn mô hình học tập tốt thì chúng ta phải giảm giá trị phần dư này đi. Để thực hiện điều này, cây học tập thứ 2 sẽ được thiết lập để ước lượng các giá trị của phần dư đó (không phải là giá trị y). Tương tự như trên, khi ước lượng phần dư $G_1(X)$, cây học tập thứ 2 sẽ ước lượng được giá trị $f_2(X)$ tạo ra phần dư $G_2(X)$. Để ước lượng phần dư $G_2(X)$, ta lại tiếp tục tạo ra cây học tập thứ 3. Quá trình lặp cứ tiếp tục như vậy. Cuối cùng thì giá trị ước lượng sẽ là $\sum f_n(X)$. Để nâng cao hiệu suất làm việc của mô hình XGBoost, hàm mất mát sẽ được thêm vào và có dạng sau:

$$J(£)=L(£)+Q(£) \quad (3)$$

Trong đó các tham số của mô hình được huấn luyện ký hiệu là £, L là hàm mất mát; Q là thành phần được thêm vào thường được gọi là regularization nhằm đánh giá mức độ phức tạp của mô hình. Việc thêm vào thành phần regularization giúp làm hài hòa các tham số thu được của mô hình học máy và tránh hiện tượng mô hình trở lên quá khớp (overfitting). Theo kinh nghiệm, việc sử dụng hàm mục tiêu được chuẩn hóa như trong công thức (3) sẽ giúp mô hình được lựa chọn có xu hướng sử dụng các hàm đơn giản và có thể dự đoán. Mô hình càng đơn giản sẽ cho phép tránh hiện tượng quá khớp càng tốt. Do dựa vào mô hình học tập dạng cây, giá trị ước đoán cuối cùng sẽ là:

$$y_i^n = \sum_{i=1}^{n} f_n(X_i) \quad (4)$$

Hàm mất mát ở vòng lặp thứ t có dạng:

$$J = \sum_{i=1}^{n} L(y_i, y_i^t) + \sum_{i=1}^{n} Q(f_n) \quad (5)$$

Giá trị ước lượng đầu ra y_i ở vòng lặp thứ t, y_i^t được tính như sau:

$$y_i^{()} = \sum_{i=1}^{n} f_n(X_i) = y_i^{t-1} + f_t(X_i) \quad (6)$$

Giá trị regularization $Q(f_n)$ có thể sử dụng công thức sau để xác định:

$$Q_n = \gamma^T + \frac{1}{2}\mu \sum_{i=1}^{t} W_i^2 \quad (7)$$

Trong đó γ là độ phức tạp của các lá trong cây quyết định; T là số là trong cây quyết định; μ là hệ số phóng đại hàm phạt; W là véc tơ điểm số cho các lá. Khi đó phân tích bậc 2 Taylor sẽ được sử dụng trong hàm mất mát ở thuật toán XGBoost thay thế cho phân tích bậc nhất được sử dụng trong thuật toán GTB.

Giả thiết rằng hàm tối ưu cho quá trình học tập là hàm MSE, khi đó hàm mục tiêu sẽ được viết thành:

$$J_i^{()} = \sum_{i=1}^{n} \left[g_i W_{qi} + \frac{1}{2}\left(h_i W_{q()}^2\right) \right] + \gamma^T + \frac{1}{2}\gamma \sum_{i=1}^{T} W_i^2 \quad (8)$$

Trong công thức (8), các hằng số đã được loại bỏ q() là hàm số dùng để gán dữ liệu cho là tương ứng; g_i và h_i là đạo hàm bậc nhất và bậc hai của hàm mất mát MSE. Hàm mất mát được xác định bằng tổng của các giá trị mất mát cho từng mẫu do mỗi mẫu chỉ tương ứng với 01 lá cho nên hàm mất mát có thể được xác định bằng tổng các giá trị mất mát của từng lá. Do đó, công thức (8) được viết lại như sau:

$$J^{(\ )} \approx \sum_{i=1}^{T} \left[\left(\sum_{i_{dj}} g_i \right) W_j + \frac{1}{2}\left(\sum_{i_{dj}} h_i + \gamma \right) W_j^2 \right] = \sum_{i=1}^{T} \left[G_j W_j + \frac{1}{2}\left(h_j + \gamma\right) W_j^2 \right] + \gamma^T \quad (9)$$

Khi đó, bài toán tối ưu của hàm mất mát có thể chuyển thành bài toán tìm giá trị nhỏ nhất của hàm bậc 2. Nói cách khác, sau khi phân chia một nút nhất định trong cây ra quyết định, sự thay đổi hiệu suất của mô hình có thể đánh giá dựa trên hàm mất mát. Nếu hiệu suất của mô hình được cải thiện sau khi thực hiện sự phân chia nút này, sự phân chia đó sẽ được chấp nhận, ngược lại việc tách nút sẽ dừng lại.

2.3. Phương pháp đánh giá dự báo

- Sai số quân phương trung bình (*Root Mean Square Error - RMSE*):

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(F_i - O_i\right)^2} \quad (10)$$

- Hệ số tương quan Pearson (r):

$$r = \frac{\sum_{i=1}^{n}(O_i - \bar{O})(F_i - F)}{\sqrt{\sum_{i=1}^{n}(O_i - \bar{O})^2 (F_i - \bar{F})^2}} \quad (11)$$

Trong đó N là độ dài chuỗi số liệu, F_i và O_i là các biến trong tập dữ liệu ứng với giá trị dự báo và quan trắc ở thời điểm i, $\bar{F}$ là giá trị dự báo trung bình của chuỗi F, $\bar{O}$ là giá trị quan trắc trung bình của chuỗi O.

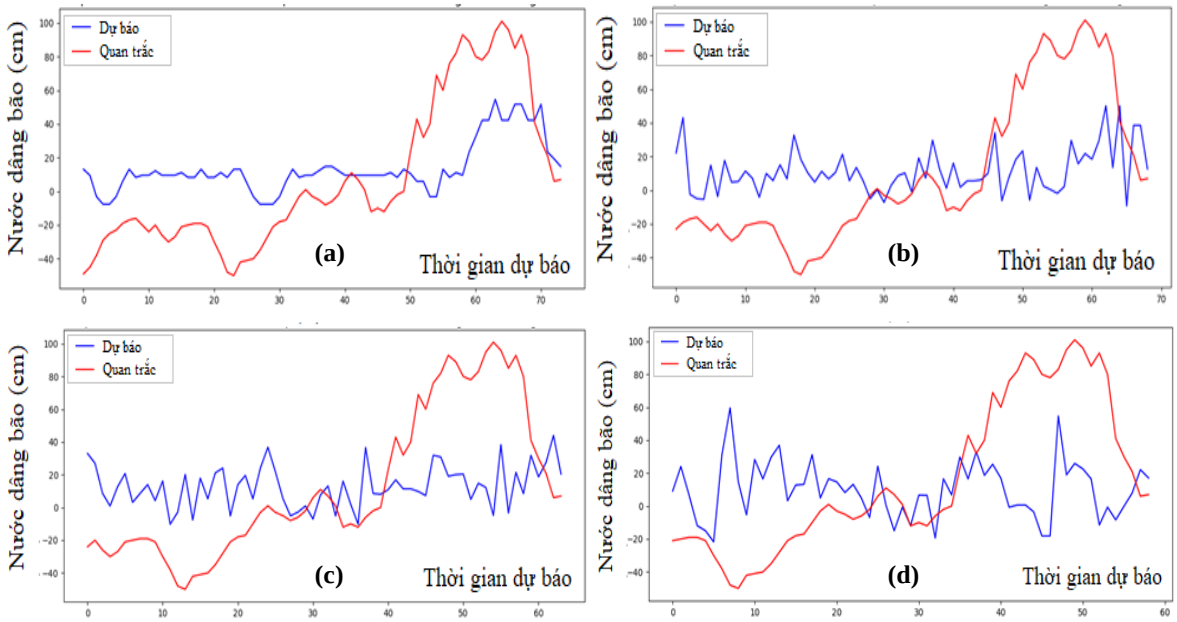
Mô hình XGBoost dự được xây dựng dựa trên dữ liệu dạng chuỗi thời gian tuần tự và liên tục, do đó cần phải được xử lý các giá trị mực nước bị khuyết thiếu. Để lấp đầy các giá

trị khuyết thiếu về mực nước tại trạm, trong nghiên cứu này sử dụng phương pháp nội suy tuyến tính.

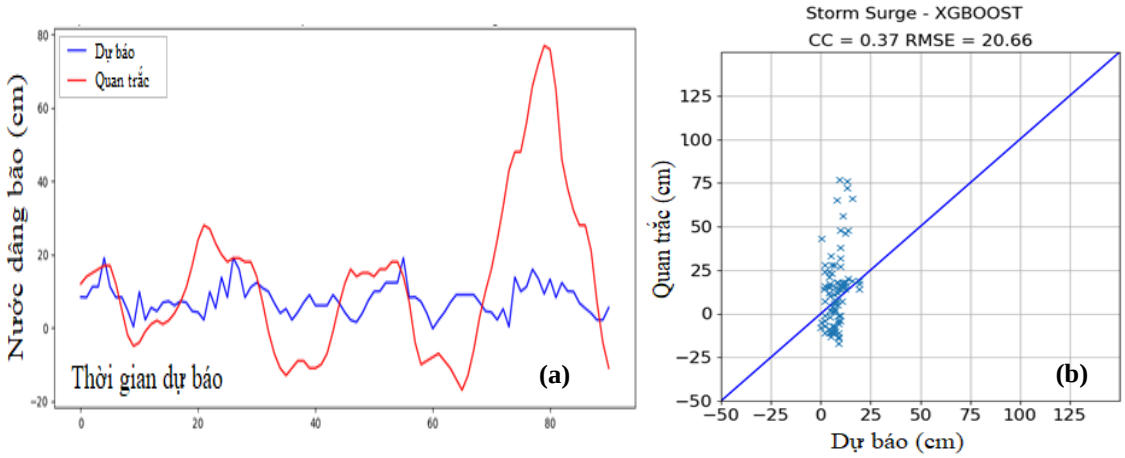
3. Kết quả và thảo luận

3.1. Mô hình XGBoost đơn biến

Đối với mô hình đơn biến (chỉ sử dụng số liệu quan trắc mực nước tại trạm), diễn biến độ cao nước dâng tính toán và thực đo tại Hòn Dáu trong bão Duksuri (9/2014) thể hiện trên hình 3 cho thấy sai số dự báo lớn ở tất cả các thời hạn dự báo, 06, 12, 18 và 24 giờ, với hệ số tương quan R và sai số bình phương trung bình RMSE, tương ứng là 0,67 (39,3 cm), 0,19 (43,1 cm), 0,19 (43,8 cm) và -0,09 (49,8 cm). Kết quả dự báo đối với bão Wutip (9/2013) cũng có sai số lớn ở tất cả các thời hạn dự báo (Hình 4). Như vậy, có thể thấy trường hợp mô hình XGBoost đơn biến chỉ sử dụng độ cao mực nước quan trắc trạm để dự báo nước dâng cho thời hạn 06, 12, 18 và 24 giờ sẽ không cho kết quả tin cậy, do bởi với chỉ độ cao mực nước quan trắc không phản ánh được xu thế nước dâng mà còn cần những yếu tố tương quan khác.



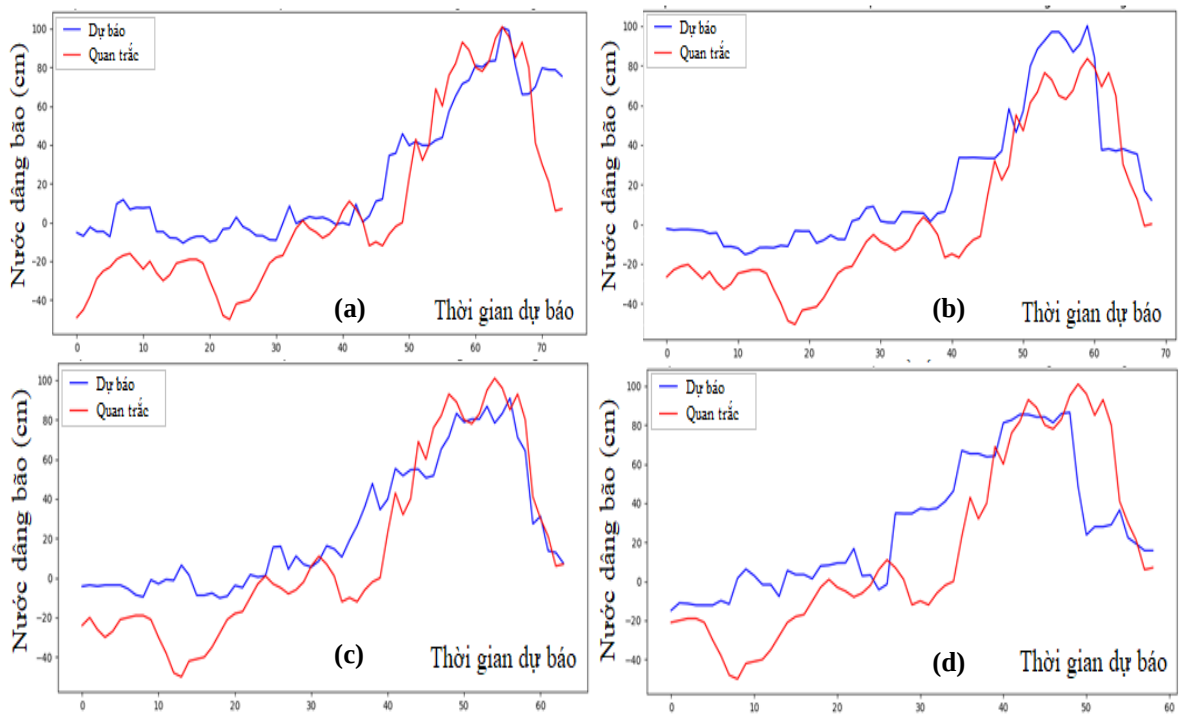
Hình 3. So sánh độ cao nước dâng dự báo bằng mô hình XGBoost đơn biến và quan trắc trong bão Duksuri (9/2017) tại Hòn Dáu tại các thời hạn dự báo: (a) 6 giờ, (b) 12 giờ, (c) 18 giờ, (d) 24 giờ.



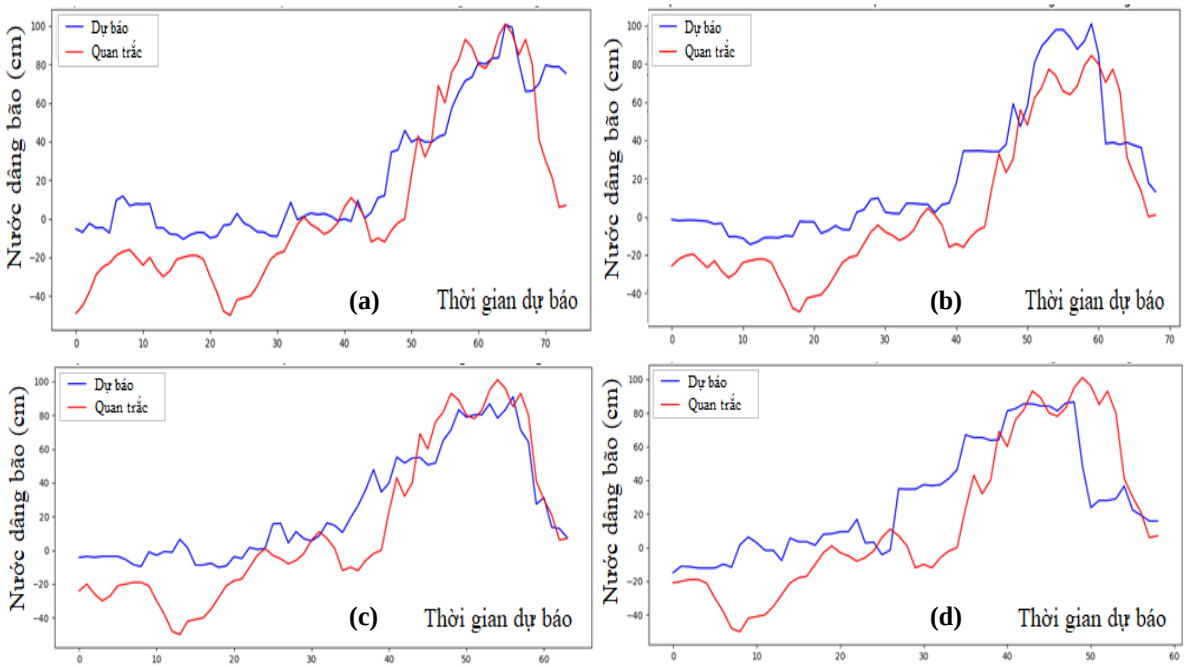
Hình 4. So sánh độ lớn nước dâng dự báo bằng mô hình XGBoost đơn biến và quan trắc tại Hòn Dáu trong bão Wutip (9/2013) thời hạn 6 giờ: (a) Diễn biến độ cao nước dâng; (b) Biểu đồ phân tán nước dâng do bão.

3.2. Mô hình XGBoost đa biến I

Đối với mô hình XGBoost đa biến I, trong trường hợp ngoài số liệu quan trắc mực nước tại trạm Hòn Dáu, các tham số bão (kinh độ và vĩ độ tâm bão; khí áp tại tâm bão; vận tốc gió cực đại; tốc độ di chuyển; hướng di chuyển). Biểu đồ so sánh độ lớn nước dâng dự báo và quan trắc tại Hòn Dáu trong bão Duksuri (9/2014) với các hạn dự báo 06, 12, 18 và 24 giờ trên hình 5 cho thấy kết quả dự báo độ lớn nước dâng đã được cải thiện. Hệ số tương quan R (và chỉ số RMSE) trong trường hợp này với thời hạn dự báo 06, 12, 18 và 24 giờ, tương ứng là 0,88 (27,0 cm), 0,88 (26,7 cm), 0,93 (24,5 cm) và 0,86 (28,6 cm).



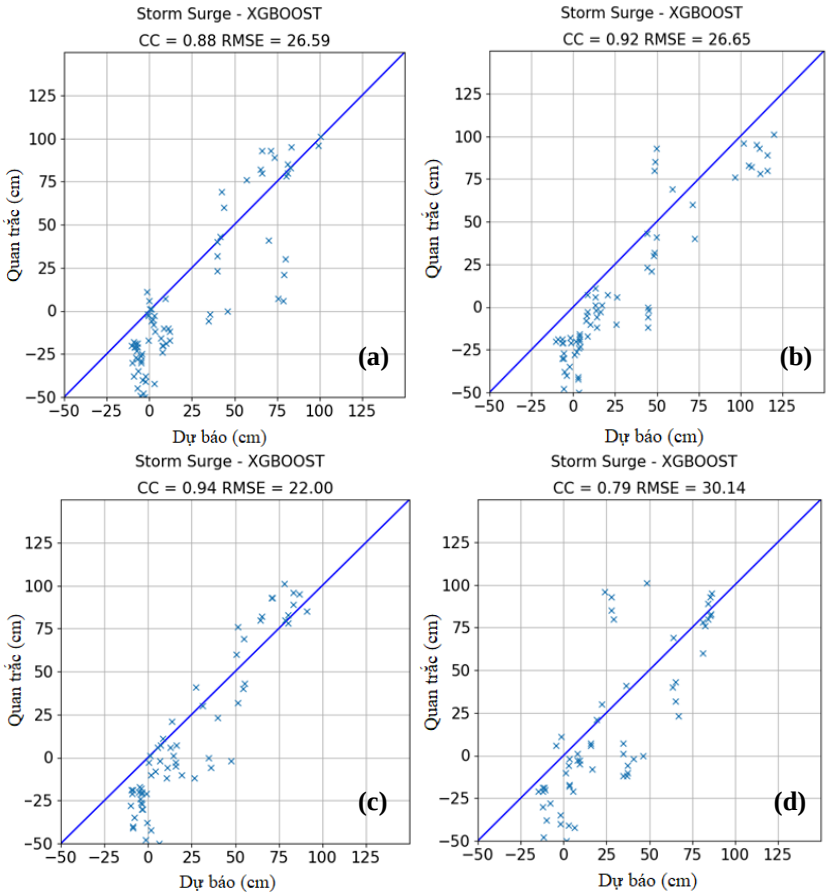
Hình 5. So sánh độ cao nước dâng dự báo bằng mô hình XGBoost đa biến I và quan trắc trong bão Duksuri (9/2017) tại Hòn Dáu theo các thời hạn dự báo: (a) 6 giờ, (b) 12 giờ, (c) 18 giờ, (d) 24 giờ.



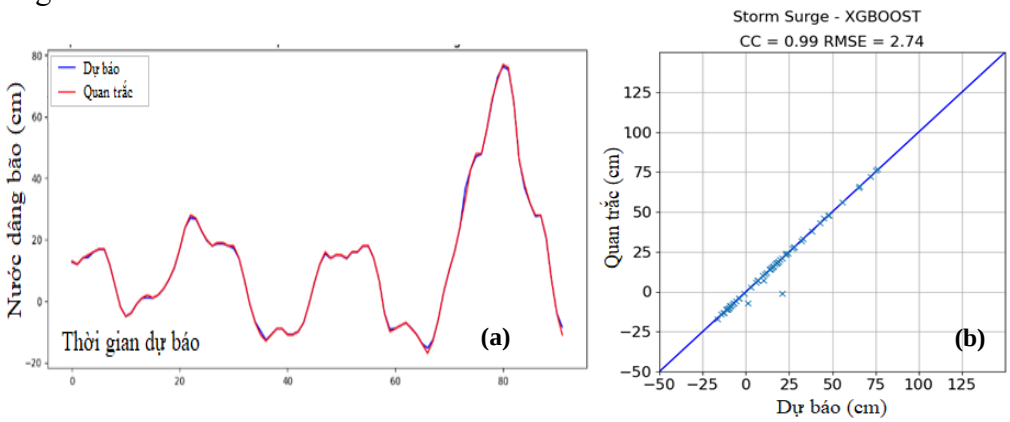
Hình 6. So sánh độ cao nước dâng dự báo bằng mô hình XGBoost đa biến II và quan trắc trong bão Duksuri (9/2017) tại Hòn Dáu theo các thời hạn dự báo, (a) 6 giờ, (b) 12 giờ, (c) 18 giờ, (d) 24 giờ.

3.3. Mô hình XGBoost đa biến II

Đối với mô hình XGBoost đa biến II, ngoài các tham số được sử dụng ở mô hình XGBoost đa biến I, đã sử dụng thêm số liệu quan trắc tại trạm, gồm mực nước quan trắc, vận tốc gió và khí áp. Kết quả dự báo và so sánh với quan trắc cho các hạn dự báo 06, 12, 18 và 24 giờ thể hiện trên hình 6, biểu đồ phân tán thể hiện trên hình 7. So sánh kết quả dự báo với mô hình XGBoost đa biến I, sai số dự báo tại thời hạn 06 giờ là tương đương, thời hạn 12 và 18 giờ có sai số nhỏ hơn, với hệ số tương quan và RMSE tương ứng là 0,92 (26, 7 cm) và 0,94 (22 cm). Tuy nhiên, sai số dự báo thời hạn 24 giờ lớn hơn kết quả của mô hình XGBoost đa biến I với R = 0,79 và RMSE = 30,14. Trong khi đó, với trường hợp bão Wutip (9/2013), kết quả so sánh giữa dự báo và quan trắc độ lớn nước dâng tại Hòn Dáu trên hình 8 cho thấy mô hình có độ tin cậy cao cả ở thời hạn 06 và 24 giờ.



Hình 7. Biểu đồ phân tán giữa độ lớn nước dâng dự báo bằng mô hình XGBoost đa biến II và quan trắc tại Hòn Dáu trong bão Duksuri (9/2017) thời hạn: 6 giờ (a), 12 giờ (b), 18 giờ (c) và 24 giờ (d).

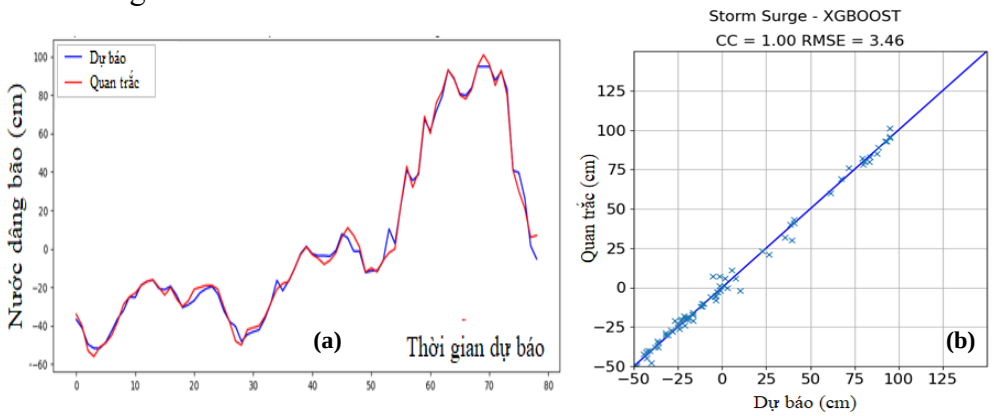


Hình 8. So sánh độ lớn nước dâng dự báo bằng mô hình XGBoost đa biến II và quan trắc tại Hòn Dáu trong bão Wutip (9/2013) thời hạn 24 giờ: (a) Diễn biến độ cao nước dâng, (b) Biểu đồ phân tán nước dâng do bão.

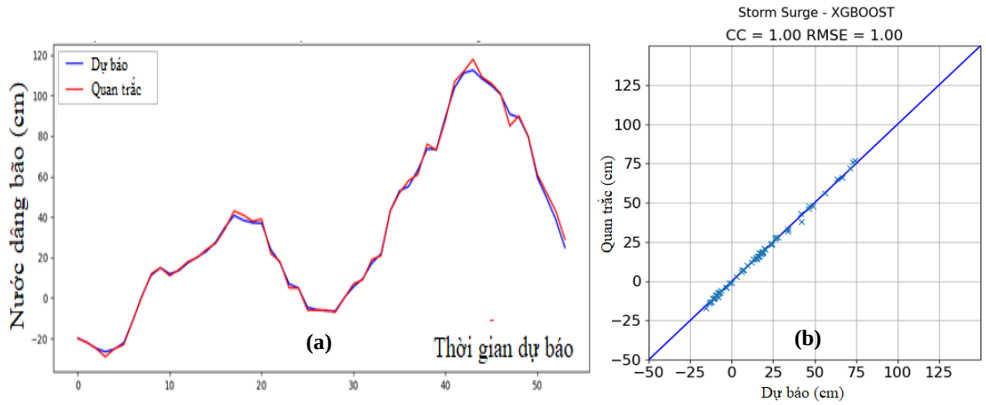
3.4. Mô hình XGBoost sử dụng số liệu chéo

Mô hình XGBoost sử dụng số liệu chéo, bao gồm mực nước quan trắc tại Hòn Dáu, vận tốc gió, khí áp quan trắc tại trạm Hòn Ngư và Sơn Trà và các tham số bão như trường hợp

Mô hình XGBoost đơn biến I và II. Kết quả dự báo độ cao nước dâng thời hạn 24 giờ với bão Duksiri (9/2017) và Wutip (9/2013) thể hiện trên hình 9 và 10 tương ứng. Có thể thấy rằng với mô hình sử dụng số liệu chéo có độ tin cậy với thời hạn dự báo 24 giờ cao, sai số RMSE chỉ khoảng từ 1-3 cm.



Hình 9. So sánh độ lớn nước dâng dự báo bằng Mô hình XGBoost dữ liệu chéo và quan trắc tại Hòn Dáu trong bão Duksuri (9/2017) thời hạn 24 giờ: (a) diễn biến độ cao nước dâng, (b) Biểu đồ phân tán nước dâng do bão.



Hình 10. So sánh độ lớn nước dâng dự báo bằng Mô hình XGBoost dữ liệu chéo và quan trắc tại Hòn Dáu trong bão Wutip (9/2013) thời hạn 24 giờ: (a) diễn biến độ cao nước dâng, (b) Biểu đồ phân tán nước dâng do bão.

Qua kết quả phân tích ở trên cho thấy, ngoài mô hình XGBoost đơn biến, các mô hình XGBoost đa biến I, mô hình XGBoost đa biến II và mô hình XGBoost sử dụng dữ liệu chéo đều có thể ứng dụng để dự báo nước dâng do bão cho thời hạn 06, 12, 18 và 24 giờ, độ tin cậy quả dự báo phần lớn đạt trên 80%, tương đương và cao hơn các mô hình dự báo số trị đang dự báo nghiệp vụ tại Việt Nam. Do vậy, tuỳ thuộc vào hiện trạng số liệu quan trắc ở thời điểm dự báo, dự báo viên có thể lựa chọn mô hình phù hợp để dự báo nước dâng do bão tại trạm Hòn Dáu.

4. Kết luận

Trong nghiên cứu này, mô hình XGBoost được ứng dụng để xây dựng công cụ dự báo nước dâng do bão tại trạm Hòn Dáu cho các hạn dự báo 06, 12, 18 và 24 giờ. Mô hình XGBoost được thiết lập với 04 phương án sử dụng dữ liệu khác nhau (04 mô hình) với tên gọi: Mô hình XGBoost đơn biến, mô hình XGBoost đa biến I, mô hình XGBoost đa biến II và mô hình XGBoost sử dụng dữ liệu chéo (sử dụng số liệu quan trắc gió, khí áp tại Hòn Ngư và Sơn Trà). Bộ dữ liệu độ cao mực nước quan trắc tại trạm Hòn Dáu, vận tốc gió, khí áp quan trắc tại trạm Hòn Dáu, Hòn Ngư và Sơn Trà, các tham số bão (vị trí tâm bão, khí áp tại tâm bão, vận tốc gió cực đại, tốc độ di chuyển và hướng di chuyển) trong 28 cơn bão ảnh hưởng tới khu vực trạm Hòn Dáu giai đoạn 2002-2021 được thu thập để xây dựng và kiểm

định mô hình XGBoost. Trong đó, khoảng 80% dữ liệu được đưa vào huấn luyện và 20% dữ liệu, tương ứng với số liệu trong 02 cơn bão (Duksuri, tháng 9/2017 và Wutip, tháng 9/2013) dùng để kiểm định các mô hình. Kết quả cho thấy, mô hình XGBoost đơn biến cho độ tin cậy thấp ở tất cả các thời hạn dự báo. Trong khi đó, hai mô hình XGBoost đa biến và mô hình sử dụng dữ liệu chéo đều cho kết quả tin cậy cao, với phần lớn hệ số tương quan giữa dự báo và quan trắc đều trên 80%. Trong đó, mô hình XGBoost sử dụng dữ liệu chéo cho độ tin cậy cao nhất với thời hạn dự báo 24 giờ. Tuy nhiên, các mô hình được xây dựng cần tiếp tục được thử nghiệm thêm với nhiều cơn bão trước khi triển khai ứng dụng trong dự báo nghiệp vụ. Ngoài ra, ứng dụng mô hình XGBoost vào xây dựng công cụ dự báo nước dâng do bão ở các trạm khí tượng hải văn khác cũng cần triển khai thực hiện để nâng cao năng lực dự báo hải văn tại Việt Nam.

Đóng góp của tác giả: Xây dựng ý tưởng nghiên cứu: B.M.H., N.B.T.; Thu thập, xử lý số liệu: P.K.N., B.M.H., P.V.T.; Phân tích kết quả: N.B.T., P.K.N., B.M.H.; Viết bản thảo bài báo: B.M.H., N.B.T.; Chỉnh sửa bài báo: B.M.H., N.B.T., P.V.T.

Lời cảm ơn: Nghiên cứu này được tài trợ bởi đề tài nghiên cứu khoa học cấp Bộ Tài nguyên và Môi trường mã số: TNMT.2022.06.04. Tập thể tác giả xin chân thành cảm ơn.

Lời cam đoan: Tập thể tác giả cam đoan bài báo này là công trình nghiên cứu của tập thể tác giả, chưa được công bố ở đâu, không được sao chép từ những nghiên cứu trước đây; không có sự tranh chấp lợi ích trong nhóm tác giả.

Tài liệu tham khảo

1. Ca, V.T.; Hiếu, P.Đ.; Hiển, N.X.; Đạo, N.X. Mô hình dự báo nước dâng do bão có tính đến thủy triều. *Tạp chí Khí tượng Thủy văn* **2008**, *568*, 25–33.
2. Ou, S.H.; Liao, J.M.; Tsai, C.Y.; Hsu, T.W. Numerical studies of typhoon-induced storm surge using POM and finite element depth-averaged model in Taiwan. Proceedings 4th Chinese-German Joint Symposium on Hydraulic and Ocean Engineering, Darmstadt, Germany, 2008.
3. Thủy, N.B.; Ngọc, P.K.; Tiến, D.Đ.; Tiến, T.Q.; Hole, L.R.; Kristensen, N.M.; Röhrs, J. Mô hình Roms 2D dự báo nước dâng do bão và gió mùa tại Việt Nam. *Tạp chí Khí tượng Thủy văn* **2016**, *665*, 36–40.
4. Qin, G.; Fang, Z.; Zhao, S.; Meng, Y.; Sun, W.; Yang, G.; Wang, L.; Feng, T. Storm Surge Inundation Modulated by Typhoon Intensities and Tracks: Simulations Using the Regional Ocean Modeling System (ROMS). *J. Mar. Sci. Eng.* **2023**, *11*, 1112. https://doi.org/10.3390.
5. Li, Z.; Li, S.; Hu, P.; Mo, D.; Li, J.; Du, M.; Yan, J.; Hou, Y.; Yin, B. Numerical study of storm surge-induced coastal inundation in Laizhou Bay, China. *Phys. Oceanogr.* **2022**, *9*, 1–14. https://doi.org/10.3389/fmars.2022.952406
6. Taflanidis, A.A.; Kennedy, A.B.; Westerink, J.J.; Smith, J.; Cheung, K.F.; Hope, M.; Tanaka, S. Rapid assessment of wave and surge risk during landfalling hurricanes: Probabilistic approach. *J. Waterway Port Coastal Ocean Eng.* **2012**, *139(3)*, 171–182. https://doi.org/10.1061/(ASCE)WW.1943-5460.0000178.
7. Chiến, Đ.Đ.; Thủy, N.B.; Sáo, N.T.; Thái, T.H.; Kim, S. Nghiên cứu tương tác sóng và nước dâng do bão bằng mô hình số trị. *Tạp chí Khí tượng Thủy văn* **2014**, *647*, 19–24.
8. Cường, H.Đ.; Thủy, N.B.; Hưởng, N.V.; Tiến, D.Đ. Đánh giá nguy cơ bão và nước dâng do bão tại ven biển Việt Nam. *Tạp chí Khí tượng Thủy văn* **2018**, *684*, 29–36.
9. Thái, T.H.; Trí, Đ.Q.; Hoàng, Đ.V. Nghiên cứu mô phỏng tác động của sóng và nước dâng bão khu vực ven biển miền Trung. *Tạp chí Khí tượng Thủy văn* **2018**, *687*, 1–14.

10. Thủy, N.B. Nghiên cứu lựa chọn mô hình dự báo nghiệp vụ nước dâng do bão vào dự báo nghiệp vụ tại Việt Nam. Báo cáo tổng kết đề tài nghiên cứu khoa học cấp Bộ. Hà Nội, 2016.
11. Cát, V.M.; Lân, V.V. Mô phỏng nước dâng do bão và xây dựng bản đồ ngập lụt đảo Phú Quốc. *Tạp chí Khoa học Kỹ thuật thủy lợi và môi trường* **2017**, *56*, 16–23.
12. Hiển, N.X. Nghiên cứu nước dâng do bão có tính đến ảnh hưởng của sóng và áp dụng cho vùng ven biển Hải Phòng. Luận án tiến sĩ. Hà Nội, 2013.
13. Sztobryn, M. Forecast of storm surge by means of artificial neural network. *J. Sea Res.* **2003**, *49(4)*, 317–322.
14. You, S.; Seo, J.W. Storm surge prediction using an artificial neural network model and cluster analysis. *Geology Nat. Hazards* **2009**, 53996115.
15. Lee, T.L. 2006. Neural network prediction of a storm surge. *Ocean Eng.* **2006**, *33(3)*, 483–494.
16. Lee, T.L. Predictions of typhoon storm surge in Taiwan using artificial neural networks. *Ocean Eng.* **2009**, *40(11)*, 1200–1206.
17. Kim, S.; Matsumi, Y.; Pan, S.; Mase, H. 2016. A real-time forecast model using artificial neural network for after- runner storm surges on the Tottori coast, Japan. *Ocean Eng.* **2016**, *122(6)*, 44–53.
18. Kim, S.; Pan, S.; Mase, H. Artificial neural network-based storm surge forecast model: Practical application to Sakai Minato, Japan. *J. Ocean Res.* **2019**, 199134715.
19. Chao, W.T.; Young, C.C.; Hsu, T.W.; Liu, W.C. Long-lead-time prediction of storm surge using artificial neural networks and effective typhoon parameters: revisit and deeper insight. *J. Ocean Res.* **2020**, *12(9)*, 2394.
20. Pacheva, B.; Arorab, P.; del-Castillo-Negretea, C.; Valsetha, E.; Dawsona, C. A framework for flexible peak storm surge prediction. *Coastal Eng.* **2023**, 1–31.
21. Pacheva, B.; Valsetha, E.; Dawsona, C. Learning storm surge with gradient boosting. *Ocean Modell.* **2022**, 1–14.
22. Sun, H.; Wang, J.; Ye, W. A Data Augmentation-Based Evaluation System for Regional Direct Economic Losses of Storm Surge Disasters. *Int. J. Environ. Res. Public Health* **2021**, *18*, 2918.
23. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, Aug. 2016, pp. 785–794.
24. Du, X.; Li, X.; Zhang, S.; Zhao, T.; Hou, Q.; Jin, X.; Zhan, J. High-accuracy estimation method of typhoon storm surge disaster loss under small sample conditions by information diffusion model coupled with machine learning models. *Int. J. Disaster Risk Reduct.* **2022**, *82*, 103307.
25. Osman, A.I.A.; Ahmed, A.N.; Chow, M.F.; Huang, Y.F.; El-Shafie, A. Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. *Ain Shams Eng. J.* **2021**, *12(2)*, 1545–1556.
26. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **2001**, *29(5)*, 1189–1232. https://doi.org/10.1214/aos/1013203451.

Initial results of testing the XGBoost algorithm to predict storm surge at Hon Dau station

Bui Manh Ha[1], Nguyen Ba Thuy[1]*, Pham Khanh Ngoc[1], Pham Van Tien[2]

[1] National Center for Hydro-Meteorological Forecasting; manhamhc@gmail.com; thuybanguyen@gmail.com; ngocpkchibo@gmail.com
[2] Institute of Meteorology, Hydrology and Climate Change; phamvantien@gmail.com

Abstract: In this study, the high-level gradient boosting algorithm XGBoost (Extreme Gradient Boosting, hereinafter referred to as the XGBoost model) is applied to build a storm surge forecasting tool at Hon Dau. The XGBoost model is built with 4 different data usage options (04 models): univariate XGBoost model, multivariate XGBoost model I, multivariate XGBoost model II and XGBoost model using cross-sectional data. A data set of 28 storms affecting Hon Dau station in the period 2002-2021 was collected to build models and test forecast results. Test results of the XGBoost model predicting storm surges show that the univariate XGBoos model has low reliability at all forecast horizons. Meanwhile, the two multivariate XGBoos models and the model using cross-sectional data both give highly reliable results, with most correlation coefficients between forecasts and observations above 80%. The results of the study serve as the basis for selecting storm surge forecasting tools at Hon Dau depending on the current status of meteorological and oceanographic monitoring data.

Keywords: Storm surge forecasting; XGBoost; Machine Learning; AI.