NGHIÊN CỨU ỨNG DỤNG MÔ HÌNH HYBRID (HYPE - RF) TRONG HIỆU CHỈNH DÒNG CHẢY VỀ HỒ CHỨA PLEIKRONG

Vũ Văn Lân[1, 2*], Vũ Minh Cát[2], Bùi Du Dương[3]

Tóm tắt: *Nghiên cứu này đề xuất phương pháp kết hợp giữa mô hình thủy văn HYPE và thuật toán học máy Random Forest (RF) nhằm hiệu chỉnh sai số dòng chảy mô phỏng tại lưu vực hồ chứa Pleikrong, tỉnh Kon Tum. Mô hình HYPE được sử dụng để mô phỏng dòng chảy theo bước thời gian ngày và giờ, dựa trên dữ liệu khí tượng, thủy văn và viễn thám. Sai số giữa dòng chảy mô phỏng và thực đo được huấn luyện bởi mô hình RF để tối ưu kết quả đầu ra. Kết quả cho thấy mô hình lai HYPE–RF cải thiện đáng kể độ chính xác mô phỏng, với các chỉ số NSE, CC và KGE đều tăng rõ rệt. Cụ thể, NSE tăng từ 0.54 lên 0.80 đối với chuỗi dòng chảy ngày, và từ 0.45 lên 0.83 đối với chuỗi dòng chảy thời đoạn giờ. Phương pháp đề xuất cho thấy tiềm năng lớn trong ứng dụng cho các hệ thống dự báo thủy văn chính xác, đặc biệt trong điều kiện thiếu dữ liệu quan trắc.*

Từ khóa: Hiệu chỉnh sai số, HYPE, Random Forest, mô hình lai, dòng chảy, học máy, Pleikrong.

1. ĐẶT VẤN ĐỀ

Mô phỏng tính toán dòng chảy là một trong những nhiệm vụ trọng tâm của ngành thủy văn hiện đại, phục vụ trực tiếp cho việc quản lý tài nguyên nước, vận hành điều tiết hiệu quả hồ chứa, cảnh báo lũ, và ứng phó biến đổi khí hậu (Beven, 2012). Các mô hình thủy văn như SWAT, MIKE SHE, HBV, VIC và gần đây là HYPE đã được áp dụng rộng rãi trong mô phỏng dòng chảy ở quy mô và điều kiện lưu vực khác nhau. Tuy nhiên, nhiều nghiên cứu đã chỉ ra rằng sai số giữa dòng chảy mô phỏng và thực đo khi áp dụng các mô hình trên còn khá lớn, đặc biệt mô phỏng cho dòng chảy lũ do thiếu dữ liệu hoặc số liệu có độ chính xác không cao (Refsgaard, 1997), (B. Arheimer, C. Donnelly, and G. Lindström, 2020).

Nguyên nhân gây ra sai số có thể do nhiều yếu tố, gồm sai số do dữ liệu khí tượng – thủy văn, cấu trúc mô hình thủy văn và khả năng mô hình hóa các quan hệ phi tuyến phức tạp giữa các yếu tố của quá trình mưa – dòng chảy (K. Beven and A. Binley, 1992), (M. P. Clark et al, 2011). Do đó, việc nâng cao kết quả mô phỏng thông qua hiệu chỉnh sai số đầu ra đã đang được các chuyên gia và cơ quan xây dựng mô hình quan tâm. Một trong những hướng tiếp cận nổi bật là tích hợp mô hình thủy văn với mô hình học máy và được biết là dạng mô hình lai, nhằm tận dụng ưu điểm của cả hai cách tiếp cận - mô hình cơ chế vật lý và mô hình dữ liệu (A. Mosavi, et al, 2020).

Phương pháp học sai số (error modeling) – trong đó mô hình học máy, học sai số giữa dòng chảy thực đo và dòng chảy mô phỏng đã chứng minh hiệu quả cao trong nhiều nghiên cứu (Abrahart et al, 2004), (Dawson et al , 2001) (Fang et al, 2022). Thay vì can thiệp vào cấu trúc mô hình vật lý, cách tiếp cận này đóng vai trò như một tầng hiệu chỉnh đầu ra, sử dụng mô hình học máy để "học" các thành phần sai lệch còn lại và cộng ngược trở lại để hiệu chỉnh kết quả. Các mô hình học máy như Random Forest (RF), Support Vector Regression (SVR), mạng nơ-ron nhân tạo (ANN), và đặc biệt là LSTM đã được áp dụng rộng rãi trong lĩnh vực này.

Mosavi et al. (2020) đã thực hiện một tổng quan hệ thống trên hơn 80 nghiên cứu ứng dụng học máy trong hiệu chỉnh thủy văn, cho thấy RF và XGBoost là các mô hình nổi bật nhờ khả năng học hiệu quả từ dữ liệu không đầy đủ, nhiễu và phân bố bất định. Bên cạnh đó, các mô hình lai kết hợp giữa mô hình vật lý và học máy (như SWAT–ANN, NAM–RF, MIKE SHE–LSTM) của nhóm nghiên cứu (Kratzert et al, 2019) được xem là hướng tiếp cận triển vọng, vừa đảm bảo độ chính xác mô phỏng, vừa duy trì các nguyên lý vật lý . Các nghiên cứu gần đây (Shrestha et al, 2015), (Liu et al, 2017), (Mosavi et al, 2020) cũng khẳng định rằng mô hình lai có thể cải thiện đáng kể chất lượng của kết quả mô phỏng (thông qua các chỉ tiêu đánh giá sai số không có đơn vị, như: NSE, RMSE, KGE) và tăng khả năng phản ứng trước các biến động bất thường trong chuỗi dòng chảy mô phỏng.

Hiện nay, mặc dù mô hình HYPE đã được ứng dụng rộng rãi trong mô phỏng dòng chảy ở nhiều quy mô và điều kiện lưu vực khác nhau, tuy nhiên các nghiên cứu tích hợp HYPE với mô hình học máy, đặc biệt ở quy mô thời đoạn giờ vẫn còn rất hạn chế (SMHI, 2022), (Lindström, G et al, 2010), (Arheimer, B et al, 2020). Đặc biệt tại các lưu vực miền núi Tây Nguyên, nơi dữ liệu quan trắc còn thưa thớt và hệ thống thủy văn chịu ảnh hưởng mạnh bởi các quá trình phi tuyến và biến động ngắn hạn (T. M. Nguyen et al, 2020). Việc nghiên cứu kết hợp giữa hai mô hình này

[1] *Trường Đại học Tài nguyên và Môi trường Hà Nội*
[2] *Trường Đại học Thủy lợi*
[3] *Trung tâm Quy hoạch và Điều tra tài nguyên nước quốc gia*
[*] *Tác giả liên hệ*

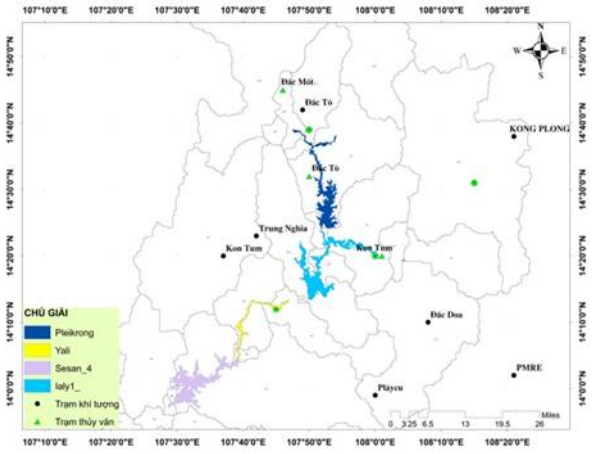
trong bối cảnh trên có ý nghĩa quan trọng nhằm nâng cao chất lượng mô phỏng dòng chảy và phục vụ điều hành hồ chứa hiệu quả hơn.

Việc lựa chọn kết hợp mô hình HYPE với thuật toán Random Forest (RF) trong nghiên cứu này xuất phát từ đặc điểm dữ liệu và mục tiêu hiệu chỉnh dòng chảy. Mô hình HYPE, với cơ chế vật lý bán phân bố, có khả năng mô phỏng các quá trình thủy văn từ mưa - dòng chảy trên lưu vực, nhưng vẫn tồn tại sai số do đơn giản hóa các quá trình phi tuyến và hạn chế về tham số (Lindström, G et al, 2010). Trong khi mô hình RF là một mô hình học máy dựa trên ensemble, có khả năng xử lý tốt các mối quan hệ phi tuyến và dữ liệu nhiễu, đồng thời ít nhạy cảm với việc tinh chỉnh siêu tham số so với các mạng nơ-ron nhân tạo (ANN) hay mô hình chuỗi thời gian như LSTM (Breiman, 2001). Hơn nữa, mô hình RF giúp giảm thiểu hiện tượng quá khớp thông qua kỹ thuật bagging, ổn định với dữ liệu hạn chế và cho phép đánh giá tầm quan trọng của các biến đầu vào, từ đó giúp hiểu rõ hơn ảnh hưởng của các yếu tố khí tượng – thủy văn đến sai số mô phỏng. So với các mô hình ANN, LSTM hay SVM, thì mô hình RF triển khai nhanh hơn, dễ hiệu chỉnh và cho kết quả đáng tin cậy với dữ liệu biến động mạnh, đặc biệt là dòng chảy theo giờ tại các lưu vực miền núi như Tây Nguyên.

Do đó việc xây dựng mô hình kết hợp giữa HYPE và RF để hiệu chỉnh dòng chảy về hồ chứa Pleikrong là rất cần thiết trong việc nâng cao chất lượng mô phỏng dòng chảy từ mô hình thủy văn tại khu vực nghiên cứu. Mục tiêu của nghiên cứu là cải thiện độ chính xác của kết quả mô phỏng dòng chảy thông qua cấu trúc mô hình lai (HYPE -RF), qua đó tạo ra bộ dữ liệu dòng chảy có độ chính xác hơn theo bước thời gian ngày và giờ.

2. PHẠM VI, PHƯƠNG PHÁP VÀ DỮ LIỆU PHỤC VỤ NGHIÊN CỨU

2.1. Phạm vi nghiên cứu



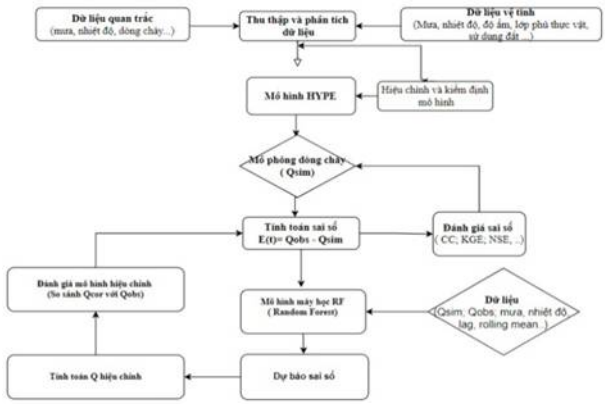
Hình 1. Khu vực nghiên cứu

Hồ thủy điện Pleikrông nằm trên nhánh sông Krông Pôkô, thuộc hệ thống sông Sê San, thuộc địa phận xã Sa Bình, huyện Sa Thầy và xã Kroong, thị xã Kon Tum (Kon Tum). Hệ thống các sông suối chảy về hồ chứa Pleikrông bao gồm sông Krông PôKô với diện tích 3216 km^2 , Lsông = 125.6 km; suối Đak Rô Long với diện tích lưu vực 335 km^2 và chiều dài 34.5 km, suối Đak Tơ Can với diện tích lưu vực 3.14 km^2 và chiều dài sông 39.0 km và suối Đak Psi với diện tích lưu vực 869 km^2 và độ dài sông 62.0 km.

2.2. Phương pháp nghiên cứu

Nghiên cứu này được triển khai theo hướng tiếp cận lai, trong đó mô hình thủy văn vật lý HYPE được kết hợp với mô hình học máy nhằm hiệu chỉnh sai số còn lại trong mô phỏng dòng chảy về hồ chứa. Toàn bộ phương pháp được chia thành hai giai đoạn chính: (i) Thiết lập, hiệu chỉnh mô hình HYPE và mô phỏng dòng chảy về hồ chứa Pleikrong, và (ii) Xây dựng mô hình hiệu chỉnh sai số bằng trí tuệ nhân tạo. Sơ đồ tổng quan phương pháp nghiên cứu được trình bày trong hình 2.

Công thức tổng quát sai số giữa dòng chảy thực đo va mô phỏng được xác định: $E(t) = Q_{obs}(t) - Q_{sim}(t)$ trong đó $Q_{obs}(t)$ là dòng chảy thực đo; $Q_{sim}(t)$ là dòng chảy mô phỏng từ mô hình thủy văn. Mô hình học máy sẽ học mối quan hệ: $e_t = f(X_t)$ với e_t là sai số tại thời điểm t; X_t là tập các đặc trưng tại thời điểm t (bao gồm dữ liệu dòng chảy mô phỏng, dòng chảy thực đo, mưa, nhiệt độ, chỉ số thời gian,...). Sau khi học được sai số ta hiệu chỉnh kết quả mô phỏng: $Q_{corr} = Q_{sim} + \hat{e}$ trong đó $\hat{e}$ là sai số dự báo bởi mô hình học máy.



Hình 2. Sơ đồ phương pháp nghiên cứu

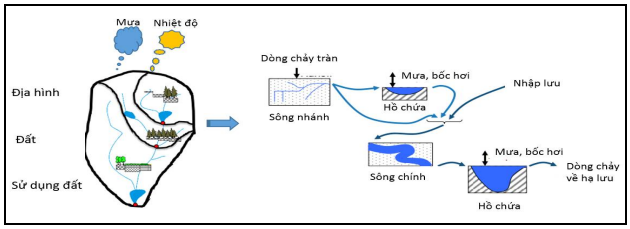
a) Giới thiệu mô hình thủy văn HYPE

Mô hình HYPE được phát triển bởi Viện Khí tượng thủy văn Thụy Điển (SMHI). Đây là một mô hình thủy văn thông số bán phân bố về dòng chảy và chất lượng nước, chạy dưới hệ điều Window hoặc Linux. Cấu trúc mô hình HYPE dựa trên cách tiếp cận đa lưu vực cho phép mô hình hóa đồng thời nhiều lưu vực con hay lưu vực sông được chia thành nhiều tiểu lưu vực và mỗi tiểu lưu vực sẽ sử dụng HYPE để tổng hợp các đơn vị thủy văn (HRU) thành đường quá trình của tiểu lưu vực đó, dựa trên đặc điểm mưa, địa hình và sử dụng đất (Tien L.T, et al, 2020).

Việc tính toán dòng chảy trong mô hình HYPE bao gồm việc tổng hợp các đóng góp từ dòng chảy bề mặt,

dòng chảy sát mặt và dòng chảy cơ bản (nước ngầm). Phương trình dòng chảy trong mô hình HYPE có thể được biểu diễn như sau: $Q_{total} = Q_{surface} + Q_{subsurface} + Q_{base}$

Trong đó: Q_{total}: Tổng dòng chảy tại một điểm bất kỳ trên sông; $Q_{surface}$: Dòng chảy bề mặt; $Q_{surface} = f(P_{excess}, landuse, slope)$;$Q_{subsurface}$: Đóng góp từ nước trong tầng đất và các lớp trung gian; $Q_{subsurface} = k_{sub} * S_{soil}$;(k_{sub}: Hệ số dòng chảy tầng ngầm; S_{soil}: Lượng nước lưu trữ trong đất có sẵn để tạo dòng chảy ngang); Q_{base}: Đóng góp từ dòng chảy cơ bản (nước ngầm); $Q_{base} = k_{gw} * S_{gw}$ (k_{gw}: Hệ số suy giảm nước ngầm; S_{gw}: Lượng nước ngầm lưu trữ).

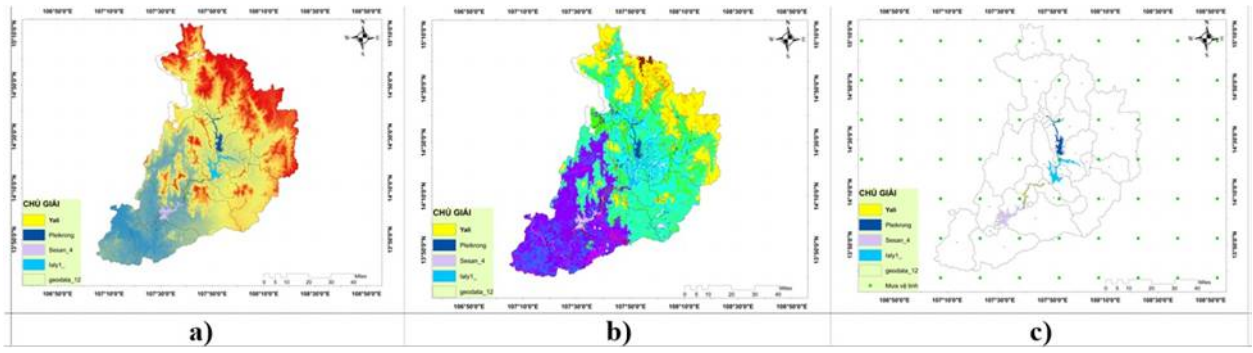


Hình 3. Các quá trình diễn toán trong mô hình HYPE

Một trong những yếu tố quan trọng ảnh hưởng đến hiệu quả của mô hình mưa–dòng chảy là độ tin cậy của dữ liệu mưa vệ tinh. Các nguồn dữ liệu GPM-IMERG (Version 6) và V_forced đã được nhiều nghiên cứu trước đây chứng minh có độ chính xác phù hợp để áp dụng trong mô phỏng mưa–dòng chảy bằng các mô hình thủy văn, đặc biệt tại các khu vực miền núi thiếu trạm đo mưa mặt đất (Bằng, 2020), (Lan.V.V, Cat.V.M, Duong.B.D, 2024). Do đó, trong nghiên cứu này, dữ liệu mưa GPM-IMERG v6 được sử dụng để mô phỏng dòng chảy theo bước thời gian giờ bằng mô hình HYPE, trong khi dữ liệu mưa V_forced được dùng cho mô phỏng dòng chảy theo bước thời gian ngày. Dữ liệu nhiệt độ được thu thập tương ứng theo thời gian giờ và ngày nhằm đảm bảo tính đồng bộ với dữ liệu mưa vệ tinh. Các dữ liệu nền khác như địa hình và sử dụng đất được giữ nguyên cho cả hai bước thời gian mô phỏng. Các thông tin chi tiết về bộ dữ liệu đầu vào phục vụ thiết lập mô hình HYPE được trình bày trong bảng sau.

Bảng 1. Các thông số dữ liệu đầu vào mô hình HYPE

Loại dữ liệu	Nguồn dữ liệu	Thời gian thu thập dữ liệu	Độ phân giải theo không gian	Độ phân giải theo thời gian
Mưa vệ tinh	Global Precipitation Measurement Integrated MultiSatellitE Retrievals for GPM (GPM-IMERG) v6	2001-2022	0.1^0	1 giờ
Mưa vệ tinh	V_forced (merge product from five data source GPM, GSMAP, MSWEP, ERA5)	2001-2023	0.1^0	1 ngày
Địa hình	Sentinel-1 SAR and SRTM DEM	2002	30 m	-
Nhiệt độ	Nhiệt độ ngày trung bình, dữ liệu toàn cầu GEE	2001-2022	0.1^0	1 giờ/ ngày
Sử dụng đất	Ảnh vệ tinh Sentinel-2	2023	30 m	-
Lưu lượng dòng chảy ngày	Trạm thủy văn Đăk Mốt	2001- 2022	-	1 ngày
Lưu lượng dòng chảy giờ	Lưu lượng về hồ chứa Pleikrong (https://hochuathuydien.evn.com.vn/login.aspx)	2011-2022	-	1 giờ



Hình 4. Dữ liệu đầu vào mô hình (a) Dữ liệu địa hình; b) Dữ liệu sử dụng đất; c) Dữ liệu mưa vệ tinh)

b) Cơ sở lý thuyết mô hình Random Forest (RF)

Random Forest (RF) là một thuật toán học máy thuộc nhóm ensemble learning, được đề xuất bởi (Breiman, 2001) kết hợp nhiều cây quyết định để tăng độ chính xác và giảm thiểu overfitting. Nguyên lý hoạt động của RF dựa trên hai đặc điểm chính: bootstrap aggregating (bagging) và random feature selection.

Cơ chế hoạt động: Với mô hình RF, từ tập dữ liệu gốc D={(x1, y1),(x2, y2),...,(xn,yn)} , nhiều tập con huấn luyện $D^b \subset D$ được tạo bằng phương pháp bootstrap (lấy mẫu ngẫu nhiên có hoàn lại). Mỗi tập con này được dùng để huấn luyện một cây quyết định. Trong quá trình xây dựng mỗi cây, tại mỗi nút chia, một tập con ngẫu nhiên các đặc trưng (gồm m<M, với M là tổng số đặc trưng đầu vào) được lựa chọn, sau đó tiến hành tìm đặc trưng tối ưu nhất theo tiêu chí giảm thiểu sai số (đối với hồi quy) hoặc tăng tính thuần nhất (gini/information gain đối với phân loại).

Cấu trúc và huấn luyện mô hình RF: Mỗi cây trong RF được xây dựng độc lập trên tập con bootstrap, sau đó kết hợp dự đoán của tất cả cây để tổng hợp kết quả:

Hồi quy: Dự báo cuối cùng là trung bình dự báo từ B cây: $\hat{f}(x) = \frac{1}{B}\sum_{b-1}^{B} f_b(x)$

Phân loại: Dự báo là lớp nhận được nhiều lựa chọn nhất: $\hat{y} = \arg\max \sum_{b-1}^{B} \Pi(f_b(x) = c)$

Trong đó, Π là hàm chỉ thị (bằng 1 nếu đúng, ngược lại bằng 0), f_b (x) là kết quả của cây thứ b , và c là lớp ứng viên.

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Hiệu chỉnh và kiểm định mô hình HYPE

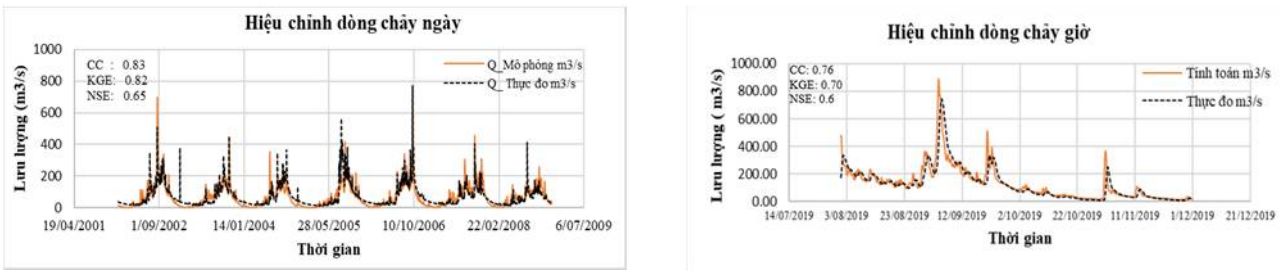
a) Hiệu chỉnh:

Trong nghiên cứu này, chuỗi số liệu dòng chảy ngày thực đo tại trạm Đăk Mốt trong giai đoạn 2002–2008 được sử dụng kết hợp với bộ số liệu vệ tinh tương ứng trong cùng thời kỳ để tiến hành hiệu chỉnh bộ thông số của mô hình thủy văn HYPE. Mô hình HYPE tích hợp sẵn công cụ tự động hiệu chỉnh (auto-calibration tool), cho phép truy tìm tổ hợp thông số tối ưu nhằm đảm bảo độ phù hợp cao nhất giữa giá trị mô phỏng và quan trắc. Quá trình hiệu chỉnh được thực hiện ở bước thời gian ngày, với kết quả đánh giá hiệu suất mô hình thông qua các chỉ tiêu thống kê gồm hệ số tương quan (CC), hệ số Kling–Gupta (KGE) và hệ số Nash–Sutcliffe (NSE). Kết quả hiệu chỉnh cho thấy mô hình đạt được CC = 0.83, KGE = 0.82 và NSE = 0.65, phản ánh mức độ tương quan và độ tin cậy cao giữa lưu lượng mô phỏng và lưu lượng thực đo tại trạm Đăk Mốt.

Sau khi hoàn thiện quá trình hiệu chỉnh, mô hình HYPE được chi tiết hóa để mô phỏng dòng chảy theo bước thời gian giờ. Kết quả mô phỏng được so sánh với chuỗi số liệu thực đo về lưu lượng dòng chảy vào hồ chứa Pleikrong trong giai đoạn từ tháng 7 đến tháng 12 năm 2019, được thu thập từ hệ thống quan trắc và vận hành hồ thủy điện. Các chỉ tiêu đánh giá chất lượng mô phỏng cho thấy CC = 0.76, KGE = 0.70 và NSE = 0.60, chứng tỏ mô hình HYPE có khả năng tái hiện tương đối tốt biến động dòng chảy thực tế ở thang thời gian ngắn hơn.

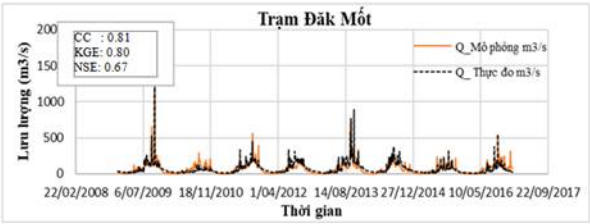
Bảng 2. Bộ thông số mô hình HYPE sau khi hiệu chỉnh và khoảng giá trị cho phép

TT	Thông số	Ký hiệu	Hiệu chỉnh theo bước thời gian ngày	Hiệu chỉnh theo bước thời gian giờ	Đơn vị	Min	Max
1	Thoát hơi nước của đất	wcfc	0.473	0.473	-	0.05	0.5
		wcwp	0.13	0.13	%	0.05	0.5
2	Độ rỗng của đất	wcep	0.15	0.15	%	0.05	0.5
3	Độ thấm của đất	mperc	113.4	4.73	mm/ngày	5	120
4	Hệ số trễ dòng chảy mặt	rrcs	0.34	0.00642	-	0	0.6
5	Hệ số dòng chảy ngầm	macrate	0.29	-	-	0.05	0.5
		mactrinf	3.75	0.156	mm/ngày	0	100
		mactrsm	0.35	-	-	0	1.0
6	Hệ số bốc hơi của đất	cevp	0.17	0.01	mm/ngày/^{0}C	0.15	0.3



Hình 5. So sánh dòng chảy mô phỏng và thực đo trong giai đoạn hiệu chỉnh mô hình

b) Kiểm định: Nghiên cứu sử dụng dữ liệu vệ tinh cùng thời gian với chuỗi số liệu thực đo ngày trạm thủy văn Đăk Mốt giai đoạn từ 2009 đến 2016 để kiểm định mô hình thủy văn. Phương pháp thực hiện tương tự như khi hiệu chỉnh. Kết quả kiểm định được thể hiện trong hình 6 với các hệ số CC = 0.81; KGE = 0.8 và NSE = 0.67.



Hình 6. Quá trình lưu lượng mô phỏng ngày và thực đo giai đoạn kiểm định mô hình

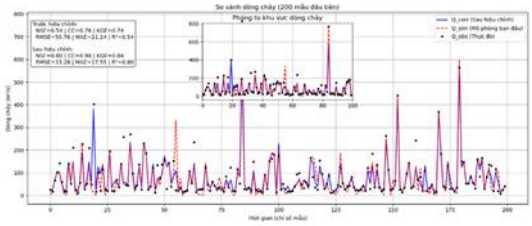
3.2. Xây dựng mô hình RF

Trong nghiên cứu này, quy trình xây dựng mô hình Random Forest (RF) được triển khai gồm ba giai đoạn: huấn luyện, hiệu chỉnh và kiểm định, trước khi ứng dụng cho giai đoạn dự báo sai số dòng chảy. Nghiên cứu đã sử dụng chuỗi dữ liệu ngày từ năm 2002–2008 để huấn luyện mô hình, trong đó các dữ liệu đầu vào của mô hình RF bao gồm dòng chảy mô phỏng từ HYPE (Q_{sim}), dòng chảy thực đo (Q_{obs}), lượng mưa, nhiệt độ, các đặc trưng thời gian (giá trị trung bình ngày, trung bình tháng, trung bình năm), cùng với các đặc trưng thủy văn mở rộng như độ trễ (Q_{sim_lag1}, Q_{sim_lag2}) và trung bình trượt (Q_{sim_roll3}, Q_{sim_roll7}). Chuỗi dữ liệu thời đoạn ngày từ năm 2009 đến 2016 dùng để kiểm định mô hình RF. Sau khi hiệu chỉnh đạt kết quả tốt, nghiên cứu ứng dụng mô hình RF dự báo sai số dòng chảy thời đoạn ngày cho giai đoạn từ 2017 đến 2022.

Trong giai đoạn hiệu chỉnh các tham số của mô hình RF được tối ưu bằng phương pháp Randomized Search kết hợp kiểm định chéo k-fold (k = 5), trong đó các tham số quan trọng như số cây (n_estimators), độ sâu tối đa (max_depth), số đặc trưng chọn tại mỗi lần chia (max_features), số mẫu tối thiểu để tách nhánh (min_samples_split) và số mẫu tối thiểu tại nút lá (min_samples_leaf) được hiệu chỉnh để đảm bảo cân bằng giữa độ chính xác và khả năng khái quát hóa (bảng 3).Nghiên cứu áp dụng mô hình tính toán cho chuỗi dòng chảy ngày giai đoạn kiểm định (2009–2016) nhằm đánh giá khả năng dự báo sai số dòng chảy của mô hình dựa trên các chỉ số NSE, KGE, CC, R^2 và RMSE, MAE. Sau khi mô hình RF có được bộ tham số tối ưu sẽ ứng dụng dự báo sai số dòng chảy phục vụ cho việc hiệu chỉnh chuỗi dòng chảy ngày cho giai đoạn 2017–2022.

Bảng 3. Tham số mô hình RF sử dụng trong nghiên cứu

Tham số	Ý nghĩa	Giá trị mặc định	Giá trị sử dụng
n_estimators	Số lượng cây quyết định trong rừng	100	500
max_depth	Độ sâu tối đa của cây	không giới hạn	10
min_samples_split	Số mẫu tối thiểu để chia một nút	2	5
min_samples_leaf	Số mẫu tối thiểu ở nút lá	1	2
max_features	Số lượng đặc trưng được chọn tại mỗi lần chia	tất cả	"sqrt"
random_state	Hạt giống ngẫu nhiên để tái lập kết quả	không giới hạn	42

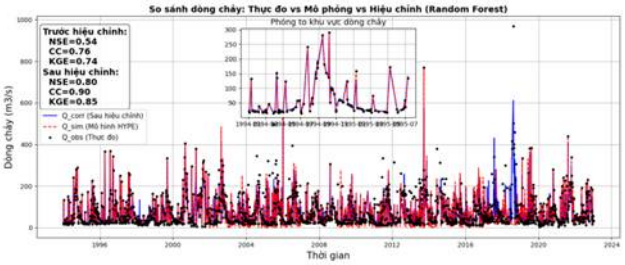


Hình 7. Kết quả hiệu chỉnh mô hình Random Forest

Sau khi đã hoàn thành quá trình hiệu chỉnh và kiểm định mô hình RF trong giai đoạn từ năm 2002 đến 2016 cho kết quả tương đối tốt với các chỉ số đánh giá mô hình KGE, CC, NSE, RMSE, MAE đều đã cải thiện so với kết quả mô phỏng dòng chảy bằng mô hình HYPE. Nghiên cứu đã sử dụng bộ thông số mô hình RF để dự báo sai số dòng chảy ngày trong giai đoạn từ 2017 đến 2022, sau đó so sánh đánh giá kết quả dự báo sai số của mô hình RF với giá trị thực đo tại các trạm kiểm tra.

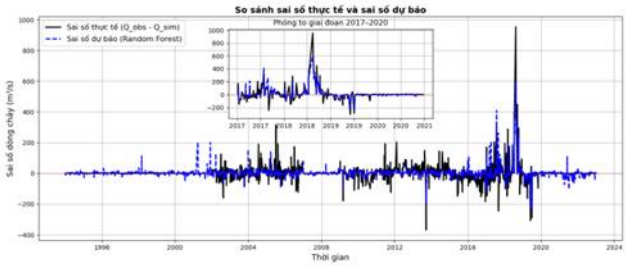
3.3. Kết quả hiệu chỉnh dòng chảy ngày

Nghiên cứu ứng dụng mô hình RF để hiệu chỉnh dòng chảy mô phỏng bằng mô hình HYPE dựa trên việc sử dụng các đặc trưng khí tượng và dòng chảy mô phỏng để dự báo sai số. Từ đó cập nhật vào chuỗi dữ liệu dòng chảy từ mô hình thủy văn HYPE nhằm làm tăng độ chính xác của mô hình. Kết quả tính toán được thể hiện qua hình sau đây:



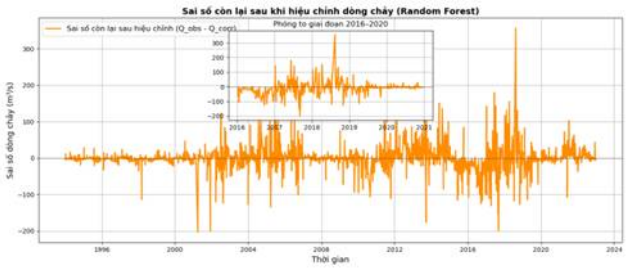
Hình 8. Biểu đồ đánh giá kết quả hiệu chỉnh dòng chảy ngày

Kết quả so sánh giữa dòng chảy ngày mô phỏng (Q_{sim}) và thực đo cho thấy mô hình HYPE bước đầu đã mô phỏng được xu thế của dòng chảy, tuy nhiên vẫn còn sai số đáng kể, đặc biệt trong các giai đoạn dòng chảy cực trị. Lưu lượng dòng chảy thường mô phỏng chưa chính xác về các đỉnh lũ. Sau khi hiệu chỉnh bằng mô hình học máy, dòng chảy hiệu chỉnh (Q_{corr}) đã được cải thiện rõ rệt về độ chính xác, dữ liệu dòng chảy đã bám sát dữ liệu quan trắc hơn, đặc biệt tại các đỉnh và pha dao động mạnh. Kết quả hiệu chỉnh dòng chảy ngày cho thấy mô hình học máy có khả năng học và khắc phục sai số phi tuyến của mô hình thủy văn HYPE.



Hình 9. Biểu đồ so sánh giữa sai số thực tế và mô hình dự báo (RF)

Kết quả so sánh giữa sai số thực tế ($E = Q_{obs} - Q_{sim}$) và sai số dự báo (Ê) cho thấy mô hình học máy tái hiện khá tốt đặc điểm dao động sai số của dòng chảy trong toàn bộ chuỗi thời gian. Mô hình Random Forest (RF) thể hiện khả năng học được xu hướng, biên độ và tần suất dao động của sai số, đặc biệt trong các giai đoạn dòng cảy ổn định hoặc biến động nhỏ. Mặc dù vẫn còn sai lệch tại một số đỉnh cực trị, nhưng mô hình RF mô phỏng tương đối chính xác các đặc trưng phi tuyến của sai số. Độ tương đồng cao giữa E và Ê phản ánh hiệu quả huấn luyện của mô hình, đồng thời khẳng định tính phù hợp của phương pháp học máy trong hiệu chỉnh sai số dòng chảy.



Hình 10. Kết quả tính toán dự báo sai số sau khi hiệu chỉnh

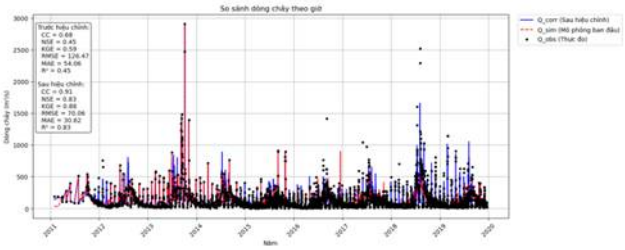
Kết quả so sánh giữa ba chuỗi dữ liệu Q_{obs}, Q_{sim} và Q_{corr} cho thấy hiệu suất mô hình được cải thiện rõ rệt sau khi áp dụng hiệu chỉnh sai số bằng mô hình học máy. Trước hiệu chỉnh, mô hình HYPE đạt NSE = 0.54, CC = 0.76 và KGE = 0.74, RMSE = 50.76, MAE =21.24 cho thấy khả năng mô phỏng xu thế dòng chảy ở mức trung bình, nhưng còn hạn chế trong việc mô phỏng dao động ngắn hạn và các cực trị. Sau hiệu chỉnh, kết quả đánh giá độ chính xác của mô hình tăng đáng kể (NSE = 0.80, CC = 0.9, KGE = 0.85). Hệ số NSE đã tăng 26% điều đó cho thấy mô hình học máy đã nâng cao rõ rệt khả năng mô phỏng đặc tính về biên độ và thời điểm đỉnh lũ.

Bảng 4. Hiệu suất mô hình theo dữ liệu dòng chảy ngày (1994–2022)

Các chỉ số	Mô hình HYPE	Mô hình HYPE-RF	Cải thiện (%)
CC	0.76	0.9	**14**
KGE	0.74	0.85	**11**
NSE	0.54	0.8	**26**
RMSE	50.76	33.28	**17.48**
MAE	21.24	17.55	**3.69**

3.2. Kết quả hiệu chỉnh dòng chảy giờ

Nghiên cứu cũng đã sử dụng mô hình Random Forest để hiệu chỉnh dòng chảy giờ cho khu vực nghiên cứu từ kết quả mô phỏng bằng mô hình thủy văn HYPE. Kết quả hiệu chỉnh dòng chảy giờ được thể hiện qua hình sau:



Hình 12. Kết quả hiệu chỉnh dòng chảy giờ bằng mô hình Random Forest (RF)

Bảng 5. Hiệu suất mô hình theo dữ liệu dòng chảy giờ (2011–2022)

Các chỉ số	Mô hình HYPE	Mô hình HYPE-RF	Cải thiện (%)
CC	0.68	0.91	**23**
KGE	0.59	0.88	**29**
NSE	0.45	0.83	**38**
RMSE	126.47	70.06	**56.41**
MAE	54.06	30.62	**23.44**

Kết quả mô phỏng dòng chảy giờ bằng mô hình HYPE: NSE = 0.45, CC = 0.68 và KGE = 0.59, RMSE =126.47, MAE =54.06 cho thấy khả năng mô phỏng xu thế dòng chảy ở mức khá, nhưng còn hạn chế về biên độ và pha thời gian của dòng chảy cực trị. Sau khi hiệu chỉnh bằng mô hình học máy, các chỉ số đánh giá mô hình được cải thiện đáng kể (NSE = 0.83, CC = 0.91, KGE = 0.88, RMSE=70.06, MAE=30.62), qua đó cho thấy khả năng bù đắp sai số hệ thống và phi tuyến vốn chưa được mô hình thủy văn HYPE thể hiện đầy đủ, nhất là trong các khoảng thời gian dòng chảy biến động nhanh.

So với kết quả mô phỏng theo ngày (NSE tăng từ 0.45 lên 0.83), trong khi đó hiệu chỉnh dữ liệu dòng chảy theo giờ cho thấy hiệu suất vượt trội hơn, với mức tăng NSE đạt 38 % so với 26 % ở chuỗi ngày. Điều này khẳng định lợi thế của mô hình học máy trong khai thác các đặc trưng chi tiết của dữ liệu tần suất cao, nhờ vào khả năng học phi tuyến và thích ứng linh hoạt với các tín hiệu ngắn hạn mà mô hình vật lý truyền thống thường chưa đạt được.

4. KẾT LUẬN VÀ KIẾN NGHỊ

4.1. Kết luận

Nghiên cứu đã phát triển và kiểm chứng hiệu quả của mô hình lai HYPE–RF trong hiệu chỉnh sai số dòng chảy mô phỏng tại lưu vực hồ chứa Pleikrong. Mô hình thủy văn HYPE được sử dụng để mô phỏng chuỗi dòng chảy ngày và giờ, dựa trên dữ liệu khí tượng – thủy văn và viễn thám. Sau đó, thuật toán học máy Random Forest (RF) được áp dụng để học và hiệu chỉnh sai số còn lại giữa dòng chảy mô phỏng và thực đo. Phương pháp học sai số này cho phép cải thiện đáng kể độ chính xác mô phỏng mà không cần thay đổi cấu trúc bên trong của mô hình thủy văn.

Kết quả tính toán cho thấy mô hình HYPE–RF nâng cao rõ rệt các chỉ số hiệu suất mô phỏng. Với dữ liệu dòng chảy ngày, chỉ số NSE tăng từ 0.54 lên 0.8 (tăng 26 % so với mô hình thủy văn HYPE), trong khi với dữ liệu dòng chảy giờ, NSE tăng từ 0.45 lên 0.83 (tăng 38 %). Tương tự, các chỉ số CC và KGE cũng cải thiện mạnh mẽ ở cả hai chuỗi dữ liệu dòng chảy ngày và giờ.

Từ kết quả hiệu chỉnh dòng chảy ngày và giờ cho thấy hiệu chỉnh dòng chảy giờ mang lại hiệu suất cao hơn so với dòng chảy ngày, do mô hình học máy có khả năng nắm bắt các đặc trưng dao động ngắn hạn và phản ứng với các biến động đột ngột trong chuỗi dữ liệu. Điều này phản ánh lợi thế của mô hình học máy RF trong việc xử lý dữ liệu có tần suất cao, nhờ năng lực học phi tuyến và thích ứng với tín hiệu thời gian chi tiết.

Phương pháp lai giữa mô hình HYPE–RF thể hiện khả năng phục hồi dữ liệu dòng chảy trong các giai đoạn thiếu quan trắc, nhờ vào việc ước lượng sai số từ các đặc trưng khí tượng – thủy văn đầu vào. Điều này đặc biệt có ý nghĩa trong bối cảnh nhiều lưu vực tại Việt Nam và trên thế giới còn thiếu dữ liệu quan trắc liên tục.

4.2. Kiến nghị

Nghiên cứu mới chỉ sử dụng mô hình RF trong việc hiệu chỉnh dòng chảy vì vậy trong nghiên cứu tiếp theo, nghiên cứu mở rộng đánh giá bằng cách so sánh HYPE–RF với các cấu hình HYPE–LSTM, HYPE–XGBoost nhằm kiểm định mức độ cải thiện hiệu suất và năng lực phản ứng với dữ liệu dòng chảy cực trị.

TÀI LIỆU THAM KHẢO

Bằng, N. L. (2020). *Nghiên cứu đánh giá sản phẩm mưa từ nhiệm vụ đo mưa toàn cầu (GPM) cho miền Bắc Việt Nam*. Khoa học kỹ thuật Thủy Lợi và Môi trường, 110-115.

Lan.V.V, Cat.V.M, Duong.B.D. (2024). *Đánh giá khả năng sử dụng dữ liệu mưa vệ tinh để mô phỏng dòng chảy bằng mô hình thủy văn HYPE, áp dụng cho lưu vực sông Sê San.* Tạp chí Khoa học tài nguyên và môi trường, 3-12.

A. Mosavi, et al. (2020). *Flood prediction using machine learning models: A revie.* Water, 1427.

Abrahart et al. (2004). *Neural network modelling of nonlinear hydrological relationships.* Hydrology and Earth System Sciences, 478–486.

Arheimer, B et al. (2020). *Prolonged experience of large-scale hydrological modelling in Sweden: HYPE model development and applications.* Hydrology Research, 20–38.

B. Arheimer, C. Donnelly, and G. Lindström. (2020). *Large-scale hydrological modelling in Sweden.* Hydrology Research, 20–38.

Beven, K. (2012). *Rainfall–Runoff Modelling.* Wiley-Blackwell.

Breiman. (2001). *Random Forests.* Machine Learning, 5–32.

Dawson et al . (2001). *Hydrological modelling using artificial neural networks.* Progress in Physical Geography, 80–108.

Fang et al. (2022). *Combining physically-based models and machine learning for hydrologic prediction: Current progress and challenges.* Water Resources Research, 58.

K. Beven and A. Binley. (1992). *The future of distributed models.* Hydrological Processes, 279–298.

Kratzert et al. (2019). *Towards learning universal, regional, and local hydrological behaviors via machine learning.* Hydrology and Earth System Sciences, 5089–5110.

Lindström, G et al. (2010). *Development and testing of the HYPE hydrological model: A water quality model for different spatial scales.* Hydrology Research, 295–319.

Liu et al. (2017). *Correction of SWAT-simulated streamflow using ANN model.* Water Resources Management, 4527–4541.

M. P. Clark et al. (2011). *A unified approach for process-based hydrologic modelling.* Hydrological Processes, 2554–2577.

Mosavi et al. (2020). *Flood prediction using machine learning models: Literature review.* Water, 1427.
Refsgaard, J. C. (1997). *Parameterisation, calibration and validation of distributed hydrological models.* Journal of Hydrolog, 69–97.
Shrestha et al. (2015). *Machine learning approaches for rainfall-runoff modelling.* Hydrological Sciences Journal, 505–518.
SMHI. (2022). *Retrieved from Model water download the code and learn how to use hype,* HypeWeb: https://hypeweb.smhi.se/model-water
T. M. Nguyen et al. (2020). *Evaluation of satellite rainfall in Central Highlands Vietnam.* Journal of Hydrometeorology, 765–781.
Tien L.T, et al. (2020). *Streamflow prediction in "geopolitically ungauged" basins using satellite observations and regionalization at subcontinental scale.* Journal of Hydrology, 588.

Abstract:

APPLICATION OF A HYBRID MODEL (HYPE-RF) FOR STREAMFLOW ERROR CORRECTION AT PLEIKRONG RESERVOIR

This study proposes a hybrid approach that integrates the HYPE hydrological model with the Random Forest (RF) machine-learning algorithm to correct simulation errors of streamflow in the Pleikrong reservoir basin, Kon Tum Province. The HYPE model is employed to simulate daily and hourly streamflow based on meteorological, hydrological, and remote-sensing data. The discrepancies between simulated and observed streamflow are trained using the RF model to optimize the final outputs. The results indicate that the hybrid HYPE–RF framework substantially enhances simulation accuracy, with notable improvements in NSE, CC, and KGE. Specifically, NSE increases from 0.54 to 0.80 for daily streamflow and from 0.45 to 0.83 for hourly streamflow. The proposed method demonstrates strong potential for application in high-accuracy hydrological forecasting systems, particularly under conditions of limited observational data.

Keywords: Error correction, HYPE, Random Forest, hybrid model, streamflow, machine learning, Pleikrong.

Ngày nhận bài: 12/8/2025
Ngày chấp nhận đăng: 26/11/2025