DỰ BÁO MỰC NƯỚC TẠI TRẠM KHÁNH AN, TỈNH AN GIANG BẰNG MÔ HÌNH HỌC SÂU LSTM *(LONG SHORT - TERM MEMORY)*

NGUYỄN THỊ MỸ TRUYỀN[1], TRẦN NGỌC CHÂU[2], LƯU VĂN NINH[3]

[1] Khoa Công nghệ thông tin, Trường Đại học An Giang, Đại học Quốc gia TP. Hồ Chí Minh

[2] Khoa Kỹ thuật - Công nghệ - Môi trường, Trường Đại học An Giang, Đại học Quốc gia TP. Hồ Chí Minh

[3] Đài Khí tượng Thủy văn tỉnh An Giang

Tóm tắt

Dự báo mực nước là một công cụ quan trọng trong quản lý tài nguyên nước, đặc biệt đối với các vùng chịu ảnh hưởng mạnh của lũ lụt vào mùa mưa và khan hiếm nước vào mùa khô như khu vực đồng bằng sông Cửu Long (ĐBSCL). Tại đây, diễn biến mực nước có tính chu kỳ rõ rệt, chịu tác động đồng thời từ chế độ triều và dòng chảy thượng nguồn nên đòi hỏi các phương pháp dự báo có khả năng xử lý dữ liệu chuỗi thời gian phi tuyến và vượt trội hơn so với phương pháp truyền thống. Nghiên cứu này áp dụng mô hình học sâu LSTM (Long Short - Term Memory) để dự báo mực nước tại trạm Khánh An, tỉnh An Giang, nơi có vị trí quan trắc quan trọng trong hệ thống thủy văn sông Hậu. Hai kịch bản được thiết lập, sử dụng chuỗi dữ liệu đầu vào 24 giờ và 48 giờ để dự báo mực nước 6 giờ tiếp theo và quá trình huấn luyện được thực hiện với các giá trị epochs khác nhau (50, 100, 200, 300). Kết quả cho thấy mô hình đạt hiệu quả tốt nhất với dữ liệu đầu vào 48 giờ và 300 epochs, khi sai số bình phương trung bình trên tập kiểm tra (RMSE) đạt 6,894 và hệ số R^2 lên đến 0,997. Mô hình mô phỏng tốt các thời điểm cực trị và là công cụ hiệu quả dự báo mực nước theo mùa tại trạm Khánh An. Mô hình có tiềm năng ứng dụng rộng rãi trong cảnh báo lũ và quản lý hạn hán tại những khu vực chịu ảnh hưởng bởi biến đổi khí hậu và biến động dòng chảy như vùng ĐBSCL nói chung, tỉnh An Giang nói riêng.

Từ khoá: *Dự báo mực nước, mô hình học máy, LSTM, Khánh An.*

Ngày nhận bài: *30/5/2025;* ***Ngày sửa chữa:*** *10/6/2025;* ***Ngày duyệt đăng:*** *26/6/2025.*

Forecasting Water level at Khanh An station, An Giang province using LSTM deep learning model

Abstract

Water level forecasting is a vital tool in water resource management, especially for regions affected by flooding during the rainy season and water scarcity during the dry season, such as the Mekong Delta area. In this area, water level fluctuations exhibit a seasonal cycle, influenced simultaneously by tidal regimes and upstream flows. As such, forecasting methods must be capable of handling nonlinear time series data and outperform traditional approaches. This study employs a deep learning model, specifically the Long Short-Term Memory (LSTM) model, to forecast water levels at Khanh An station, situated in An Giang Province, a key monitoring gauge in the Hau River hydrological system. Two forecasting scenarios were developed, using 24-hour and 48-hour input sequences to predict the next 6 hours of water levels. The model was trained with number of epochs (50, 100, 200, and 300). Results showed that the model performs best with 48-hour input data and 300 epochs, achieving a Root Mean Square Error (RMSE) of 6.894 and a coefficient of determination (R^2) of 0.997 on the test set. The model accurately simulates extreme conditions and serves as an effective tool for seasonal water level forecasting at Khanh An station. It holds strong potential for broader application in flood warning, and drought management in regions significantly impacted by climate change and flow variability, such as the Mekong Delta in general and An Giang Province in particular.

Keywords: *Water level forecasting, machine learning model, LSTM, Khanh An.*

JEL Classifications: *Q50, Q51, Q54.*

1. GIỚI THIỆU

ĐBSCL là khu vực diễn ra hiện tượng ngập lụt tự nhiên thường niên trên diện rộng. Theo nghiên cứu, nguyên nhân chính chủ yếu dẫn tới hiện trạng này là do lượng mưa cao, trung bình hàng năm, tại ĐBSCL lượng mưa dao động khoảng từ 1.500 - 2.000 mm, kết hợp với một lượng tuyết tan đáng kể từ Tây Tạng, lượng mưa ở thượng - hạ Lào và Campuchia chảy về, tạo nên các trận lũ lụt. Bên cạnh đó, khi có sự tập trung các yếu tố, bao gồm nước lũ từ thượng nguồn, triều cường ở biển Đông và mưa liên tục tại khu vực thì ĐBSCL sẽ xảy ra ngập lụt cực lớn (Trang P. , 2016). Mỗi năm vào mùa lũ, tất cả các hoạt động thường ngày của con người, nền nông nghiệp, công nghiệp, kinh tế và cả cơ

sở hạ tầng đều bị ảnh hưởng nặng nề bởi lũ. Ngược lại, đến mùa khô, lũ rút làm cho mực nước xuống thấp gây ra tình trạng khô hạn và thiếu nước trầm trọng.

Tỉnh An Giang, thuộc ĐBCSL có tình hình ngập lụt khá phức tạp và diễn biến bất thường do lũ từ thượng nguồn sông Mêkông, mưa to nội khu vực và triều cường. Lũ tại An Giang thuộc khu vực Nam bộ thường diễn ra từ tháng 6 đến tháng 11 theo Đài Khí tượng thủy văn An Giang, trùng với thời gian diễn ra mùa mưa từ đầu tháng 5 đến giữa tháng 11, chiếm khoảng 90% lượng mưa cả năm (Ninh L. , 2017). Vì vậy, vào khoảng thời gian này tỉnh không chỉ xảy ra ngập lụt, ngập úng cục bộ ở vùng trũng thấp, vùng ven sông, kênh rạch và khu vực không có hệ thống đê bao mà còn có khả năng cao xuất hiện tình trạng sạt lở bờ sông gây thiệt hại về người, tài sản, đặc biệt là ảnh hưởng nghiêm trọng đến hoạt động nông nghiệp. Để chủ động trong việc ứng phó với lũ, triều cường, khô hạn, tỉnh đã có nhiều giải pháp và đầu tư chi phí cho các hệ thống quản lý, cảnh báo, theo dõi tình hình mực nước trên sông. Đồng thời, rà soát, cập nhật chính sách, kế hoạch trong phòng, chống, giảm thiệt hại bởi lũ lụt tại địa phương. Tuy nhiên, các giải pháp vẫn chưa thực sự hiệu quả trong việc dự báo và kiểm soát lũ trong tương lai, một phần do sự hạn chế về áp dụng công nghệ hiện đại.

Hiện nay, nhiều giải pháp mới, tiềm năng liên quan đến công nghệ được nghiên cứu ứng dụng để hỗ trợ hiệu quả công tác dự báo mực nước trong kiểm soát lũ lụt. Sự phối hợp giữa công nghệ số hóa cùng với các mô hình học máy đang trở thành những công cụ đắc lực và rất mạnh mẽ trong việc hỗ trợ giải quyết hiệu quả các vấn đề phức tạp liên quan đến tính toán và dự báo. Theo kết quả nghiên cứu về dự báo mực nước gần đây cho thấy, các mô hình học máy là công cụ tiềm năng rất lớn, bởi vì mô hình dự báo có thể được xây dựng nhanh chóng, dễ dàng và không đòi hỏi phải có sự hiểu biết sâu về các quá trình vật lý ẩn đằng sau. Bên cạnh đó, khả năng tính toán, hiệu chỉnh, kiểm định nhanh hơn so với các mô hình vật lý truyền thống và cách sử dụng ít phức tạp hơn (Mekanik F., 2013).

Nhóm nghiên cứu Thư T., 2019 đã chứng minh rằng mô hình LSTM tối ưu hơn các mạng nơ-ron truyền thống khác khi xử lý vấn đề liên quan đến dự đoán chuỗi thời gian với bộ dữ liệu quan trắc mực nước đặt tại 4 trạm trên sông Mêkông giai đoạn 2012 - 2016. Atashi V., 2022 nghiên cứu sử dụng phương pháp máy học sâu để dự đoán lũ bao gồm các mô hình Seasonal Autoregressive Integrated Moving Average (SARIMA), Random Forest (RF) và Long Short - Term Memory (LSTM). Kết quả nhận định rằng, mô hình LSTM vượt trội hơn hẳn các mô hình còn lại. Tại Yeojubo, tỉnh Gyeonggi-do, Hàn Quốc đã thực hiện dự báo mực nước nhằm phục vụ trong công tác quản lý lũ lụt bằng cách sử dụng mô hình LSTM kết hợp với gated recurrent unit (GRU). Dữ liệu đầu vào được sử dụng trong mô hình này là dữ liệu khí tượng, gồm các bộ dữ liệu về mực nước ngược và xuôi dòng, nhiệt độ, độ ẩm, lượng mưa (Minwoo C., 2022).

Hơn nữa, Yu L., 2021 đã dự báo mực nước sông bằng cách ứng dụng mô hình LSTM, kết quả cho thấy, mô hình này có thể dự báo một cách hiệu quả trong điều kiện khoảng thời gian luân phiên là 30 phút và giai đoạn dự báo trong 2 giờ tại 5 vị trí lấy số liệu. Một nghiên cứu của Hiền L., 2018 cũng sử dụng mô hình LSTM với dữ liệu mô phỏng là mực nước theo giờ tại các trạm thủy văn, dự báo từ 1 giờ - 5 giờ. Mô hình không bao gồm các dữ liệu về khí hậu, địa hình và cho kết quả dự báo có độ chính xác cao. Ngoài ra, tại sông Cấm, Hải Phòng đã được thực hiện dự báo mực nước bằng mô hình Long Short - Term Memory Neural Network (LSTM), một dạng của Mạng nơ-ron hồi quy (Recurrent Neural Network) với dữ liệu đầu vào là mực nước tại các trạm thủy văn. Mô hình cho ra kết quả có độ chính xác cao, đủ độ tin cậy để ứng dụng vào dự báo mực nước trong quản lý (Hùng H., 2021). Một nghiên cứu khác áp dụng ba mô hình gồm Support Vector Regression (SVR), LSTM và mô hình kết hợp giữa LSTM với SVR để dự báo mực nước với các dữ liệu có sẵn như lượng mưa, lượng mưa tích lũy, độ cao mực nước sông. Nghiên cứu chỉ ra rằng mô hình LSTM cho ra kết quả đạt tỷ lệ lỗi thấp nhất, tuy nhiên không thể nắm bắt được những thay đổi nhanh chóng trong bộ dữ liệu (Punyanuch B., 2022).

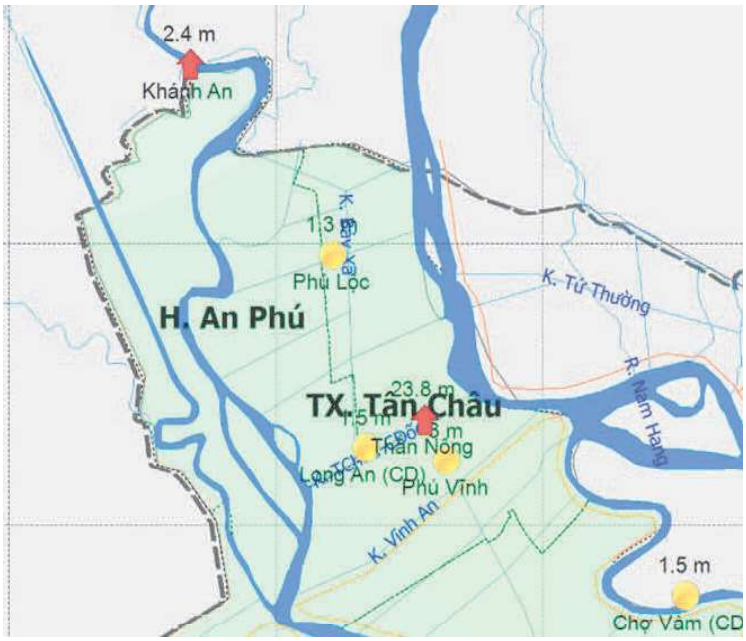
Giải pháp dự báo mực nước bằng mô hình học máy được đề xuất ở đây không đòi hỏi phải có sự hiểu biết sâu về các quá trình vật lý phức tạp bên trong như các mô hình truyền thống. Dựa vào dữ liệu thu thập được từ các trạm quan trắc trên sông, công tác dự báo trở nên đơn giản hơn, kết quả có độ chính xác cao, nhanh chóng, hiệu quả, giúp tiết kiệm chi phí và nhận biết sớm được tình hình mực nước trong thời gian sắp tới để kịp thời có những giải pháp ứng phó chủ động nhằm phát huy tối đa hiệu quả kinh tế địa phương. Thêm vào đó, tỉnh An Giang hiện đang có nguồn dữ liệu số được tạo ra từ các trạm quan trắc ngày càng lớn, độ tin cậy cao và trở thành nguồn tài nguyên dữ liệu quý giá trong chuyên môn. Do đó, việc thực hiện nghiên cứu "Dự báo mực nước tại trạm Khánh An, huyện An Phú, tỉnh An Giang bằng mô hình học máy" sẽ góp phần tìm ra những giải pháp dự báo mực nước hiệu quả, tin cậy về mặt khoa học

trong công tác dự báo thông qua các chỉ số về độ đo lỗi Root Mean Square Error (RMSE) và coefficient of determination (R^2) của mô hình học sâu LSTM.

2. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Đối tượng nghiên cứu

Nghiên cứu được thực hiện trên đối tượng dữ liệu mực nước theo giờ trong 8 năm (từ năm 2016 - 2023), tại trạm quan trắc Khánh An thuộc huyện An Phú, tỉnh An Giang. Vị trí của trạm Khánh An nằm ở đầu nguồn sông Hậu chảy qua địa phận tỉnh An Giang. Thông tin mực nước tại trạm này giữ vai trò quan trọng đối với vùng hạ lưu sông Hậu cũng như các vùng khác trong tỉnh An Giang (Hình 1).



Hình 1. Trạm quan trắc tại Khánh An, huyện An Phú, tỉnh An Giang

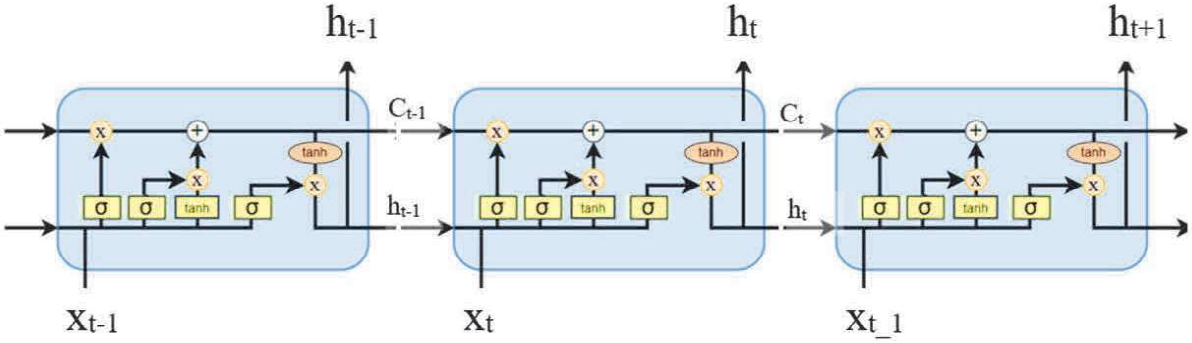
2.2. Phương pháp nghiên cứu

a. Mô hình máy học

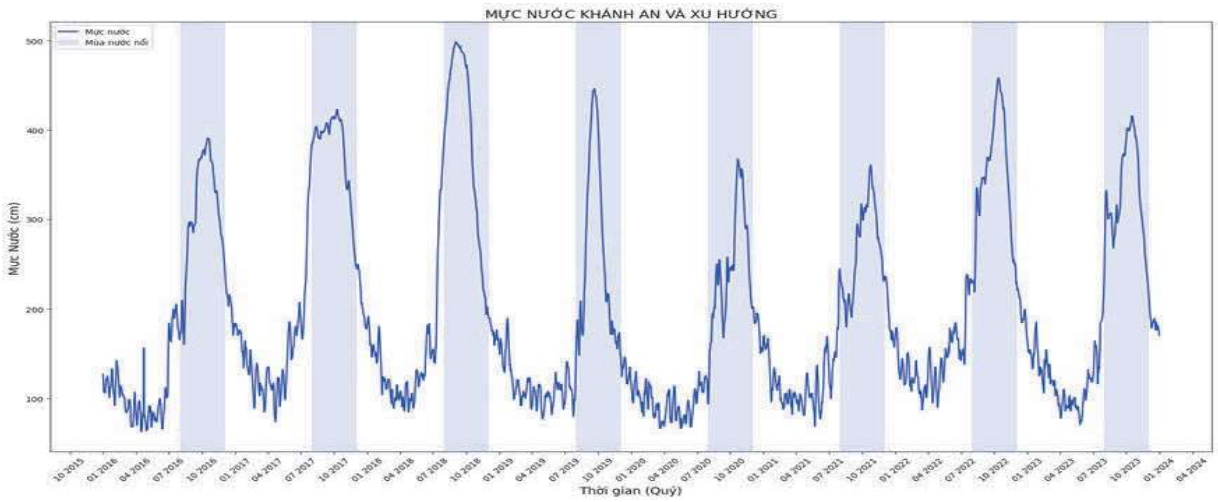
Học máy là một lĩnh vực của trí tuệ nhân tạo, tập trung vào việc phát triển các thuật toán và mô hình giúp máy tính tự động học hỏi, cải thiện hiệu suất từ dữ liệu hoặc kinh nghiệm mà không cần được lập trình chỉ dẫn. Học máy hướng tới việc giúp máy móc trở nên thông minh hơn trong việc dự đoán, ra quyết định hoặc xử lý vấn đề phức tạp nhờ vào bộ dữ liệu đủ lớn, đủ tin cậy mà không cần con người can thiệp quá nhiều vào quá trình xử lý.

Mạng nơ-ron hồi quy (RNN) là mô hình học máy xử lý dữ liệu theo chuỗi bằng cách sử dụng thông tin từ quá khứ để dự đoán hiện tại. RNN gồm ba lớp: Đầu vào, ẩn và đầu ra, trong đó, lớp ẩn có khả năng ghi nhớ tạm thời thông tin trước đó. Tuy nhiên, RNN gặp khó khăn với các phụ thuộc xa do giới hạn bộ nhớ. Mạng LSTM (Long Short - Term Memory) là phiên bản cải tiến của RNN, được Hochreiter và Schmidhuber đề xuất năm 1997. LSTM khắc phục hạn chế của RNN bằng cách tự động học và ghi nhớ các phụ thuộc xa. Nhờ cơ chế chọn lọc thông tin cần nhớ hoặc quên, LSTM hoạt động hiệu quả với dữ liệu chuỗi thời gian, có khả năng ghi nhớ lâu dài mà không cần huấn luyện đặc biệt.

Trạng thái tế bào là yếu tố cốt lõi trong kiến trúc LSTM, hoạt động như một băng chuyền truyền tải thông tin xuyên suốt qua các bước thời gian mà không bị biến đổi. LSTM điều chỉnh thông tin trong trạng thái tế bào thông qua ba cổng chính: Cổng quên (forget gate), cổng đầu vào và cổng đầu ra. Các cổng này sử dụng hàm sigmoid để quyết định mức độ thông tin được giữ lại hay loại bỏ, với đầu ra nằm trong khoảng [0, 1]. Nhờ vậy, LSTM kiểm soát hiệu quả việc ghi nhớ và loại bỏ thông tin qua từng thời điểm, đảm bảo duy trì những đặc trưng quan trọng trong quá trình học. Một mô-đun trong mạng gồm 3 cổng như vậy nhằm mục đích kiểm soát trạng thái tế bào (Hình 2).



Hình 2. Băng chuyền giữa các tế bào trong mô hình LSTM



Hình 3. Diễn biến mực nước tại trạm Khánh An từ năm 2016 - 2023

b. Mô phỏng mực nước bằng mô hình máy học

Thu thập dữ liệu mực nước:

Việc lựa chọn dữ liệu đầu vào cho mô hình LSTM là rất quan trọng và quyết định tính hiệu quả của việc dự báo. Dựa trên nhiều phân tích, kết quả nghiên cứu liên quan, nhóm đã sử dụng bộ dữ liệu mực nước thu được theo từng giờ tại trạm quan trắc Khánh An, một bộ dữ liệu đơn biến khoảng hơn 70.000 dòng làm dữ liệu đầu vào cho mô hình.

Bộ dữ liệu mực nước từ năm 2016 - 2023 được chia thành 2 tập dữ liệu với tỷ lệ 80%, dùng để huấn luyện các mô hình máy học (từ ngày 1/1/2016 - 25/5/2022) và phần còn lại là tập kiểm tra (testing set) 20%, dùng để đánh giá tính hiệu quả của mô hình.

Xây dựng mô hình học máy, huấn luyện và đánh giá kết quả mô hình:

Nhóm nghiên cứu chọn ngôn ngữ lập trình Python chạy trên Google Colab để triển khai dự đoán mực nước bằng mô hình LSTM bởi đây là công cụ, là nền tảng thuận lợi để sử dụng các gói thư viện mã nguồn mở như Keras và Sklearn.

Quá trình huấn luyện mạng sẽ thực hiện điều chỉnh các trọng số của mô hình nhằm tìm ra bộ trọng số tối ưu sao cho giá trị hàm mất mát đạt giá trị nhỏ nhất và để xác định các chỉ số về độ đo lỗi RMSE và R^2. Thước đo quan trọng để đánh giá mô hình là sai số trung bình phương (RMSE) và hệ số xác định (R^2).

Hệ số RMSE có giá trị càng nhỏ thì mô hình càng tốt, được tính toán theo công thức sau:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (1)$$

Trong đó: yi: Giá trị thực tế;

$\hat{y}_i$: Giá trị dự đoán; n: Tổng số điểm dữ liệu.

Hệ số R^2 có giá trị nằm ở khoảng từ 0 - 1. Nếu giá trị kết quả càng gần 1 thì càng tốt, giá trị mô phỏng tương đồng với giá trị đo thực và ngược lại, giá trị càng gần 0 càng kém, giá trị mô phỏng tương đồng với giá trị đo thực. Công thức tính toán R^2 như sau (Dongfen R., 2025):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{\sum_{i=1}^{n}(\bar{y}_i - \hat{y}_i)} \qquad (2)$$

Trong đó: yi: Giá trị thực tế;

$\hat{y}_i$: Giá trị dự đoán;

$\bar{y}_i$: Giá trị trung bình;

n: Tổng số điểm dữ liệu.

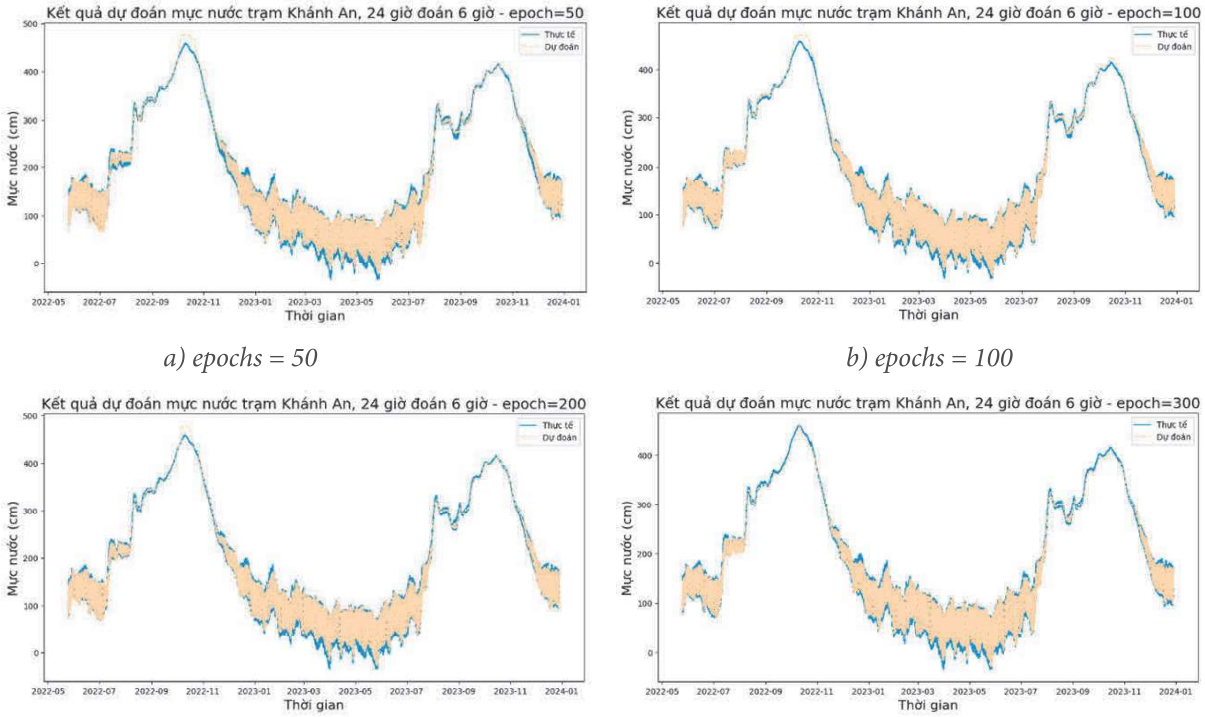
Hiệu chỉnh kết quả mô hình: Sau khi huấn luyện mô hình, các chỉ số RMSE và R^2 được xem xét để tiến hành điều chỉnh các thông số bên trong mô hình để đạt kết quả tốt nhất trong dự đoán.

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Diễn biến mực nước từ năm 2016 - 2023 tại trạm Khánh An, tỉnh An Giang

Trong suốt giai đoạn từ năm 2016 - 2023 (Hình 3), mực nước tại trạm Khánh An thể hiện rõ một chu kỳ thủy văn lặp lại hàng năm và bị ảnh hưởng trực tiếp từ hệ thống sông Mêkông cũng như chế độ triều. Mỗi năm, mực nước bắt đầu tăng từ khoảng tháng 6 và đạt đỉnh vào khoảng tháng 8 - 11, đây là giai đoạn triều cường kết hợp với mùa lũ, tạo nên những đỉnh mực nước cao nhất trong năm, thường đạt trên 300 - 500 cm, có thể gây ra tình trạng ngập lụt ở các khu vực trũng nếu kết hợp với mưa lớn và hệ thống thoát nước kém.

Ngược lại, vào các tháng đầu năm (tháng 1 - 3) và cuối năm (tháng 9 -12), mực nước thường giảm xuống mức thấp nhất, chỉ dao động từ khoảng 60 - 120 cm. Đây là thời kỳ triều kiệt thường rơi vào mùa khô, thể hiện tình trạng thiếu hụt nguồn nước mặt, đặc biệt ảnh



a) epochs = 50 *b) epochs = 100*

c) epochs = 200 *d) epochs = 300*

Hình 4. Kịch bản dữ liệu đầu vào 24 giờ dự đoán 6 giờ tiếp theo

hưởng đến sản xuất nông nghiệp và sinh hoạt ở vùng ven sông, kênh rạch (Hình 3).

Vào năm 2019 và 2023, mực nước tại trạm Khánh An có đỉnh triều cường rất cao, cho thấy ảnh hưởng từ lũ lớn hoặc hiện tượng thời tiết cực đoan. Trong khi đó, giai đoạn 2021 - 2022, đỉnh lũ, mực nước triều cường đều ở mức thấp, dẫn đến hạn hán và dòng chảy về hạ lưu bị suy giảm, điều này có thể liên quan đến hoạt động điều tiết nước ở thượng nguồn và biến đổi khí hậu. Nhìn chung, chu kỳ mực nước tại Khánh An mang tính mùa vụ rõ rệt, lặp lại hàng năm, nhưng có sự dao động về biên độ giữa các năm. Việc theo dõi mực nước triều cường, triều kiệt qua nhiều năm là cơ sở quan trọng để dự báo lũ và xây dựng kế hoạch ứng phó thiên tai hiệu quả hơn trong bối cảnh biến đổi khí hậu đang ngày càng rõ rệt.

3.2. Dự báo mực nước bằng mô hình máy học LSTM tại trạm Khánh An, tỉnh An Giang

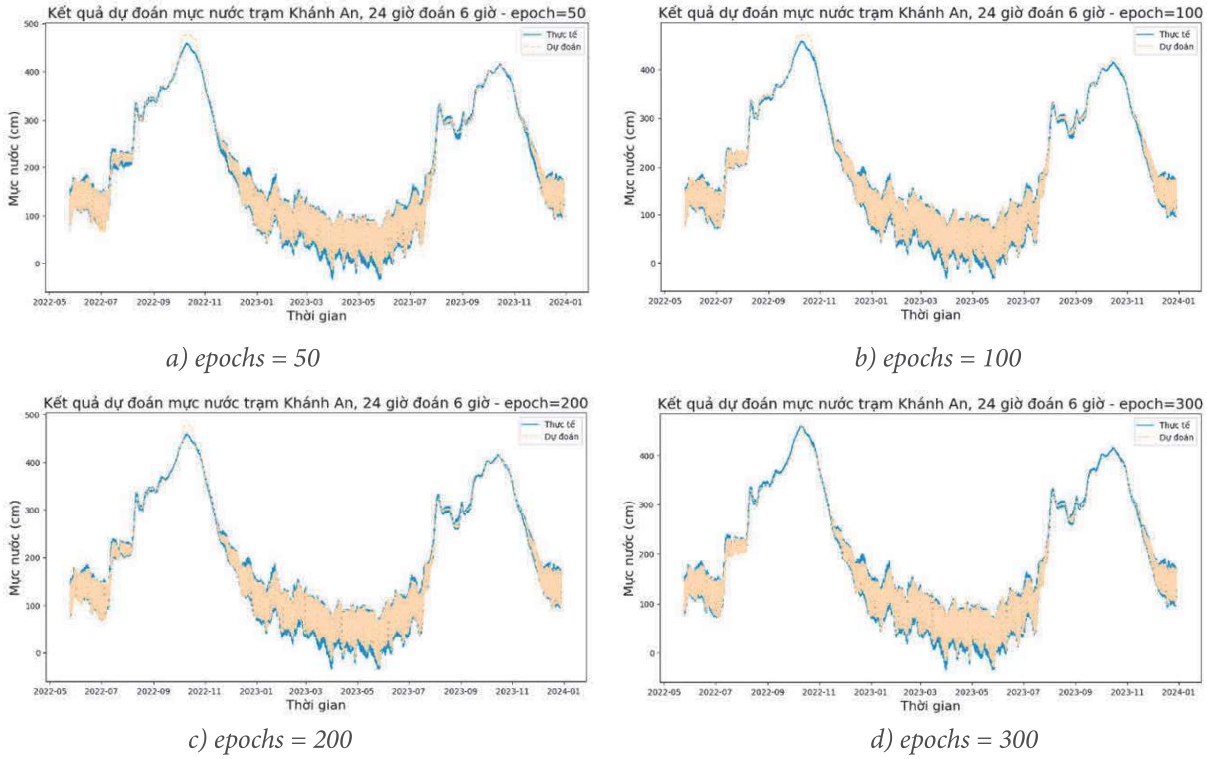
Trong nghiên cứu sử dụng mô hình LSTM để dự báo mực nước, độ dài chuỗi dữ liệu đầu vào đóng vai trò quan trọng, do đặc tính chu kỳ ngày đêm và thủy triều của mực nước. Do đó, chuỗi 24 giờ, 48 giờ được lựa chọn để dự báo 6 giờ tiếp theo, phù hợp với yêu cầu cảnh báo sớm và ngắn hạn. Khung thời gian này giúp cân bằng giữa độ ổn định về thủy văn, tránh quá khớp về mặt thuật toán. Mô hình 24 giờ giúp ghi nhớ biến động ngắn hạn, trong khi 48 giờ hỗ trợ nhận diện xu hướng dài hơn. Quá trình huấn luyện sử dụng các giá trị khác nhau của số epoch (50, 100, 200, 300) để đánh giá hiệu suất theo độ sâu huấn luyện, tuy nhiều epoch giúp mô hình học tốt hơn, nhưng nếu quá mức có thể gây quá khớp và giảm khả năng tổng quát hóa.

Do đó, Nghiên cứu đã tiến hành mô phỏng mực nước với 2 kịch bản: (1) Dùng dữ liệu đầu vào của 24 giờ để dự đoán 6 giờ tiếp theo; (2) Dùng dữ liệu đầu vào của 48 giờ để dự đoán 6 giờ tiếp theo. Đồng thời nghiên cứu thực hiện mô phỏng mực nước với các tham số epochs ở các giá trị khác nhau (50, 100, 200, 300).

Kết quả chạy mô hình LSTM trên bộ dữ liệu mực nước tại trạm Khánh An từ năm 2016 - 2023 cụ thể với tập huấn luyện 80% tạm tính từ ngày 1/1/2016 - 25/5/2022. Phần dữ liệu còn lại cho đến ngày 31/12/2023 sử dụng để kiểm tra và đánh giá kết quả chạy mô hình. Kết quả được thể hiện qua các biểu đồ biểu diễn giá trị dự đoán so với các giá trị thực đo từ tháng 5/2022 - 12/2023 (Hình 4 và Hình 5).

3.2.1. Kịch bản dữ liệu đầu vào 24 giờ dự đoán 6 giờ tiếp theo

Mô hình học máy LSTM đã được áp dụng để dự báo mực nước tại trạm Khánh An, tỉnh An Giang với kịch bản dữ liệu đầu vào 24 giờ để dự đoán 6 giờ tiếp theo. Kết quả mô phỏng trên bộ dữ liệu từ tháng



a) epochs = 50 *b) epochs = 100*

c) epochs = 200 *d) epochs = 300*

Hình 5. Kịch bản dữ liệu đầu vào 48 giờ dự đoán 6 giờ tiếp theo

5/2022 -12/2023 cho thấy khả năng của mô hình và sự ảnh hưởng của số lượng epochs.

Mô hình học sâu LSTM cho thấy khả năng mô phỏng mực nước hiệu quả tại trạm Khánh An trong điều kiện thủy văn theo mùa, nhờ khả năng ghi nhớ chuỗi thời gian dài hạn và xử lý dữ liệu phi tuyến. Kết quả từ bốn kịch bản kiểm tra (Hình 4a - 4d) cho thấy hiệu suất của mô hình tăng dần theo số lượng epochs. Đặc biệt, mô hình đạt độ chính xác cao nhất tại 300 epochs (Hình 4d), thể hiện qua sự tương quan gần như đồng nhất giữa giá trị mực nước mô phỏng và giá trị mực nước thực tế. Đây là giai đoạn mà mô hình hội tụ ổn định, giảm thiểu được sai số dự báo. Mức độ khớp giữa dữ liệu dự báo và dữ liệu thật đặc biệt rõ rệt tại các thời điểm triều cường, triều kiệt, thời điểm có độ dao động lớn, nhạy cảm với chất lượng dữ liệu mô phỏng.

Từ các kết quả trên (Hình 4), có thể kết luận rằng mô hình LSTM là công cụ dự báo hiệu quả đối với bài toán mô phỏng mực nước trong điều kiện thủy văn theo mùa tại trạm Khánh An, tỉnh An Giang với bộ dữ liệu đầu vào 24 giờ. Việc lựa chọn số lượng epochs phù hợp (300 epochs) là cần thiết để đảm bảo hiệu suất mô hình và độ tin cậy của quá trình dự báo.

3.2.2. Kịch bản dữ liệu đầu vào 48 giờ dự đoán 6 giờ tiếp theo

Nhìn chung với sự khác biệt giữa dữ liệu thực tế và dữ liệu dự báo mực nước thì mô hình LSTM thể hiện khả năng dự báo mực nước khá tốt ở cả hai kịch bản (24 giờ và 48 giờ). Tuy nhiên, độ chênh lệch rõ rệt hơn tại các điểm cực trị như đỉnh và đáy của mực nước.

Trong kịch bản sử dụng dữ liệu đầu vào 48 giờ để dự báo mực nước 6 giờ tiếp theo tại trạm Khánh An, mô hình học sâu LSTM cho thấy hiệu suất mô phỏng tăng dần theo số lượng epochs (Hình 5a - 5d), với số epochs lần lượt là 50, 100, 200 và 300. Sai số dự báo giảm rõ rệt; độ tương thích giữa giá trị mô phỏng và giá trị mực nước thực tế ngày càng cao. Khi số lần lặp lại (epochs) là 50 (Hình 5a), sự tương thích giữa mực nước thực tế và dự báo là thấp nhất trong khoảng thời gian từ tháng 9 - 11/2022 và năm 2023. Khi tăng epochs lên 100 (Hình 5b), độ tương thích có cải thiện nhưng vẫn còn thấp ở các khoảng thời gian tương tự.

Riêng tại epochs = 300 (hình 5d), dữ liệu mô phỏng và dữ liệu thực tế gần trùng nhau, thể hiện sự khớp giữa 2 bộ dữ liệu tốt nhất trong 4 giá trị epochs, mô phỏng chính xác cả các điểm cực trị như đỉnh triều, đáy triều, thường khó dự báo do tính phi tuyến và dao động mạnh của chuỗi mực nước. Điều này cho thấy mô hình không chỉ học tốt xu hướng chính mà còn tái hiện được cấu trúc dao động chi tiết trong chuỗi dữ liệu.

Tổng thể, kịch bản đầu vào 48 giờ giúp mô hình khai thác tốt hơn các đặc trưng thủy văn theo mùa và

Bảng 1. Các giá trị RMSE, R^2 trong quá trình huấn luyện và kiểm tra mô hình máy học LSTM theo 2 kịch bản khác nhau

Kịch bản	epochs	RMSE (Train)	RMSE (Test)	R^2 (Test)	Hội tụ
24 đoán 6	50	8,433	8,879	0,995	37 epochs
	100	8,055	8,709	0,995	70 epochs
	200	7,325	7,427	0,997	74 epochs
	300	7,470	7,264	0,997	94 epochs
48 đoán 6	50	7,429	8,099	0,996	50 epochs
	100	7,005	7,122	0,997	74 epochs
	200	7,033	7,369	0,997	66 epochs
	300	**6,623**	**6,894**	0,997	68 epochs

Ghi chú: Train RMSE là độ sai lệch trên tập huấn luyện; Test RMSE là độ sai lệch trên tập kiểm tra

với thời gian dài hạn trong dữ liệu mực nước. Kết quả mô phỏng ở epochs = 300 chứng minh đây là cấu hình huấn luyện tối ưu do đảm bảo mô hình hội tụ ổn định. Do đó, việc sử dụng mô hình LSTM với đầu vào 48 giờ là hướng tiếp cận hiệu quả cho các ứng dụng dự báo mực nước ngắn hạn, hỗ trợ ra quyết định trong công tác quản lý lũ, triều cường, điều tiết nguồn nước ở khu vực ĐBSCL nói chung, sông Hậu (tỉnh An Giang) nói riêng.

3.3. Đánh giá khả năng dự báo của mô hình máy học LSTM với bộ dữ liệu mực nước

Nghiên cứu đã chứng minh khả năng ứng dụng của mô hình LSTM trong dự báo mực nước tại trạm Khánh An, tỉnh An Giang. Cả hai kịch bản đầu vào 24 giờ và 48 giờ đều cho thấy giá trị mô phỏng có xu hướng sát với mực nước thực tế, đặc biệt khi số lượng epochs được điều chỉnh phù hợp. Đối với kịch bản dữ liệu đầu vào 48 giờ, 300 epochs được xác định là ngưỡng hiệu quả cao cho mô hình. Để có một so sánh định lượng và khách quan hơn giữa hai kịch bản, cần có các chỉ số đánh giá hiệu suất cụ thể như RMSE, R^2 cho cả hai kịch bản trên cùng một tập dữ liệu kiểm tra.

Dựa trên các chỉ số thống kê được trình bày trong Bảng 1 về hiệu suất mô hình học sâu LSTM ứng dụng cho bài toán mô phỏng và dự báo mực nước tại trạm Khánh An cho thấy, bộ giá trị RMSE-test và RMSE-train khá tương đồng và phù hợp. Đây là tín hiệu rất tốt, thể hiện mô hình không bị hiện tượng quá khớp (overfiting). Giá trị R^2 trong quá trình kiểm tra mô hình đều lớn hơn 0,99 ở tất cả mọi trường hợp, mô hình có khả năng dự đoán cực kỳ tốt, gần như hoàn hảo trên tập dữ liệu chưa từng được sử dụng trong quá trình huấn luyện. Đồng thời, kết quả đưa ra đánh giá chi tiết về hai kịch bản huấn luyện với đầu vào lần lượt là 24 giờ và 48 giờ, trong khi cùng dự báo mực nước cho 6 giờ tiếp theo.

Ở kịch bản 1 (đầu vào 24 giờ), mô hình cho thấy xu hướng cải thiện hiệu suất khi tăng số epochs huấn luyện, với RMSE trên tập kiểm tra giảm từ 8,879 (50 epochs) xuống còn 7,264 (300 epochs) và hệ số xác định R^2 tăng từ 0,995 lên 0,997. Điều này phản ánh khả năng học của LSTM khi được cung cấp thời gian huấn luyện đủ dài để khai thác các đặc trưng có chu kỳ trong dữ liệu thủy văn. Tuy nhiên, mặc dù cải thiện đáng kể, độ sai số RMSE vẫn còn cao hơn so với kịch bản thứ hai.

Ở kịch bản 2 (đầu vào 48 giờ), kết quả mô phỏng mực nước thể hiện độ chính xác vượt trội. Tại 300 epochs, mô hình đạt RMSE chỉ còn 6,894 và R^2 tương tự kịch bản 1 (0,997), có thể thấy mô hình có khả năng mô phỏng gần như toàn bộ biến động của mực nước thực tế. Ngoài ra, RMSE và R^2 trên cả tập huấn luyện, kiểm tra đều ổn định, cho thấy mô hình không bị hiện tượng quá khớp (overfitting), đồng thời khai thác hiệu quả mối quan hệ phụ thuộc thời gian dài hạn trong chuỗi dữ liệu.

Từ các kết quả này có thể kết luận rằng, việc sử dụng chuỗi đầu vào 48 giờ giúp tăng cường khả năng trích xuất đặc trưng động lực học của hệ thống thủy văn, nhờ đó mô hình LSTM nâng cao hiệu quả dự báo. Do đó, kịch bản tối ưu trong bối cảnh này là mô hình LSTM với dữ liệu đầu vào 48 giờ và huấn luyện 300 epochs, phù hợp cho các ứng dụng dự báo mực nước ngắn hạn trong công tác quản lý lũ và vận hành hệ thống thủy lợi.

4. KẾT LUẬN

Mô hình máy học LSTM thể hiện khả năng mô phỏng mực nước rất tốt tại trạm Khánh An với chuỗi dữ liệu có tính chu kỳ theo mùa, đặc biệt trong điều kiện biến động mạnh vào mùa lũ và triều cường.

Trong cả hai kịch bản đầu vào 24 giờ và 48 giờ, hiệu suất mô hình tăng dần theo số lượng epochs, trong đó kịch bản 48 giờ đầu vào với 300 epochs cho kết quả tốt nhất với RMSE thấp (6,894) và R^2 cao (0,997), phản ánh khả năng dự báo chính xác, ổn định. Các giá trị sai số RMSE trên tập huấn luyện và kiểm tra tương đối đồng đều, cùng với hệ số R^2 luôn trên 0,99, cho thấy mô hình hội tụ tốt, không xảy ra hiện tượng quá khớp (overfitting).

Kết quả dự báo thể hiện mô hình bám sát các đặc trưng thủy văn phức

Đầu nguồn sông Hậu, đoạn chảy qua địa phận tỉnh An Giang

tạp, mô phỏng chính xác các điểm cực trị (đỉnh triều, đáy triều), cho thấy tiềm năng ứng dụng cao trong quản lý nguồn nước và cảnh báo sớm thiên tai.

Tuy nhiên, nghiên cứu chưa thực hiện tối ưu toàn diện về cấu trúc mô hình (số lớp, batch-size, learning-rate…) và hàm kích hoạt, dẫn đến giới hạn trong việc giảm sai số về mức thấp hơn nữa.

Vì vậy, nghiên cứu tiếp theo có thể tích hợp thêm các hướng cải tiến mô hình như tối ưu siêu tham số (hyperparameter tuning), thử nghiệm mô hình kết hợp (LSTM-Attention hoặc LSTM-CNN) và bổ sung dữ liệu khí tượng (mưa, dòng chảy thượng nguồn) cho biến đầu vào để cải thiện chất lượng dự báo và giảm giá trị sai số, đặc biệt tại thời điểm mực nước biến động mạnh như mùa lũ.

Lời cảm ơn: *Nhóm tác giả xin cảm ơn Trường Đại học An Giang, Đại học Quốc gia TP. Hồ Chí Minh đã tạo điều kiện cho nhóm thực hiện Đề tài nghiên cứu "Ứng dụng mô hình máy học dự báo mực nước tại trạm Khánh An và Châu Đốc", cùng đơn vị phối hợp hỗ trợ dữ liệu - Đài Khí tượng Thủy văn tỉnh An Giang*■

TÀI LIỆU THAM KHẢO

1. Phạm Thị Huyền Trang và Trương Văn Tuấn (2016). Lũ lụt ở ĐBSCL: Nguyên nhân và giải pháp. Tạp chí Khoa học Đại học Sư phạm TP. Hồ Chí Minh, số 3 (81).

2. Đài Khí tượng Thủy văn tỉnh An Giang. Kiến thức thủy văn.

3. Lưu Văn Ninh và Nguyễn Minh Giám (2017). Đặc điểm khí hậu tỉnh An Giang. Tạp chí Khí tượng Thủy văn, số tháng 12.

4. Mekanik, F., Imteaz, M. A., Gato-Trinidad, S., và Elmahdi, A. (2013). Multiple regression and Artificial Neural Network for long-term rainfall forecasting using large scale climate modes. Journal of Hydrology, 503, 11–21. https://doi.org/10.1016/j.jhydrol.2013.08.035.

5. Trần Nguyễn Minh Thư, Nguyễn Hồng Hải và Phạm Trường An. (2019). Dự báo mực nước sông Mekong sử dụng LSTM và dữ liệu quan trắc thượng nguồn. Hội nghị khoa học công nghệ quốc gia lần thứ XII (FAIR). https://doi.org/10.15625/vap.2019.00016.

6. Atashi, V., Gorji, H. T., Shahabi, S. M., Kardan, R., & Lim, Y. H. (2022). Water Level Forecasting Using Deep Learning Time-Series Analysis: A Case Study of Red River of the North. Water, 14 (12), 1971. https://doi.org/10.3390/w14121971.

7. Minwoo Cho, Changsu Kim, Kwanyoung Jung và Hoekyung Jung (2022). Water Level Prediction Model Applying a Long Short-Term Memory (LSTM)–Gated Recurrent Unit (GRU) Method for Flood Prediction. Water, 14 (14), 2221 https://doi.org/10.3390/w14142221.

8. Yu Liu, Hao Wang, Wenwen Feng & Haocheng Huang. (2021). Short Term Real-Time Rolling Forecast of Urban River Water Levels Based on LSTM: A Case Study in Fuzhou City, China. Environmental Research and Public Health.

9. Lê Xuân Hiền và Hồ Việt Hùng (2018). Ứng dụng mạng LSTM để dự báo mực nước tại trạm Quang Phục và Cửa Cấm, Hải Phòng, Việt Nam. Tạp chí Khoa học kỹ thuật thủy lợi và môi trường, số 62 (9/2018).

10. Hồ Việt Hùng (2021). Dự báo mực nước sông Cấm, TP. Hải Phòng bằng mô hình Mạng nơ-ron LSTM. Tạp chí Khoa học và Công nghệ thủy lợi, số 64 – 2021.

11. Punyanuch Borwarnginna, Jason H. Hagab và Worapan Kusakunniran (2022). Predicting river water height using deep learning-based features. ICT Express, 8 (4). https://doi.org/10.1016/j.icte.2022.03.012.

12. Dongfeng Ren, Qian Hu và Tengda Zhang (2025). EKLT: Kolmogorov-Arnold attention-driven LSTM with Transformer model for river water level prediction. Journal of Hydrology, 649 (2025). https://doi.org/10.1016/j.jhydrol.2024.132430.