

# NGHIÊN CỨU ỨNG DỤNG HỌC SÂU ĐỂ DỰ BÁO DÒNG CHẢY NGÀY ĐẾN TRẠM ĐỒNG TRĂNG TRÊN SÔNG CÁI - NHA TRANG

Lê Hữu Minh Quân, Nguyễn Đức Hạnh, Nguyễn Văn Anh Hoàng  
Đại học Khoa học Tự nhiên - Đại học Quốc gia Hà Nội

Ngày nhận bài: 12/9/2024; ngày chuyển phản biện: 13/9/2024; ngày chấp nhận đăng: 10/10/2024

**Tóm tắt:** Tính toán dự báo lưu lượng nước trong sông có vai trò quan trọng trong các nghiên cứu về thủy văn, quy hoạch và quản lý tài nguyên nước, hay các bài toán về cân bằng nước lưu vực. Trong một vài thập kỷ trở lại đây, trí tuệ nhân tạo hay cụ thể hơn là học sâu đã và đang được ứng dụng rất nhiều trong các lĩnh vực, trong đó có nghiên cứu dự báo dòng chảy nhằm nâng cao chất lượng dự báo. Nghiên cứu này ứng dụng mô hình học sâu LSTM kết hợp với biến đổi trung bình trượt để thử nghiệm phân tích chuỗi thời gian và dự báo lưu lượng 1 ngày đến trạm Đồng Trăng trên sông Cái - Nha Trang. Kết quả tính toán này được so sánh với một mô hình phân tích chuỗi thời gian phổ biến là mô hình SARIMA. Từ đó cho thấy việc sử dụng mô hình học sâu LSTM kết hợp với biến đổi trung bình trượt đã làm nâng cao chất lượng dự báo lên rất nhiều so với mô hình SARIMA và so với bản thân mô hình LSTM (khi không sử dụng kết hợp với biến đổi trung bình trượt). Điều này được khẳng định thông qua so sánh đánh giá các chỉ số như MSE, NSE, MAPE, KGE của các phương án dự báo.

**Từ khóa:** Học sâu, học máy, dự báo lưu lượng, LSTM, SARIMA.

## 1. Mở đầu

Dự báo chính xác và đáng tin cậy dòng chảy ngày là rất cần thiết trong quy hoạch và quản lý tài nguyên nước, cũng như trong các ứng dụng liên quan đến nước như nâng cao hiệu quả phát điện của nhà máy thủy điện, quy hoạch và quản lý nông nghiệp, và kiểm soát lũ [1], [2], [3]. Vì vậy, dự báo lưu lượng luôn được quan tâm nghiên cứu từ trước đến nay [4], [5], đặc biệt là trong bối cảnh biến đổi khí hậu, các vấn đề liên quan đến hiện tượng thủy tai diễn biến ngày càng phức tạp, gây khó khăn trong các công tác quy hoạch và quản lý [6].

Dự báo dòng chảy vẫn luôn là một trong những lĩnh vực khó khăn của khoa học thủy văn, đặc biệt là khi hệ thống nghiên cứu được đặc trưng bởi một quá trình động lực học phi tuyến tính, chịu sự tác động của nhiều điều kiện phức tạp, mang tính ngẫu nhiên khác nhau. Hiện nay, có rất nhiều các phương pháp khác nhau

để dự báo dòng chảy ngày. Các phương pháp này có thể phân thành hai hướng tiếp cận là: Các phương pháp dựa trên quá trình (process-driven methods) như các mô hình quan niệm, tổng hợp, mô hình thủy văn, thủy lực; và các phương pháp dựa dữ liệu (data-driven methods) như các mô hình tự hồi quy, các mô hình máy học, các mô hình mạng nơ ron thần kinh [7]. Các phương pháp dựa dữ liệu xác định sự kết nối về mặt toán học giữa các đầu vào và đầu ra. Một trong những phương pháp dựa dữ liệu phổ biến là hướng tiếp cận các mô hình chuỗi thời gian, đây cũng là hướng tiếp cận của nghiên cứu này.

Các kỹ thuật thống kê thông thường được sử dụng cho mô hình hóa chuỗi thời gian trong thủy văn như Auto-Regressive (AR), Auto-Regressive Moving Average (ARMA), Auto-Regressive Integrated Moving Average (ARIMA) hay Seasonal Autoregressive Integrated Moving Average (SARIMA). Mô hình ARIMA được xem là phương pháp phổ biến nhất cho dự báo chuỗi thời gian [8]. Mô hình SARIMA là một sự mở rộng của mô hình ARIMA, xem xét các điều

Liên hệ tác giả: Lê Hữu Minh Quân  
Email: lehuuminhquan\_t66@hus.edu.vn

chính theo mùa (S), được chứng minh là cung cấp dự báo tốt hơn so với mô hình ARIMA trong hầu hết các trường hợp [9].

Những năm gần đây, các phương pháp học máy và học sâu ngày càng phổ biến, được phát triển mạnh mẽ với hàng loạt cải tiến, cung cấp thêm nhiều hướng tiếp cận trong việc xử lý dự báo các đặc trưng của ngành thủy văn bao gồm Artificial Neural Network (ANN) [10], [11], Fuzzy Inference Systems (FIS) [12], [13], Recurrent Neural Network [14] hay Long Short Term Memory (LSTM) [15], và một số nghiên cứu có thể kể đến như [16], [17], [18] đều cho thấy các phương pháp này đem lại hiệu quả tốt trong việc dự báo dòng chảy ngày tại các khu vực nghiên cứu của mình. RNN thuộc một lớp của hệ thống mạng thần kinh nhân tạo, được cho là hiệu quả trong xử lý nhiều vấn đề liên quan đến dự báo trong nhiều nghiên cứu khác nhau [19]. Mặc dù vậy, điểm yếu cố hữu của RNN là gặp vấn đề trong việc học các điểm phụ thuộc dài hạn, theo nghiên cứu của Bengio đã chỉ ra [20]. Mô hình LSTM, một dạng đặc biệt của RNN được tạo ra bởi Hochreiter và Schmidhuber vào năm 1997 nhằm giải quyết vấn đề này [21].

Bên cạnh đó, một đặc điểm chung của các lưu vực sông miền Trung Việt Nam là sông thường ngắn và dốc khiến cho dòng chảy trong mùa lũ có những biến đổi đột ngột, lũ lên nhanh và xuống nhanh. Điều này cũng gây khó khăn cho việc áp dụng các mô hình phân tích chuỗi thời gian như SARIMA, LSTM vào dự báo dòng chảy. Các mô hình này thường khó dự đoán được tốt trong các thời kỳ dòng chảy lớn (đỉnh lũ). Một trong những phương pháp xử lý chuỗi thời gian, làm giảm đi những biến đổi đột ngột đó là sử dụng phép biến đổi trung bình trượt (MA) [22].

Nghiên cứu này tập trung vào ứng dụng mô hình học sâu LSTM kết hợp phép biến đổi trung bình trượt trong dự báo dòng chảy ngày, thử nghiệm với số liệu của một trạm ở miền Trung Việt Nam (trạm Đồng Trăng trên sông Cái - Nha Trang). Kết quả thử nghiệm được đánh giá và so sánh với một mô hình chuỗi thời gian truyền thống là SARIMA dựa trên 5 chỉ số đánh giá là Sai số tuyệt đối trung bình (MSE), Hiệu suất Nash-Sutcliffe (NSE), Hệ số xác định ( $R^2$ ), Sai số phần trăm tuyệt đối trung bình (MAPE) và hiệu

suất Kling Gupta (KGE).

## 2. Phương pháp nghiên cứu

### 2.1. Phép biến đổi trung bình trượt - MA(q)

Nghiên cứu này sử dụng trung bình trượt đơn giản, là một công thức tính trung bình phổ biến của các điểm dữ liệu trước đó trong dữ liệu chuỗi thời gian [22]. Mỗi điểm trong chuỗi thời gian dữ liệu có trọng số như nhau, và được thể hiện với công thức:

$$MA(q) = \frac{P_i + P_{i-1} + \dots + P_{i-(q-1)}}{n} \quad (1)$$

Trong đó,  $P_i$  đại diện cho điểm dữ liệu tại thời gian thứ  $i$  (ngày) và  $q$  đại diện cho số điểm dữ liệu dùng trong tính toán trung bình trượt. Ở nghiên cứu này lựa chọn sử dụng chỉ số  $q = 5$  dựa theo kết quả tính toán trên tập dữ liệu đã được xử lý trung bình trượt sao cho đường quá trình không còn những điểm biến đổi đột ngột mà vẫn giữ được những dao động của những ngày lưu lượng nhỏ.

### 2.2. Mô hình SARIMA

Chữ viết tắt SARIMA có ý nghĩa đại diện cho yếu tố mùa vụ (S), tự hồi quy (AR) dựa trên mối quan hệ giữa quan sát hiện tại và các quan sát trước đó, tích hợp (I), sự khác biệt của các quan sát thực tế để làm cho chuỗi thời gian trở nên dừng (stationary), và trung bình trượt (MA), độ trễ của các lỗi dự báo của mô hình trung bình trượt [19]. Các kỹ thuật mô hình hóa chuỗi thời gian tự hồi quy cổ điển như AR; ARMA; ARIMA; SARIMA có thể được sử dụng như các phương pháp thay thế cho mô hình hóa dòng chảy một biến [20]. SARIMA vượt trội hơn so với các phương pháp khác vì nó có thể xử lý được cả các đặc trưng về xu hướng tiềm năng và chu kỳ trong chuỗi dòng chảy [23]. Các hệ số của mô hình bao gồm 3 hệ số không mùa vụ ( $p, d, q$ ) và 4 hệ số mùa vụ ( $P, D, Q, s$ ), trong đó  $s$  là thời gian của mùa vụ [24]. Các thông số còn lại sẽ được giải thích trong Bảng 1 dựa theo [25].

SARIMA hoạt động hiệu quả khi được ứng dụng cho các bài toán có dữ liệu chuỗi thời gian đủ dài và ổn định [26]. Trong nghiên cứu này, mô hình SARIMA được sử dụng như một mô hình cơ sở để so sánh các kết quả, đánh giá tính

ứng dụng của việc áp dụng mô hình LSTM kết hợp với biến đổi trung bình trượt trong việc dự

báo lưu lượng ngày đến trạm Đồng Trăng trên sông Cái - Nha Trang, tỉnh Khánh Hòa.

Bảng 1. Các siêu tham số của mô hình SARIMA

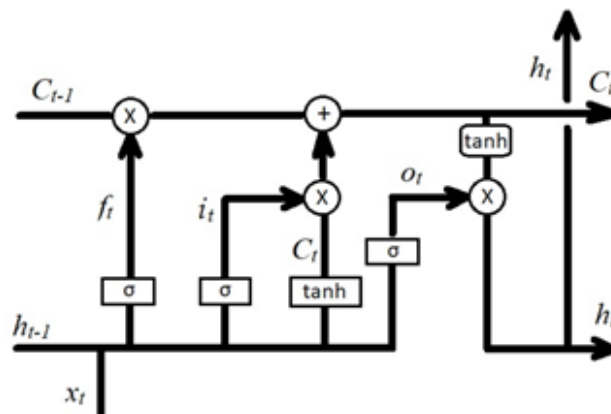
Siêu tham số	Ý nghĩa	Khoảng giá trị
P	Số bậc của phần tự hồi quy mùa vụ: Số lượng giá trị trong quá khứ của chu kỳ mùa vụ được sử dụng để dự đoán giá trị hiện tại	[0-10]
D	Số bậc của quá trình sai phân mùa vụ: Số lần lấy sai phân theo chu kỳ mùa vụ để làm cho chuỗi thời gian trở nên tĩnh	[0-10]
Q	Số bậc của phần di động trung bình mùa vụ: Số lượng nhiễu trong quá khứ của chu kỳ mùa vụ được sử dụng để dự đoán giá trị hiện tại	[0-10]
p	Số bậc của phần tự hồi quy: Số lượng các giá trị trong quá khứ của chuỗi thời gian được sử dụng để dự đoán giá trị hiện tại	[1-7]
d	Số bậc của quá trình sai phân: Số lần lấy sai phân để làm cho chuỗi thời gian trở nên tĩnh (stationary)	[1-7]
q	Số bậc của phần di động trung bình: Số lượng nhiễu (error terms) trong quá khứ được sử dụng để dự đoán giá trị hiện tại	[1-7]

### 2.3. Mô hình LSTM

Một mạng nơ-ron truyền thống coi tất cả các đầu vào và đầu ra là độc lập với nhau, đối với dự báo chuỗi thời gian, việc sử dụng một mạng như vậy sẽ không hợp lý. Một giải pháp thay thế là sử dụng phương pháp LSTM, được phát triển bởi Hochreiter & Schmidhuber (1997), và được chứng minh là hiệu quả trong dự báo chuỗi thời gian, từ đó nó được ứng dụng để dự báo mực nước, lưu lượng dòng chảy trong thủy văn [15], [27].

Động lực phát triển LSTM là để loại bỏ vấn đề gradient biến mất xảy ra với RNN khi xử lý các phụ thuộc dài hạn. RNN tiêu chuẩn bao gồm một chuỗi các module lặp lại của mạng nơ-ron,

trong đó mỗi module bao gồm một cấu trúc, và được cho là khá đơn giản [19]. LSTM được thiết kế đặc biệt để tránh vấn đề phải phụ thuộc dài hạn, giúp việc ghi nhớ này trở thành hành vi tự động thay vì phải cố gắng để học. Cấu tạo của LSTM có dạng chuỗi, tuy nhiên mỗi nút mạng sẽ bao gồm 4 neurol tương tác đặc biệt với nhau thay vì chỉ 1 như của RNN [28]. Điểm cốt lõi của LSTM là ô trạng thái (cell state), đường thẳng cắt ngang qua phần trên của sơ đồ hoạt động trong Hình 1. Các thông số trạng thái bao gồm trạng thái ẩn h (hidden state), đầu vào tại thời điểm t, sau khi xử lý qua các hàm kích hoạt sigmoid ( $\sigma$ ), tanh và các phép toán sẽ cho ra kết quả là trạng thái ô C và trạng thái ẩn h tại thời điểm t, được sử dụng cho các nút mạng t+1 tiếp theo [29].



Hình 1. Sơ đồ hoạt động của mạng LSTM

Bảng 2. Các siêu tham số của mô hình LSTM

Siêu tham số	Ý nghĩa	Khoảng giá trị
batch_size	Số lượng mẫu mà mạng nơ-ron sẽ xử lý đồng thời trước khi thực hiện một lần cập nhật trọng số.	[1-50]
look_back	Xác định số lượng bước thời gian trước đó được sử dụng để dự đoán giá trị tương lai.	[1-5]
epochs	Một vòng lặp đầy đủ qua toàn bộ tập dữ liệu huấn luyện, mỗi epoch bao gồm việc đưa tất cả các mẫu dữ liệu qua mô hình một lần	[10-50]

Trên Bảng 2 là các siêu tham số của mô hình LSTM và khoảng giá trị thường được dùng của các siêu tham số này [30].

Trong rất nhiều nghiên cứu gần đây [15], [17], [25], [27] cho thấy rằng mô hình LSTM hiệu quả hơn nhiều mô hình khác trong bài toán dự báo dòng chảy, tuy nhiên vẫn còn những hạn chế khi dự báo phần dòng chảy cao (đỉnh lũ). Trong nghiên cứu này, các kết quả tính toán bằng mô hình LSTM kết hợp với phép biến đổi trung bình trượt cũng được đánh giá so với việc sử dụng mô hình LSTM đơn lẻ để cho thấy được hiệu quả của việc kết hợp với phép biến đổi trung bình trượt trong dự báo dòng chảy ngày đến trạm Đồng Trăng trên sông Cái - Nha Trang, tỉnh Khánh Hòa, đặc biệt là đối với những phần dòng chảy lớn.

Các siêu tham số của các mô hình SARIMA và LSTM (Bảng 1 và Bảng 2) đóng một vai trò quan trọng ảnh hưởng trực tiếp đến hiệu quả của mô hình [31]. Để có thể lựa chọn giá trị tốt nhất cho các siêu tham số này, có thể có nhiều phương pháp khác nhau như tìm kiếm theo lưới (Grid search), tìm kiếm ngẫu nhiên (Random search), thuật toán tối ưu bayes (Bayesian optimization algorithm). Nghiên cứu này sử dụng thuật toán Grid search, là thuật toán đơn giản, dễ dàng thực hiện và nhanh chóng có kết quả nhất [32] để tìm kiếm các bộ siêu tham số tốt nhất cho các mô hình. Đồng thời, để hạn chế việc mô hình dự đoán quá khớp với dữ liệu (overfitting) [33], phương pháp grid search được kết hợp với thuật toán kiểm tra chéo (Cross Validation) [34] với tham số cv = 5 là giá trị thường được sử dụng trong nhiều nghiên cứu [35], [36].

#### 2.4. Chỉ số đánh giá mô hình

Nghiên cứu này sử dụng 5 chỉ số nhằm xác định hiệu suất của mô hình gồm NSE,  $R^2$ , MAPE,

MSE và KGE thường được sử dụng trong nhiều nghiên cứu khác như [17], [31], [37], [38]

NSE [31] tính toán hiệu quả của một mô hình bằng cách so sánh tổng bình phương chênh lệch giữa các giá trị quan sát được và dự đoán với tổng bình phương chênh lệch giữa các giá trị quan sát được và giá trị trung bình của chúng. Giá trị NSE càng gần 1, mô hình càng chính xác.

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

Hệ số  $R^2$  biểu thị mức độ mà sự biến thiên quan sát được giải thích bởi mô hình, và nằm trong khoảng [0, 1] [39] với điểm  $R^2$  càng gần 1 mô hình sẽ càng tốt. Công thức của  $R^2$  được xác định như sau:

$$R^2 = \frac{(\sum_{i=1}^n (O_i - \bar{O})(P_i - \bar{P}))^2}{\sum_{i=1}^n (O_i - \bar{O})^2 \sum_{i=1}^n (P_i - \bar{P})^2} \quad (3)$$

Mean absolute percentage error (MAPE), còn được gọi là độ lệch phần trăm tuyệt đối trung bình, là một thước đo độ chính xác của phương pháp dự báo trong thống kê từng được sử dụng trong nghiên cứu của [40]. MAPE càng gần 0 càng tốt và được định nghĩa bởi công thức sau:

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{O_i - P_i}{O_i} \right| \quad (4)$$

Một trong những chỉ số thường được sử dụng nhất trong nhiệm vụ dự đoán là MSE [41], trung bình bình phương sai khác giữa các giá trị thực tế và giá trị dự báo. Phạm vi giá trị của MSE từ 0 đến  $+\infty$ , và một kết quả dự đoán tốt nhất là khi MSE = 0. MSE được tính qua Công thức 5:

$$MSE = \frac{1}{n} \sum_{i=1}^n (O_i - P_i)^2 \quad (5)$$

KGE có thể được sử dụng để đánh giá độ phù hợp giữa đầu ra mô hình như mực nước, dữ liệu dòng chảy với dữ liệu quan trắc [41]. Giá trị KGE dao động trong khoảng từ  $-\infty$  (không phù hợp) đến 1 (phù hợp nhất). KGE được tính theo Công thức 6 sau:

$$KGE = 1 - \sqrt{(CC - 1)^2 + \left(\frac{\sigma_P}{\sigma_O} - 1\right)^2 + \left(\frac{\bar{P}}{\bar{O}} - 1\right)^2} \quad (6)$$

$$CC = \frac{\sum_{i=1}^n (O_i - O_m) * (P_i - P_m)}{\sqrt{\sum_{i=1}^n (O_i - O_m)^2} * \sqrt{\sum_{i=1}^n (P_i - P_m)^2}}$$

Trong đó:  $P_i$ ,  $O_i$  là giá trị dự báo, quan trắc thứ  $i$ ;  $\bar{P}$ ,  $\bar{O}$  là giá trị trung bình của dự báo, quan trắc; và  $\sigma_P$ ,  $\sigma_O$  là độ lệch chuẩn của giá trị dự báo, quan trắc.

### 3. Khu vực nghiên cứu và dữ liệu sử dụng

#### 3.1. Khu vực nghiên cứu

Nghiên cứu này tập trung vào lưu vực sông

Cái - Nha Trang, còn được gọi là sông Phú Lộc hoặc sông Cù, thuộc tỉnh Khánh Hòa, Việt Nam. Sông Cái - Nha Trang là con sông dài nhất ở tỉnh Khánh Hòa, với chiều dài gần 80 km, và đóng vai trò rất quan trọng trong phát triển kinh tế - xã hội của tỉnh [42] (xem Hình 2). Diện tích toàn lưu vực là khoảng 1.904 km<sup>2</sup>, trong đó phần diện tích tính đến trạm thủy văn Đồng Trăng là khoảng 1.400 km<sup>2</sup>.

Lưu vực nằm trong khu vực khí hậu nhiệt đới gió mùa, nhưng có những đặc điểm khí hậu đặc biệt: Khí hậu ôn hòa với tính chất đại dương, và được phân chia thành hai mùa rõ rệt: Mùa mưa từ khoảng tháng 10 đến tháng 12 và mùa khô trong các tháng còn lại. Mặc dù mùa mưa chỉ kéo dài ngắn ngủi, nhưng nó lại chiếm hơn 50% tổng lượng mưa hàng năm. Nhiệt độ trung bình hàng năm khoảng 26,7°C. Đặc điểm của sông suối trong lưu vực nghiên cứu là khá dốc và ngắn, phân bố dày với mật độ khoảng 0,6 đến 1 km/km<sup>2</sup>. Sông Cái chảy qua nhiều khu vực nông nghiệp và công nghiệp của tỉnh, cung cấp nước cho các hoạt động nông nghiệp, công nghiệp và sinh hoạt của tỉnh Khánh Hòa.



Hình 2. Bản đồ lưu vực sông Cái - Nha Trang và vị trí trạm Đồng Trăng

### 3.2. Dữ liệu sử dụng

Dữ liệu sử dụng trong nghiên cứu này là dữ liệu dòng chảy ngày từ ngày 01/01/1983 đến ngày 31/12/2022 tại trạm Đồng Trăng trên sông Cái Nha Trang. Bộ dữ liệu kéo dài 39 năm này đủ đại diện cho những diễn biến thủy văn tại lưu vực nghiên cứu và được phân tích, xem xét xu hướng, tính mùa và tính ổn định, kiểm định giả thiết thống kê, sau đó được chuẩn hóa dạng Min - Max sao cho chúng nằm trong khoảng [0,1] trước khi được dùng làm đầu vào cho các mô hình tính toán. Trong nghiên cứu này, hai mô hình SARIMA và LSTM đều được áp dụng cho bài toán tự tương quan, lấy chính dữ liệu lưu lượng ở các bước thời gian trước làm biến đầu vào để tính toán dự đoán cho bước thời gian sau.

Toàn bộ dữ liệu được chia làm 2 phần như sau:

Phần thứ nhất từ 01/01/1983 đến 31/12/2021 dùng cho việc huấn luyện (training) các mô hình.

Phần thứ hai từ ngày 01/01/2022 đến 31/12/2022 được sử dụng để thử nghiệm dự báo lưu lượng 1 ngày tại trạm Đồng Trăng bằng các mô hình đã được huấn luyện.

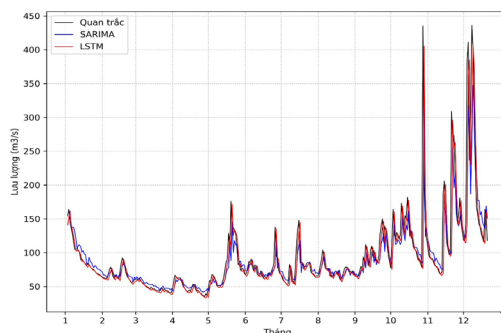
### 4. Kết quả

Giá trị các siêu tham số tốt nhất tìm được bằng phương pháp Grid search của các mô hình được trình bày như trong Bảng 3.

Kết quả sử dụng các mô hình SARIMA và LSTM với các bộ siêu tham số đã được lựa chọn ở trên thử nghiệm dự báo dòng chảy ngày trạm Đồng Trăng trên sông Cái - Nha Trang cho năm 2022 được thể hiện như trên Hình 3. Qua đó ta có thể thấy rằng, cả mô hình SARIMA và LSTM đều cho kết quả dự báo khá tốt, trong đó mô hình LSTM luôn cho kết quả dự báo dòng chảy cao (đỉnh lũ) tốt hơn mô hình SARIMA. Trong đó phần mùa kiệt (nước thấp) thì mô hình SARIMA dự báo tốt hơn, còn về phần mùa lũ (nước cao) thì mô hình LSTM dự báo tốt hơn.

Bảng 3. Bộ siêu tham số mô hình SARIMA và LSTM tốt nhất lựa chọn được bằng phương pháp Grid search

Mô hình	SARIMA						LSTM		
Siêu tham số	p	d	q	P	D	Q	batch_size	look_back	epochs
Giá trị	2	0	2	0	0	0	10	1	20



Hình 3. Kết quả thử nghiệm dự báo dòng chảy ngày (năm 2022) trạm Đồng Trăng trên sông Cái Nha Trang

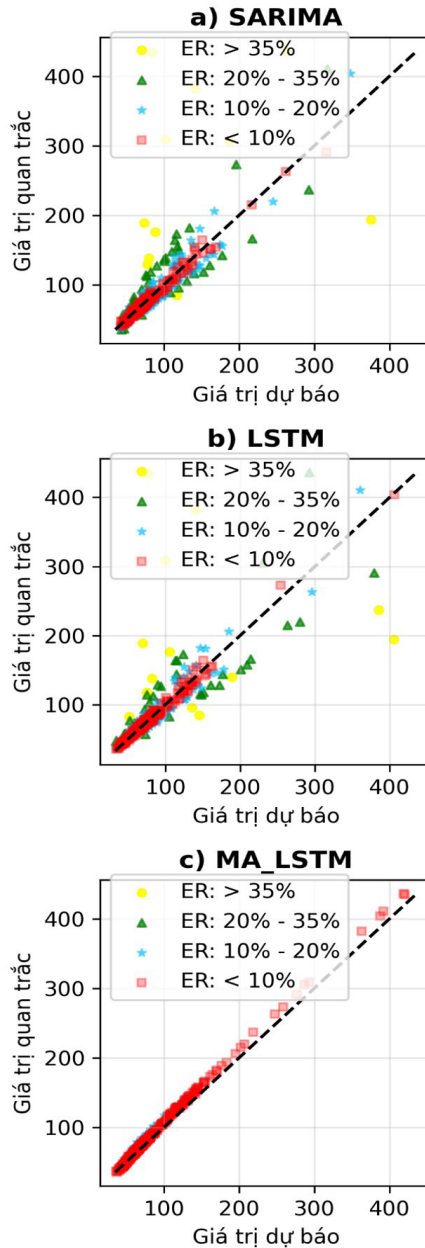
Tuy nhiên, đối với các thời kỳ đỉnh lũ lớn (cuối tháng 10 đến nửa đầu tháng 12) của biểu đồ, có thể thấy kết quả của mô hình LSTM có xu hướng thiên thấp. Do đặc điểm lưu lượng đang từ thấp tăng lên cao đột ngột, gây khó khăn cho việc học của mô hình LSTM. Để khắc phục hạn chế này, phương pháp trung bình trượt - MA(q) (xem phần 2.1) để biến đổi làm trơn các đỉnh,

giảm những biến đổi đột ngột của chuỗi thời gian trước khi đưa vào tính toán bằng mô hình LSTM. Nghiên cứu đã thực hiện phép biến đổi lấy trung bình trượt với nhiều giá trị siêu tham số q khác nhau và kết quả với q = 5 cho hiệu quả tốt nhất. Hình 4 a, b, c lần lượt so sánh cho thấy sự tương quan giữa lưu lượng quan trắc và dự báo bằng mô hình SARIMA, LSTM và LSTM xử

lý trung bình trượt (MA\_LSTM) tương ứng. Qua Hình 4a, 4b, có thể thấy rằng hầu hết các điểm dữ liệu cho thấy sai số phân tán không đều, mô hình kém ổn định và xuất hiện những sai số lớn hơn 35%. SARIMA mặc dù đã có bước xử lý trung bình trượt, tuy nhiên kết quả vẫn chưa đủ độ tin cậy. Bên cạnh đó, mô hình MA\_LSTM (xem Hình 4c) có kết quả dự báo chính xác hơn mô hình SARIMA và LSTM rất nhiều, không có kết quả dự báo nào sai số tương đối lớn hơn

20%, đặc biệt là các kết quả dự báo cho các thời kỳ đỉnh lũ đã chính xác hơn rất nhiều với sai số tương đối đều nhỏ hơn 10%.

Qua đó, có thể khẳng định rằng mô hình LSTM kết hợp xử lý trung bình trượt trên dữ liệu tập Test vượt trội hơn cả LSTM và SARIMA về tính ổn định, chính xác và độ tin cậy trong việc dự đoán lưu lượng. Điều này cũng được thể hiện thông qua việc đánh giá hiệu suất của ba mô hình qua các chỉ số ở Bảng 4.



Hình 4. Biểu đồ sai số tương đối (ER) dự báo bằng các mô hình: a) SARIMA, b) LSTM, c) MA\_LSTM

Bảng 4. So sánh các chỉ số đánh giá của 3 mô hình

Mô hình	SARIMA	LSTM	MA_LSTM
MSE	1123	1212	48
NSE	0,65	0,63	0,98
R2	0,65	0,66	0,99
MAPE	11,67	10,63	4,93
KGE	0,72	0,80	0,91

Qua Bảng 4 có thể thấy rằng mô hình LSTM kết hợp xử lý trung bình trượt cho thấy hiệu suất tốt nhất trên tất cả các chỉ số và mốc thời gian dự báo, đặc biệt vượt trội ở các mốc thời gian ngắn ngày. Mô hình SARIMA và LSTM khá tương đồng, không có sự khác biệt rõ ràng trong các chỉ số đánh giá. Chỉ số MSE của 2 mô hình này gấp hơn 20 lần so với MA\_LSTM, các chỉ số NSE và  $R^2$  chỉ ở ngưỡng khoảng 0,65, sai số tuyệt đối trung bình lớn hơn 10%. Phương án kết hợp hoàn toàn có thể được ứng dụng vào thực tế trong dự báo lưu lượng 1 ngày khi mà chỉ số NSE,  $R^2$ , và KGE đều nằm trên ngưỡng 0,9, MAPE nhỏ hơn 5%.

## 5. Kết luận

Việc dự báo chính xác dòng chảy ngày là vô cùng quan trọng, nhằm đánh giá đúng tác động mà các trận lũ hay hạn mang lại để đưa ra các biện pháp phù hợp và kịp thời, đồng thời hỗ trợ các công tác quản lý tài nguyên nước, sử dụng trong tưới tiêu và sinh hoạt hàng ngày. Đối với các lưu vực vừa và nhỏ, có đặc điểm dòng chảy biến động nhanh và lớn như các lưu vực sông thuộc miền Trung Việt Nam thì vấn đề dự báo dòng chảy càng trở nên khó khăn. Gần đây đã có

nhều nghiên cứu ứng dụng mô hình học sâu vào trong nghiên cứu giải quyết bài toán này. Nghiên cứu này dựa trên kết quả tính toán thử nghiệm với số liệu của trạm Đồng Trăng trên sông Cái Nha Trang cho thấy rằng:

- Theo hướng tiếp cận phân tích chuỗi thời gian truyền thống, sử dụng mô hình phổ dụng là SARIMA cho kết quả đánh giá không cao (MSE = 1123, NSE = 0,65,  $R^2$  = 0,65, MAPE = 11,67, KGE = 0,72).

- Theo hướng tiếp cận học sâu: nếu sử dụng thuật toán LSTM (được đánh giá cao trong nhiều nghiên cứu) cũng không giúp nâng cao đáng kể các kết quả tính toán dự báo (MSE = 1212, NSE = 0,63,  $R^2$  = 0,66, MAPE = 10,63, KGE = 0,80). Tuy nhiên nếu sử dụng kết hợp thuật toán LSTM và phép biến đổi trung bình trượt thì có thể giúp nâng cao kết quả dự báo lên rất nhiều (MSE = 48, NSE = 0,98,  $R^2$  = 0,99, MAPE = 4,93, KGE = 0,91).

Nghiên cứu này cũng có thể là cơ sở cho việc nghiên cứu áp dụng kết hợp các thuật toán máy học, học sâu và các phép biến đổi hoặc các mô hình khác nhằm nâng cao chất lượng dự báo dòng chảy cho các lưu vực sông.

**Đóng góp của từng tác giả trong bài báo:** Xây dựng ý tưởng nghiên cứu: Nguyễn Đức Hạnh; Xử lý số liệu: Lê Hữu Minh Quân; Thu thập tài liệu: Nguyễn Đức Hạnh, Lê Hữu Minh Quân; Viết bản thảo bài báo: Lê Hữu Minh Quân, Nguyễn Văn Anh Hoàng; Chỉnh sửa bài báo: Lê Hữu Minh Quân, Nguyễn Đức Hạnh.

**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. Bassam Bou-Fakhreddine et al. (2018), "Daily river flow prediction based on Two-Phase Constructive Fuzzy Systems Modeling: A case of hydrological - meteorological measurements asymmetry", *Journal of Hydrology*, 558, 255–265, ISSN 0022-1694. <https://doi.org/10.1016/j.jhydrol.2018.01.035>
2. Delafrouz, H., Ghaheri, A. & Ghorbani, M.A. (2018), "A novel hybrid neural network based on



- phase space reconstruction technique for daily river flow prediction*", *Soft Comput*, 22, 2205-2215. <https://doi.org/10.1007/s00500-016-2480-8>
3. Chen, X. et al. (2020), "The importance of short lag-time in the runoff forecasting model based on long short-term memory", *Journal of Hydrology*, 589, 125359, ISSN 0022-1694. <https://doi.org/10.1016/j.jhydrol.2020.125359>
  4. Twedt, T.M. (1977), *National Weather Service extended streamflow prediction [USA]*, Proceedings of the Western Snow Conference.
  5. Day, G.N. (1985), "Extended streamflow forecasting using NWSRFS", *J. Water Resour. Plann. Manage.* 111, 157-170.
  6. Troin, M. et al. (2021), "Generating Ensemble Streamflow Forecasts: A Review of Methods and Approaches Over the Past 40 Years", *Water Resour. Res.* 57(7). <https://doi.org/10.1029/2020wr028392>
  7. Amiri, E. (2015), "Forecasting daily river flows using nonlinear time series models", *Journal of Hydrology* 2015, 527, 1054-1072, ISSN 0022-1694. <https://doi.org/10.1016/j.jhydrol.2015.05.048>.
  8. Masum, S. et al. (2018), "Multi-step Time Series Forecasting of Electric Load Using Machine Learning Models". In: Rutkowski, L., Scherer, R., Korytkowski, M., Pedrycz, W., Tadeusiewicz, R., Zurada, J. (eds) *Artificial Intelligence and Soft Computing*. ICAISC 2018. Lecture Notes in Computer Science, vol 10841. Springer, Cham. [https://doi.org/10.1007/978-3-319-91253-0\\_15](https://doi.org/10.1007/978-3-319-91253-0_15)
  9. He, K. (2021), "Using SARIMA-CNN-LSTM approach to forecast daily tourism demand", *J. Hosp. Tour. Manag.* 49, 25-33. ISSN 1447-6770. <https://doi.org/10.1016/j.jhtm.2021.08.022>
  10. Dibike, Y.B., Solomatine, D.P. (001), "River flow forecasting using artificial neural networks", *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere* 2001, 26(1), 1-7, ISSN 1464-1909. [https://doi.org/10.1016/S1464-1909\(01\)85005-X](https://doi.org/10.1016/S1464-1909(01)85005-X)
  11. Pulido-Calvo, I.; Portela, M.M. (2007), "Application of neural approaches to one-step daily flow forecasting in Portuguese watersheds", *J. Hydrol.* 332, 1-15. <https://doi.org/10.1016/j.jhydrol.2006.06.01>
  12. Huamaní, I.R.L. (2011), "Daily reservoir inflow forecasting using fuzzy inference systems", *Proceedings of the 2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011)*, Taipei, Taiwan, pp. 2745-2751, doi:10.1109/FUZZY.2011.6007690.
  13. Lesot, M.J. (2008), "Fuzzy Prototypes: From a Cognitive View to a Machine Learning Principle" In: Bustince, H.; Herrera, F.; Montero, J. (eds.) *Fuzzy Sets and Their Extensions: Representation, Aggregation and Models*. *Studies in Fuzziness and Soft Computing*, vol. 220. Springer, Berlin, Heidelberg, pp. 431-452. [https://doi.org/10.1007/978-3-540-73723-0\\_22](https://doi.org/10.1007/978-3-540-73723-0_22)
  14. Chang, L.-C. Y.-M. (2004), "A two-step-ahead recurrent neural network for stream-flow forecasting", *Hydrol. Process.* 18, 81-92. <https://doi.org/10.1002/hyp.1313>
  15. Li, J.; Yuan, X. (2023), "Daily Streamflow Forecasts Based on Cascade Long Short-Term Memory (LSTM) Model over the Yangtze River Basin", *Water* 2023, 15, 1019. <https://doi.org/10.3390/w15061019>
  16. Vizi, Z.; Batki, B.; Rátki, L.; et al. (2023), "Water level prediction using long short-term memory neural network model for a lowland river: a case study on the Tisza River, Central Europe", *Environ. Sci. Eur.* 35, 92. <https://doi.org/10.1186/s12302-023-00796-3>
  17. Nifa, K. et al. (2023), "Deep Learning Approach with LSTM for Daily Streamflow Prediction in a Semi-Arid Area: A Case Study of Oum Er-Rbia River Basin, Morocco", *Water* 2023, 15, 262. <https://doi.org/10.3390/w15020262>
  18. Behmard Sabzipour, R. et al. (2023), "Comparing a long short-term memory (LSTM) neural network with a physically-based hydrological model for streamflow forecasting over a Canadian catchment", *Journal of Hydrology*, Volume 627, Part A, December 2023, 130380, <https://doi.org/10.1016/j.jhydrol.2023.130380>
  19. Masum, S.; Liu, Y.; Chiverton, J. (2018), "Multi-step Time Series Forecasting of Electric Load Using

- Machine Learning Models". In: Rutkowski, L., Scherer, R., Korytkowski, M., Pedrycz, W., Tadeusiewicz, R., Zurada, J. (eds) Artificial Intelligence and Soft Computing. ICAISC 2018. Lecture Notes in Computer Science, vol 10841. Springer, Cham. [https://doi.org/10.1007/978-3-319-91253-0\\_15](https://doi.org/10.1007/978-3-319-91253-0_15)*
20. Bengio, Y.; Simard, P.; Frasconi, P. (1994), "Learning long-term dependencies with gradient descent is difficult", *IEEE Trans. Neural Netw.* 5(2), 157-166.
  21. Hochreiter, S.; Schmidhuber, J. (1997), "Long short-term memory", *Neural Comput.* 9(8), 1735-1780.
  22. Hansun, S. (2013), "A new approach of moving average method in time series analysis", *In Proceedings of the 2013 Conference on New Media Studies (CoNMedia)*, Tangerang, Indonesia, pp. 1-4. doi: 10.1109/CoNMedia.2013.6708545.
  23. Moeeni, H.; Bonakdari, H.; Ebtehaj, I. (2017), "Monthly reservoir inflow forecasting using a new hybrid SARIMA genetic programming approach", *J. Earth Syst. Sci.*, 126, 18. <https://doi.org/10.1007/s12040-017-0798-y>
  24. Do, P.; Chow, C.W.K.; Rameezdeen, R.; et al. (2022), "Wastewater inflow time series forecasting at low temporal resolution using SARIMA model: a case study in South Australia", *Environ. Sci. Pollut. Res.* 29, 70984–70999. <https://doi.org/10.1007/s11356-022-20777-y>
  25. Dubey, A.K. et al. (2021), "Study and analysis of SARIMA and LSTM in forecasting time series data", *Sustain Energy Technol. Assess.* 47. <https://doi.org/10.1016/j.seta.2021.101474>
  26. Dimri, T. et al. (2020), "Time series analysis of climate variables using seasonal ARIMA approach", *J. Earth Syst. Sci.* 129, 149. <https://doi.org/10.1007/s12040-020-01408-x>
  27. Zhao, X. et al. (2024), "A Comprehensive Review of Methods for Hydrological Forecasting Based on Deep Learning", *Water*, 16, 1407. <https://doi.org/10.3390/w16101407>
  28. Colah's blog (2015), *Understanding LSTM Networks*, Online: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>
  29. Nguyễn Công Thành, Nguyễn Tiền Giang (2022), "Xây dựng mô hình máy học LSTM (Long Short-Term Memory) phục vụ công tác dự báo mực nước tại trạm đo mực nước Đại Ngãi", *Tạp chí Khoa học Việt Nam*, 740(1), 98–104. doi:10.36335/VNJHM.2022(740(1)).98-104
  30. Li, H. (2020), "Ultra-Short-Term Load Demand Forecast Model Framework Based on Deep Learning", *Energies* 2020, 13, 4900. <https://doi.org/10.3390/en13184900>
  31. Moriasi, D.N. et al. (2015), "Hydrologic and water quality models: Performance measures and evaluation criteria", *Trans. ASABE* 2015, 58, 1763-1785. <https://elibrary.asabe.org/abstract.asp?aid=46548&t=3&dabs=Y&redir=&redirType=>
  32. Raghavendra, S.N.; Deka, P.C. (2014), "Support vector machine applications in the field of hydrology: A review", *Appl. Soft Comput.* 19, 372-386. ISSN 1568-4946. <https://doi.org/10.1016/j.asoc.2014.02.002>
  33. Tran Thai et al. (2022), "Application of Long Short-Term Memory neural network for time series prediction of flow rate at My Thuan hydrology station, Tien river", *Sci. Technol. Dev. J. Nat. Sci.* 6(1). DOI:10.32508/stdjns.v6i1.1129
  34. Zanetti, M. et al. (2022), "Spatio-temporal cross-validation to predict pluvial flood events in the Metropolitan City of Venice", *J. Hydrol.*, 612, Part B, 128150. ISSN 0022-1694.
  35. Ali, A.; Gravino, C. (2021), "An empirical comparison of validation methods for software prediction models", *J. Softw. Evol. Proc.*, 33, e2367. <https://doi.org/10.1002/smr.2367>
  36. Kohavi, R. (2001), "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 2001.
  37. Jain, A.; Kumar, A.M. (2007), "Hybrid neural network models for hydrologic time series forecasting", *Appl. Soft Comput.* 7, 585-592. ISSN 1568-4946. <https://doi.org/10.1016/j.asoc.2006.03.002>
  38. Li, L.; Jun, K.S. (2022), "A Hybrid Approach to Improve Flood Forecasting by Combining a Hydrodynamic Flow Model and Artificial Neural Networks", *Water*, 14, 1393. <https://doi.org/10.3390/w14091393>
  39. Nifa, K. et al. (2023), "Deep Learning Approach with LSTM for Daily Streamflow Prediction in a

*Semi-Arid Area: A Case Study of Oum Er-Rbia River Basin, Morocco*, *Water*, 15, 262. <https://doi.org/10.3390/w15020262>

40. Nguyen Duc Hanh et al. (2023), "Modelling Stage-discharge Relationships for Hydrological Stations in the Da River Basin using First-kind Chebyshev Polynomial Approximation", *VNU Journal of Science: Earth and Environmental Sciences*, 39, 16-30. <https://doi.org/10.25073/2588-1094/vnuees.4956>
41. Dang Tran Duc, M. et al. (2023), "A comparative analysis of regression equations for rating curve development at a gauging station in Da river, Northern Vietnam", *Viet Nam Journal of Hydrometeorology*, 14, 53-69. doi:10.36335/VNJHM.2023(14).53-69
42. Nguyễn Thị Phương và cộng sự (2017), "Ứng dụng mô hình SWAT đánh giá tác động của biến đổi khí hậu đến dòng chảy lũ trạm Đồng Trăng, lưu vực sông Cái, Nha Trang tỉnh Khánh Hòa", *Tạp chí Khí tượng Thủy văn*, 682, 15-22.

## APPLICATION OF DEEP LEARNING METHODS FOR DAILY STREAMFLOW FORECASTING AT THE DONG TRANG STATION ON THE CAI RIVER - NHA TRANG

Le Huu Minh Quan, Nguyen Duc Hanh, Nguyen Van Anh Hoang  
*Hanoi University of Science - Viet Nam National University*

*Received: 12/9/2024; Accepted: 10/10/2024*

**Abstract:** *Forecasting river flow is crucial in hydrological studies, water resource planning and management, and watershed water balance problems. In recent decades, artificial intelligence, specifically deep learning, has been extensively applied across various fields, including streamflow forecasting, to improve prediction quality. This study applies the LSTM deep learning model combined with moving average transformation to experiment with time series analysis and daily flow forecasting at the Dong Trang station on the Cai River - Nha Trang. The results are compared with a popular time series analysis model, SARIMA. The findings indicate that using the LSTM deep learning model combined with moving average transformation considerably enhances forecast quality compared to the SARIMA model and the LSTM model alone (without moving average transformation). This is confirmed by comparing evaluation metrics such as MSE, NSE, MAPE, and KGE of the forecasting approaches.*

**Keywords:** *Deep learning, machine learning, flow forecasting, LSTM, SARIMA.*