TẠP CHÍ
KHÍ TƯỢNG THỦY VĂN

Bài báo khoa học

Xây dựng mô hình mạng nơ-ron hồi quy dự báo độ cao sóng có nghĩa tại trạm Cồn Cỏ, Quảng Trị, Việt Nam

Trần Hồng Thái[1], Mai Văn Khiêm[2], Nguyễn Bá Thủy[2*], Bùi Mạnh Hà[2], Phạm Khánh Ngọc[2]

[1] Tổng cục Khí tượng Thủy văn; tranthai.vkttv@gmail.com
[2] Trung tâm Dự báo khí tượng thủy văn quốc gia; maikhiem77@gmail.com; thuybanguyen@gmail.com; manhhamhc@gmail.com; ngocpkchibo@gmail.com

*Tác giả liên hệ: thuybanguyen@gmail.com; Tel.: +84-975853471
Ban Biên tập nhận bài: 8/2/2022; Ngày phản biện xong: 1/4/2022; Ngày đăng bài: 25/4/2022

Tóm tắt: Những năm gần đây, trí tuệ nhân tạo (*AI - Artificial Intelligence*) đã được ứng dụng trong mọi lĩnh vực của đời sống, xã hội trong đó có lĩnh vực dự báo khí tượng thủy văn biển. Nghiên cứu này trình bày các kết quả trong việc sử dụng mạng bộ nhớ ngắn dài (*LSTM – Long Short Term Memory*) một phiên bản cải tiến của mạng nơ-ron hồi quy (*RNN – Recurrent Neural Network*) để xây dựng mô hình dự báo sóng tại trạm hải văn Cồn Cỏ theo các hạn dự báo 06, 12, 18 và 24 giờ. Số liệu quan trắc tại trạm được phân tích, tính toán các đặc tính thống kê và mối tương quan giữa các yếu tố để lựa chọn yếu tố đầu vào cho mô hình. Qua phân tích thống kê, hai mô hình đã được xây dựng, đó là mô hình đơn biến (chỉ sử dụng yếu tố độ cao sóng) và mô hình 02 biến (sử dụng độ cao sóng và vận tốc gió). Cả mô hình được xây dựng được sử dụng để dự báo độ cao sóng có nghĩa theo các hạn dự báo 06, 12, 18 và 24 giờ. Kết quả cho thấy, mặc dù mô hình hai biến cho độ tin cậy cao hơn, tuy nhiên cả hai mô hình chỉ đáp ứng với thời hạn dự báo 06 giờ, với độ tin cậy dự báo của mô hình đạt được lớn nhất với hệ số tương quan $R^2 = 0,582$, do bởi chất lượng số liệu quan trắc còn hạn chế về tần suất quan trắc và độ tin cậy.

Từ khóa: Dự báo sóng biển; Machine Learning; LSTM; AI, RNN.

1. Mở đầu

Các thông tin về dự báo sóng có ý nghĩa đặc biệt quan trọng công tác phòng tránh thiên tại mà còn với các hoạt động kinh tế - xã hội ven bờ và trên biển, nó ảnh hưởng trực tiếp đến việc lên kế hoạch xây dựng lịch trình di chuyển của tàu thuyền trên biển, xây dựng các cảng/để biển, hoạt động đánh bắt hải sản và tìm kiếm cứu nạn…. Các thông tin về dự báo sóng càng có ý nghĩa hơn trong điều kiện thời tiết nguy hiểm như bão, áp thấp nhiệt đới và gió mùa mạnh. Hiện nay dự báo sóng được thực hiện chủ yếu bởi các mô hình số trị thế hệ thứ 3 như WAM [1], WAVEWATCH III [2] và SWAN [3]. Cả 3 mô hình này đều dựa trên việc giải phương trình cân bằng tác động sóng [4]. Mặc các mô hình đều cho kết quả dự báo tương đối tốt, tuy nhiên vẫn còn hạn chế bởi các nhân tố sau: (1) Sự phát triển của sóng gió chủ yếu dựa trên các tham số thực nghiệm; (2) Đầu vào của mô hình chủ yếu là trường gió dự báo từ các mô hình khí tượng; (3) Độ phân không gian còn hạn chế do bởi thiếu năng lực tính toán. Ngoài ra, việc sử dụng các mô hình số trị còn có nhược điểm là yêu cầu dung lượng lưu trữ và năng lực tính toán lớn để có thể cho dự báo chi tiết. Chi phí tính toán cao này là nguyên nhân giới hạn độ phân giải về không gian và thời gian tính toán của các mô hình.

Thời gian gần đây, các phương pháp dự báo sóng theo hướng sử dụng phương pháp máy học đang được các nhà khoa học trên thế giới tích cực nghiên cứu và phát triển. Nhiều các

nghiên cứu đã chỉ ra rằng việc sử dụng phương pháp máy học không yêu cầu cơ sở hạ tầng tính toán có hiệu năng cao, chi phí tính toán rẻ, có thể thực hiện tức thời với chuỗi dữ liệu quá khứ có sẵn, thời gian tính toán nhanh hơn các mô hình vật lý rất nhiều với độ chính xác tương đương. Một phương pháp máy học điển hình được sử dụng trong dự báo sóng là phương pháp ứng dụng công nghệ mạng thần kinh nhân tạo (*ANN - Artificial Neural Network*), trong đó mạng nơ-ron hồi quy (*RNN – Recurrent Neural Network*) với phiên bản cải tiến của nó là mạng bộ nhớ ngắn dài (*LSTM – Long Short Term Memory*) [5] được sử dụng phổ biến trong dự báo chuỗi thời gian nói chung (timeseries data) và dự báo sóng nói riêng. Một ứng dụng gần đây của LSTM được thực hiện bởi [6] để dự báo sóng và so sánh với kết quả dự báo từ mô hình số trị, tác giả nhận thấy rằng mô hình LSTM tạo ra các dự báo chính xác hơn so với mô hình số trị. [7] cũng áp dụng mô hình LSTM để dự tính độ cao sóng có nghĩa tại một số trạm phao trên toàn cầu và mô hình LSTM của họ có thể cung cấp dự báo độ cao sóng có nghĩa chính xác hơn khi so sánh với một số mô hình học máy khác. [8] đã áp dụng RNN bao gồm một mô hình LSTM đơn giản và một mô hình mã hóa và giải mã LSTM (encoder – and – decoder LSTM) để dự đoán một biến (độ cao sóng có nghĩa) và hai biến (độ cao sóng có nghĩa và năng lượng rối). [9] đã trình bày một cách tiếp cận tích hợp việc sử dụng mạng LSTM và phương pháp phân tích thành phần chính (PCA) để dự đoán độ cao sóng có nghĩa. Phương pháp PCA của họ được sử dụng để trích xuất các thành phần chính từ một tập hợp các tín hiệu đầu vào trong khi LSTM được áp dụng để tránh sự độc lập lâu dài trong quá trình dự báo. [10] đã phát triển một LSTM phức hợp (convolutional LSTM) và một trình mã hóa tự động giảm nhiễu để dự báo thời tiết biển bao gồm cả thông số sóng gió.

Ở Việt Nam dự báo sóng chủ yếu dựa trên kết quả của mô hình số trị. Các kết quả dự báo của mô hình SWAN được thiết lập tại Trung tâm Dự báo khí tượng thủy văn quốc gia là nguồn tham khảo chính để đưa ra các bản tin dự báo sóng hằng ngày tại Việt Nam. Mô hình SWAN hiện tại được thiết lập chạy trên hệ thống máy tính hiệu năng cao (HPC) của Tổng cục khí tượng thủy văn với độ phân giải xấp xỉ 4km × 4km với thời gian dự báo là 10 ngày, bước thời gian dự báo là 03 giờ với thời gian tính toán khoảng 30–40 phút. Mặc dù đã được thiết lập trên HPC nhưng do tài nguyên tính toán vẫn còn hạn chế nên chưa thể chi tiết hơn cho các khu vực ven bờ, quanh đảo. Chưa có khả năng đáp ứng dự báo phục vụ chi tiết cho các khu vực kinh tế trọng điểm như: bãi tắm, tuyến hàng hải, đảo du lịch, giàn khoan … Vì vậy, việc nghiên cứu và xây dựng được một chương trình dự báo sóng sử dụng phương pháp máy học sẽ là một giải pháp tích cực trong việc khắc phục các vấn đề về tài nguyên tính toán cũng như thiếu hụt số liệu quan trắc sóng biển tại Việt Nam.

Trong nghiên cứu này, 02 mô hình sử dụng mạng bộ nhớ ngắn dài – LSTM sẽ được xây dựng để dự báo độ cao sóng có nghĩa tại trạm Cồn Cỏ là:

(1) Mô hình đơn biến: Tức là chỉ sử dụng chuỗi số liệu quan trắc độ cao sóng có nghĩa làm đầu vào để huấn luyện mô hình.

(2) Mô hình 02 biến: Sử dụng chuỗi số liệu quan trắc độ cao sóng có nghĩa và vận tốc gió tại độ cao 10m làm đầu vào để huấn luyện mô hình

Các mô hình sau khi được huấn luyện sẽ được thử nghiệm dự báo theo các hạn dự báo 06, 12, 18 và 24 giờ. Các giá trị dự báo của mô hình sau đó sẽ được so sánh lại với các giá trị thực đo để đánh giá khả năng ứng dụng của các mô hình trong thực tế.

2. Phương pháp nghiên cứu

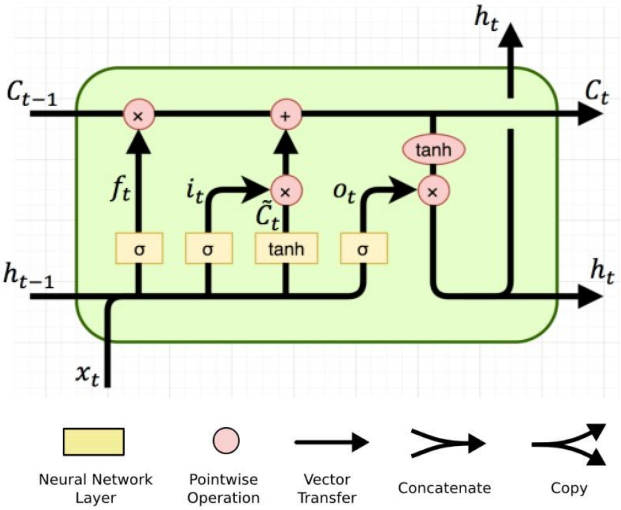
2.1 Mạng bộ nhớ ngắn dài – Long Short Term Memory (LSTM)

Mô hình mạng LSTM (*Long short term memory*) là một dạng mô hình Recurrent neural network (*RNN*) mà mạng network của nó được tổng hợp từ nhiều các đơn vị *Long short term memory*. *LSTM* đã được giới thiệu lần đầu tiên vào năm 1997 bởi Sepp Hochreiter (lĩnh vực y sinh và học máy) và Jurgen Schmidhuber (lĩnh vực trí tuệ nhân tạo) [5], sau đó công trình đã

được phát triển bởi [11] vào năm 2000 bằng việc đưa thêm forget gate vào cấu trúc ban đầu của mạng *LSTM*.

LSTM đã đạt được hiệu quả cao trong các mô hình xử lý ngôn ngữ tự nhiên, nhận diện chữ viết tay và dành chiến thắng trong cuộc thi ICDAR được tổ chức vào năm 2009. *LSTM* cũng là một trong những thành phần chính của mạng đạt được 17,7% phoneme error rate (một chỉ số được dùng để đo mức độ sai khác giữa các âm) trên bộ dữ liệu âm thanh TIMIT. Các hãng công nghệ lớn như *Google, Facebook, Apple, Microsoft* đều sử dụng *LSTM* như một nên tảng trong các ứng dụng nhận diện giọng nói của mình [12].

Khác với mạng RNN chuẩn chỉ có một tầng mạng nơ-ron, mạng LSTM có tới 4 tầng tương tác với nhau một cách rất đặc biệt. Mạng LSTM rất phù hợp với các bài toán phân loại và dự báo dựa trên dữ liệu dạng chuỗi thời gian bởi vì model có khả năng ghi nhớ tức thời các sự kiện xảy ra ở gần nó. LSTM được thiết kế để giải quyết sự bùng nổ và triệt tiêu gradient, hiện tượng mà khiến cho các mô hình truyền thống của RNNs có thể gặp phải.



Hình 1. Sơ đồ của một đơn vị mạng LSTM (*Long short term memory*).

Ở sơ đồ trên, mỗi một đường mang một véc-tơ từ đầu ra của một nút tới đầu vào của một nút khác. Các hình trong màu hồng biểu diễn các phép toán như phép cộng véc-tơ chẳng hạn, còn các ô màu vàng được sử dụng để học trong các từng mạng nơ-ron. Các đường hợp nhau kí hiệu việc kết hợp, còn các đường rẽ nhánh ám chỉ nội dung của nó được sao chép và chuyển tới các nơi khác nhau. Đầu vào là input trạng thái t x_t, trạng thái ẩn (hidden state) của trạng thái t - 1 là h_{t-1} và trạng thái ô (cell state) của trạng thái t - 1 là C_{t-1}và đầu ra là hidden state của trạng thái t là h_t và cell state của trạng thái t là C_t. Khởi đầu với $h_0 = 0$ và $C_0 = 0$, các hàm được định nghĩa như sau:

Với $\boldsymbol{W} \in \mathbb{R}^{m \times n}$, $\boldsymbol{U} \in \mathbb{R}^{m \times n}$, $\boldsymbol{b} \in \mathbb{R}^{m \times n}$: các ma trận hệ số và vectơ sai số được học trong quá trình training. Với n là số chiều của x và m là số chiều của các vectơ kích hoạt. σ và $tanh$ lần lượt là 2 hàm kích hoạt sigmoid và tanh [13-17].

Trong nghiên cứu này, thư viện phần mềm mã nguồn mở TensorFlow của Google, các thư viện Numpy, Pandas, Keras cùng với ngôn ngữ lập trình Python 3.6 đã được sử dụng để thiết lập mô hình LSTM.

2.2. *Phương pháp đánh giá*

Để đánh giá mức độ phù hợp giữa các giá trị dự báo độ cao sóng từ mô hình với các quan trắc thực tế, chúng tôi sử dụng một loạt các chỉ số sau:

- Hệ số tương quan R^2 là thước đo độ chặt chẽ của mối quan hệ tuyến tính giữa bộ giá trị thực đo và mô phỏng hay cho biết mô hình đang nghiên cứu phù hợp với dữ liệu ở mức bao

nhiêu %. Mục đích của mô phỏng khi hệ số tương quan được sử dụng là để hàm mục tiêu cực đại hoá tới 1. Tuy nhiên, khả năng đạt giá trị tuyệt đối khó có thể đạt được nên giá trị R^2 thường được chấp nhận khi đạt trên 0,5 [17].

$$R^2 = \frac{(\sum_{i=1}^{n} O_i - \bar{O})(P_i - \bar{P})^2}{\left(\sqrt{\sum_{i=1}^{n}(O_i - \bar{O})^2}\right)\left(\sqrt{\sum_{i=1}^{n}(P_i - \bar{P})^2}\right)^2} \quad (1)$$

- Sai số bình phương trung bình (*Root Mean Square Error - RMSE*): là căn bậc hai của MSE và là thước đo của biên độ sai số.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2} \quad (2)$$

Chỉ số RMSE cho biết biên độ trung bình của sai số dự báo, nhưng không cho biết hướng của độ lệch.

- NSE (*Nash Sutcliffe Efficiency – hệ số Nash*): Hệ số hiệu quả: được sử dụng để đo mức độ liên kết giữa các giá trị thực đo và mô phỏng.

$$NSE = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2} \quad (3)$$

Trong đó P_i là giá trị dự báo, O_i là giá trị quan trắc, $\bar{P}$ là giá trị trung bình của chuỗi giá trị dự báo và $\bar{O}$ là giá trị trung bình của chuỗi giá trị quan trắc.

2.3. Dữ liệu đầu vào của mô hình

Tập dữ liệu trong nghiên cứu này là số liệu quan trắc sóng và gió tại trạm hải văn Cồn Cỏ có tọa độ 17°10'N - 107 ° 21' E bao gồm các yếu tố vận tốc gió (m/s), và hướng gió ở độ cao 10m trên bề mặt biển, độ cao sóng có nghĩa (m) và hướng sóng. Các yếu tố được quan trắc 4obs/ngày vào các thời điểm 1 giờ, 7 giờ, 13 giờ và 19 giờ (giờ Việt Nam). Riêng dữ liệu sóng (độ cao và hướng) không có số liệu quan trắc lúc 1 giờ, do đó 1 ngày chỉ có 3obs. Thêm vào đó, quan trắc sóng chủ yếu được ước lượng bằng mắt nên tính chính xác không cao nhưng đây là nguồn số liệu quan trắc đáng tin cậy nhất mà nhóm nghiên cứu thu thập được. Dữ liệu quan trắc tại trạm Cồn Cỏ được thu thập từ 1 giờ ngày 01 tháng 7 năm 2016 đến 19 giờ ngày 30 tháng 6 năm 2021.

Mô hình LSTM được xây dựng dựa trên chuỗi thời gian tuần tự và liên tục, chính vì vậy trước khi tiến hành xây dựng mô hình số liệu cần phải được xử lý các giá trị khuyết thiếu (missing data), trong đó yếu tố chủ yếu cần phải xử lý là yếu tố sóng. Trong nghiên cứu này phương pháp nội suy tuyến tính sẽ được áp dụng để lấp đầy các giá trị khuyết thiếu trong tập dữ liệu. Qua bảng mô tả thống kê các yếu tố của tập dữ liệu trong Bảng 1 thấy rằng số quan trắc độ cao sóng ban đầu thống kê được chỉ là 5478 sau khi nội suy là 7304 giá trị nhưng độ lệch so với giá trị trung bình chỉ thay đổi 0,01 đơn vị (0,52 xuống 0,51 đơn vị) các giá trị khác như giá trị trung bình, giá trị nhỏ nhất, giá trị lớn nhất và trung vị là không thay đổi. Nội suy số liệu cũng không làm thay đổi xu thế cũng như tính chu kỳ của chuỗi số liệu như được thể hiện trên hình 2.

Bảng 1. Mô tả thống kê các yếu tố quan trắc trong tập dữ liệu trước và sau khi xử lý giá trị khuyết thiếu.

Mô tả thống kê	Các yếu tố ban đầu				Các yếu tố sau khi nội suy			
	Độ cao sóng	**Hướng sóng**	**Vận tốc gió**	**Hướng gió**	**Độ cao sóng**	**Hướng sóng**	**Vận tốc gió**	**Hướng gió**
Số quan trắc	5478	5171	7304	7109	7304	7304	7304	7304
Trung bình	0,85	139,81	3,71	152.83	0,85	139.85	3,71	152,73
Giá trị nhỏ nhất	0	0	0	0	0	0	0	0
Độ lệch	0,52	100,49	2.14	101,82	0,51	94,80	2,14	101,05
25%	0,5	45	2	90	0,5	45	2	90

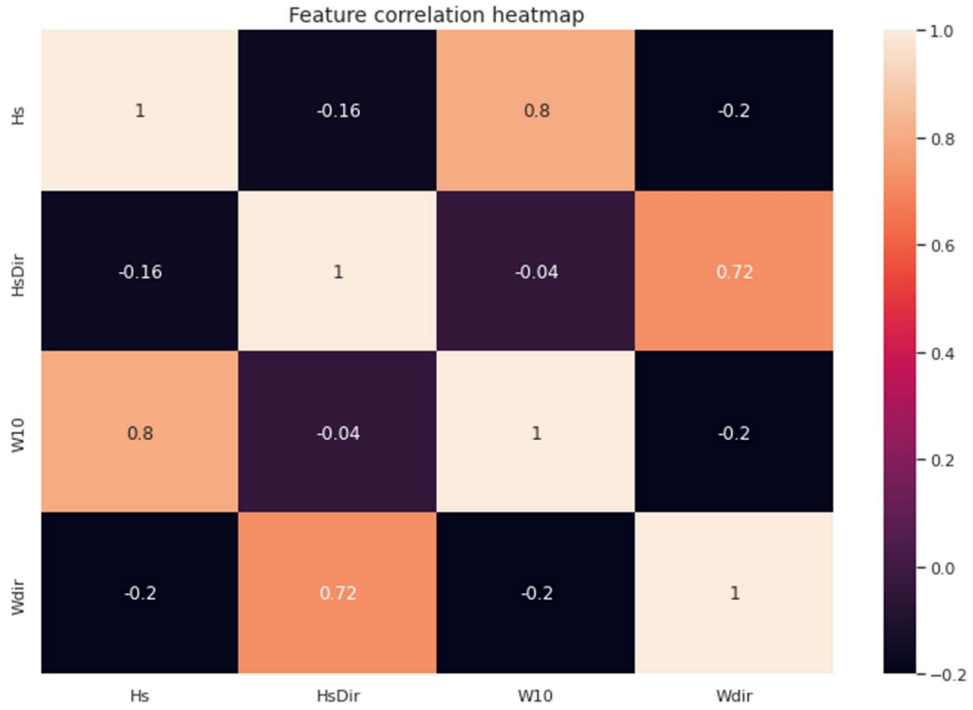
Mô tả thống kê	Các yếu tố ban đầu				Các yếu tố sau khi nội suy			
	Độ cao sóng	Hướng sóng	Vận tốc gió	Hướng gió	Độ cao sóng	Hướng sóng	Vận tốc gió	Hướng gió
50%	0,75	135	3	135	0,75	135	3	135
75%	1	225	5	225	1	225	5	225
Giá trị lớn nhất	6	315	22	337,5	6	315	22	337,5

Hình 2. Biến thiên độ cao sóng và vận tốc gió tại trạm Cồn Cỏ (giai đoạn 7/2016 – 6/2021) trước (a) và sau (a) khi đã nội suy tuyến tính.

Như đã trình bày ở trên, dữ liệu quan trắc tại trạm Cồn Cỏ bao gồm các yếu tố vận tốc gió (m/s), hướng gió, độ cao sóng (m) và hướng sóng. Mục đích của nghiên cứu này là dự báo độ cao sóng sử dụng mô hình học máy, chính vì vậy biến mục tiêu của nghiên cứu sẽ là độ cao sóng có nghĩa. Để lựa chọn các tham số đầu vào cho một mô hình học máy, trước hết cần phải phân tích tương quan thống kê giữa các yếu tố với nhau (ví dụ: Fan và cộng sự, 2020; Kim và cộng sự, 2020b).

Trên hình 3 thể hiện hệ số tương quan Pearson (r) giữa các yếu tố (biến) quan trắc thu thập được tại trạm hải văn Cồn Cỏ. Có thể thấy rằng ngoài việc tương quan với chính nó thì độ cao sóng có nghĩa có tương quan rất lớn với giá trị vận tốc gió tại độ cao 10m với r = 0,8. Trong khi đó, độ cao sóng lại có tương quan nghịch rất thấp với các biến hướng gió và hướng sóng với r = -0,2 và r = -0,16 tương ứng. Như vậy các tham số được lựa chọn cho mô hình học máy chỉ có thể là độ cao sóng có nghĩa và vận tốc gió tại độ cao 10 m.

Tập dữ liệu tại trạm Cồn Cỏ sẽ được chia làm 2 phần, phần thứ nhất là chuỗi số liệu quan trắc từ 01 giờ ngày 01/7/2016 đến 19 giờ ngày 31/12/2020 (90% chuỗi số liệu) được dùng để huấn luyện (training) mô hình và phần thứ 2 là chuỗi số liệu quan trắc từ 01 giờ ngày 1/1/2021 đến 19 giờ ngày 30/6/2021 (10% chuỗi số liệu) được sử dụng để đánh giá (test) mô hình.



Hình 3. Ma trận hệ số tương quan Pearson giữa các biến.

Bảng 2. Lựa chọn các tham số tại trạm Cồn Cỏ cho mô hình học máy.

TT	Biến	Mô tả	Sử dụng
1	Hs	Độ cao sóng có nghĩa (m)	Có (Biến mục tiêu)
2	HsDir	Hướng sóng	Không
3	W10	Vận tốc gió tại độ cao 10m	Có
4	Wdir	Hướng gió tại độ cao 10m	Không

Trong nghiên cứu này mô hình học máy sử dụng mạng LSTM được xây dựng để dự báo độ cao sóng có nghĩa lần lượt cho từng trường hợp: 06 giờ, 12 giờ, 18 giờ và 24 giờ. Các mô hình dự báo sẽ được xây dựng theo các kịch bản như sau:

- Xây dựng mô hình dự báo độ cao sóng có nghĩa sử dụng tập dữ liệu đơn biến, tức là sử dụng chính giá trị độ cao sóng làm tham số đầu vào để huấn luyện và xây dựng mô hình dự báo theo (1) thời hạn dự báo 06 giờ; (2) thời hạn dự báo 12 giờ; (3) thời hạn dự báo 18 giờ và (4) thời hạn dự báo 24 giờ. Sau đây sẽ gọi là mô hình (kịch bản) lần lượt là **CC111**, **CC112**, **CC113** và **CC114** tương ứng.

- Xây dựng mô hình dự báo độ cao sóng có nghĩa sử dụng tập dữ liệu đa biến, tức là sử dụng các yếu tô độ cao sóng có nghĩa và vận tốc gió làm tham số đầu vào để huấn luyện và xây dựng mô hình dự báo theo (5) thời hạn dự báo 06 giờ; (6) thời hạn dự báo 12 giờ; (7) thời hạn dự báo 18 giờ và (8) thời hạn dự báo 24 giờ. Sau đây sẽ gọi là mô hình (kịch bản) **CC121**, **CC122**, **CC123** và **CC124** tương ứng.

Bảng 3. Tổng hợp các mô hình dự báo dựa theo các kịch bản.

TT		Mô hình	Tham số đầu vào (X)	Tham số dự báo (Y)	Hạn dự báo (giờ)
Mô hình một		**CC111**	Hs_{t-1}	Hs_t	06
		CC112	Hs_{t-1}	Hs_t	12

TT	Mô hình	Tham số đầu vào (X)	Tham số dự báo (Y)	Hạn dự báo (giờ)
Mô hình hai biến	**CC113**	Hs_{t-1}	Hs_t	18
	CC114	Hs_{t-1}	Hs_t	24
	CC121	Hs_{t-1}, $W10_{t-1}$	Hs_t	06
	CC122	Hs_{t-1}, $W10_{t-1}$	Hs_t	12
	CC123	Hs_{t-1}, $W10_{t-1}$	Hs_t	18
	CC124	Hs_{t-1}, $W10_{t-1}$	Hs_t	24

2.4. Tối ưu hóa các siêu tham số

Việc lựa chọn một bộ siêu tham số phù hợp ảnh hưởng đáng kể đến hiệu suất của mô hình. Trong nghiên cứu này, các siêu tham số liên quan đến mô hình sử dụng mạng LSTM được hiệu chỉnh sẽ là: mini-batch size, tỷ lệ Dropout, số đơn vị lớp ẩn, early stopping và số Epochs. Các siêu tham số sẽ được lựa chọn bằng phương pháp tìm kiếm ngẫu nhiên để tìm được bộ tham số thích hợp nhằm tối ưu hóa mô hình mạng LSTM. Cấu hình để tìm kiếm các siêu tham số sẽ được thiết kế theo các mô hình dựa theo các kịch bản đã được lựa chọn. Để khảo sát ảnh hưởng của các siêu tham số đã được lựa chọn đến hiệu suất của mô hình, hệ số tương quan R^2 và sai số quân phương RMSE sẽ được sử dụng để đánh giá. Các siêu tham số phù hợp nhất cho mô hình dự báo sẽ được lựa chọn dựa vào hệ số tương quan lớn nhất và sai số quân phương nhỏ nhất đạt được trong quá trình huấn luyện và xác thực mô hình mô hình. Phạm vi của các siêu tham số để tiến hành tìm kiếm ngẫu nhiên và các siêu tham số phù hợp được lựa chọn được thể hiện trên Bảng 3.

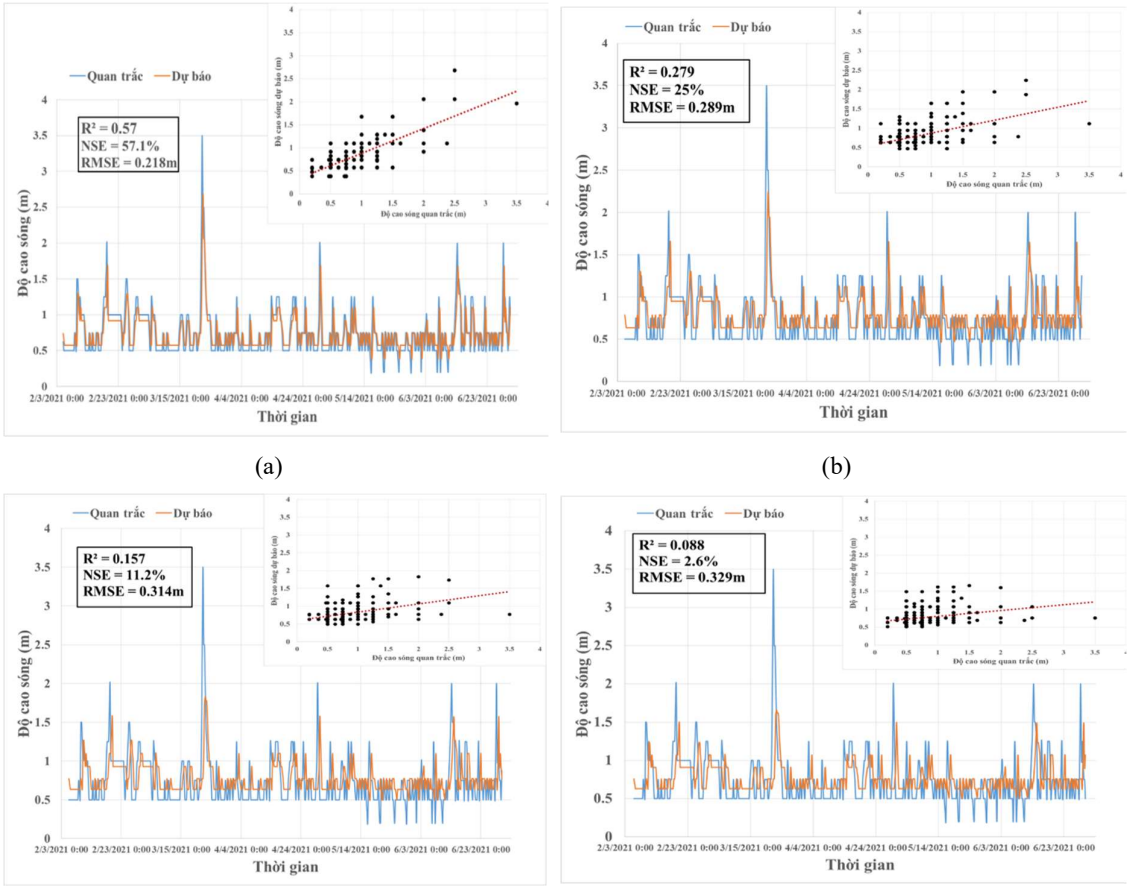
Bảng 4. Phạm vi và các siêu tham số phù hợp được lựa chọn cho các mô hình.

Siêu tham số	Phạm vi điều chỉnh	Mô hình	
		CC111-CC112-CC113-CC114	CC121-CC122-CC123-CC124
Số đơn vị ẩn	[100, 200]	100	100
Tỷ lệ Dropout	[0,25, 0,5]	0,25	0,5
Early stopping	[10, 50]	50	50
Epochs	[10, 100, 200, 500, 1000]	200	200
Batch_size	[32, 64, 128, 256, 512]	64	32

3. Đánh giá mô hình dự báo sóng sử dụng mạng LSTM

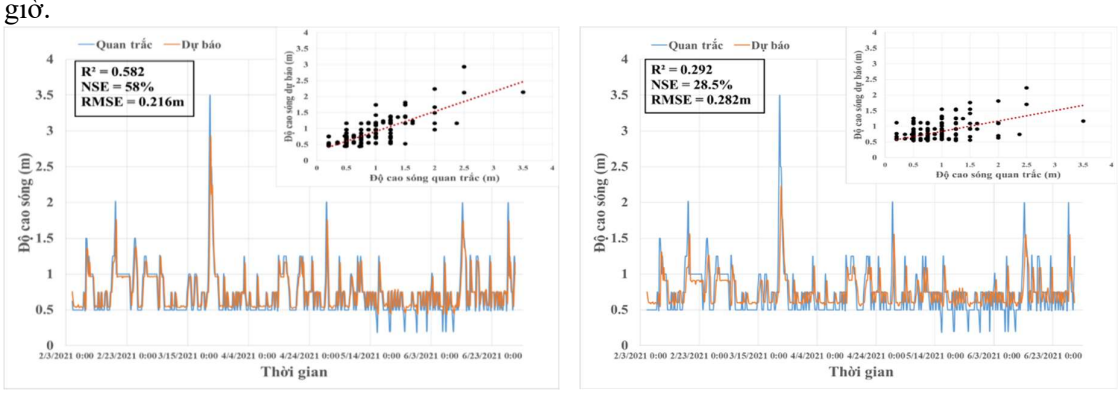
Các mô hình LSTM đã được huấn luyện sử dụng các siêu tham số đã được lựa chọn sẽ được đánh giá lại (kiểm định) theo các hạn dự báo 06, 12, 18 và 24 giờ, bằng cách so sánh kết quả dự báo của mỗi mô hình với giá trị độ cao sóng thực đo.

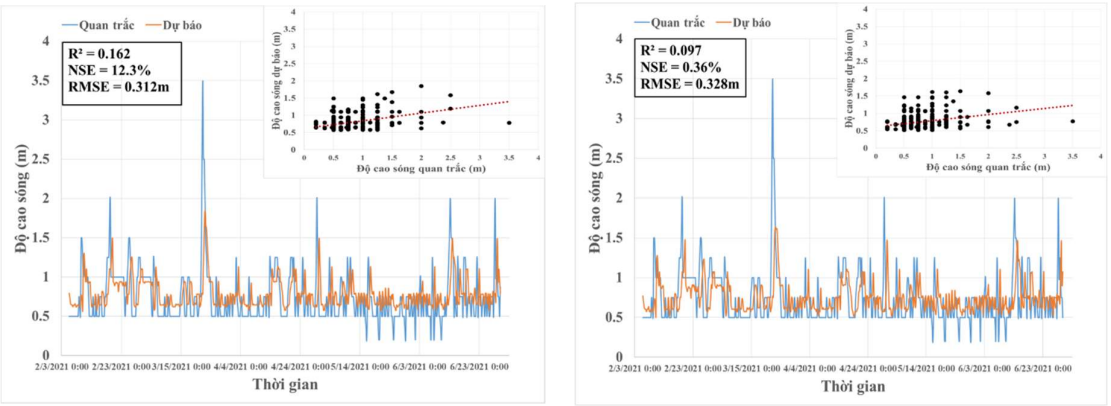
Trên các hình 3 (a, b, c và d) là kết quả so sánh giữa giá trị dự báo của mô hình sử dụng chuỗi số liệu đơn biến (chỉ sử dụng độ cao sóng) theo các hạn dự báo 06 (CC111), 12 (CC112), 18 (CC113) và 24 giờ (CC114) tương ứng với giá trị thực đo. Có thể thấy rằng với mô hình này hạn dự báo 06 giờ cho độ tin cậy của dự báo lớn nhất với hệ số tương quan bình phương R^2 và chỉ số NSE tương ứng là 0,57 và 57,1% và sai số quân phương nhỏ nhất RMSE = 0,218m. Hạn dự báo 24 giờ có độ tin cậy thấp nhất với R^2 = 0,088 và NSE = 2,6%, đồng thời sai số cũng tăng lên với RMSE = 0,329 m. Từ hạn dự báo 12 giờ, độ tin cậy của mô hình chỉ còn R^2 = 0,279 tức là đã giảm xuống dưới 0,5. Điều này có nghĩa rằng mô hình không đủ độ tin cậy với các hạn dự báo 12, 18 và 24 giờ, hay nói cách khác mô hình không thể sử dụng để dự báo với hạn dự báo lớn hơn 06 giờ.



Hình 4. So sánh giữa giá trị quan trắc với kết quả dự báo của mô hình được xây dựng từ chuỗi số liệu đơn biến (chỉ sử dụng độ cao sóng có nghĩa) theo các hạn dự báo (a) 06 giờ, (b) 12 giờ, (c) 18 giờ và (d) 24 giờ.

Với mô hình được xây dựng từ chuỗi số liệu 2 biến (sử dụng độ cao sóng và vận tốc gió), kết quả so sánh giá trị dự báo với thực đo theo các hạn dự báo 06, 12, 18 và 24 giờ (hình 4 - a, b, c, d) cũng cho thấy rằng hạn dự báo 06 giờ cho độ tin cậy lớn nhất và sai số nhỏ nhất với R^2 = 0,582, NSE = 58% và RMSE = 0,216m. Hạn dự báo 24 giờ cho độ tin cậy thấp nhất với R^2 = 0,097, NSE = 3,6% và RMSE = 0,328m. Từ hạn dự báo 12 giờ, độ tin cậy của mô hình dự báo cũng giảm chỉ còn R^2 = 0,292 tức là dưới 0,5. Như vậy mô hình được xây dựng từ chuỗi số liệu 02 biến cũng không thể sử dụng để dự báo độ cao sóng với các hạn dự báo lớn hơn 06 giờ.

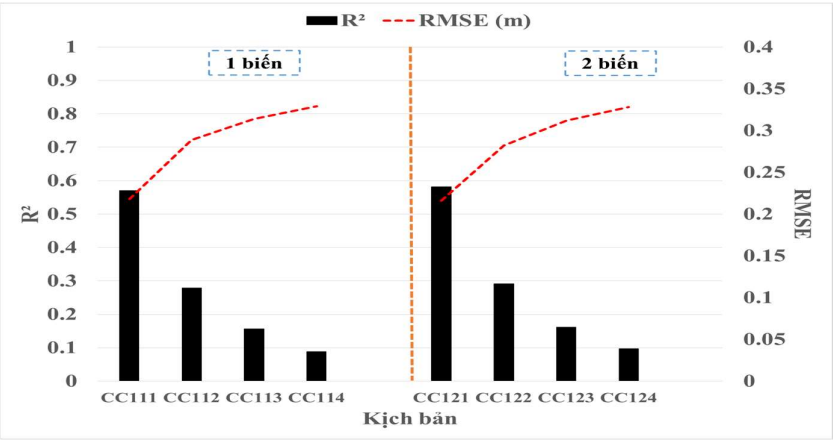




(c) (d)

Hình 5. So sánh giữa giá trị quan trắc với kết quả dự báo của mô hình được xây dựng từ chuỗi số liệu 02 biến (sử dụng độ cao sóng có nghĩa và vận tốc gió) theo các hạn dự báo (a) 06 giờ, (b) 12 giờ, (c) 18 giờ và (d) 24 giờ.

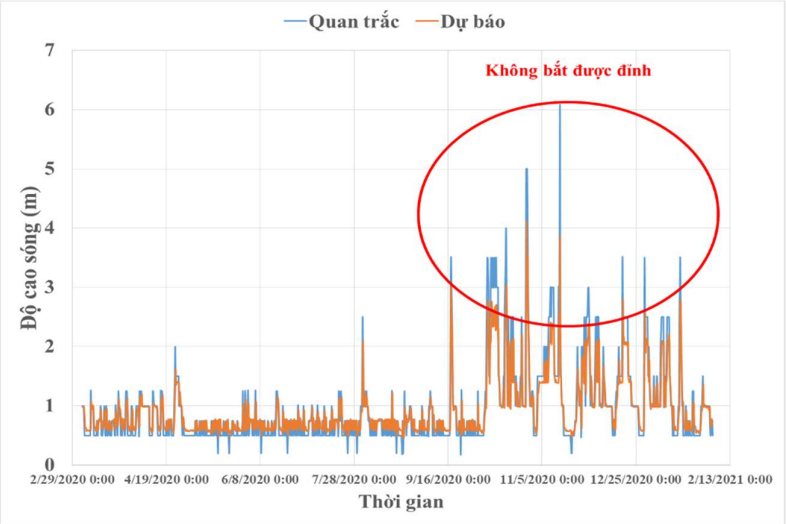
Dựa vào các kết quả so sánh có thể thấy rằng, các mô hình đều có chung xu thế là độ tin cậy của mô hình giảm và sai số dự báo tăng khi hạn dự báo càng xa. Với tập dữ liệu tại tạm Cồn Cỏ, mô hình sử dụng chuỗi số liệu 02 biến có độ tin cậy cao hơn mô hình sử dụng chuỗi số liệu đơn biến. Nhưng chênh lệch giữa độ tin cậy của mô hình đơn biến (R^2 = 0,57) và mô hình 02 biến (R^2 = 0,582) là không lớn, điều này cho thấy ảnh hưởng của vận tốc gió tới kết quả dự báo trong mô hình này là không nhiều. Nhưng cũng có thể thấy rằng, việc xem xét đầy đủ ảnh hưởng của các yếu tố đầu với tới yếu tố cần được dự báo sẽ làm tăng độ chính xác của mô hình dự báo.



Hình 6. Đánh giá dự báo của các mô hình được xây dựng với chuỗi số liệu đơn biến, 02 biến và 03 biến theo các hạn dự báo 06 giờ, 12 giờ, 18 giờ và 24 giờ.

Thêm vào đó, các mô hình được xây dựng dựa trên tập dữ liệu tại trạm Cồn Cỏ được đánh giá là đủ độ tin cậy để dự báo nhưng độ chính xác của mô hình dự báo vẫn còn khá thấp so với mong đợi (độ tin cậy của mô hình lớn nhất cũng chỉ đạt R^2 = 0,582). Nguyên nhân của việc này một phần là do chất lượng của số liệu quan trắc. Một nguyên nhân khác nữa là trong quá trình huấn luyện, mô hình đã không dự báo được các giá trị cực trị của chuỗi số liệu (Hình 6). Các giá trị cực trị này không thường xuyên xảy ra, nó thường xảy ra khi có hiện tượng thời tiết như bão hoặc gió mùa mạnh. Điều này đã không được phân tích đầy đủ trong quá trình xử lý số liệu ban đầu đã dẫn đến sai số của mô hình dự báo. Một cách khắc phục là trong quá trình phân tích, xử lý số liệu, các giá trị cực trị có thể được phân tách ra thành các sóng thành

phần từ chuỗi số liệu ban đầu (có thể sử dụng phương pháp biến đổi Wavelet). Sau đó các sóng thành phần được phân tách này cũng có thể được coi như là một biến để tham gia vào quá trình huấn luyện mô hình.



Hình 7. Không dự báo đúng các giá trị cực trị trong quá trình huấn luyện mô hình.

4. Kết luận

Trong nghiên cứu này, các mô hình sử dụng mạng LSTM đã được xây dựng để dự báo độ cao sóng có nghĩa tại trạm Cồn Cỏ, Quảng Trị, Việt Nam theo các hạn dự báo 06, 12, 18 và 24 giờ. Khả năng dự báo của các mô hình được xây dựng dựa vào các yếu tố đầu vào và bộ siêu tham số thích hợp đã được lựa chọn, đánh giá bằng cách so sánh các giá trị dự báo của mô hình với quan trắc. Từ các kết quả so sánh, đánh giá, các kết luận chính của nghiên cứu được đưa ra như sau:

- Độ chính xác của mô hình dự báo phụ thuộc rất lớn và chất lượng của chuỗi số liệu và các yếu tố được lựa chọn làm đầu vào. Mô hình sử dụng chuỗi số liệu 02 biến (độ cao sóng và vận tốc gió) cho độ tin cậy lớn hơn với mô hình 1 biến (chỉ sử dụng độ cao sóng), tuy nhiên độ chính xác được cải thiện không nhiều, có nghĩa là khi xem xét tới vận tốc gió độ tin cậy tăng không đáng kể với hệ số tương quan cho mô hình 1 biến và 2 biến tương ứng là $R^2 = 0,57$ và $R^2 = 0,582$.

- Hầu hết các kịch bản dự báo đều cho độ tin cậy của các mô hình dự báo cao nhất là ở thời hạn 06 giờ, sai số tăng dần khi hạn dự báo càng xa. Do vậy, để có thể dự báo tốt cho các thời hạn xa hơn cần sử dụng kết hợp giữa số liệu quan trắc và kết quả dự báo từ mô hình số trị như là các biến đầu vào cho mô hình học máy như các nghiên cứu trước đã đề cập.

- Các giá trị cực trị trong chuỗi số liệu cũng ảnh hưởng tới độ chính xác của mô hình dự báo. Nghiên cứu đã chỉ ra rằng trong quá trình huấn luyện mô hình, các điểm cực trị đã không được dự báo lại một cách chính xác hay nói cách khác mô hình đã không học được hoặc tìm ra được các trọng số tại các điểm cực trị này. Các điểm cực trị là các giá trị không thường xuyên xảy ra, chúng thường xuất hiện khi có các điều kiện thời tiết bất thường như bão, áp thấp nhiệt hay gió mùa mạnh. Do đó, trong bước phân tích và xử lý số liệu đầu vào cần phân tách riêng rẽ các cực trị này thành các sóng thành phần từ chuỗi số liệu ban đầu và coi các sóng thành phần này như là một biến đầu vào để tham gia vào quá trình huấn luyện mô hình.

- Qua đánh giá độ tin cậy của mô hình cho thấy nếu có chuỗi số liệu quan trắc đủ dài và tin cậy để đảm bảo xác định các đặc tính thống kê sâu hơn thì hoàn toàn có thể xây dựng được mô hình dự báo sóng ứng dụng phương pháp học máy áp dụng được trong thực tế.

Đóng góp của tác giả: Xây dựng ý tưởng nghiên cứu: T.H.T., M.V.K., N.B.T.; Điều tra, khảo sát, phân tích số liệu: N.B.T., P.K.N., B.M.H.; Viết bản thảo bài báo: N.B.T., P.K.N.; Chỉnh sửa bài báo: T.H.T., M.V.K., N.B.T., P.K.N.

Lời cảm ơn: Nghiên cứu này được tài trợ bởi đề tài nghiên cứu khoa học cấp Bộ Tài nguyên và Môi trường mã số TNMT.2022.06.04. Tập thể tác giả xin chân thành cảm ơn.

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

Tài liệu tham khảo

1. WAMDI Group. The WAM Model—A Third Generation Ocean Wave Prediction Model. *J. Phys. Oceanogr.* **1988**, *18*, 1775-1810.
2. Tolman, H.L. User manual and system documentation of WAVEWATCH III TM version 3.14. Technical note, MMAB Contribution, 2019.
3. Booij, N.; Ris, R.C.; Holthuijsen, L.H. A third-generation wave model for coastal regions: Model description and validation. *J. Geophys. Res. Oceans* **1999**, *104(C4)*, 7649-7666.
4. Kim, S.K.; Takedab, M.; Mase, H.M. GMDH-based wave prediction model for one-week nearshore waves using one-week forecasted global wave data. *Appl. Ocean Res.* **2021**, *117*, 102859.
5. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1999**, *9 (8)*, 1735–1780.
6. Kagemoto, H. Forecasting a water-surface wave train with artificial intelligence-a case study. *Ocean. Eng.* **2020**, *207*, 107380.
7. Fan, S.; Xiao, N.; Dong, S. A novel model to predict significant wave height based on long short-term memory network. *Ocean. Eng.* **2020**, *205*, 107298.
8. Pirhooshyaran, M.; Snyder, L.V. Forecasting, hindcasting and feature selection of ocean waves via recurrent and sequence-to-sequence networks. *Ocean. Eng.* **2020**, *207*, 107424.
9. Ni, C.; Ma, X. An integrated long-short term memory algorithm for predicting polar westerlies wave height. *Ocean. Eng.* **2020**, *215*, 107715.
10. Kim, K.S.; Lee, J.B.; Roh, M.I.; Han, K.M.; Lee, G.H. Prediction of ocean weather based on denoising autoen-coder and convolutional lstm. *J. Mar. Sci. Eng.* **2020**, *8*, 805
11. Gers, F.A.; Schmidhuber J.; Cummins F. Learning to forget: Continual prediction with LSTM. *Neural comput.* **2000**, *12(10)*, 2451-2471.
12. https://machinelearningcoban.com/.
13. Hùng, H.V.; Tuấn, H.V. Sử dụng mạng nơ-ron nhân tạo dự báo mực nước sông chịu ảnh hưởng của thủy triều. *Tạp chí Khoa học và Công nghệ Thủy lợi* **2019**, *52*.
14. https://nguyentruonglong.net/giai-thich-chi-tiet-ve-mang-long-short-term-memory-lstm.html
15. https://colah.github.io/posts/2015-08-Understanding-LSTMs/
16. https://github.com/phamdinhkhanh/LSTM/blob/master/LSTM.ipynb
17. https://phantichspss.com/r-binh-phuong-r-binh-phuong-hieu-chinh-cong-thuc-y-nghia-cach-tinh-thu-cong-va-cach-tinh-bang-spss.html
18. Mengning, W.; Christos, S.; Zhen, G. Multi-Step-Ahead Forecasting of Wave Conditions Based on a Physics-Based Machine Learning (PBML) Model for Marine Operations. *J. Mar. Sci. Eng.* **2020**, *8(12)*, 992.

19. Trung, T.D.; Vinh, T.N.; Kim, J. Improving the Accuracy of Dam Inflow Predictions Using a Long Short-Term Memory Network Coupled with Wavelet Transform and Predictor Selection. *Mathematics* **2021**, *9(5)*, 551.
20. Báo cáo tổng hợp đề tài nghiên cứu khoa học cấp nhà nước: Nghiên cứu cơ sở khoa học và giải pháp ứng dụng trí tuệ nhân tạo để nhận dạng, hỗ trợ dự báo và cảnh báo một số hiện tượng khí tượng thủy văn nguy hiểm trong bối cảnh biến đổi khí hậu tại Việt Nam. 2020. Chủ nhiệm đề tài Ths. Ngô Văn Mạnh.
21. Hiền, L.X.; Hùng, H.V.; Lee, G. Xây dựng mô hình mạng nơ-ron hồi quy dựa trên phần mềm mã nguồn mở để dự báo lưu lượng dòng chảy. *Tuyển tập Hội nghị Khoa học thường niên năm* **2018**. ISBN: 978-604-82-2548-3.
22. Deepthi, I.G.; Dwarakish, G.S. Wave Prediction Using Neural Networks at New Mangalore Port along West Coast of India. *Aquat. Procedia* **2015**, *4*, 143-150.
23. Deo, M.C.; Naidu, C.S. Real time wave forecasting using neural networks. *Ocean. Eng.* **1998**, *26(3),* 191-203.
24. James, S.C.; Zhang, Y.; O'Donncha, F. A machine learning framework to forecast wave conditions. *Coastal Eng*. **2018**, *137*, 1–10.

Building a regression neural network model to predict significant wave heights at Con Co station, Quang Tri, Vietnam

Tran Hong Thai[1], Mai Van Khiem[2], Nguyen Ba Thuy[2*], Bui Manh Ha[2], Pham Khanh Ngoc[2]

[1] Viet Nam Meteorological and Hydrological Administration (VNMHA); tranthai.vkttv@gmail.com
[2] National Center for Hydro-Meteorological Forecasting; maikhiem77@gmail.com; thuybanguyen@gmail.com; manhhamhc@gmail.com; ngocpkchibo@gmail.com

Abstract: In recent years, artificial intelligence (AI) has been applied to many different sectors and industries, including marine hydrometeorological forecasting. The application of Long Short-Term Memory (LSTM) technique which is an improved version from Recurrent Neural Network (RNN) and its results for wave prediction at the Con Co station in Quang Tri province, Vietnam is presented in this paper. The observations of wave height have been analyzed with the statistical characteristics and correlation with the observed parameters to select the inputs for training the prediction model. Two models have been built based on the numbers of inputs that are univariate model and two-variable model. These models were implemented to predict the significant wave height with the prediction periods of 06, 12, 18 and 24h. The results show that these developed models are good to compute significant wave height for the period of 6 hours only. At that predict period, the accuracy of prediction is $R^2 = 0{,}582$ for two-variables. The quality of wave observation data at Con Co station is not reliable and detailed, which is the cause of large forecasting errors at longer forecast periods.

Key words: Wave forecasting; Machine Learning; LSTM; AI, RNN.