Tạp chí Khoa học Đại học Cần Thơ
Số chuyên đề: Khoa học tự nhiên Đồng bằng sông Cửu Long
ISSN 1859-2333 | e-ISSN 2815-5599

DOI:10.22144/ctujos.2026.198

PHÂN TÍCH VÀ DỰ BÁO LƯỢNG MƯA TẠI ĐÀ LẠT VỚI CÁC MÔ HÌNH THỐNG KÊ VÀ MÔ HÌNH LAI HỌC SÂU

Văn Thị Huỳnh Như[1], Thân Thị Hương Mai[1], Ôn Bích Gia Hân[1] và Trần Phước Lộc[2,*]
[1]*Sinh viên ngành Thống kê Khóa 49, Trường Khoa học Tự nhiên, Đại học Cần Thơ, Việt Nam*
[2]*Trường Khoa học Tự nhiên, Đại học Cần Thơ, Việt Nam*
**Tác giả liên hệ (Corresponding author): tploc@ctu.edu.vn*

Thông tin chung (Article Information)

Nhận bài (Received): 08/06/2026
Sửa bài (Revised): 30/06/2026
Duyệt đăng (Accepted): 26/08/2026

Title: *Rainfall analysis and forecasting in Da Lat using statistical and hybrid deep learning models*

Author(s): *Van Thi Huynh Nhu[1], Than Thi Huong Mai[1], On Bich Gia Han[1] and Tran Phuoc Loc[2,*]*

Affiliation(s): *Faculty of Mathematics, College of Natural Sciences, Can Tho University, Viet Nam*

TÓM TẮT

Trong nghiên cứu này, chuỗi số liệu lượng mưa trung bình tháng tại trạm khí tượng Đà Lạt trong giai đoạn 01/2002–12/2024 đã được sử dụng nhằm đánh giá và dự báo xu thế biến động lượng mưa. Mười giá trị thiếu (3,62%) được xử lý bằng nội suy tuyến tính sau khi kiểm tra cơ chế thiếu dữ liệu. Bảy mô hình dự báo được xem xét, gồm bốn mô hình đơn lẻ (Holt–Winters, SARIMA, Prophet và LSTM) và ba mô hình lai kết hợp với LSTM trên phần dư (SARIMA+LSTM, HW+LSTM và Prophet+LSTM). Mô hình LSTM được cải tiến thông qua kỹ thuật ensemble với hai hạt giống ngẫu nhiên, bổ sung đặc trưng mùa vụ dạng sin/cos và cửa sổ trễ 24 tháng. Kết quả cho thấy Prophet+LSTM đạt sai số MAE thấp nhất (61,84 mm) trong các mô hình khảo sát, đồng thời phần dư thỏa mãn các giả định về nhiễu trắng và phân phối chuẩn. Kết quả dự báo từ các mô hình nghiên cứu cung cấp cơ sở khoa học cho quản lý tài nguyên nước và giảm nhẹ rủi ro thiên tai tại khu vực cao nguyên Lâm Viên.

Từ khóa: *Dự báo lượng mưa, LSTM, mô hình lai, Prophet, SARIMA*

ABSTRACT

This study used monthly average rainfall data from the Da Lat meteorological station for the period January 2002 to December 2024 to analyze and forecast rainfall variability. Ten missing observations (3.62%) were imputed using linear interpolation after assessing the missing-data mechanism. Seven forecasting models were evaluated, including four standalone models (Holt–Winters, SARIMA, Prophet, and LSTM) and three residual-based hybrid models (SARIMA+LSTM, HW+LSTM, and Prophet+LSTM). The LSTM component was enhanced through an ensemble strategy with two random seeds, the incorporation of sinusoidal seasonal features (sine/cosine), and a 24-month lookback window. The results showed that the Prophet+LSTM hybrid model achieved the lowest MAE (61.84 mm) among the models considered. In addition, the residuals satisfied the assumptions of white noise and approximate normality, indicating adequate model performance. Forecast results from research models provide a scientific basis for water-resource management and disaster-risk mitigation in the Lam Vien Plateau region.

Keywords: *Hybrid model, LSTM, Prophet, rainfall forecasting, SARIMA*

1. GIỚI THIỆU

Lượng mưa là biến khí hậu có tính biến động cao theo cả không gian lẫn thời gian, là yếu tố đầu vào không thể thiếu trong các mô hình thủy văn, nông nghiệp và đánh giá rủi ro thiên tai. Trong bối cảnh biến đổi khí hậu, tần suất và cường độ của các sự kiện mưa cực đoan trên phạm vi toàn cầu đang gia tăng, làm trầm trọng thêm thách thức trong công tác quản lý tài nguyên nước và phòng chống thiên tai (IPCC, 2023).

Tại Việt Nam, chế độ mưa mang đặc trưng nhiệt đới gió mùa với lượng mưa trung bình năm dao động từ khoảng 1.500 mm ở các vùng đồng bằng đến trên 3.000 mm tại một số khu vực miền núi và ven biển. Sự tương tác giữa địa hình, hoàn lưu gió mùa Đông Nam Á và các hệ thống thời tiết như bão, áp thấp nhiệt đới tạo nên tính biến động cao và khó dự báo của lượng mưa. Riêng năm 2024, thiệt hại do thiên tai liên quan đến mưa, bão và ngập lụt ước tính hơn 90.264.894 triệu đồng (Cục Quản lý đê điều và phòng, chống thiên tai, 2024), đặt ra yêu cầu cấp thiết đối với công tác quan trắc và dự báo.

Nằm trên cao nguyên Lâm Viên ở độ cao trung bình 1.500 m so với mực nước biển, Đà Lạt chịu tác động đặc biệt của chế độ mưa nhiệt đới núi cao, với địa hình dốc làm gia tăng nguy cơ sạt lở đất và lũ quét. Kết quả phân tích chuỗi số liệu giai đoạn 2002 – 2024 tại trạm Đà Lạt cho thấy tổng lượng mưa năm có xu hướng tăng tuyến tính. Các trận mưa cực đoan vượt ngưỡng lịch sử đã dẫn đến hàng loạt sự cố sạt lở nghiêm trọng tại đường Hoàng Hoa Thám (tháng 6/2023) và các khu vực dân cư lân cận trong giai đoạn 2024, gây thiệt hại lớn về người và tài sản. Việc phát triển mô hình dự báo lượng mưa với độ chính xác cao là nhu cầu thực tiễn nhằm tối ưu hóa công tác quản lý rủi ro thiên tai và nông nghiệp công nghệ cao của địa phương.

Hai nhóm mô hình chính đã được ứng dụng rộng rãi tại nhiều trạm khí tượng Việt Nam. Nhóm mô hình thống kê truyền thống như ARIMA, SARIMA, Holt–Winters có ưu điểm rõ ràng về tính tường minh và yêu cầu dữ liệu khiêm tốn. Huynh et al. (2024) áp dụng SARIMA cho chuỗi 45 năm tại 12 trạm Đồng bằng sông Cửu Long để dự báo xu thế lượng mưa năm, đồng thời xác định mối liên hệ giữa biến động lượng mưa với hiện tượng El Niño/La Niña. Dimri et al. (2020) chứng minh hiệu quả của SARIMA trong phân tích chuỗi thời gian lượng mưa và nhiệt độ tại khu vực núi cao (lưu vực sông Bhagirathi, Uttarakhand, Ấn Độ) với cấu trúc mùa vụ rõ rệt.

Nhóm mô hình học sâu như LSTM (Hochreiter & Schmidhuber, 1997), GRU và Transformer khắc phục hạn chế tuyến tính của nhóm trước nhờ khả năng học các phụ thuộc dài hạn phi tuyến. Kratzert et al. (2018) chứng minh LSTM vượt trội so với các mô hình truyền thống trong mô hình mưa–dòng chảy. Nguyen et al. (2023) áp dụng thành công LSTM cho dự báo dòng chảy ở Đồng bằng sông Cửu Long. Tuy nhiên, LSTM đòi hỏi nhiều dữ liệu huấn luyện và khó diễn giải về mặt thống kê. Can và Lộc (2024) phân tích và dự báo lưu lượng nước ở trạm đo Tân Châu, An Giang với các mô hình thống kê và học máy. Xu hướng phát triển mô hình lai (hybrid), kết hợp khả năng mô hình hóa tuyến tính của thống kê với sức mạnh phi tuyến của học sâu, đang được quan tâm. Zhang (2003) đề xuất khung lai ARIMA–ANN nguyên thủy, Wu et al. (2021) phát triển dạng Wavelet–ARIMA–LSTM và Cong et al. (2025) tích hợp ECMWF S2S với Random Forest và XGBoost cho dự báo mưa cực đoan tại bảy vùng khí hậu Việt Nam.

Mặc dù các hướng tiếp cận đã có những tiến bộ đáng kể, nhưng vẫn chưa có nghiên cứu nào được thực hiện nhằm so sánh trực diện và hệ thống giữa cả ba nhóm mô hình thống kê, học sâu và lai trên cùng bộ dữ liệu tại khu vực cao nguyên Lâm Viên. Trong nghiên cứu này, việc đánh giá hiệu năng của bảy cấu trúc mô hình thông qua sáu chỉ số sai số (MAE, RMSE, MAPE, NRMSE, R^2, Theil U) và bổ sung kiểm định Diebold–Mariano đã tiến hành để xác định ý nghĩa thống kê của khác biệt giữa các cặp mô hình và chẩn đoán phần dư bằng Ljung–Box và Shapiro–Wilk. Cấu trúc mô hình lai được nâng cấp bằng kỹ thuật ensemble nhiều hạt giống và bổ sung đặc trưng mùa (sin/cos tháng) ở đầu vào LSTM. Đóng góp chính của nghiên cứu gồm: (i) cung cấp bằng chứng thực nghiệm về hiệu năng các mô hình tại khu vực cao nguyên Lâm Viên, (ii) đề xuất quy trình hybrid Prophet+LSTM nâng cấp có khả năng diễn giải tốt và sai số tối ưu và (iii) cung cấp dự báo lượng mưa 12 tháng năm 2025 kèm khoảng tin cậy 95% phục vụ công tác quản lý thực tiễn.

2. DỮ LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Dữ liệu nghiên cứu

Bộ dữ liệu lượng mưa trung bình tháng tại trạm khí tượng Đà Lạt, tỉnh Lâm Đồng, giai đoạn 01/2002 – 12/2024 (276 quan sát) được thu thập từ Hệ thống cơ sở dữ liệu quốc gia về khí hậu và thống kê khí tượng của Tổng cục Khí tượng Thủy văn Việt Nam (mã dữ liệu: V0109) (Cục Thống kê, 2025). Dữ liệu được tổ chức dưới dạng bảng với ba biến: năm

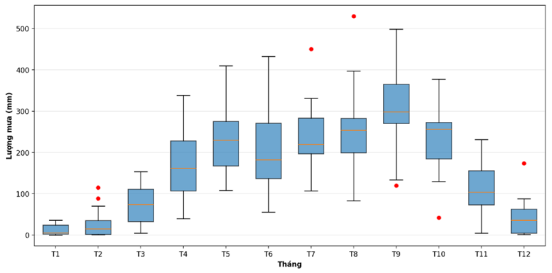
(year), tháng (month) và lượng mưa (mm). Trong 276 quan sát, có 10 giá trị thiếu (chiếm 3,62%) phân bố rải rác theo thời gian, không hình thành đoạn thiếu liên tục dài.

Đặc trưng cơ bản của chuỗi được tóm tắt ở Bảng 1, thống kê theo từng tháng dựa trên các quan sát hợp lệ trước khi điền giá trị thiếu.

Bảng 1. Thống kê mô tả lượng mưa theo tháng tại trạm Đà Lạt (2002 – 2024, trước khi điền giá trị thiếu)

Tháng	N	Mean	SD	Min	Max
1	18	11,98	12,44	0,1	35,7
2	20	27,48	33,45	0,2	115,1
3	23	73,73	48,66	4,9	153,7
4	23	169,21	78,63	39,5	338,5
5	23	232,03	83,62	108,2	410,0
6	23	207,00	102,90	55,6	432,8
7	23	239,67	75,44	106,7	450,1
8	23	249,69	95,18	83,0	530,0
9	23	305,15	89,08	120,0	498,7
10	23	236,79	79,69	42,0	377,7
11	23	111,95	61,49	5,0	231,0
12	21	40,67	41,73	0,9	173,9

Ghi chú: N: số quan sát hợp lệ, Mean: trung bình, SD: độ lệch chuẩn, Max: giá trị lớn nhất, Min: giá trị nhỏ nhất, đơn vị đo: mm.



Hình 1. Biểu đồ hộp lượng mưa trung bình theo tháng tại Đà Lạt (2002 – 2024)

Bảng 1 và Hình 1 cho thấy lượng mưa tại Đà Lạt phân hóa rõ rệt theo mùa: mùa khô (tháng 12–3) có lượng mưa thấp (trung bình < 75 mm/tháng) và phương sai nhỏ, mùa mưa (tháng 5–10) lượng mưa cao (trung bình > 200 mm/tháng) với phương sai lớn và xuất hiện các giá trị cực trị (> 450 mm). Tháng 9 là đỉnh mưa với trung bình 305,15 mm và biên độ rộng (120,0 – 498,7 mm). Sự khác biệt phương sai lớn giữa các tháng cho thấy chuỗi có cấu trúc mùa vụ rõ rệt nhưng tính ngẫu nhiên cao trong các tháng cao điểm.

2.2. Xử lý giá trị thiếu

Mười giá trị thiếu phân bố rải rác theo thời gian, không hình thành đoạn thiếu liên tục dài. Để xác định cơ chế thiếu dữ liệu, mô hình hồi quy logistic đã được sử dụng với biến phụ thuộc là trạng thái dữ liệu ($M = 1$ nếu quan sát hợp lệ, $M = 0$ nếu thiếu) và các biến độc lập là year và month:

$$\text{logit}(P(M = 1)) = \beta_0 + \beta_1 year + \beta_2 month. \quad (1)$$

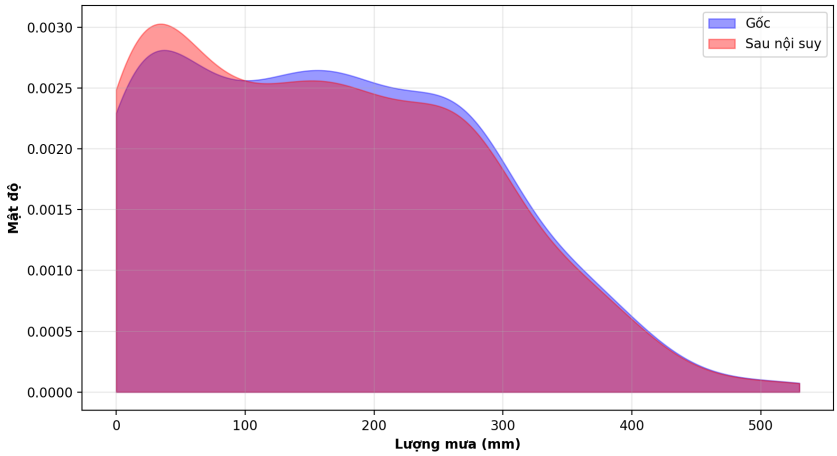
Kết quả hồi quy từ mô hình (1): $\beta_2 = 0{,}3188$ ($p = 0{,}011$) có ý nghĩa thống kê và $\beta_1 = -0{,}034$ ($p = 0{,}491$) không có ý nghĩa thống kê với mức ý nghĩa 5%. Kết quả cho thấy cơ chế thiếu dữ liệu không hoàn toàn ngẫu nhiên (MCAR) và có liên quan đến tháng quan sát, do đó giả thiết thiếu ngẫu nhiên (MAR) được xem là hợp lý cho mục đích xử lý dữ liệu thiếu. Tuy nhiên, do tỷ lệ thiếu thấp (3,62%) và các giá trị thiếu phân bố rải rác nên *nội suy tuyến tính* được sử dụng để điền giá trị thiếu:

$$y_t = y_{t_1} + \frac{t - t_1}{t_2 - t_1}(y_{t_2} - y_{t_1}). \quad (2)$$

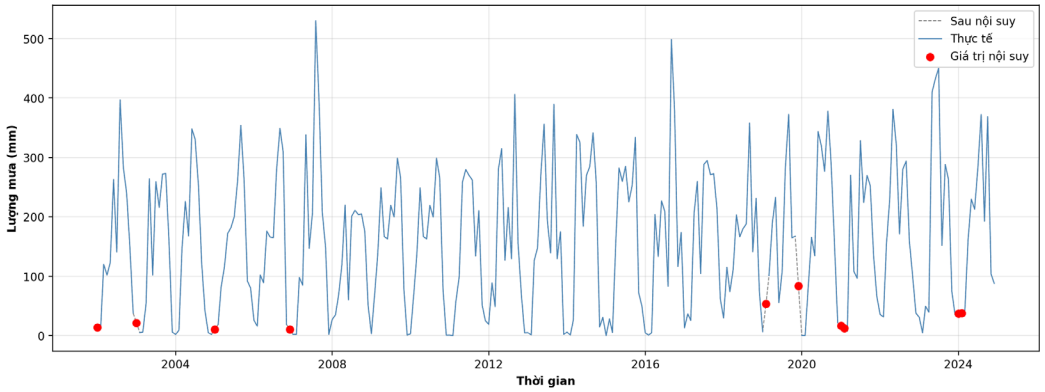
Trong đó, $t_1$và t_2 là hai thời điểm hợp lệ gần nhất trước và sau vị trí thiếu. Kết quả nội suy được thể hiện trực quan ở Hình 3, các điểm đỏ tương ứng các vị trí được điền giá trị. Các phương pháp điền giá trị thiếu khác, ví dụ Kalman (Moritz & Bartz-Beielstein, 2017) cũng có thể được xem xét sử dụng. Để đánh giá mức độ bảo toàn cấu trúc thống kê sau nội suy, các chỉ số mô tả của chuỗi trước và sau khi điền giá trị thiếu đã được so sánh (Bảng 2) và kết quả ước lượng hàm mật độ xác suất của lượng mưa trước và sau nội suy tuyến tính bằng phương pháp ước lượng mật độ kernel (Kernel Density Estimation - KDE) với hạt nhân Gaussian đã được trình bày tại Hình 2.

Bảng 2. So sánh chỉ số thống kê của chuỗi lượng mưa trước và sau nội suy tuyến tính

Chỉ số	Trước nội suy	Sau nội suy	Chênh lệch
Mean	163,91	159,05	−4,86
SD	118,86	119,43	+0,57
Min	0,10	0,10	0,00
Max	530,00	530,00	0,00
Median	158,25	152,80	−5,45



Hình 2. Biểu đồ ước lượng hàm mật độ xác suất của lượng mưa trước và sau nội suy tuyến tính



Hình 3. Chuỗi lượng mưa Đà Lạt trước (xanh) và sau nội suy tuyến tính (xám đứt)

Ghi chú: Chấm đỏ là giá trị nội suy.

Các chỉ số mô tả thay đổi không đáng kể: trung bình giảm 4,86 mm (~3%), độ lệch chuẩn tăng nhẹ 0,57 mm, các giá trị Min/Max được bảo toàn hoàn toàn. Các đường mật độ ước lượng (Hình 2) trước và sau nội suy gần như trùng khít, cho thấy việc nội suy tuyến tính không làm thay đổi đáng kể phân bố thực nghiệm của chuỗi.

2.3. Các mô hình dự báo

2.3.1. Holt–Winters

Mô hình Holt–Winters (Holt, 2004; Winters, 1960) là kỹ thuật làm trơn mũ ba thành phần (mức độ, xu hướng và mùa vụ) cho chuỗi có cả xu hướng và mùa vụ. Do dữ liệu lượng mưa trung bình tháng thể hiện đặc điểm mùa vụ tương đối ổn định theo thời gian nên dạng cộng đã được sử dụng. Các phương trình cập nhật được xác định như sau:

$$\hat{y}_{t+h|t} = L_t + hT_t + S_{t+h-m(k+1)}. \quad (3)$$

Trong đó:

$$L_t = \alpha\,(Y_t - S_{t-m}) + (1-\alpha)(L_{t-1} + T_{t-1}),$$

$$T_t = \beta\,(L_t - L_{t-1}) + (1-\beta)T_{t-1},$$

$$S_t = \gamma\,(Y_t - L_t) + (1-\gamma)S_{t-m}.$$

Với Y_t là giá trị quan sát tại thời điểm t; L_t, T_t và S_t lần lượt là thành phần mức độ, xu hướng và mùa vụ; α, β và γ $(0 < \alpha, \beta, \gamma < 1)$ là các tham số làm trơn; $m = 12$ (chu kỳ mùa năm); h là bước dự báo; k là chỉ số chu kỳ.

2.3.2. SARIMA

Mô hình $SARIMA(p,d,q)(P,D,Q)_m$ theo phương pháp Box–Jenkins (Box et al., 2015) mở rộng mô hình $ARIMA(p,d,q)$ nhằm mô tả đồng thời các thành phần tự hồi quy, trung bình trượt và cấu trúc mùa vụ của chuỗi thời gian. Dạng tổng quát của mô hình được biểu diễn bởi:

$$\Phi(B^m)\varphi(B)(1-B)^d(1-B^m)^D y_t = \Theta(B^m)\theta(B)\varepsilon_t. \quad (4)$$

Trong đó: B là toán tử lùi; $\varphi(B)$, $\theta(B)$ tương ứng các đa thức AR, MA bậc p, q; Φ, Θ là các đa thức AR, MA mùa vụ bậc P, Q; d, D là bậc sai phân thường và sai phân mùa; $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. Mô hình SARIMA này thu gọn thành $ARIMA(p, d, q)$ khi không tồn tại thành phần mùa vụ, tức là $P = Q = D = 0$.

2.3.3. Prophet

Mô hình Prophet (Taylor & Letham, 2018) là một mô hình dự báo chuỗi thời gian theo cách tiếp cận phân rã cộng, được thiết kế để xử lý tốt các chuỗi có xu hướng phi tuyến, tính mùa vụ đa chu kỳ và các điểm thay đổi đột ngột. Mô hình Prophet phân rã chuỗi thời gian Y_t thành bốn thành phần:

$$y_t = g(t) + s(t) + \hbar(t) + \varepsilon_t. \quad (5)$$

Trong đó: $g(t)$ là thành phần xu hướng, $s(t)$ là thành phần mùa vụ, $\hbar(t)$ là thành phần ngày lễ và sự kiện đặc biệt, $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ là nhiễu trắng.

Thành phần xu hướng $g(t)$ được mô hình hóa theo một trong hai dạng: mô hình tăng trưởng logistic khi chuỗi có giới hạn bão hòa:

$$g(t) = \frac{L}{1 + \exp(-k(t - m))}, \quad (6)$$

hoặc mô hình tuyến tính từng khúc (piecewise linear) khi chuỗi không có giới hạn bão hòa:

$$g(t) = (k + a(t)^T \delta)t + (m + a(t)^T \gamma). \quad (7)$$

Trong đó: L là giá trị bão hòa, k là tốc độ tăng trưởng, m là tham số dịch chuyển, δ là vector thay đổi tốc độ tại các điểm thay đổi, $a(t)$ là vector chỉ số xác định điểm thay đổi nào đang hoạt động tại thời điểm t.

Thành phần mùa vụ $s(t)$ được xấp xỉ bằng chuỗi Fourier bậc N:

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos \frac{2\pi n t}{P} + b_n \sin \frac{2\pi n t}{P} \right). \quad (8)$$

Trong đó: P là chu kỳ mùa, N là bậc của chuỗi Fourier kiểm soát độ mịn của mùa vụ, a_n, b_n là các tham số cần ước lượng.

Thành phần ngày lễ và sự kiện đặc biệt $\hbar_t$ được mô hình hóa như sau:

$$\hbar_t = Z(t)\, k. \quad (9)$$

Trong đó, $Z(t) = [\, I(t \epsilon D_1), \ldots, I(t \epsilon D_k)]$ là vector chỉ số nhị phân xác định thời điểm t có thuộc khoảng ảnh hưởng của sự kiện thứ j hay không, D_j là tập hợp các ngày thuộc sự kiện j và vùng lân cận của nó, $k = (k_1, \ldots, k_K)^T$ là vector tham số biểu thị mức độ ảnh hưởng của từng sự kiện lên chuỗi, với K là tổng số sự kiện đặc biệt được xác định trước. Các tham số của mô hình được ước lượng theo khung suy luận Bayes sử dụng phương pháp xấp xỉ MAP (Maximum a posteriori).

2.3.4. LSTM (Long Short-Term Memory)

Mạng LSTM (Hochreiter & Schmidhuber, 1997) là một biến thể của mạng nơ-ron hồi quy (Recurrent Neural Network – RNN), được thiết kế nhằm khắc phục hạn chế suy giảm hoặc bùng nổ gradient khi mô hình hóa các phụ thuộc dài hạn trong chuỗi thời gian. Khác với RNN truyền thống, LSTM sử dụng cơ chế cổng để điều tiết quá trình ghi nhớ, loại bỏ và truyền tải thông tin qua các thời điểm.

Tại thời điểm t, với vector đầu vào x_t và trạng thái ẩn trước đó h_{t-1} và trạng thái ô nhớ trước đó c_{t-1}, các thành phần của LSTM được xác định như sau:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i), \quad (10)$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f), \quad (11)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o), \quad (12)$$

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c), \quad (13)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \quad (14)$$

$$h_t = o_t \odot \tanh(c_t). \quad (15)$$

Trong đó, i_t, f_t và o_t lần lượt là cổng vào, cổng quên và cổng ra tại thời điểm t; $\tilde{c}_t$ là trạng thái ô nhớ ứng viên; c_t là trạng thái ô nhớ hiện tại; h_t là trạng thái ẩn đầu ra. Các ma trận W_i, W_f, W_o, W_c biểu thị trọng số liên kết giữa đầu vào và các cổng tương ứng; U_i, U_f, U_o, U_c là các ma trận trọng số liên kết với trạng thái ẩn trước đó; còn b_i, b_f, b_o, b_c là các vector độ lệch cần được ước lượng trong quá trình huấn luyện. Hàm $\sigma(\cdot)$ là hàm sigmoid, $\tanh(\cdot)$ là hàm tiếp tuyến hyperbolic và $\odot$ là phép nhân từng phần tử (Hadamard product).

2.3.4. Mô hình lai phân tách phần dư

Mô hình lai phân tách phần dư được xây dựng theo hướng tiếp cận của Zhang (2003), giả định chuỗi thời gian gồm hai thành phần độc lập: thành phần tuyến tính L_t và thành phần phi tuyến N_t,

$$y_t = L_t + N_t. \quad (16)$$

Trước hết, một mô hình dự báo thống kê $\mathcal{M}$ (ví dụ, Holt–Winters, SARIMA hoặc Prophet) được sử dụng để ước lượng thành phần tuyến tính $\hat{L}_t$. Phần dư

$$e_t = y_t - \hat{L}_t, \quad (17)$$

sau đó được đưa vào mạng LSTM nhằm học cấu trúc phi tuyến còn lại và tạo dự báo $\hat{N}_t$. Dự báo cuối cùng của mô hình lai được xác định bởi

$$\hat{y}_t = \hat{L}_t + \hat{N}_t. \quad (18)$$

Trong nghiên cứu này, để nâng cao khả năng dự báo khi dữ liệu chuỗi ngắn, mô hình LSTM trên phần dư được sử dụng phương pháp tìm kiếm lưới và cải tiến theo ba hướng: (i) sử dụng ensemble gồm hai mô hình LSTM huấn luyện với các hạt giống ngẫu nhiên khác nhau và lấy trung bình dự báo nhằm giảm phương sai do khởi tạo; (ii) bổ sung các biến mùa vụ $\sin(2\pi m/12)$ và $\cos(2\pi m/12)$, với m là tháng trong năm, để biểu diễn chu kỳ mùa một cách tường minh và (iii) sử dụng cửa sổ trễ (lookback) 24 tháng nhằm khai thác thông tin từ hai chu kỳ mùa liên tiếp.

2.4. Tiêu chí đánh giá và kiểm định

Giả sử $\{y_t, t = 1,2,\dots,T\}$, là chuỗi thời gian quan sát và $\hat{y}_t$ là giá trị dự báo tương ứng, gọi n $(n \le T)$ là quy mô mẫu của tập dữ liệu được phân tích. Để định lượng và so sánh hiệu quả dự báo giữa các mô hình, các tiêu chí sai số sau được sử dụng.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|, \quad (19)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}, \quad (20)$$

$$MAPE = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|, \quad (21)$$

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} \cdot 100\%, \quad (22)$$

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2}, \quad (23)$$

$$U = \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}}{\sqrt{\frac{1}{n}\sum_{t=1}^{n}{y_t}^2} + \sqrt{\frac{1}{n}\sum_{t=1}^{n}{\hat{y}_t}^2}}. \quad (24)$$

Về quy tắc đánh giá hiệu năng, các chỉ số sai số định lượng bao gồm MAE, RMSE, MAPE và NRMSE phản ánh mức độ lệch giữa giá trị thực tế y_t và giá trị dự báo $\hat{y}_t$. Do đó, mô hình có các giá trị sai số này càng nhỏ sẽ tốt hơn (chẳng hạn, mô hình có chỉ số MAE nhỏ hơn cho thấy sai số tuyệt đối thấp hơn và độ chính xác cao hơn). Hệ số xác định R^2 có giá trị nằm trong khoảng $[0, 1]$, giá trị càng gần 1 biểu thị mô hình giải thích biến động dữ liệu thực tế càng hiệu quả. Đối với chỉ số bất định U-Theil (nằm trong khoảng $[0, 1]$), giá trị càng tiến gần về 0 thể hiện năng lực dự báo càng tối ưu.

Kiểm định Diebold–Mariano (Diebold & Mariano, 1995) được sử dụng để đánh giá liệu sự khác biệt về độ chính xác dự báo giữa hai mô hình có ý nghĩa thống kê hay không. Gọi $e_{1,t}$ và $e_{2,t}$ là sai số dự báo của mô hình 1 và 2 tại thời điểm t. Chuỗi sai biệt hàm mất mát bình phương được xác định là $d_t = e_{1,t}^2 - e_{2,t}^2$. Kiểm định đặt ra giả thiết $H_0: E(d_t) = 0$, tức hai mô hình có độ chính xác dự báo tương đương nhau; và đối thiết $H_1: E(d_t) \neq 0$, tức độ chính xác dự báo giữa hai mô hình có sự khác biệt. Thống kê DM được định nghĩa:

$$DM = \frac{\bar{d}}{\sqrt{\frac{\hat{V}}{n_T}}}. \quad (25)$$

Trong đó: $\bar{d}$ là trung bình của d_t, $\hat{V}$ là ước lượng phương sai dài hạn theo phương pháp HAC (Heteroskedasticity and Autocorrelation Consistent) và n_T là số quan sát tập kiểm tra.

Với mức ý nghĩa $\alpha = 0{,}05$, nếu p-value $\ge 0{,}05$, chưa đủ bằng chứng để bác bỏ H_0; do đó, hai mô hình được xem là chưa có khác biệt có ý nghĩa thống kê về độ chính xác dự báo. Ngược lại, nếu p-value $< 0{,}05$ thì H_0 bị bác bỏ, cho thấy sai số dự báo kỳ vọng giữa hai mô hình khác nhau có ý nghĩa thống kê. Khi đó, dấu của thống kê DM được sử dụng để xác định mô hình có sai số nhỏ hơn: $DM < 0$ cho thấy mô hình 1 có sai số bình phương kỳ vọng nhỏ hơn mô hình 2, trong khi $DM > 0$ cho thấy mô hình 2 có sai số bình phương kỳ vọng nhỏ hơn mô hình 1.

Trong nghiên cứu này, khi so sánh Prophet+LSTM với từng mô hình đối chứng, Prophet+LSTM được quy ước là mô hình 1 và mô hình còn lại là mô hình 2. Theo quy ước này, nếu p-value $< 0{,}05$ và $DM < 0$, có thể kết luận Prophet+LSTM có độ chính xác dự báo tốt hơn mô hình đối chứng trong cặp so sánh tương ứng. Ngược lại, nếu p-value $\ge 0{,}05$ thì chưa đủ cơ sở thống kê để kết luận hai mô hình khác nhau về độ chính xác dự báo, ngay cả khi các chỉ số sai số như MAE hoặc RMSE có chênh lệch về mặt số học.

Ngoài ra, phần dư của mô hình tốt nhất được chẩn đoán bằng: (i) kiểm định Ljung–Box ở các lag 6, 12, 18 với H_0: phần dư là nhiễu trắng; (ii) kiểm định Shapiro–Wilk và Jarque–Bera với H_0: phần dư có phân phối chuẩn. Ngoài ra, tính dừng của chuỗi gốc được đánh giá bằng kiểm định ADF với H_0: chuỗi không dừng và kiểm định KPSS với H_0: chuỗi dừng.

2.5. Quy trình phân tích dữ liệu

Bộ dữ liệu 276 quan sát được chia theo tỷ lệ 80/20: 220 quan sát huấn luyện (01/2002 – 04/2020) và 56 quan sát kiểm tra (05/2020 – 12/2024). Đối với SARIMA, Holt–Winters và Prophet: sử dụng trực tiếp chuỗi gốc. Đối với LSTM: chuẩn hóa Z-score, lookback = 12 cho LSTM đơn lẻ và lookback = 24 cho LSTM trên phần dư. Toàn bộ quá trình được thực hiện trên Python 3.10 với TensorFlow 2.21, Keras, statsmodels và Prophet. Hạt giống ngẫu nhiên = 9.999 cho mô hình nền; hạt giống ensemble = (101, 202) cho LSTM phần dư. Các siêu tham số chi tiết của các mô hình được trình bày ở Bảng 3.

Bảng 3. Siêu tham số chính của các mô hình

Mô hình	Tham số	Giá trị
Holt–Winters	Dạng / chu kỳ	Cộng (additive)/ $m = 12$
SARIMA	Bậc / Mùa	$(p,d,q)(P,D,Q)_m$ chọn theo AIC/ $m = 12$
Prophet	yearly_seasonality / cps / sps / n_changepoints	10 (Fourier) / 0,05 / 10,0 / 25
LSTM đơn lẻ	Lookback / units / dropout / dense / epochs / batch	12 / 64 / 0,2 / 32 (ReLU) / 40 / 16
LSTM phần dư	Lookback / units / dropout / epochs / batch	24 / 32 / 0,2 / 30 / 16
Ensemble	Số hạt giống / phương pháp tổng hợp	2 (101, 202) / trung bình
Đặc trưng đầu vào	LSTM phần dư	[phần dư chuẩn hóa, $\sin(2\pi m/12), \cos(2\pi m/12)$]

Việc lựa chọn siêu tham số được thực hiện dựa trên đặc điểm mùa vụ của chuỗi lượng mưa tháng, kích thước mẫu hiện có và mục tiêu cân bằng giữa độ chính xác dự báo với khả năng tổng quát hóa của mô hình. Đối với các mô hình thống kê, chu kỳ mùa được xác định là $m = 12$, phù hợp với cấu trúc mùa vụ theo năm của dữ liệu lượng mưa tháng; riêng mô hình SARIMA được lựa chọn bậc theo tiêu chí AIC nhằm hạn chế tính chủ quan trong xác định cấu trúc mô hình. Đối với LSTM đơn lẻ, lookback = 12 được sử dụng để mô hình tiếp nhận thông tin của một chu kỳ mùa đầy đủ. Trong khi đó, đối với LSTM trên phần dư, lookback = 24 được lựa chọn nhằm khai thác thông tin từ hai chu kỳ mùa liên tiếp sau khi thành phần xu hướng và mùa vụ chính đã được mô hình nền loại bỏ. Lựa chọn này giúp LSTM phần dư nhận diện tốt hơn các cấu trúc phi tuyến còn lại giữa các năm, đồng thời vẫn bảo toàn số lượng mẫu huấn luyện đủ lớn. Nếu lookback = 12 được sử dụng thì mô hình có thể chưa khai thác đầy đủ quan hệ liên mùa giữa các năm; ngược lại, nếu lookback = 36 được sử dụng thì làm giảm đáng kể số mẫu huấn luyện hiệu quả trong bối cảnh chuỗi chỉ gồm 276 quan sát, từ đó có thể làm tăng nguy cơ kém ổn định hoặc quá khớp.

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Kiểm định sơ bộ và đặc trưng dữ liệu

Với mức ý nghĩa 5%, kiểm định ADF cho chuỗi trả về $p = 0{,}0003$, bác bỏ giả thiết không dừng; KPSS cho $p = 0{,}10$ không bác bỏ giả thiết dừng. Hai kiểm định nhất quán xác nhận chuỗi đã dừng quanh trung bình, do đó không cần sai phân thường. Kiểm định Ljung–Box (lag 12) với $p < 0{,}0001$ cho thấy chuỗi có tự tương quan mạnh, đây là tín hiệu cấu trúc mùa vụ rõ rệt, phù hợp với việc sử dụng mô hình mùa vụ.

Sau khi điền giá trị thiếu, chuỗi đầy đủ gồm 276 quan sát (23 quan sát/tháng × 12 tháng). Bảng 4 trình bày thống kê mô tả lượng mưa theo tháng sau khi điền missing, có thể so sánh trực tiếp với Bảng 1 để thấy mức độ thay đổi nhỏ ở các tháng có giá trị thiếu (tháng 1, 2, 12).

Bảng 4. Thống kê mô tả lượng mưa theo tháng tại Đà Lạt sau khi điền giá trị thiếu

Tháng	N	Mean	SD	Min	Max
1	23	13,72	12,26	0,1	36,97
2	23	28,42	31,82	0,2	115,1
3	23	73,73	48,66	4,9	153,7
4	23	169,21	78,63	39,5	338,5
5	23	232,03	83,62	108,2	410,0
6	23	207,00	102,90	55,6	432,8
7	23	239,67	75,44	106,7	450,1
8	23	249,69	95,18	83,0	530,0
9	23	305,15	89,08	120,0	498,7
10	23	236,79	79,69	42,0	377,7
11	23	111,95	61,49	5,0	231,0
12	23	41,24	41,34	0,9	173,9

Việc phân rã chuỗi thời gian bằng phương pháp cộng tính cho thấy chuỗi gồm bốn thành phần chính: Observed (chuỗi gốc), Trend (xu hướng), Seasonal (mùa vụ) và Random (ngẫu nhiên) (Hình 4). Thành phần mùa vụ thể hiện chu kỳ 12 tháng rất rõ rệt với

biên độ ổn định qua các năm. Thành phần xu hướng cho thấy mức tăng nhẹ trong giai đoạn nghiên cứu, phù hợp với các phân tích về biến đổi khí hậu khu vực Tây Nguyên. Thành phần ngẫu nhiên có biên độ tương đối nhỏ ở các tháng khô và lớn hơn ở các tháng mưa cực đại, phản ánh tính ngẫu nhiên của các sự kiện mưa cường độ cao.

3.2. So sánh hiệu năng giữa các mô hình dự báo

Bảng 5 trình bày sáu chỉ số đánh giá hiệu năng của bảy mô hình trên tập kiểm tra gồm 56 quan sát (05/2020 – 12/2024). Hai cấu trúc lai dẫn đầu, Prophet+LSTM và HW+LSTM, cấu hình LSTM hoàn toàn giống nhau (ensemble hai hạt giống + đặc trưng sin/cos tháng + lookback 24) được sử dụng cho phép so sánh công bằng.

Khi xét theo tiêu chí sai số tuyệt đối, mô hình lai Prophet+LSTM đạt mức MAE thấp nhất (61,84 mm). Khi so với Prophet đơn lẻ, cấu trúc lai này giúp giảm nhẹ 0,53% sai số so với Prophet đơn lẻ (62,17 mm) và chỉ số MAPE (75,47%) tối ưu hơn so với các cấu trúc lai còn lại. Ngược lại, cấu trúc HW+LSTM đạt mức RMSE thấp nhất (81,99 mm) và chỉ số NRMSE% tốt nhất (18,41%). Tuy nhiên, với giá trị NRMSE% của cả hai mô hình lai đều nằm trong khoảng 18% (dưới ngưỡng 20%), chúng đều đạt cấp độ chính xác "Tốt" đối với một chuỗi dữ liệu khí tượng có cường độ biến động mạnh như lượng mưa trạm Đà Lạt. Thêm vào đó, hệ số xác định R^2 đạt mức 0,59 (HW+LSTM là 0,599 và Prophet+LSTM là 0,594) cùng với chỉ số Theil U ở mức thấp tương đương (0,197) đồng loạt xác nhận hai cấu trúc lai có khả năng giải thích phương sai và hiệu năng định lượng tương đương nhau.

Sự dịch chuyển nhẹ giữa các chỉ số xuất phát từ bản chất toán học: HW+LSTM kiểm soát tốt hơn các sai số lớn tại các điểm mưa cực đoan do cơ chế bình phương sai số của RMSE/NRMSE, trong khi Prophet+LSTM ổn định hơn trên toàn bộ chuỗi để tối ưu MAE/MAPE. Ngoài ra, kết quả kiểm định Diebold–Mariano (DM) tại mục 3.3 tiếp theo cũng xác nhận sự chênh lệch sai số giữa hai mô hình lai là không có ý nghĩa thống kê ($p > 0,05$).

Tuy nhiên, đối với mục tiêu dự báo mở rộng cho tương lai, cấu trúc Prophet+LSTM vẫn mang lại tính thuyết phục cao hơn để làm mô hình nòng cốt nhờ hai ưu thế kỹ thuật vượt trội: (i) tiếp cận phân rã Bayesian của Prophet cho phép tự động thích ứng linh hoạt với các điểm thay đổi xu hướng bất thường (changepoints) thay vì giả định tuyến tính máy móc của HW và (ii) cơ chế hậu nghiệm MCMC từ dữ liệu quá khứ giúp Prophet+LSTM cung cấp khoảng tin cậy có sự phân hóa mùa vụ thực tế, mức khoảng tin cậy hẹp ở mùa khô thay vì dải sai số tịnh tiến mở rộng theo thời gian của HW.

Đáng chú ý, mô hình LSTM đơn lẻ ghi nhận sai số cao nhất (MAE = 72,27 mm; NRMSE = 20,72%) đi kèm chỉ số MAPE bùng nổ tới 100,98%. Hiện tượng này bắt nguồn từ việc LSTM đơn lẻ không có cơ chế mô hình hóa mùa vụ tường minh, trong khi 220 quan sát huấn luyện chưa đủ để mạng học cấu trúc chu kỳ 12 tháng từ chuỗi giá trị thô; hệ quả là mạng có xu hướng dự báo lệch về giá trị trung bình tổng thể (~159 mm) ở các tháng khô, tạo sai số tuyệt đối 30 – 60 mm tại các quan sát có lượng mưa thực tế dưới 1 mm. Khi mẫu số trong công thức MAPE tiệm cận 0 thì chỉ một quan sát đơn lẻ cũng đủ tạo phần trăm sai số hàng nghìn %, đẩy MAPE trung bình toàn chuỗi vượt ngưỡng 100%. Hiệu ứng này được giảm thiểu rõ rệt ở các cấu trúc lai do thành phần thống kê (Holt–Winters, SARIMA, Prophet) đã hấp thụ phần lớn cấu trúc mùa vụ, đồng thời đặc trưng $\sin(2\pi m/12)$ và $\cos(2\pi m/12)$ bổ sung ở đầu vào LSTM giúp mạng nhận diện chính xác pha mùa khô, đưa MAPE về khoảng 75 – 80%. Tuy vậy, MAPE của các mô hình nhìn chung vẫn duy trì ở mức cao (> 70%), đặc điểm cố hữu của chuỗi mưa nhiệt đới có nhiều giá trị mùa khô tiệm cận 0, trong khi sai số tuyệt đối toàn cục ở mức 62 mm phản ánh giới hạn chung của dự báo đơn biến, vấn đề được thảo luận sâu hơn ở phần Hạn chế. Hình 5 minh họa các đường dự báo của bảy mô hình nghiên cứu so với giá trị thực tế trên tập kiểm tra.

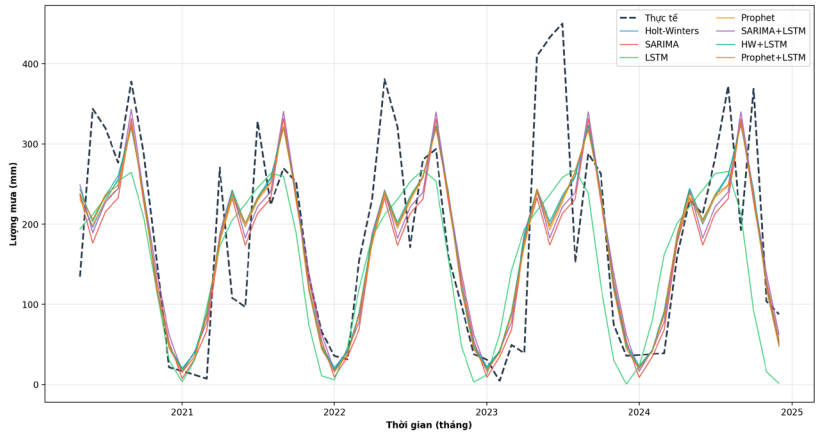
Bảng 5. Kết quả đánh giá hiệu năng các mô hình trên tập kiểm tra

Hạng	Mô hình	MAE	RMSE	MAPE%	NRMSE%	R^2	Theil U
1	**Prophet+LSTM ★**	**61,84**	**82,53**	**75,47**	**18,53**	**0,594**	**0,197**
2	HW + LSTM	61,91	81,99	78,45	18,41	0,599	0,198
3	Prophet	62,17	83,29	74,52	18,70	0,587	0,199
4	Holt-Winters	62,23	82,68	77,39	18,57	0,593	0,197
5	SARIMA + LSTM	65,27	86,26	79,21	19,37	0,557	0,205
6	SARIMA	65,68	87,95	71,21	19,75	0,539	0,213
7	LSTM	72,27	92,24	100,98	20,72	0,493	0,228

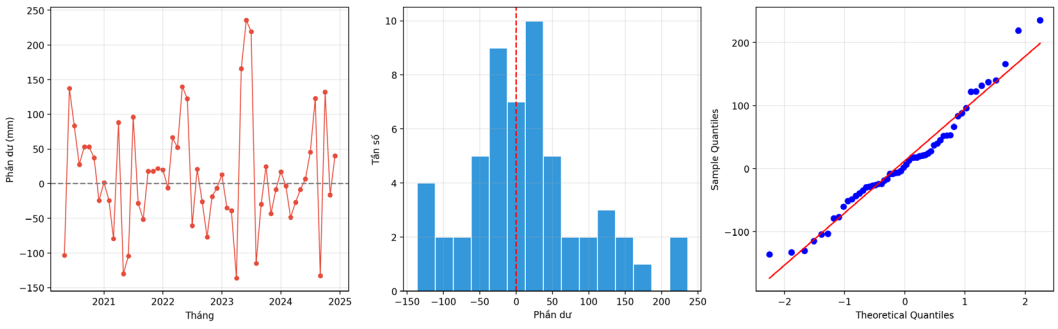
Ghi chú: ★ mô hình tốt nhất theo MAE. Kết quả được sắp xếp theo MAE tăng dần.



Hình 4. Phân rã chuỗi lượng mưa Đà Lạt 2002 – 2024 theo mô hình cộng tính



Hình 5. Dự báo của bảy mô hình so với giá trị thực tế trên tập kiểm tra



(a) (b) (c)

Hình 6. Chẩn đoán phần dư mô hình Prophet+LSTM

Ghi chú: (a) phần dư theo thời gian, (b) phân phối phần dư và (c) Q-Q plot.

3.3. Kiểm định Diebold–Mariano

Kiểm định DM được sử dụng để xác định mức ý nghĩa thống kê của khác biệt sai số giữa Prophet+LSTM và sáu mô hình còn lại (Bảng 6).

Bảng 6. Kết quả kiểm định Diebold–Mariano so với Prophet+LSTM

Mô hình	DM	p-value	Ý nghĩa
Holt-Winters	−0,259	0,7957	ns
SARIMA	−2,381	0,0172	** Prophet+ LSTM tốt hơn
LSTM	−1,303	0,1924	ns
Prophet	−1,444	0,1487	ns
SARIMA + LSTM	−2,430	0,0151	** Prophet+ LSTM tốt hơn
HW+ LSTM	+0,850	0,3953	ns

*Ghi chú: ** $p < 0,05$; * $p < 0,1$; ns: không có ý nghĩa thống kê.*

Mô hình Prophet+LSTM vượt trội có ý nghĩa thống kê ($p < 0,05$) so với SARIMA và SARIMA+LSTM. Đối với Prophet đơn lẻ, Holt-Winters và HW+LSTM, chênh lệch sai số là không có ý nghĩa thống kê, phù hợp với chênh lệch tuyệt đối nhỏ ($< 0,5$ mm) ghi nhận ở Bảng 5. Điều này phản ánh giới hạn lý thuyết của nhóm mô hình mùa vụ tuyến tính trên chuỗi mưa nhiệt đới có mùa vụ rất rõ rệt: phần dư sau khi loại bỏ thành phần mùa và xu hướng gần với nhiễu trắng, khiến lớp học sâu chỉ có thể đóng góp cải tiến biên.

3.4. Kiểm tra phần dư của mô hình

Kết quả chẩn đoán phần dư của mô hình Prophet+LSTM được trình bày trong Bảng 7 và Hình 6. Các kiểm định Ljung–Box tại nhiều độ trễ đều không bác bỏ giả thuyết phần dư là nhiễu trắng ($p > 0,05$), trong khi các kiểm định Shapiro–Wilk và Jarque–Bera cho thấy không có bằng chứng để bác bỏ giả thiết phần dư tuân theo phân phối chuẩn. Việc quan sát trực quan trên đồ thị phần dư theo thời gian, histogram và Q–Q plot cũng cho thấy phần dư dao động ngẫu nhiên quanh giá trị 0, không còn xu hướng hay cấu trúc mùa vụ rõ rệt. Những kết quả này chứng tỏ mô hình Prophet+LSTM đã mô tả tốt động thái của chuỗi lượng mưa, đồng thời trích xuất hiệu quả cả thành phần tuyến tính và phi tuyến, khiến phần thông tin còn lại chủ yếu mang tính ngẫu nhiên và khó dự báo.

Bảng 7. Kết quả chẩn đoán phần dư mô hình Prophet+LSTM

Kiểm định	Thống kê	p-value	Kết luận
Ljung-Box (lag 6)	1,262	0,9737	Phần dư là nhiễu trắng
Ljung-Box (lag 12)	9,129	0,6918	Phần dư là nhiễu trắng
Ljung-Box (lag 18)	20,989	0,2799	Phần dư là nhiễu trắng
Shapiro-Wilk	0,967	0,1241	Phần dư có phân phối chuẩn
Jarque-Bera	2,860	0,2394	Phần dư có phân phối chuẩn

3.5. Dự báo lượng mưa năm 2025

Mô hình Prophet+LSTM được huấn luyện lại trên toàn bộ 276 quan sát và dự báo lượng mưa 12 tháng năm 2025. Khoảng dự báo xấp xỉ 95% được ước lượng từ độ lệch chuẩn phần dư huấn luyện, với cận dưới chặn tại 0 (do lượng mưa không âm). Kết quả thể hiện ở Bảng 8 và Hình 7.

Bảng 8. Dự báo lượng mưa tháng tại trạm Đà Lạt năm 2025 (mm)

Tháng	Dự báo điểm	KTC dưới 95%	KTC trên 95%
01/2025	25,90	0,00	161,46
02/2025	42,53	0,00	178,09
03/2025	89,69	0,00	225,24
04/2025	185,08	49,53	320,64
05/2025	239,29	103,74	374,85
06/2025	215,17	79,61	350,72
07/2025	246,81	111,26	382,37
08/2025	255,80	120,24	391,35
09/2025	311,41	175,85	446,96
10/2025	243,42	107,87	378,97
11/2025	120,85	0,00	256,40
12/2025	52,52	0,00	188,07

Việc dự báo bám sát quy luật mùa vụ tự nhiên của khu vực: cực tiểu vào các tháng mùa khô (tháng 1: 25,9 mm), tăng dần vào mùa mưa với đỉnh đạt 311,41 mm ở tháng 9 (chu kỳ thông thường của khu vực cao nguyên Tây Nguyên). Biên độ khoảng tin cậy phản ánh độ không chắc chắn của dự báo (~136 mm hai phía), phù hợp với độ biến động lịch sử của chuỗi.

3.6. Thảo luận

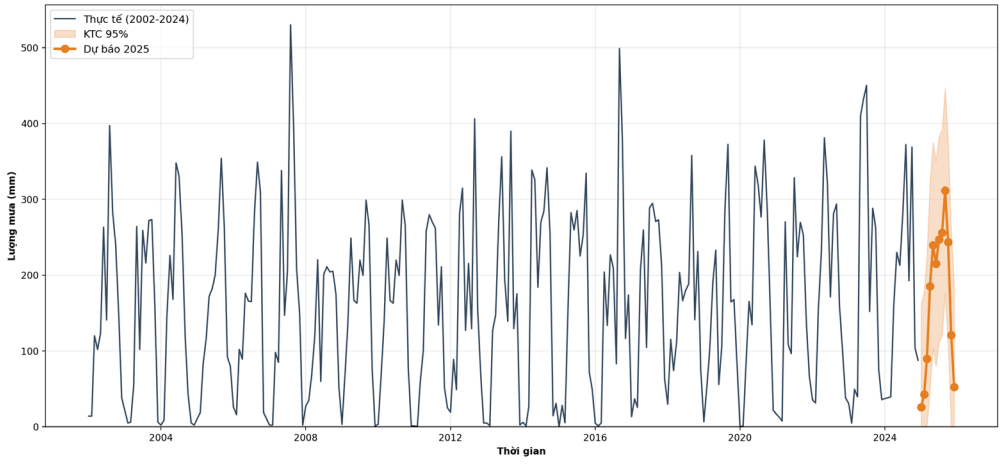
Kết quả thu được từ nghiên cứu này phù hợp với các nghiên cứu trước đó. Huynh et al. (2024) đạt

MAE ~ 50 – 80 mm với SARIMA ở Đồng bằng sông Cửu Long, trong khi Wu et al. (2021) báo cáo cải tiến 5 – 15% của mô hình lai Wavelet–ARIMA–LSTM so với ARIMA đơn lẻ. Trong nghiên cứu này, mô hình lai mang lại cải tiến biên nhỏ (~0,5% MAE), khác biệt được giải thích bởi chu kỳ mùa rất rõ rệt và biên độ cực đoan của mưa Đà Lạt khiến phần dư sau khi tách mùa và xu hướng có ít cấu trúc phi tuyến để LSTM khai thác (xác nhận bởi Ljung–Box $p > 0{,}28$ trên phần dư).

Đóng góp phương pháp luận của nghiên cứu nằm ở: (i) khung so sánh trực diện ba nhóm mô hình với sáu chỉ số sai số và kiểm định DM, làm rõ giới hạn cải thiện thực tế, (ii) đề xuất cấu trúc LSTM phần dư nâng cấp bằng ensemble với đặc trưng mùa, áp dụng đồng nhất cho ba mô hình lai để so sánh công bằng.

3.7. Hạn chế của nghiên cứu

Nghiên cứu còn một số hạn chế. Thứ nhất, dữ liệu đơn biến chưa kết hợp các biến ngoại sinh quan trọng (nhiệt độ, độ ẩm, vận tốc gió, áp suất khí quyển, ENSO/SOI) ảnh hưởng đến thời tiết ở Đà Lạt. Thứ hai, sai số tuyệt đối (MAE ≈ 62 mm) so với lượng mưa trung bình ~159 mm vẫn còn cao, phản ánh giới hạn lý thuyết của dự báo lượng mưa đơn biến. Thứ ba, kết quả từ một trạm quan trắc duy nhất chưa thể tổng quát hóa cho toàn bộ khu vực cao nguyên Lâm Viên. Thứ tư, sự khác biệt giữa Prophet+LSTM và các mô hình thân cận (HW+LSTM, Prophet, Holt-Winters) không có ý nghĩa thống kê, việc lựa chọn mô hình nên cân nhắc thêm các yếu tố vận hành như tính diễn giải và chi phí tính toán.



Hình 7. Dự báo lượng mưa trạm Đà Lạt năm 2025 (Prophet+LSTM) kèm khoảng tin cậy 95%

4. KẾT LUẬN

Nghiên cứu được thực hiện đã giúp so sánh hệ thống bảy cấu trúc mô hình dự báo lượng mưa tháng tại trạm Đà Lạt (276 quan sát, 2002 – 2024), gồm bốn mô hình đơn lẻ (Holt–Winters, SARIMA, Prophet, LSTM) và ba mô hình lai phân tách phần dư với LSTM nâng cấp (ensemble 2 hạt giống, đặc trưng mùa sin/cos, lookback 24). Việc đánh giá đa chỉ số (MAE, RMSE, MAPE, NRMSE, R^2, Theil U) kết hợp kiểm định Diebold–Mariano và chẩn đoán phần dư cho thấy mô hình lai Prophet+LSTM đạt sai số MAE thấp nhất (61,84 mm), tốt hơn có ý nghĩa thống kê so với SARIMA và SARIMA+LSTM ($p < 0{,}05$). Phần dư của Prophet+LSTM có phân phối chuẩn và gần với nhiễu trắng, khẳng định mức độ phù hợp của mô hình.

Dự báo cho 12 tháng năm 2025 bám sát quy luật mùa vụ tự nhiên với đỉnh mưa tháng 9 đạt 311,41 mm và cực tiểu tháng 1 đạt 25,90 mm, cung cấp cơ sở khoa học cho công tác quản lý tài nguyên nước và phòng chống thiên tai tại Đà Lạt – Lâm Đồng.

Các hướng nghiên cứu có thể phát triển tiếp theo bao gồm: (i) tích hợp biến ngoại sinh (nhiệt độ, độ ẩm, ENSO/SOI) qua SARIMAX hoặc LSTM đa đầu vào; (ii) thử nghiệm Transformer và cơ chế Attention và (iii) sử dụng hàm mất mát Quantile để xử lý các sự kiện cực đoan.

LỜI CẢM TẠ

Đề tài này được tài trợ bởi Đại học Cần Thơ, Mã số: TSV2026-167.

TÀI LIỆU THAM KHẢO

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). Wiley.

Can, L. T., & Lộc, T. P. (2024). Dự báo chuỗi thời gian với một số mô hình học máy và ứng dụng. *Tạp chí Khoa học Đại học Cần Thơ*, *60*(Số chuyên đề: Khoa học tự nhiên), 32–39. https://doi.org/10.22144/ctujos.2024.341

Cong, T., Ha, T.-H.-G., Vu, G.-L., Bui, H.-N., Dao, N.-Q.-H., Le, M.-H., Dinh, C.-M., Cong, M.-P., & Doan, Q.-V. (2025). Advancing subseasonal extreme rainfall forecasting in Vietnam using machine learning. *SOLA*, *21*, 273–282. https://doi.org/10.2151/sola.2025-034

Cục Quản lý đê điều và phòng, chống thiên tai. (2024). *Công bố số liệu thiệt hại do thiên tai gây ra trong phạm vi cả nước năm 2024*. http://phongchongthientai.mard.gov.vn/Pages/cong-bo-so-lieu-thiet-hai-do-thien-tai-gay-ra-trong-pham-vi-ca-nuoc-nam-2024.aspx

Cục Thống kê. (2025). *Cơ sở dữ liệu thống kê – Chỉ tiêu lượng mưa trung bình tháng tại trạm khí tượng Đà Lạt (mã dữ liệu: V0109).* https://www.nso.gov.vn/

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263. https://doi.org/10.1080/07350015.1995.10524599

Dimri, T., Ahmad, S., & Sharif, M. (2020). Time series analysis of climate variables using seasonal ARIMA approach. *Journal of Earth System Science*, *129*, Article 149. https://doi.org/10.1007/s12040-020-01408-x

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, *20*(1), 5–10. https://doi.org/10.1016/j.ijforecast.2003.09.015

IPCC. (2023). Summary for Policymakers. In H. Lee, & J. Romero (Eds.), *Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 1–34). IPCC.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005–6022. https://doi.org/10.5194/hess-22-6005-2018

Huynh, M. V. T., Tran, T. V., Nguyen, N. D. G., Bui, T. B. L., Nguyen, T. T., Ngo, P. C., Meraj, G., Kumar, P., Le, V. T., Doan, V. D., Nguyen, V. T., Downes, N. K., Bhuyan, M. S., Kanga, S., & Singh, S. K. (2024). Modelling and predicting annual rainfall over the Vietnamese Mekong Delta (VMD) using SARIMA. *Discover Geoscience*, *2*(1), 19. https://doi.org/10.1007/s44288-024-00018-0

Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: Time series missing value imputation in R. *The R Journal*, *9*(1), 207–218.

https://doi.org/10.32614/RJ-2017-009Nguyen, H. D., Pham Van, C., Nguyen, Q.-H., & Bui, Q.-T. (2023). Daily streamflow prediction based on the long short-term memory algorithm: A case study in the Vietnamese Mekong Delta. *Journal of Water and Climate Change*, *14*(4), 1247–1267. https://doi.org/10.2166/wcc.2023.419

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, *72*(1), 37–45. https://doi.org/10.1080/00031305.2017.1380080

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, *6*(3), 324–342. https://doi.org/10.1287/mnsc.6.3.324

Wu, X., Zhou, J., Yu, H., Liu, D., Xie, K., Chen, Y., Hu, J., Sun, H., & Xing, F. (2021). The development of a hybrid Wavelet-ARIMA-LSTM model for precipitation amounts and drought analysis. *Atmosphere*, *12*(1), 74. https://doi.org/10.3390/atmos12010074

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159–175. https://doi.org/10.1016/S0925-2312(01)00702-0