TẠP CHÍ
KHÍ TƯỢNG THỦY VĂN

Bài báo khoa học

Xây dựng mô hình máy học LSTM (Long Short-Term Memory) phục vụ công tác dự báo mặn tại trạm đo mặn Đại Ngãi

Nguyễn Công Thành[1], Nguyễn Tiền Giang[2*]

[1] Đài Khí tượng Thủy văn khu vực Nam bộ, Tổng cục Khí tượng Thủy văn, Bộ Tài nguyên và Môi trường; jackynguyen.kttv@gmail.com
[2] Khoa Khí tượng, Thủy văn và Hải dương học, Trường Đại học Khoa học Tự Nhiên, ĐHQGHN; giangnt@vnu.edu.vn.

***Tác giả liên hệ: giangnt@vnu.edu.vn; Tel.: +84–912800896

Ban Biên tập nhận bài: 8/8/2022; Ngày phản biện xong: 23/8/2022; Ngày đăng bài: 25/8/2022

Tóm tắt: Hiện nay, máy học hay học máy ML (*Machine Learning*) có lẽ đã không còn quá xa lạ và đã được ứng dụng vào rất nhiều lĩnh vực trong đời sống. Dự báo khí tượng thủy văn cũng không nằm ngoài sự đổi mới với việc xây dựng và ứng dụng các mô hình máy học này. Bài báo trình bày kết quả của nghiên cứu xây dựng một mô hình mạng bộ nhớ dài–ngắn LSTM (*Long Short–Term Memory*), là một dạng đặc biệt của mạng nơ–ron hồi quy (*RNN–Recurrent Neural Network*) để dự báo độ mặn tại trạm đo mặn Đại Ngãi, tỉnh Sóc Trăng. Số liệu sử dụng cho mô hình là số liệu quan trắc độ mặn cao nhất trong ngày tại trạm từ năm 2002–2021. Kết quả thiết lập mô hình cho các chỉ số đánh giá RMSE và NSE tốt (NSE > 0,9 với hầu hết các trường hợp), làm tiền đề cho việc ứng dụng mô hình máy học vào công tác dự báo xâm nhập mặn tại các trạm trên khu vực đồng bằng Sông Cửu Long.

Từ khóa: Dự báo xâm nhập mặn; Mô hình LSTM; Đại Ngãi; Sóc Trăng; Machine Learning.

1. Mở đầu

Trong những năm gần đây, dưới tác động của việc xây dựng các đập thủy điện ở thượng nguồn sông Mekong, chế độ dòng chảy trong hệ thống sông suối, kênh rạch tại Đồng bằng sông Cửu Long (ĐBSCL) đã có những thay đổi. Đồng thời, nước biển dâng do biến đổi khí hậu (BĐKH) tại các cửa sông Cửu Long, sự hạ thấp đáy sông do khai thác cát, sụt giảm bùn cát đến do hồ chứa thượng nguồn trữ lại, gia tăng sử dụng nước nội vùng đã và đang làm xâm nhập mặn ngày càng lấn sâu vào trong nội đồng ảnh hưởng lớn đến đời sống sinh hoạt và sản xuất của người dân [1–3]. Các nghiên cứu dự báo xâm nhập mặn gần đây thường sử dụng bộ mô hình Mike [4–6] và đã thu được kết quả tương đối tốt. Tuy nhiên cần yêu cầu dữ liệu đầu vào nhiều (đặc biệt là việc cập nhật dữ liệu địa hình, mặt cắt, công trình thủy lợi), cần kiểm định hiệu chỉnh và năng lực tính toán lớn.

Với sự phát triển của các thuật toán máy học trong thời gian gần đây đã cung cấp thêm hướng tiếp cận mới với việc xử lý và dự báo chuỗi thời gian đạt được độ chính xác cao. Có thể kể đến mô hình máy học truyền thống như ARIMA cho kết quả tương đối tốt với việc dự báo độ mặn [7]. Một phương pháp tiếp cận mới hơn nhằm khắc phục những nhược điểm của các mô hình máy học truyền thống là các mạng học sâu (*Deep Learning*). Điển hình là mạng nơ–ron hồi quy (*RNN–Recurrent Neural Network*) và phiên bản mở rộng của nó là

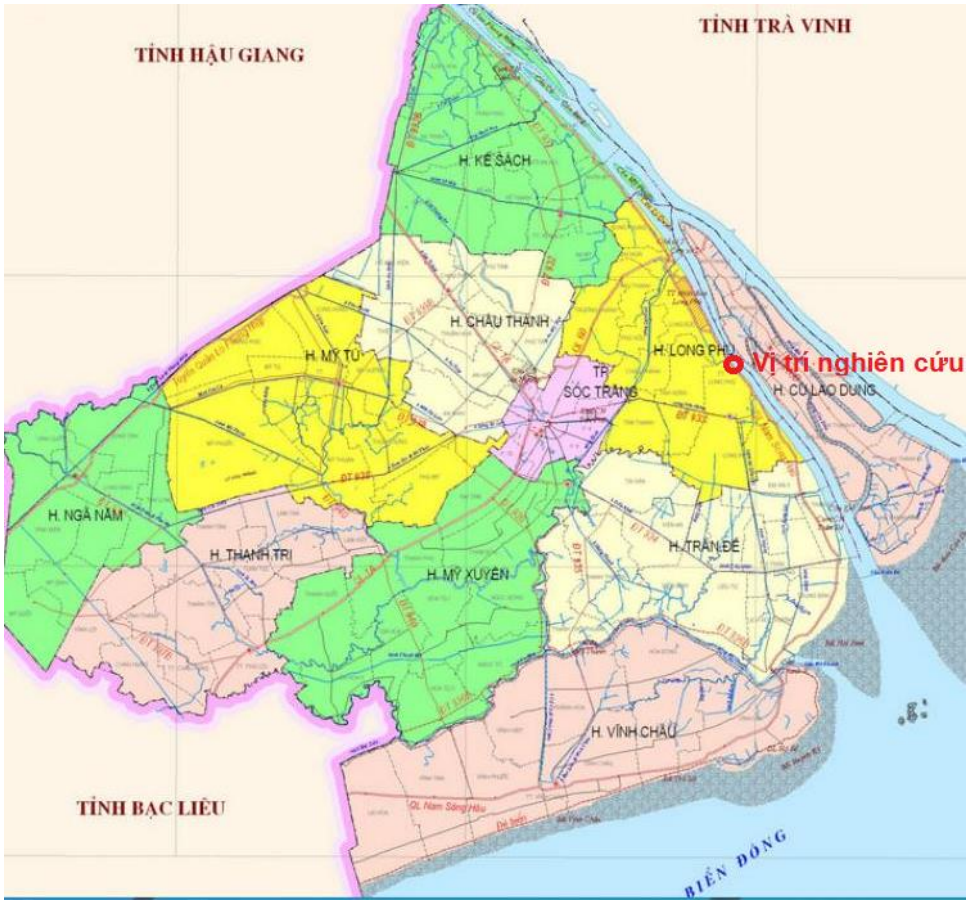
mạng bộ nhớ dài–ngắn LSTM (*Long Short–Term Memory*) được sử dụng nhiều trong các bài toán dự báo chuỗi thời gian [8] với kết quả khả quan nhờ có khả năng ghi nhớ các bước và không bị ảnh hưởng nhiều khi số liệu đầu vào bị thiếu.

Bài báo này trình bày kết quả nghiên cứu xây dựng mạng bộ nhớ dài–ngắn LSTM để dự báo độ mặn tại trạm đo mặn Đại Ngãi, tỉnh Sóc Trăng dựa trên chuỗi số liệu quan trắc quá khứ tại trạm, từ đó đánh giá khả năng ứng dụng mô hình vào trong thực tế.

2. Phương pháp nghiên cứu và số liệu sử dụng

2.1. Khu vực nghiên cứu

Tỉnh Sóc Trăng nằm ở cửa Nam sông Hậu, chịu ảnh hưởng của khí hậu nhiệt đới gió mùa, có mùa khô và mùa mưa rõ rệt hằng năm. Địa hình tỉnh Sóc trăng thấp trũng với hệ thống kênh rạch chằng chịt, nhiều vùng đất nhiễm mặn, phèn. Đây là địa phương cuối nguồn sông Hậu cũng là vùng cửa sông Mekong, do đó tác động của BĐKH và nước biển dâng có nguy cơ cao hơn so với các tỉnh bên trong nội đồng. Nếu mực nước biển dâng cao thêm 1 m thì có khoảng 43,7% diện tích tỉnh Sóc Trăng sẽ bị ngập mặn và tác động đến hơn 450.000 người, tương đương 35% tổng dân số của tỉnh Sóc Trăng. Trong các ngành kinh tế, nông nghiệp sẽ là đối tượng bị ảnh hưởng nhiều nhất, trong đó dịch bệnh trên cây trồng do tác động của quá trình xâm nhập mặn thời gian qua là biểu hiệu rõ nhất và nghiêm trọng đến ngành sản xuất nông nghiệp tỉnh Sóc Trăng. Ngành sản xuất nông nghiệp chiếm vị trí quan trọng hàng đầu trong nền kinh tế tỉnh Sóc Trăng. Hiện nay tỷ lệ dân số nông nghiệp và lao động nông nghiệp của tỉnh khá lớn (chiếm khoảng 72% dân số và 63% lao động) là nguồn thu nhập chính của trên 70% dân số của tỉnh [9]. Trạm đo mặn Đại Ngãi nằm ở cửa Nam sông Hậu thuộc thị trấn Đại Ngãi, huyện Long Phú, tỉnh Sóc Trăng (Hình 1).



Hình 1. Bản đồ hành chính tỉnh Sóc Trăng và vị trí nghiên cứu.

2.2. Số liệu sử dụng

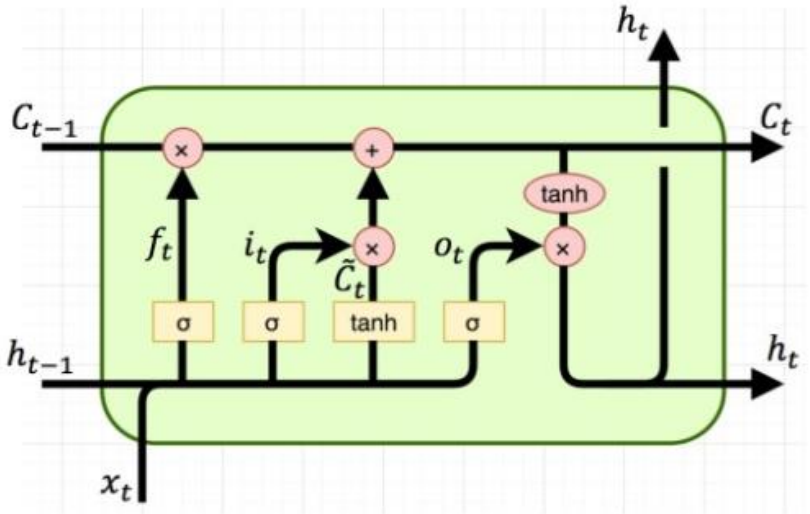
Trong bài báo này, số liệu được sử dụng là số liệu quan trắc độ mặn cao nhất (g/l) trong ngày tại trạm đo Đại Ngãi qua 20 năm (2002–2021) trong các tháng mùa kiệt. Các giá trị quan trắc không liên tục nên đã được tách mẫu theo mùa (tháng 1 – tháng 6) hàng năm. Số liệu được loại bỏ các giá trị NaN, sau đó được chuẩn hóa dạng Logarit. Toàn bộ dữ liệu được chia làm 3 phần: 70% cho tập huấn luyện (*training*), 15% cho tập kiểm chứng (*validation*) và 15% cho tập kiểm tra (*testing*).

2.3. Mạng LSTM

Mạng LSTM được cải tiến từ mạng thần kinh hồi quy (*RNN–Recurrent Neural Network*) nhằm khắc phục những nhược điểm về phụ thuộc xa (*Long–term Dependency*) của mạng RNN truyền thống. LSTM được giới thiệu bởi [10] và càng ngày càng được cải tiến [11].

Về mặt lý thuyết, RNN có khả năng xử lý các phụ thuộc theo thời gian (*temporal dependencies*) bằng việc sử dụng bộ nhớ ngắn hạn và dựa trên việc xác định (luyện) các tham số một cách hiệu quả [12]. Tuy nhiên, đáng tiếc trong thực tế RNN không thể giải quyết các phụ thuộc theo thời gian khi chuỗi số liệu có các phụ thuộc xa (*long–term dependencies*). Vấn đề này đã được nghiên cứu khá sâu bởi [13–14]. Trong các công bố của mình, họ đã tìm được những lý do để giải thích tại sao RNN không thể học được một các hiệu quả.

LSTM có cấu trúc dạng chuỗi các nút mạng như RNN, nhưng cấu trúc bên trong thì lại phức tạp hơn, bao gồm 4 tầng tương tác với nhau (Hình 2). Điểm đặc biệt của mạng LSTM nằm ở trạng thái ô C (*cell state*), nơi lưu trữ các trọng số dài hạn của mô hình. Các thông số trạng thái ô C, trạng thái ẩn h (*hidden state*), đầu vào tại thời điểm t x_t được đưa vào nút mạng. Sau khi được xử lý qua các hàm kích hoạt sigmoid σ, tanh và các phép toán véc–tơ, kết quả đầu ra là trạng thái ô C và trạng thái ẩn h tại thời điểm t sẽ được sử dụng cho nút mạng t+1 tiếp theo [15].



Hình 2. Cấu trúc một nút mạng trong mạng LSTM.

2.4. Các chỉ số đánh giá chất lượng mô hình

Để đánh giá hiệu quả dự báo độ mặn của mô hình tại trạm Đại Ngãi, nghiên cứu này sử dụng các chỉ số đánh giá: NSE (*Nash–Sutcliffe efficiency coefficient*) [16] và lỗi trung bình bình phương gốc (*RMSE–Root Mean Squared Error*) [17]. Chỉ số NSE biểu thị mức độ liên kết giữa các giá trị thực đo và mô phỏng, dao động từ $-\infty$ đến 1, giá trị càng gần 1 thì độ chính xác của mô hình càng cao [18]. Chỉ số RMSE thể hiện sự chênh lệch giữa các giá trị

dự đoán và giá trị quan trắc được, giá trị này càng thấp thì mô hình càng tốt (dao động từ 0 đến ∞).

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{n}(F_i - O_i)^2}{\sum_{i=1}^{n}(\bar{O} - O_i)^2} \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(F_i - O_i)^2} \quad (2)$$

Trong đó F_i là giá trị dự báo; O_i là giá trị quan trắc; $\bar{O}$ là trung bình giá trị quan trắc; n là số mẫu.

2.5. Thiết lập mô hình

Trong nghiên cứu này, mô hình LSTM sử dụng tập dữ liệu đơn biến là độ mặn cao nhất quan trắc được trong ngày tại trạm Đại Ngãi được chia thành các tập đầu vào 4 ngày, 8 ngày và 15 ngày và đánh giá cho 1 ngày tiếp theo. Sau đây sẽ gọi là W4, W8 và W15.

Các siêu tham số đóng một vai trò quan trọng ảnh hưởng trực tiếp đến hiệu quả của mô hình. Vì vậy trong nghiên cứu này, các siêu tham số được tối ưu hóa bằng phương pháp tìm kiếm ngẫu nhiên (*Random Search*). Vì các mẫu dữ liệu quan trắc còn hạn chế nên để tránh tình trạng overfitting (dữ liệu dự báo quá khớp với dữ liệu quan trắc), trong nghiên cứu này đã đưa thêm bước kiểm định chéo k–fold (*k–fold cross validation*) [19] nhằm chia tập training thành k phần, ở mỗi lần train, mô hình sẽ chọn 1 phần làm dữ liệu đánh giá (*validation*) và k–1 phần còn lại làm dữ liệu huấn luyện (*training*). Kết quả cuối cùng sẽ là trung bình cộng kết quả đánh giá của k lần train, giúp cho việc đánh giá mô hình khách quan hơn. Phương pháp này có sẵn trong thư viện scikit–learn là RandomSearchCV. Các siêu tham số sau khi được tối ưu được trình bày như Bảng 1.

Bảng 1. Các siêu tham số tối ưu cho các mô hình.

Siêu tham số	Phạm vi	Mô hình W4	Mô hình W8	Mô hình W15
Số đơn vị ẩn LSTM	[16, 32, 64, 128, 256]	32	256	64
Dropout	[0.1, 0.2, 0.25, 0.5]	0,2	0,5	0,2
Learning_rate	[0.01, 0.005, 0.001]	0,005	0,005	0,005
Batch_size	[16, 32, 64, 128, 256]	16	128	16
Epochs	[100, 200]	200	200	200

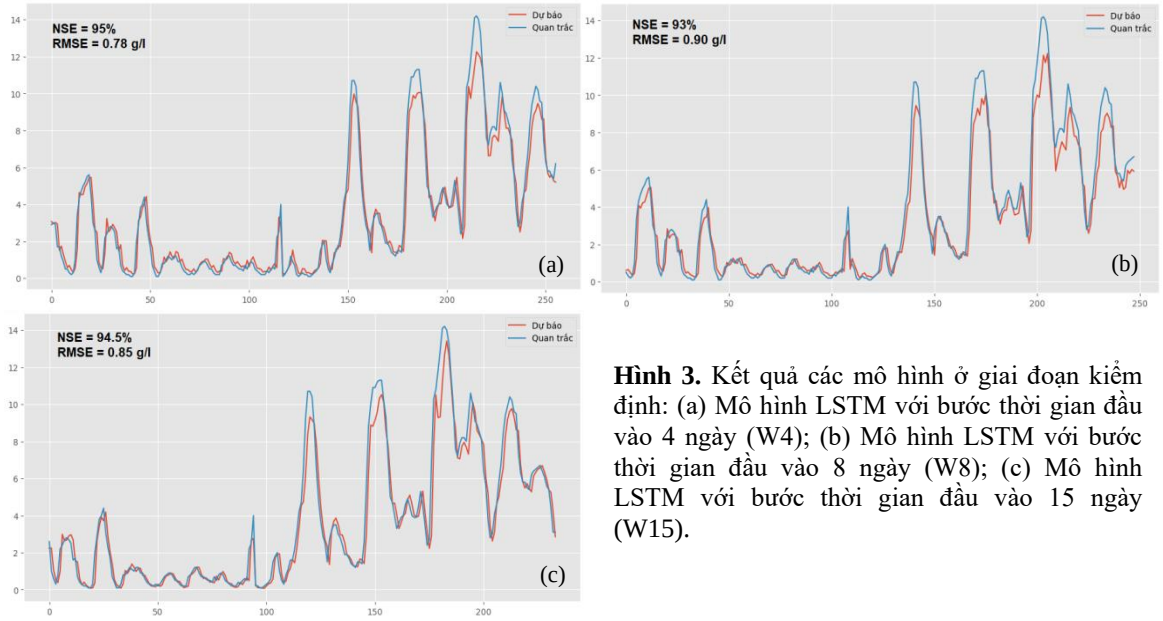
3. Kết quả và thảo luận

Các mô hình LSTM với các bước thời gian đầu vào khác nhau được huấn luyện với bộ siêu tham số đã được lựa chọn, ta sẽ tiến hành kiểm định các mô hình với chuỗi số liệu trong tập kiểm định và kiểm tra. Các chỉ số đánh giá mô hình được liệt kê như Bảng 2.

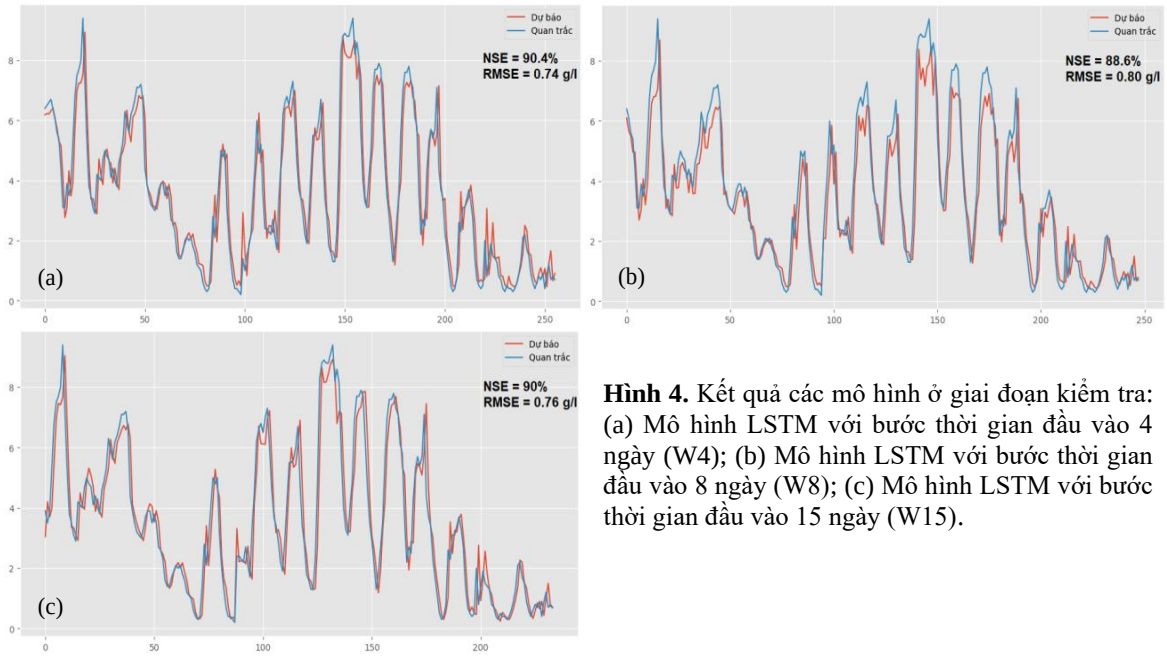
Bảng 2. Các chỉ số đánh giá các mô hình LSTM.

	Mô hình W4		Mô hình W8		Mô hình W15	
	Kiểm định	Kiểm tra	Kiểm định	Kiểm tra	Kiểm định	Kiểm tra
NSE	95%	90,4%	93%	88,6%	94,5%	90%
RMSE	0,78g/l	0,74g/l	0,90g/l	0,80g/l	0,85g/l	0,76g/l

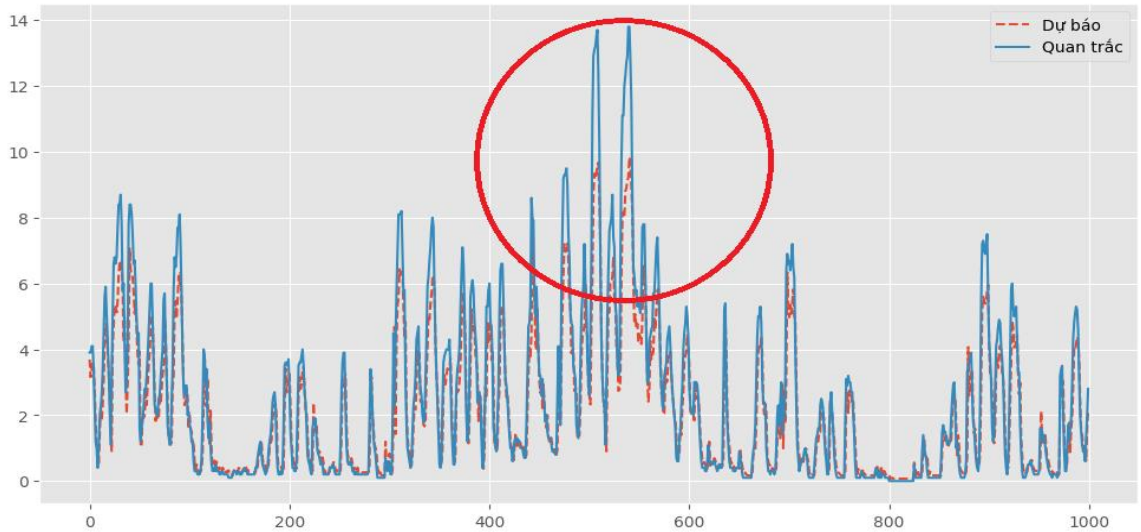
Từ các chỉ số đánh giá NSE và RMSE cho thấy các mô hình cho kết quả tương đồng rất cao giữa giá trị quan trắc và giá trị dự báo. Trong đó với bước thời gian đầu vào là 4 ngày cho kết quả tốt nhất (NSE lần lượt là 0,95 đối với tập số liệu kiểm định và 0,90 đối với tập kiểm tra; sai số bình phương RMSE cũng nhỏ nhất). Tuy nhiên mô hình đã không dự đoán đúng các giá trị cực trị (Hình 5). Với chuỗi số liệu đầu vào có sự biến đổi phức tạp như độ mặn, việc chuẩn hóa bằng phương pháp logarit hóa có lẽ là không đủ và cần thêm bước xử lý hoặc các phương pháp chuẩn hóa khác.



Hình 3. Kết quả các mô hình ở giai đoạn kiểm định: (a) Mô hình LSTM với bước thời gian đầu vào 4 ngày (W4); (b) Mô hình LSTM với bước thời gian đầu vào 8 ngày (W8); (c) Mô hình LSTM với bước thời gian đầu vào 15 ngày (W15).



Hình 4. Kết quả các mô hình ở giai đoạn kiểm tra: (a) Mô hình LSTM với bước thời gian đầu vào 4 ngày (W4); (b) Mô hình LSTM với bước thời gian đầu vào 8 ngày (W8); (c) Mô hình LSTM với bước thời gian đầu vào 15 ngày (W15).



Hình 5. Mô hình không bắt được các giá trị cực trị trong quá trình huấn luyện.

4. Kết luận

Qua các kết quả thu được từ nghiên cứu, bước đầu cho thấy mô hình LSTM có thể xử lý và dự báo chuỗi số liệu khá tốt. Bên cạnh đó, việc dự báo các giá trị cực trị của mô hình còn hạn chế và kết quả khi thay đổi bước thời gian đầu vào cũng có thay đổi. Như vậy, việc chuẩn hóa số liệu ban đầu và lựa chọn bước thời gian làm đầu vào có thể cải thiện hiệu suất của mô hình. Bên cạnh đó, độ mặn cũng bị ảnh hưởng bởi nhiều yếu tố khác như gió, nhiệt độ, chế độ triều, dòng chảy... nhưng nghiên cứu mới chỉ sử dụng một biến đầu vào là độ mặn cao nhất theo ngày. Trong tương lai, nghiên cứu này cần bổ sung thêm các phương pháp xử lý số liệu đầu vào và thử nghiệm mô hình với chuỗi số liệu đa biến để đạt được hiệu quả dự báo tốt hơn.

Đóng góp của tác giả: Xây dựng ý tưởng nghiên cứu: N.T.G., N.C.T.; Xử lý số liệu: N.C.T.; Thiết lập mô hình: N.C.T.; N.T.G.; Viết bản thảo bài báo: N.C.T.; Chỉnh sửa bài báo: N.T.G.

Lời cảm ơn: Nghiên cứu này có sự hỗ trợ về mặt dữ liệu và phương pháp luận từ đề tài mã số ĐTĐL.CN–50/18 do Bộ Khoa học và Công nghệ tài trợ. Bài báo được sự góp ý, vi chỉnh bởi TS. Nguyễn Hữu Duy.

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

Tài liệu tham khảo

1. Nguyen, H.T.; Gupta, A.D. Assessment of water resources and salinity intrusion in the Mekong Delta. *Water Int.* **2001**, *26(1)*, 86–95. https://doi.org/10.1080/02508060108686889.
2. Tran, A.D.; Hoang, L.P.; Bui, M.D.; Rutschmann, P. Simulating future flows and salinity intrusion using combined one–and two–dimensional hydrodynamic modelling–the case of Hau River, Vietnamese Mekong Delta. *Water* **2018**, *10(7)*, 897. https://doi.org/10.3390/w10070897.
3. Doan, V.B.; Kantoush, S.A.; Saber, M.; Mai, N.P.; Maskey, S.; Phong, D.T.; Sumi, T. Long–term alterations of flow regimes of the Mekong River and adaptation strategies for the Vietnamese Mekong Delta. *J. Hydrol. Reg. Stud.* **2020**, *32*, 100742. https://doi.org/10.1016/j.ejrh.2020.100742.
4. Lam, Đ.H.; Phương, N.H.; Đạt, N.Đ.; Giang, N.T. 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 Thuỷ văn* **2022**, *740(1)*, 38–49.
5. Trí, Đ.Q. 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* **2016**, *671*, 39–46.
6. Dũng, Đ.V.; Phương, T.Đ.; Oanh, L.T.; Công, T.T. 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* **2018**, *693*, 48–58.
7. Thai, T.T.; Liem, N.D.; Luu, P.T.; Yen, N.T.M.; Yen, T.T.H.; Quang, N.X.; Tan, L.V.; Hoai, P.N. Performance evaluation of Auto–Regressive Integrated Moving Average models for forecasting saltwater intrusion into Mekong river estuaries of Vietnam. *VN J. Earth Sci.* **2021**, 1–15. https://doi.org/10.15625/2615-9783/16440.
8. Thái, T.H.; Khiêm, M.V.; Thủy, N.B.; Hà, B.M.; Ngọc, P.K. Xây dựng mô hình mạng nơ–ron hồi quy dự báo độ cao sóng có nghĩa tại trạm Cồn Cỏ, Quảng Trị, Việt Nam. *Tạp chí Khí tượng Thủy văn* **2022**, *EME4*, 73–84.
9. Điệp, N.T.H.; Huội, D.; Cần, N.T. Đánh giá tác động của xâm nhập mặn do biến đổi khí hậu trên hiện trạng canh tác lúa tại tỉnh Sóc Trăng. *Tạp chí Khoa học Trường Đại học Cần Thơ* **2017**, 137–143. Doi:10.22144/ctu.jsi.2017.062.

10. Hochreiter, S.; Schmidhuber, J. Long Short–Term Memory. *Neural Comput.* **1997**, *9(8)*, 1735–1780.
11. Yao, K.; Cohn, T.; Vylomova, K.; Duh, K.; Dyer, C. Depth–Gated Recurrent Neural Networks, 2015, pp.1–5. https://arxiv.org/pdf/1508.03790v2.pdf.
12. Koutnik, J.; Greff, K.; Gomez, F.; Schmidhuber, J. A Clockwork RNN, 2014. https://arxiv.org/pdf/1402.3511v1.pdf.
13. Hochreiter. Untersuchungen zu dynamischen neuronalen Netzen, 1991. https://people.idsia.ch/~juergen/SeppHochreiter1991ThesisAdvisorSchmidhuber.pdf.
14. Bengio, S.; Bengio, Y. Taking on the curse of dimensionality in joint distributions using neuralnetworks. IEEE Trans. Neural Networks, Special issue on Data Mining and KnowledgeDiscovery, **2000**, *11(3)*, 550–557.
15. https://colah.github.io/posts/2015–08–Understanding–LSTMs/
16. Nash, J.E.; Sutcliffe, J.V. River Flow Forecasting Through Conceptual Models Part Ia Discussion of Principles. *J. Hydrol.* **1970**, *10*, 282–290.
17. https://en.wikipedia.org/wiki/Root–mean–square_deviation.
18. Kato, T.; Goda, H. Formation and maintenance processes of a stationary band–shaped heavy rainfall observed in Niigata on 4 August 1998. *J. Meteor. Soc. Japan* **2001**, *79*, 899–294.
19. Cross–validation (statistics). https://en.wikipedia.org/wiki/Cross-validation_(statistics).

Building LSTM (Long Short–Term Memory) machine learning model for water salinity forecasting in Dai Ngai

Nguyen Cong Thanh[1], Nguyen Tien Giang[2*]

[1] Southern Regional Hydrometeorological Center; jackynguyen.kttv@gmail.com
[2] Faculty of Hydrology, Meteorology & Oceanography, VNU University of Science, VNU–HN; giangnt@vnu.edu.vn

Abstract: Today, machine learning (ML) has been applied to many fields and hydrometeorological forecasting is one of them. This paper presents the results of building a LSTM (Long Short–Term Memory) model, which is a special form of recurrent neural network (RNN–Recurrent Neural Network) to predict salinity concentration at Dai Ngai gauging station, Soc Trang province. The input data series used is the observed highest daily salinity from 2002–2021. Results obtained during model validation and testing give very good values of RMSE and NSE (NSE > 0.9 in almost all setups), which shows the great potential of using LSTM models for water salinity forecasting in the Mekong Delta.

Keywords: Salinity forcasting; LSTM; Dai Ngai; Soc Trang; Machine Learning.