

Nghiên cứu ứng dụng mô hình LSTM (Long Short-Term Memory) trong dự báo độ mặn từ chuỗi dữ liệu đa biến tại trạm Đại Ngãi, Sóc Trăng

Bùi Duy Quỳnh^{1*}, Hà Thị Hằng¹, Lưu Thị Diệu Chinh², Trần Xuân Trường^{3,4}

¹Khoa Cầu đường - Trường Đại học Xây dựng Hà Nội

²Khoa Công trình thủy - Trường Đại học Xây dựng Hà Nội

³Khoa Trắc địa- Bản đồ và Quản lý đất đai - Trường Đại học Mỏ - Địa chất

⁴Nhóm nghiên cứu Công nghệ Địa tin học trong Khoa học Trái đất (GES) - Trường Đại học Mỏ - Địa chất

TÓM TẮT

Xâm nhập mặn đang đặt ra những thách thức đáng kể đối với các khu vực ven biển trên toàn thế giới. Các mô hình dự đoán xâm nhập mặn có thể cung cấp thông tin quý giá trong chủ động lên kế hoạch phòng chống, giảm thiểu tác động và ảnh hưởng của xâm nhập mặn. Nghiên cứu này ứng dụng mô hình học sâu mạng bộ nhớ dài-ngắn LSTM (Long Short-Term Memory) - một dạng đặc biệt của mạng nơron tái phát (RNN-Recurrent Neural Network) - để dự báo độ mặn dựa trên việc phân tích chuỗi dữ liệu đa biến, bao gồm: dữ liệu độ mặn, mực nước và lượng mưa tại trạm Đại Ngãi từ năm 2011 đến năm 2022. Độ tin cậy của mô hình dự báo độ mặn được đánh giá dựa trên các chỉ số R², MAE và RMSE cho các tập dữ liệu kiểm chứng và dữ liệu kiểm tra. Kết quả đánh giá cho thấy mô hình dự báo độ mặn có độ tin cậy cao, với giá trị R² > 95,8, giá trị MAE dao động từ 0,1 g/L đến 0,29 g/L và giá trị RMSE cũng rất nhỏ, thay đổi trong phạm vi từ 0,34 g/L đến 0,54 g/L. Kết quả thực nghiệm này cho thấy việc áp dụng mô hình học sâu trên cơ sở sử dụng chuỗi dữ liệu đa biến trong dự báo độ mặn có ý nghĩa rất lớn trong việc cảnh báo sớm cũng như lên kế hoạch phòng tránh, thực thi các giải pháp ngăn mặn lâu dài. Ngoài ra, kết quả thực nghiệm này cũng có thể áp dụng hiệu quả cho các trạm quan trắc độ mặn khác trong khu vực.

Từ khóa: Dự báo độ mặn, chuỗi dữ liệu đa biến, LSTM, trạm Đại Ngãi.

1. Đặt vấn đề

Xâm nhập mặn là hiện tượng xảy ra khi nước biển xâm nhập vào các khu vực nước ngọt thông qua hệ thống sông, suối (Ippen và cộng sự, 1961) và thường xảy ra ở các khu vực ven biển và đồng bằng nơi có sự giao thoa giữa nước biển và nước ngọt (Mahmuduzzaman và cộng sự, 2014). Đồng bằng sông Cửu Long (ĐBSCL) được coi là một trong những vùng bị ảnh hưởng nặng nề trên thế giới trước các thiên tai như hạn hán, xâm nhập mặn (Xiao và cộng sự, 2021). Hiện nay, với sự gia tăng của mực nước biển dâng do biến đổi khí hậu (BĐKH), việc trữ nước tại các đập thủy điện đầu nguồn, sự hạ thấp đáy sông do khai thác cát, sự gia tăng sử dụng nước nội vùng đã và đang làm cho tình trạng xâm nhập mặn ngày càng trở nên nghiêm trọng, nước biển lấn sâu vào nội đồng, gây ảnh hưởng lớn đến đời sống sinh hoạt và sản xuất tại khu vực ĐBSCL (Duong Tran Anh và cộng sự, 2018; Doan Van Binh và cộng sự, 2020; Nguyen Huu Thoi, 2001). Để dự báo xâm nhập mặn, một số nghiên cứu thường sử dụng bộ mô hình MIKE và thu được kết quả tương đối tốt (Đặng Hoàng Lam và các cộng sự, 2022; Đặng Văn Dũng và các cộng sự, 2018; Đoàn Quang Trí, 2016); tuy nhiên, các mô hình này thường yêu cầu số lượng dữ liệu đầu vào lớn và chi tiết, như dữ liệu địa hình, mặt cắt sông, các công trình thủy lợi, hệ số nhám,... để kiểm định mô hình và cần nhiều thời gian tính toán.

Ngày nay, các mô hình trí tuệ nhân tạo đã cung cấp thêm một cách tiếp cận mới trong việc xây dựng các mô hình dự báo độ mặn dựa trên chuỗi số liệu theo thời gian và đạt được độ chính xác cũng như độ tin cậy cao. Các mô hình trí tuệ đã từng được áp dụng nhiều trong dự báo độ mặn và cho kết quả tốt, như mô hình trung bình Bayes (BMA) (Kairong Lin và các cộng sự, 2019; Nguyen, Thu Thi Minh và các cộng sự, 2023), ARIMA (Tran Thanh Thai và các cộng sự, 2021), rừng ngẫu nhiên (RF), máy vector hỗ trợ (SVM), mạng nơron Elman (ENN) (Pengyu Lu và các cộng sự, 2021) hoặc các mô hình kết hợp Boruta - mạng nơron nhân tạo (B-ANN), Boruta - hồi quy vector hỗ trợ (B-SVR) (Abbasi, Mahdi và các cộng sự, 2022). Song song với các mô hình học máy truyền thống, mô hình học sâu (Deep Learning) mà điển hình là mạng nơron hồi quy (RNN- Recurrent Neural Network) với phiên bản mở rộng là mạng bộ nhớ dài ngắn LSTM (Long Short-Term Memory) đã được sử dụng trong các bài toán về dự báo chuỗi thời gian, như dự đoán lưu lượng xả nước hàng ngày trong lưu vực (Gauch, M. và các cộng sự, 2021), dự báo độ mặn (Nguyễn Công Thành và Nguyễn Tiên Giang, 2022), đánh giá sự dịch chuyển của cầu (Thai Ha Vu, Ngoc Quang Vu và Nguyen Van Thieu, 2024), hay dự báo độ cao sóng (Trần

* Tác giả liên hệ

Email: quynhbdd@huce.edu.vn

Hồng Thái và các cộng sự, 2022). Hầu hết các nghiên cứu này đều khẳng định tính hiệu quả của mô hình LSTM trong dự báo, cho kết quả khả quan nhờ khả năng ghi nhớ các bước và không bị ảnh hưởng nhiều khi số liệu đầu vào bị thiếu. Trong một nghiên cứu dự báo độ mặn mới đây nhất của Thành và cộng sự (Nguyễn Công Thành và Nguyễn Tiền Giang, 2022) vào năm 2023, nghiên cứu này chỉ sử dụng một biến dữ liệu đầu vào duy nhất, đó là dữ liệu độ mặn cao nhất trong ngày tại trạm Đại Ngãi trong 6 tháng mùa khô (từ tháng 1 đến tháng 6) năm 2002 đến năm 2021 để dự báo độ mặn. Kết quả của nghiên cứu này cho thấy mô hình LSTM mặc dù có thể xử lý và dự báo chuỗi dữ liệu theo chuỗi thời gian khá tốt, nhưng mô hình không dự đoán được các giá trị cực trị trong quá trình huấn luyện. Trong khi đó, ưu điểm nổi trội của mô hình LSTM là xử lý dữ liệu đa biến theo chuỗi thời gian (S. Alhirmizy và B. Qader, 2019). Đối với dự báo xâm nhập mặn, mô hình LSTM có thể xử lý tốt dữ liệu đa biến, nơi mà nhiều biến số có thể ảnh hưởng qua lại lẫn nhau theo thời gian. Bằng cách này, LSTM không chỉ phân tích từng biến số riêng lẻ mà còn học được các mối quan hệ phức tạp giữa chúng, qua đó, giúp cải thiện độ chính xác của mô hình (Zhang, Xiaoying và các cộng sự, 2023).

Chính vì vậy, nghiên cứu này đề xuất và trình bày kết quả ứng dụng mô hình bộ nhớ dài-ngắn LSTM trong dự báo độ mặn tại trạm đo mặn Đại Ngãi, tỉnh Sóc Trăng dựa trên chuỗi dữ liệu đa biến, đa thời gian trong quá khứ, bao gồm độ mặn, mực nước, lượng mưa. Trên cơ sở đó đánh giá hiệu quả, độ tin cậy và khả năng áp dụng rộng rãi mô hình này cho các trạm quan trắc độ mặn khác trong khu vực.

2. Cơ sở lý thuyết và phương pháp nghiên cứu

2.1. Khu vực nghiên cứu

Trạm thủy văn Đại Ngãi ($9^{\circ}44'4,60''N$, $106^{\circ}4'27,03''E$) là một trong các trạm quan trắc khí tượng của tỉnh Sóc Trăng, nằm ở khu vực cửa Trần Đề của nhánh Sông Hậu, thuộc thị trấn Đại Ngãi, huyện Long Phú, tỉnh Sóc Trăng (Hình 1). Hiện nay, tại trạm Đại Ngãi quan trắc nhiều thông tin thủy văn như mực nước, độ mặn, đo mưa tự động,... phục vụ cung cấp thông tin khí tượng thủy văn cho khu vực cũng như Ủy ban hội sông Mê Kông vì sự phát triển bền vững của khu vực này.



Hình 1. Vị trí trạm thủy văn Đại Ngãi, tỉnh Sóc Trăng

2.2. Số liệu sử dụng

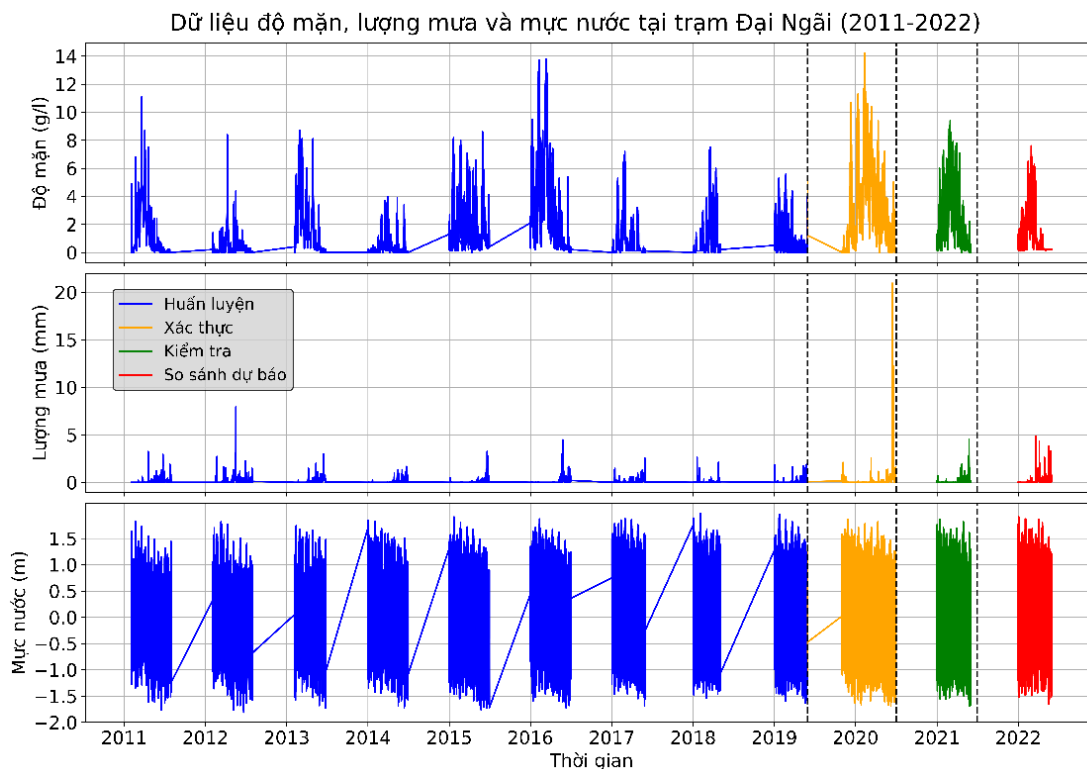
Trong nghiên cứu này, số liệu được sử dụng gồm: Số liệu quan trắc mặn (g/L), số liệu mực nước (cm) được đo liên tục (2 h/lần) trong ngày tại trạm Đại Ngãi trong 12 năm (2011-2022) ở các tháng mùa khô (từ tháng 01 đến tháng 6); số liệu lượng mưa thu thập từ dữ liệu vệ tinh và được lấy từ trang web của dự án POWER (Dự đoán tài nguyên năng lượng toàn cầu) của NASA, <https://power.larc.nasa.gov/data-access-viewer/>.

Toàn bộ dữ liệu được chia thành 4 phần: 70% cho tập huấn luyện (training), 10% cho tập xác thực (validation), 10% cho tập kiểm tra (testing) và 10% cho tập so sánh dự báo (prediction) được thể hiện trong Hình 2. Số liệu từ 01/01/2011 tới 31/05/2019 dùng để huấn luyện mô hình, số liệu từ 01/11/2019 đến 30/6/2020 dùng để kiểm chứng, số liệu từ 01/01/2021 đến 31/5/2021 dùng để kiểm tra và phần còn lại từ 01/01/2022 đến 31/5/2022 dùng để kiểm tra kết quả dự báo trong tương lai.

2.3. Phương pháp nghiên cứu

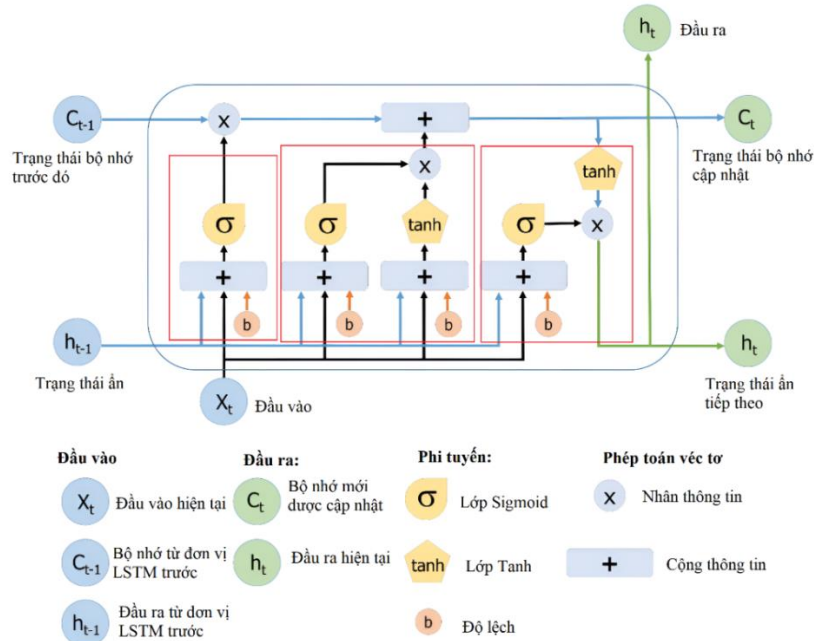
a. Mạng LSTM

Mô hình mạng LSTM được xây dựng dựa trên mạng nơron hồi quy RNN do Hochreiter và Schmidhuber đề xuất năm 1997 (Hochreiter, S. và Schmidhuber, J., 1997) và ngày càng được cải tiến (Yao, Kaisheng và các cộng sự, 2015) với mục đích khắc phục các thiếu sót của RNN bao gồm sự phụ thuộc lâu dài trong RNN và hiện tượng gradient biến mất và bùng nổ. Về lý thuyết, RNN có khả năng tự xử lý các phụ thuộc theo thời gian (temporal dependencies) bằng việc sử dụng bộ nhớ ngắn hạn và dựa trên việc xác định các tham số một cách hiệu quả. Tuy nhiên, trong thực tế RNN không thể tự giải quyết các phụ thuộc theo thời gian khi chuỗi số liệu có các phụ thuộc xa (long-term dependencies) (Bengio, Samy và Bengio, Yoshua, 2000).



Hình 2. Dữ liệu thủy văn tại trạm Đại Ngãi từ 2011-2022 (Nhóm tác giả thu thập)

LSTM có cấu trúc giống như RNN, nhưng cấu trúc bên trong thì lại phức tạp hơn nhằm khắc phục các vấn đề của RNN (Hình 3). Có ba cổng trong LSTM với các chức năng khác nhau: cổng đầu vào (kiểm soát mối quan hệ trạng thái và dữ liệu đầu vào), cổng đầu ra (kiểm soát quan hệ trạng thái và đầu ra) và cổng quên (kiểm soát thang thời gian bộ nhớ).



Hình 3. Cấu trúc của mạng LSTM

Công quen sử dụng hàm sigmoid áp dụng cho dữ liệu đầu vào hiện tại và trạng thái ẩn trước đó để quyết định sự đóng góp của chúng ở các trạng thái tiếp theo, được biểu diễn qua công thức (1). Đây chính là công có vai trò chia khóa để giải quyết các vấn đề tồn tại của mô hình RNN truyền thống. Công vào được trộn giữa hàm sigmoid (xác định các giá trị cần cập nhật) và hàm tanh (tạo ra các giá trị mới) để áp dụng cho trạng thái đầu vào và trạng thái ẩn để tính toán trạng thái hiện tại của của ô LSTM, biểu diễn qua công thức (2) và (3). Đầu ra của hai công cuối cùng được sử dụng cho tính trạng thái ẩn tiếp theo của ô, biểu diễn qua công thức (4) và (5).

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

$$i_t = s(W_i \cdot [h_{t-1}, X_t] + b_i) \quad (2)$$

$$C_t = f_t \times c_{t-1} + i_t \times \tanh(W_c \cdot [h_{t-1}, X_t] + b_c) \quad (3)$$

$$o_t = s(W_o \cdot [h_{t-1}, X_t] + b_o) \quad (4)$$

$$h_t = o_t \times \tanh(C_t) \quad (5)$$

b. Thiết lập mô hình

Dữ liệu gồm độ mặn, lượng mưa và mực nước đo 2 h/lần tại trạm Đại Ngãi, được sử dụng làm dữ liệu đa biến đầu vào cho mô hình LSTM. Trong nghiên cứu này, mô hình LSTM áp dụng độ dài chuỗi đầu vào trong quá khứ kí hiệu $w = 84$ (tương ứng với dữ liệu 7 ngày), $w = 168$ (14 ngày) và $w = 360$ (30 ngày) để đánh giá cho 2 h tiếp theo, được kí hiệu là $w84$, $w168$ và $w360$. Ngoài ra, mô hình LSTM còn bị ảnh hưởng của các tham số đầu vào, vì vậy, việc lựa chọn các tham số này rất quan trọng. Do đó, trong nghiên cứu này đã sử dụng phương pháp tìm kiếm ngẫu nhiên (Random Search) dựa trên tập dữ liệu huấn luyện và tập dữ liệu kiểm chứng để tối ưu hóa các tham số cho từng loại kích thước cửa sổ trượt. Random search sẽ áp dụng các bộ thông số khác nhau của các mô hình được thiết lập, qua đó tìm được bộ thông số tối ưu nhất thông qua bộ dữ liệu huấn luyện và kiểm chứng. Quá trình thực hiện tìm kiếm cũng như chạy mô hình được thực hiện trên nền tảng Google Colab bằng ngôn ngữ Python kết hợp các thư viện có sẵn scikit-learn. Các siêu tham số này được trình bày trong Bảng 1 dưới đây.

Bảng 1. Các siêu tham số của mô hình LSTM

Siêu tham số	Phạm vi	Mô hình		
		$w = 84$ (7 ngày)	$w = 168$ (14 ngày)	$w = 360$ (30 ngày)
Số đơn vị ẩn LSTM	[16,32,64,128,256]	16	32	128
Dropout	[0,1, 0,2, 0,3, 0,5]	0,5	0,3	0,5
Batch size	[16,32,64,128,256]	256	256	256
Epochs	[10, 50, 100, 500]	50	50	50

c. Các chỉ số đánh giá độ tin cậy của mô hình

Để đánh giá độ tin cậy của mô hình dự báo LSTM, nghiên cứu này sử dụng một số chỉ số sau, bao gồm: hệ số xác định R^2 (Di Bucchianico, Alessandro, 2008; Wright, Sewall, 1921), sai số trung phương (RMSE) (Nevitt, Jonathan và Hancock, Gregory R., 2000), độ lệch tuyệt đối (MAE) (Sammur, Claude và Webb, Geoffrey I., 2010), giá trị phương sai được giải thích Ex_VAR (explained variance) (Nevitt, Jonathan và Hancock, Gregory R., 2000). Các giá trị RMSE và MAE đạt giá trị tối ưu khi bằng 0; giá trị R^2 và giá trị Ex_VAR đạt tối ưu khi bằng 100%. Các chỉ số này được tính theo công thức sau:

$$R^2 = \left(1 - \frac{\sum_{i=1}^n (y_{tt,i} - y_{db,i})^2}{\sum_{i=1}^n (y_{tt,i} - \bar{y}_{tt})^2}\right) \times 100 \quad (6)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_{tt,i} - y_{db,i})^2} \quad (7)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_{db,i} - y_{tt,i}| \quad (8)$$

$$Ex_VAR = 100 \times \left(1 - \frac{var(y_{tt,i} - y_{db,i})}{var(y_{tt,i})}\right) \quad (9)$$

3. Kết quả và thảo luận

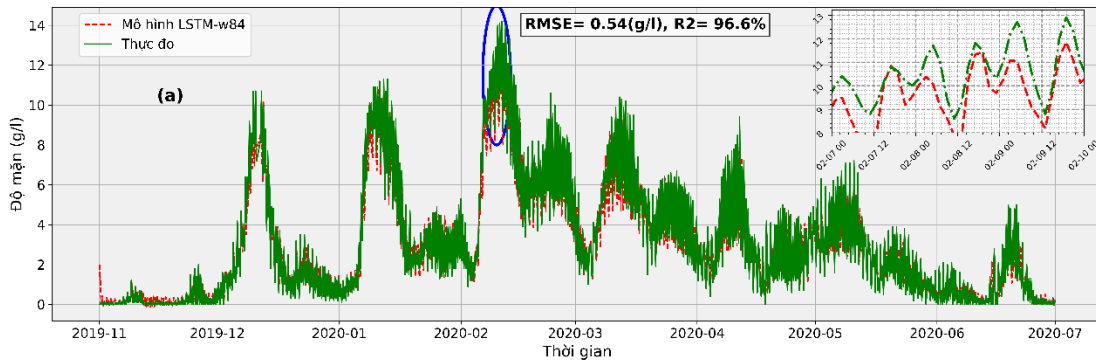
Với 3 mô hình LSTM có các khoảng dữ liệu đầu vào khác nhau (w84, w168 và w360) được huấn luyện với bộ tham số trong Bảng 1, sẽ tiến hành đánh giá các mô hình này dựa trên bộ dữ liệu trong tệp kiểm chứng và kiểm tra. Các chỉ số đánh giá mô hình được thống kê trong Bảng 2.

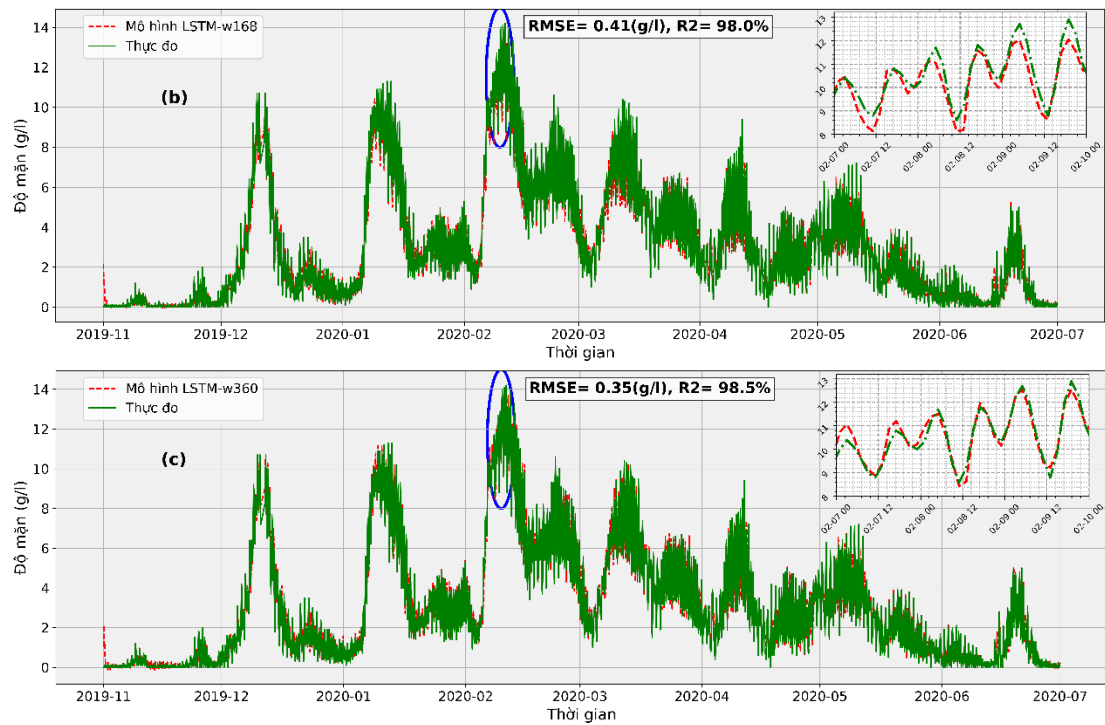
Dựa trên các chỉ số trong Bảng 2 cho thấy, các mô hình có độ chính xác rất cao với hệ số xác định R^2 đều lớn hơn 95% ở cả tệp kiểm chứng và tệp kiểm tra. Các chỉ số như RMSE và MAE đều nhỏ hơn 1 g/L.

Bảng 2. Các chỉ số đánh giá mô hình LSTM

Chỉ số	Mô hình w84		Mô hình w168		Mô hình w360	
	Kiểm chứng	Kiểm tra	Kiểm chứng	Kiểm tra	Kiểm chứng	Kiểm tra
RMSE (g/l)	0,54	0,43	0,41	0,34	0,35	0,39
MAE (g/L)	0,29	0,19	0,17	0,12	0,12	0,10
R^2 (%)	96,58	95,82	98,02	97,29	98,48	97,77
Var_ex (%)	96,77	95,96	98,05	97,35	98,48	97,72

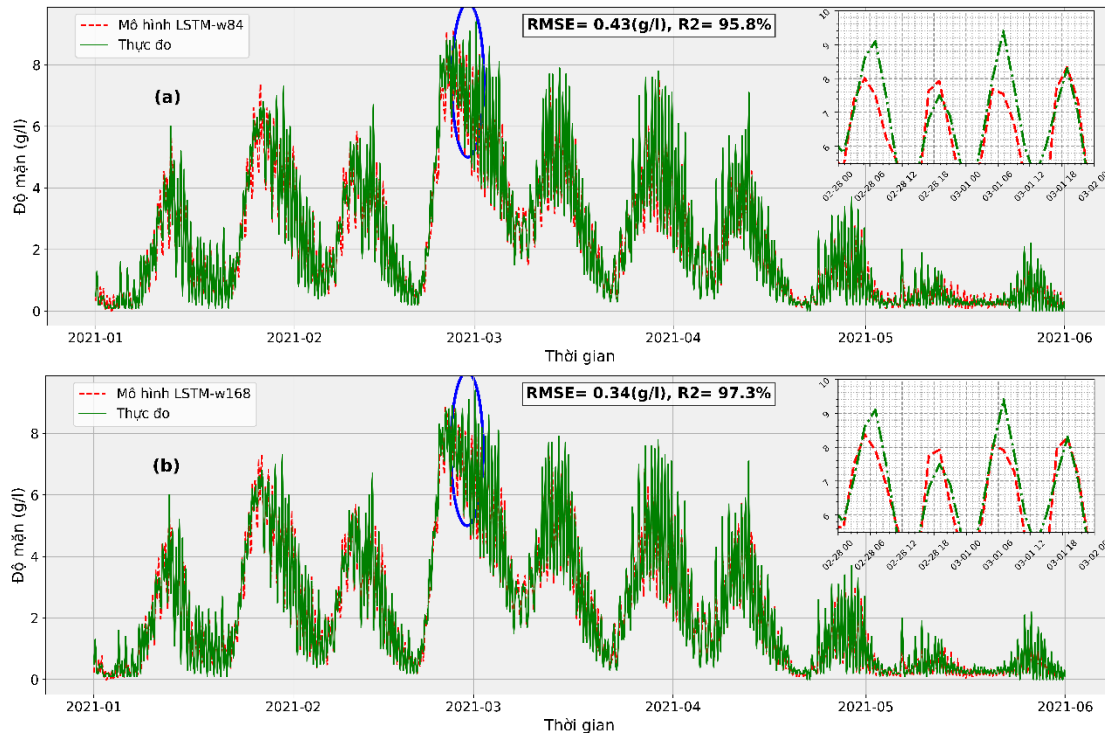
Hình 4 trình bày kết quả của các tệp kiểm chứng với các khoảng dữ liệu đầu vào khác nhau được dự báo bằng mô hình LSTM. Mặc dù, chuỗi dữ liệu độ mặn có sự biến đổi phức tạp, tuy nhiên, các kết quả cho thấy có sự tương đồng rất cao giữa giá trị thực đo và giá trị dự báo từ mô hình, kể cả tại các vị trí đỉnh mặn của chuỗi dữ liệu. Có thể thấy, mô hình LSTM-w360 có giá trị dự báo tại các vị trí đỉnh mặn sát với giá trị thực đo nhất trong 3 mô hình (khu vực phóng to Hình 4).

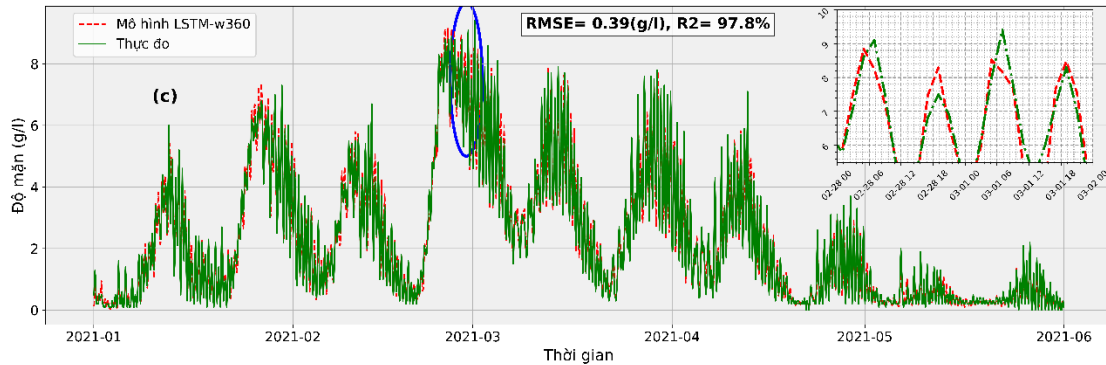




Hình 4. Kết quả các mô hình với các khoảng dữ liệu w84, w168 và w360 của tập kiểm chứng

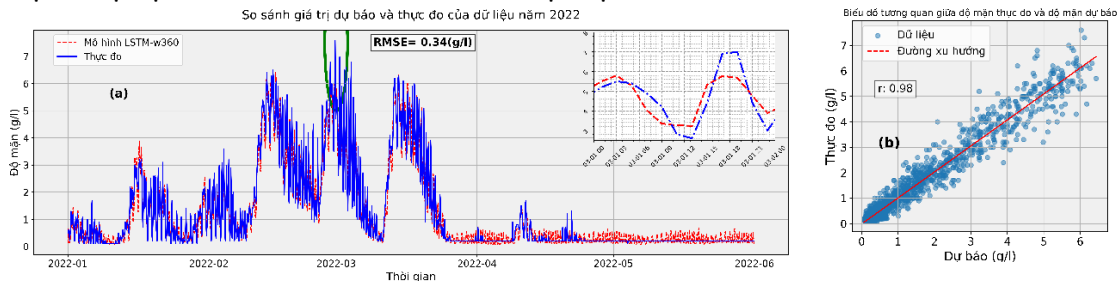
Tương tự, việc so sánh giữa kết quả của mô hình và dữ liệu thực tế đối với tập dữ liệu kiểm tra thể hiện trong Hình 5. Các kết quả cho thấy có sự tương đồng rất cao giữa giá trị thực đo và giá trị dự báo từ mô hình, kể cả tại các vị trí đỉnh mặn của chuỗi dữ liệu. Có thể thấy, mô hình LSTM-w360 có giá trị dự báo tại các vị trí đỉnh mặn sát với giá trị thực đo nhất trong các mô hình LSTM-w84, LSTM-w168 và LSTM-w360 (khu vực phóng to trong Hình 5).





Hình 5. Kết quả các mô hình với các khoảng dữ liệu w84, w168 và w360 của tập kiểm tra

Việc các mô hình LSTM có thể dự báo được trong tương lai dựa trên dữ liệu lịch sử, trong nghiên cứu đã tiến hành dự báo độ mặn cho năm 2022 và so sánh với số liệu thực đo của năm 2022.



Hình 6. Kết quả giữa mô hình dự báo và dữ liệu thực đo năm 2022

Với kết quả mô hình LSTM-w360 cho độ chính xác cao nhất nên được sử dụng làm mô hình dự báo độ mặn cho năm 2022. Kết quả dự báo được thể hiện trong Hình 6, cho thấy độ chính xác dự báo rất cao với sai số RMSE = 0,34 g/L (Hình 6a) và hệ số tương quan $r = 0,977$ (Hình 6b).

4. Kết luận

Các kết quả đạt được trong nghiên cứu này cho thấy mô hình LSTM có thể xử lý hiệu quả chuỗi số liệu đa biến, đa thời gian (độ mặn, mực nước và lượng mưa) trong dự báo độ mặn. Nghiên cứu này đã sử dụng mô hình LSTM với kích thước các chuỗi dữ liệu đầu vào khác nhau, bao gồm LSTM-w84, LSTM-w168, LSTM-w360 để dự báo độ mặn tại trạm Đại Ngãi, Sóc Trăng. Trong đó, mô hình LSTM-w360 có hiệu suất dự báo tốt nhất so với các mô hình còn lại (LSTM-w84 và LSTM-w168). Tuy nhiên, các mô hình này cần được thử nghiệm mở rộng hơn trong thời gian tới với sự tham gia của các yếu tố khác như gió, nhiệt độ, lưu lượng dòng chảy,... cũng như áp dụng dự báo độ mặn tại các trạm quan trắc khác trong khu vực để kiểm chứng độ chính xác và tính hiệu quả của mô hình LSTM này.

Lời cảm ơn

Nghiên cứu này được tài trợ bởi đề tài khoa học và công nghệ cấp bộ “Nghiên cứu phương pháp đánh giá rủi ro xâm nhập mặn có xét tới ảnh hưởng của biến đổi khí hậu” mã số CT.2022.01.XDA.04.

Tài liệu tham khảo

- Di Bucchianico, Alessandro (2008). "Coefficient of Determination (R^2)", *Encyclopedia of Statistics in Quality and Reliability*, <https://doi.org/10.1002/9780470061572.eqr173>.
- Duong Tran Anh và cộng sự (2018). "Simulating future flows and salinity intrusion using combined one- and two-dimensional hydrodynamic modelling-the case of Hau River, Vietnamese Mekong Delta", *Water (Switzerland)*, 10(7).
- Nevitt, Jonathan & Hancock, Gregory R. (2000). "Improving the root mean square error of approximation for nonnormal conditions in structural equation modeling", *The Journal of Experimental Educational*, tr. 251-268.
- Sammut, Claude & Webb, Geoffrey I. (2010). "Mean Absolute Error", *Encyclopedia of Machine Learning*, Springer Reference, tr. 652.
- Wright, Sewall (1921). "Correlation and causation", *Journal of Agricultural Research*, XX(7), tr. 557-585.
- Abbasi, Mahdi et al. (2022). "Long-term streamflow prediction using hybrid svr-ann based on bayesian model averaging", *Journal of Hydrologic Engineering*. 27(11).

- Bengio, Samy & Bengio, Yoshua, IEEE Transactions on Neural Networks (2000). "Taking on the curse of dimensionality in joint distributions using neural networks", 11(3), tr. 550-557.
- Đặng Hoàng Lam và cộng sự (2022). "Xây dựng mô hình MIKE 11 phục vụ công tác dự báo thủy văn và xâm nhập mặn tỉnh Bến Tre", *Tạp chí Khí tượng thủy văn*, 740(1), tr. 38-49.
- Đặng Văn Dũng và cộng sự (2018). "Khai thác mô hình MIKE 11 trong dự báo, cảnh báo xâm nhập mặn vùng Đồng bằng sông Cửu Long", *Tạp chí Khí tượng thủy văn*, 693, tr. 48-58.
- Đoàn Quang Trí (2016). "Ứng dụng mô hình Mike 11 mô phỏng và tính toán xâm nhập mặn cho khu vực Nam bộ", *Tạp chí Khí tượng Thủy văn*, 11, tr. 39-46.
- Doan Van Binh et al. (2020). "Long-term alterations of flow regimes of the Mekong River and adaptation strategies for the Vietnamese Mekong Delta", *Journal of Hydrology: Regional Studies*, 32, tr. 100742.
- Gauch, M. et al. (2021). "Rainfall-runoff prediction at multiple timescales with a single Long Short-Term Memory network", *Hydrol. Earth Syst. Sci.*, 25(4), tr. 2045-2062.
- Hochreiter, S. và Schmidhuber, J. (1997). "Long Short-Term Memory", *Neural Computation*, 9(8), tr. 1735-1780.
- Ippen, Arthur T & Harleman, Donald RF (1961). *One-dimensional analysis of salinity intrusion in estuaries*, Waterways Experiment Station.
- Kairong Lin et al. (2019). "Modeling saltwater intrusion using an integrated Bayesian model averaging method in the Pearl River Delta", *Journal of Hydroinformatics*, 21(6), tr. 1147-1162.
- Mahmuduzzaman, Mohammed et al. (2014). "Causes of salinity intrusion in coastal belt of Bangladesh", 4(4A), tr. 8-13.
- Nguyễn Công Thành và Nguyễn Tiền Giang (2022). "Xây dựng mô hình máy học LSTM (Long Short Term Memory) phục vụ công tác dự báo mặn tại trạm đo mặn Đại Ngãi", *Tạp chí Khí tượng Thủy văn*, 740(1).
- Nguyen Huu-Thoi và Ashim Das Gupta (2001). "Assessment of Water Resources and Salinity Intrusion in the Mekong Delta", *Water International*, 26(1), tr. 86-95.
- Nguyen, Thu Thi Minh và cộng sự. (2023), "Mapping of Top-Soil Salinity Zoning in the Coastal Area of Ben Tre province, Vietnam", *International Journal of Intelligent Systems Applications in Engineering*, 11(4s).
- Pengyu Lu et al. (2021). "An integrated framework of input determination for ensemble forecasts of monthly estuarine saltwater intrusion", *Journal of Hydrology*, 598.
- S. Alhirmizy & B. Qader (2019). Multivariate Time Series Forecasting with LSTM for Madrid, Spain pollution, *2019 International Conference on Computing and Information Science and Technology and Their Applications (ICCISTA)*, tr. 1-5.
- Thai Ha Vu, Ngoc Quang Vu & Nguyen Van Thieu (2024). "Chapter 24 - Spatial prediction of bridge displacement using deep learning models: A case study at Co Luy bridge", *Applications of Artificial Intelligence in Mining and Geotechnical Engineering*, Elsevier, tr. 437-461.
- Trần Hồng Thái và cộng sự (2022). "Xây dựng mô hình mạng nơron hồi quy dự báo độ cao sóng có nghĩa tại trạm Cồn Cỏ, Quảng Trị, Việt Nam", *Tạp chí Khí tượng thủy văn*.
- Tran Thanh Thai et al. (2021). "Performance evaluation of Auto-Regressive Integrated Moving Average models for forecasting saltwater intrusion into Mekong river estuaries of Vietnam", *Vietnam Journal of Earth Sciences*, 43(4), tr. 428-443.
- Xiao, Han et al. (2021). "Saltwater intrusion into groundwater systems in the Mekong Delta and links to global change", 12(3), tr. 342-352.
- Yao, Kaisheng et al. (2015). "Depth-gated recurrent neural networks", *arXiv preprint arXiv:1503.03790*, 9, tr. 98.
- Zhang, Xiaoying et al. (2023). "Advance prediction of coastal groundwater levels with temporal convolutional and long short-term memory networks", *Hydrology and Earth System Sciences*, 27(1), tr. 83-96.

ABSTRACT

Research on the application of the LSTM (Long Short-Term Memory) model in salinity forecasting from multivariate data series at Dai Ngai station, Soc Trang province

Bui Duy Quynh^{1*}, Ha Thi Hang¹, Luu Thi Dieu Chinh², Tran Xuan Truong^{3,4}

¹Faculty of Bridges and Roads, Hanoi University of Civil Engineering

²Faculty of Hydraulic Engineering, Hanoi University of Civil Engineering

³Faculty of Geodesy and Land Management, Hanoi University of Mining and Geology

⁴Geomatics in Earth Sciences Research Group, Hanoi University of Mining and Geology

Saltwater intrusion is posing significant challenges for coastal areas worldwide. Predictive models of saltwater intrusion can provide valuable information for proactive planning, mitigating impacts, and reducing the effects of saltwater intrusion. This study applies the Long Short-Term Memory (LSTM) deep learning model, a special type of Recurrent Neural Network (RNN), to forecast salinity based on multivariate time series analysis, including salinity data, water level, and rainfall at the Dai Ngai station from 2011 to 2022. The reliability of the salinity prediction model is assessed based on the R^2 , MAE, and RMSE indicators for validation and test datasets. The evaluation results show that the salinity prediction model is highly reliable, with an R^2 value > 95.8 , MAE values ranging from 0.1 g/L to 0.29 g/L, and RMSE values also being very small, varying from 0.34 g/L to 0.54 g/L. These experimental results indicate that applying deep learning models using time series data in salinity prediction is highly significant for early warning, planning, and implementing long-term saltwater intrusion prevention measures. These experimental results can also be effectively utilized at other regional salinity monitoring stations.

Keywords: Salinity forecasting, multivariate data series, LSTM, Dai Ngai station.