

ỨNG DỤNG MÔ HÌNH HỌC MÁY XGBOOST VÀ LIGHTGBM TRONG VIỆC DỰ BÁO MỨC NƯỚC TRIỀU TRÊN SÔNG SÀI GÒN - ĐỒNG NAI

Đặng Đồng Nguyên, Lê Thị Hòa Bình,
Phùng Tấn Phương, Phạm Hồng Đức

Phân hiệu trường Đại học Thủy lợi tại tỉnh Bình Dương

Tóm tắt: Trong những năm gần đây, trí tuệ nhân tạo (AI) đã được sử dụng rộng rãi, thay thế dần cho các mô hình thủy động lực học trong việc dự báo mực nước, lưu lượng trên các sông nhằm cảnh báo sớm nguy cơ lũ lụt. Nghiên cứu này áp dụng một số mô hình học máy để dự báo mực nước tại vị trí các trạm quan trắc trên hệ thống sông Sài Gòn – Đồng Nai. Kết quả của nghiên cứu chỉ ra rằng có xu thế gia tăng đáng kể dữ liệu mực nước tại Nhà Bè, Phú An và Thủ Dầu Một trong khi đó xu thế gia tăng không đáng kể được ghi nhận tại trạm Vũng Tàu. Mô hình XGBoost và LightGBM được đánh giá có độ tin cậy cao để dự báo mực nước tại 4 trạm đo mực nước. Ngoài ra đề tài cũng tích hợp mô hình XGBoost và LightGBM vào trong nền tảng website cung cấp thông tin về mực nước triều dự báo của cả 2 mô hình trong 10 ngày tiếp theo tại các trạm. Kết quả dự báo mực nước triều cung cấp những thông tin hữu ích trong việc phòng tránh ngập úng cho đô thị ven sông Sài Gòn – Đồng Nai cũng như là vận hành công trình thủy phục vụ cho nông nghiệp và giao thông thủy.

Từ khóa: LightGBM, Machine Learning, Mực nước, Sài Gòn – Đồng Nai, Thủy triều, XGBoost

Summary: In recent years, artificial intelligence has been widely used, replacing hydrodynamic models in forecasting water levels, flows on rivers to warn of the risk of flooding. In this research, several machine learning models are applied to forecast water levels at the locations of monitoring stations on the Sai Gon-Dong Nai river. The results of the study show that there is a significant increase in water level at Nha Be, Phu An and Thu Dau Mot while the insignificant increase trend is recorded at Vung Tau station. The XGBoost and LightGBM models are evaluated to have high reliability to forecast water levels at 4 water level measurement stations. In addition, the research also integrates the XGBoost and LightGBM models into the website platform providing information on the predicted tide level of both models in the next 10 days at the stations. The results of the tide level forecast provide useful information in preventing flooding for the urban areas along the Sai Gon-Dong Nai river as well as operating water resources system for agriculture and water transport.

Keywords: LightGBM, Machine Learning, Sai Gon-Dong Nai river, Tidal level, Water level, XGBoost

1. ĐẶT VẤN ĐỀ

Hệ thống sông Sài Gòn – Đồng Nai là hệ thống sông lớn thứ hai ở Nam Bộ, chỉ đứng

sau sông Cửu Long. Hạ lưu của lưu vực này bao trùm toàn bộ khu kinh tế trọng điểm phía Nam với nhiều đô thị lớn và khu công nghiệp. Mặc dù cơ sở hạ tầng và hệ thống tiêu thoát nước đã được nâng cấp, đầu tư khá lớn, tuy nhiên, tình hình ngập lụt vẫn diễn ra trên các khu đô thị của khu vực hạ lưu sông Sài Gòn –

Ngày nhận bài: 03/12/2023

Ngày thông qua phản biện: 10/01/2024

Ngày duyệt đăng: 05/02/2024

Đồng Nai, xu thế này được đánh giá gia tăng khá nhiều trong những năm gần đây. Các nhà khoa học và các tổ chức đã nghiên cứu thực trạng ngập úng ở hạ lưu sông Sài Gòn – Đồng Nai và đưa ra một số nguyên nhân chính đó là sự gia tăng mưa cực đoan, lũ từ các hồ chứa thượng nguồn đổ về, sụt lún đất, thủy triều biển Đông cao cộng hưởng với quá trình đô thị hóa nhanh trên diện rộng [1-4].

Khu vực hạ lưu sông Sài Gòn – Đồng Nai có địa hình tương đối thấp, hệ thống kênh rạch chằng chịt và dễ bị ảnh hưởng bởi mực nước biển dâng. Nghiên cứu của Việt [5] đánh giá sự thay đổi mực nước trên hệ thống sông Sài Gòn – Đồng Nai cho thấy rằng biên độ mực nước giữa cấp tần suất 0,1% và 99,9% tại trạm Vũng Tàu tăng 7 cm trong giai đoạn 1980-2014, cùng với mức biến dạng thủy triều lớn nhất xảy ra được ghi nhận tại các trạm Nhà Bè và Phú An. Một nghiên cứu khác của Giang, Quang [6] chỉ ra rằng hầu hết các trạm mực nước hạ lưu sông Sài Gòn -Đồng Nai có sự gia tăng từ 0.17 đến 1.8 cm/năm và các yếu tố ảnh hưởng đến sự gia tăng này chủ yếu đến từ quá trình đô thị hóa và vận hành công trình thủy lợi của khu vực nghiên cứu. Trong khi đó, nghiên cứu của nhóm tác giả Thủy và Tiến [7] về nước dâng trong các đợt triều cường tại ven biển Đông Nam Bộ kết luận rằng hầu hết các tháng 1, 2, 10, 11 và 12 đều xuất hiện mực nước lớn nhất cao hơn 4,0 m, trong đó nước dâng lớn trên 40 cm chủ yếu xuất hiện trong tháng 10 và 11.

Việc ứng dụng mô hình thủy văn, thủy lực trong mô phỏng dòng chảy và dự báo mực nước sông, mực nước triều, cảnh báo ngập lụt cho vùng hạ lưu sông Sài Gòn – Đồng Nai cũng đã được thực hiện và phát triển bởi các nhà nghiên cứu trong nước. Các mô hình phổ biến được sử dụng cho công việc này có thể kể đến như MIKE, SWMM, UTIDE [8-11]. Tuy

nhiên, các mô hình này đòi hỏi một lượng lớn dữ liệu đầu vào như lượng mưa, địa hình, mặt cắt ngang lòng sông, các công trình cầu cống trên hệ thống sông. Bên cạnh đó, mô hình cần trải qua các bước hiệu chỉnh và kiểm định dựa trên các tài liệu thực đo, điều này mất khá nhiều thời gian và đòi hỏi kinh nghiệm của người thiết lập mô hình. Chính vì thế, vài năm trở lại đây, trí tuệ nhân tạo (Artificial Intelligence – AI) đã được sử dụng rộng rãi, thay thế dần cho các mô hình thủy động lực học trong việc dự báo mực nước, lưu lượng trên các sông nhằm cảnh báo sớm nguy cơ lũ lụt [12, 13]. Điển hình như nhóm tác giả Toàn và Nhân [14] đã áp dụng mô hình học máy (Machine learning) vào mô phỏng dòng chảy trên lưu vực sông Sài Gòn – Đồng Nai. Nhóm tác giả Phan và Nguyen [15] đã kết hợp mô hình thống kê và mô hình học máy để dự báo mực nước tại các trạm quan trắc trên sông Hồng. Tương tự, các ứng dụng của mô hình học máy vào việc dự báo mực nước cũng được tìm thấy trong các nghiên cứu tại Ba Lan, Hàn Quốc, Campuchia [16-18]. Mục đích của bài báo này là áp dụng mô hình học máy để dự báo mực nước tại vị trí các trạm quan trắc trên hệ thống sông Sài Gòn – Đồng Nai, bao gồm trạm Thủ Dầu Một, Phú An, Nhà Bè và Vũng Tàu.

2. SỐ LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Số liệu thu thập

Trong nghiên cứu này, số liệu mực nước từ năm 1988 đến 2014 tại 4 trạm đo Thủ Dầu Một (Bình Dương), Phú An, Nhà Bè (Thành phố Hồ Chí Minh) và Vũng Tàu (Bà Rịa - Vũng Tàu) sẽ được sử dụng để phân tích tính toán.

2.2. Phương pháp nghiên cứu

Kiểm định Mann-Kendall

Kiểm định phi tham số Mann-Kendall [19, 20], thường được sử dụng rộng rãi để phân tích các xu hướng đơn điệu trong chuỗi dữ liệu. Kết quả từ kiểm định Mann-Kendall cho biết giá trị của Tau (Tau value), nó cho biết chuỗi số liệu có xu hướng tăng hay giảm. Với giá trị $Tau > 0$, chuỗi số liệu thể hiện xu thế tăng, ngược lại khi $Tau < 0$, chuỗi số liệu thể hiện xu thế giảm.

Mô hình XGBoost và LightGBM

Mô hình XGBoost [21] là một trong những mô hình học máy được sử dụng phổ biến trong việc dự đoán và phân loại. XGBoost là viết tắt của Extreme Gradient Boosting Regressor, được xây dựng dựa trên kỹ thuật Gradient Boosting. Gradient Boosting là một phương pháp học máy tập trung vào việc xây dựng một loạt các mô hình dự đoán đơn giản, mỗi mô hình được tạo ra bằng cách tìm kiếm các trọng số tối ưu của các đặc trưng đầu vào để tối thiểu hóa sai số. LightGBM [22] là một thuật toán được phát triển bởi Microsoft Research Asia dựa trên phương pháp cây quyết định tăng

cường (Gradient Boosting Decision Tree). Mô hình này có một số ưu điểm như tốc độ huấn luyện và hiệu quả tính toán cao, có thể sử dụng được với các bài toán dự đoán có số lượng dữ liệu lớn với độ chính xác cao.

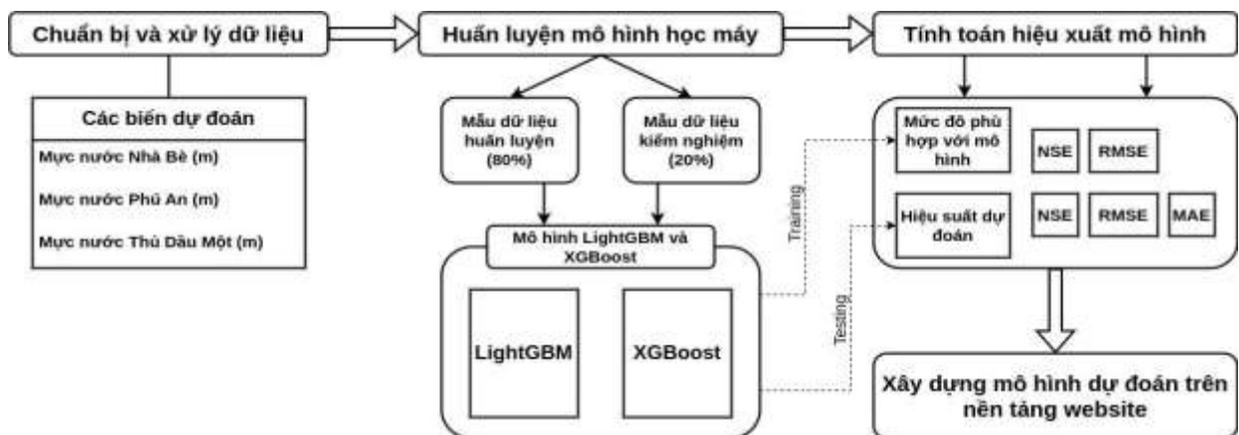
Trong nghiên cứu này, kỹ thuật dự báo mực nước triều trên sông Sài Gòn – Đồng Nai được mô tả qua 04 bước sau:

Bước 1: Chuẩn bị và xử lý dữ liệu

Các dữ liệu về mực nước triều thời đoạn giờ từ năm 1988 đến 2014 được thu thập từ Đài Khí tượng thủy văn Nam bộ. Dữ liệu mực nước triều tại Vũng Tàu trong 10 ngày được sử dụng từ mô hình dự báo của Glass [23]. Tiền xử lý dữ liệu (trend, outlier, missing) để tiến hành trước khi xây dựng mô hình dự đoán mực nước trên lưu vực sông Sài Gòn – Đồng Nai.

Bước 2: Huấn luyện mô hình học máy

Sau khi có bộ dữ liệu được chuẩn hóa, dữ liệu được chia thành 2 tập huấn luyện và kiểm tra với tỷ lệ 80/20 một cách ngẫu nhiên.



Hình 1: Sơ đồ khối mô tả ứng dụng mô hình XGBoost và LightGBM để dự báo mực nước triều

Bước 3: Đánh giá mô hình

Việc đánh giá mô hình được thực dựa trên các chỉ số sau: Root Mean Squared Error (RMSE) – sai số bình phương gốc là độ lệch chuẩn của lỗi dự đoán, cho biết mức độ tập trung dữ liệu

xung quanh dòng phù hợp nhất. RMSE được sử dụng trong các mô hình học máy dự báo để xác minh kết quả. RMSE càng nhỏ, mức độ chính xác càng tốt. Giá trị của RMSE được tính theo công thức:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_{i-true} - y_{i-pred})^2} \quad (1)$$

Mean Absolute Error (MAE) - sai lệch giữa giá trị dự đoán và giá trị thực tế trong các bài toán dự đoán hoặc hồi quy MAE đo lường sự tương tự trung bình giữa giá trị dự đoán và giá trị thực tế dựa trên giá trị tuyệt đối của sai lệch giữa chúng. MAE càng thấp, mô hình càng chính xác trong dự đoán giá trị. Công thức tính MAE như sau:

$$MAE = \frac{\sum_{i=1}^N |y_{i-true} - y_{i-pred}|}{N} \quad (2)$$

Nash-Sutcliffe (Coefficient of Efficiency hoặc NSE) là một phương pháp thống kê được sử dụng để đánh giá hiệu suất của mô hình dự đoán trong mô phỏng thủy văn và các lĩnh vực liên quan khác. NSE thường có giá trị trong khoảng $[-\infty, 1]$, và mô hình tốt nhất có NSE gần bằng 1. Công thức tính chỉ số NSE như sau:

$$NSE = 1 - \frac{\sum_{i=1}^N (y_{i-true} - y_{i-pred})^2}{\sum_{i=1}^N (y_{i-true} - y_{average})^2} \quad (3)$$

Trong đó: y_{i-true} là giá trị thực tế, y_{i-pred} là giá trị dự đoán, $y_{average}$ là giá trị trung bình, N là tổng số lượng mẫu.

Bước 4: Triển khai mô hình dự đoán trên nền tảng Website

Để triển khai mô hình dự báo mực nước triều, đề tài sử dụng các kỹ thuật như sau: Phần ngôn ngữ lập trình sử dụng Python, JavaScript. Framework sử dụng Reactjs và Google Maps API. Cơ sở dữ liệu sử dụng MongoDB. Phần giao diện được thiết kế để có thể tương tác và sử dụng các tính năng của website một cách dễ dàng và thuận tiện và đặc biệt website đảm bảo hoạt động ổn định, bảo mật.

3. KẾT QUẢ NGHIÊN CỨU

Dựa vào kết quả kiểm định Mann-Kendall, có thể kết luận rằng các trạm Nhà Bè, Phú An và Thủ Dầu Một đều có xu hướng gia tăng rõ rệt trong chuỗi dữ liệu mực nước, trong khi đó trạm Vũng Tàu không có xu hướng tăng đáng kể. Giá trị p và Tau của các trạm đo mực nước được thể hiện ở Bảng 1.

Bảng 1: Kết quả thể hiện giá trị p và Tau tại các trạm mực nước

STT	Trạm	$p.value$	Tau
1	Vũng Tàu	6.555168×10^{-2}	0.253261
2	Nhà Bè	1.803432×10^{-7}	0.715100
3	Phú An	1.591409×10^{-9}	0.826211
4	Thủ Dầu Một	8.039722×10^{-10}	0.840456

Trạm Vũng Tàu có giá trị $p = 0.06555$ và $Tau = 0.25356$. Giá trị p lớn hơn mức ý nghĩa thống kê 0.05, cho thấy không có xu hướng đáng kể trong dữ liệu tại trạm này. Trong khi đó, các trạm Nhà Bè, Phú An và Thủ Dầu Một

ghi nhận giá trị p rất nhỏ. Giá trị p nhỏ hơn mức ý nghĩa thống kê 0.05, cho thấy có xu hướng gia tăng đáng kể trong chuỗi dữ liệu mực nước triều tại các trạm này. Ví dụ, tại trạm Phú An có giá trị p xấp xỉ là $1.591409 \times$

10^{-9} trong khi đó trạm Thủ Dầu Một là 8.039722×10^{-10} .

Giá trị của các chỉ số MAE, RMSE và NSE

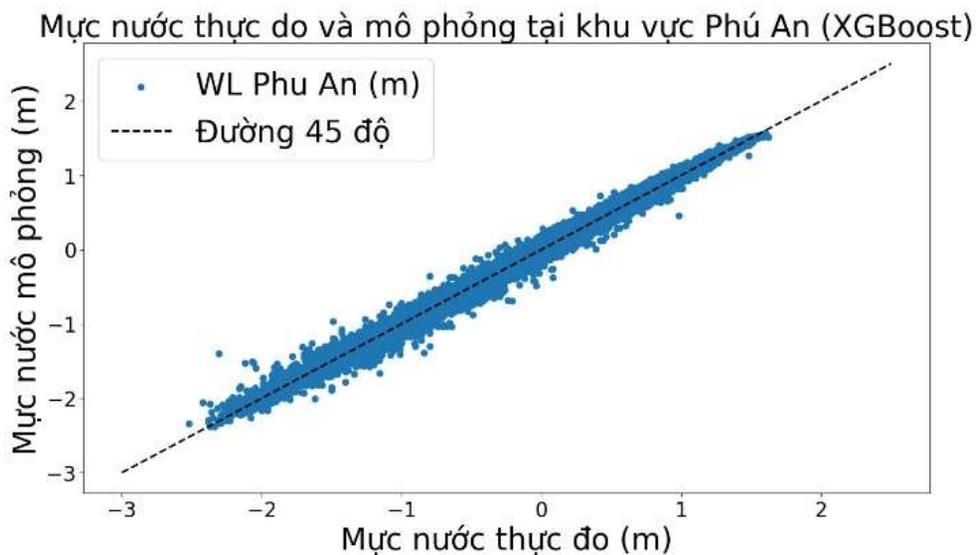
được sử dụng để đánh giá sai số của mô hình XGBoost và LightGBM được hiển thị theo Bảng 2.

Bảng 2: Kết quả đánh giá sai số mô hình

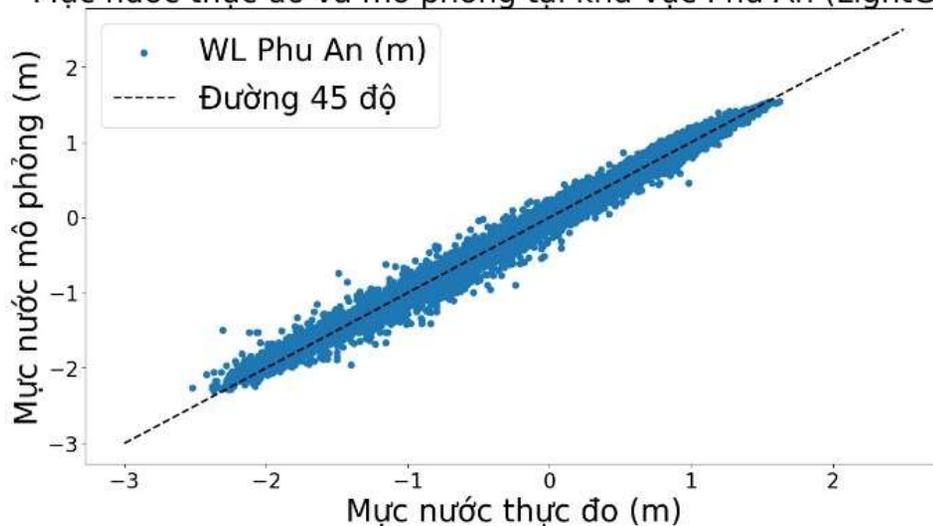
Mô hình	MAE	RMSE	NSE
Trạm Nhà Bè			
XGBoost	0.05	0.06	0.97
LightGBM	0.05	0.07	0.95
Trạm Phú An			
XGBoost	0.03	0.05	0.96
LightGBM	0.04	0.06	0.97
Trạm Thủ Dầu Một			
XGBoost	0.03	0.04	0.95
LightGBM	0.03	0.04	0.97

Kết quả cho thấy rằng cả hai mô hình XGBoost và LightGBM có độ chính xác khá cao tại hầu hết các trạm. Đặc biệt kết quả tại trạm Thủ Dầu Một cho thấy rằng đối với mô hình XGBoost thì giá trị của MAE là 0.03, RMSE là 0.04 và NSE là 0.95. Các giá trị này

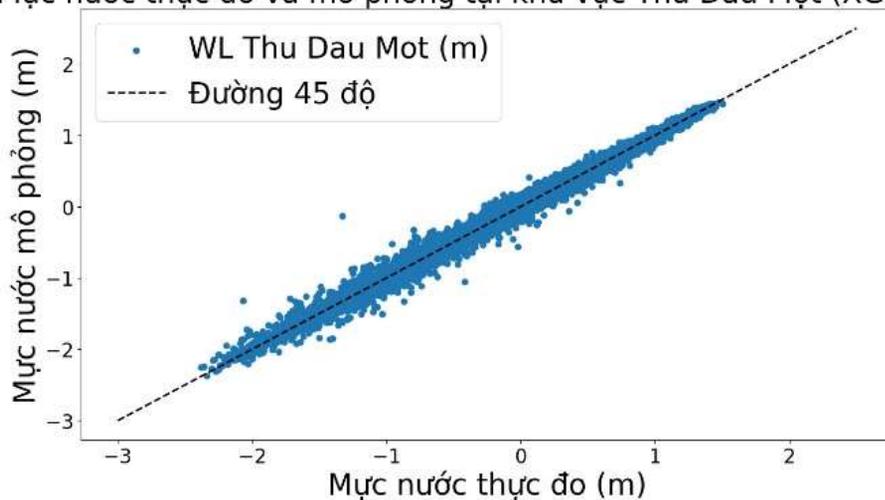
cho thấy mức độ chênh lệch giữa kết quả dự báo và dữ liệu quan trắc thực tế khá nhỏ. Hình 2 thể hiện biểu đồ so sánh mực nước triều thực đo và mô phỏng của 2 mô hình học máy tại các trạm quan trắc.



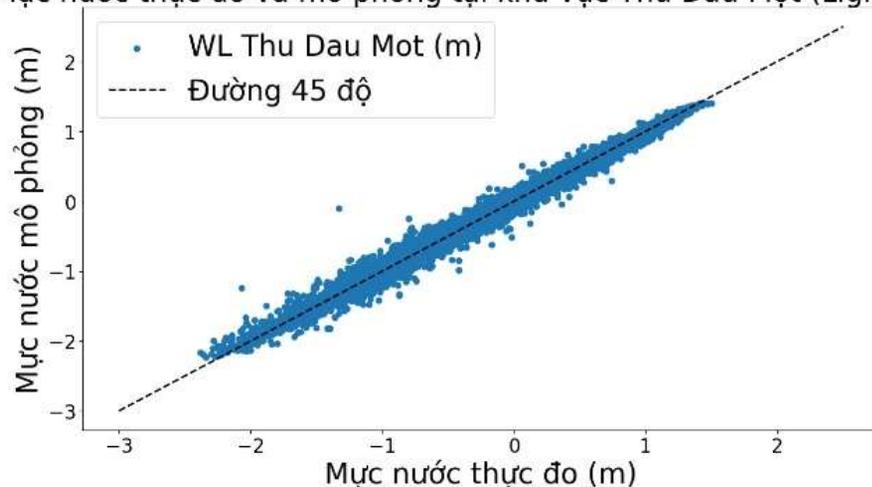
Mức nước thực đo và mô phỏng tại khu vực Phú An (LightGBM)

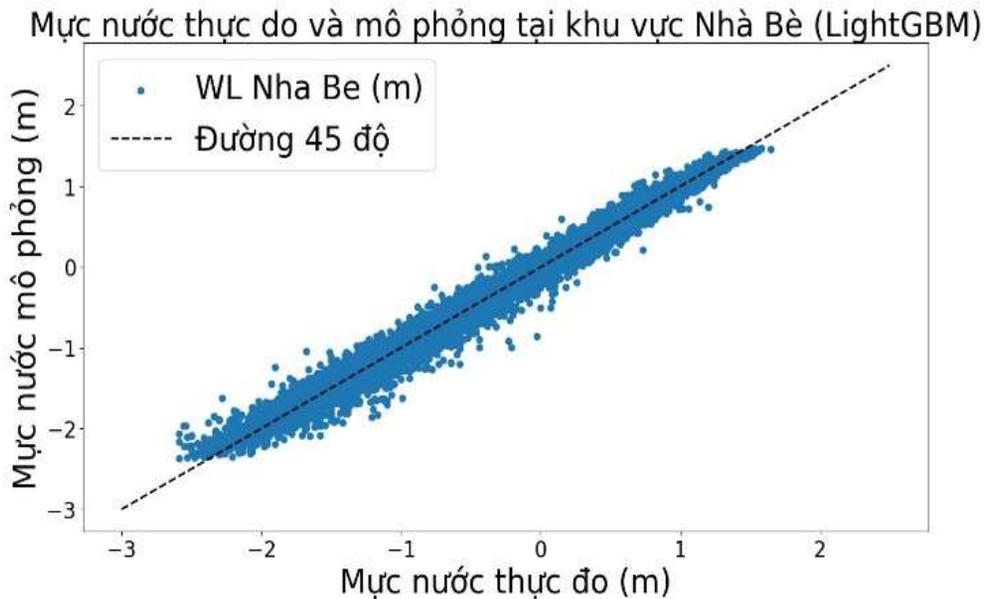
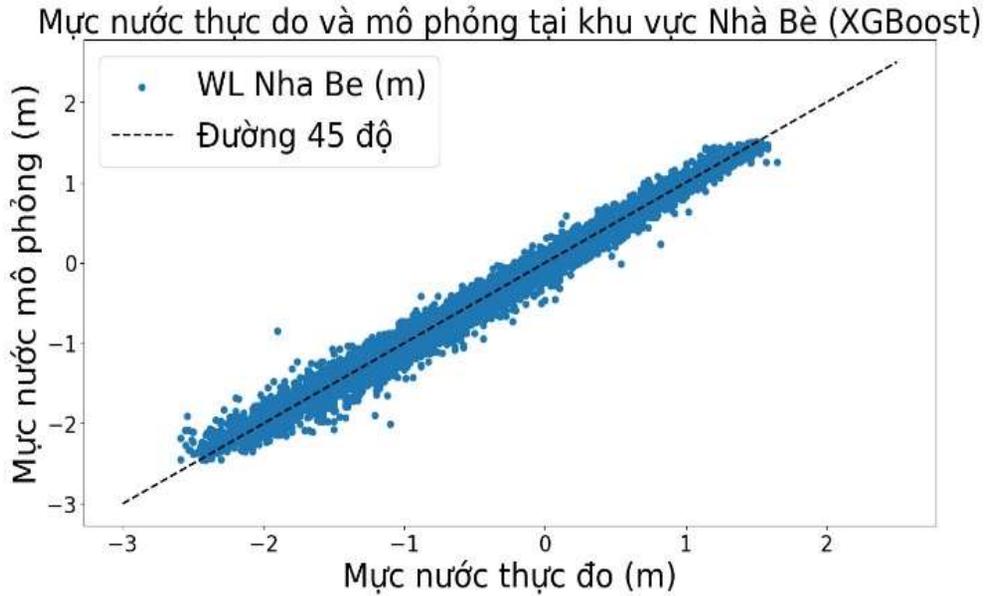


Mức nước thực đo và mô phỏng tại khu vực Thủ Dầu Một (XGBoost)



Mức nước thực đo và mô phỏng tại khu vực Thủ Dầu Một (LightGBM)

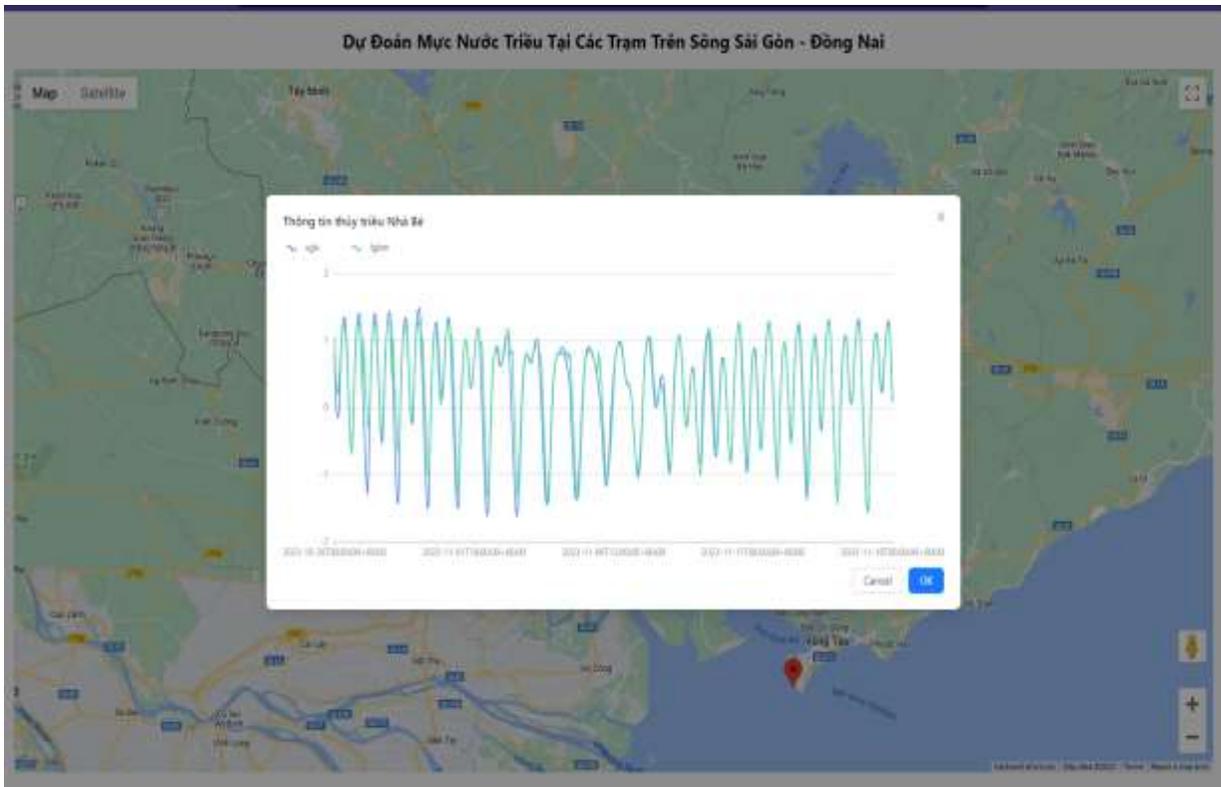




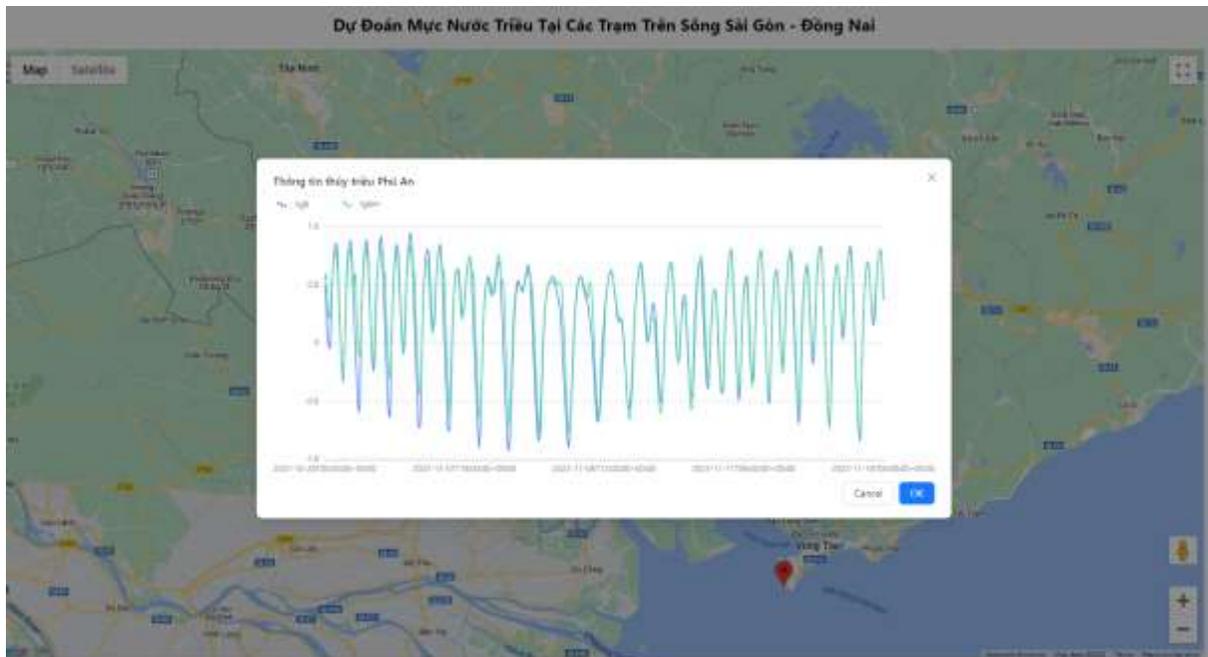
Hình 2: So sánh mức nước thực đo và mô phỏng của 2 mô hình học máy

Kết quả dự báo mực nước triều của 4 trạm trong 10 ngày tiếp theo được thể hiện trực tuyến tại địa chỉ website “<http://phuongnamdts.com:3137/>”. Giao diện website thiết kế gồm các tính năng thể hiện 4

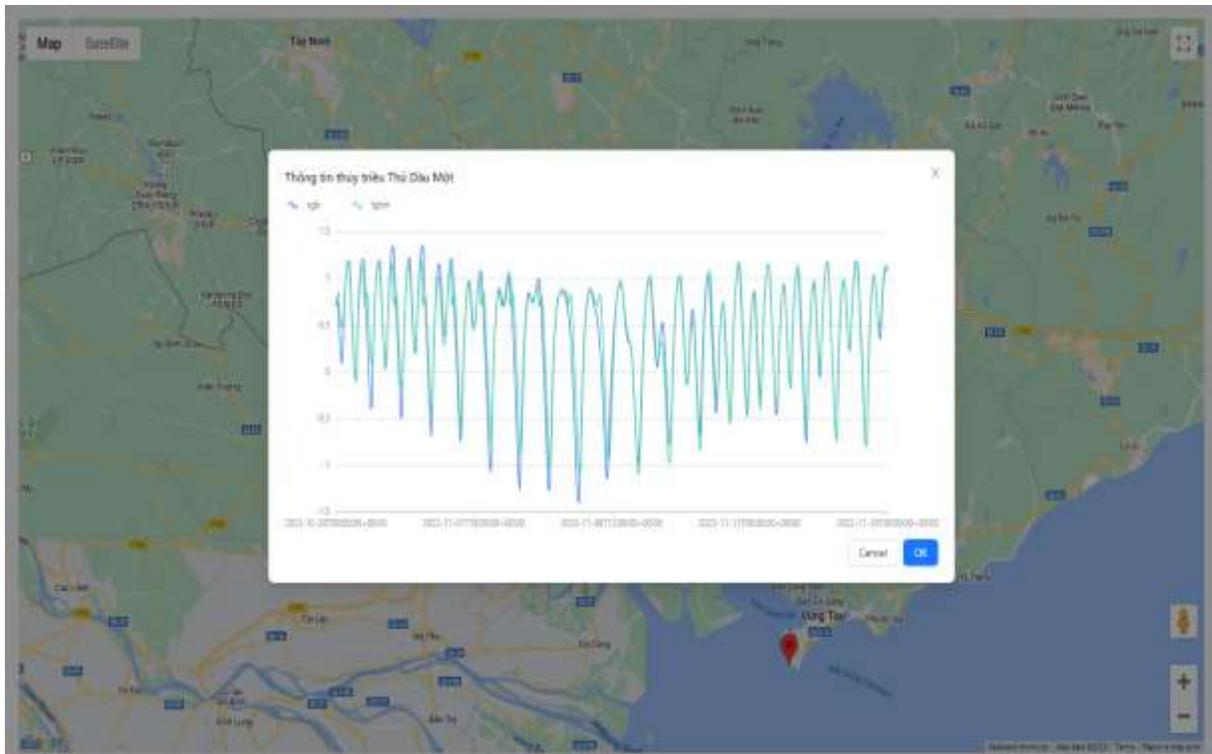
vị trí của trạm đo trên nền bản đồ Google Map. Tại mỗi vị trí thể hiện đường quá trình mực nước dự báo cả 2 mô hình XGBoost và LightGBM. Kết quả thể hiện bởi các Hình 3, 4 và 5.



Hình 3: Dữ liệu mực nước triều dự báo tại trạm Nhà Bè



Hình 4: Dữ liệu mực nước triều dự báo tại trạm Phú An



Hình 5: Dữ liệu mực nước triều dự báo tại trạm Thủ Dầu Một

4. KẾT LUẬN

Nghiên cứu đã thu thập và xử lý được dữ liệu mực nước triều trên sông Sài Gòn – Đồng Nai tại 4 trạm đo là Vũng Tàu, Nhà Bè, Phú An và Thủ Dầu Một trong 26 năm. Kiểm định Mann-Kendall đã được sử dụng để đánh giá xu thế gia tăng chuỗi mực nước triều tại 4 trạm nghiên cứu. Kết quả chỉ ra rằng dữ liệu mực nước tại Nhà Bè, Phú An và Thủ Dầu Một có xu thế gia tăng đáng kể, trong khi đó tại trạm Vũng Tàu ghi nhận xu thế gia tăng không đáng kể.

Mô hình XGBoost và LightGBM được xây dựng để dự báo mực nước tại 4 trạm đo mực nước. Kết quả đánh giá sai số mô hình chỉ ra rằng cả 2 mô hình có độ tin cậy cao trong việc dự đoán mực nước triều thông qua các chỉ số RMSE, MAE và NSE. Kết quả cũng cho thấy

rằng mô hình XGBoost cho kết quả dự đoán tốt hơn so với mô hình LightGBM.

Ngoài ra nghiên cứu cũng tích hợp mô hình XGBoost và LightGBM trên nền tảng website để thể hiện kết quả dự báo mực nước triều trực tuyến. Website này cung cấp thông tin về mực nước triều dự báo của cả 2 mô hình trong 10 ngày tiếp theo tại các trạm.

Kết quả của nghiên cứu bước đầu cho thấy tính khả thi và tiềm năng của việc sử dụng mô hình học máy để dự báo mực nước triều trực tuyến. Kết quả dự báo mực nước triều sẽ cung cấp những thông tin hữu ích trong việc đề xuất các giải pháp chủ động phòng tránh ngập úng cho các đô thị ven sông Sài Gòn – Đồng Nai cũng như là vận hành công trình thủy phục vụ cho nông nghiệp và giao thông thủy.

TÀI LIỆU THAM KHẢO

- [1] Giám, N.M., et al., *Những nguyên nhân chính tác động đến ngập Thành phố Hồ Chí Minh*. Tạp chí Khí tượng thủy văn, 2023.
- [2] Storch, H. và N.K. Downes, *A scenario-based approach to assess Ho Chi Minh City's urban development strategies against the impact of climate change*. Cities, 2011. **28**(6): p. 517-526.
- [3] Học, Đ.X., *Nguyên nhân và các giải pháp chống ngập úng ở TP Hồ Chí Minh*. Khoa học Kỹ thuật thủy lợi và Môi trường, 2009. **24**.
- [4] ADB. (Asian Development Bank) *Ho Chi Minh City Adaptation to Climate Change: Summary Report*. 2010; Available from: <https://www.adb.org/publications/ho-chi-minh-city-adaptation-climate-change-summary-report>.
- [5] Việt, L.V., *Ảnh hưởng của biến đổi khí hậu và quá trình đô thị hóa đến mực nước trên hệ thống sông Sài Gòn-Đồng Nai*. Khí tượng thủy văn, 2016. **07**.
- [6] Giang, N.N.H., et al., *Statistical and hydrological evaluations of water dynamics in the lower Sai Gon-Dong Nai River, Vietnam*. 2022. **14**(1): p. 130.
- [7] Thủy, N.B. và T.Q. Tiên, *Nghiên cứu nước dâng trong các đợt triều cường tại ven biển đông Nam Bộ*. Khí tượng thủy văn, 2017. **683**(11).
- [8] Lê, T.Q., et al., *Mô phỏng và xây dựng bản đồ ngập lụt cho hạ lưu hệ thống sông Đồng Nai*. Khí tượng thủy văn, 2022(747): p. 9-20.
- [9] Hoàng, T.T., et al., *Xây dựng hệ thống mô hình dự báo, cảnh báo ngập cho Thành phố Thủ Đức*. 2021. **5**(SI2).
- [10] Nữ, H.T.T., et al., *Ứng dụng mô hình thủy văn đô thị mô phỏng mức độ ngập do gia tăng mực nước triều và khả năng thoát nước cho hệ thống kênh Tân Hóa-Lò Gốm ở thành phố Hồ Chí Minh*. Khí tượng thủy văn, 2022. **740**: p. 22-35.
- [11] Tín, N.V., et al., *Ứng dụng phần mềm UTIDE dự báo mực nước triều ở khu vực ven Nam Bộ*. Khí tượng thủy văn, 2022(734): p. 50-63.
- [12] Tuấn, H.V. và H.V. Hùng, *Sử dụng mạng nơ-ron nhân tạo dự báo mực nước sông chịu ảnh hưởng của thủy triều*. Khoa học và công nghệ Thủy lợi, 2019. **52**: p. 108-116.
- [13] Cảnh, D.T., *Nghiên cứu, ứng dụng mạng nơ ron nhân tạo để dự báo, chỉnh biên tài liệu mực nước sông không bị ảnh hưởng bởi thủy triều*. Science Journal of Natural Resources Environment, 2021(36): p. 64-74.
- [14] Toàn, T.Q. và N.T.N. Nhẫn, *Mô phỏng dữ liệu dòng chảy bằng mô hình chi tiết hóa động lực kết hợp với thuật toán học máy: áp dụng cho lưu vực sông Sài Gòn-Đồng Nai*. Khoa học và công nghệ Thủy lợi, 2021. **66**.
- [15] Phan, T.T.H. và X.H. Nguyen, *Combining statistical machine learning models with ARIMA for water level forecasting: The case of the Red river*. Advances in Water Resources, 2020. **142**: p. 103656.

- [16] Boueshagh, M. và M. Hasanlou, *Estimating water level in the Urmia Lake using satellite data: a machine learning approach*. The International Archives of the Photogrammetry, Remote Sensing Spatial Information Sciences, 2019. **42**: p. 219-226.
- [17] Choi, C., et al., *Development of water level prediction models using machine learning in wetlands: A case study of Upo wetland in South Korea*. Water, 2019. **12**(1): p. 93.
- [18] Zhu, S., et al., *Forecasting of water level in multiple temperate lakes using machine learning models*. Hydrology, 2020. **585**: p. 124819.
- [19] Mann, H.B., *Nonparametric Tests Against Trend*. Econometrica, 1945. **13**(3): p. 245-259.
- [20] Kendall, M.G., *Rank correlation methods*. 1962.
- [21] Chen, T. và C. Guestrin. *Xgboost: A scalable tree boosting system*. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016.
- [22] Ke, G., et al., *Lightgbm: A highly efficient gradient boosting decision tree*. Advances in neural information processing systems, 2017. **30**.
- [23] Glass, S. *API Tidal Prediction*. 2023 06/112023]; Available from: <https://docs.stormglass.io/#/authentication>.