ÁP DỤNG MÔ HÌNH TRÍ TUỆ NHÂN TẠO VÀO DỰ BÁO LƯU LƯỢNG ĐẾN HỒ LƯU VỰC SÔNG BA

Cao Hoàng Hải[1], Trần Anh Phương[1], Thái Quỳnh Như[1], Trần Mạnh Cường[1]

Tóm tắt: *Trong nghiên cứu này, hai mô hình AI là Random Forest (RF) và Support Vector Machine (SVM/SVR) đã được áp dụng thử nghiệm đối với một hồ chứa lớn - hồ Sông Hinh trên lưu vực Sông Ba, Việt Nam. Ba trường hợp tính toán là dự báo lưu lượng trung bình 3 ngày, 7 ngày và 1 tháng (tương ứng với ngắn, trung và dài hạn) đến hồ sử dụng số liệu khí tượng, thủy văn trong khu vực đã được xây dựng để kiểm nghiệm khả năng dự báo của hai mô hình RF và SVR. Kết quả cho thấy, cả hai mô hình đều đưa ra kết quả dự báo với độ chính xác cao thể hiện qua chỉ số NSE trung bình đạt trên 0,8, đặc biệt trong một số trường hợp tính toán như dự báo lưu lượng trung hạn và dài hạn, chỉ số NSE trung bình trên 0,9. Trong 2 mô hình được thử nghiệm thì mô hình SVR nhìn chung cho kết quả tốt nhất đối với dự báo ngắn và dài hạn, trong khi đó mô hình RF lại cho thấy sự vượt trội ở dự báo trung hạn. Các mô hình AI thử nghiệm đều không dự báo chính xác một cách đồng nhất dòng chảy lũ do các mô hình không được huấn luyện tập trung vào dự báo dòng chảy lũ mà ưu tiên vào quá trình dòng chảy. Bên cạnh đó, việc lựa chọn số liệu đầu vào có độ tương quan cao với lưu lượng đến hồ đóng vai trò quan trọng trong việc nâng cao hiệu quả dự báo của mô hình. Đây hoàn toàn có thể là một phương án bổ sung cho công tác dự báo lưu lượng tới hồ bên cạnh các cách tiếp cận đang được sử dụng hiện nay.*

Từ khóa: *AI, ML, SVR, RF, Sông Ba.*

Ban Biên tập nhận bài: 5/7/2019 Ngày phản biện xong: 22/8/2019 Ngày đăng bài: 25/09/2019

1. Đặt vấn đề

nguyên nước. Các công trình này được xây dựng nhằm cung cấp nước cho sản xuất công nghiệp, nông nghiệp và sinh hoạt kết hợp với cắt và giảm lũ hạ du. Việc quản lý hiệu quả công trình hồ chứa nước sẽ đem lại lợi ích lớn cho công tác phòng chống thiên tai và phát triển kinh tế, xã hội trong vùng. Do đó, việc nâng cao chất lượng dự báo lưu lượng tới hồ chứa là một trong những vấn đề được nhiều nhà khoa học cũng như các nhà quản lý nước trong nước và trên thế giới quan tâm.

Cho đến nay, trong các nghiên cứu về dự báo lưu lượng vào hồ chứa nói riêng, hay dự báo hoặc mô phỏng lưu lượng/quá trình mưa-dòng chảy đều sử dụng các mô hình thủy văn phân bố hay bán phân bố khác nhau. Các mô hình loại này được xây dựng để mô phỏng đặc tính vật lý và quá trình của dòng chảy. Do khả năng mô phỏng có độ chính xác cao các quá trình vật lý và phân tích độ nhạy cảm một cách toàn diện [1], hơn nữa các mô hình này rất hữu ích cho các nhà khoa học trong việc giải thích được toàn bộ quá trình ẩn đằng sau [2], do đó các mô hình loại này được áp dụng khá rộng rãi ở nhiều khu vực trên thế giới Tuy nhiên, việc sử dụng các mô hình này thường yêu cầu một lượng dữ liệu chi tiết về đặc tính của lưu vực như các số liệu thông tin địa lý, mưa, dòng chảy, địa chất… Bên cạnh nó việc hiệu chỉnh và kiểm định mô hình cũng rất phức tạp và đòi hỏi nhiều thời gian, kinh nghiệm và kiến thức của người chạy mô hình đối với từng lưu vực cụ thể. Vì vậy, khả năng áp dụng loại mô hình này ở nhiều khu vực và trong các bài toán dự báo thời đoạn ngắn vẫn còn bị hạn chế [3].

Những hạn chế của các mô hình truyền thống nêu trên đã khuyến khích sự phát triển của các mô hình dựa vào số liệu (*data-driven models*),

[1]*Viện Khoa học tài nguyên nước*
Email: hoanghaicao90@gmail.com

mà phổ biến nhất gần đây có thể kể đến là phương pháp máy học (Machine Learning - ML). Các mô hình ML là công cụ tiềm năng trong việc dự báo dòng chảy do các mô hình này có thể được xây dựng dựa nhanh chóng, dễ dàng, không đòi hỏi phải có sự hiểu biết về các quá trình vật lý ẩn đằng sau. Ngoài ra, lượng dữ liệu yêu cầu tối thiểu, cùng với khả năng tính toán, hiệu chỉnh và kiểm định nhanh hơn so với các mô hình vật lý truyền thống, và cách sử dụng ít phức tạp hơn là những ưu điểm lớn mà các mô hình dựa vào số liệu mang lại [4].

Trong lĩnh vực thủy văn và tài nguyên nước nói chung, và trong các bài toán về mô phỏng, dự báo dòng chảy vào hồ nói riêng, các mô hình trí tuệ nhân tạo như *Artificial Neural Network* (ANN) đã được ứng dụng từ những năm 90 [5], [6]. Tuy nhiên, trong những năm trở lại đây, với sự phát triển vượt bậc của ngành khoa học máy tính cùng với sự quan tâm mạnh mẽ của cộng đồng khoa học với các vấn đề liên quan đến dữ liệu lớn (*big data*), các mô hình trí tuệ nhân tạo, máy học ngày càng được sử dụng rộng rãi hơn.Hiện nay, bên cạnh ANN, *Random Forest* (RF) và *Support Vector Machine* (SVM) là hai mô hình ML được sử dụng khá rộng rãi trong các nghiên cứu về dự báo dòng chảy [7].

RF là phương pháp học máy có giám sát qua việc xây dựng một tập hợp nhiều cây quyết định (*decision tree - DT*) và sử dụng trung bình kết quả của các cây quyết định trên [8]. Các nghiên cứu so sánh khả năng dự báo của RF với các mô hình khác như ANN, SVM của [9] hay với mô hình Prophet của apacharalampous và Tyralis (2018) đều cho thấy RF cho kết quả tốt hơn các mô hình khác, đặc biệt là trong khả năng dự báo sự thay đổi gián đoạn của dòng chảy. Li và nnk [10] và Obringer và Nateghi [11] cũng đã thử nghiệm RF trong dự báo mực nước hồ với nhiều trường hợp tính toán khác nhau bao gồm dự báo thời gian thực. Kết quả cho thấy mô hình RF cho kết quả tốt khi sử dụng số liệu mực nước có độ trễ 4 ngày và trung bình tuần trước đó làm đầu vào tính toán [10], và ở khu vực thành thị thì RF cho kết quả dự báo tốt hơn các mô hình ML khác [11]. Một số nghiên cứu khác sửu dụng mô hình cùng nguồn gốc với RF như Decision Tree hay CART cũng cho kết quả tương tự khi khẳng định thuật toán RF/DT/CART cho kết quả tối ưu hơn khi dự báo dòng chảy ví dụ như [12] với nghiên cứu về dự báo dòng chảy trung bình tháng ở sông Coruh, vùng Đông Biển Đen, Thổ Nhĩ Kỳ; Senthil Kumar và nnk [13] với nghiên cứu so sánh khả năng của các thuật toán MLR, ANN, fuzzy logic và DT trong dự báo dòng chảy ở thượng lưu hồ chứa lưu vực Sutlej, Ấn Độ; Galelli và Castelletti [14] với nghiên cứu đánh giá khả năng dự báo của phương pháp DT và ANN trong dự báo dòng chảy ở lưu vực Marina, Singapore; và Yang và nnk [15] với nghiên cứu so sánh thuật toán DT cơ bản và thuật toán RF trong việc dự báo dòng xả từ hồ chứa cho 9 lưu vực khác nhau ở California, Mỹ đồng thời thử nghiệm sự phù hợp của các mô hình DT trong việc khát quát hóa các vấn đề về mô phỏng dòng chảy.

Tương tự như RF, SVM, một thuật toán học máy có giám sát được đề xuất bởi Vapnik (1963), cũng là một mô hình được sử dụng phổ biến trong dự báo dòng chảy. Mô hình này cho thấy tiềm năng cao trong dự báo dòng chảy ngắn hạn và dài hạn [16-17]. Khi so sánh với các phương pháp khác, mô hình SVM với các biến thể LS-SVR hay SVR cho kết quả tốt hơn và cho thấy khả năng dự báo dòng chảy chính xác với nhiều loại dữ liệu khác nhau [18-20]. Việc áp dụng mô hình SVM/SVR cho dự báo dòng chảy, dòng xả lũ của hồ cũng đươc nghiên cứu ở trên nhiều lưu vực ở Trung Quốc ví dụ như nghiên cứu của [21] về dự báo dòng xả thời đoạn dài của hồ thủy điện Manwan, hay nghiên cứu của Guo và nnk [22] về dự báo dòng chảy tới khu vực đập Tam Hiệp trên sông Dương Tử. Các nghiên cứu này đều đưa ra kết luận rằng mô hình SVR có khả năng dự báo chính xác dòng chảy, đặc biệt là khi áp dụng các biện pháp làm giảm nhiễu số liệu đầu vào.

Như vậy, có thể thấy SVM và RF đã được chứng minh là hai mô hình ML có khả năng dự báo lưu lượng dòng chảy có độ chính xác cao.

Do đó, nghiên cứu này được thực hiện nhằm mục đích áp dụng và so sánh khả năng dự báo của hai mô hình này ở các trường hợp tính toán khác nhau với các điều kiện số liệu khác nhau qua đó tìm ra được mô hình phù hợp cho công tác dự báo lưu lượng đến hồ chứa trên lưu vực Sông Ba.

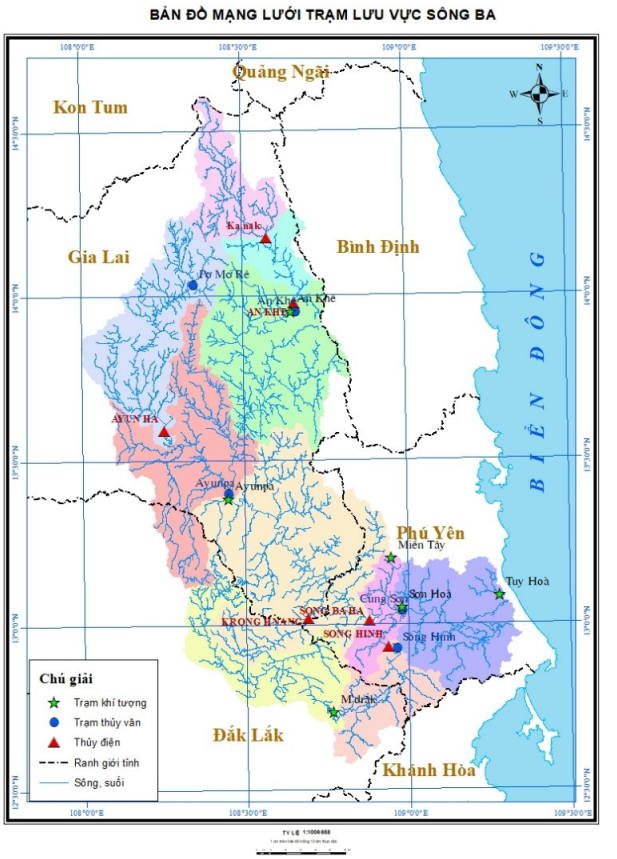
2. Phương pháp nghiên cứu và thu thập tài liệu

2.1. Giới thiệu về khu vực nghiên cứu

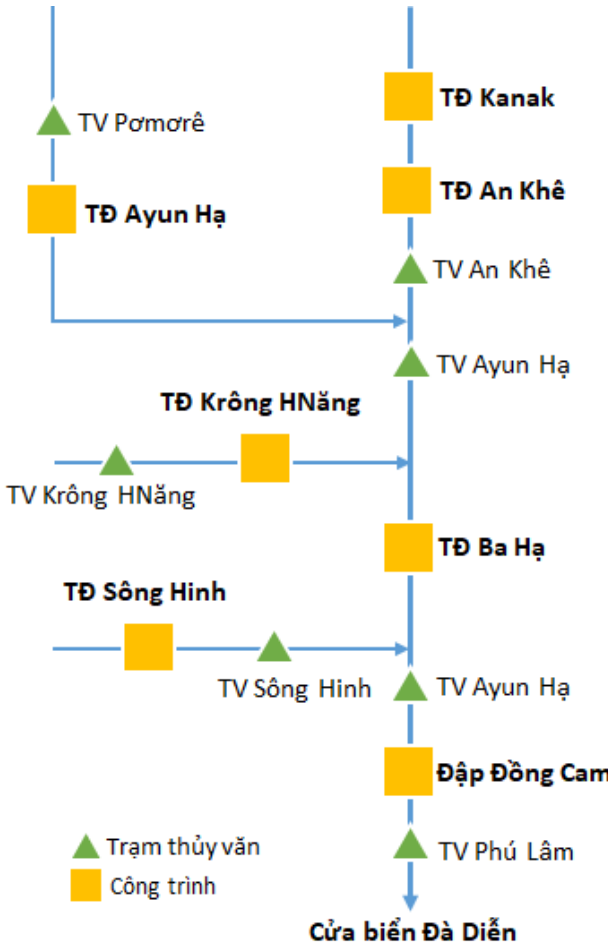
Lưu vực Sông Ba là một trong chín lưu vực sông lớn ở Việt Nam với diện tích 13.900 km2. Sông Ba nằm trong ranh giới hành chính của 20 huyện thị và 1 thành phố thuộc các tỉnh: Gia Lai, Đắc Lắc, Kon Tum, Phú Yên. Trong đó, có một huyện thuộc tỉnh Kon Tum là huyện Kông Plong, 10 huyện thị thuộc tỉnh Gia Lai là: K'bang thị xã An Khê, Đăk Pơ, Kông Chrô, Đăk Đoa, Mang Yang, Chư Sê, Ayun Pa, Krông Pa, Ea Pa, 4 huyện thuộc tỉnh Đắc Lắc là: Ea Hleo, Krông Hnăng, Ea Kar, Ma Đ'răk và 5 huyện thuộc tỉnh Phú Yên là: Sơn Hoà, Sông Hinh, Phú Hoà, Tuy hoà, thành phố Tuy Hoà..

Với ảnh hưởng của dãy Trường Sơn, lưu vực sông Ba cũng là nơi có điều kiện khí tượng, thủy văn tương đối phức tạp. Ở khu vực Tây Trường Sơn, tổng lượng mưa trung bình năm đều nhỏ hơn 2000 mm, và biến đổi trong khoảng từ 1192 - 2186mm với mùa mưa kéo dài 6 tháng từ tháng V đến tháng X trùng với mùa gió mùa Tây Nam hoạt động. Trong khi đó ở khu vực Đông Trường Sơn, mùa mưa chỉ từ 3-4 tháng từ tháng IX đến tháng XI hoặc XII hàng năm cùng với thời kỳ gió mùa Đông Bắc và bão muộn trên Biển Đông với lượng mưa trung bình đạt 1700 - 2000mm. Khu vực trung gian là khu vực có lượng mưa ít nhất (1294-1618mm) do chịu tác động qua lại của khí hậu Tây và Đông Trường Sơn.

Lưu vực Sông Ba cũng là nơi có mạng lưới sông suối dầy đặc với 36 sông cấp 1, 54 sông cấp 2, 14 sông cấp 3 và một số sông cấp 4. Các sông suối thuộc lưu vực sông Ba thường hẹp và sâu, độ dốc sông suối lớn nên có tiềm năng lớn về nguồn thuỷ năng. Do đó, đã có rất nhiều hồ chứa thủy điện, thủy lợi đã được xây dựng để phục vụ khai tác tài nguyên nước và tài nguyên năng lượng trên lưu vực. Năm công trình hồ chứa lớn trên lưu vực bao gồm An Khê - Kanak, Ayun hạ, Krong Hnăng, sông Ba hạ, sông Hinh (Hình 2). Trong đó, Hồ Sông Hinh, với diện tích 772km^2 và dung tích 323 triệu m^3 để thử nghiệm khả năng dự báo lưu lượng đến hồ của các mô hình AI, do trong hệ thống hồ chứa trong lưu vực, đây là hồ chứa độc lập, không chịu tác động của điều tiết liên hồ chứa, và có số liệu quan trắc lưu lượng đến hồ tương đối đầy đủ để phục vụ cho quá trình huấn luyện và kiểm tra mô hình.



Hình 1. Mạng lưới trạm khí tượng, thủy văn trên lưu vực sông Ba



Hình 2. Sơ đồ hệ thống hồ chứa thủy điện chính trên lưu vực sông Ba

2.2. Giới thiệu về mô hình trí tuệ nhân tạo

2.2.1.Mô hình Support Vector Regression

Mô hình Support Vector Regression (SVR) là mô hình với cơ chế hồi quy của mô hình Support Vector Machine (SVM) - một thuật toán học máy có giám sát được đề xuất lần đầu tiên bởi Vladimir N. Vapnik [23] và được sử dụng rộng rãi trong việc giải quyết các bài toán phi tuyến tính. Thuật toán SVM bao gồm hai bước chính. Đầu tiên, dữ liệu đầu vào sẽ được ánh lên không gian nhiều chiều hơn sử dụng các kernel trick, ở đó việc tìm kiếm siêu phẳng tối ưu được chứng minh là đơn giản hơn [24]. Sau đó, thuật toán sẽ tìm kiếm siêu phẳng để phân tách dữ liệu thông qua việc đánh giá khoảng cách từ các điểm dữ liệu ánh xạ đến siêu phẳng này.

Với tập dữ liệu huấn luyện là $\{X_i, Y_i\}_{i=1}^{I}$, trong đó *I* là số lượng điểm dữ liệu, giả sử có một hàm f(x) tồn tại miêu tả mối quan hệ phi tuyến giữa biến x_i và y_i như sau:

$$f(x)=(w.\varphi(x))+b \quad (1)$$

Trong đó φ(x) là hàm ánh xạ dữ liệu đầu vào lên không gian đa chiều; w là vectơ trọng số, và b là hệ số thiên lệch [25]. Như vậy, để tìm ra siêu phẳng, cần phải tối đa hóa được khoảng cách (*margin*) giữa các lớp dữ liệu với nhau theo w và b, như ở phương trình dưới đây:

$$\min\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{I}\xi_i + \xi_i^*\right) \quad (2)$$

Với điều kiện

$$\begin{cases} y_i - (w.\varphi(x) + b) \le \epsilon + \xi_i \\ (w.\varphi(x) + b) - y_i \le \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, \ \ i = 1, \dots, I \end{cases} \quad (3)$$

Trong đó C > 0, được xác định bởi người lập trình, là hằng số điều chỉnh sự đánh đổi giữa giá trị của hàm mục tiêu sự hy sinh; ξ_i và ξ_i^* là các biến bù, xác định khoảng cách giới hạn cho phép từ biến dung sai ϵ. Áp dụng nhân tử Lagrange vào phương trình số (1), ta có:

$$f(x) = \sum_{i=i}^{I}(a_i - a_i^*)K(x, x_i) + b \quad (4)$$

Trong đó a_i và a_i^* là các nhân tử Lagrange, K là hàm nhân (kernel function). Khai triển dạng toàn phương của phương trình (3) như sau:

$$W(a_i, a_i^*) = \sum_{i=1}^{I} y_i(a_i - a_i^*) - \epsilon\sum_{i=1}^{I}(a_i + a_i^*) - \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{I}(a_i - a_i^*)(a_j - a_j^*)K(x_i, x_j) \quad (5)$$

Với điều kiện:

$$\sum_{i=1}^{I}(a_i - a_i^*) = 0 \quad (6)$$
$$0 \le a_i \le C, \ \ i = 1, \dots, I$$
$$0 \le a_{i^*} \le C, \ \ i = 1, \dots, I$$

Các hàm nhân phổ biến là Linear, Polynomial và Gaussian và Sigmoid, đã được thử nghiệm trong nghiên cứu này có phương trình lần lượt như sau:

$$\begin{cases} K(x,x_i) = x.x_i \\ K(x,x_i) = (\gamma(x.x_i)+r)^d \\ K(x,x_i) = \exp(-\gamma|x-x_i|^2) \\ K(x,x_i) = \tanh(\gamma(x.x_i)+r \end{cases} \quad (7)$$

2.2.2. Mô hình Random Forest

Trong những năm gần đây, cây ra quyết định (decision tree) là một trong những mô hình học máy được sử dụng rất rộng rãi do sự đơn giản trong việc thiết lập và khả năng giải thích của nó. Tuy nhiên, mô hình này vẫn tồn tại một số hạn chế ví dụ như ovcrfitting hay nhạy cảm với số lượng dữ liệu [26]. Random Forest (RF) là một trong những phương pháp được đề xuất để giải quyết các vấn đề nêu trên. Đây một mô hình học có giám sát sử dụng cho các bài toán phân loại và hồi quy được đề xuất bởi Breiman vào năm 2001 [8]. RF 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.

Giả sử có một tập dữ liệu đầu vào $X=x_1,x_2,x_3,\ldots,x_n$ trong đó n là số chiều dữ liệu hay số biến dự báo. Một mô hình RF sẽ là một tập hợp T cây $T_1(X),T_2(X),T_3(X),\ldots,T_n(X)$. Kết quả dự báo của các cây ra quyết định này là $\hat{Y}_1, \hat{Y}_2, \hat{Y}_3, \ldots, \hat{Y}_n$. Đối với bài toán hồi quy, kết quả cuối cùng của mô hình RF sẽ là trung bình của tất cả các kết quả dự báo của các cây trên.

Việc phát triển các cây ra quyết định (*tree growing*) được thực hiện nguyên tắc chia ra các tập huấn luyện ban đầu ra các tập huấn luyện nhỏ hơn, và trong mỗi lần phân chia chỉ một số biến dự báo được lựa chọn một cách ngẫu nhiên. Các cây ra quyết định được phát triển mở rộng liên tục mà không bị cắt tỉa (*pruning*) đến một giới hạn (*stoping criteria*) định trước bởi lập trình viên. Các giới hạn dừng phát triển cây thường được sử dụng là *Root Mean Squared Error, Gini Diversity Index*, hay *Mean Square Error*. Sau đó, các cây có kết quả dự báo thấp sẽ bị loại bỏ, và chỉ những cây có giá trị dự báo đủ điều kiện được lựa chọn trong mô hình RF cuối cùng. Việc lựa chọn ngẫu nhiên các biến dự báo và tập hợp kết quả của các cây ra quyết định sẽ loại bỏ được vấn đề overfitting của mô hình cây ra quyết định đơn lẻ [8], [27].

2.3. Lựa chọn số liệu đầu vào

Lựa chọn số liệu đầu vào (*feature selection*) là một bước rất quan trọng trong việc xây dựng mô hình AI hay ML. Mục tiêu của việc lựa chọn các biến đầu vào cho mô hình bao gồm: cải thiện hiệu quả dự báo của mô hình, tăng tốc độ tính toán của mô hình, và để hiểu rõ hơn các quá trình ẩn đằng sau [28].

Với mục tiêu xây dựng và đánh giá khả năng dự báo của các mô hình AI cụ thể là hai mô hình SVM và RF trong dự báo lưu lượng đến hồ chứa sông Hinh, các mô hình này lần lượt được thử nghiệm đối với dự báo lưu lượng trung bình 3 ngày, trung bình 7 ngày và trung bình 1 tháng tương ứng với với dự báo lưu lượng ngắn hạn, trung hạn và dài hạn trong các bài toán dự báo. Ba trường hợp này sau đây ký hiệu là TH1, TH2 và TH3.

Các số liệu mưa và bốc hơi trung bình ngày tại các trạm Củng Sơn, Tuy Hòa, Sơn Hòa, và Mdrak cùng với số liệu lưu lượng vào hồ Sông Hinh (sau đây gọi là trạm Sông Hinh) từ năm 11/1999 (năm bắt đầu vận hành hồ) đến năm 12/2017 đã được tổng hợp.

Do không có tiêu chuẩn chung cho việc lựa chọn số liệu trong các mô hình AI, trong nghiên cứu này, hệ số tương quan r giữa các biến mưa, bốc hơi và dòng chảy ở kỳ trước so với dòng chảy ở kỳ dự báo (Q_t) được xem xét. Các biến được chọn là các biến có hệ số tương quan $r \geq 0.5$. Tuy nhiên, vẫn có một số ngoại lệ như đối với số liệu bốc hơi có $r \leq 0.5$ vẫn được lựa chọn một cách chủ quan để đảm báo số chiều của bộ dữ liệu đầu vào và cũng để tăng khả năng giải thích của mô hình.

Các biến đầu vào được chọn tương ứng với các trường hợp tính toán như sau:

TH1: Sử dụng số liệu mưa và bốc hơi của kỳ tính toán ($P_{(t)}$ và $E_{(t)}$) tại các trạm Tuy Hòa, Sơn Hòa, Mdrak, Củng Sơn, số liệu lưu lượng trung bình của 2 kỳ trước đó tại trạm Sông Hinh ($Q_{(t-2)}$ và $Q_{(t-1)}$), và số liệu lưu lương lớn nhất và nhỏ nhất của kỳ trước đó ($Q_{max(t-1)}$, $Q_{min(t-1)}$)

TH2: Sử dụng số liệu mưa và bốc hơi của kỳ tính toán ($P_{(t)}$ và $E_{(t)}$) tại các trạm Tuy Hòa, Sơn Hòa, Mdrak, Củng Sơn, số liệu lưu lượng trung bình, lớn nhất và nhỏ nhất của kỳ trước đó tại trạm Sông Hinh ($Q_{(t-1)}$,$Q_{max(t-1)}$, $Q_{min(t-1)}$).

TH3: Sử dụng số liệu mưa và bốc hơi của kỳ tính toán ($P_{(t)}$ và $E_{(t)}$) tại các trạm Tuy Hòa, Sơn Hòa, Mdrak, Củng Sơn, số liệu lưu lượng trung bình của kỳ trước đó tại trạm Sông Hinh ($Q_{(t-1)}$)

Số liệu ở kỳ trước nêu trên được hiểu như sau: Giả sử ta có chuỗi số liệu lưu lượng đến hồ Sông Hinh có giá trị y_i,$y_{(i+1)}$,$y_{(i+2)}$,…,y_n. Đối với TH1, chuỗi số liệu ban đầu sẽ được chuyển thành chuỗi số liệu 3 ngày bằng cách tính giá trị trung bình của 3 giá trị kế tiếp nhau tạo nên chuỗi số liệu mới Y_j,$Y_{(j+1)}$,$Y_{(j+2)}$…,Y_N với Y_j=mean (y_i,$y_{(i+1)}$,$y_{(i+2)}$), $Y_{(j+1)}$=mean($y_{(i+3)}$,$y_{(i+4)}$,$y_{(i+5)}$)… Như vậy, theo như trình bày ở trên, giả sử muốn dự báo lưu lượng tại thời điểm dự báo t có giá trị $Y_{(j+2)}$ ta phải sử dụng số liệu lưu lượng tại thời điểm kỳ trước (t-1) và (t-2) tương ứng với hai giá trị Y_j và $Y_{(j+1)}$. Cách tính toán này được thực hiện tương tự cho các trường hợp khác.

2.4. Phương pháp đánh giá mô hình

2.4.1. Chỉ số thống kê

Để đánh giá hiệu quả dự báo của của các mô hình, nghiên cứu này đã sử dụng các chỉ số đánh giá mô hình bao gồm *Nash - Sutcliffe Efficiency* (NSE) [29] và chỉ số sai số căn quân phương (*RMSE - Root Mean Square Error*).

NSE là chỉ số thống kê thường được sử dụng để đánh giá chất lượng của các mô hình thủy văn. Chỉ số này được tính toán theo công thức sau:

$$NSE = \left[\frac{\sum_{i=1}^{n}\left(Y_i^{obs} - Y_i^{sim}\right)^2}{\sum_{i=1}^{n}(Y_i^{obs} - Y^{mean})^2}\right] \quad (8)$$

Trong đó Y_i^{obs} là giá trị dòng chảy thực đo tại thời điểm i; Y_i^{sim} là giá trị dòng chảy tính toán/mô phỏng tại thời điểm i; Y^{mean} là giá trị trung bình của dòng chảy thực đo; n là tổng số giá trị thực đo.

NSE có giá trị trong khoảng -"∞" đến 1, với NSE = 1 là giá trị tối ưu nhất, chỉ ra sự tương đồng tuyệt đối giữa giá trị thực đo và tính toán. Trong khi đó, NSE ≤ 0 chỉ ra rằng kết quả mô phỏng/tính toán là không chấp nhận được. Theo Moriasi và nnk (2007), chỉ số NSE ≥ 0,5 được gọi là chấp nhận được đối với các mô hình dự báo theo tháng. Trong nghiên cứu này, khoảng giá trị này cũng được áp dụng cho cả ba trường hợp tính toán.

Tương tự như NSE, RMSE cũng được nhiều nghiên cứu về áp dụng mô hình dự báo sử dụng. RMSE cũng là được sử dụng như là một hàm mục tiêu để tối ưu hóa các mô hình AI. Công thức tính toán chỉ số RMSE như sau:

$$RMSE = \sqrt{\left(\frac{\sum_{i=1}^{n}\left(Y_i^{obs} - Y_i^{sim}\right)^2}{n}\right)} \quad (9)$$

Các chỉ số này được sử dụng để đánh giá quá trình dòng chảy và dòng chảy theo hai mùa khô và mùa mưa. Ngoài ra, kết quả dự báo đỉnh lũ tiêu biểu của các năm trong thời gian kiểm tra cũng được đánh giá theo tỷ lệ thay đổi giữa giá trị dự báo và giá trị thực đo. Các đánh giá này nhằm mục đích so sánh khả năng dự báo của hai mô hình AI qua đó lựa chọn được mô hình phù hợp cho các trường hợp tính toán.

2.5. Thiết lập mô hình

Để áp dụng mô hình RF và SVR trong dự báo lưu lượng tới hồ Sông Hinh, nghiên cứu này đã sử dụng thư viện *Scikit-learn* chạy trên nền Python 3.6.

Bộ số liệu đầu vào của 2 mô hình ở 3 trường hợp tính toán được chia làm hai phần: huấn luyện và kiểm tra. Số liệu từ tháng 11/1999 - 31/12/2013 được dùng để huấn luyện các mô hình, phần còn lại của bộ số liệu từ 01/01/2014 - 31/12/2017 được dùng để kiểm nghiệm mô hình.

Do số liệu phần huấn luyện tương đối ngắn và để tránh tình trạng overfit của mô hình, nghiên cứu đã sử dụng phương pháp kiểm định chéo nhiều lớp (*k-fold cross validation*) do tính đơn giản và hiệu quả cao trong việc sử dụng. Đầu tiên, số liệu huấn luyện sẽ được chia làm k phần nhỏ. Sau đó, một phần của bộ số liệu được giữ lại để kiểm tra, các phần còn lại (k-1) sẽ được sử dụng để huấn luyện. Quá trình này diễn ra liên tục cho đến khi tất cả các phần được sử dụng làm số liệu kiểm tra. Nếu kết quả dự báo ở mỗi phần

là tốt và tương đồng nhau thì mô hình sẽ phù hợp để áp dụng cho dữ liệu kiểm tra nêu trên. Thực tế triển khai cho thấy, việc thay đổi giá trị k không mang lại kết quả khác biệt đáng kể, do đó các giá trị k = 15, 10, 5 được sử dụng cho TH1, TH2 và TH3 theo thứ tự đó.

Nhằm đánh giá hiệu quả của các mô hình, các thông số chính của hai mô hình sẽ được tối ưu bằng công cụ *GridSearchCV* sẵn có trong thư viện *scikit-learn*. *GridSearchCV* sẽ áp dụng các bộ thông số khác nhau của các mô hình được thiết lập trước lập trình viên qua đó tìm được bộ thông số tối ưu của các mô hình. Số lần kiểm định chéo k của phương pháp *k-fold validation* cũng được thiết lập trong công cụ này.

3. Kết quả và thảo luận

Sau khi được hiệu chỉnh bằng *GridSearchCV*, các thông số tối ưu của mô hình được trình bày trong Bảng 1 dưới đây.

Bảng 1. Các thông số tối ưu của các mô hình trong 3 trường hợp tính toán

Thông số	Mô hình SVR			Thông số	Mô hình RF		
	TH1	TH2	TH3		TH1	TH2	TH3
kernel	rbf	rbf	rbf	*n_estimators*	50	50	50
gamma	0,01	0,01	0,01	*max_depth*	8	8	15
C	5	5	10				
epsilon	0,1	0,1	0,1				

Sau khi có được bộ thông số tối ưu, các mô hình được áp dụng cho bộ dữ liệu kiểm tra từ tháng 01/2014 đến tháng 12/2017. Đây là chuỗi dữ liệu mà mô hình chưa "nhìn thấy" (*unseen data*), do đó kết quả dự báo của mô hình trên chuỗi dữ liệu này sẽ được dùng để đánh giá hai mô hình thử nghiệm trong nghiên cứu. Các nội dung đánh giá bao gồm: (i) đánh giá kết quả dự báo quá trình dòng chảy; (ii) đánh giá kết quả dự báo theo mùa; (iii) đánh giá kết quả dự báo đỉnh lũ tiêu biểu.

3.1.1 Kết quả dự báo quá trình dòng chảy

Kết quả tính toán cho thấy diễn biến dòng chảy trong giai đoạn kiểm tra được cả hai mô hình dự báo với độ chính xác cao (Hình 2). Các chỉ số thống kê đều đạt mức tốt với NSE dao động từ 0,84 - 0,93 và RMSE dao động từ 31,98 đến 60,24 (Bảng 2). Có thể thấy rằng các mô hình cho kết quả dự báo chính xác hơn ở TH2 và TH3.

Nhìn vào chi tiết, có thể thấy ở TH1, các giá trị đỉnh lũ dự báo lại chưa đạt được độ chính xác cao, đặc biệt là đối với các đỉnh lũ ở cuối năm 2015 trở đi. Mặt khác, dòng chảy cạn được các mô hình dự báo khá tốt, đặc biệt là ở mô hình SVR. Ở TH2 và TH3, các giá trị đỉnh lũ đã được dự báo chính xác hơn, mặc dù vẫn có nhưng sai số đáng kể (TH3).

Bảng 2. Tổng hợp kết quả đánh giá khả năng dự báo quá trình dòng chảy của hai mô hình

	NSE			RMSE		
	TH1	TH2	TH3	TH1	TH2	TH3
SVR	0,85	0,89	0,93	53,37	45,65	30,88
RF	0,84	0,92	0,91	60,24	40,91	31,98

3.1.2. Kết quả dự báo theo mùa

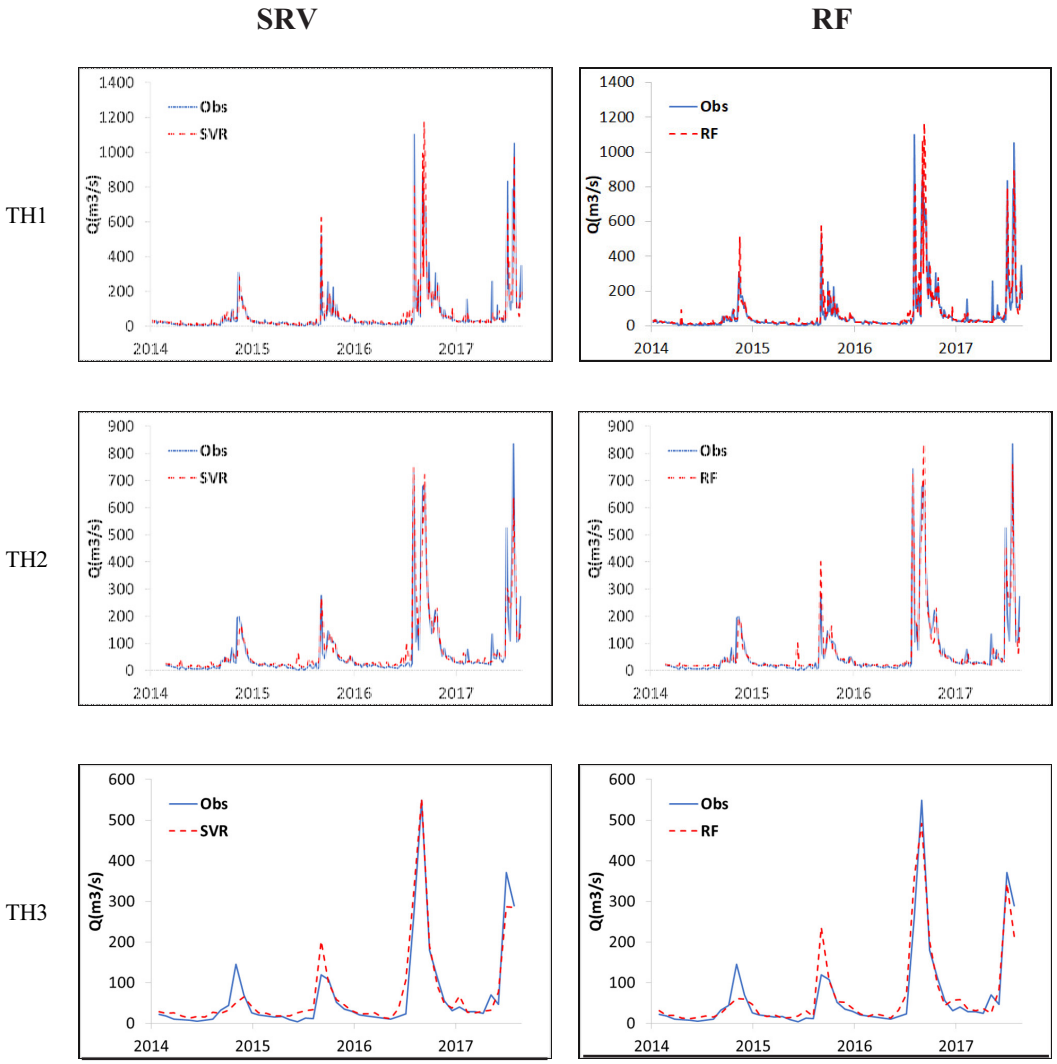
Trên thực tế, việc đánh giá kết quả dự báo theo mùa được thực hiện cho hai giai đoạn: tháng 1 - 5, giai đoạn khô hạn nhất và tháng 9 - 12, giai đoạn xảy ra nhiều trận lũ nhất, của giai đoạn kiểm định (2104 - 2017), sau đây gọi là mùa mưa và mùa khô.

Bảng 3. Tổng hợp kết quả đánh giá khả năng dự báo theo mùa của hai mô hình

Mùa		NSE			RMSE		
	Mô hình	TH1	TH2	TH3	TH1	TH2	TH3
Mùa khô	SVR	0,89	0,90	0,93	15,24	14,68	11,01
	RF	0,85	0,86	0,90	18,08	18,17	12,79
Mùa mưa	SVR	0,81	0,87	0,90	89,82	67,29	48,45
	RF	0,79	0,90	0,88	95,77	57,10	53,78

Kết quả kiểm nghiệm cho thấy, các mô hình cho kết quả dự báo mùa khô tốt hơn so với mùa mưa, thể hiện qua chỉ số NSE đều trên 0,85 và RMSE đều nhỏ hơn 20m^3/s, trong khi đó RMSE cho mùa mưa đều ở mức tương đối cao từ 48,45 - 95,77 m^3/s (Bảng 3).

Trong cả ba trường hợp, mô hình SVR chiếm ưu thế khi cho kết quả dự báo tốt hơn, đặc biệt là trong mùa khô, chỉ duy nhất ở TH2 mô hình RF có kết quả dự báo tốt hơn trong mùa mưa.



Hình 3. Kết quả dự báo lưu lượng vào hồ của hai mô hình SVR và RF theo 3 trường hợp tính toán trong giai đoạn kiểm tra từ 01/2014 - 12/2017

3.1.3. Kết quả dự báo đỉnh lũ tiêu biểu

Do tầm quan trọng của công tác phòng chống lũ lụt và phục vụ cho công tác điều tiết hồ chứa và phát điện, bên cạnh việc dự báo được xu hướng tổng dòng chảy, dòng chảy theo mùa, thì việc dự báo được chính xác cường độ hay giá trị của đỉnh lũ là một yếu tố rất quan trong trong đánh giá hiệu quả của một mô hình. Trong nghiên cứu này, nhóm lựa chọn 4 trận lũ tiêu biểu tương ứng với 4 đỉnh lũ trong 4 năm từ 2014 - 2017 để so sánh kết quả của các mô hình. Kết quả so sánh được trình bày ở Bảng 4.

Bảng 4. So sánh độ lớn đỉnh lũ tiêu biểu dự báo và thực đo

Năm	Mô hình	Trường hợp tính toán								
		TH1			TH2			TH3		
		Thực đo [m^3/s]	Tính toán [m^3/s]	Sai số (%)	Thực đo [m^3/s]	Tính toán [m^3/s]	Sai số (%)	Thực đo [m^3/s]	Tính toán [m^3/s]	Sai số (%)
2014	RF	307,439	257,091	-16,38	199,184	193,186	-3,01	145,261	61,334	-57,78
	SVR		124,737	-59,43		139,15	-30,14		49,9	-65,65
2015	RF	524,383	573,473	9,36	277,011	401,55	44,96	119,099	236,616	98,67
	SVR		624,384	19,07		266,074	-3,95		202,11	69,7
2016	RF	1101,694	680,78	-38,21	742,589	730,153	-1,67	548,305	492,147	-10,24
	SVR		819,289	-25,63		747,43	0,65		551,954	0,67
2017	RF	1050,814	889,213	-15,38	835,633	759,355	-9,13	370,447	343,731	-7,21
	SVR		982,701	-6,48		635,931	-23,90		285,952	-22,81

Theo kết quả tính toán sai số của đỉnh lũ dự báo, có thể thấy rằng chưa có mô hình nào thể hiện sự vượt trội về khả năng dự báo đỉnh lũ, khi sai số của từng mô hình đối với từng đỉnh lũ và trường hợp dự báo lại tương đối khác nhau. Ở TH1, các đỉnh lũ ở năm 2015 và 2017 được dự báo khá chính xác với sai số từ -6,48% - 19,07%, trong khi các đỉnh lũ ở năm 2014 và 2016 chưa được dự báo tốt. TH2 là trường hợp có kết quả dự báo tốt nhất với sai số khá nhỏ từ-0,65% của mô hình SVR và -1,67% của mô hình RF ở năm 2016, hay sai số -3,95% của SVR ở năm 2015 và -3,01% của RF ở năm 2014. Ở TH3, mặc dù các mô hình cho kết quả dự báo rất tốt ở các đỉnh lũ năm 2016 và 2017, nhưng ở hai năm đầu của chuỗi dữ liệu kiểm tra, các mô hình đều đưa ra kết quả dự báo có độ sai số cao đáng kể. Điều này có thể là do mô hình đang trong quá trình warm up.

Nhìn chung, xét về dự báo lưu lượng đỉnh lũ, mô hình RF cho kết quả dự báo tốt hơn SVR ở TH1 và TH2 với sai số tuyệt đối là 19,83% và 14,69%, trong khi SVR là sự lựa chọn tốt hơn ở TH3 với sai số tuyệt đối là 20,39%.

4. Kết luận

Nghiên cứu đã bước đầu thử nghiệm thành công hai mô hình AI là SVR và RF trong dự báo lưu lượng đến hồ, áp dụng cho hồ Sông Hinh thuộc lưu vực sông Ba. Ba trường hợp tính toán là dự báo báo dòng chảy trung bình 3 ngày, trung bình 7 ngày và trung bình 1 tháng tương ứng với dự báo ngắn hạn, trung hạn và dài hạn, đã được thử nghiệm Kết quả cho thấy, cả hai mô hình ở cả ba trường hợp đều cho kết quả có độ chính xác khá cao đặc biệt là đối với trường hợp dự báo lưu lượng trung bình 7 ngày và 1 tháng. Trong 2 mô hình được thử nghiệm thì mô hình SVR nhìn chung cho kết quả tốt nhất đối với dự báo ngắn và dài hạn, trong khi đó mô hình RF lại cho thấy sự vượt trội ở dự báo trung hạn. Đối với dự báo theo mùa, các mô hình cho kết quả dự báo tốt trong cả mùa khô (tháng 1-5) và mùa mưa (tháng 9-12) với kết quả nhỉnh hơn trong mùa khô một điểm đáng chú ý là, các mô hình AI đều không dự báo chính xác một cách đồng nhất dòng chảy lũ. Lý do của hiện tượng này là các mô hình không được huấn luyện tập trung vào dự báo dòng chảy lũ mà ưu tiên vào quá trình dòng chảy. Kết quả tính toán ở trường hợp dự báo dòng chảy trung bình 3 ngày có độ chính xác thấp hơn đáng kể so với hai trường hợp còn lại, điều này là do ở bước thời gian này sự dao

động trong dữ liệu cao hơn các trường hợp dữ liệu trung bình tuần hay tháng. Kết quả dự báo có thể được cải thiện nếu có dữ liệu có độ dài và chất lượng tốt hơn.

Ngoài ra, việc lựa chọn dữ liệu đầu vào phù hợp là yếu tố rất quan trọng quyết định nên hiệu quả dự báo của mô hình. Trong đó, dòng chảy trong quá khứ đóng là một trong những biến đầu vào quan trọng. Bên cạnh đó, số lượng dữ liệu đầu vào (số lượng features) cần phải đủ lớn để hỗ trợ cho mô hình AI trong việc khái quát hóa được mối quan hệ giữa biến đầu vào và dòng chảy đầu ra.

Dựa trên những phân tích và đánh giá đã thực hiện, nhóm nghiên cứu đề xuất sử dụng mô hình SVR cho dự báo lưu lượng trung bình 3 ngày và 1 tháng, và RF cho dự báo lưu lượng trung bình 1 tuần. Tuy nhiên, đối với từng trường hợp dự báo, hay yêu cầu dự báo (đỉnh lũ, dòng chảy mùa khô, dòng chảy mùa mưa) có thể sử dụng các mô hình khác thay thế cho các mô hình được đề xuất do hiệu quả dự báo là khá tương đồng như đã phân tích ở các phần trên.

Như vậy, bên cạnh các phương pháp truyền thống, các mô hình AI như SVR và RF sẽ cung cấp một công cụ mới, hiệu quả để bổ trợ cho công tác quản lý và vận hành hồ chứa nói chung và hồ sông Hinh nói riêng. Tuy nhiên, việc ứng dụng trong tương lai phụ thuộc rất lớn vào điều kiện và chất lượng số liệu ở địa điểm áp dụng.

Lời cảm ơn: *Kết quả nghiên cứu thể hiện trong bài báo này là một phần sản phẩm của đề tài nghiên cứu khoa học và công nghệ cấp sơ sở: "Nghiên cứu cơ sở khoa học áp dụng công nghệ trí tuệ nhân tạo dự báo lưu lượng vào hồ chứa áp dụng cho lưu vực sông Ba". Nhóm tác giả xin chân thành cảm ơn Ban Lãnh đạo Viện Khoa học tài nguyên nước đã tạo điều kiện để chúng tôi thực hiện nghiên cứu này.*

Tài liệu tham khảo

1. Elsafi, S.H., (2014), *Artificial Neural Networks (ANNs) for flood forecasting at Dongola Station in the River Nile*, Sudan. Alexandria Eng. J., 53 (3), 655-662.

2. VanderKwaak, J.E, Loague K., (2001), *Hydrologic-Response simulations for the R-5 catchment with a comprehensive physics-based model*. Water Resour. Res., 37 (4), 999-1013.

3. Nayak, P.C, Sudheer K.P, Rangan, D.M, Ramasastri, K.S., (2005), *Short-term flood forecasting with a neurofuzzy model.* Water Resour. Res., 41 (4).

4. Mosavi, A., Ozturk, P., (2018), *Flood Prediction Using Machine Learning, Literature Review*. Water, 1-40, 2018.

5. Jain, S.K., Das, A., Srivastava, D.K., (1999), *Application of ANN for Reservoir Inflow Prediction and Operation*, J. Water Resour. Plan. Manag., 125 (5), 263-271.

6. Maier, H.R., Dandy, G.C., (1996), *The Use of Artificial Neural Networks for the Prediction of Water Quality Parameters,* Water Resour. Res., 32 (4), 1013-1022.

7. Mosavi, A., Rabczuk, T., Varkonyi-Koczy, A.R., (2018), *Reviewing the Novel Machine Learning Tools for Materials Design*, Springer, 50-58.

8. Breiman, L., (2001), *Random Forests*, Statistics ,45 (1), 1-33.

9. Yang, T., Asanjan, A.A., Welles E., Gao, X., Sorooshian, S., Liu, X., (2017), *Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information,* Water Resour. Res., 53 (4), 2786-2812.

10. Li, B., Yang, G., Wan, R., Dai, X., Zhang, Y., (2016), *Comparison of random forests and other statistical methods for the prediction of lake water level: a case study of the Poyang Lake in China,* Hydrol. Res., 47 (S1), 69-83.

11. Obringer, R., Nateghi, R., (2018), *Predicting Urban Reservoir Levels Using Statistical Learn-*

ing Techniques, Sci. Rep., 8 (1), 5164.

12. Erdal, H.I., Karakurt, O., (2013), *Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms*, J. Hydrol., 477, 119-128.

13. Senthil Kumar, M.K., Goyal, A.R., Ojha, C.S.P., Singh, R.D., Swamee, P.K., (2013), *Application of artificial neural network, fuzzylogic and decision tree algorithms for modelling of streamflow at Kasol in India*, Water Sci. Technol., 68 (12), 2521-2526.

14. Galelli, S., Castelletti, A., (2013), *Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling*, Hydrol. Earth Syst. Sci., 17 (7), 2669-2684

15. Yang, T., Gao, X., Sorooshian, S., Li, X., (2016), *Simulating California reservoir operation using the classification and regression-tree algorithm combined with a shuffled cross-validation scheme*, Water Resour. Res., 52 (3), 1626-1651.

16. Asefa, T., Kemblowski, M., McKee, M., Khalil, A., (2006), *Multi-time scale stream flow predictions: The support vector machines approach*, J. Hydrol., 318 (1-4), 7-16.

17. Londhe, S., Gavraskar, S., (2018), *Stream Flow Forecasting using Least Square Support Vector Regression,* Soft Comput. Civ. Eng., 2 (2), 56-88.

18. Adnan, R.M., Yuan, X., Kisi, O., Adnan, M., Mehmood, A., (2018), *Stream Flow Forecasting of Poorly Gauged Mountainous Watershed by Least Square Support Vector Machine, Fuzzy Genetic Algorithm and M5 Model Tree Using Climatic Data from Nearby Station*, Water Resour. Manag., 32 (14),469-4486.

19. Maity, R., Bhagwat, R., Bhatnagar, A., (2010), *Potential of support vector regression for prediction of monthly streamflow using endogenous property*, Hydrol. Process., 24 (7), 917-923.

20. Rafidah, A., Suhaila, Y., (2013), *Modeling River Stream Flow Using Support Vector Machine*, Appl. Mech. Mater., 315, 602-605.

21. Lin, J., Cheng, C., Chau, K., (2006), *Using support vector machines for long-term discharge prediction Using support vector machines for long-term discharge prediction,* Hydrol. S, 51(4), 599-612.

22. Guo, J., Zhou, J., Qin, H., Zou, Q., Li, Q., (2011), *Monthly streamflow forecasting based on improved support vector machine model*, Expert Syst. Appl., 38(10), 13073-13081.

23. Vapnik V. N., (1995), *The nature of statistical learning theory*. *Springer.*

24. Boser, B.E., Guyon, I.M., Vapnik, V.N., (1992), *A training algorithm for optimal margin classifiers, in Proceedings of the fifth annual workshop on Computational learning theory - COLT* , 92, 144-152.

25. Vapnik, V.N., (1999), *An overview of statistical learning theory, IEEE Transactions on Neural Networks*, 10 (5), 988-999.

26. Gupta, B., Rawat, A., Jain, A., Arora, A., Dhami, N., (2017), *Analysis of Various Decision Tree Algorithms for Classification in Data Mining*, Int. J. Comput. Appl., 163 (8), 15-19.

27. Ahmad, M.W., Mourshed, M., Rezgui, Y., (2017), *Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption*, Energy Build., 147, 77-89.

28. Guyon, I., Elisseeff, A., (2003), *An Introduction to Variable and Feature Selection*. J. Mach. Learn. Res., 3 (3), 1157-1182.

29. Nash, J.E., Sutcliffe, J.V., (1970), *River Flow Forecasting Through Conceptual Models Part I-a Discussion of Principles*. J. Hydrol., 10, 282-290.

30. Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Binger, R.L., Harmel, R.D., Veith, T.L., (2007), *Model evaluation guidelines for systematic quantification of accuracy in watershed simulations*, Trans. ASABE, 50 (3), 885-900.

APPLICATION OF ARTIFICIAL INTELLIGENCE MODELS FOR RESERVOIR INFLOW PREDICTION IN BA RIVER BASIN

Cao Hoang Hai[1], Tran Anh Phuong[1], Thai Quynh Nhu[1], Tran Manh Cuong[1]

[1]Water Resources Institute

Abstract: *In this study, two AI models namely Random Forest (RF) and Support Vector Regression (SVR) are tested for its capabilities in predicting inflow to Hinh River Reservoir in Ba River Basin, Vietnam. Three calculation scenarios are adopted including prediction of mean 3 day inflow, mean 7 day inflow, and mean 1 month inflow (corresponding to short-, mid-, and long-term prediction) to test and compare the performance of the two models. The results show that, both models present a high accuracy prediction results with mean NSE of over 0.8, particularly in mid- and long-term scenarios, NSE values are over 0.9. Of the two models, SVR, in general, yields better production results in short and long term scenario, while regarding mid-term inflow, RF is predominant one. The tested models do not show a consistent peak flows prediction since they are not trained specifically on extreme flow values, but rather focus on total flow process. On top of that, the selection of highly correlated inputs or features play an important role in improving the prediction performance of the models. Overall, these 2 models can be valuable alternatives to the existing reservoir inflow prediction approach.*

Keywords: *AI, ML, SVR, RF, Ba River, reservoir inflow prediction.*