

ỨNG DỤNG MÔ HÌNH HỌC MÁY DỰ BÁO LƯU LƯỢNG ĐẾN HỒ TẢ TRẠCH

TS. Tạ Quang Chiếu^{1*}

¹*Trường Đại học Thủy lợi*

*Tác giả liên hệ: *Tạ Quang Chiếu, quangchieu.ta@tlu.edu.vn*

THÔNG TIN CHUNG

Ngày nhận bài: 31/05/2023

Ngày nhận bài sửa: 13/06/2023

Ngày duyệt đăng: 22/06/2023

TỪ KHOẢ

Dự báo lưu lượng;

Học máy;

Hồ Tả Trạch;

Hồi quy

TÓM TẮT

Trong nghiên cứu này, các tác giả nghiên cứu, ứng dụng ba mô hình học máy hồi quy *LR*, *RFR* và *LGBMR* dự đoán lưu lượng đến hồ Tả Trạch của tỉnh Thừa Thiên Huế. Bộ dữ liệu thu được gồm lưu lượng đến hồ, lượng mưa 3 trạm Tả Trạch, Thượng Nhật, Nam Đông. Các kết quả chỉ ra rằng mô hình *LR* cho kết quả tốt nhất với các tiêu chí R^2 là 0.97, RMSE là 58.31, MAE là 27.68. Ngoài ra, mô hình *RFR* cũng cho kết quả tốt với MAX-ERROR là 1587.43.

ABSTRACT

In this study, the authors have proposed three regression models *LR*, *RFR* and *LGBMR* to predict the flow in Ta Trach reservoir of Thua Thien Hue province. The collected data set included the flow to the lake, the rainfall at 3 stations of Ta Trach, Thuong Nhat, and Nam Dong. The results showed that the *LR* model gave the best results with the criteria R^2 of 0.97, RMSE of 58.31, MAE of 27.68. In addition, the *RFR* model also gave good results with a MAX-ERROR of 1587.43.

1. GIỚI THIỆU

Hồ Tả Trạch (Hình 1) là công trình thủy lợi, thủy điện đa mục tiêu, xây dựng trên dòng chính là sông Tả Trạch đây là một phụ lưu của sông Hương. Hồ Tả Trạch thuộc xã Dương Hòa, thị xã Hương Thủy, tỉnh Thừa Thiên Huế và cách thành phố Huế 24 km về phía đông nam. Tổng dung tích hồ là 646 triệu m^3 do Ban Quản lý đầu tư và xây dựng thủy lợi 5 (thuộc Bộ Nông nghiệp và Phát triển nông thôn) quản lý.

Hồ Tả Trạch là một trong những công trình quan trọng đối với Quốc gia, là một trong bốn hồ thủy lợi lớn nhất cả nước. Tháng 2/2017 Thủ tướng Chính phủ có Quyết định 166 đưa hồ Tả Trạch vào danh sách công trình quan trọng liên quan đến an ninh quốc gia, mới giao cho Bộ NN-PTNT quản lý.

Hồ không chỉ cung cấp điện năng cho vùng miền và còn góp phần giảm cát lũ, giảm ngập lụt sâu trong thành phố Huế, các địa phương ở vùng hạ du được an toàn. Hồ Tả Trạch có vai trò rất quan trọng đầu tiên là chống lũ tiêu mẫn và lũ sớm, giảm lũ chính cho hệ thống sông Hương, giảm thiểu ngập úng cho thành phố Huế; cấp nước cho sinh hoạt và công nghiệp cho bà con thành phố Huế và các vùng lân cận. Cung cấp nguồn nước tưới ổn định cho 34.782 ha đất canh tác thuộc vùng đồng bằng sông Hương; bổ sung nguồn nước ngọt cho hạ lưu sông Hương để đầy mẫn, cải thiện môi trường vùng đầm phá, phục vụ nuôi trồng thủy sản với lưu lượng 25,0 m^3/s .

Hồ Tả Trạch được thiết kế với mục tiêu điều tiết lưu lượng về hạ du là chủ yếu nên khả năng cắt lũ của hồ này là rất lớn. Từ khi đi vào vận hành hồ, lũ hạ du sông Hương được giảm đi đáng kể.



Hình 1. Hồ Tả Trạch, tỉnh Thừa Thiên Huế

Với tính phức tạp của chế độ mưa tại Thừa Thiên Huế, việc vận hành hồ trong mùa mưa theo đúng quy trình, đảm bảo các mục tiêu: an toàn tuyệt đối cho các công trình hồ chứa, góp phần giảm lũ cho hạ du và đảm bảo hiệu quả phát điện hiện nay vẫn còn nhiều khó khăn. Chính vì sự quan trọng của Hồ Tả Trạch mà công tác dự báo lưu lượng đến hồ sẽ góp phần hỗ trợ công tác vận hành đảm bảo an toàn công trình và chống lũ cho hạ du.

Hiện nay công tác dự báo lưu lượng đến hồ chứa (Kim, Lee, & Kim, 2022), (Saad Mawlood Saab, Allawi, Sherif, & El-Shafie, 2022) là công việc dự báo trong tương lai một cách có khoa học về trạng thái biến đổi các yếu tố thủy văn. Tuy nhiên sự biến đổi này là một quá trình tự nhiên phức tạp, chịu tác động của nhiều yếu tố. Tính biến động của các yếu tố này phụ thuộc vào cả không gian và thời gian nên gây khó khăn rất lớn cho quá trình dự báo. Thêm nữa, do thiếu các trạm quan trắc cần thiết và thiếu sự kết hợp giữa các ngành liên quan cho nên dữ liệu quan trắc thực tế thường là không đầy đủ, không mang tính chất đại diện.

Do biến đổi khí hậu khắc nghiệt, nhu cầu phân tích chính xác tài nguyên nước ngày càng tăng để cung cấp nước ổn định và giảm thiểu thiệt hại do lũ lụt. Trong số các đối tượng nghiên cứu khác nhau, lượng lưu lượng vào đập là một yếu tố quan trọng trong việc thiết lập các kế hoạch đối phó với hạn hán, lũ lụt và vận hành đập. Các yếu tố chính ảnh hưởng đến lượng lưu lượng vào là các yếu tố khí hậu, bao gồm lượng mưa có ảnh hưởng lớn

nhất, nhiệt độ và tốc độ gió, cũng như các yếu tố địa hình như diện tích lưu vực và độ cao của dốc.

Nhiều phương pháp dự báo đã được đưa ra dựa trên mô hình vật lý, mô hình tham số như HBV, TOPMODEL, SHE, SWAT [4],... từ lâu đã được sử dụng để mô phỏng và dự đoán các sự kiện thủy văn. Các mô hình dựa trên tính chất vật lý có độ chính xác cao trong việc dự đoán, tuy nhiên chúng thường yêu cầu một lượng dữ liệu đầu vào rất lớn, bao gồm dữ liệu quan trắc, địa mạo, thủy văn,... và đòi hỏi nhiều thời gian trong việc thiết lập hiệu chỉnh mô hình, mô phỏng/tính toán. Do đó, việc dự báo trong thời gian ngắn hoặc theo thời gian thực, gần thực khó có thể áp dụng các mô hình này. Hơn nữa, việc phát triển các mô hình dựa trên tính chất vật lý yêu cầu người dùng phải có kiến thức chuyên sâu và kiến thức chuyên môn liên quan đến các thông số thủy văn, thủy lực, kinh nghiệm hiểu biết liên quan đến lưu vực cần tính toán. Vì vậy, đây là một thách thức lớn với các cơ quan chức năng. Điều này cho thấy nhu cầu ngày càng lớn về một hệ thống dự báo đơn giản mà chính xác để nâng cao hiệu quả trong công tác dự báo và cảnh báo lũ.

Tuy nhiên, gần đây công nghệ trí tuệ nhân tạo đã được sử dụng ngày càng nhiều trong các trường hợp để phân tích lưu lượng vào hồ. Các mô hình dựa trên học máy là các thuật toán máy tính tự cải thiện thông qua kinh nghiệm hoặc các bộ dữ liệu trong quá khứ một cách tự động. Dự báo lượng nước đổ về hồ bằng mô hình Perceptron đa lớp (MLP) trong vận hành hồ chứa. Kỹ thuật này đã được áp dụng trong một số nghiên cứu trước đây như (C., A.J., & S, 2016), (G.F., G.R., & P.Y, 2010), (Yang, Asanjan, Welles, Sorooshian, & Liu, 2017). Các chỉ số hiện tượng khí hậu đã được giới thiệu trực tiếp trong các mô hình máy học để dự báo lưu lượng vào hồ. Các kết quả đã được cung cấp để hỗ trợ ra quyết định trong các hoạt động của hồ chứa. Tương tự, (Lee, Kim, Jung, & Yoon,

2020) đã phát triển mô hình thống kê để dự báo lưu lượng vào đập. Nghiên cứu này đã sử dụng lưu lượng vào hồ hàng tháng và mười bốn chỉ số khí hậu, chẳng hạn như chỉ số dao động phương nam (SOI) Dao động phương Nam El Niño (ENSO), v.v. Các phương pháp này, hồi quy tuyến tính bội (MLR), máy vector hỗ trợ (SVM) và mạng lưới thần kinh nhân tạo (ANN), được sử dụng để dự báo lưu lượng hồ chứa hàng tháng ở Hàn Quốc. Các chỉ số khí hậu hàng tháng và bộ dữ liệu dòng chảy vào hồ chứa từ năm 1998 đến 2012 được sử dụng cho giai đoạn đào tạo và các bộ dữ liệu này từ năm 2013 đến 2016 được áp dụng cho giai đoạn thử nghiệm. Các phương pháp tập hợp, tính trung bình mô hình bayes (BMA), tính trung bình mô hình đơn giản (SMA) và dự báo (NF), được so sánh giữa hiệu suất của các mô hình. Kết quả cho thấy BMA chính xác và hữu ích hơn SMA và NF. (Hong & al, 2020) đã điều tra hiệu suất của sáu thuật toán để dự báo lượng dòng chảy vào đập sông Soyang. Kim et al. (Kim, et al., 2019) sử dụng mô hình mạng lưới thần kinh nhân tạo (ANN) để dự đoán lưu lượng vào của các đập bằng cách áp dụng dữ liệu khí tượng trong các khu vực nghiên cứu và lượng mưa lưu vực được tính toán bằng cách sử dụng mạng lưới Thiessen. Nghiên cứu này cho thấy mô hình sử dụng tất cả các trạm mưa trong mạng Thiessen hoạt động tốt hơn so với chỉ sử dụng các trạm trong lưu vực hoặc ngoài lưu vực. Kim et al. (Kim, Shon, Joo, Jang, & Shin, 2011) đã phân tích lượng mưa trung bình và dữ liệu lưu lượng vào đập Chungju ở lưu vực sông Hàn bằng cách áp dụng mô hình ANN bao gồm thuật toán lan truyền ngược. Nghiên cứu này cho thấy có sự cải thiện đáng kể về độ chính xác của mô hình bao gồm cả hệ số tương quan (CC) khi tiền xử lý dữ liệu được thực hiện. Mok và cộng sự (Mok, Choi, & Moon, 2020) đã áp dụng Trí nhớ ngắn hạn dài (LSTM) và mô hình ANN để dự đoán lưu lượng vào mỗi giờ của đập Yongdam. Trong nghiên cứu

này, các siêu tham số LSTM (trình tự, thứ nguyên ẩn, tốc độ học và phép lặp) đã được tối ưu hóa và độ chính xác của mô hình được cải thiện bằng cách áp dụng lưu lượng vào đập và lượng mưa làm biến đầu vào Lee và cộng sự. (Lee, Kim, & Han, 2020) đã thực hiện đánh giá định lượng bằng cách điều chỉnh và mô phỏng các biến đầu vào cho lưu vực sông Taehwa bằng cách sử dụng mạng nơ-ron tái phát (RNN), mạng nơ-ron trễ thời gian (TDNN) và các mô hình ngoại sinh tự hồi quy phi tuyến tính (NARX). Nghiên cứu này đã cải thiện hiệu suất (NSE) từ 0,530 lên 0,988 bằng cách điều chỉnh tham số thời gian trễ của mô hình Chang et al. (Chang, Hsu, & Chang, 2019) giới thiệu những tiến bộ gần đây về học máy trong dự đoán và quản lý lũ lụt, đồng thời trình bày một cách tiếp cận học thuật đối với mô hình liên quan đến rủi ro lũ lụt. Chang et al. (Chang, et al., 2014) khám phá hiệu quả của nhiều nguồn mưa để ước tính lượng mưa đa cảm biến dựa trên đồng hóa và thực hiện dự báo lượng mưa trước nhiều bước dựa trên lượng mưa được đồng hóa. Chakravarti và cộng sự. (Chakravarti, Joshi, & Panjiar, 2015) đã chứng minh rằng mô hình ANN có thể là một công cụ đầy hứa hẹn để cung cấp hiểu biết sâu sắc từ các mối quan hệ đã học cũng như mô hình hóa chính xác các quy trình phức tạp thông qua so sánh lượng lưu lượng được tạo ra bởi mô phỏng lượng mưa trong phòng thí nghiệm và lượng lưu lượng dự đoán của mô hình ANN. Kao et al. (Kao, Zhou, Chang, & Chang, 2020) lần đầu tiên đề xuất Bộ mã hóa-giải mã dựa trên bộ nhớ ngắn hạn dài (LSTM-ED) để dự đoán lũ trước nhiều bước. Shen và cộng sự. (Shen, et al., 2018) gợi ý rằng các nhà khoa học thủy văn xem xét nghiên cứu sử dụng khai thác dữ liệu dựa trên học sâu để bổ sung cho các phương pháp truyền thống. Tokar et al. (Tokar & Markus, 2000) đã so sánh và phân tích các mô hình khái niệm và mô hình ANN, khác nhau đối với từng lưu vực. Sau khi so sánh mô hình Watbal cho sông

Fraser, mô hình tính toán độ ẩm đất Sacramento (SACSMA) cho sông Raccoon và lượng mưa-lưu lượng theo khái niệm đơn giản (SCRR) cho Sông Little Patuxent, Colorado, Hoa Kỳ, với mô hình ANN, kết quả là mô hình ANN cùng với mô hình khái niệm hiện tại có thể được sử dụng để dự báo lưu lượng mưa. Chen và cộng sự. (Chen, Wang, & Tsou, 2013) đã so sánh và phân tích dữ liệu lượng mưa và lưu lượng mỗi giờ sau khi 27 cơn bão đổ bộ từ năm 2005 và 2009 tại lưu vực sông Linbien, Đài Loan, bằng cách áp dụng mô hình hồi quy thông thường và mô hình ANN cùng với khái niệm về Lan truyền ngược. Trong đánh giá thống kê, mô hình ANN cho kết quả tốt hơn so với mô hình phân tích hồi quy thông thường. Coulibaly và cộng sự. (Coulibaly, Anctil, & Bobee, 2001) dự đoán lưu lượng vào của các đập đa năng bằng cách áp dụng lượng mưa, tuyết rơi, lưu lượng vào và nhiệt độ làm biến đầu vào của bốn mô hình: Multilayer Perceptron (MLP), Mạng nơ-ron có độ trễ đầu vào (IDNN), RNN và Mạng nơ-ron hiện tại có độ trễ thời gian (TDRNN).

Những hạn chế của các mô hình dựa trên tính chất vật lý được đề cập ở trên đã tạo tiền đề cho việc sử dụng các mô hình theo định hướng dữ liệu nâng cao (advanced data-driven models như: mô hình học máy (Machine Learning model), mô hình học sâu (Deep learning model). Một lý do nữa cho sự phổ biến của các mô hình như vậy là chúng có thể hình thành tính phi tuyến tính về mặt số học, chỉ dựa trên dữ liệu về lịch sử mà không yêu cầu kiến thức về các quá trình vật lý cơ bản. Các mô hình dự báo theo hướng dữ liệu sử dụng phương pháp học máy là những công cụ đầy hứa hẹn vì chúng được xây dựng, phát triển nhanh hơn với yêu cầu đầu vào tối thiểu.

Bài toán dự báo lưu lượng nước về hồ Tả Trạch dựa vào các dữ liệu quan sát được trong quá khứ và các yếu tố tác động đến lưu lượng nước về hồ. Dự báo lưu lượng về hồ sẽ chịu tác động của nhiều

yếu tố ảnh hưởng đến kết quả dự báo như mưa, lưu lượng, địa hình, thảm phủ thực vật, độ ẩm, khí hậu và tác động của con người trên lưu vực. Trong những yếu tố đó thì yếu tố mưa là quan trọng nhất, các yếu tố về địa chất, thổ nhưỡng, thảm phủ thực vật ít thay đổi.

Trong bài báo này, nhóm nghiên cứu đề xuất và ứng dụng mô hình học máy sử dụng các phương pháp hồi quy (LR, RFR và LGBMR) để dự báo lưu lượng đến hồ Tả Trạch dựa trên số liệu mưa quan trắc và số liệu giám sát lưu lượng nước về hồ trong quá khứ.

Các kết quả dự báo đạt được sẽ giúp công tác hỗ trợ ra quyết định vận hành hồ chứa. Các kết quả này cũng có thể mở rộng cho bài toán dự báo lưu lượng đến các hồ chứa khác phục vụ công tác chỉ đạo điều hành phòng, tránh và giảm nhẹ thiên tai.

Cấu trúc bài báo gồm: Phần 2, Trình các phương pháp học máy (hồi quy). Phần 3, Giới thiệu các tiêu chí đánh giá hiệu suất của các mô hình. Phần 4, Kết quả và Thảo luận. Phần cuối cùng là Kết luận.

2. PHƯƠNG PHÁP

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 định hướng dữ liệu này được gọi là "giai đoạn học hồi", 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:

$$[(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)] = \{x_i, y_i\}_{i=1}^n \quad (2)$$

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

2.1. 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 $x_1, x_2, \dots, x_m$ trong không gian n chiều (biến độc lập, thuộc tính), có giá trị tương ứng của biến phụ thuộc (cần dự báo) là $y_1, y_2, \dots, y_m$. 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):

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

2.2. 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 2). 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:

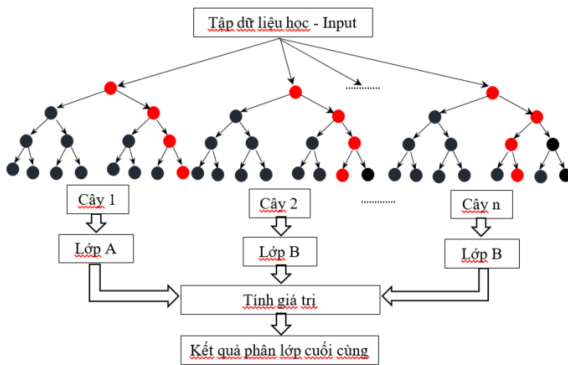
1. Trên cơ sở của phương pháp bootstrap, một tập hợp con các mẫu được tạo ngẫu nhiên với các mẫu thay thế từ tập dữ liệu ban đầu;

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

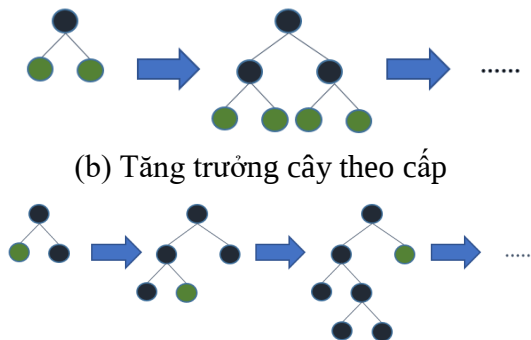
3. Mỗi cây hồi quy cung cấp kết quả dự đoán. Khi tất cả các cây hồi quy đã đạt đến kích thước tối đa, dự đoán cuối cùng được xác định là giá trị trung bình của các kết quả từ tất cả các cây hồi quy (Guo và nnk, 2021) như trong công thức (5):

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

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



(a) Cây Quyết định



(b) Tăng trưởng cây theo cấp

Hình 2. Phương pháp hồi quy RFR và LGBMR

2.3. Light Gradient Boosted 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 2) 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).

3. CÁC TIÊU CHÍ ĐÁNH GIÁ HIỆU SUẤT CỦA CÁC MÔ HÌNH

Để đánh giá hiệu quả của mô hình, chúng tôi sử dụng các chỉ số như R^2 , RMSE, MAE. Ngoài ra, bài toán đề xuất sử dụng thêm tiêu chí Max Error Value để đánh giá sai số trong dự đoán lưu lượng đến hồ.

3.1. Hệ số tương quan R^2 (R^2 Coefficients)

Hệ số xác định (R^2) là một số từ 0 đến 1 đo lường mức độ hiệu quả của một mô hình thống kê dự đoán kết quả. Hệ số được tính toán dựa trên tỷ lệ giữa tổng bình phương của số dư giữa giá trị thực tế và giá trị dự đoán và tổng bình phương sai của các giá trị thực tế. Hệ số R^2 càng cao thì mô hình càng có khả năng dự đoán chính xác. Công thức của R^2 được trình bày như sau:

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

Trong đó H_i^{mea} và H_i^{pre} lần lượt là giá trị thực tế và giá trị dự đoán của lưu lượng; $\bar{H}^{mea}$ và $\bar{H}^{pre}$ lần lượt là trung bình của lưu lượng đến hồ thực tế và lưu lượng đến hồ dự đoán.

3.2. Căn bậc hai của trung bình sai số bình phương (Root Mean Squared Error - RMSE)

Căn bậc hai của trung bình sai số bình phương (RMSE) đo lường sai số tuyệt đối giữa các giá trị dự đoán và giá trị thực tế. RMSE thường được sử dụng làm chỉ số đánh giá trong các bài toán khi có yêu cầu cao trong việc tối giản sai số đối với các giá trị ngoại lai (outliers). RMSE của các giá trị ngoại lai sẽ được khuếch đại hơn (thông qua bình phương sai số) so với các giá trị thông thường trong dữ liệu thực tế. Giá trị của RMSE càng nhỏ thì mô hình càng có tính chính xác cao. Công thức của RMSE được trình bày như sau:

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

Trong đó H_i^{mea} và H_i^{pre} lần lượt là giá trị thực tế và giá trị dự đoán của mực nước; H_i^{mea} và H_i^{pre} lần lượt là trung bình của lưu lượng đến hồ thực tế và lưu lượng đến hồ dự đoán.

3.3. Trung bình sai số tuyệt đối (Mean Absolute Error - MAE)

Trung bình sai số tuyệt đối (MAE) đo lường sai số tuyệt đối giữa các giá trị dự đoán và giá trị thực tế. MAE thường được sử dụng làm chỉ số đánh giá trong các bài toán khi có yêu cầu cao trong việc tối giản sai số đối với toàn bộ các điểm dữ liệu thực tế. Giá trị của MAE càng nhỏ thì mô hình càng có tính chính xác cao. Công thức của MAE được trình bày như sau:

$$MAE = \frac{\sum_{i=1}^n |H_i^{mea} - H_i^{pre}|}{n}$$

Trong đó H_i^{mea} và H_i^{pre} lần lượt là giá trị thực tế và giá trị dự đoán của mực nước; H_i^{mea} và H_i^{pre} lần lượt là trung bình của lưu lượng đến hồ thực tế và lưu lượng đến hồ dự đoán.

3.4. Sai số tuyệt đối lớn nhất (Maximum Error)

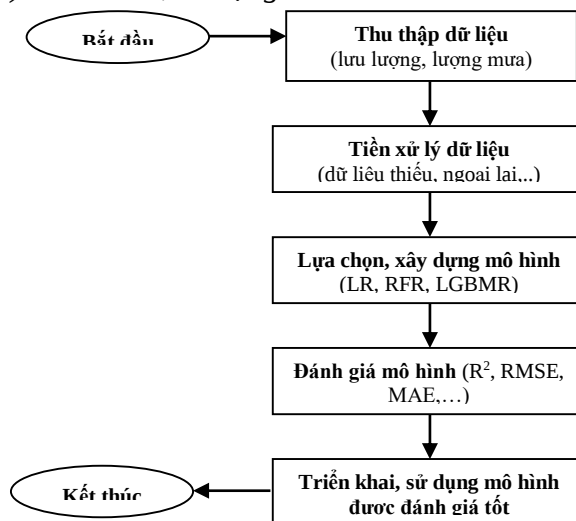
ME (Max-error) Là giá trị sai số tuyệt đối lớn nhất (sự sai khác lớn nhất) giữa lưu lượng đến hồ thực tế và lưu lượng đến hồ dự đoán.

$$Max.error = ME = \max_i |H_i^{mea} - H_i^{pre}|$$

4. THU THẬP DỮ LIỆU VÀ TRIỂN KHAI BÀI TOÁN

4.1. Các bước triển khai bài toán

Các bước chính để triển khai bài toán gồm: 1) Thu thập dữ liệu, 2) Tiền xử lý dữ liệu, 3) Lựa chọn, xây dựng mô hình, 4) Đánh giá mô hình và 5) Triển khai, sử dụng mô hình.

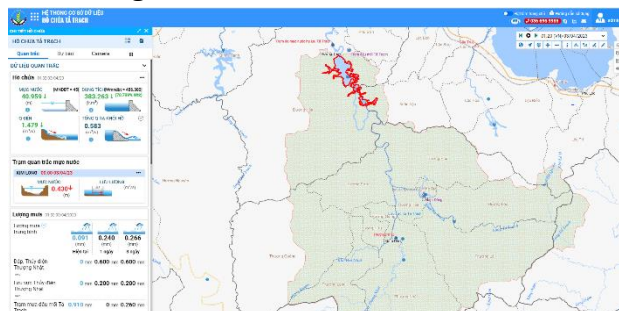


Hình 3. Các bước chính triển khai bài toán

Ở bước đầu tiên, các biến được thu thập bao gồm thời gian, lưu lượng vào hồ, lượng mưa của 3 trạm (Tả Trạch, Thượng Nhật, Nam Đông), các dữ liệu thu thập được trong thời gian từ tháng 01/2017 đến tháng 04/2023, gồm hơn 55000 dữ liệu. Bước thứ hai, thực hiện tiền xử lý dữ liệu với các dữ liệu thiếu, dữ liệu ngoại lai, làm sạch dữ liệu, tích hợp dữ liệu, biến đổi dữ liệu, thu giảm dữ liệu. Bước thứ ba, lựa chọn, xây dựng (huấn luyện) các mô hình học máy cho bài toán. Bước thứ tiếp theo, đánh giá hiệu suất của các mô hình được lựa chọn. Cuối cùng, mô hình tốt nhất và thời gian thực hiện tối ưu được lựa chọn và ứng dụng như trong Hình 3.

4.2. Thu thập dữ liệu

Số liệu được sử dụng tính toán trong bài báo được khai thác trực tuyến trên hệ thống quản lý giám sát số liệu theo thời gian thực hồ Tả Trạch (Hình 4) tại địa chỉ <https://hotatrach.vn> do Ban Quản lý Đầu tư và Xây dựng Thủy lợi 5 quản lý. Hệ thống hồ Tả Trạch được trang bị hệ thống quan trắc tự động số liệu được truyền về cơ sở dữ liệu theo thời gian thực.



Hình 4. Hệ thống cơ sở dữ liệu giám sát hồ chứa Tả Trạch

Số liệu được khai thác trên hệ thống thông qua báo cáo với tần suất dữ liệu 1h/số liệu quan trắc. Các dữ liệu được sử dụng được thể hiện trong Bảng 1.

Bảng 1. Thông tin lưu lượng và các trạm đo mưa

STT	Số liệu quan trắc	Đơn vị tính	Tần suất dữ liệu	Thời gian dữ liệu
1	Lưu lượng đến hồ	m ³ /s	1h/lần	03/2017 - nay
2	Trạm đo mưa đầu mối hồ Tả Trạch	mm/h	1h/lần	01/2017 - 04/2023
3	Trạm đo mưa đập thủy điện Thượng Nhật (Trạm đo lưu vực hồ Tả Trạch)	mm/h	1h/lần	01/2017 - 04/2023
4	Trạm đo mưa Nam Đông (Trạm đo lưu vực hồ Tả Trạch)	mm/h	1h/lần	01/2017 - 04/2023

4.3. Lựa chọn độ dài dữ liệu để xây dựng và đánh giá mô hình

Trước khi triển khai huấn luyện mô hình, chúng tôi đã thực hiện tiền xử lý dữ liệu với các dữ liệu thiếu bằng cách bù dữ liệu và tùy theo thời điểm để bổ sung dữ liệu.

Dữ liệu tập train (train set) dùng để xây dựng mô hình được sử dụng từ 01/2017-12/2021 chiếm gần 80% gồm các lượng mưa của 3 trạm Tả Trạch, Thượng Nhật, Nam Đông và lưu lượng đến hồ Tả Trạch.

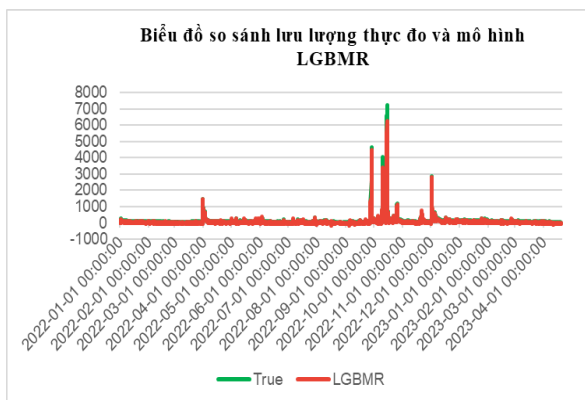
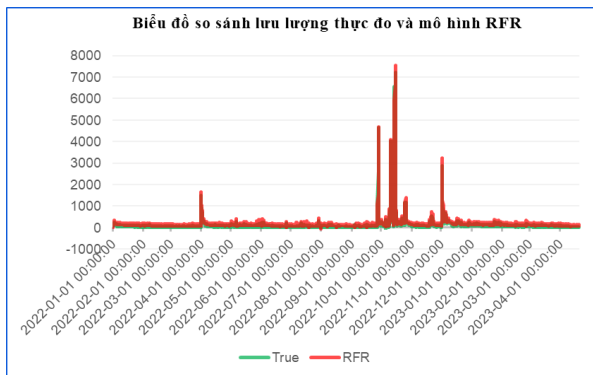
Trong khi đó dữ liệu tập test (test set) từ 01/2022 đến 04/2023 chiếm hơn 20% được sử dụng đánh giá mô hình gồm dữ liệu lượng mưa của 3 trạm Tả Trạch, Thượng Nhật, Nam Đông và để dự báo lưu lượng đến hồ Tả Trạch.

Để lựa chọn và đánh giá được mô hình, nhóm nghiên cứu đã thực nghiệm chạy các mô hình với các bộ tham số khác nhau cùng với các tiêu chí đánh giá mô hình R^2 , RMSE, MAE,... để có được bộ tham số tốt nhất cho mô hình.

5. KẾT QUẢ

Hình 5. Biểu đồ so sánh lưu lượng thực đo và mô hình LR

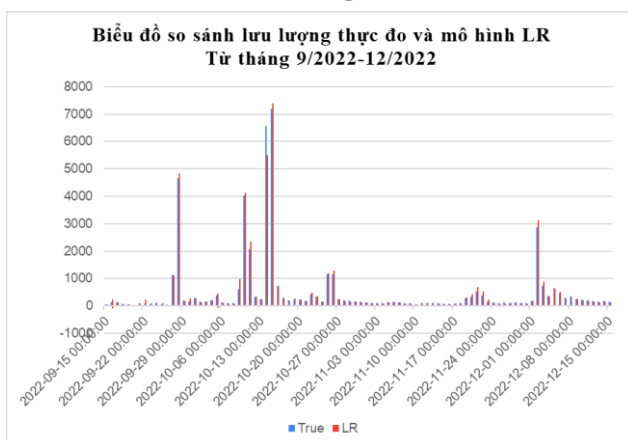
Hình 5. Biểu đồ so sánh lưu lượng thực đo và mô hình LR



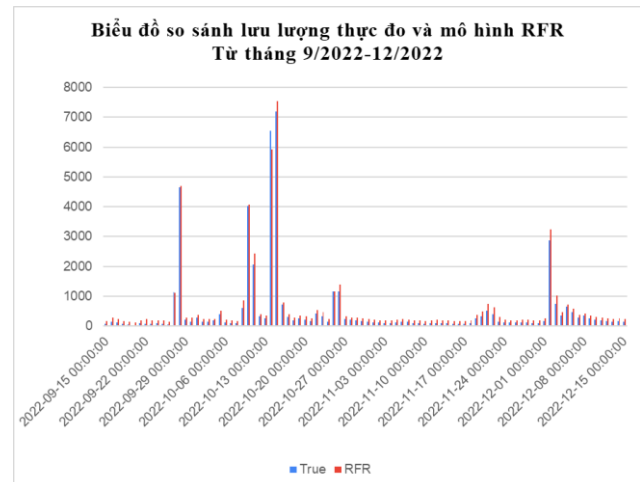
Hình 6. Biểu đồ so sánh lưu lượng thực đo và mô hình RFR

Kết quả dự đoán lưu lượng đến hồ Tả Trạch sử dụng 03 mô hình hồi quy *LR*, *RFR* và *LGBMR* so với giá trị thực đo từ tháng 1/2022 đến 4/2023 được thể hiện trong các **Error! Reference source not found.**, **Error! Reference source not found.**, **Error! Reference source not found.**, 7

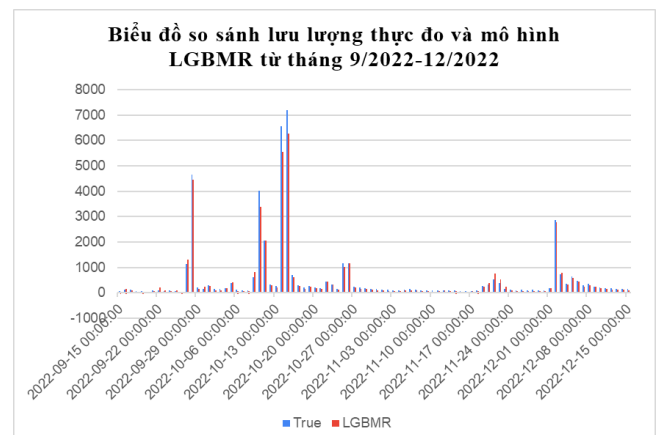
Hình 7. Biểu đồ so sánh lưu lượng thực đo và mô hình LGBMR



Hình 7. Biểu đồ so sánh lưu lượng thực đo và mô hình LR từ tháng 9/2022-12/2022



Hình 8. Biểu đồ so sánh lưu lượng thực đo và mô hình RFR từ tháng 9/2022-12/2022



Hình 9. Biểu đồ so sánh lưu lượng thực đo và mô hình LGBMR từ tháng 9/2022-12/2022

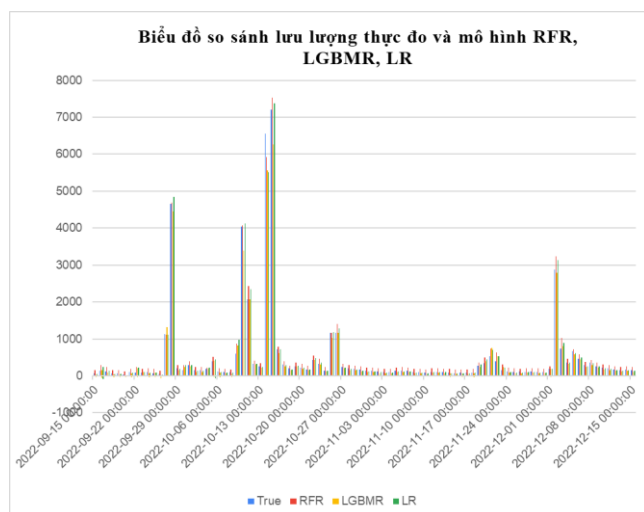
Trong các Hình 8, Hình 9, Hình 10 là kết quả dự đoán lưu lượng đến hồ Tả Trạch của 03 mô hình hồi quy *LR*, *RFR* và *LGBMR* so với giá trị thực đo ở thời điểm mùa mưa lũ từ tháng 09/2022 đến 12/2022.

Trong công tác phòng chống và giảm nhẹ thiên tai do lũ lụt thì việc dự báo chính xác được thời gian mà lưu lượng về hồ lớn nhất là hết sức quan trọng, và là một trong các tiêu chí đánh giá hiệu quả của mô hình. Từ các kết quả của từng mô hình được so sánh với giá trị thực đo tại Hình 5, 6, 7. Trong quá trình chạy thực nghiệm các mô hình, nhóm nghiên cứu đã thử nghiệm với các bộ tham

số khác nhau và với các tham số mặc định đang cho kết quả tốt hơn. Các kết quả của 3 mô hình được thể hiện trên biểu đồ (Hình 11) và cùng với các tiêu chí đánh giá hiệu quả của mô hình tại (Bảng 2. Số liệu trung bình với các tiêu chí đánh giá mô hình) chỉ ra rằng mô hình LR cho kết quả tốt nhất trong 3 mô hình được đề xuất. Tại thời điểm 02h ngày 15/10/2022 với lưu lượng đến hồ lớn nhất đo được là 7198.97 m³/s trong khi đó kết quả của 3 mô hình LR, RFR, LGBMR lần lượt là: 7281.69 m³/s, 7534.01 m³/s, 5737.68 m³/s. Với các kết quả đạt được trong bài báo, chúng tôi đã tham chiếu đến kết quả bài báo của tác giả Jiyeong Hong và cộng sự (Hong & al, 2020). Kết quả chỉ ra rằng phương pháp chúng tôi áp dụng có một số chỉ số đánh giá mô hình tốt hơn, ví dụ như kết quả R² cho giá trị là 0.97.

Bảng 2. Số liệu trung bình với các tiêu chí đánh giá mô hình

	R ²	RMSE	MAE	MAX ERROR
LGBM				
R	0.83	101.47	90.14	1676.01
RFR	0.93	67.90	41.80	1587.43
LR	0.97	58.31	27.68	2155.74



Hình 10. Biểu đồ so sánh lưu lượng thực đo và mô hình RFR, LGBMR, LR từ tháng 9/2022-12/2022

6. KẾT LUẬN

Trong nghiên cứu này, các tác giả đã nghiên cứu, áp dụng ba mô hình học máy dự đoán lưu lượng đến hồ Tả Trạch của tỉnh Thừa Thiên Huế dựa trên phương pháp hồi quy *LR*, *RFR* và *LGBMR*. Bộ dữ liệu thu được gồm lưu lượng đến hồ, lượng mưa 3 trạm Tả Trạch, Thượng Nhật, Nam Đông. Dữ liệu huấn luyện mô hình được sử dụng từ 01/2017 đến tháng 12/2021, dữ liệu để kiểm tra mô hình từ tháng 01/2022 đến 04/2023. Các kết quả chỉ ra rằng mô hình LR cho kết quả tốt nhất với các tiêu chí R² là 0.97, RMSE là 58.31, MAE là 27.68. Ngoài ra, mô hình RFR cũng cho kết quả tốt với MAX-ERROR là 1587.43 thể hiện sự sai khác tuyệt đối lớn nhất giữa lưu lượng đến hồ dự báo và lưu lượng đến hồ thực tế sai khác nhỏ.

Trong tương lai, nhóm nghiên cứu sẽ thực hiện bài toán dự báo lưu lượng với các phương pháp học sâu và tối ưu bộ tham số. Tiếp đến chúng tôi sẽ so sánh các kết quả các mô hình học máy đã thực hiện này với các mô hình mới.

TÀI LIỆU THAM KHẢO

- C., C., A.J., A., & S, B.-S. (2016). Inflow forecasting using artificial neural networks for reservoir operation. *Proc. Int. Assoc. Hydrol*, 209–214.
- Chakravarti, A., Joshi, N., & Panjiar, H. (2015). Rainfall Runoff Analysis Using the Artificial Neural Network. *Indian J. Sci. Technol*, 1–7.
- Chang, F., Chiang, Y., Tsai, M., Shieh, M., Hsu, K., & Sorooshian, S. (2014). Watershed

- rainfall forecasting using neuro-fuzzy networks with the assimilation of multi-sensor information. *J. Hydrol*, 374–384.
- Chang, F., Hsu, K., & Chang, L. (2019). Flood Forecasting Using Machine Learning Methods. *MDPI: Basel*.
- Chen, S., Wang, Y., & Tsou, I. (2013). Using artificial neural network approach for modelling rainfall–runoff due to typhoon. *J. Earth Syst*, 399–405.
- Coulibaly, P., Anctil, F., & Bobee, B. (2001). Multivariate reservoir inflow forecasting using temporal neural networks. *J. Hydrol. Eng*, 367–376.
- G.F., L., G.R., C., & P.Y, H. (2010). Effective typhoon characteristics and their effects on hourly reservoir inflow forecasting. *Adv. Water Resour*, 887–898 .
- Hong, J., & al, e. (2020). Development and evaluation of the combined machine learning models for the prediction of dam inflow. *Water* , 29-37.
- Kao, I., Zhou, Y., Chang, L., & Chang, F. (2020). Exploring a Long Short-Term Memory based Encoder-Decoder framework for multi-step-ahead flood forecasting. *J. Hydrol*.
- Kim, B.-J., Lee, Y.-T., & Kim, B.-H. (2022). A Study on the Optimal Deep Learning Model for Dam Inflow Prediction. *Water*, 27-66.
doi:<https://doi.org/10.3390/w14172766>
- Kim, M., Shon, T., Joo, J., Jang, Y., & Shin, H. (2011). Forecasting of Short-term Runoff with Artificial Neural Network with Pre-processing Techniques. *Joint Fall Conference & Water Korea*. Daejeon, Korea: Preceedings of the Joint Fall Conference.
- Kim, S., Kim, K., Hwang, S., Park, J., Lee, J., & Kang, M. (2019). Influence of Rainfall observation Network on Daily Dam Inflow using Artificial Neural Networks. *J. Korean Soc. Agric. Eng*, 63–74.
- Lee, D., Kim, H., Jung, I., & Yoon, J. (2020). Monthly reservoir inflow forecasting for dry period using teleconnection indices: a statistical ensemble approach. *Appl*, 34-70.
- Lee, J., Kim, H., & Han, K. (2020). Linkage of Hydrological Model and Machine Learning for Real-time Prediction of River Flood. *J. Korean Soc. Civ. Eng*, 303–314.
- Mok, J., Choi, J., & Moon, Y. (2020). Prediction of Multipose Dam Inflow using Deep Learning. *J. Korea Water Resour*, 97–105.
- Saad Mawlood Saab, F. O., Allawi, M. F., Sherif, M., & El-Shafie, A. (2022). Utilizing deep learning machine for inflow forecasting in two different environment regions: a case study of a tropical and semi-arid region. *Applied Water Science*, 12, 27-41.
- Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F., . . . al, e. (2018). HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community. *Hydrol. Earth Syst*, 5639–5656.

- Tokar, A., & Markus, M. (2000). Precipitation-runoff modeling using artificial neural networks and conceptual models. *J. Hydrol. Eng*, 156–161.
- Yang, T., Asanjan, A., Welles, E. G., Sorooshian, S., & Liu, X. (2017). Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. *Water Resour. Res*, 2786–2812 .