

ỨNG DỤNG MÔ HÌNH HỌC MÁY XGBOOST NHẬN DẠNG LŨ LỚN TRÊN SÔNG ĐÀ

Lương Hữu Dũng⁽¹⁾, Ngô Thị Thủy⁽¹⁾, Lương Tuấn Trung⁽¹⁾, Trịnh Thu Phương⁽²⁾

⁽¹⁾Viện Khoa học Khí tượng Thủy văn và Biến đổi khí hậu

⁽²⁾Ủy hội sông Mê Công Việt Nam

Ngày nhận bài: 11/5/2025; ngày chuyển phản biện: 12/5/2025; ngày chấp nhận đăng: 30/5/2025

Tóm tắt: Nhận dạng lũ lớn trên các lưu vực sông đóng vai trò quan trọng trong công tác phòng chống thiên tai và đảm bảo sự phát triển bền vững của lưu vực. Hiện nay, với sự phát triển vượt bậc của khoa học máy tính, các phương pháp số như mô hình toán và trí tuệ nhân tạo đang được ứng dụng rộng rãi trong các nghiên cứu dự báo và nhận dạng các đặc trưng lũ lớn trên lưu vực Sông Hồng. Tuy nhiên, để xây dựng được mô hình học máy có thể nhận dạng, dự báo được những trận lũ lớn với độ chính xác cao vẫn còn gặp nhiều thách thức về nguồn dữ liệu và các quan hệ phi tuyến phức tạp. Nghiên cứu này được thực hiện với mục đích xây dựng được thuật toán học máy đủ mạnh, có thể dự báo, nhận dạng được dòng chảy đỉnh lũ đến Hồ Sơn La thông qua Trạm thủy văn Tạ Bú nằm ở hạ lưu hồ trong những trận lũ lớn từ nguồn dữ liệu ngắn (49 số liệu), không có sự liên tục về thời gian. Mô hình học máy tiên tiến XGBoost và các kỹ thuật xử lý dữ liệu đầu vào được sử dụng cho kết quả huấn luyện đạt trên 97% trong tất cả 1.000 lần chạy và kết quả kiểm tra đạt khoảng 70%. Dự báo thử nghiệm cho trận lũ năm 2024 cũng cho kết quả dự báo đỉnh lũ chính vụ khá tốt, với sai số dưới 5%. Kết quả nhận dạng đỉnh lũ sớm và lũ muộn chấp nhận được với sai số đỉnh lũ trung bình ngày khoảng 14%.

Từ khóa: Nhận dạng lũ, dự báo lũ, Sông Đà, học máy, XGBoost.

1. Giới thiệu chung

Sông Đà là sông lớn nhất trong hệ thống Sông Hồng, chảy qua địa phận của ba quốc gia là Trung Quốc, Lào và Việt Nam với tổng diện tích lưu vực là 52.900 km², trong đó phần lãnh thổ phía Trung Quốc chiếm khoảng 47%, Lào chiếm khoảng 2% còn lại là Việt Nam khoảng 51% [1]. Đoạn sông ở Việt Nam dài 527 km (có tài liệu ghi 543 km [2]) có 10 phụ lưu cấp 1, trong đó có 4 sông nhánh chính với diện tích lưu vực trên 2.000 km²: Nậm Pô, Nậm Na, Nậm Mực, Nậm Mu; hai phụ lưu nhỏ hơn là Nậm Sập và Nậm Bú (Hình 1).

Do đặc điểm về địa hình và hướng núi nên lượng mưa trên lưu vực phân bố không đều theo không gian và thời gian. Lượng mưa tập trung chủ yếu từ tháng 5-10 chiếm 85-88% tổng lượng mưa cả năm, vậy nên dòng chảy vào mùa

lũ cũng chiếm khoảng 80% tổng dòng chảy cả năm. Lượng nước năm trên Sông Đà tại Trạm Hòa Bình đạt trung bình 55,4 tỷ m³. Trên lưu vực Sông Đà có 3 hồ chứa lớn: Lai Châu, Sơn La, Hòa Bình đều là các hồ chứa lợi dụng tổng hợp. Dòng chảy qua Hồ Hòa Bình đóng góp đến 50% cho dòng chảy hạ lưu và Sông Đà đóng góp khoảng 37% tổng lượng nước Sông Hồng tại Sơn Tây. Là lưu vực có lượng mưa khá lớn, địa hình đa dạng, nên lũ lớn hình thành trên lưu vực Sông Đà là khá thường xuyên. Chính vì vậy, việc nhận dạng, dự báo lũ lớn đóng vai trò quan trọng trong việc nâng cao hiệu quả phòng lũ của hệ thống công trình trên lưu vực Sông Đà.

Nhận dạng lũ là cách thức nhận biết những dấu hiệu về khả năng hình thành, nhận biết đặc tính của một đợt lũ hay một trận lũ một cách định lượng, gồm các đặc trưng chính: Đỉnh lũ, tổng lượng lũ và dạng quá trình lũ [3]. Nhận dạng lũ đã được thực hiện tại các trạm thủy văn trên các lưu vực sông chính tại Việt Nam. Nhận

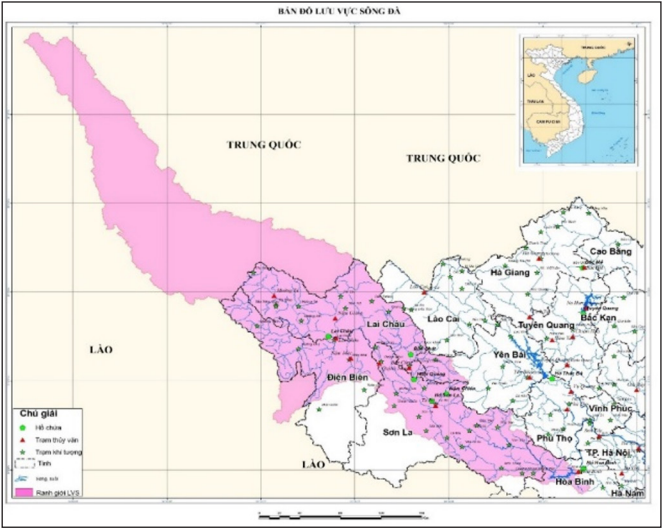
Liên hệ tác giả: Lương Hữu Dũng

Email: dungluonghuu@gmail.com

dạng lũ từ xa đến gần theo quy mô khái quát và chi tiết. Đối với nhận dạng càng xa, các thông tin thường ở dạng khái quát và có độ tin cậy ở mức tương đối. Nhận dạng lũ càng gần, các thông tin sẽ chi tiết hơn, có độ tin cậy cao hơn. Tiêu chí nhận dạng lũ là nhận biết, dự đoán khái quát được trước các giá trị đặc trưng của lũ (mức nước/ lưu lượng lớn nhất, nhỏ nhất, trung bình) và nhận dạng chi tiết được quá trình mức nước/ lưu lượng lũ. Nhận dạng lũ được thực hiện theo các thời hạn từ hạn ngắn (24-48 giờ) đến hạn vừa (5-10 ngày) và hạn dài (tháng, mùa). Các nghiên cứu được thực hiện trên cơ sở sử dụng các phương pháp từ thống kê truyền thống [4] cho đến các mô hình toán thủy văn từ mô hình thông số tập trung [5] đến mô hình thông số phân bố [6]. Với sự tiến bộ của công nghệ máy tính, các phương pháp số, sử dụng trí tuệ nhân tạo (học máy, học sâu) được sử dụng để nhận dạng lũ trên các sông ở Việt Nam [3], [7]. Tuy

nhien những phương pháp này về cơ bản mới huấn luyện tốt ở tập dữ liệu liên tục về thời gian và thường sử dụng thuật toán tự hồi quy (như NARX) với tập dữ liệu dài (thường lên đến hàng trăm số liệu) để dự báo cho tương lai.

Chính vì vậy, các mô hình xây dựng được còn nhiều hạn chế khi dự báo thử nghiệm cho chuỗi dữ liệu ngắn nhất là khi các dữ liệu không có sự liên tục về thời gian như lưu lượng đỉnh lũ của các trận lũ lớn trong khoảng 40-50 năm gần đây. Nghiên cứu này được thực hiện để xây dựng mô hình học máy tiên tiến với mục tiêu nhận dạng, dự báo được lưu lượng đỉnh lũ đến vị trí Trạm thủy văn Tạ Bú (nằm ở hạ lưu thủy điện Sơn La) trên Sông Đà từ dữ liệu khí tượng, thủy văn của các trạm lân cận, từ đó dự báo được đỉnh lũ đến Hồ thủy điện Sơn La. Phương pháp học máy được sử dụng trong nghiên cứu là mô hình XGBoost (Extreme Gradient Boosting).



Hình 1. Bản đồ và mạng lưới trạm khí tượng thủy văn lưu vực Sông Đà

2. Phương pháp thực hiện

2.1. Mô hình học máy XGBoost

XGBoost (Extreme Gradient Boosting) là một trong những thuật toán học máy (học có giám sát) mạnh mẽ và phổ biến nhất hiện nay, đặc biệt hiệu quả trong các bài toán phân loại và hồi quy có dữ liệu lớn và phức tạp. Thuật toán này được đề xuất bởi Chen và Guestrin [8], dựa trên khung Boosting truyền thống nhưng được cải tiến đáng kể về mặt hiệu suất tính toán

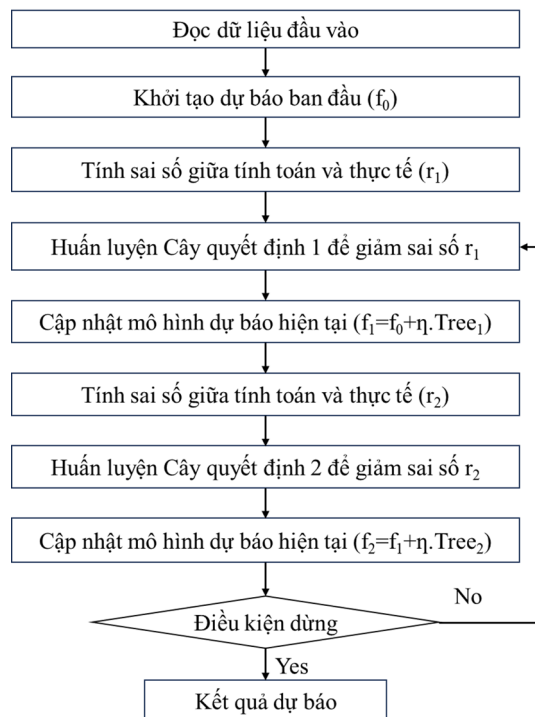
và khả năng tổng quát hóa mô hình. XGBoost sử dụng một phương pháp tối ưu hóa độ dốc (gradient) theo hình thức cộng dồn, kết hợp nhiều cây quyết định yếu để xây dựng một mô hình mạnh thông qua việc giảm thiểu hàm tổn thất bằng cách thêm từng cây một cách tuần tự. Điểm nổi bật của XGBoost là khả năng xử lý tốt dữ liệu thiếu, tự động lựa chọn đặc trưng quan trọng và khả năng song song hóa trong quá trình huấn luyện, nhờ đó cải thiện đáng kể tốc

độ so với các thuật toán Boosting truyền thống như AdaBoost hay Gradient Boosting Machine (GBM) [9].

Ngoài ra, XGBoost tích hợp cơ chế chuẩn hóa (regularization) thông qua các hệ số phạt (L1 và L2) vào hàm mục tiêu, từ đó giúp mô hình tránh được hiện tượng “quá khớp” (overfitting) - một vấn đề phổ biến trong học máy. Nhiều nghiên cứu đã chứng minh hiệu quả vượt trội của

XGBoost trong các cuộc thi học máy, chẳng hạn như các cuộc thi trên nền tảng Kaggle, nhờ khả năng khai thác tốt cả dữ liệu dạng bảng lẫn khả năng mở rộng mô hình [10].

XGBoost xây dựng mô hình theo cách lặp, trong đó mỗi cây quyết định mới được huấn luyện để giảm thiểu sai số (residual) của tổ hợp các cây trước đó theo tỷ lệ học η (learning rate) (Hình 2).



Hình 2. Sơ đồ khối quy trình huấn luyện của XGBoost

2.2. Dữ liệu và xử lý dữ liệu

Để tiến hành nhận dạng lũ lớn đến Hồ Sơn La, nghiên cứu tiến hành xây dựng mô hình học máy nhận dạng dòng chảy lũ đến Trạm thủy văn Tạ Bú, nằm ngay sau hồ. Mô hình dự báo tại trạm Tạ Bú được thiết lập thông qua số liệu quan trắc dòng chảy tự nhiên đến trạm và dữ liệu mưa ngày trong các đợt mưa, lũ lớn từ năm 1961 đến 2009, là thời điểm trước khi Hồ thủy điện Sơn La đi vào vận hành (Bảng 1). Các dữ liệu này được tổng hợp thành 3 nhóm dữ liệu bao gồm lũ sớm, lũ chính vụ, lũ muộn. Trong đó dữ liệu nhận dạng lũ sớm là các trận lũ xảy ra trong khoảng từ 15/6 đến 19/7, lũ chính vụ từ 20/7 đến 21/8, lũ muộn từ 22/8 đến 15/9.

Có thể thấy rằng dữ liệu có được từ năm 1961 đến 2009 về cơ bản là đầy đủ nhưng để phục vụ xây dựng mô hình học máy thì 49 dữ liệu là quá ít. Ngoài ra, tập dữ liệu bao gồm các loại dữ liệu khác nhau với biên độ thay đổi khác nhau. Ví dụ, dữ liệu lưu lượng chân lũ sớm biến thiên từ 248-2070, dữ liệu mưa 1 ngày lớn nhất của Tuần Giáo từ 27-189,... Bên cạnh đó, tập dữ liệu bao gồm các “cụm biến đầu vào” là mưa 1, 3, 5 ngày lớn nhất của các trạm đo mưa lân cận nhau nên có thể có tương quan cao. Việc sử dụng dữ liệu thô, chưa qua xử lý sẽ có thể dẫn đến thất bại trong việc học của mô hình cũng như đưa ra kết quả nhận dạng không chính xác. Chính vì vậy, nghiên cứu tiến hành xử lý dữ liệu dùng để học máy theo các bước sau.

Bước 1: Sàng lọc các biến có tương quan tốt và độc lập với nhau để nhận dạng lũ Trạm thủy văn Tạ Bú. Chỉ tiêu để sàng lọc là hệ số tương quan Pearson và được thực hiện bằng công cụ Data Analysis trong Excel 2019. Kết quả thu được các biến quan trọng đóng góp vào nhận dạng đỉnh lũ đến Trạm thủy văn Tạ Bú (Bảng 1).

Bước 2: Chuẩn hóa dữ liệu sử dụng các kỹ

thuật xây dựng biến tương tác giữa các trạm khí tượng, thêm biến thống kê tổng hợp và dùng biến phi tuyến (logarithm hoặc căn bậc 2) để làm mượt giá trị lớn và sử dụng thuật toán chuẩn hóa để loại bỏ các giá trị ngoại lai làm nhiễu quá trình học máy. Tất cả các kỹ thuật này được đưa vào bộ mã XGBoost lập trình bằng ngôn ngữ Python.

Bảng 1. Dữ liệu khí tượng thủy văn sử dụng trong nhận dạng lũ đến trạm Tạ Bú

Loại dữ liệu		Lũ sớm	Lũ chính vụ	Lũ muộn
Thủy văn	Lưu lượng đỉnh lũ	Tạ Bú	Tạ Bú	Tạ Bú
	Lưu lượng chân lũ	Tạ Bú	Tạ Bú	Tạ Bú
Khí tượng	Mưa 3 ngày lớn nhất	Điện Biên Mường Tè	Mường Tè Tam Đường	Điện Biên Mường Tè Sìn Hồ
	Mưa 5 ngày lớn nhất	Tuần Giáo Lai Châu Sìn Hồ Tam Đường Quỳnh Nhai Than Uyên Mù Cang Chải	Điện Biên	Tuần Giáo Lai Châu Tam Đường Quỳnh Nhai Than Uyên Mù Cang Chải

2.3. Xây dựng mô hình học máy XGBoost

Mô hình học máy XGBoost được xây dựng trên ngôn ngữ lập trình Python. Mô hình được xây dựng để học với tập dữ liệu gồm 49 số liệu với một số tính năng như sau:

- Sử dụng 80% dữ liệu để huấn luyện và 20% dữ liệu để kiểm tra;
- Tạo các đặc trưng mở rộng (feature engineering) như biến thống kê tổng hợp (max) và dùng biến phi tuyến (logarithm hoặc căn bậc 2);
- Chuẩn hóa dữ liệu bằng thuật toán RobustScaler;
- Sử dụng thuật toán Randomized Search Cross-validation để tối ưu hóa bộ thông số mô hình XGBoost;
- Chạy XGBoost 1000 lần với random split nhằm giảm thiểu tính ngẫu nhiên của mô hình học máy;
- Lựa chọn mô hình tốt nhất dựa trên kết quả tính toán hệ số tương quan R^2 của tập kiểm tra (test) với ràng buộc sai khác R^2 giữa tập huấn luyện (train) và kiểm tra $\Delta R^2 < 0,3$. Trong đó R^2

được tính theo công thức sau:

$$R^2 = 1 - \frac{\sum (DB - TĐ)^2}{\sum (TĐ - TB)^2}$$

Trong đó: *TĐ* dùng để chỉ số liệu lưu lượng thực đo tại trạm Tạ Bú (m³/s); *TB* là lưu lượng trung bình (m³/s); *DB* là lưu lượng dự báo từ mô hình XGBoost (m³/s).

3. Kết quả nhận dạng lũ đến Hồ Sơn La

3.1. Kết quả huấn luyện và kiểm tra dòng chảy đến trạm Tạ Bú

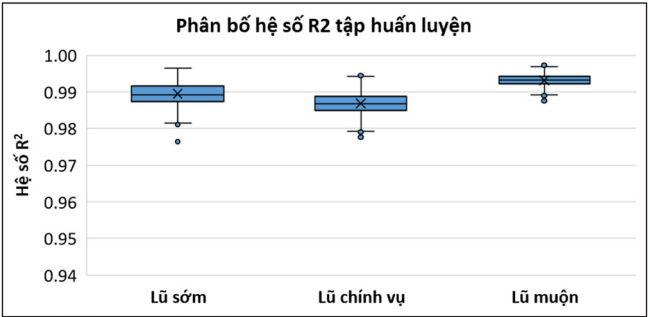
Mô hình học máy XGBoost cho kết quả huấn luyện và kiểm tra lũ sớm với chuỗi dữ liệu ngắn, gồm 49 số liệu, khá tốt với cả nhận dạng lũ sớm, lũ chính vụ và lũ muộn. Kết quả học sâu 1.000 lần chạy cho thấy hệ số tương quan R^2 của tập huấn luyện thường đạt trên 97% (Hình 3). Trong đó, huấn luyện nhận dạng lũ muộn cho kết quả ổn định và tốt nhất, trong khi huấn luyện nhận dạng lũ chính vụ cho kết quả kém hơn.

Để tránh xảy ra tình trạng “quá khớp” trong học máy khi kết quả huấn luyện cho độ chính

xác rất cao nhưng kết quả kiểm tra lại rất kém, nghiên cứu sử dụng điều kiện ràng buộc $\Delta R^2 < 0,3$. Kết quả thiết lập mô hình XGBoost được thể hiện trong Bảng 2 và Hình 4, Hình 5.

Như vậy, có thể thấy rằng với mô hình XGBoost được huấn luyện rất tốt ($R^2 > 0,97$) thì kết quả kiểm tra cho hệ số R^2 đạt trên 70% với sai số tuyệt đối trung bình nhỏ hơn 15% giá trị

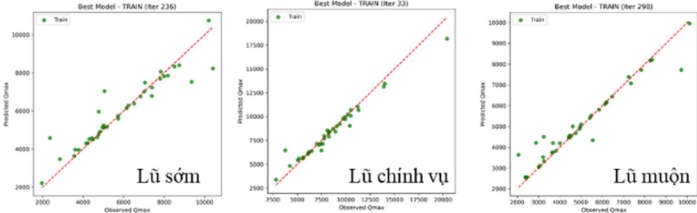
lưu lượng đỉnh lũ trung bình. Đối với tập kiểm tra của lũ muện, kết quả tính toán từ mô hình cho kết quả thiên lớn, tuy nhiên, sai số này vẫn ở trong khoảng cho phép (trung bình 13,5%). Như vậy, có thể thấy rằng mô hình XGBoost đã cho kết quả huấn luyện và kiểm tra rất tốt với chuỗi dữ liệu ngắn trong nhận dạng lũ đến trạm Tạ Bú.



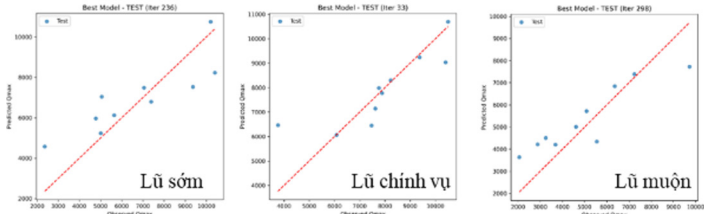
Hình 3. Kết quả huấn luyện XGBoost nhận dạng lũ đến trạm Tạ Bú 1.000 lần

Bảng 2. Kết quả thiết lập mô hình học máy XGBoost tốt nhất

Kết quả thiết lập		Lũ sớm	Lũ chính vụ	Lũ muện
Hiệu suất mô hình	R^2 tập train	0,99	0,98	0,99
	R^2 tập test	0,79	0,71	0,77
	MAE tập test (m ³ /s)	874,8	630,9	702,1
Thông số chính mô hình XGBoost	colsample_bytree	0,644	0,800	0,644
	gamma	2,222	None	4,444
	learning_rate	0,254	0,05	0,117
	max_depth	4	5	9
	subsample	0,644	0,800	0,689



Hình 4. Kết quả huấn luyện mô hình XGBoost nhận dạng lũ đến trạm Tạ Bú



Hình 5. Kết quả kiểm tra mô hình XGBoost nhận dạng lũ đến trạm Tạ Bú

3.2. Kết quả dự báo thử nghiệm dòng chảy đến Hồ Sơn La

Do Hồ thủy điện Sơn La đã đi vào hoạt động nên để kiểm định lại mô hình nhận dạng lũ đã xây dựng, nghiên cứu tiến hành dự báo thử nghiệm dòng chảy lũ đến Hồ Sơn La trong mùa lũ năm 2024. Do diện tích khống chế đến Trạm thủy văn Tạ Bú (45.900 km²) và Hồ thủy điện Sơn

La (43.760 km²) có sự chênh lệch không đáng kể (khoảng 5%) nên dòng chảy lũ đến Hồ Sơn La (Qdb_SL) sẽ được tính từ dòng chảy dự báo đến Tạ Bú (Qdb_TB) với tỷ lệ diện tích (k=0,95). Trong đó Qdb_TB được nhận dạng từ mô hình XGBoost đã xây dựng. Kết quả nhận dạng đỉnh lũ các thời kỳ trong năm 2024 tại Hồ Sơn La được thể hiện trong Bảng 3.

Bảng 3. Kết quả dự báo thử nghiệm mùa lũ năm 2024 đến Hồ Sơn La

Hiệu suất mô hình	Lũ sớm	Lũ chính vụ	Lũ muộn
Lưu lượng đỉnh lũ đến Hồ Sơn La	3.652	8.712	6.256
Lưu lượng đỉnh lũ dự báo	4.161	8.963	5.394
Sai số (%)	13,9	2,9	13,8

Kết quả nhận dạng thử nghiệm cho thấy, mô hình XGBoost dự báo đỉnh lũ đến Hồ Sơn La cho các thời kỳ khác nhau khá tốt. Trong đó mô hình nhận dạng tốt nhất ở thời kỳ lũ chính vụ với sai số dưới 5%. Trong thời kỳ lũ sớm và lũ muộn, sai số đỉnh lũ trung bình ngày dưới 15%.

4. Kết luận

Như vậy, bài báo đã nhận dạng được dòng chảy lũ đến Hồ Sơn La thông qua xây dựng được mô hình nhận dạng đỉnh lũ đến Trạm thủy văn Tạ Bú. Mô hình học máy XGBoost thể hiện được hiệu suất mạnh mẽ trong việc huấn luyện, kiểm tra và kiểm định trong các thời kỳ lũ sớm, lũ

chính vụ và lũ muộn với chuỗi dữ liệu ngắn và các biến có quan hệ phi tuyến phức tạp. Kết quả huấn luyện cho thấy hiệu suất mô hình luôn đạt trên 97% với 1.000 lần chạy khác nhau. Với tập kiểm tra, hiệu suất mô hình cũng đạt tới trên 70%. Kết quả dự báo thử nghiệm cho Hồ Sơn La với mùa lũ năm 2024 cũng cho sai số đỉnh lũ giữa dự báo và thực đo dưới 5% với lũ chính vụ và dưới 15% với lũ sớm và lũ muộn.

Từ việc nhận dạng được đỉnh lũ của các kỳ lũ khác nhau đến Hồ Sơn La là cơ sở nhận dạng lũ lớn tạo sự chủ động trong phòng chống thiên tai và từ có xây dựng được kế hoạch vận hành đảm bảo hiệu quả sử dụng nước.

Đóng góp của các tác giả: Xây dựng ý tưởng, phân tích kết quả và viết chính bài báo: Lương Hữu Dũng; Xây dựng code python, phân tích kết quả và tham gia viết bài báo: Ngô Thị Thủy; Tham gia viết code, xây dựng ý tưởng: Lương Tuấn Trung; Xây dựng ý tưởng và chuẩn bị dữ liệu đầu vào: Trịnh Thu Phương.

Lời cảm ơn: Nghiên cứu này là sản phẩm của Đề tài “Nghiên cứu khả năng chống chịu của hệ thống công trình phòng chống lũ, xây dựng công nghệ nhận dạng lũ lớn, tổ hợp lũ bất lợi trên lưu vực Sông Hồng và đề xuất giải pháp ứng phó” mã số ĐTĐL.CN-47/22.

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

Tài liệu tham khảo

1. Trần Thanh Xuân (2016), *Mạng lưới và tài nguyên nước sông Việt Nam những biến đổi và thách thức*, Nhà xuất bản Khoa học và Kỹ thuật.

2. Bộ Tài nguyên và Môi trường (2020), *Quyết định 1757/QĐ-BTNMT ngày 11 tháng 8 năm 2020, Về việc ban hành danh mục nguồn nước liên tỉnh và danh mục nguồn nước liên quốc gia (nguồn nước mặt)*.

3. Trịnh Thu Phương (2021), *Nghiên cứu cơ sở khoa học và thực tiễn nhận dạng lũ lớn trên lưu vực Sông Hồng góp phần nâng cao hiệu quả điều hành hệ thống liên hồ chứa*, Luận án Tiến sĩ Thủy văn, Viện Khoa học Khí tượng thủy văn và Biến đổi khí hậu.

4. Đặng Ngọc Tĩnh (2001), *Nghiên cứu Xây dựng phương án dự báo hạn ngắn lũ miền Trung trên máy vi tính*, Đề tài nghiên cứu khoa học công nghệ cấp Tổng cục Khí tượng thủy văn.
5. Trịnh Thu Phương (2017), *Nghiên cứu xây dựng công nghệ nhận định lũ lớn và dòng chảy mùa cạn trên lưu vực Sông Hồng nhằm nâng cao hiệu quả vận hành liên hồ chứa*, Đề tài nghiên cứu khoa học công nghệ cấp Bộ.
6. Bùi Đình Lập (2016), *Nghiên cứu xây dựng công nghệ dự báo dòng chảy lũ đến các hồ chứa trên hệ thống Sông Hồng*, Đề tài nghiên cứu khoa học công nghệ cấp Bộ.
7. Lương Hữu Dũng (2019), *Nghiên cứu xây dựng công nghệ dự báo tài nguyên nước mặt phục vụ lập kế hoạch sử dụng nước cho lưu vực sông Ba và sông Kôn ở Nam Trung Bộ và Tây Nguyên*, Báo cáo tổng kết Đề tài.
8. Chen, T., & Guestrin, C. (2016), "XGBoost: A scalable tree boosting system", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. <https://doi.org/10.1145/2939672.2939785>.
9. Friedman, J. H. (2001), "Greedy function approximation: A gradient boosting machine", *Annals of Statistics*, 29(5), 1189-1232. DOI:10.1214/aos/1013203451.
10. Nielsen, D. (2016), "Tree boosting with XGBoost - Why does XGBoost win "every" machine learning competition?", *NTNU Data Mining Report*, 1-13.

APPLICATION OF MACHINE LEARNING FOR LARGE FLOOD IDENTIFICATION IN THE DA RIVER

Luong Huu Dung⁽¹⁾, Ngo Thi Thuy⁽¹⁾, Luong Tuan Trung⁽¹⁾, Trinh Thu Phuong⁽²⁾

⁽¹⁾The Viet Nam Institute of Meteorology, Hydrology and Climate Change

⁽²⁾Mekong River Commission

Received: 11/5/2025; Accepted: 30/5/2025

Abstract: Severe flood identification plays a critical role in disaster prevention and the sustainable development of river basins. In recent years, with the rapid advancement of computer science, numerical methods such as mathematical modeling and artificial intelligence have been widely applied in studies related to flood forecasting and the identification of flood characteristics in the Red River basin. However, developing a machine learning model capable of accurately identifying and forecasting major floods remains challenging due to limitations in data availability and complex nonlinear relationships among variables. This study aims to develop a robust machine learning algorithm capable of identifying and predicting flood peak discharges to the Son La reservoir using the Ta Bu station during severe flood events, using a limited dataset (49 records) with discontinuous time series. The advanced machine learning model XGBoost, along with input data preprocessing techniques, was employed. The training results achieved over 97% accuracy across all 1.000 runs, while the testing accuracy reached approximately 70%. Model-based forecasts for the 2024 flood season also showed promising performance in predicting the major flood peak, with an error margin of less than 5%. The model's performance in identifying early and late flood peaks was acceptable, with an average daily peak error of around 14%.

Keywords: Flood identification, flood forecast, Da River, machine learning, XGBoost.