DỰ BÁO MỰC NƯỚC TRÊN SÔNG KIẾN GIANG SỬ DỤNG PHƯƠNG PHÁP HỒI QUY

Đinh Nhật Quang[1], Tạ Quang Chiểu[1], Đào Thị Huệ[1], Nguyễn Thị Kim Ngân[1]

Tóm tắt: *Mô hình dự báo sự thay đổi của mực nước sông gần đây được sử dụng như một công cụ hỗ trợ cho các nhà quản lý trong việc đề xuất các giải pháp thích ứng và giảm nhẹ rủi ro thiên tai do lũ. Các mô hình định hướng dữ liệu sử dụng phương pháp học máy đã trở thành một cách tiếp cận hấp dẫn và hiệu quả để mô phỏng và dự báo biến động mực nước sông. Trong nghiên cứu này, các mô hình dựa trên phương pháp hồi quy tuyến tính (LR), Random Forest Regression (RFR) và Light Gradient Boosting Machine Regression (LGBMR) được xây dựng để dự đoán mực nước hàng ngày trên sông Kiến Giang dựa trên bộ dữ liệu thu thập từ năm 1977 đến năm 2020. Các chỉ số thống kê R^2, NSE, MAE và RMSE được tính toán để kiểm tra độ tin cậy của ba mô hình đề xuất. Kết quả nghiên cứu chỉ ra hiệu quả của các thuật toán hồi quy trong việc dự báo mực nước lũ, đặc biệt là phương pháp hồi quy tuyến tính với các chỉ số R^2, NSE, MAE và RMSE lần lượt là 0,959; 0,958; 6,67 cm và 12,2 cm.*

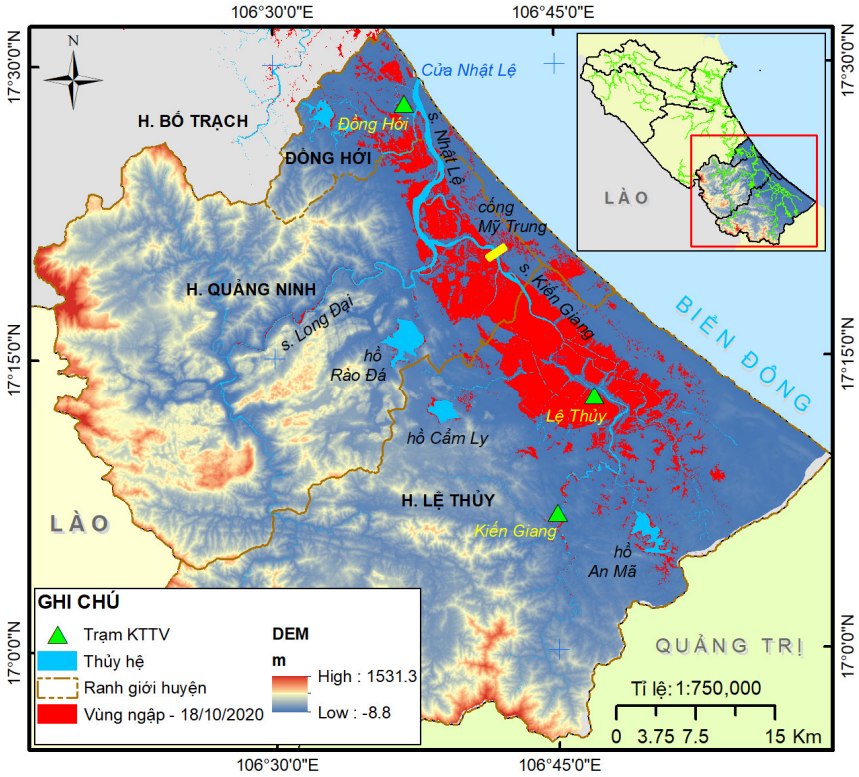
Từ khoá: Dự báo mực nước, học máy, phương pháp hồi quy, sông Kiến Giang.

1. ĐẶT VẤN ĐỀ

Là một trong hai phụ lưu lớn của sông Nhật Lệ, sông Kiến Giang chảy qua huyện Lệ Thủy và Quảng Ninh, tỉnh Quảng Bình với chiều dài 69 km (Hình 1) (Nguyễn Đức Lý và nnk, 2013). Hàng trăm năm qua, kể từ khi hình thành vùng đất Lệ Thủy và Quảng Ninh, nơi đây được coi là "rốn lũ" của tỉnh Quảng Bình. Đặc biệt, đợt mưa lũ lịch sử trong tháng 10 năm 2020 đã nhấn chìm vùng đồng bằng châu thổ nhỏ hẹp dưới chân dãy Trường Sơn với trên 50.000 nhà dân bị ngập sâu và hàng chục thôn, bản bị cô lập. Trong trận lũ đặc biệt lớn từ ngày 16 đến 22 tháng 10 năm 2020, đỉnh lũ trên sông Kiến Giang tại trạm Lệ Thủy lên tới 4,88 m, trên mức báo động III và vượt 0,97 m so với đỉnh lũ lịch sử năm 1979.

Dự báo mực nước sông chính xác là một thành phần quyết định trong hệ thống cảnh báo lũ sớm và đóng vai trò quan trọng trong việc giảm nhẹ thiên tai do lũ. Nhìn chung, có hai cách tiếp cận chính được sử dụng để thiết lập các mô hình dự báo mực nước trên sông. Cách tiếp cận đầu tiên chủ yếu phụ thuộc vào các mô hình dựa trên tính chất vật lý (*physically-based models*), chẳng hạn như bộ mô hình MIKE HYDRO River, HEC-HMS, SOBEK, EFDC, v.v. Các mô hình dựa trên tính chất vật lý có độ chính xác cao trong việc dự đoán mực nước, tuy nhiên chúng thường yêu cầu một lượng dữ liệu đầu vào rất lớn, bao gồm dữ liệu địa hình, khí tượng, thủy văn, hải văn, v.v. và đòi hỏi nhiều thời gian mô phỏng. Do đó, các mô hình này không phù hợp cho việc dự báo trong thời gian ngắn hoặc theo thời gian thực, gần thực. Hơn nữa, việc phát triển các mô hình dựa trên tính chất vật lý thường yêu cầu người dùng phải có kiến thức chuyên sâu và kiến thức chuyên môn liên quan đến các thông số thủy văn và thủy lực (Atashi và nnk, 2022).

[1] *Trường Đại học Thủy lợi*



Hình 1. Hệ thống sông Kiến Giang và các trạm khí tượng, thủy văn trong khu vực

Một cách tiếp cận thay thế những hạn chế của mô hình truyền thống nêu trên là sử dụng các mô hình theo định hướng dữ liệu (*data-driven models*), dựa trên việc thu thập và phân tích mối quan hệ thống kê giữa các dữ liệu đầu vào và đầu ra. Mô hình học máy (*Machine Learning* - ML) đã được sử dụng để dự báo ngập từ những năm 1990 và là một trong những thư viện, nền tảng (*frameworks*) phổ biến nhất trong phương pháp định hướng dữ liệu. Các nghiên cứu gần đây đã chỉ ra rằng các mô hình học máy là công cụ tiềm năng trong việc dự báo mực nước do chúng có thể được xây dựng nhanh chóng, dễ dàng và không đòi hỏi phải có sự hiểu biết về các quá trình vật lý ẩn đằng sau. Ngoài ra, khả năng tính toán, hiệu chỉnh và kiểm định nhanh hơn so với các mô hình vật lý truyền thống và cách sử dụng ít phức tạp hơn là những ưu điểm lớn mà các mô hình học máy dựa vào số liệu mang lại (Mekanik và nnk, 2013).

Trong bài báo này, các tác giả đề xuất và phát triển các mô hình sử dụng các phương pháp hồi quy (*LR*, *RFR* và *LGBMR*) để dự báo mực nước lũ tại một trạm đại diện cho vùng đồng bằng sông Kiến Giang, đó là trạm Lệ Thủy.

2. PHƯƠNG PHÁP NGHIÊN CỨU VÀ SỐ LIỆU THU THẬP

2.1. Các phương pháp hồi quy

Hồi quy là phương pháp toán học trong thống kê để phân tích mối liên hệ giữa đại lượng cần dự báo theo thời gian thông qua số liệu thống kê được trong quá khứ. Trong nghiên cứu này, ba kỹ thuật hồi quy của học máy đã được áp dụng để xây dựng các mô hình định hướng dữ liệu. Quá trình chính khi xây dựng các mô hình này được gọi là "*giai đoạn học hỏi*", trong đó mối quan hệ giữa các biến đầu vào và đầu ra của hệ thống được xây dựng (Guo và nnk, 2021):

$$y = f(x) \quad (1)$$

với các dữ liệu có sẵn:

$[(x_1, y_1),(x_2, y_2),...(x_n, y_n)]= [[x]_i, y_i]_{i=1}^{n}$ (2)

trong đó x là vectơ đầu vào, y là đầu ra mong muốn, n là số lượng dữ liệu và f là hàm hồi quy.

2.1.1. Hồi quy tuyến tính

Mô hình hồi quy là để xác định mối quan hệ giữa biến phụ thuộc y với một hay nhiều biến độc lập x; mối quan hệ giữa biến phụ thuộc và biến độc lập là tuyến tính. Mô hình hồi quy tuyến tính (*LR*) có dạng:

$$y = \alpha + \beta x \quad (3)$$

với α là chặn (*intercept*) và β là độ dốc (*slope*).

Xét tập dữ liệu gồm m phần tử $x_1, x_2, \ldots, x_m$ trong không gian n chiều (biến độc lập, thuộc tính), có giá trị tương ứng của biến phụ thuộc (cần dự báo) là $y_1, y_2, \ldots, y_m$. Các tham số α và β của mô hình được ước lượng từ bộ dữ liệu quan sát bằng phương pháp bình phương nhỏ nhất (*least squares*):

$$\text{Min}\left(\sum_{i=1}^{m}[[y_i - (\alpha + \beta x_i)]^2)\right]$$

Giá trị dự báo cho phần tử mới x dựa vào công thức (4):

$$\hat{y} = \alpha + \beta x \quad (4)$$

2.1.2. Random Forest Regression

Random Forest Regression (*RFR*) đề xuất bởi Breiman (2001) là một trong những phương pháp học có giám sát (*supervised learning*) sử dụng cho các bài toán phân loại và hồi quy. *RFR* là một phương pháp học tổng hợp, tập hợp kết quả từ các cây ra quyết định đơn lẻ, từ đó nâng cao hiệu quả dự báo thông qua hình thức biểu quyết đa số hay trung bình kết quả tùy theo từng bài toán cụ thể (Hình). Về bản chất *RFR* sử dụng kỹ thuật có tên gọi là *bagging* - kỹ thuật cho phép lựa chọn một nhóm nhỏ các thuộc tính tại mỗi nút của cây phân lớp để phân chia thành các mức tiếp theo. Do đó, *RFR* có khả năng phân chia không gian tìm kiếm rất lớn thành các không gian tìm kiếm nhỏ hơn, nhờ thế thuật toán có thể thực hiện việc phân loại một cách nhanh chóng và dễ dàng. Đối với bài toán hồi quy, kết quả cuối cùng của mô hình *RFR* sẽ là trung bình của tất cả các kết quả dự báo của các cây. Thuật toán *RFR* được tóm tắt như sau:

1. Trên cơ sở của phương pháp bootstrap, một tập hợp con các mẫu được tạo ngẫu nhiên với các mẫu thay thế từ tập dữ liệu ban đầu;

2. Các mẫu bootstrap này được sử dụng để xây dựng cây hồi quy (Hình). Tiêu chuẩn tối ưu được sử dụng để chia nút của cây hồi quy thành hai nút con. Thủ tục đệ quy được thực trên mỗi nút con cho đến khi kết thúc;

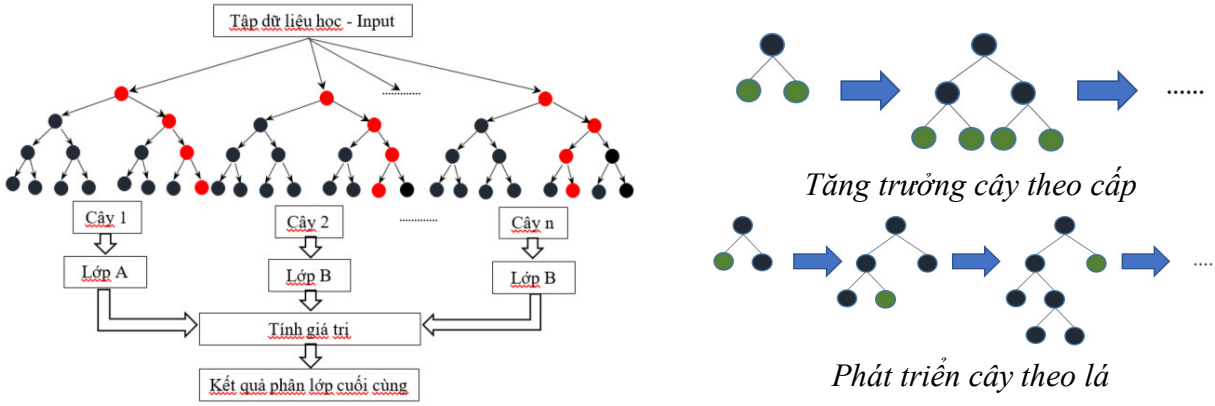
3. Mỗi cây hồi quy cung cấp kết quả dự đoán. Khi tất cả các cây hồi quy đã đạt đến kích thước tối đa, dự đoán cuối cùng được xác định là giá trị trung bình của các kết quả từ tất cả các cây hồi quy (Guo và nnk, 2021) như trong công thức (5):

$$f^{RFR}(x) = \frac{1}{tr}\sum_{tr=1}^{N_{trees}} \tilde{h}_{tr}(x) \quad (5)$$

trong đó tr là số cây, N_{tree} là kích thước tối đa của cây và $\tilde{h}_{tr}$ biểu thị dự đoán của mỗi cây hồi quy.

2.1.3. Light Gradient Boosting Machine Regression

Light Gradient Boosted Machine Regression (*LGBMR*) được phát triển dựa trên cây quyết định (*Decision Tree*). Thuật toán dựa trên biểu đồ histogram chia tách biến liên tục thành các nhóm khác nhau. Nó sử dụng phương pháp phát triển cây theo lá (*leaf-wise tree growth*) (Hình) thay vì phương pháp tăng trưởng cây theo cấp (*level-wise tree growth*, được sử dụng bởi hầu hết các phương pháp dựa trên cây quyết định khác) để tăng hiệu quả của mô hình, giảm mức sử dụng bộ nhớ và cải thiện thời gian tính toán (Guo và nnk, 2021).



Hình 2. Phương pháp hồi quy RFR (trái) và LGBMR (phải)

2.2. Các tiêu chí đánh giá độ tin cậy của các mô hình hồi quy

Để đánh giá mức độ dự báo chính xác của các mô hình hồi quy, bốn tiêu chí đánh giá sau được sử dụng:

1. Hệ số xác định (R^2)

$$R^2 = \left[\frac{\sum_{i=1}^{n}\left[\left(H_i^{mea} - \bar{H}^{mea}\right)\right]\left(H_i^{pre} - \bar{H}^{pre}\right)}{\sqrt{\sum_{i=1}^{n}\left(H_i^{mea} - \bar{H}^{mea}\right)^2 \sum_{i=1}^{n}\left(H_i^{pre} - \bar{H}^{pre}\right)^2}}\right]^2 \quad (5)$$

2. Hệ số hiệu quả (Nash-Sutcliffe efficiency - *NSE*)

$$NSE = 1 - \frac{\sum_{i=1}^{n}\left(H_i^{mea} - H_i^{pre}\right)^2}{\sum_{i=1}^{n}\left(H_i^{mea} - \bar{H}^{mea}\right)^2} \quad (6)$$

3. Sai số tuyệt đối trung bình (Mean Absolute Error - *MAE*)

$$MAE = \frac{\sum_{i=1}^{n}\left|H_i^{mea} - H_i^{pre}\right|}{n} \quad (7)$$

4. Lỗi trung bình bình phương gốc (Root Mean Square Error - *RMSE*)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(H_i^{mea} - H_i^{pre}\right)^2}{n}} \quad (8)$$

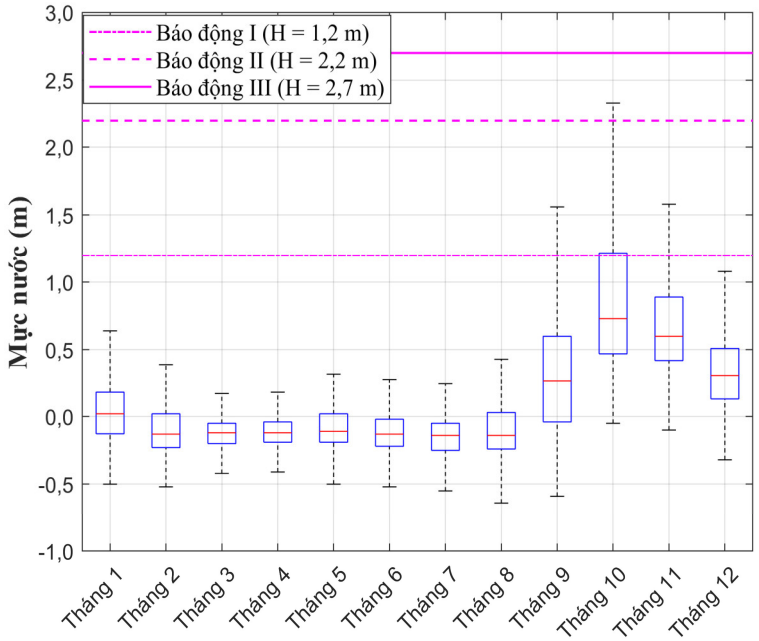
Trong đó: H_i^{mea} và H_i^{pre} là mực nước sông thực đo và dự đoán tại thời điểm i; $\bar{H}^{mea}$ và $\bar{H}^{pre}$ là giá trị trung bình của mực nước sông thực đo và dự đoán. Giá trị R^2 nằm trong khoảng từ 0 đến 1; trong đó giá trị 1 biểu thị các giá trị dự đoán bằng với giá trị thực đo. Giá trị của *NSE* nằm trong khoảng từ $-\infty$ đến 1; và càng gần 1 thì khả năng dự đoán của mô hình càng tốt. *MAE* và *RMSE* là hai chỉ số hiển thị sai số của mô hình với giá trị tối ưu bằng không.

2.3. Vùng nghiên cứu và số liệu thu thập

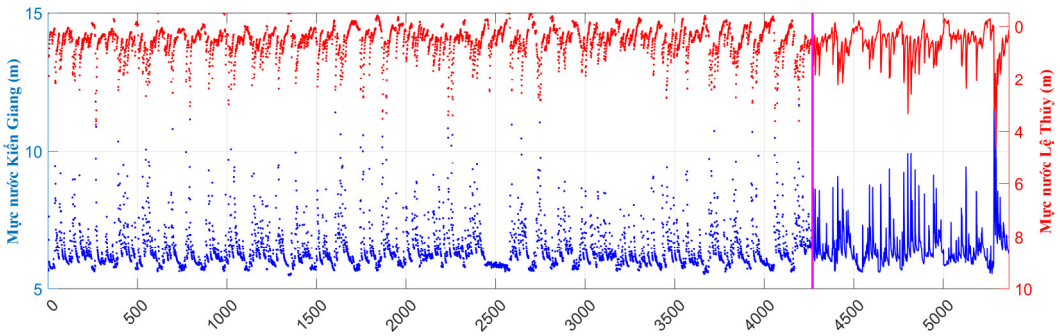
Lưu vực sông Nhật Lệ bao gồm 03 nhánh sông chính là Kiến Giang, Long Đại và Nhật Lệ với tổng diện tích của toàn lưu vực là 2.612 km^2 (Hình). Trên lưu vực sông Kiến Giang có một số hồ thủy lợi loại vừa là An Mã, Cẩm Ly và Rào Đá với dung tích phòng lũ nhỏ (chỉ khoảng 22,1; 6,9 và 11,6 triệu m^3), còn lại là các hồ chứa nhỏ nên tác động của hoạt động vận hành hồ chứa đến mực nước hạ lưu là không đáng kể. Trên lưu vực nghiên cứu hiện chỉ có 03 trạm đo mưa và mực nước (Kiến Giang, Lệ Thủy và Đồng Hới)

có dữ liệu đo đạc liên tục đến nay; trong đó số liệu theo giờ chỉ có trong một thời đoạn ngắn trong từng đợt lũ. Trong nghiên cứu này bộ số liệu quan trắc mưa và mực nước ngày tại 03 trạm trong giai đoạn 1977-2020 đã được thu thập. Biểu đồ hộp râu (*Box and Whisker plot*) trong Hình cho thấy sự tăng đáng kể mực nước sông Kiến Giang thường xảy ra từ tháng 9 đến tháng 12 - thời điểm mùa lũ trong khu vực nghiên cứu (Nguyễn Đức Lý và nnk, 2013).

Do nghiên cứu chỉ tập trung vào bài toán dự báo mực nước ngày trên sông trong mùa lũ nên 5.368 bộ số liệu mưa và mực nước ngày tại 03 trạm trong mùa lũ (tháng 9 - tháng 12) trong 44 năm đã được sử dụng. Để thiết lập và đánh giá các mô hình học máy, bộ dữ liệu trên được chia thành bộ dữ liệu huấn luyện và kiểm định, trong đó 80% (giai đoạn 1977–2011) dành cho huấn luyện (các chấm trong Hình 3) và 20% còn lại dùng để kiểm định (nét liền).



Hình 3. Biến động mực nước ngày tại trạm Lệ Thủy trong các tháng trong giai đoạn 1977-2020 và ba mức báo động trên sông



Hình 4. Số liệu mực nước thực đo trong mùa lũ tại trạm Kiến Giang (xanh) và Lệ Thủy (đỏ)

3. KẾT QUẢ VÀ THẢO LUẬN

Ba kỹ thuật học máy dựa trên phương pháp hồi quy được sử dụng để xây dựng mô hình định hướng dữ liệu nhằm dự báo mực nước tại trạm Lệ Thủy trên sông Kiến Giang. Bốn tiêu chí R^2, *NSE*, *MAE* và *RMSE* được sử dụng để đánh giá mô hình và lựa chọn được bộ dữ liệu đầu vào tối ưu cũng như phương pháp hồi quy tin cậy để dự

báo mực nước lũ trên sông. Các mô hình học máy hồi quy được phát triển trong môi trường Python 3.7.

3.1. Bộ dữ liệu đầu vào tối ưu để dự báo mực nước tại trạm Lệ Thủy

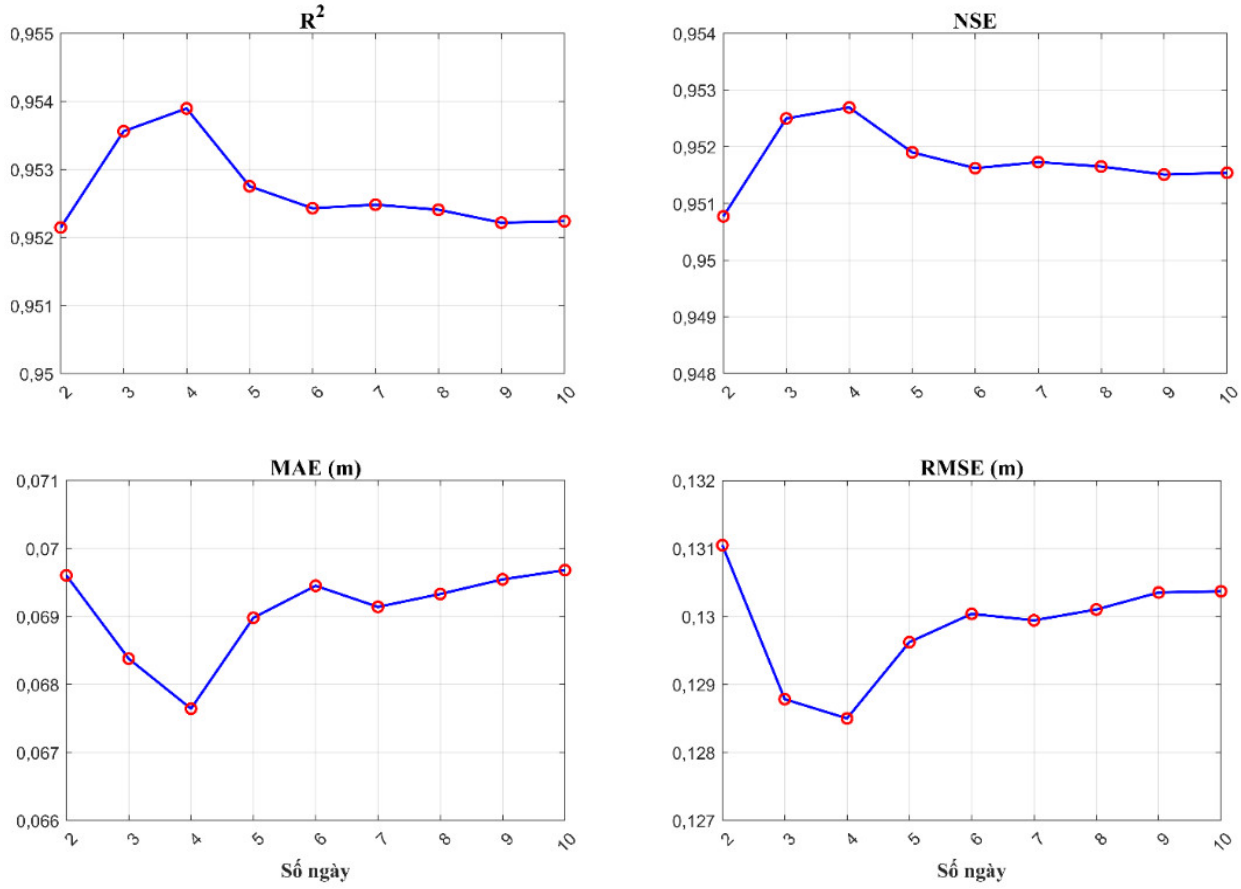
3.1.1. Lựa chọn độ dài dữ liệu trong quá khứ

Các mô hình theo định hướng dữ liệu được thiết lập dựa trên mối quan hệ thống kê giữa dữ liệu đầu vào và đầu ra để dự đoán các giá trị trong tương lai. Mực nước ngày dự báo $\hat{H}^{LT}_{t+1}$ tại trạm Lệ Thủy cho ngày tiếp theo (t+1) được coi là một hàm của các biến đầu vào (mưa và mực nước ngày) trong k ngày trước đó và có thể được biểu diễn như sau:

$$\hat{H}^{LT}_{t+1} = \int(R_{tt-1,t-2\ldots t-k}; H_{tt-1,t-2\ldots t-k}) \quad (9)$$

Trong đó: R và H là lượng mưa và mực nước ngày tại ba trạm; và t là thời điểm/ngày hiện tại.

Trước khi huấn luyện và kiểm định mô hình, phương pháp hồi quy tuyến tính được sử dụng với tất cả bộ số liệu của 6 thông số đầu vào (mưa và mực nước ngày tại 03 trạm) để phân tích và lựa chọn độ dài dữ liệu trong quá khứ (k) tối ưu trong dự đoán mực nước. Kết quả phân tích cho thấy việc sử dụng dữ liệu từ 3 đến 10 ngày trước để dự báo mực nước cho ngày hôm sau cho kết quả tốt với giá trị R^2 và *NSE* lần lượt dao động trong khoảng 0,9522÷0,9539 và 0,9508÷0,9527 (Hình 5). Đặc biệt với k =3 thì cả 4 tiêu chí đánh giá độ tin cậy của mô hình đều đạt kết quả tốt nhất với sai số *MAE* và *RMSE* lần lượt bằng là 6,76 cm và 12,85 cm. Do đó, nhóm nghiên cứu lựa chọn bộ dữ liệu đầu vào trong 4 ngày (tại các thời điểm t, t-1, t-2 và t-3) để dự báo mực nước ở thời điểm t+1 trong các trường hợp tính toán ở bước tiếp theo.



Hình 5. Các tiêu chí đánh giá của mô hình hồi quy tuyến tính với 6 biến đầu vào

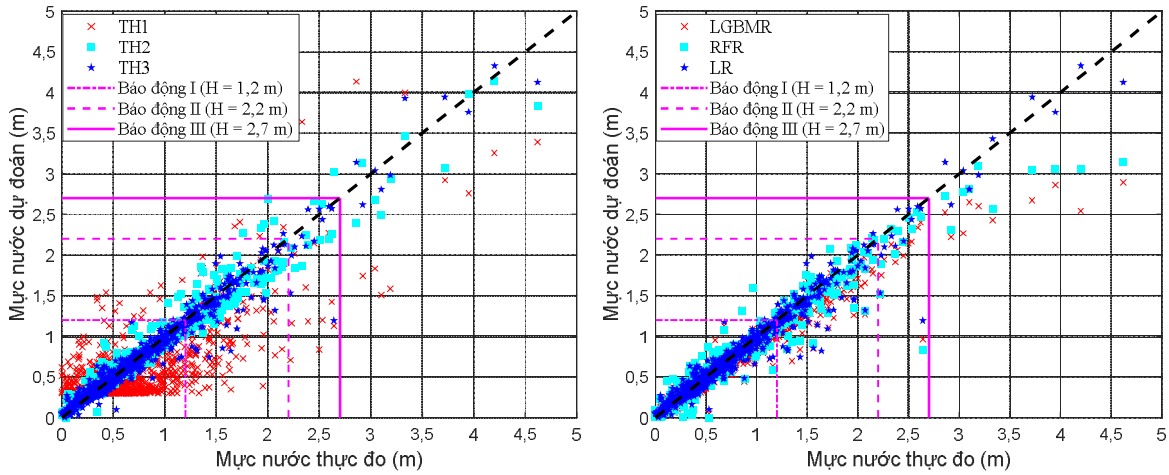
3.1.2. Lựa chọn bộ dữ liệu đầu vào tối ưu

Phương pháp hồi quy tuyến tính tiếp tục được sử dụng để phân tích và lựa chọn bộ dữ liệu đầu vào tối ưu cho bài toán dự báo mực nước $\hat{H}^{LT}_{t+1}$ tại trạm Lệ Thủy tại thời điểm t+1. Ba tổ hợp tính toán đã được phân tích như trong Bảng 2: i) tổ hợp TH1 chỉ xét đến các dữ liệu mưa ngày tại 03 trạm từ thời điểm t-3 tới t; ii) tổ hợp TH2 ngoài dữ liệu mưa giống như TH1 còn xét thêm yếu tố đầu vào là mực nước ngày tại 03 trạm trong thời đoạn [t-3; t]; và iii) tổ hợp TH3 như TH2 nhưng có xét thêm dữ liệu mưa dự báo R_{t+1} tại các trạm trong thời điểm t+1. Trong nghiên cứu này, mưa thực đo tại thời điểm t+1 được sử dụng như mưa dự báo; tuy nhiên trong thực tế vận hành sẽ sử dụng số liệu mưa dự báo từ các đài khí tượng thủy văn.

Bảng 2. Các tổ hợp tính toán nhằm phân tích và lựa chọn bộ đầu vào tối ưu

Tổ hợp dữ liệu đầu vào	Các vector đầu vào		Biến đầu ra
	Mưa tại 03 trạm	*Mực nước tại 03 trạm*	
TH1	R_t, R_{t-1}, R_{t-2}, R_{t-3}	-	$\hat{H}_{t+1}$
TH2	R_t, R_{t-1}, R_{t-2}, R_{t-3}	H_t, H_{t-1}, H_{t-2}, H_{t-3}	$\hat{H}_{t+1}$
TH3	R_{t+1}, R_t, R_{t-1}, R_{t-2}, R_{t-3}	H_t, H_{t-1}, H_{t-2}, H_{t-3}	$\hat{H}_{t+1}$

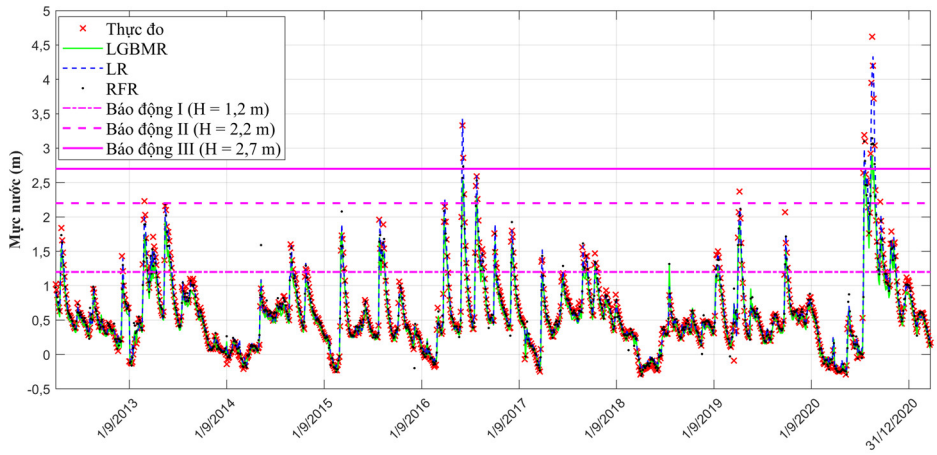
Biểu đồ phân tán trong Hình cho thấy, các điểm trong tổ hợp TH3 (màu xanh đậm) bám sát với đường 1:1 (đường y=x, màu đen nét đứt) hơn so với các điểm trong tổ hợp TH1 (màu đỏ) và tổ hợp TH2 (màu xanh nhạt). Tuy nhiên với TH2, các điểm mực nước trên báo động 3 (H = 2,7 m) phân tán khá rộng; có nhiều điểm mực nước dự đoán thiên thấp so với mực nước thực đo, thể hiện việc dự đoán đỉnh lũ của mô hình chưa tốt. Sự khác biệt lớn giữa mực nước sông mô phỏng và thực đo được quan sát thấy trong TH1 khi các điểm (màu đỏ) phân tán rộng và xa so với đường 1:1; đặc biệt trong trường hợp mực nước thấp. Điều này cho thấy nếu không xét đến mực nước trong quá khứ cũng như lượng mưa dự báo thì mô hình sẽ không bắt được chân và đỉnh của đường quá trình mực nước trong sông. Do đó, nhóm nghiên cứu đã sử dụng bộ số liệu đầu vào trong tổ hợp TH3 để huấn luyện và kiểm định các mô hình học máy dự báo mực nước theo ba phương pháp hồi quy *LR*, *RFR* và *LGBMR* trong phần tiếp theo.



Hình 6. Biểu đồ phân tán mực nước thực đo và dự đoán với 3 tổ hợp tính toán (trái) và 3 mô hình hồi quy (phải)

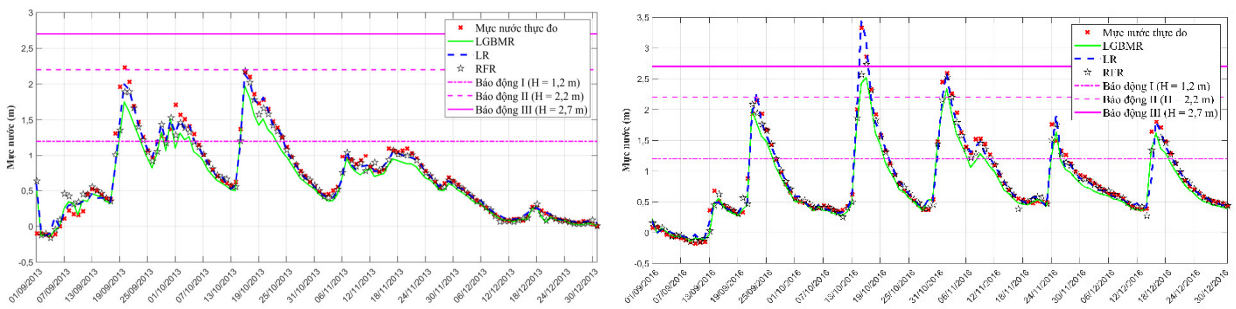
3.2. Kết quả dự báo mực nước theo ba phương pháp hồi quy

Chất lượng và kết quả dự đoán mực nước tại trạm Lệ Thủy sử dụng 03 mô hình hồi quy *LR*, *RFR* và *LGBMR* được thể hiện trong Hình 6. Mực nước dự đoán bằng mô hình hồi quy *LR* cho ra kết quả tốt nhất với các điểm bám sát đường 1:1. Hai mô hình *RFR* và *LGBMR* dự đoán mực nước dưới mức báo động III cũng khá tốt; tuy nhiên không bắt được đỉnh lũ khi mực nước trên 3 m (mực nước dự báo thiên thấp so với giá trị thực đo).



Hình 7. So sánh mực nước thực đo và dự đoán tại trạm Lệ Thủy với 3 MH hồi quy

Trong công tác phòng chống và giảm nhẹ thiên tai do lũ lụt thì việc dự báo chính xác được đỉnh lũ và thời gian đạt đỉnh là hết sức quan trọng, và là một trong các tiêu chí đánh giá hiệu quả của mô hình. Kết quả so sánh đường quá trình mực nước dự báo của 03 mô hình hồi quy với giá trị thực đo cho thấy mô hình hồi quy tuyến tính chiếm ưu thế khi dự báo xu thế và đỉnh lũ chính xác hơn (Hình 7). Mô hình *LGBMR* đánh giá thấp và thường không dự báo được các đỉnh lũ khi mực nước cao (ví dụ trên mức báo động III). Hình 8 so sánh kết quả dự báo mực nước của ba mô hình học máy với giá trị mực nước quan trắc trong mùa lũ năm 2013 và 2016. Trong trận lũ từ 15-24/10/2013, đỉnh lũ thực đo là 2,160 m; mô hình *LR* dự báo là 2,135 m (sai số 25 cm hay 1,15%); mô hình *RFR* và *LGBMR* dự báo lần lượt là 2,185 m và 1,975 m. Trong trận lũ từ 13-20/10/2016, sai số đỉnh lũ của 3 mô hình *LR*, *RFR* và *LGBMR* lần lượt là 9,9 cm (3,0%); 76,3 cm (22,9%) và 90,3 cm (27,1%). Nói chung các mô hình dự đoán tốt xu hướng đường quá trình lũ, tuy nhiên vẫn chưa dự đoán hoàn toàn chính xác được giá trị đỉnh và thời gian lũ lên. Nguyên nhân có thể là do việc sử dụng dữ liệu đầu vào theo ngày, hoặc nghiên cứu chưa xem xét một số các yếu tố khác có ảnh hưởng tới quá trình hình thành lũ như địa hình khu vực, thảm phủ mặt đất, điều kiện độ ẩm ban đầu.



Hình 8. So sánh mực nước thực đo và dự đoán tại Lệ Thủy trong mùa lũ năm 2013 và 2016

4. KẾT LUẬN

Trong nghiên cứu này, các tác giả đã xây dựng ba mô hình dự đoán mực nước theo ngày tại trạm Lệ Thủy trên sông Kiến Giang dựa trên phương pháp hồi quy *LR*, *RFR* và *LGBMR*. Bộ dữ liệu về lượng mưa và mực nước ngày tại 3 trạm trong mùa lũ của 44 năm đã được sử dụng để huấn luyện và kiểm định các mô hình. Kết quả nghiên cứu chỉ ra việc sử dụng bộ dữ liệu đầu vào gồm lượng mưa và mực nước ngày trong thời đoạn [*t*-3, *t*] và lượng mưa dự báo tại ba trạm khí tượng thủy văn trong khu vực để dự báo mực nước lũ tại trạm Lệ Thủy ở thời điểm *t*+1 cho kết quả tốt nhất khi mô hình có thể dự báo tốt cả chân lẫn đỉnh của đường quá trình mực nước. Các chỉ số thống kê R^2, *NSE, MAE và RMSE* cho thấy việc ứng dụng mô hình định hướng dữ liệu là hoàn toàn khả thi và đáng tin cậy trong việc dự đoán mực nước; trong đó mô hình dự đoán mực nước bằng phương pháp hồi quy tuyến tính cho kết quả tốt hơn so với hai phương pháp hồi quy còn lại. Trong các nghiên cứu tiếp theo, các tác giả sẽ xem xét đến việc bổ sung một số yếu tố đầu vào khác như dòng chảy, mực nước triều, mưa tại một số trạm lân cận, hay mưa dự báo từ các đài KTTV (thay vì sử dụng mưa thực đo tại thời điểm *t*+1), cũng như xây dựng mô hình dự báo mực nước cho các trạm thủy văn và trạm "ảo" khác dọc theo sông Kiến Giang và Nhật Lệ. Ngoài ra, các kỹ thuật học máy khác, như thuật toán học sâu, cũng sẽ được áp dụng nhằm cải thiện chất lượng dự báo mực nước trong tương lai.

LỜI CẢM ƠN

Nghiên cứu này được hỗ trợ bởi Đề tài nghiên cứu khoa học và phát triển công nghệ cấp bộ "Nghiên cứu ứng dụng giải pháp công nghệ số chuyển đổi hình thức cảnh báo lũ cho cộng đồng, xây dựng thí điểm cảnh báo lũ trên lưu vực sông Nhật Lệ, tỉnh Quảng Bình". Nhóm tác giả chân thành cảm ơn Ban chủ nhiệm đề tài đã tạo điều kiện tốt nhất để hoàn thành nghiên cứu này.

TÀI LIỆU THAM KHẢO

Nguyễn Đức Lý, Ngô Hải Dương, & Nguyễn Đại. (2013). *Khí hậu và thủy văn tỉnh Quảng Bình*. Nhà xuất bản Khoa học Kỹ thuật.

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

Breiman, L. (2001). *Random Forests.* Mach. Learn., 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Guo, W.-D., Chen, W.-B., Yeh, S.-H., Chang, C.-H., & Chen, H. (2021). *Prediction of River Stage Using Multistep-Ahead Machine Learning Techniques for a Tidal River of Taiwan.* Water, 13(7), 920. https://doi.org/10.3390/w13070920

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

Abstract:

PREDICTION OF WATER LEVEL IN KIEN GIANG RIVER USING REGRESSION-BASED MODELS

A reliable model to predict the water levels in a river is crucial for better planning to mitigate any risk associated with flooding. Data-driven models using machine learning (ML) techniques have become an attractive and effective approach to model and analyze river stage dynamics. In this study, three regression-based models, including Linear Regression (LR), Random Forest Regression (RFR) and Light Gradient Boosting Machine Regression (LGBMR) were developed and compared to predict the daily water levels in Kien Giang river based on collected data from 1977 to 2020. Four evaluation criteria, i.e., R^2, NSE, MAE, and RMSE, were employed to examine the reliability of the proposed models. The results show the high accuracy of the proposed models in predicting water levels, especially the LR model. The LR model outperforms the RFR and LGBMR models with the values of R^2, NSE, MAE and RMSE are 0.959, 0.958, 6.67 cm and 12.2 cm respectively.

Keywords: Water level prediction, machine learning, regression techniques, Kien Giang river.

Ngày nhận bài: 06/9/2022

Ngày chấp nhận đăng: 30/9/2022