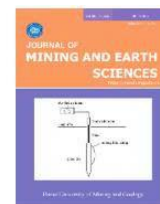




Journal of Mining and Earth Sciences

Website: <https://jmes.humg.edu.vn>



Application of the LSTM (Long Short-Term Memory) model for tidal forecasting: A case study at the Ba Ria - Vung Tau tide gauge station, Vietnam



Quoc Dinh Nguyen Huynh¹, Trong Gia Nguyen^{2,3,*}

¹ Ho Chi Minh City University of Natural Resources and Environment, Ho Chi Minh City, Việt Nam

² Faculty of Geomatics and Land Administration, Hanoi University of Mining and Geology, Hanoi, Vietnam

³ Geodesy and Environment Research Group, Hanoi University of Mining and Geology, Hanoi, Vietnam

ARTICLE INFO

Article history:

Received 6th Jun. 2025

Revised 6th Sept. 2025

Accepted 20th Sept. 2025

Keywords:

Deep learning,
Forecasting,
LSTM,
Tide.

ABSTRACT

In recent years, the application of tidal research has markedly supported and enhanced various socio-economic activities in coastal regions, particularly port operations, aquaculture, and disaster prevention and mitigation. Accurate tidal water level forecasting plays a vital role in spatial planning, coastal infrastructure management, and early flood warning systems. This study proposes the application of a Long Short-Term Memory (LSTM) recurrent neural network model to forecast daily tidal water levels at the Ba Ria - Vung Tau station, based on observed data from 1999 to 2022. The input data were preprocessed through three main steps: outlier removal using the Interquartile Range (IQR) method, Min-Max normalization, and time series decomposition into trend, seasonal, and residual components. The model was trained on the 1999÷2021 dataset, tested with 2022 data, and used to forecast tidal levels for 2023 using recursive forecasting. The results show that the LSTM model achieved high performance with low error on the test dataset ($MSE = 0.0039$ cm, $MAE = 0.0449$ cm, $R^2 = 0.9443$). The model effectively captured the semi-diurnal tidal cycle, demonstrated fast convergence, low error, and high stability, highlighting its potential for integration into automated tidal forecasting and early warning systems for coastal management. However, the current model uses only univariate data (tidal water level) without incorporating meteorological and oceanographic variables such as wind, pressure, temperature, or wave data. Future studies should expand toward multivariate models to improve generalization and forecasting accuracy.

Copyright © 2025 Hanoi University of Mining and Geology. All rights reserved.

*Corresponding author

E - mail: nguyengiatrong@humg.edu.vn

DOI: 10.46326/JMES.2025.66(5).09



Tạp chí Khoa học Kỹ thuật Mỏ - Địa chất

Trang điện tử: <https://tapchi.humg.edu.vn>



Nghiên cứu ứng dụng mô hình LSMT (Long Short-Term Memory) dự báo thủy triều: trường hợp nghiên cứu tại trạm đo mực nước Bà Rịa - Vũng Tàu, Việt Nam

Huỳnh Nguyễn Đình Quốc¹, Nguyễn Gia Trọng^{2,3,*}

¹ Trường Đại học Tài nguyên Môi trường, TP.HCM, Việt Nam

² Khoa Trắc địa - Bản đồ và Quản lý đất đai, Trường Đại học Mỏ - Địa chất, Hà Nội, Việt Nam

³ Nhóm nghiên cứu Trắc địa cao cấp - môi trường, Trường Đại học Mỏ - Địa chất, Hà Nội, Việt Nam

THÔNG TIN BÀI BÁO

Quá trình:

Nhận bài 6/6/2025

Sửa xong 6/9/2025

Chấp nhận đăng 20/9/2025

Từ khóa:

Dự báo,
LSTM,
Học sâu,
Thủy triều.

TÓM TẮT

Những năm gần đây, nhờ ứng dụng các nghiên cứu về thủy triều, nhiều hoạt động kinh tế - xã hội ở khu vực ven biển như vận hành cảng, nuôi trồng thủy sản và phòng chống thiên tai đã được hỗ trợ và cải thiện rõ rệt. Việc dự báo chính xác mực nước thủy triều đóng vai trò then chốt trong quy hoạch không gian, vận hành công trình ven biển và cảnh báo sớm nguy cơ ngập lụt. Nghiên cứu này đề xuất mô hình mạng nơ-ron hồi tiếp (LSTM) nhằm dự báo mực nước thủy triều hàng ngày tại trạm Bà Rịa - Vũng Tàu, dựa trên chuỗi số liệu quan trắc thực đo từ năm 1999 đến 2022. Dữ liệu đầu vào được xử lý qua ba bước chính: loại bỏ ngoại lai bằng phương pháp IQR, chuẩn hóa theo thang Min-Max, và phân tích tách thành phần chuỗi thời gian thành xu hướng, chu kỳ mùa vụ và phần nhiễu. Mô hình được huấn luyện trên dữ liệu từ 1999÷2021, kiểm thử với năm 2022 và dự báo cho năm 2023 theo phương pháp cuốn trượt (recursive forecasting). Kết quả cho thấy mô hình LSTM đạt hiệu suất cao với sai số cho tập dữ liệu kiểm thử thấp ($MSE = 0,0039$ cm, $MAE = 0,0449$ cm, $R^2 = 0,9443$). Mô hình học tốt đặc trưng chu kỳ bán nhật của thủy triều, hội tụ nhanh, sai số thấp và khả năng dự báo ổn định. Điều này cho thấy tiềm năng ứng dụng của LSTM trong các hệ thống dự báo mực nước tự động phục vụ cảnh báo sớm và quản lý vùng ven biển. Tuy nhiên, mô hình hiện chỉ sử dụng dữ liệu đơn biến (mực nước), chưa tích hợp các yếu tố khí tượng - hải văn như gió, áp suất, nhiệt độ hoặc sóng biển. Việc mở rộng theo hướng mô hình đa biến là cần thiết để tăng tính tổng quát và độ chính xác trong các nghiên cứu tiếp theo.

© 2025 Trường Đại học Mỏ - Địa chất. Tất cả các quyền được bảo đảm.

*Tác giả liên hệ

E - mail: nguyengiatrong@humg.edu.vn

DOI: 10.46326/JMES.2025.66(5).09

1. Mở đầu

Thủy triều là một hiện tượng tự nhiên quan trọng, ảnh hưởng trực tiếp đến hoạt động giao thông thủy, nuôi trồng thủy sản, khai thác cảng biển và phòng chống thiên tai tại các khu vực ven biển. Việc dự báo mực nước thủy triều chính xác và kịp thời đóng vai trò thiết yếu trong công tác lập kế hoạch và ứng phó với các rủi ro ven biển, đặc biệt trong bối cảnh biến đổi khí hậu và nước biển dâng đang diễn ra ngày càng rõ rệt.

Sự thay đổi mực nước biển là một trong những biểu hiện rõ rệt nhất của biến đổi khí hậu toàn cầu. Trong những năm gần đây, các mô hình học máy đã được ứng dụng rộng rãi nhằm cải thiện độ chính xác trong việc phân tích và dự báo mực nước biển dựa trên dữ liệu chuỗi thời gian.

Xie và nnk. (2024) đã đề xuất mô hình học máy lai SSA-LSTM, kết hợp phân tích phổ đơn nhất (SSA) để phân rã chuỗi thời gian mực nước biển thành các thành phần (xu hướng, chu kỳ, nhiễu) và dự báo từng phần bằng LSTM. Mô hình được áp dụng cho dữ liệu AVISO (1993÷2021) tại khu vực ven biển Trung Quốc, đạt độ chính xác cao với $R^2 = 0,98$, MAE = 4,16 mm, RMSE = 5,34 mm, vượt trội so với LSTM, EMD-LSTM và CEEMDAN-LSTM. Tuy nhiên, sai số tuyệt đối vẫn còn khá lớn, hạn chế khả năng ứng dụng trong các bài toán yêu cầu độ chính xác cao như cảnh báo rủi ro hoặc quy hoạch vùng ven biển nhạy cảm.

Uluocak (2025) đề xuất mô hình lai LSTM-CNN để dự báo mực nước biển toàn cầu đến năm 2050, sử dụng dữ liệu mực nước biển (1880÷2022) và nhiệt độ không khí từ CMIP6. Mô hình đạt hiệu suất cao (MSE = 0,4644 mm, NSE = 0,9994) và vượt trội GRU, CNN, GRU-CNN theo kiểm định Diebold và nnk. (1995). Dự báo mực nước biển năm 2050 là 313,81 mm, tăng khoảng 83 mm so với năm 2022. Tuy nhiên, sự chênh lệch lớn giữa các mô hình (250,65÷399,48 mm) cho thấy độ bất định cao, đồng thời mô hình chưa xét đến các yếu tố như băng tan, dòng hải lưu và chưa triển khai theo vùng, hạn chế tính ứng dụng thực tiễn.

Makky và nnk. (2024) áp dụng hai mô hình học máy là XGBoost và MLP để dự đoán mực nước biển tại Hawaii dựa trên dữ liệu vệ tinh Jason-2 năm 2019. Kết quả cho thấy XGBoost đạt hiệu quả cao hơn với hệ số xác định $R^2 = 0,9520$ so với MLP là 0,9255, ngoài ra mô hình XGBoost đạt RMSE là 0,1832, trong khi MLP đạt 0,2451 cho thấy mô

hình XGBoost có khả năng xử lý dữ liệu phi tuyến mạnh mẽ hơn mô hình MLP.

Braakmann-Folgmann và nnk. (2017) phát triển mô hình CNN-RNN dự báo bất thường mực nước biển (SLA) ở Thái Bình Dương, kết hợp đặc trưng không gian và thời gian cho kết quả chính xác hơn hồi quy tuyến tính và ổn định theo thời gian. Đây là một trong những nghiên cứu đầu tiên ứng dụng CNN-RNN cho SLA nhưng chỉ kiểm nghiệm tại Thái Bình Dương, chưa so sánh với các mô hình học sâu khác và thiếu phân tích siêu tham số.

Nghiên cứu của Raj và nnk. (2022) đề xuất mô hình lai SSA-BiLSTM nhằm dự báo mực nước biển và sóng thủy triều tại Nam Thái Bình Dương. Mô hình sử dụng SSA để phân rã tín hiệu và BiLSTM để học các yếu tố đặc trưng chuỗi thời gian hai chiều. Kết quả đạt được cho thấy SSA-BiLSTM có hiệu suất cao với sai số $R^2 = 0,985$ và giảm RMSE 8% so với các mô hình LSTM, BiLSTM và SSA-LSTM. Tuy nhiên, mô hình chưa được kiểm chứng cho dự báo dài hạn và dữ liệu ngoài mẫu làm hạn chế khả năng tổng quát hóa.

Chi (2022) đề xuất mô hình lai kết hợp giữa SARIMA và MLP dự báo mực nước biển trung bình hàng tháng tại Grand Isle, Louisiana trong giai đoạn 1978÷2020. Thông qua phân tích chuỗi thời gian, nghiên cứu làm rõ ảnh hưởng của xu thế và yếu tố mùa vụ đến biến động mực nước biển. Kết quả cho thấy mô hình MLP vượt trội hơn SARIMA với sai số MSE lần lượt là 0,0015 và 0,00245, chứng minh khả năng nắm bắt đặc trưng phi tuyến của MLP trong dữ liệu thời gian. Tuy nhiên, nghiên cứu còn giới hạn ở phạm vi địa lý hẹp, chưa so sánh với các mô hình học sâu như LSTM hoặc GRU và thiếu phân tích chi tiết về xu thế dài hạn cũng như tác động của yếu tố biến đổi khí hậu.

Le và nnk. (2024) ứng dụng mô hình GRU để cải thiện dự báo chuỗi thời gian, khắc phục hạn chế của mô hình RNN truyền thống với cấu trúc đơn giản và huấn luyện nhanh hơn LSTM. Kết quả cho thấy giảm batch size và tăng số epoch giúp cải thiện độ chính xác (giảm RMSE, MAE; tăng F-score). Tuy nhiên, nghiên cứu chưa so sánh GRU với các mô hình học sâu khác như LSTM, CNN-GRU và chưa phân tích đầy đủ ảnh hưởng của các siêu tham số như learning rate, hay số lớp GRU.

Nguyen và nnk. (2025) xây dựng mô hình học sâu dùng mạng GRU để dự báo độ cao điểm GNSS tại trạm CPHU, với dữ liệu xử lý qua Gamit/Globk.

Mô hình huấn luyện trên chuỗi từ 2019÷2022, thử hai thuật toán Adam, SGD và hai hàm mất mát MSE, Huber. Kết quả tốt nhất là GRU + Adam + MSE ($R^2 \approx 0,5452$; MAE = 4,24 mm), trong khi SGD và Huber làm giảm độ chính xác. Mô hình hiệu quả hơn với cửa sổ trượt nhỏ (lag = 10) cho dự báo ngắn hạn nhưng còn hạn chế về giải thích, siêu tham số, đặc trưng đầu vào và kiểm tra tổng quát trên dữ liệu mới.

Trong nhiều thập kỷ qua, các mô hình thống kê truyền thống như ARIMA và Fourier đã được sử dụng để dự báo mực nước biển, có ưu điểm là dễ hiểu, yêu cầu ít dữ liệu và phù hợp với điều kiện ổn định. Theo Pugh (2002) cho thấy mực nước triều thường biến đổi theo quy luật điều hòa, có thể dự báo hiệu quả khi xác định đúng các tham số từ dữ liệu lịch sử. Ngoài ra, phương pháp truyền thống không phản ánh tốt các yếu tố phi tuyến và hiện tượng bất thường như gió, áp suất, dòng chảy (Abubakar và nnk., 2019). Ngược lại, mạng nơ-ron hồi tiếp LSTM nổi bật với khả năng ghi nhớ chuỗi thời gian dài, có khả năng học các mối quan hệ phức tạp từ dữ liệu lớn, mang lại độ chính xác cao hơn trong dự báo, đặc biệt trong môi trường thủy văn phức tạp, cần nhiều dữ liệu có các yếu tố phi tuyến hoặc biến đổi không theo quy luật cũ. Từ đó cho thấy mô hình LSTM phù hợp hơn trong điều

kiện dữ liệu thay đổi phức tạp, dài hạn trong khi phương pháp truyền thống thích hợp với môi trường ổn định.

Bài báo xây dựng và đánh giá mô hình LSTM dự báo mực nước thủy triều hàng ngày dựa trên dữ liệu 1999÷2022, huấn luyện đến năm 2021, kiểm thử năm 2022 và dự báo cho năm 2023. Quy trình bao gồm tiền xử lý, xây dựng mô hình, đánh giá và trực quan kết quả. Mục tiêu là tạo ra mô hình chính xác, ứng dụng thực tiễn trong cảnh báo sớm và quản lý ven biển Việt Nam.

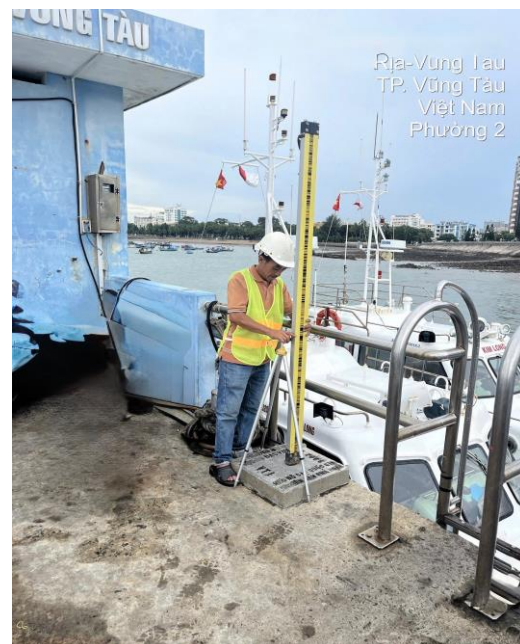
2. Phạm vi dữ liệu và phương pháp nghiên cứu

2.1. Phạm vi dữ liệu nghiên cứu

Trạm nghiệm triều Vũng Tàu là một trạm quan trắc thủy văn ven biển quan trọng và lâu đời tại Việt Nam, đặt tại Vũng Tàu - cửa ngõ ra biển của vùng Đông Nam Bộ (Hình 1). Với vị trí chiến lược và hệ thống thiết bị hiện đại, trạm ghi nhận mực nước biển liên tục, phục vụ theo dõi thủy triều, dự báo thiên tai và quy hoạch đô thị ven biển. Dữ liệu dài hạn từ trạm nghiệm triều có giá trị trong nghiên cứu biến đổi khí hậu, sụt lún đất và mô hình dự báo mực nước. Trạm nghiệm triều hiện đang giữ vai trò chủ lực trong mạng lưới quan trắc ven biển của Việt Nam hiện nay.



a. Bản đồ Việt Nam.



b. Trạm nghiệm triều khu vực Bà Rịa - Vũng Tàu.

Hình 1. Bản đồ Việt Nam (a) và trạm nghiệm triều ở khu vực Bà Rịa - Vũng Tàu thuộc TP. Hồ Chí Minh (b).

Dữ liệu nghiệm triều trạm Bà Rịa - Vũng Tàu được thu thập trong giai đoạn 1999÷2022 thông qua máy đo mực nước tự động triều ký tự ghi. Cấu trúc dữ liệu gồm các giá trị mực nước theo thời gian được ghi nhận định kỳ 1 giờ/lần bằng các cảm biến áp suất đặt dưới nước hoặc cảm biến radar/siêu âm lắp trên cao. Dữ liệu được lưu trữ, số hóa và truyền về trung tâm qua mạng viễn thông sau đó được hiệu chuẩn định kỳ để đảm bảo độ chính xác. Các giá trị mực nước được sử dụng theo dõi thủy triều, cảnh báo thiên tai và nghiên cứu khoa học. Tập dữ liệu triều được sử dụng trong nghiên cứu là chuỗi thời gian đo độ cao mực nước biển tại khu vực Vũng Tàu, Việt Nam.

2.2. Phân tích dữ liệu dựa trên các đặc trưng thống kê

Tập dữ liệu nghiên cứu được cung cấp bởi Trung tâm Khí tượng Thủy văn Nam Bộ là chuỗi dữ liệu theo thời gian khảo sát độ cao mực nước biển so với độ cao Hòn Dấu (ký hiệu H) tại khu vực Vũng Tàu, Việt Nam được thu thập liên tục từ ngày 01/01/1999 đến ngày 31/12/2022 với tần suất khảo sát trung bình 1 mẫu/ngày. Tổng cộng có 8.765 trị đo khảo sát hợp lệ trong giai đoạn 24 năm. Dữ liệu này đại diện cho biến động thủy triều

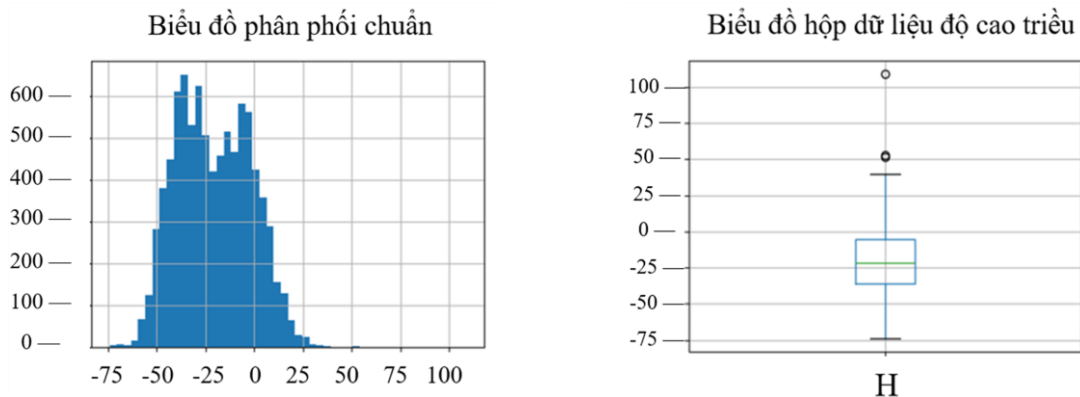
tại một trong những cảng biển lớn và nhộn nhịp nhất miền nam Việt Nam. Thống kê trên tập dữ liệu triều nhận được kết quả ở Bảng 1.

Qua kết quả thống kê ở Bảng 1 trên cho thấy biến động thủy triều ở khu vực Bà Rịa - Vũng Tàu chịu ảnh hưởng của chế độ bán nhật triều không đều với biên độ triều lớn nhất đạt $108,90 - (-74,20) = 183,1$ cm, mức biến thiên triều trong năm có thể lên đến gần 2 m. Mực nước triều dao động quanh giá trị trung bình với độ lệch chuẩn nhỏ (± 19 cm), phân bố gần đối xứng cho thấy triều lên và triều xuống diễn ra cân bằng và ổn định theo chu kỳ. Kết quả phân tích phân vị cho thấy 50% giá trị dữ liệu nằm trong khoảng từ -36,00 cm đến -5,50 cm, dữ liệu không quá phân tán và ít xuất hiện các giá trị cực đoan. Độ lệch (0,18) gần bằng 0 cho thấy phân bố dữ liệu gần như đối xứng, biểu thị rằng triều lên và triều xuống có tần suất và độ cao tương đối cân bằng.

Biểu đồ Hình 2 cho thấy phân phối mực nước có dạng gần chuẩn, hơi nghiêng phải nhẹ ($\text{skew} = 0,18$), không có đỉnh nhọn ($\text{kurtosis} \text{ âm}$) và tập trung chủ yếu trong khoảng -50 cm đến 12 cm. Biểu đồ hộp bên cạnh cho thấy có một số điểm ngoại lệ phía cao, vượt ngưỡng 50 cm.

Bảng 1. Dữ liệu triều thống kê.

Thống kê	Giá trị
Số trị đo khảo sát	8.765
Trung bình (mean)	-20,637 cm
Độ lệch chuẩn (std)	19,040 cm
Min - Max	-74,20÷108,90 cm
Phân vị 25÷75%	-36,00÷-5,50 cm
Độ lệch (Skewness)	0,18
Độ nhọn (Kurtosis)	-0,54



Hình 2. Biểu đồ phân phối chuẩn và biểu đồ hộp (boxplot) của dữ liệu độ cao triều.

2.3. Tiền xử lý và làm sạch dữ liệu

Do dữ liệu thu thập thực tế thường tồn tại nhiều hoặc giá trị bất thường, nhóm nghiên cứu đã tiến hành xử lý các trị khảo sát ngoại lai (outlier) bằng phương pháp xác định khoảng giá trị hợp lý dựa trên quy tắc IQR (Interquartile Range) (Tukey, 1977). Các điểm nằm ngoài khoảng: $[Q1 - 1.5 * IQR, Q3 + 1.5 * IQR]$ được xem là ngoại lệ và được loại bỏ khỏi tập huấn luyện. Cụ thể: các điểm có giá trị độ cao triều H lớn hơn 50 cm đã bị loại bỏ; các trị đo thứ 5.419 (108,9 cm), 8.751 (51,5 cm) và 8.752 (52,6 cm) chiếm dưới 1% tổng số mẫu giúp cải thiện chất lượng huấn luyện và giảm sai số trong dự báo (Hình 3, 4).

Kết quả ở Hình 3 cho thấy phân phối dữ liệu trở nên tập trung và cân đối hơn, không còn các giá trị vượt quá 50 cm khi loại bỏ trị ngoại lai.

Việc loại bỏ outlier không chỉ giúp mô hình hội tụ nhanh hơn mà còn cải thiện đáng kể hiệu suất dự báo, thể hiện qua độ giảm của sai số MSE và MAE trong giai đoạn kiểm thử.

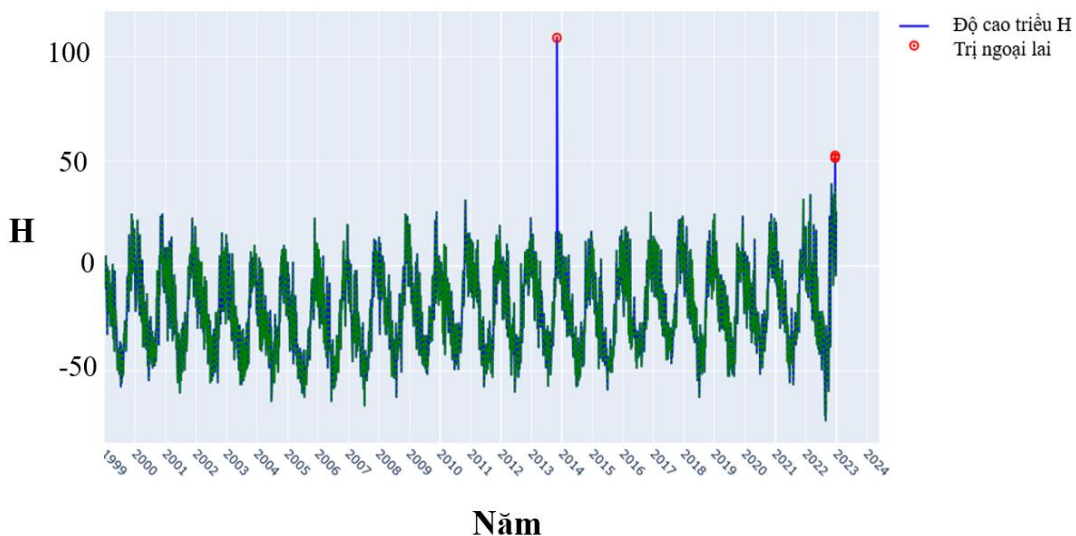
Biểu đồ ở Hình 5 thể hiện chuỗi gốc đi kèm với hai đường trung bình trượt ngắn hạn (MA7 - trung bình trượt 7 ngày) và dài hạn (MA30 - trung bình trượt 30 ngày), cùng với đường xu hướng tuyến tính (OLS). Đường MA7 bám sát dao động ngắn ngày của thủy triều trong khi MA30 phản ánh chuyển động trung hạn ổn định hơn. Đường xu hướng có hướng đi lên cho thấy có xu hướng

Biểu đồ hộp của dữ liệu mực nước triều (H) trước và sau loại bỏ trị ngoại lai



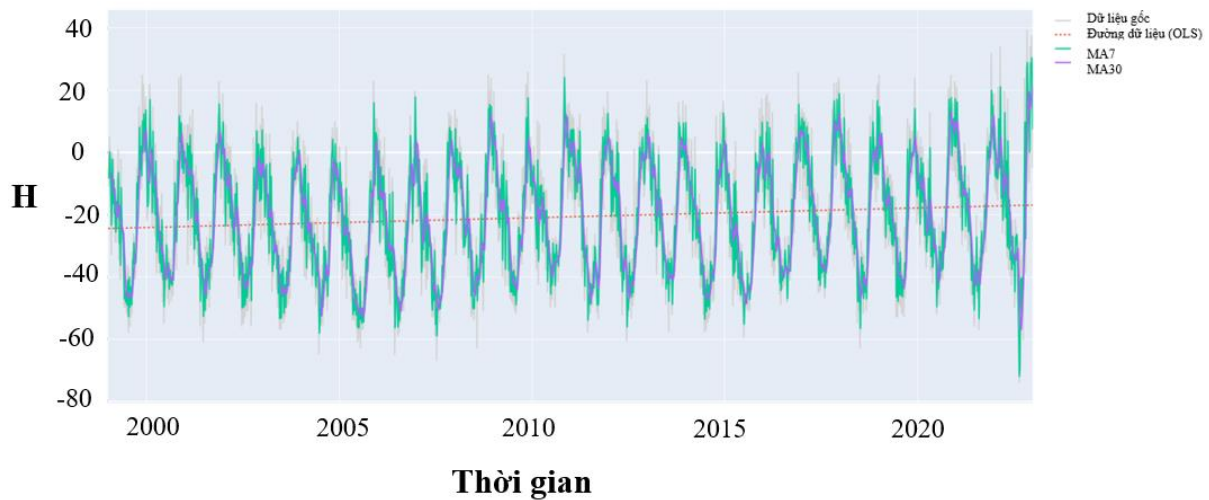
Hình 3. Biểu đồ hộp trước và sau khi loại bỏ trị ngoại lai outlier.

Chuỗi dữ liệu thời gian của độ cao mực nước triều H



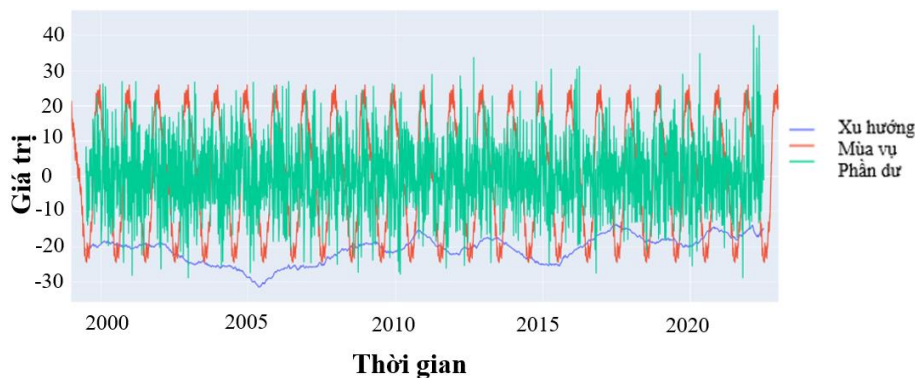
Hình 4. Minh họa trực quan chuỗi thời gian gốc và các trị đo ngoại lai outlier.

Đường xu hướng, MA7, MA30 trên chuỗi dữ liệu thời gian



Hình 5. Biểu đồ chuỗi dữ liệu độ cao triều (loại trị outlier) và các đường trung bình trượt ngắn hạn MA7 và dài hạn MA30.

Thành phần chuỗi thời gian (Xu hướng/ Mùa vụ/ Phần dư)



Hình 6. Biểu đồ phân tích dữ liệu thành phần chuỗi thời gian.

tăng rõ rệt ở chu kỳ dài hạn tuy nhiên chuỗi dữ liệu có tính dao động tăng theo thời gian.

Hình 6 trên trình bày kết quả phân rã chuỗi theo mô hình hồi tiếp tuyến tính (Additive Decomposition), trong đó:

- Thành phần Xu hướng Trend (đường màu xanh lam): biểu diễn xu thế dài hạn của dữ liệu theo thời gian. Quan sát từ biểu đồ, xu hướng biến thiên chậm theo thời gian, có giai đoạn giảm nhẹ khoảng năm 2005 sau đó phục hồi và có dấu hiệu tăng nhẹ từ sau năm 2010 đến 2022. Điều này cho thấy dữ liệu có xu hướng biến đổi trong dài hạn, không hoàn toàn ổn định, có thể liên quan đến hiện tượng tự nhiên hoặc thay đổi môi trường.

- Thành phần Mùa vụ Seasonal (đường màu đỏ): phản ánh các biến động mang tính chất định kỳ (chu kỳ mùa vụ). Thành phần Seasonal có dạng sóng điều hòa rõ ràng, được lặp đi lặp lại hàng năm qua đó cho thấy dữ liệu có tính mùa vụ rất cao - có thể là ảnh hưởng của chu kỳ khí hậu, thủy triều hoặc hiện tượng thiên nhiên khác. Ngoài ra, biên độ tương đối ổn định và cao cho thấy tính lặp lại mạnh mẽ.

- Thành phần Phần dư Residual (đường màu xanh lá): phần còn lại sau khi đã loại bỏ xu hướng và mùa vụ khỏi dữ liệu gốc. Biểu đồ này thể hiện một số điểm bất thường (spikes) xuất hiện, đặc biệt là các năm 2020÷2022 cho thấy khả năng có hiện tượng bất thường trong giai đoạn này.

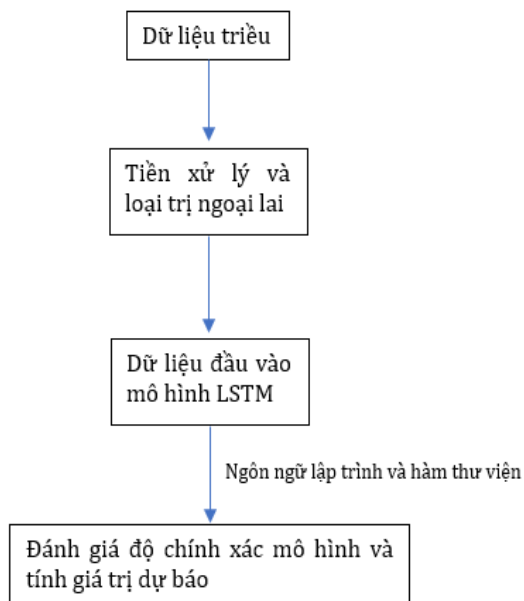
Như vậy, dựa vào đồ thị trên cho thấy dữ liệu có tính mùa vụ mạnh mẽ với chu kỳ đều đặn qua các năm. Như vậy, xu hướng tổng thể có dao động nhưng nhìn chung có dấu hiệu tăng nhẹ về sau, phần dư chứa nhiều dao động khó dự đoán, thể hiện ảnh hưởng của các yếu tố ngẫu nhiên hoặc chưa quan sát được. Phân tích trên cho thấy chuỗi thủy triều tại Vũng Tàu có đặc điểm rõ ràng về mùa vụ, rất phù hợp với mô hình học sâu như LSTM vốn có khả năng học phụ thuộc dài hạn và phi tuyến tính hiệu quả.

3. Mô hình LSTM (Long Short-Term Memory)

Để dự báo độ cao mực nước biển tại Vũng Tàu, nghiên cứu áp dụng kiến trúc mạng nơ-ron hồi tiếp LSTM - một biến thể của RNN (Recurrent Neural Network) thiết kế để xử lý và dự báo dữ liệu chuỗi thời gian (time series), đặc biệt là những chuỗi có mối quan hệ phụ thuộc dài hạn. Phương pháp nghiên cứu được thể hiện trên sơ đồ Hình 7.

3.1. Chuẩn hóa dữ liệu

Do giá trị tập dữ liệu mực nước tại trạm nghiệm triều có biên độ dao động rộng (khoảng từ -74,2 cm đến 108,9 cm) được khảo sát từ 01/01/1999 đến 31/12/2022, vì vậy áp dụng chuẩn hóa MinMaxScaler để đưa dữ liệu về



Hình 7. Phương pháp dự báo giá trị độ cao triều bằng mô hình LSTM.

khoảng [0, 1] giúp cải thiện tốc độ hội tụ và tính ổn định của mô hình. Ngoài ra, dữ liệu được chia thành hai tập:

- Tập huấn luyện train: thời gian từ 01/01/1999 đến 31/12/2021.
- Tập kiểm thử test: từ 01/01/2022 đến 31/12/2022.

3.2. Tạo dữ liệu theo dạng chuỗi trượt

Mô hình LSTM yêu cầu đầu vào theo dạng chuỗi thời gian cố định. Ở đây, mỗi mẫu dữ liệu được tạo từ cửa sổ trượt gồm 30 ngày liên tiếp (tương ứng window_size = 30), với giá trị dự báo là mực nước ở ngày kế tiếp. Sau khi tiền xử lý, số lượng mẫu thu được là:

$X_{train}.shape = (8367, 30, 1)$

$X_{test}.shape = (368, 30, 1)$

3.3. Kiến trúc mô hình LSTM

Để mô hình hóa và dự báo chuỗi thời gian mực nước biển tại trạm nghiệm triều khu vực Bà Rịa - Vũng Tàu, nghiên cứu này đã sử dụng mạng nơ-ron hồi tiếp LSTM, một kiến trúc được thiết kế đặc biệt để xử lý dữ liệu có tính phụ thuộc theo thời gian dài và khắc phục hiện tượng mất thông tin trong các mạng RNN truyền thống (Hochreiter và nnk, 1997).

Dữ liệu đầu vào bao gồm giá trị mực nước triều tại thời điểm t là H_t . Mô hình dự đoán giá trị H_{t+1} dựa trên một cửa sổ trượt n bước: $\hat{H}_{t+1} = f(H_{t-n+1}, \dots, H_t)$ trong đó mô hình học được hàm ánh xạ f thông qua quá trình huấn luyện bằng chuỗi dữ liệu.

Mô hình gồm hai tầng LSTM nối tiếp nhau với 64 và 32 đơn vị ẩn, kèm theo lớp Dropout để hạn chế quá khớp (overfitting) (Brownlee, 2017). Lớp đầu ra là Dense với một nút duy nhất nhằm dự đoán giá trị mực nước tiếp theo.

Tổng số tham số huấn luyện là 29.345 được tối ưu bằng thuật toán Adam (Kingma và nnk, 2014).

Tại mỗi bước thời gian t , một đơn vị LSTM thực hiện các bước tính toán sau (Hochreiter và nnk, 1997):

Cổng quên:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

Cổng đầu vào:

$$\begin{cases} i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \end{cases} \quad (2)$$

Cập nhật trạng thái bộ nhớ:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (3)$$

Cổng đầu ra:

$$\begin{cases} o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t = o_t * \tanh(C_t) \end{cases} \quad (4)$$

Trong đó: x_t - đầu vào tại thời điểm t , h_t - trạng thái ẩn, C_t - trạng thái bộ nhớ, σ - hàm kích hoạt sigmoid, $*$ - phép nhân từng phần tử.

Trong nghiên cứu này, mô hình LSTM được xây dựng như sơ đồ trong Hình 8.

3.4. Chiến lược huấn luyện và dừng sớm

Khi tính toán thực nghiệm, mô hình được huấn luyện tối đa 100 epoch với kích thước batch-size = 32. Để tránh hiện tượng overfitting, EarlyStopping được áp dụng dựa trên chỉ số val_loss, với ngưỡng lần lượt được chọn patience = 10 và patience = 20 (ngừng huấn luyện nếu mô hình không cải thiện).

Trong các bài toán dự báo và hồi quy, các chỉ số đánh giá hiệu suất mô hình thường được sử

dụng bao gồm Sai số bình phương trung bình (Mean Squared Error - MSE), Sai số tuyệt đối trung bình (Mean Absolute Error - MAE (Goodfellow và nnk, 2016)).

Sai số trung bình được định nghĩa bởi công thức:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (5)$$

Trong đó: y_i - giá trị thực tế, $\hat{y}_i$ - giá trị dự báo và n - số lượng điểm dữ liệu. MSE đo lường độ lớn trung bình của sai số dự đoán, trong đó các sai số được bình phương để đảm bảo không có giá trị âm và nhấn mạnh các sai số lớn.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (6)$$

Giá trị MAE đo lường sai số dự báo trung bình mà không nhấn mạnh các sai số lớn như trong MSE. MAE mang tính trực quan và không bị ảnh hưởng bởi các giá trị ngoại lai.

Ngôn ngữ lập trình được lựa chọn để xây dựng chương trình máy tính là ngôn ngữ Python (Ceder, 2018) với các hàm thư viện được sử dụng như pandas (Chen, 2017), numpy (Gupta & Bagchi, 2024), matplotlib (Tosi, 2009), tensorflow (Karim, 2018).

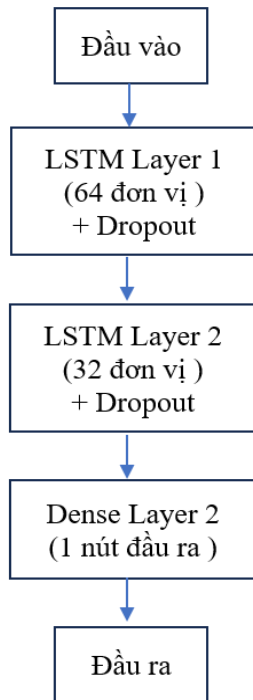
4. Kết quả và thảo luận

Sau quá trình huấn luyện mô hình LSTM, tiến hành đánh giá hiệu quả của mô hình trên tập kiểm thử test (dữ liệu năm 2022) và sử dụng mô hình đã huấn luyện để dự báo mực nước cho giai đoạn 01/2023 đến 12/2023. Phân tích được chia thành ba phần: hiệu suất huấn luyện, khả năng dự báo ngắn hạn và khả năng dự báo dài hạn.

4.1. Hiệu suất huấn luyện và hội tụ mô hình

Từ tập dữ liệu huấn luyện cho thấy khi thay đổi giá trị ngưỡng thì sai số của mô hình gần như không thay đổi.

Đồ thị Hình 9 và 10 trình bày diễn biến của hàm mất mát trên tập huấn luyện và tập dữ liệu thực tế theo từng epoch theo hai trường hợp ngưỡng 10 và 20 cho thấy cả hai mô hình đều đạt được sự hội tụ nhanh chóng với hàm mất mát giảm mạnh trong giai đoạn đầu và sau đó duy trì ổn định. Đường mất mát không dao động mạnh và nằm gần đường mất mát huấn luyện cho thấy mô hình có khả năng học tốt và khái quát hóa ổn định.



Hình 8. Xây dựng mô hình LSTM.

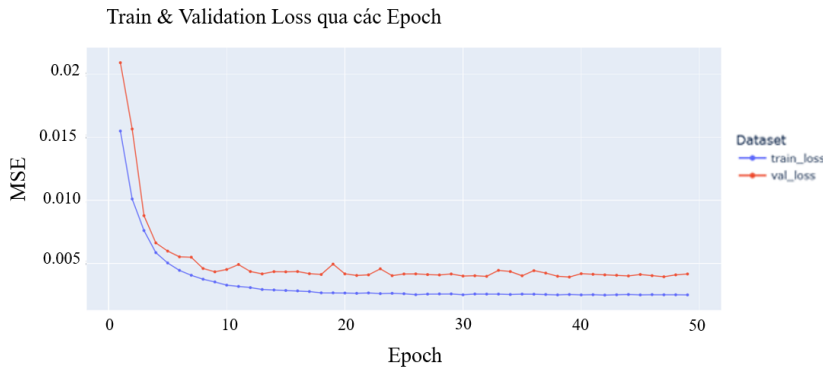
Để đánh giá định lượng hiệu suất của mô hình, các chỉ số sai số như MAE (Mean Absolute Error), RMSE (Root Mean Squared Error) và hệ số xác định R^2 được sử dụng khi tính toán thực nghiệm cho tập dữ liệu huấn luyện nhận được kết quả trong Bảng 2 bên dưới.

Các chỉ số đánh giá trong Bảng 2 cho thấy mô hình đạt độ chính xác cao, đặc biệt mô hình có hàm mất mát trong hình 10 có giá trị sai số thấp hơn và hệ số R^2 cao hơn phản ánh khả năng giải thích phương sai trong dữ liệu tốt hơn.

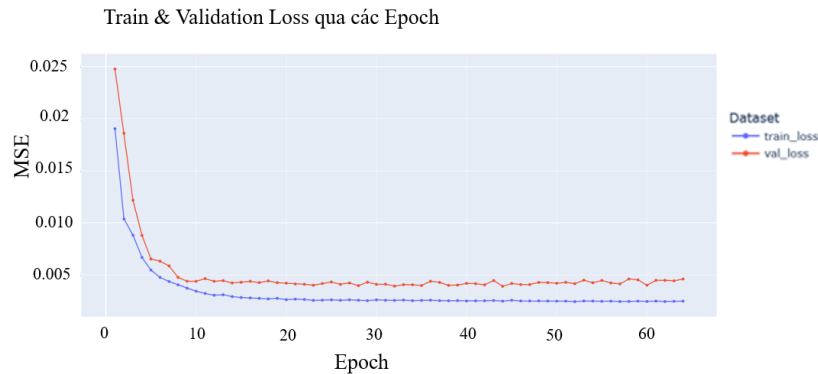
Qua đó cho thấy mô hình LSTM có tiềm năng cao trong việc dự báo chuỗi thời gian mực nước biển, với độ chính xác tốt và khả năng tổng quát hóa ổn định. Mô hình trong biểu đồ thứ hai có thể được lựa chọn là phương án tối ưu cho ứng dụng thực tế (trường hợp ngưỡng bằng 20).

4.2. Đánh giá mô hình trên tập kiểm thử (năm 2022)

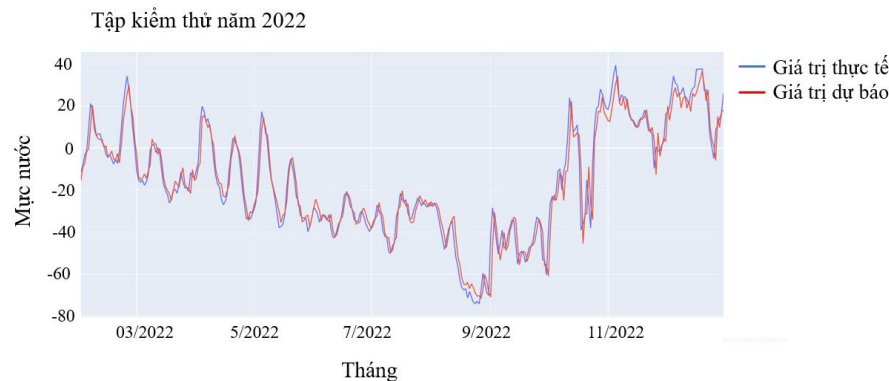
Dựa vào biểu đồ Hình 11 giữa mực nước thực tế và giá trị dự báo trên tập kiểm thử năm 2022



Hình 9. Biểu đồ hàm mất mát của tập huấn luyện và tập dữ liệu thực tế theo epoch cho trường hợp ngưỡng bằng 10.



Hình 10. Biểu đồ hàm mất mát của tập huấn luyện và tập dữ liệu thực tế theo epoch cho trường hợp ngưỡng bằng 20



Hình 11. Biểu đồ giữa dữ liệu thực tế và dữ liệu dự báo của tập kiểm thử năm 2022.

Bảng 2. Thống kê hiệu suất của mô hình LSTM với ngưỡng bằng 10 và 20.

TT	Ngưỡng	MAE (cm)	MSE (cm)	R^2
1	10	0,0355	0,0459	0,9402
2	20	0,0352	0,0453	0,9494

cho thấy đường dự báo (màu đỏ) bám rất sát theo đường thực tế (màu xanh), thể hiện khả năng học của mô hình như sau:

- Mô hình dự báo đúng các chu kỳ lên/xuống theo mùa vụ trong năm.
- Mặc dù có một số độ lệch nhẹ tại các điểm đỉnh/thấp cực trị, sai số tổng thể vẫn giữ ở mức ổn định.

4.3. Dự báo mực nước cho năm 2023 (01/2023÷12/2023)

Dựa trên mô hình đã huấn luyện, thực hiện dự báo mực nước biển hàng ngày cho giai đoạn từ 01/01/2023 đến 31/12/2023 bằng phương pháp dự báo cuốn trượt (recursive forecasting) đối với trường hợp ngưỡng bằng 20 (Hình 12) cho thấy:

- Mực nước dự báo có xu hướng tăng nhẹ trong đầu năm 2023 và nhanh chóng ổn định quanh mức 12 cm.
- Chuỗi kết quả có độ mượt cao do sử dụng đầu ra mô hình làm đầu vào liên tục, phản ánh tính ổn định cao.

Mặc dù mô hình dự báo khá ổn định, các ảnh hưởng từ yếu tố thời tiết, thủy triều thiên văn, hoặc biến đổi khí hậu có thể khiến dự báo lệch trong thực tế do mô hình hiện tại chỉ dùng dữ liệu đơn biến (mực nước H).

5. Kết luận

Bài báo này đã xây dựng và đánh giá một mô hình học sâu dựa trên mạng nơron hồi tiếp LSTM nhằm dự báo sự thay đổi mực nước biển tại khu vực Bà Rịa - Vũng Tàu cũng như cho khu vực Đông Nam Bộ, Việt Nam khi sử dụng chuỗi dữ liệu độ

cao triều đơn biến kéo dài từ năm 1999 đến 2022. Mô hình được huấn luyện trên dữ liệu 23 năm (1999÷2021) và kiểm thử trên năm 2022, sau đó sử dụng để dự báo mực nước cho giai đoạn 01/2023 đến 12/2023.

Các kết quả thực nghiệm cho thấy:

- Mô hình LSTM có khả năng học tốt chu kỳ thủy triều, đạt sai số cho tập dữ liệu kiểm thử thấp ($MSE = 0,0039$ cm; $MAE = 0,0449$ cm; $R^2 = 0,9443$).

- Dự báo mực nước cho hai năm tiếp theo có xu hướng ổn định và phù hợp với đặc điểm thủy triều khu vực tại Bà Rịa - Vũng Tàu.

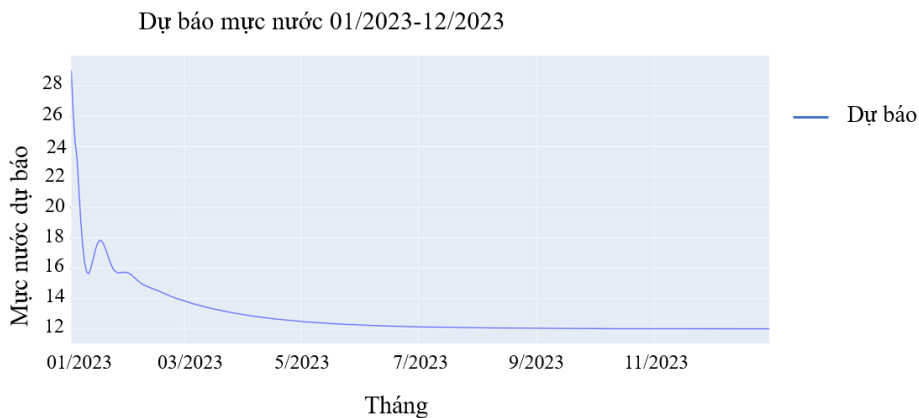
- Quy trình tiền xử lý, làm sạch outlier và chuẩn hóa dữ liệu đóng vai trò quan trọng trong việc tăng độ chính xác.

Kết quả trên cho thấy tiềm năng ứng dụng của mô hình LSTM trong các hệ thống cảnh báo sớm và hỗ trợ ra quyết định trong quản lý ven biển, đặc biệt trong bối cảnh biến đổi khí hậu ngày càng diễn biến phức tạp.

Tuy đạt được độ chính xác cao, nghiên cứu vẫn tồn tại hạn chế: mô hình chỉ sử dụng dữ liệu đơn biến (độ cao mực nước triều H), chưa khai thác các yếu tố ảnh hưởng như: gió, áp suất khí quyển, sóng,... Trong thời gian tới, nội dung này cần phải được tiếp tục nghiên cứu.

Đóng góp của tác giả

Huỳnh Nguyễn Định Quốc - đánh giá các kết quả thực nghiệm, biên tập bản thảo, xử lý dữ liệu thành phần và kiểm soát lỗi của chương trình; Nguyễn Gia Trọng - lên ý tưởng thiết kế, viết phần phương pháp, xử lý dữ liệu thành phần và hoàn thiện bản thảo.



Hình 12. Dự báo mực nước triều cho khoảng thời gian 01/2023 đến 12/2023.

Tài liệu tham khảo

- Abubakar, A. G., Mahmud, M. R., Tang, K. K. W., Hussaini, A., & Md Yusuf, N. H. (2019). A review of modelling approaches on tidal analysis and prediction. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42, 23-34. <https://doi.org/10.5194/isprs-archives-XLII-4-W16-23-2019>.
- Braakmann-Folgmann, A., Roscher, R., Wenzel, S., Uebbing, B., & Kusche, J. (2017). *Sea level anomaly prediction using recurrent neural networks* (arXiv:1710.07099v1) [Preprint]. arXiv. <https://arxiv.org/abs/1710.07099>.
- Brownlee, J. (2017). Long Short-Term Memory Networks With Python. *Machine Learning Mastery*.
- Ceder, N. (2018). *The quick Python book*: Simon and Schuster.
- Chen, D. Y. (2017). *Pandas for everyone: Python data analysis*: Addison-Wesley Professional.
- Chi, Y. N. (2022). Time series modeling and forecasting of monthly mean sea level (1978–2020): SARIMA and multilayer perceptron neural network. *International Journal of Data Science*, 3(1), 45–61. <https://doi.org/10.33140/IJDS.03.01.05>.
- Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263. <https://doi.org/10.1080/07350015.1995.10524599>.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Regularization for deep learning. *Deep learning*, 1, 224-270.
- Gupta, P. & Bagchi, A. (2024). Introduction to NumPy. In *Essentials of Python for Artificial Intelligence and Machine Learning* (pp. 127-159): Springer.
- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>.
- Karim, M. R. (2018). *TensorFlow: Powerful Predictive Analytics with TensorFlow: Predict valuable insights of your data with TensorFlow*. Packt Publishing Ltd.
- Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.
- Le, D. T., Huynh, N. D., & Nguyen, G. T. Q. (2024). Exploring the training results of machine learning models using different batch sizes and epochs: A case study with GNSS time series data. *J. Hydro-Meteorol*, 19, 90-99.
- Nguyen, N. V., & Nguyen, T. G. (2025). Predicting elevation values using Gated Recurrent Unit (GRU) neural networks. *Journal of Mining and Earth Sciences Vol*, 66(3), 29-38.
- Makky, N., Valizadeh Kamran, K., & Karimzadeh, S. (2024). Impact of Global Warming on Water Height Using XGBOOST and MLP Algorithms. *Environmental Sciences Proceedings*, 29(1), 83. <https://www.mdpi.com/2673-4931/29/1/83>.
- Pugh, D. T. (2002). Tides, surges and mean sea-level: A handbook for engineers and scientists. John Wiley & Sons.
- Raj, N., Gharineiat, Z., Ahmed, A. A. M., & Stepanyants, Y. (2022). Assessment and prediction of sea level trend in the South Pacific Region. *Remote Sensing*, 14(4), 986. <https://doi.org/10.3390/rs14040986>.
- Tosi, S. (2009). *Matplotlib for Python developers* (Vol. 307). Birmingham, UK: Packt Publishing.
- Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2, pp. 131-160). Reading, MA: Addison-wesley.
- Uluocak, I. (2025). Comparative study of multivariate hybrid neural networks for global sea level prediction through 2050. *Environmental Earth Sciences*, 84(3), 79. <https://link.springer.com/article/10.1007/s12665-025-12090-x>.
- Xie, Y., Zhou, S., & Wang, F. (2024). Prediction analysis of sea level change in the China adjacent seas based on singular spectrum analysis and long short-term memory network. *Journal of Marine Science and Engineering*, 12(8), Article 1397. <https://doi.org/10.3390/jmse12081397>.