TẠP CHÍ
KHÍ TƯỢNG THỦY VĂN

Check for updates

Bài báo khoa học

Xây dựng các mô hình hồi quy hỗ trợ véc tơ dự báo mực nước trạm Cao Lãnh, tỉnh Đồng Tháp

Lê Xuân Hòa[1], **Nguyễn Tiền Giang**[2*]

[1] Đài Khí tượng Thủy văn tỉnh Đồng Tháp, Đài Khí tượng Thủy văn khu vực Nam Bộ, Tổng cục Khí tượng Thủy văn, Bộ Tài Nguyên và Môi trường; lexuanhoakttv@gmail.com

[2] Khoa Khí tượng Thủy văn và Hải dương học, Trường Đại học Khoa học Tự nhiên, ĐHQGHN; giangnt@vnu.edu.vn

*****Tác giả liên hệ: giangnt@vnu.edu.vn; Tel.: +84–912800896

Ban Biên tập nhận bài: 15/7/2022; Ngày phản biện xong: 23/8/2022; Ngày đăng bài: 25/8/2022

Tóm tắt: Trong nghiên cứu này, ba dạng hàm kernel: Radial basis function (RBF), tuyến tính (Linear) và Sigmoid được sử dụng trong các mô hình máy học Support Vector Regression (SVR) với ba chuỗi dữ liệu đầu vào là: mực nước cao nhất ngày (H_{max}CL); mực nước thấp nhất ngày (H_{min}CL); mực nước trung bình ngày (H_{tb}CL) trong quá khứ để dự báo mực nước tương lai trạm Cao Lãnh, tỉnh Đồng Tháp. Kết quả cho thấy, các hàm nhân trong mô hình đều đưa ra kết quả dự báo với độ chính xác cao thể hiện qua chỉ số NSE > 0,95 đối với tất các các dữ liệu đầu vào khác nhau cũng như hàm nhân khác nhau trong mô hình SVR. Trong ba chuỗi dữ liệu đầu vào và các hàm nhân được thử nghiệm thì chuỗi dữ liệu H_{max}CL cho sai số là tối ưu nhất. Kết quả nghiên cứu này là tài liệu tham khảo tốt cho việc xây dựng mô hình máy học phục vụ dự báo mực nước tương lai cho trạm thủy văn Cao Lãnh, tỉnh Đồng Tháp.

Từ khóa: SVR; RBF; Tuyến tính; Sigmoid; ML; Cao Lãnh.

1. Mở đầu

Ngày nay, các nghiên cứu về dữ liệu chuỗi thời gian đem lại những ứng dụng khá quan trọng, đảm bảo tính thực tế cao trong các lĩnh vực: tài chính, thống kê, xử lý dữ liệu, dự báo các hiện tượng thiên tai,… Một số trong đó là bài toán về dự báo chuỗi thời gian kết hợp xây dựng các dự báo thích hợp. Trong các nghiên cứu về dự báo lưu lượng, dự báo dòng chảy đều sử dụng các mô hình thủy văn phân bố hay bán phân bố khác nhau. Các mô hình này được xây dựng để mô phỏng 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]. Ngoài ra các mô hình này rất tốt cho các nhà khoa học trong việc giải thích được toàn bộ quá trình ẩn đằng sau [2]. Chính vì vậy các mô hình này được áp dụng nhiều và rộng rãi ở nhiều khu vực trên thế giới. Tuy nhiên, việc sử dụng các mô hình này cần một số dữ liệu lớn về thông tin địa lý, mưa, dòng chảy… Bên cạch đó việc hiệu chỉnh và kiểm định mô hình còn khá phức tạp đòi hỏi phải có nhiều thời gian, kinh nghiệm và kiến thức của người xây dựng, chạy mô hình cho từng lưu vực. Chính vì vậy việc sử 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]. Từ những hạn chế của các mô hình truyền thống đã khuyến khích sự phát triển của các mô hình dựa vào chuỗi số liệu mà phát triển nhất

đó là phương pháp máy học (Machine Learning – ML). Các mô hình ML là một trong những công cụ rất tiềm năng trong việc dự báo dòng chảy do các mô hình ML này có thể xây dựng một cách nhanh chóng, dễ dàng mà không cần đòi hỏi 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 các bài toán về mô phỏng, dự báo dòng chảy, các mô hình trí tuệ nhân tạo như Artificial Neural Network (ANN) đã được ứng dụng từ những năm 90 [5–6]. Nhưng những năm trở lại đây, với tiến bộ vượt bậc của các ngành khoa học máy tính cùng với sự quan tâm của cộng đồng khoa học tới 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 và đa dạng hơn. Các thuật toán ANN, Random Forest (RF) và Support Vector Machine (SVM) là ba thuật toán ML được sử dụng khá rộng rãi trong các nghiên cứu về dự báo dòng chảy [7].

SVM, một thuật toán học máy có giám sát được đề xuất bởi Vapnik (1963), 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 [8–9]. 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 [10–12]. 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 [13] về dự báo dòng xả thời đoạn dài của hồ thủy điện Manwan, hay nghiên cứu của Guo và nnk [14] về dự báo dòng chảy tới khu vực đập Tam Hiệp trên sông Dương Tử. Những nghiên cứu này đều đưa ra kết quả khẳng định rằng mô hình SVR cho khả năng dự báo dòng chảy chính xác.

Đồng Tháp là một tỉnh có nhiều hệ thống sông ngòi dầy đặc, nằm ở đầu nguồn sông Tiền thượng lưu ảnh hưởng bởi dòng chảy từ bên Campuchia còn hạ lưu thì ảnh hưởng bởi thủy triều, vì vậy việc dự báo mực nước cho tỉnh Đồng Tháp nói chung, Tp Cao Lãnh nói riêng gặp rất nhiều khó khăn, mất rất nhiều thời gian và độ chính xác chưa cao. Từ những khó khăn đó việc xây dựng một mô hình máy học để phục vụ dự báo mực nước là rất cần thiết đối với các dự báo viên. Chính vì vậy các tác giả đã nghiên cứu xây dựng các mô hình hồi quy hỗ trợ véc tơ dự báo mực nước trạm Cao Lãnh, tỉnh Đồng Tháp.

2. Phương pháp nghiên cứu và số liệu sử dụng

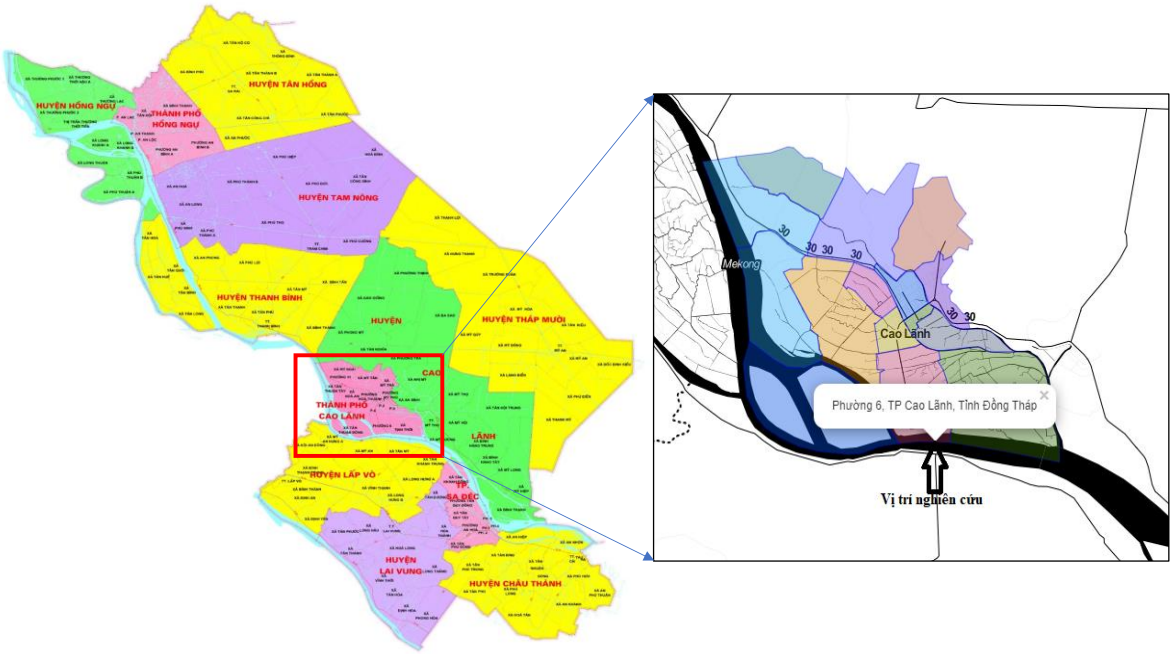
2.1. Khu vực nghiên cứu

Đồng Tháp là một trong 13 tỉnh của vùng đồng bằng sông Cửu Long, nằm ở đầu nguồn sông Tiền, lãnh thổ của tỉnh Đồng Tháp nằm trong giới hạn tọa độ 10°07’–10°58’ vĩ độ Bắc và 105°12’–105°56’ kinh độ Đông. Phía bắc giáp với tỉnh Long An, phía tây bắc giáp tỉnh Preyveng thuộc Campuchia, phía nam giáp An Giang và Cần Thơ. Tỉnh Đồng Tháp có đường biên giới quốc gia giáp với Campuchia với chiều dài khoảng 50 km từ Hồng Ngự đến Tân Hồng, với 4 cửa khẩu là Thông Bình, Dinh Bà, Mỹ Cân và Thường Phước. Hệ thống đường quốc lộ 30, 80, 54 cùng với quốc lộ N1, N2 gắn kết Đồng Tháp với thành phố Hồ Chí Minh và các tỉnh trong khu vực (Hình 1). Trạm thủy văn Cao Lãnh được đặt tại phường 6, thành phố Cao Lãnh, tỉnh Đồng Tháp, có tọa độ 10°25’0.41” vĩ độ Bắc và 105°38’38.79” kinh độ Đông, phía bắc giáp với khu dân cư, phía nam hướng ra sông Tiền, phía Đông là bến phà đang hoạt động cách trạm gần 100 m, phía tây là bãi đất trống (Hình 1).

2.2. Thuật toán SVR

Thuật toán Support Vector Regression (SVR) là thuật toán học 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 [15] 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 xạ lên

không gian nhiều chiều hơn sử dụng các hàm kernel. Sau đó, thuật toán sẽ tìm kiếm một siêu phẳng (*hyperplane*) để phân tách dữ liệu thông qua việc đánh giá khoảng cách từ các điểm dữ liệu ánh xạ đến siêu phẳng này.



Hình 1. Bản đồ hành chính tỉnh Đồng Tháp và khu vực nghiên cứu.

Ví dụ với tập dữ liệu huấn luyện là $\{X_i, Y_i\}^I_{i=1}$ trong đó I là số lượng điểm dữ liệu. Hàm ước lượng SVR có dạng

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

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 [16]. Như vậy, để tìm ra siêu phẳng, cần phải tối đa hóa được khoảng cách giữa vector gần nhất với mặt siêu phẳng 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 rằng buộc:

$$\begin{aligned} y_i - (w \times \varphi(x) + b) &\leq \varepsilon + \xi_i \\ (w \times \varphi(x) + b) - y_i &\leq \varepsilon + \xi_i^* \\ \xi_i \xi_i^* &\geq 0, i = 1, \ldots, I \end{aligned} \quad (3)$$

Trong đó C > 0, là hằng số điều chỉnh sự thay đổi giữa giá trị của hàm mục tiêu và sai số đào tạo; 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:

$$f(x) = \sum_{i=1}^{I} (a_i - a_i^*) K(x, x_i) + b \quad (4)$$

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

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

Với điều kiện:

$$\sum_{i=1}^{I} (a_i - a_i^*) = 0 \quad (6)$$

$$0 \leq a_i \leq C, i = 1, \ldots, I$$

$$0 \leq a_i^* \leq C, i = 1, \ldots, I$$

Các hàm kernel phổ biến là RBF, tuyến tính,và Sigmoid đã được thử nghiệm trong nghiên cứu này có phương trình lần lượt như sau [17]:

$$\text{Hàm RBF } K(x, x_i) = \exp(-\beta|x, x_i|^2) \quad (7)$$

$$\text{Hàm tuyến tính } K(x, x_i) = x.x_i$$

$$\text{Hàm Sigmoid } K(x, x_i) = \tanh((\gamma(x.x_i) + r)$$

2.3. Lựa chọn số liệu đầu vào

Lựa chọn số liệu đầu vào là một phần rất quan trọng trong việc xây dựng mô hình ML. Mục tiêu chính của việc lựa chọn các biến đầu vào cho mô hình gồm: Cải thiện kết quả dự báo của mô hình, tăng tốc độ tính toán, và để hiểu rõ hơn các quá trình ẩn đằng sau [18].

Với mục tiêu xây dựng và đánh giá khả năng dự báo của mô hình SVR, các hàm kernel được thử nghiệm lần lượt để dự báo mực nước tương lai trước 1 ngày cho trạm thủy văn Cao Lãnh.

Các số liệu mực nước lớn nhất ngày, nhỏ nhất ngày, trung bình ngày của trạm thủy văn Cao Lãnh từ tháng 1/2000 tới tháng 12/2020 đã được tổng hợp.

2.4. Phương pháp đánh giá mô hình

Để đánh giá được 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)* [19] và chỉ số sai số căn quân phương (*RMSE–Root Mean Square Error*) [20].

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:

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

Trong đó Y_i^{obs} là giá trị mực nước thực đo tại thời điểm i; Y_i^{sim} là giá trị mực nước tính toán/ mô phỏng tại thời điểm i; $\bar{Y}$ là giá trị trung bình của mực nước thực đo; n là độ dài chuỗi 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. Tiêu chí để đánh giá chất lượng cho chỉ số NSE có thể chia ra như sau: NSE ≤ 0,5 là xếp loại không đạt; 0,5 ≤ NSE ≤ 0,65 là xếp loại đạt yêu cầu; 0,65 ≤ NSE ≤ 0,75 là xếp loại tốt; 0,75 ≤ NSE ≤ 1 là xếp loại rất tốt [21].

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

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

Trong dó đó Y_i^{obs} là giá trị mực nước thực đo tại thời điểm i; Y_i^{sim} là giá trị mực nước tính toán/ mô phỏng tại thời điểm i.

Các chỉ số này được sử dụng để đánh giá giữa các hàm kernel trong mô hình SVR.

2.5. Thiết lập mô hình SVR

Để áp dụng mô hình SVR trong dự báo mực nước trạm thủy văn Cao Lãnh tỉnh Đồng Tháp, nghiên cứu đã sử dụng thư viện Scikit–learn chạy trên nền Python. Bộ số liệu đầu vào của mô hình được chia làm 3 phần: huấn luyện (*training*), thẩm định (*validation*) và kiểm tra (*testing*). Số liệu từ 01/01/2000 tới 12/9/2014 được dùng để huấn luyện mô hình, số liệu từ 13/9/2014 tới ngày 5/11/2017 dùng để thẩm định và phần còn lại từ 06/11/2017 tới 31/12/2020 dùng để kiểm tra.

Do số liệu phần huấn luyện còn hạn chế và để tránh tình trạng khớp quá nhiều (*overfiting*) với dữ liệu đào tạo của mô hình, nghiên cứu đã sử dụng phương pháp kiểm định chéo nhiều lớp (*k–fold cross validation*) [22]. 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, (k–1) phần còn lại sẽ được sử dụng để huấn luyện. Quá trình này diễn ra liên tục và tuần tự 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 vì vậy tác giả đã lựa chọn k = 10 thường được dùng phổ biến để áp dụng cho nghiên cứu này.

Để đánh giá hiệu quả của mô hình, các thông số chính của mô hình 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, qua đó tìm được bộ thông số tối ưu nhất của các hàm kernel trong mô hình. Từ các thiết lập đó, chúng ta chạy các hàm kernel của mô hình để tìm ra hàm kernel tối ưu nhất.

3. Kết quả và thảo luận

Sau khi chạy hiệu chỉnh GridSearchCV, các thông số tối ưu và các chỉ số đánh giá trong qua trình huấn luyên và thẩm định của các mô hình được thể hiện ở Bảng 1.

Bảng 1 cho thấy chuỗi số liệu cho máy học (training) có chỉ số tương quan rất lớn, ở mọi hàm kernel đều cho hệ số tương quan lớn hơn 0,97; thấp nhất là 0,974 ở chuỗi dữ liệu mực nước thấp nhất và cao nhất là 0,976 ở chuỗi mực nước cao nhất. Điều này cho thấy việc máy học đạt kết quả cao. Còn ở chuỗi thẩm định thì hệ số tương quan cho kết quả cũng khá tốt (R^2 > 0,95). Tương quan thấp nhất là 0,952 của hàm rbf của chuỗi dữ liệu mực nước thấp nhất và tương quan cao nhất là chuỗi dữ liệu mực nước lớn nhất với hàm rbf.

Bảng 1. Giá trị các thông số tối ưu của các hàm kernel trong mô hình SVR.

	HmaxCL			**HminCL**			**HtbCL**		
	RBF	**Linear**	**Sigmoid**	**RBF**	**Linear**	**Sigmoid**	**RBF**	**Linear**	**Sigmoid**
C	1000,0	501,18	1000,0	1000,0	3,98107	501,187	501,187	3,98107	251,188
Gamma	0,002	0,004	0,004	0,002	0,001	0,008	0,004	0,001	0,0158
Epsilon	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1
R^2 (máy học)	0,976	0,976	0,976	0,974	0,975	0,976	0,975	0,975	0,975
R^2(thẩm định)	0,962	0,958	0,960	0,952	0,954	0,955	0,958	0,958	0,958

Sau khi có được bộ thông số tối ưu này, ta sử dụng để chạy kiểm tra các hàm kernel của mô hình trong chuỗi dữ liệu từ 06/11/2017 tới 31/12/2020. Dữ liệu dùng để kiểm tra này mô hình chưa sử dụng nên ta dùng chuỗi này để đánh giá các hàm kernel trong mô hình của các chuỗi dữ liệu khác nhau. Kết quả được thể hiện trong bảng 2.

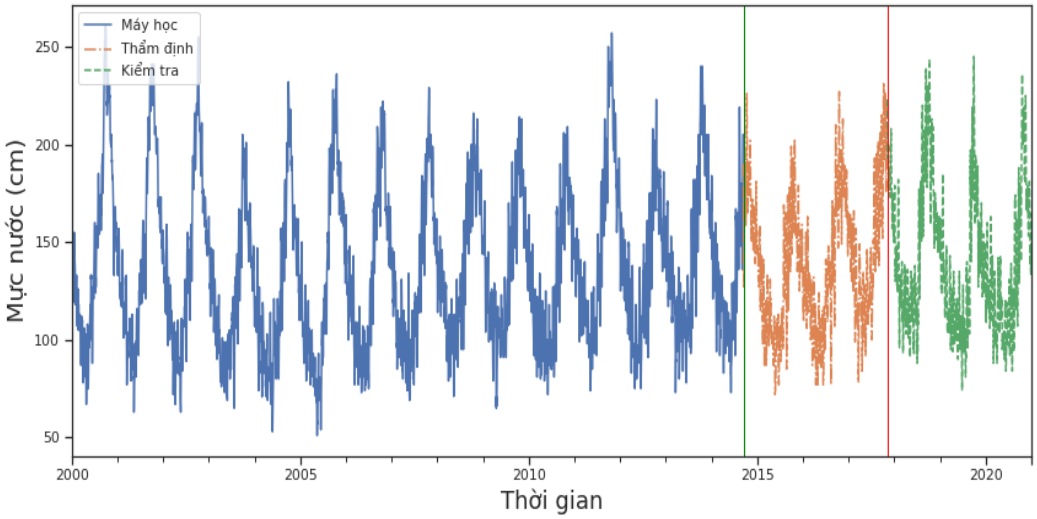
Bảng 2. Sai số của các hàm kernel trong mô hình SVR.

	HmaxCL			**HminCL**			**HtbCL**		
	RBF	**Linear**	**Sigmoid**	**RBF**	**Linear**	**Sigmoid**	**RBF**	**Linear**	**Sigmoid**
NSE	0,959	0,957	0,958	0,950	0,952	0,953	0,958	0,959	0,959
RMSE	7,37	7,71	7,45	12,43	12,16	12,03	8,17	8,16	8,15

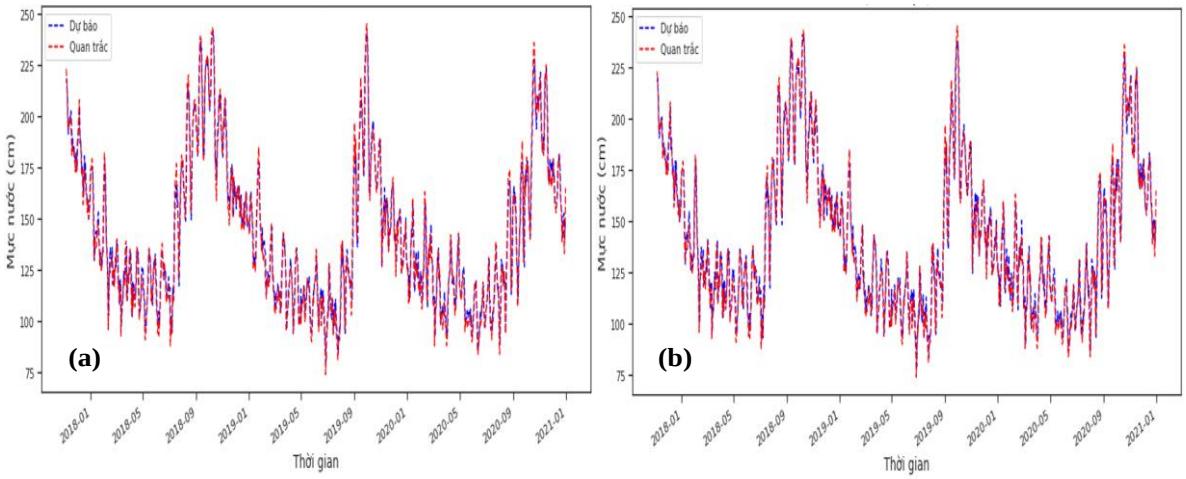
3.1. Kết quả hiệu chỉnh và kiểm định mô hình SVR với chuỗi dữ liệu mực nước cao nhất ngày trạm thủy văn Cao Lãnh

Kết quả bảng 2 cho thấy chỉ số NSE của các hàm kernel trong mô hình đều lớn hơn 0,95, kết quả này là rất tốt, và thấy được sự tương đồng cao giữa giá trị mực nước lớn nhất ngày thực đo và tính toán. Từ kết quả tính toán chỉ số NSE và RMSE cho ta thấy được hàm kernel RBF cho sai số thấp nhất, tối ưu nhất và cho tương quan cao nhất giữa các hàm kernel trong

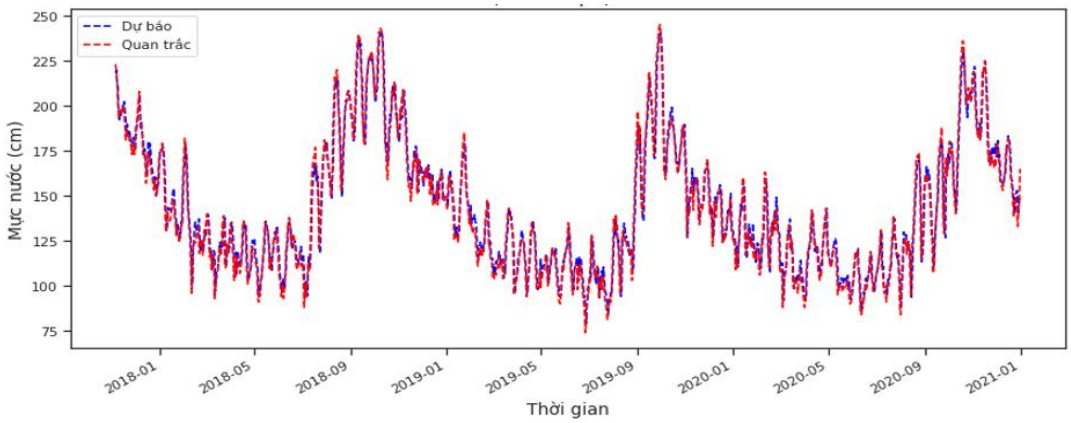
mô hình cho chuỗi dữ liệu mực nước lớn nhất ngày. Một số hình ảnh so sánh giữa chuỗi thực đo và tính toán giữa các hàm kernel trong mô hình được trình bày trong các Hình 2–4.



Hình 2. Chuỗi dữ liệu mực nước cao nhất ngày tại trạm Cao Lãnh từ năm 2000–2020.



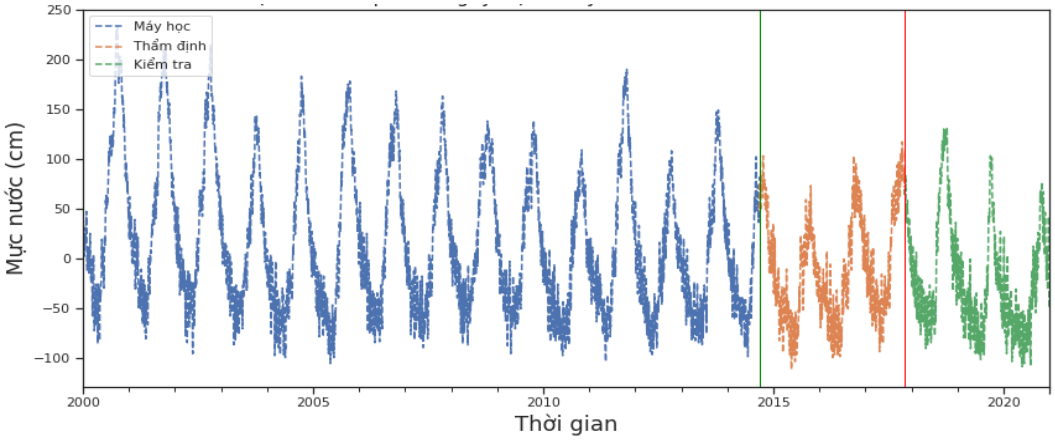
Hình 3. Kết quả dự báo của hàm kernel RBF trong mô hình với số liệu mực nước cao nhất ngày thực đo trong giai đoạn kiểm tra từ 06/11/2017 tới 31/12/2020 sử dụng (a) hàm kernel RBF và (b) hàm kernel tuyến tính.



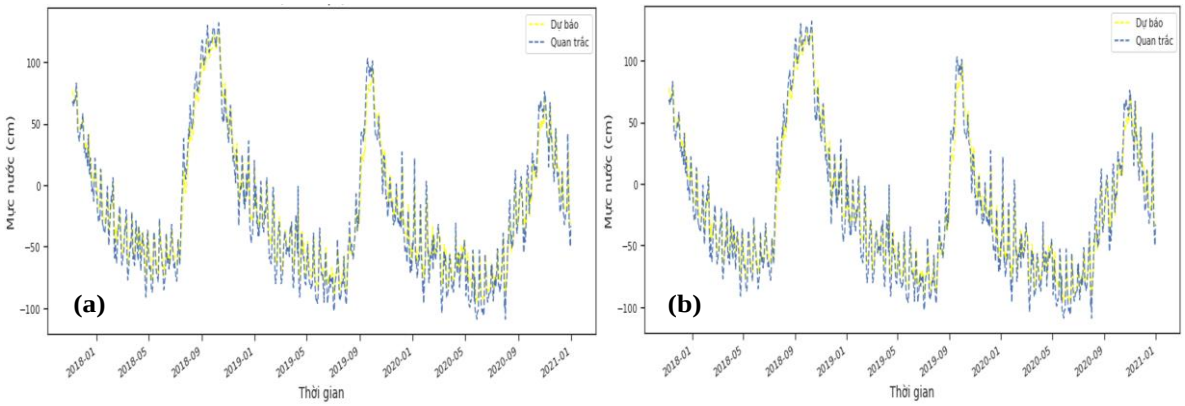
Hình 4. Kết quả dự báo của hàm kernel Sigmoid trong mô hình với số liệu mực nước cao nhất ngày thực đo trong giai đoạn kiểm tra từ 06/11/2017 tới 31/12/2020.

3.2. Kết quả hiệu chỉnh và kiểm định mô hình SVR với chuỗi dữ liệu mực nước thấp nhất ngày trạm thủy văn Cao Lãnh

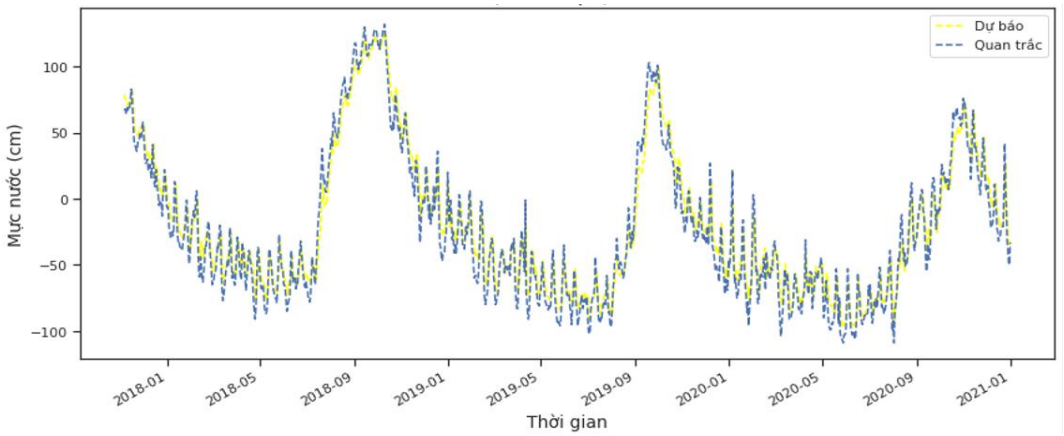
Kết quả về chuỗi dữ liệu mực nước thấp nhấy ngày cho thấy chỉ số NSE của các hàm kernel lớn hơn 0,95; thấp nhất là 0,95 đối với hàm rbf và cao nhất là 0,953 ở hàm sigmoid (Bảng 2). Cho thấy sự tương đồng rất cao giữa giá trị HminCL thực đo và tính toán. Ở chuỗi dữ liệu này thì cho ta thấy hàm sigmoid cho sai số thấp hơn hàm rbf và hàm tuyến tính. Kết quả so sánh giữa mực nước dự báo và thực đo quan trắc trong giai đoạn kiểm tra từ 06/11/2017 tới 31/12/2020 được trình bày trong Hình 5–7.



Hình 5. Chuỗi dữ liệu mực nước thấp nhất ngày tại trạm Cao Lãnh từ năm 2000–2020.



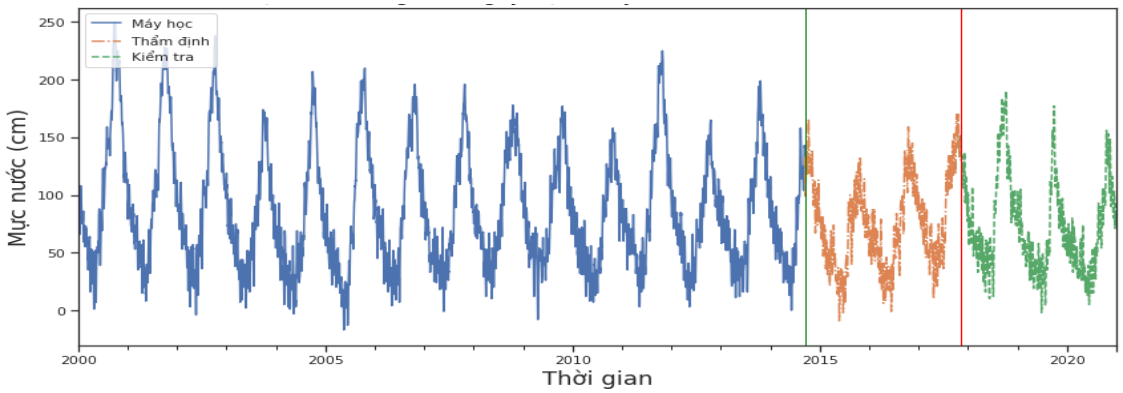
Hình 6. Kết quả dự báo của hàm kernel RBF trong mô hình với số liệu mực nước thấp nhất ngày thực đo trong giai đoạn kiểm tra từ 06/11/2017 tới 31/12/2020 sử dụng (a) hàm kernel RBF và (b) hàm kernel tuyến tính.



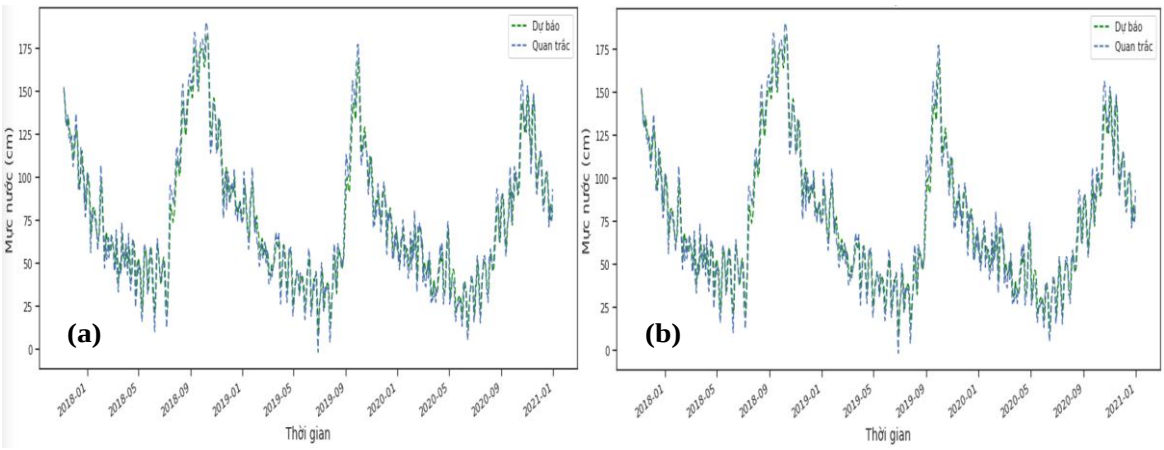
Hình 7. Kết quả dự báo của hàm kernel Sigmoid trong mô hình với số liệu mực nước thấp nhất ngày thực đo trong giai đoạn kiểm tra từ 06/11/2017 tới 31/12/2020.

3.3. Kết quả hiệu chỉnh và kiểm định mô hình SVR với chuỗi dữ liệu mực nước trung bình ngày trạm thủy văn Cao Lãnh

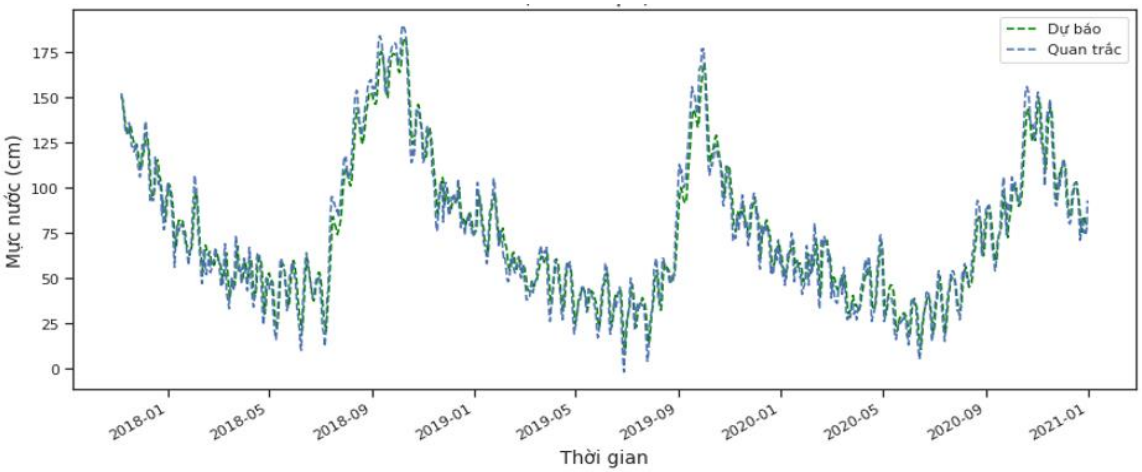
Kết quả ở Bảng 2 cho thấy chỉ số NSE có sự tương đồng rất cao giữa giá trị thực đo và tính toán, ở chuỗi dữ liệu Htb này thì chỉ số NSE của hàm rbf cho kết quả 0,958 thấp hơn hàm tuyến tính và sigmoid đều có chỉ số NSE = 0,959. Như vậy ở chuỗi dữ liệu mực nước trung bình ngày thì hàm sigmoid cho sai số thấp hơn và có sự tương đồng lớn hơn hàm rbf và tuyến tính. Kết quả so sánh giữa mực nước dự báo và quan trắc trong giai đoạn kiểm tra 06/11/2017 tới 31/12/2020 được trình bày trong Hình 8–10.



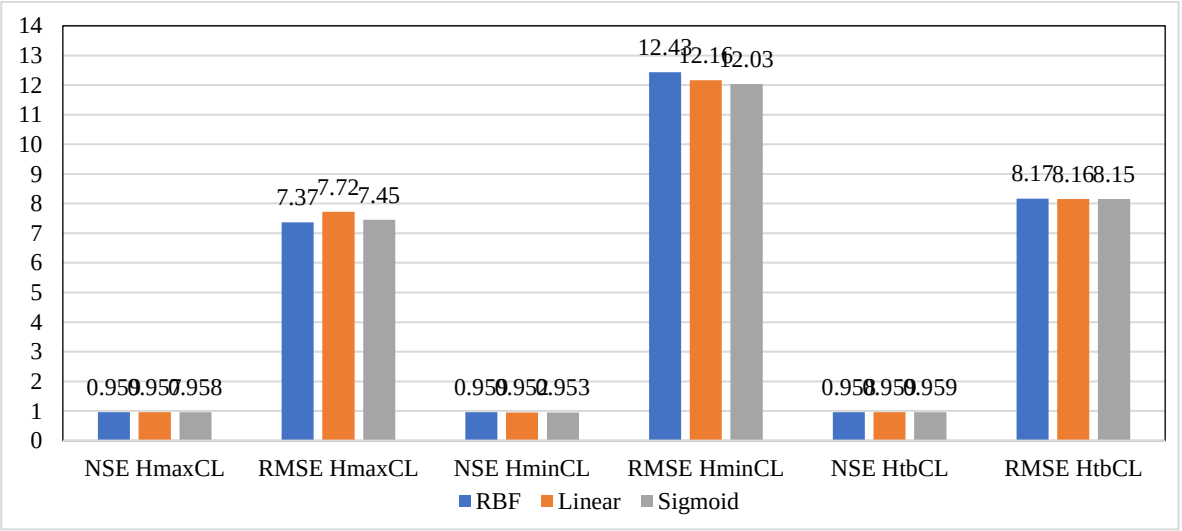
Hình 8. Chuỗi dữ liệu mực nước trung bình ngày tại trạm Cao Lãnh từ năm 2000–2020.



Hình 9. Kết quả dự báo với số liệu mực nước trung bình ngày thực đo trong giai đoạn kiểm tra từ 06/11/2017 tới 31/12/2020 sử dụng: (a) hàm kernel RBF và (b) hàm tuyến tính.



Hình 10. Kết quả dự báo của hàm kernel Sigmoid trong mô hình với số liệu mực nước trung bình ngày thực đo trong giai đoạn kiểm tra từ 06/11/2017 tới 31/12/2020.



Hình 11. Sai số các hàm kernel trong mô hình SVR giữa các chuỗi dữ liệu đầu vào.

Chỉ số sai số NSE cho ta thấy dữ liệu đầu vào bất kể nào cũng như ba hàm kernel RBF, Linear, Sigmoid đều cho sai số NSE đều lớn hơn 0,95, khẳng định sự tương đồng cao giữa dữ liệu thực đo với tính toán của mô hình SVR.

Với chuỗi dữ liệu đầu vào là mực nước cao nhất ngày: sai số RMSE khá tốt ở cả ba hàm kernel trong mô hình SVR và có độ chênh lệch không đáng kể. Nhưng với hàm kernel RBF thì cho sai số thấp nhất với RMSE = 7,37.

Còn đối với chuỗi dữ liệu đầu vào là mực nước trung bình và thấp nhất ngày thì hàm kernel Sigmoid lại cho sai số tốt hơn hàm kernel Rbf và Linear.

4. Kết luận

Nghiên cứu đã bước đầu thử nghiệm thành công giữa các hàm kernel Rbf, Linear, Sigmoid trong mô hình SVR dự báo mực nước trạm thủy văn Cao Lãnh tỉnh Đồng Tháp. Ba trường hợp tính toán là dự báo mực nước cao nhất, trung bình và thấp nhất ngày với các hàm kernel khác nhau. Kết quả cho thấy, với dữ liệu đầu vào là HmaxCL thì hàm kernel Rbf cho kết quả có độ chính xác khá cao, với dữ liệu đầu vào là HtbCL và HminCL thì hàm kernel Rbf lại cho sai số không tốt bằng hàm kernel Sigmoid. Như vậy việc lựa chọn dữ liệu đầu vào và hàm kernel trong mô hình là rất quan trọng quyết định hiệu quả của việc dự báo của mô hình SVR. Dựa trên những phân tích và kết quả tính toán, chúng tôi đề xuất sử dụng dữ liệu mực nước ngày lớn nhất để làm dữ liệu đầu vào cho mô hình SVR với hàm kernel Rbf. Ngoài ra còn một số hạn chế của nghiên cứu này là các tác giả chưa đưa hết được các hàm kernel vào sử dụng để so sánh và đánh giá các hàm kernel này trong mô hình SVR. Việc chạy các hàm kernel còn rất mất thời gian do năng lực tính toán của hệ thống máy tính của các tác giả có bộ vi xử lý chưa cao.

Đóng góp của tác giả: Xây dựng ý tưởng nghiên cứu: N.T.G., L.X.H.; xử lý số liệu L.X.H.; thiết lập các mô hình: L.X.H.; N.T.G.; Viết bản thảo bài báo: L.X.H.; Chỉnh sửa bài báo: N.T.G.

Lời cảm ơn: Nghiên cứu này có sự hỗ trợ về mặt dữ liệu và phương pháp luận từ đề tài mã số NĐT.58.RU/19 do Bộ Khoa học và Công nghệ tài trợ. Bài báo được sự góp ý, chỉnh sửa bởi TS. Lê Vũ Việt Phong.

Lời cam đoan: Các tác giả cam đoan bài báo này là công trình nghiên cứu của các tác giả, chưa được công bố ở đâu, không được sao chép từ những nghiên cứu trước đây.

Tài liệu tham khảo

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

2. VanderKwaak, J.E.; Loague, K. Hydrologic–Response simulations for the R–5 catchment with a comprehensive physics–based model. Water Resour. Res. **2001**, *37(4)*, 999–1013.
3. Nayak, P.C.; Sudheer, K.P.; Rangan, D.M.; Ramasastri, K.S. Short–term flood forecasting with a neurofuzzy model. *Water Resour. Res.* **2005**, *41(4)*, W04004.
4. Mosavi, A.; Ozturk, P. Flood Prediction Using Machine Learning, Literature Review. *Water* **2018**, 1–40.
5. Jain, S.K.; Das, A.; Srivastava, D.K. Application of ANN for Reservoir Inflow Prediction and Operation. *J. Water Resour. Plan. Manag.* **1999**, *125(5)*, 263–271.
6. Maier, H.R.; Dandy, G.C. The Use of Artificial Neural Networks for the Prediction of Water Quality Parameters. *Water Resour. Res.* **1996**, *32(4)*, 1013–1022.
7. Mosavi, A.; Rabczuk, T.; Varkonyi–Koczy, A.R. Reviewing the Novel Machine Learning Tools for Materials Design. International Conference on Global Research and Education: Recent Advances in Technology Research and Education, **2018**, 50–58.
8. Asefa, T.; Kemblowski, M.; McKee, M.; Khalil, A. Multi–time scale stream flow predictions: The support vector machines approach. *J. Hydrol.* **2006**, *318(1–4)*, 7–16.
9. Londhe, S.; Gavraskar, S. Stream Flow Forecasting using Least Square Support Vector Regression, Soft Comput. *Civ. Eng.* **2018**, *2(2)*, 56–88.
10. Adnan, R.M.; Yuan, X.; Kisi, O.; Adnan, M.; Mehmood, A. 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.* **2018**, *32(14)*, 4469–4486.
11. Maity, R.; Bhagwat, R.; Bhatnagar, A. Potential of support vector regression for prediction of monthly streamflow using endogenous property. *Hydrol. Process.* **2010**, *24(7)*, 917–923.
12. Rafidah, A.; Suhaila, Y. Modeling River Stream Flow Using Support Vector Machine. *Appl. Mech. Mater.* **2013**, *315*, 602–605.
13. Lin, J.; Cheng, C.; Chau, K. Using support vector machines for long–term discharge prediction Using support vector machines for long–term discharge prediction. *Hydrol. Sci. J.* **2006**, *51(4)*, 599–612.
14. Guo, J.; Zhou, J.; Qin, H.; Zou, Q.; Li, Q. Monthly streamflow forecasting based on improved support vector machine model. *Expert Syst. Appl.* **2011**, *38(10)*, 13073–13081.
15. Vapnik, V.N. The Nature of Statistical Learning Theory. Springer, New York, 1995.
16. Vapnik, V.N. An overview of statistical learning theory. *IEEE Trans. Neural Networks* **1999**, *10(5)*, 988–999.
17. Londhe, S.N.; Gavraskar, S. Stream Flow Forecasting Using Least Square Support Vector Regression. *J. Soft Comput. Civ. Eng.* **2018**, *2-2*, 56–88.
18. Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* **2003**, *3(3)*, 1157–1182.
19. Nash, J.E.; Sutcliffe, J.V. River Flow Forecasting Through Conceptual Models Part Ia Discussion of Principles. *J. Hydrol.* **1970**, *10*, 282–290.
20. https://en.wikipedia.org/wiki/Root–mean–square_deviation.
21. Lam, Đ.H.; Phương, N.H.; Đạt, N.Đ.; Giang, N.T. Xây dựng mô hình MIKE 11 phục vụ công tác dự báo thủy văn và xâm nhập mặn tỉnh Bến Tre. *Tạp chí Khí tượng Thủy văn* **2022**, *740(1)*, 38–49.
22. Hải, C.H.; Phương, T.A.; Như, T.Q.; Cường, T.M. Áp dụng mô hình trí tuệ nhân tạo vào dự báo lưu lượng đến hồ lưu vực sông Ba. *Tạp chí Khí tượng Thủy văn* **2019**, *705*, 22–33.

Building support vector regression models for water level forecasting at Cao Lanh station, Dong Thap province

Le Xuan Hoa[1], Nguyen Tien Giang[2*]

[1] Dong Thap Province Hydrometeorological Station; lexuanhoakttv@gmail.com
[2] Faculty of Hydrology, Meteorology & Oceanography, VNU University of Science, VNU–HN; giangnt@vnu.edu.vn

Abstract: In this study, 3 kernel functions Rbf, Linear (Linear) and Sigmoid are used in the Support Vector Regression (SVR) model and 3 input data series are: daily highest water level (HmaxCL); lowest water level of the day (HminCL); average daily water level (HtbCL) in the past to forecast the future water level at Cao Lanh station, Dong Thap province. The results show that all kernel functions in the SVR models give forecast results with high accuracy as shown by the NSE index > 0.95 for all different input data. Among the 3 input data series and tested kernel functions, the predicted HmaxCL data series has smallest error. This result is a good reference for building a machine learning model for forecasting future water levels for Cao Lanh Hydrological Station, Dong Thap province.

Keywords: SVR; RBF; Linear; Sigmoid; Machine Learning; Cao Lanh.