ỨNG DỤNG MÔ HÌNH CÂY QUYẾT ĐỊNH DỰ BÁO DÒNG CHẢY TRÊN LƯU VỰC SÔNG TIÊN YÊN, QUẢNG NINH

Nguyễn Tiến Thái[1*], Trần Tuấn Thạch[1]

Tóm tắt: *Mô phỏng dự báo dòng chảy trên các lưu vực sông đóng vai trò vô cùng quan trọng trong lĩnh vực quản lý tài nguyên nước, góp phần đưa ra các quyết định về phân phối dòng chảy và kiểm soát các hình thái lũ cực đoan một cách hiệu quả. Trong nghiên cứu này, nhóm tác giả sử dụng các mô hình học máy (Machine learning-ML) dựa trên thuật toán cây quyết định (Decision Trees-DTs) bao gồm 4 mô hình: Cây quyết định (Decision Tree-DT), Rừng ngẫu nhiên (Random Forest-RF), Tăng cường độ dốc nhẹ (Light Gradient Boosting Machine-LightGBM), Cây quyết định tăng cường độ dốc (Gradient Boosting Decision Tree-GBDT) để mô phỏng dự báo cho dòng chảy trên lưu vực sông Tiên Yên, Quảng Ninh tại trạm Bình Liêu. Bộ thông số của các mô hình được tối ưu hóa dựa trên thuật toán tối ưu tìm kiếm ngẫu nhiên (RandomizedSearchCV). Kết quả đào tạo mô hình cho thấy các mô hình cây quyết định đều cho hiệu suất mô hình tương đối chính xác, trong đó mô hình GBDT cho kết quả tốt nhất (R^2=0,89; NSE=0,91; KGE=0,89 và RMSE=5,91 m^3/s), tiếp theo lần lượt là các mô hình LightGBM, RF và DT. Ở giai đoạn kiểm định hai mô hình GBDT và LightGBM tiếp tục cho thấy mức độ hiệu suất mô hình tương đương nhau, trong khi RF và DT có kết quả thấp hơn. Dựa trên các mô hình đã được đào tạo và kiểm định, dòng chảy giai đoạn (2020-2024) được dự báo dựa trên số liệu khí tượng của trạm Tiên Yên và Bình Liêu.*

Từ khóa: Mô hình cây dữ liệu, dòng chảy lưu vực, tối ưu hóa mô hình, dự báo dòng chảy.

1. GIỚI THIỆU

Ước tính dự báo dòng chảy trên các lưu vực sông đóng vai trò quan trọng trong quản lý tài nguyên nước một cách hiệu quả, đặc biệt trong công tác dự báo dòng chảy lũ, dòng chảy mùa kiệt phục vụ sản xuất nông nghiệp, sản xuất điện, ...vv. Ứng dụng các mô hình thủy văn vật lý truyền thống được sử dụng rộng rãi nhằm ước tính dòng chảy dựa trên các đặc trưng của lưu vực và điều kiện khí tượng thủy văn. Tác giả (Maryam Hafezparast et al. 2018) sử dụng mô hình MIKE NAM để mô phỏng dòng chảy trên lưu sông Sarisoo ở phía Tây Bắc Iran với số liệu đầu vào là mưa và bốc hơi cùng với bộ 9 tham số mô hình đặc trưng cho lưu vực nghiên cứu đã cho kết quả với độ tin cậy cao. Tương tự, tác giả (Sangam Shrestha et al. 2018) sử dụng mô hình SWAT sử dụng dữ liệu đầu vào là các yếu tố khí tượng (lượng mưa, nhiệt độ min và max, độ ẩm tương đối (RH), bức xạ mặt trời và tốc độ gió) và các tài liệu (sử dụng đất, lớp phủ đất, tính chất đất, và địa hình) để ước tính dòng chảy. Tuy vậy, việc ứng mô phỏng dòng chảy trên lưu vực gặp rất nhiều khó khăn và thách thức do dòng chảy là một phần của chu trình thủy văn phức tạp chịu tác động của nhiều yếu tố, đặc biệt là đặc trưng của lưu vực như địa hình, tính chất của đất, lớp thảm phủ, ...vv. Bên cạnh đó, ảnh hưởng của các hoạt động kinh tế trong việc khai thác sử dụng đất đã làm thay đổi thảm phủ cũng làm quá trình mô phỏng dòng chảy thêm phức tạp.

Với các khó khăn trong việc ứng dụng mô hình thủy văn dựa trên phương trình vật lý, các mô hình học máy (Machine learning) và học sâu (Deep learning) nổi nên như là một công cụ mạnh mẽ để ước tính dòng chảy trên các lưu vực sông thông qua việc sử dụng dữ liệu lớn để học các mô hình phức tạp và các mối quan hệ phi tuyến trong dữ liệu thủy văn. Không giống như mô hình thủy văn truyền thống, phương pháp sử dụng học máy không yêu cầu mô phỏng phương trình vật lý rõ ràng, thay vào đó, chúng sử dụng chuỗi dữ liệu lịch sử để huấn luyện các thuật toán để tăng độ chính xác của dự đoán. Các mô hình tiếp cận dữ liệu sử dụng mạng thần kinh nhân tạo ANNs (Aiyelokun Oluwatobi et al. 2018), mạng thần kinh đơn vị định kỳ có kiểm soát (GRU), dạng cải tiến của mạng bộ nhớ ngắn dài (LSTM), mô hình cây quyết định (DTs), rừng ngẫu nhiên (RF) (Mohammad Ranjbar Kabootarkhani et al. 2024) đã cho thấy khả năng mô phỏng dự báo với hiệu suất mô hình có độ chính xác cao trong lĩnh vực thủy văn. Các mô hình được đào tạo vào kiểm định theo tỷ lệ dữ liệu nhất định. Nghiên cứu của tác giả (Phạm Văn Chiến et al. 2024) cho thấy đối với mô hình học máy dữ liệu có thể chia thành giai đoạn đào tạo (Training) (70%) và giai đoạn kiểm định (Testing) (30%) cho hiệu suất mô hình khá cao so với giá trị quan trắc. Mô hình cây dữ liệu chia dữ liệu đầu vào (mưa, nhiệt độ, bốc hơi, độ ẩm…vv) thành các tập con (node / leaf) sao cho các mẫu trong cùng một vùng có đầu ra (lưu lượng Q) gần giống nhau nhất. Các mô hình DTs khác cũng đã thể hiện kết quả tốt cho dự báo dòng chảy. Các nghiên cứu khác nhau đã đánh giá hiệu quả của các mô hình DTs như: Học tăng cường độ dốc tới hạn (XGBoost) (Mahmoud F. Maghrebi, 2023), LightGBM

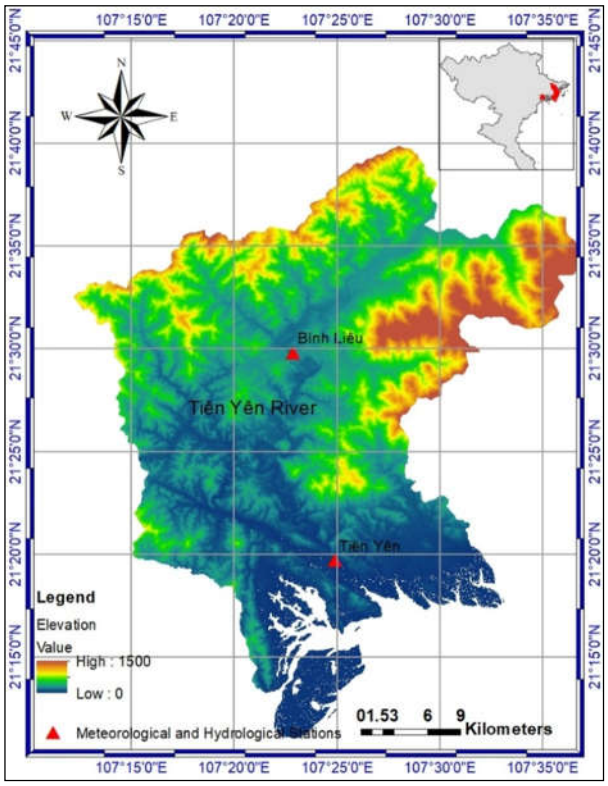
[1] *Trường Đại học Thủy lợi*
[*] *Tác giả liên hệ*

(Lekang Bian et al. 2023), GBDT (Ma et al. 2025) đều cho kết quả hiệu suất mô hình tương đối cao. Kết quả nghiên cứu cũng cho thấy mô hình đạt hiệu suất cao nhất với các bộ thông số của mô hình học máy được tối ưu hóa, thuật toán tìm kiếm ngẫu nhiên (RandomizedSearchCV) (J. Hancock et al. 2021) được tích hợp trong thư viện scikit-learn trên ngôn ngữ lập trình Python. Các nghiên cứu về tối ưu hóa bộ siêu tham số của mô hình dữ liệu góp phần cải thiện đáng kể độ chính xác của mô hình.

Trong nghiên cứu này, nhóm tác giả sử dụng các mô hình cây dữ liệu-DTs bao gồm: DT, RF, LightGBM, và GBDT để mô phỏng và dự báo dòng chảy trên lưu vực sông Tiên Yên, tỉnh Quảng Ninh. Bên cạnh đó, nghiên cứu so sánh đánh giá hiệu suất của các mô hình cây quyết định này cho lưu vực nghiên sử dụng thuật toán tối ưu xác định bộ siêu tham số cho mô hình.

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Đối tượng nghiên cứu

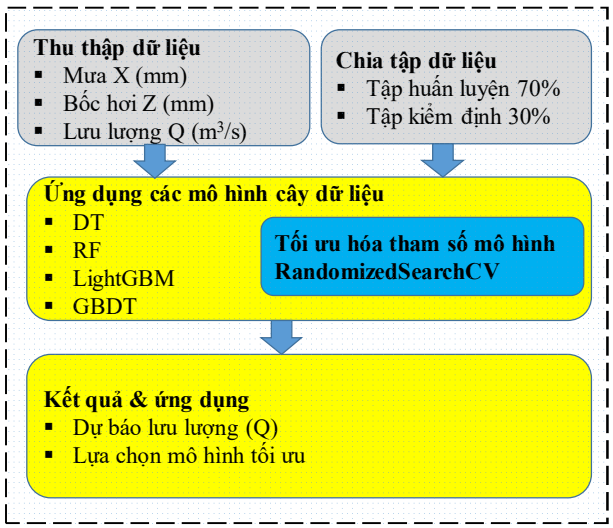


Hình 1. Lưu vực sông Tiên Yên, Quảng Ninh

Dòng chính sông Tiên Yên có chiều dài 68 km, thượng nguồn từ cửa khẩu Hoành Mô - Bình Liêu - Quảng Ninh, chảy qua thị trấn Bình Liêu, hợp với nhánh sông Tiên Mơ đổ vào và chảy xuôi xuống huyện Tiên Yên- Quảng Ninh. Bờ sông hai bên là đồi núi bao quanh hiểm trở, độ dốc bờ lớn, cấu tạo địa chất là đá, cuội sỏi, thảm thực vật chủ yếu là cây trồng như ngô, đậu tương và rừng keo trồng để lấy gỗ, lòng sông quanh co, uốn khúc. Độ rộng lòng sông trên thượng nguồn khoảng 50m-70m, xuôi xuống hạ nguồn độ rộng lớn hơn từ 80m đến 210m, độ sâu nước dao động từ 0,7-4m, vận tốc dòng chảy trung bình từ 0,01-1,2 m/s. Diện tích lưu vực của sông Tiên Yên khoảng 1.070 km^2, trong đó diện tích lưu vực tính từ trạm thủy văn Bình Liêu về phía thượng nguồn khoảng 370km^2.

2.2. Dữ liệu thu thập

Dữ liệu về khí tượng, thủy văn sử dụng trong nghiên cứu được thu thập bao gồm: mưa (1975-2024) tại trạm Bình Liêu và Tiên Yên, bốc hơi (1975-2024) tại trạm Tiên Yên, dòng chảy (1975-2019) tại trạm Bình Liêu. Dữ liệu lưu lượng, mưa và bốc hơi thu thập từ các trạm quan trắc được trải qua giai đoạn tiền xử lý để đảm bảo độ tin cậy: Các giá trị dữ liệu thiếu được xác định bằng việc nội suy; Giá trị ngoại lai được phát hiện qua phân tích thống kê Z-score và được hiệu chỉnh; Dữ liệu được chuẩn hóa bằng phương phương pháp chuẩn hóa dữ liệu (data standardization). Dữ liệu thu thập (**Bảng 1, Hình 3**) cho thấy lưu lượng bình quân của sông Tiên Yên trung bình khoảng 23,4 m^3/s. Lưu lượng đỉnh lũ lịch sử ghi nhận vào 26/9/2008 là 3260 m^3/s tương ứng với lượng mưa trong cùng thời gian tại trạm Bình Liêu (X=697mm), Tiên Yên (X=502,3mm). Dữ liệu mưa, bốc hơi (trạm Tiên Yên) và mưa (trạm Bình Liêu) được sử dụng như số liệu đầu vào để mô phỏng dòng chảy sông Tiên Yên tại trạm Bình Liêu. Số liệu từ 1/1/1975 đến 1/7/2006 được sử dụng cho quá trình đào tạo mô hình (70% dữ liệu Q), và từ 2/7/2006 đến 31/12/2019 cho quá trình kiểm định mô hình (30%).



Hình 2. Sơ đồ nghiên cứu mô phỏng dự báo dòng chảy lưu vực sông

2.3. Các mô hình cây quyết định

a) Mô hình cây quyết định- Decision Tree

Mô hình DT thuộc về lớp học có giám sát, sử dụng thuật toán hồi quy phi tuyến để mô phỏng, dự báo chuỗi dữ liệu theo thời gian. Dữ liệu sẽ được chia thành các tập con dựa trên đặc điểm của tập dữ liệu để hình thành cấu

trúc: Nút gốc (Root Node), Nút trong (Internal Nodes), cành (Branches), Nút lá (Leaf Node). Mô hình DT dạng hồi quy nhằm mục đích giảm thiểu phương sai trong mỗi tập con để xác định các phân chia tốt nhất. Giá trị phương sai của dữ liệu (MSE) được xác định như sau:

$$NSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}) \quad (1)$$

Trong đó:

y_i: Giá trị thực tế quan trắc

$\hat{y}$: Giá trị dự báo

b) Mô hình rừng ngẫu nhiên- Random Forest

RF là một mộ dạng của mô hình học tập hợp. thuật toán RF xây dựng nhiều cây DT và sau đó kết hợp kết quả của chúng lại, qua đó tăng cường mức độ chính xác và giảm tình trạng quá khớp (Overfitting) so với một DT đơn lẻ. Kết quả dự báo cuối cùng được lấy bằng trung bình của tất cả các cây riêng lẻ, được xác định theo phương trình sau:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t \quad (2)$$

Trong đó:

T: Số lượng cây hồi quy trong rừng

$\hat{y}_t$: Giá trị dự báo từ cây hồi quy thứ t

c) Mô hình cây quyết định tăng cường độ dốc - Gradient Boosting Decision Tree

GBDT là mô hình học tập hợp dựa trên cây quyết định (DT). Khác với mô hình RF đào tạo các DT độc lập và kết hợp dự đoán của chúng, GBM xây dựng lần lượt từng cây, với mỗi cây mới tập trung vào các trường hợp khó dự đoán nhất, do đó cải thiện được độ chính xác. Kết quả dự báo độ chính xác cuối cùng của mô hình GBDT theo dạng tổng quát sau:

$$y = \sum_{m=1}^{M}\alpha_m h_m \quad (3)$$

Trong đó:

M: số cây quyết định DTs

h_m: Mô hình cây thứ m (weak learner)

α_m Trọng số của mô hình con (DT)

d) Mô hình học tăng cường độ dốc nhẹ- Light Gradient Boosting Machine

LightGBM, một thuật toán học tăng cường khác, là ứng dụng hiệu quả của học tăng cường thiết kế cho bài toán hồi quy. Giống như mô hình GBDT, LightGBM xây dựng một tập hợp cây DTs theo cách tuần tự, mỗi cây sửa lỗi của cây trước đó. LightGBM là sử dụng thuật toán dự trên biểu đồ (histogram-based) nên tốc độ xử nhanh hơn. Cải tiến này cho phép LightGBM đào tạo nhanh hơn và mở rộng tốt hơn trong khi vẫn duy trì độ chính xác cao.

2.4 Đánh giá độ tin cậy của mô hình

Độ tin cậy của mô hình DTs được đánh giá dựa trên các chỉ số sai số không thứ nguyên (Coefficient of Determination (R^2); Kling-Gupta Efficiency (KGE), Nash-Sutcliffe Efficiency (NSE)) và sai số thống kê thứ nguyên (Root Mean Squared Error (RMSE)). Độ chính xác của mô hình được tính toán cho giai đoạn đào tạo và kiểm định, theo công thức:

$$KGE = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{sim}}{\sigma_{obs}} - 1\right)^2 + \left(\frac{Q_{sim}}{Q_{obs}} - 1\right)^2} \quad (4)$$

$$r = \frac{\sum_{i=1}^{N}(Q_{obs,i} - Q_{obs,m})(Q_{sim,i} - Q_{sim,m})}{\sqrt{\sum_{i=1}^{N}(Q_{obs,i} - Q_{obs,m})^2}\sqrt{\sum_{i=1}^{N}(Q_{sim,i} - Q_{sim,m})^2}} \quad (5)$$

$$NSE = 1 - \frac{\sum_{i=1}^{N}(Q_{sim,i} - Q_{obs,i})^2}{\sum_{i=1}^{N}(Q_{sim,m} - Q_{obs,m})^2} \quad (6)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Q_{obs,i} - Q_{sim,i})^2} \quad (7)$$

Trong đó Q_{obs} và Q_{sim} là giá trị lưu lượng quan trắc và giá trị tính toán; σ_{obs} và σ_{sim}độ lệch chuẩn của lưu lượng quan trắc và tính toán.

2.5. Tối ưu hóa bộ thông số của các mô hình

Trong nghiên cứu này, nhóm tác giả sử dụng thuật toán RandomizedSearchCV để tối ưu hóa thông số của mô hình, đây là thuật toán tìm kiếm ngẫu nhiên trong không gian siêu tham số (hyperparameter space). Bộ tham số tốt nhất của mô hình được xác định theo công thức sau:

$$\theta^* = argmax\ Score\ (\theta) \quad (8)$$
$$\theta \in S$$

Trong đó

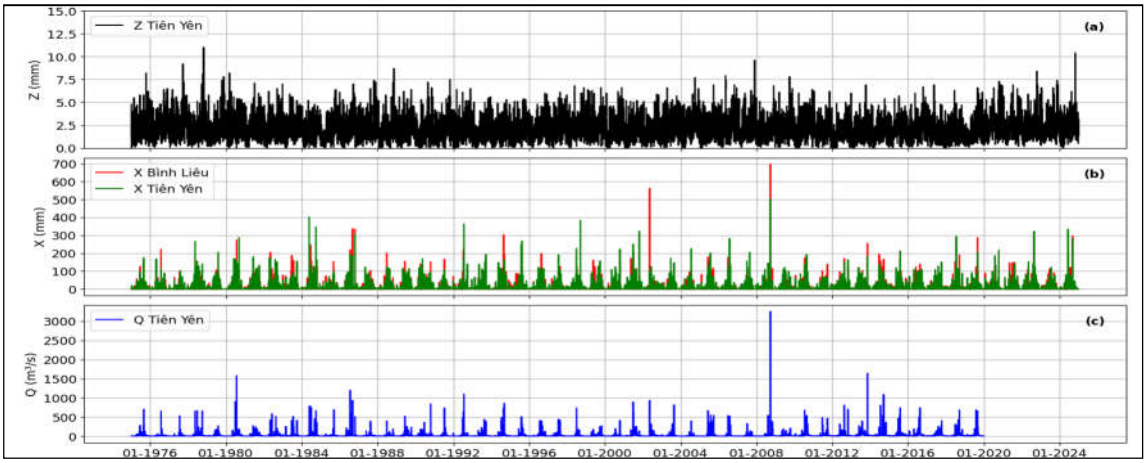
θ: Là bộ siêu tham số;

S: là tập hợp bộ tham số ngẫu nhiên từ không gian tham số

$Score\ (\theta)$: điểm đánh giá mô hình tính bằng cross-validation

Bảng 1. Dữ liệu bốc hơi, mưa, dòng chảy theo ngày thu thập ở các trạm khí tượng, thủy văn

Trạm đo	Chỉ số	Thời gian thu thập dữ liệu	Thông số thống kê			
			Min	Max	Trung bình	Độ lệch chuẩn
Tiên Yên	Bốc Hơi(mm)	1975-2024	0,0	11,0	2,2	1,2
Tiên Yên	Lượng mưa (mm)	1975-2024	0,0	502,3	6,0	20,0
Bình Liêu	Lượng mưa (mm)	1975-2024	0,0	697,0	5,3	19,0
Bình Liêu	Lưu lượng (m^3/s)	1975-2019	1,4	3260,0	23,4	65,9



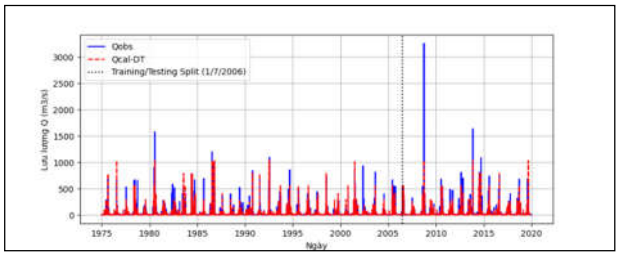
Hình 3. Bốc hơi (a), Lượng mưa (b), Dòng chảy (c) đo tại trạm Tiên Yên và Bình Liêu

Bảng 2. So sánh sai số thống kê khi sử dụng các mô hình cây quyết định, giai đoạn Training

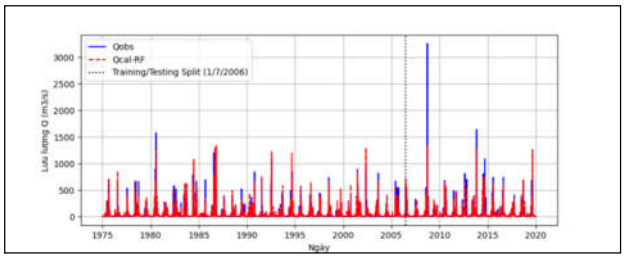
Mô hình	R	R^2	NSE	KGE	RMSE (m^3/s)
DT	0,84	0,69	0,69	0,65	39,27
RF	0,88	0,78	0,78	0,80	33,19
LightGBM	0,91	0,82	0,82	0,79	29,51
GBDT	0,94	0,89	0,91	0,89	5,91

Bảng 3. So sánh sai số thống kê khi sử dụng các mô hình cây quyết định, giai đoạn Testing

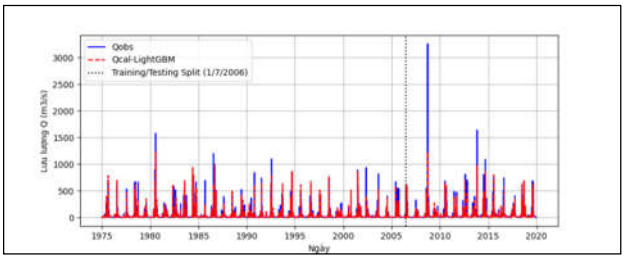
Mô hình	R	R^2	NSE	KGE	RMSE (m^3/s)
DT	0,77	0,59	0,59	0,64	34,70
RF	0,79	0,61	0,61	0,78	33,96
LightGBM	0,87	0,75	0,75	0,77	26,82
GBDT	0,88	0,75	0,75	0,70	26,93



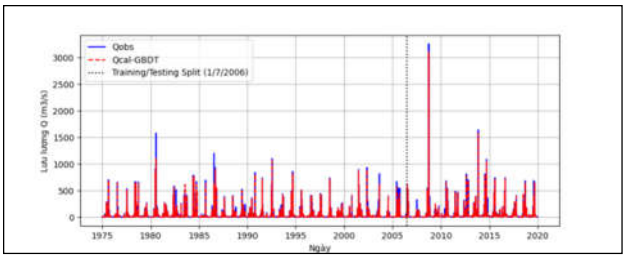
Hình 4. Mô phỏng dòng chảy sông Tiên Yên tại Bình Liêu, mô hình DT



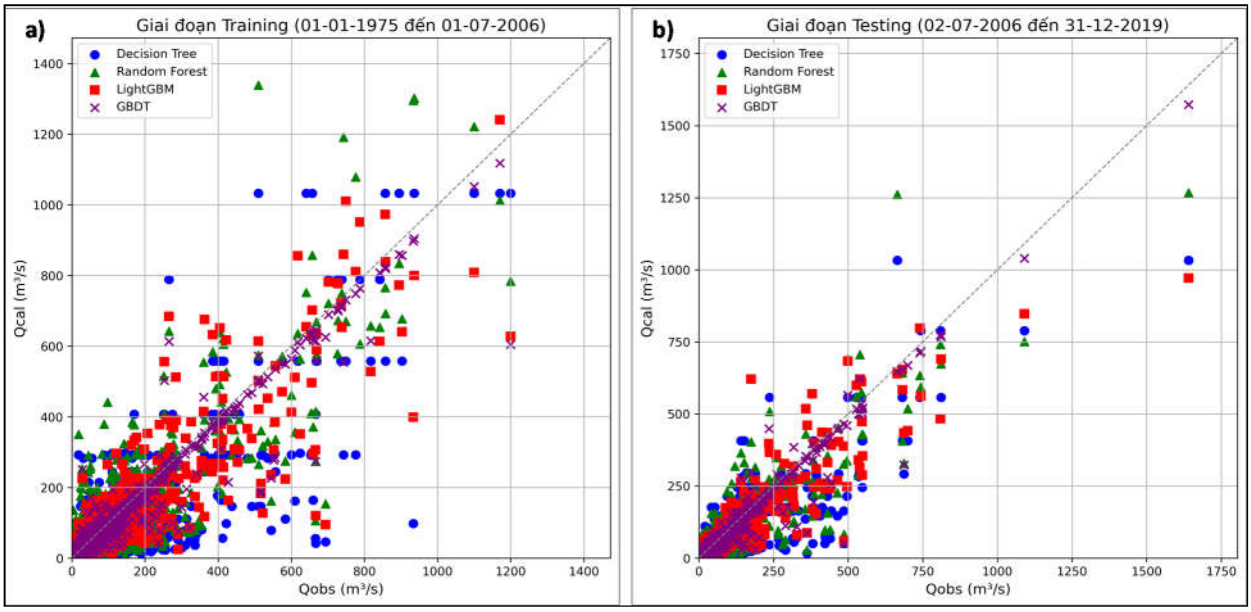
Hình 5. Mô phỏng dòng chảy sông Tiên Yên tại Bình Liêu, mô hình RF



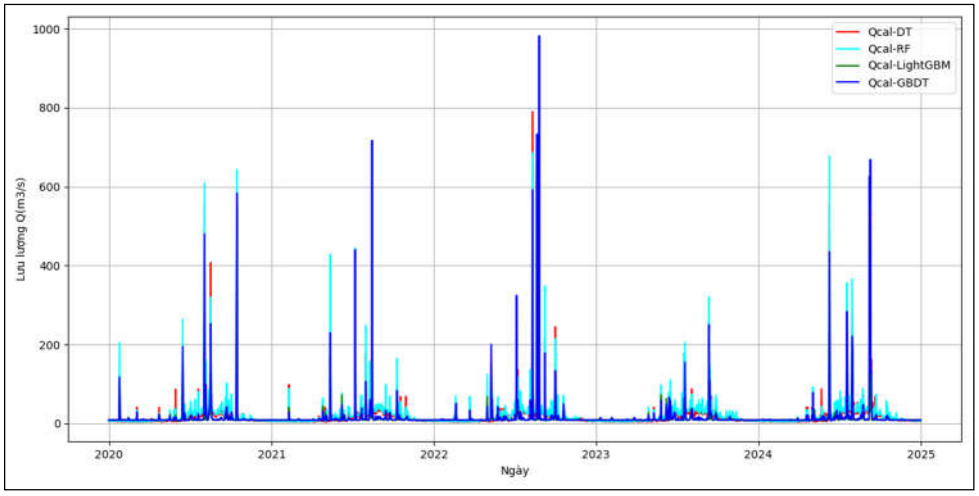
Hình 6. Mô phỏng dòng chảy sông Tiên Yên tại Bình Liêu, mô hình LightGBM



Hình 7. Mô phỏng dòng chảy sông Tiên Yên tại Bình Liêu, mô hình GBDT



Hình 8. Biểu đồ phân tán so sánh giữa lưu lượng quan trắc và mô phỏng sử dụng các mô hình DT, RF, LightGBM, và GBDT cho giai đoạn Đào tạo (Training) và Kiểm định (Testing)



Hình 9. Kết quả dự báo dòng chảy tại Bình Liêu, mô hình DT, RF, LightGBM, GBDT

Bảng 4. Thống kê kết quả dự báo dòng chảy tại Bình Liêu giai đoạn 2020-2024

Mô hình	Giai đoạn Dự báo	Lưu lượng Q (m^3/s)		
		Q_{min}	Q_{max}	Q_{tb}
DT	2020-2024	3,68	833,09	19,70
RF	2020-2024	4,18	818,38	21,30
LightGBM	2020-2024	5,36	686,14	15,94
GBDT	2020-2024	5,75	981,86	17,01

3. KẾT QUẢ NGHIÊN CỨU

Phương pháp tối ưu hóa tìm kiếm ngẫu nhiên - RandomizedSearchCV được sử dụng cho việc xác định giá trị siêu tham số tối ưu sao cho cho kết quả mô phỏng là tốt nhất cho các mô hình. Độ chính xác của mô hình được đánh giá qua các sai số thống kê (R^2, NSE, KGE, RMSE) (**Bảng 2, 3**). Kết quả mô phỏng cho thấy các mô hình đều cho hiệu suất khá cao với hệ số (R^2>0,59; NSE>0,59; KGE>0,64; RMSE<39,27 (m^3/s)). Trong giai đoạn đào tạo (1/1/1975 đến 1/7/2006), mô hình GBDT cho kết quả mô phỏng dòng chảy tốt nhất với các chỉ số đánh giá độ tin cậy của mô hình (R^2=0,89; NSE=0,91; KGE=0,89; RMSE=5,91m^3/s), tiếp theo là mô hình LightGBM, và RF. Mô hình DT cho kết quả mô phỏng với hiệu suất thấp nhất.

Tương tự, trong giai đoạn kiểm định (2/7/2006 đến 31/12/2019), mô hình GBDT cho kết quả tốt nhất với các chỉ số đánh giá độ tin cậy của mô hình (R^2=0.75; NSE=0,75; KGE=0,7; RMSE=26,93 m^3/s). Mô hình LightGBM cho kết quả mô phỏng với hiệu suất gần như mô hình GBDT. Mô hình DT và RF cho kết quả hiệu suất mô hình thấp hơn so với hai mô hình GBDT và LightGBM.

Các mô hình sau khi được đào tạo và kiểm định được sử dụng để dự báo dòng chảy sông Tiên Yên tại trạm Bình Liêu cho giai đoạn từ 1/1/2020 đếm 31/12/2024 sử dụng dữ liệu mưa và bốc hơi tại trạm Bình Liêu và Tiên Yên trong giai đoạn tương ứng. Kết quả mô phỏng dòng chảy được thể hiện trong (**Hình 4, 5, 6, 7, 8**) và (**Bảng 4)**. Các mô hình cho thấy sự tương đồng về giá trị lưu lượng trung bình, dao động (15,94 đến 21,30 m^3/s) và giá trị lưu lượng nhỏ nhất (Q_{min}=3,68 đến 5,75 m^3/s), trong khi khác biệt rất lớn về giá trị lớn nhất (Q_{max}) được ghi nhận. Mô hình DT và RF cho giá trị Q_{max} tương đồng nhau, trong khi đó mô hình GBDT cho giá trị Q_{max} cao hơn đáng kể so với mô hình LightGBM. Độ chính xác trong dự báo của mỗi mô hình phản ánh thuật toán khác nhau của mô hình cây quyết định GBDT và LightGBM học tuần tự từ lỗi, nắm quan hệ phi tuyến và cực trị, nên dự báo dòng chảy lưu vực có độ tin cậy cao, trong khi các mô hình DT và RF là tập hợp các cây đơn lẻ hoặc được tính từ sai số trung bình của các cây đơn lẻ nên độ chính xác kém hơn.

4. KẾT LUẬN

Trong nghiên cứu này, nhóm tác giả ứng dụng bốn mô hình cây quyết định DTs (DT, RF, LightGBM, và GBDT) để mô phỏng dự báo dòng chảy trên lưu vực sông Tiên Yên tại trạm Bình Liêu. Dữ liệu đầu vào được sử dụng bao gồm mưa và bốc hơi tại trạm Tiên Yên và Trạm Bình Liêu, được chia thành 2 giai đoạn, đào tạo (1/1/1975-1/7/2006) và kiểm định (2/7/2006-31/12/2019. Bộ siêu tham số của các mô hình DTs được tối ưu hóa dựa trên thuật toán tìm kiếm ngẫu nhiên RandomizedSearchCV. Kết quả đào tạo và kiểm định cho thấy, hiệu suất mô phỏng của các mô hình DTs có độ tin cậy cao với các chỉ số (R^2>0,59; NSE>0,59; KGE>0,64; RMSE<39,27 m^3/s). Trong đó mô hình GBDT cho độ chính xác cao nhất, tiếp theo là mô hình LightGBM, RF, và DT trong giai đoạn đào tạo. Mô hình GBDT và LightGBM cho kết quả độ chính xác khá tương đồng trong giai đoạn kiểm định, trong khi đó mô hình DT và RF cho kết quả chính xác thấp hơn. Dựa trên dữ liệu về mưa và bốc hơi trạm Tiên Yên và Bình Liêu giai đoạn 2020-2024, các mô hình DTs được sử dụng để dự báo dòng chảy cho giai đoạn này.

TÀI LIỆU THAM KHẢO

Phạm Văn Chiến, Nguyễn Hoàng Bách (2024), *Mô hình bộ nhớ dài - ngắn LSTM cho mô phỏng dòng chảy lưu vực sông Thu Bồn.* Tạp chí Khoa học và Công nghệ Thủy lợi số 83 – 2024

Hafezparast, M., Araghinejad, S., Fatemi, S. E., & Bressers, J. T. A. (2013). *A Conceptual Rainfall-Runoff Model Using the Auto-Calibrated NAM Models in the Sarisoo Rive*r. Hydrology: current research, 4(1), -. Article 148. https://doi.org/10.4172/2157-7587.1000148

J. Hancock and T. M. Khoshgoftaar (2021), *"Impact of Hyperparameter Tuning in Classifying Highly Imbalanced Big Data"* 2021 IEEE 22nd International Conference on Information Reuse and Integration for Data *Science (IRI)*, Las Vegas, NV, USA, pp. 348-354, doi: 10.1109/IRI51335.2021.00054.

Lekang Bian, Xueer Qin, Chenglong Zhang, Ping Guo, Hui Wu (2023), *Application, interpretability and prediction of machine learning method combined with LSTM and LightGBM-a case study for runoff simulation in an arid area*, Journal of Hydrology, Volume 625, Part B, 2023, 130091, ISSN 0022-1694, https://doi.org/10.1016/j.jhydrol.2023.130091.

Mahmoud F. Maghrebi & Sajjad M. Vatanchi (2023). *Comparison of different machine learning methods in river streamflow estimation using isovel contours and hydraulic variables*. International Journal of River Basin Management, https://doi.org/10.1080/15715124.2023.2245809

Ma, W., Zhang, X., Xie, J. et al (2025). *Prediction of non-stationary daily streamflow series based on ensemble learning: a case study of the Wei River Basin, China*. Stoch Environ Res Risk Assess 39, 509–529. https://doi.org/10.1007/s00477-024-02877-y

Mohammad Ranjbar Kabootarkhani, Soudabeh Golestani Kermani, Ammar Aldallal, Mohammad Zounemat-Kermani (2024). *Forecasting river daily discharge using decision tree and time series methods.* Proceedings of the Institution of Civil Engineers - Water Management 1 October 2024; 177 (5): 294–307. https://doi.org/10.1680/jwama.22.00079

Abstract:

APPLICATION OF DECISION TREE MODELS FOR STREAMFLOW FORECASTING IN THE TIEN YEN RIVER BASIN, QUANG NINH

The forecasting of streamflow within river basins holds paramount significance in the discipline of Hydrology and Water Resources Management, enabling the formulation of rational strategies for flow distribution and the mitigation of flood events. This study employs data-driven machine learning models, specifically those rooted in decision tree algorithms, including the Decision Tree (DT), Random Forest (RF), Light Gradient Boosting Machine (LightGBM), and Gradient Boosting Decision Tree (GBDT) to forecast streamflow for the Tien Yen River Basin, Quang Ninh province, at the Binh Lieu gauging station. The model's input data comprised meteorological variables, including rainfall data from the Binh Lieu station, as well as rainfall and evaporation data collected at the Tien Yen station from 1975 to 2024. The models' hyperparameters underwent rigorous optimization utilizing the Randomized Search Cross-Validation algorithm. The results from the training phase consistently indicate that all decision tree models achieved robust performance, with the GBDT model exhibiting the highest calibration efficiency (R^2 = 0.89, NSE = 0.91, KGE = 0.89, and RMSE = 5.91 m^3/s), followed in rank by the LightGBM, RF, and DT models. Performance assessment during the validation phase revealed that the GBDT and LightGBM models exhibited comparable predictive skill, preceding the RF and DT models. Based on the optimized and validated model results, the streamflow for the period (2020-2024) was forecasted utilizing the observed antecedent meteorological data from the Tien Yen and Binh Lieu stations.

Keywords: Decision tree model, streamflow, model optimization, streamflow forecasting.

Ngày nhận bài: 14/10/2025

Ngày chấp nhận đăng: 26/11/2025