# ỨNG DỤNG MẠNG NƠ-RON HỒI QUY ĐỂ XÂY DỰNG LẠI DỮ LIỆU DÒNG CHẢY NGÀY BỊ THIẾU

**Lê Xuân Hiền**[1]

**Tóm tắt:** *Lưu lượng sông là một trong những dữ liệu quan trọng nhất trong thủy văn bởi các dữ liệu này có thể được sử dụng cho các phân tích liên quan tới quản lý tài nguyên nước cũng như dự báo dòng chảy lũ. Việc thiếu dữ liệu dòng chảy có thể dẫn tới các phân tích khoa học không đầy đủ. Để có được những thông tin đáng tin cậy và chính xác hơn thì những dữ liệu bị thiếu này phải được lấp đầy. Mục tiêu của bài báo này là giới thiệu một cách tiệp cận hiệu quả dựa trên mô hình mạng nơ-ron hồi quy để xây dựng lại dữ liệu dòng chảy hàng ngày bị thiếu. Trạm thủy văn Lai Châu được chọn làm trạm mục tiêu cho nghiên cứu điển hình bởi đây là trạm thủy văn nằm ở thượng lưu của lưu vực sông Đà. Kết quả nghiên cứu thể hiện hiệu suất cao của mô hình mạng nơ-ron hồi quy. Với kết quả này, mô hình hoàn toàn có thể được áp dụng cho các trạm thủy văn ở thượng nguồn nơi mà thiếu các dữ liệu về dòng chảy.*
**Từ khóa**: GRU, RNN, dữ liệu dòng chảy bị thiếu, khôi phục dữ liệu.

## 1. MỞ ĐẦU

Trong thủy văn, bên cạnh các dữ liệu về lượng mưa và độ ẩm của đất, các dữ liệu về dòng chảy trên lưu vực sông đóng một vai trò rất quan trọng. Các dữ liệu này có thể được sử dụng cho công tác quản lý và vận hành tài nguyên nước, dự báo dòng chảy hoặc các phân tích liên quan tới biến đổi khí hậu. Một đặc điểm chung với các bài toán này là yêu cầu một chuỗi dữ liệu đáng tin cậy theo thời gian. Các chuỗi dữ liệu dài và liên tục sẽ cho phép các nhà khoa học có thể đưa ra các phân tích chính xác hơn về các tiến trình thủy văn đầu nguồn. Tuy nhiên, việc thu thập các dữ liệu thủy văn liên tục trong thời gian dài là một vấn đề khó khăn bởi đôi khi các dữ liệu này có thể bị thiếu hoặc mất do quá trình lưu trữ, bảo trì thiết bị hoặc cũng có thể các thiết bị đo bị hỏng do các sự kiện lũ. Đối với các trạm thủy văn ở khu vực miền núi cao hoặc ở các nước đang phát triển, việc thu thập đầy đủ các chuỗi dữ liệu dòng chảy càng trở nên khó khăn hơn. Việc thiếu dữ liệu dòng chảy trong một khoảng thời gian có thể dẫn tới các phân tích khoa học không đầy đủ. Do đó, để có được những thông tin đáng tin cậy và chính xác từ dữ liệu, những khoảng trống dữ liệu này nên được lấp đầy.

Bài toán ước tính các dữ liệu dòng chảy bị thiếu theo thời gian là một bài toán đã được nghiên cứu từ nhiều thập kỷ trước đây và cho đến hiện nay, bài toán này vẫn đang là một thách thức đáng kể với các nhà khoa học. Một số giải pháp đã được thực hiện để xây dựng lại các dữ liệu bị thiếu. Có thể kể đến như, cách tiếp cận dựa trên các phân tích hồi quy (Tencaliec et al. 2015; Woodhouse et al. 2006) hay các cách tiếp cận dựa trên mạng nơ-ron nhân tạo (Ben Aissia et al. 2017; Gao and Wang 2017; Sivapragasam et al. 2015). Cùng với đó, Harvey et al. (2012) đã chỉ ra rằng, việc sử dụng mô hình với nhiều biến đầu vào có thể đưa ra các kết quả có độ chính xác cao hơn so với việc chỉ sử dụng những mô hình hồi quy đơn giản. Tuy nhiên, trong hầu hết các nghiên cứu về xây dựng lại dữ liệu dòng chảy bị thiếu được đề cập tới ở trên, dữ liệu được ước tính là các dữ liệu dòng chảy ở hạ lưu. Điều đó có nghĩa là các nghiên cứu trước đây sử dụng các dữ liệu ở thượng nguồn như là dữ liệu đầu vào để ước tính cho dữ liệu dòng chảy bị thiếu ở hạ lưu.

Trong bài báo này, một mô hình mạng nơ-ron hồi quy (RNN - recurrent neural network) dựa trên mạng nơ-ron nhân tạo (ANN- artificial neural network) đã được xây dựng với mục đích ước tính

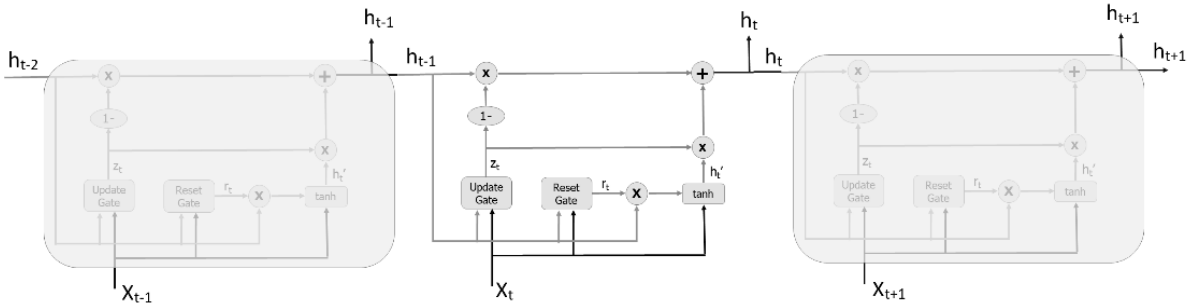
[1] *Khoa Kỹ thuật Tài nguyên nước, Trường Đại học Thủy lợi*

dữ liệu dòng chảy bị thiếu. Mô hình RNN được áp dụng trong nghiên cứu này để ước tính các dữ liệu dòng chảy bị thiếu tại các trạm thủy văn ở thượng nguồn của lưu vực sông. Đây là một trong những yếu tố quan trọng khiến cho nghiên cứu này khác biệt so với các nghiên cứu trước đây. Với mục đích đánh giá khả năng của mô hình RNN trong bài toán xây dựng lại dữ liệu dòng chảy bị thiếu, trạm thủy văn Lai Châu nằm ở thượng nguồn của lưu vực sông Đà đã được chọn làm nghiên cứu điển hình. Kết quả nghiên cứu này có thể được áp dụng để xây dựng lại dữ liệu dòng chảy bị thiếu tại các trạm thủy văn đầu nguồn khác như trạm Lào Cai hay trạm Bảo Yên, tỉnh Lào Cai, Việt Nam.

## 2. PHƯƠNG PHÁP NGHIÊN CỨU

### 2.1. Mô hình Gated Recurrent Unit (GRU)

Trong bài báo này, một mô hình mạng nơ-ron GRU đã được áp dụng để xây dựng mô hình khôi phục dữ liệu dòng chảy tại trạm Lai Châu. Mạng GRU là một dạng đặc biệt của mạng nơ-ron hồi quy, được đề xuất bởi Cho et al. (2014) để giải quyết các vấn đề về biến mất đạo hàm trong các bài toán về chuỗi thời gian. GRU cùng với LSTM (Long Short-Term Memory) là các kiến trúc mạng được sử dụng rộng rãi nhất trong các nghiên cứu về các bài toán dữ liệu tuần tự hoặc chuỗi thời gian. Về cơ bản, ý tưởng cốt lõi của RNN là sử dụng các ô bộ nhớ để lưu trữ các thông tin cần thiết từ các bước xử lý trước để đưa ra các dự báo chính xác nhất cho các bước tiếp theo. Cấu trúc của một ô bộ nhớ RNN với kiến trúc GRU được thể hiện như Hình 1.




*Hình 1. Cấu trúc của một ô nhớ GRU (GRU cell) trong mô hình*

Theo Chung et al. (2014), kiến trúc GRU không có các ô nhớ tách biệt như LSTM. Thay vì có ba lớp cổng trong mỗi ô như kiến trúc LSTM, GRU chỉ có hai lớp cổng, đó là cổng đặt lại (reset gate - $r_t$) và cổng cập nhật (update gate – $z_t$). Trong khi cổng đặt lại ($r_t$) sẽ xác định lượng thông tin cần bỏ qua từ các bộ nhớ trước thì cổng cập nhật ($z_t$) sẽ quyết định những thông tin từ bộ nhớ trước đó có thể được truyền qua nó. Chính vì vậy, kiến trúc mạng được đào tạo để có thể giữ được lượng thông tin từ các bước trước đó mà không cần loại bỏ các thông tin không liên quan tới việc dự báo. Ở bước cuối cùng trong kiến trúc mạng, đầu ra của một ô nhớ GRU hay trạng thái ẩn (hidden state – $h_t$) tại thời điểm $t$ được xác định bởi các phương trình sau:

$$z_t = \sigma\left(W_z x_t + U_z h_{t-1}\right) \quad (1)$$

$$r_t = \sigma\left(W_r x_t + U_r h_{t-1}\right) \quad (2)$$

$$h_t' = tanh\left(W_h x_t + r_t \otimes U_h h_{t-1}\right) \quad (3)$$

$$h_t = (1 - z_t) \otimes h_t' + z_t \otimes h_{t-1} \quad (4)$$

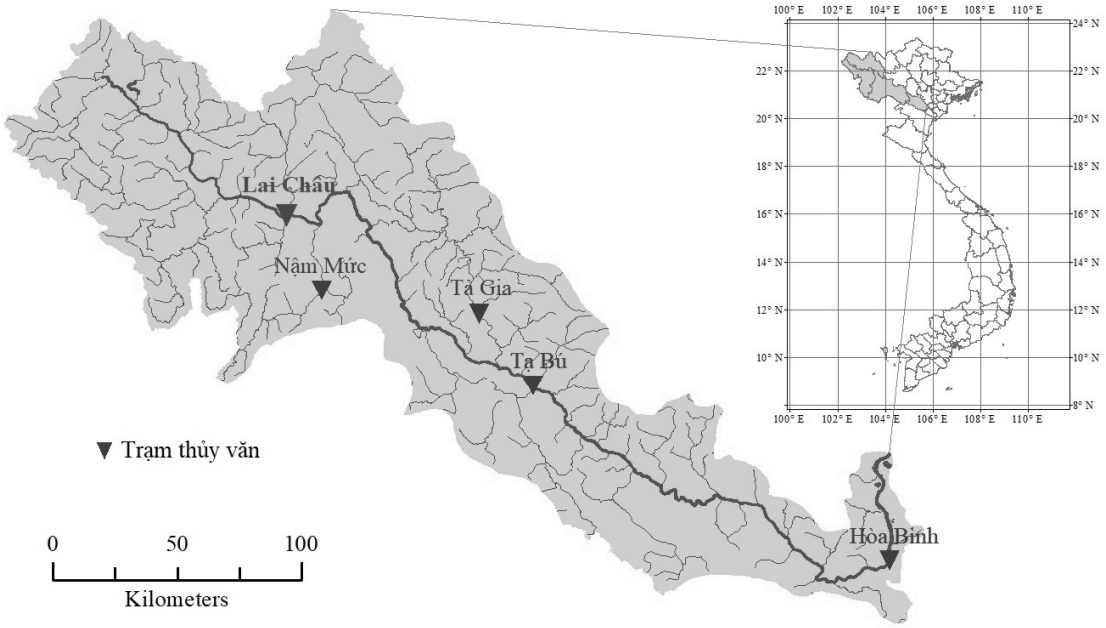
Trong các phương trình trên, $W_i$ và $U_i$ là các ma trận trọng số; $b_i$ là các hệ số; σ là hàm kích hoạt sigmoid; $r_t$ và $z_t$ là cổng đặt lại và cổng cập nhật tại bước thời gian thứ $t$; $h_t'$ là ứng viên cho giá trị lớp ẩn; và ⊗ biểu thị phép nhân các phần tử của ma trận (element-wise multiplication).

### 2.2. Khu vực nghiên cứu

Sông Đà nằm ở khu vực phía Tây Bắc, là phụ lưu lớn nhất của sông Hồng, một trong những lưu vực sông lớn nhất ở Việt Nam. Bắt nguồn từ Trung Quốc, lưu vực sông Đà trải dài theo hướng tây bắc – đông nam. Ở Việt Nam, sông Đà chảy qua các tỉnh Lai Châu, Điện Biên, Sơn La và Hòa Bình trước khi nhập vào sông Hồng ở Phú Thọ. Hiện nay, trên lưu vực sông Đà có ba đập thủy điện lớn là đập Hòa Bình

(1994), đập Sơn La (2012) và đập Lai Châu (2016) với tổng công suất lắp máy khoảng 5520 MW.

Nghiên cứu này tập trung xây dựng một mô hình mạng GRU để xây dựng lại dữ liệu dòng chảy bị thiếu hoặc bị mất trên các lưu vực sông. Thông thường, các dữ liệu dòng chảy ở thượng lưu sẽ được sử dụng làm dữ liệu đầu vào cho các mô hình để đưa ra các tính toán hoặc dự báo dòng chảy ở hạ lưu. Khác với các mô hình thủy văn thông thường cũng như mô hình dựa trên phương pháp hướng dữ liệu (data-driven method), mô hình đề xuất sử dụng dữ liệu đầu vào là dữ liệu dòng chảy ngày được quan sát tại các trạm thủy văn ở hạ lưu để tính toán và ước tính cho trạm mục tiêu ở thượng lưu. Lưu vực sông Đà được lựa chọn làm nghiên cứu điển hình và trạm thủy văn Lai Châu nằm ở thượng lưu được chọn làm trạm mục tiêu cho nghiên cứu này. Khu vực nghiên cứu bao gồm năm trạm thủy văn, trong đó có bốn trạm ở hạ lưu lần lượt là: Nậm Mức, Tà Gia, Tạ Bú, Hòa Bình; và trạm mục tiêu – Lai Châu. Sơ đồ vị trí của các trạm thủy văn trong khu vực nghiên cứu được thể hiện ở Hình 2.




*Hình 2. Sơ đồ khu vực nghiên cứu và vị trí các trạm thủy văn*

Dữ liệu dòng chảy tại 5 trạm thủy văn được thu thập từ trung tâm dự báo khí tượng thủy văn. Đây là các dữ liệu lưu lượng ngày thực đo trong 24 năm, từ 1961 đến 1984, trước khi đập thủy điện Hòa Bình được xây dựng. Các dữ liệu về lưu lượng được đo đạc với đơn vị là $m^3/s$.

**2.3. Các tiêu chí đánh giá mô hình**

Hiệu suất của mô hình được đánh giá thông qua ba trị số lần lượt là bình quân sai số tuyệt đối (MAE - mean absolute error), sai số căn quân phương (RMSE - root mean squared error), và hệ số hiệu quả Nash (NSE - Nash-Sutcliffe Efficiency). Đây là các trị số thường được sử dụng khi so sánh các giá trị thực đo với các giá trị được tính toán trong các mô hình thủy văn. Các trị số này được tính toán như sau:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|P_i - O_i\right| \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(O_i - P_i\right)^2} \quad (6)$$

$$NSE = 1 - \frac{\sum_{i=1}^{n}\left(O_i - P_i\right)^2}{\sum_{i=1}^{n}\left(O_i - \overline{O}_i\right)^2} \quad (7)$$

Trong đó: $O_i$, $\overline{O}_i$, và $P_i$ lần lượt là giá trị thực đo, giá trị thực đo trung bình và giá trị tính toán của mẫu thứ $i$ tương ứng. Mô hình cho kết quả tốt nếu các giá trị MAE, RMSE nhỏ và NSE lớn.

## 3. THIẾT LẬP THÔNG SỐ MÔ HÌNH

Mô hình mạng GRU được đề xuất cho nghiên cứu này dựa trên các thư viện phần mềm mã nguồn mở. Trong đó, Python là ngôn ngữ lập trình được lựa chọn cho nghiên cứu và các thư viện như NumPy, Pandas, Matplotlib, và TensorFlow được sử dụng để xử lý, quản lý dữ liệu và xây dựng mô hình.

Các dữ liệu thu thập được chia thành các tập dữ liệu độc lập với mục đích huấn luyện, hiệu chỉnh và kiểm định mô hình. Cụ thể, tập dữ liệu đầu tiên là chuỗi lưu lượng thực đo hàng ngày trong 23 năm (1961-1983) được sử dụng với mục đích huấn luyện và hiệu chỉnh mô hình. Trong đó, 80% dữ liệu được sử dụng cho huấn luyện và 20% dữ liệu được sử dụng cho mục đích hiệu chỉnh. Tập dữ liệu còn lại là 1 năm (1984) được sử dụng với mục đích so sánh giữa các giá trị được ước tính và các giá trị thực đo để kiểm định hiệu suất của mô hình đề xuất. Các thông số cũng như cấu trúc cơ bản của mô hình đề xuất được thể hiện ở Bảng 1.

**Bảng 1. Các thông số cơ bản của mô hình GRU**

| **Đặc trưng** | **Chi tiết** |
|---|---|
| Mục tiêu | Bổ sung lại dữ liệu dòng chảy tại trạm thủy văn Lai Châu |
| Dữ liệu đầu vào | Lưu lượng thực đo tại 5 trạm thủy văn |
| Thông số mô hình | Chiều dài chuỗi: 20<br>Hệ số học: 0,001<br>Số lượng unit: 20; 30; 50;<br>Số lượng epoch tối đa: 100.000 |

Để mô hình GRU có thể đạt hiệu suất tốt hơn cũng như học được các sự phụ thuộc từ dữ liệu hiệu quả hơn, đã có một vài sự thay đổi trong việc sắp xếp dữ liệu đầu vào cho mô hình. Thay vì sử dụng vectơ dữ liệu đầu vào chỉ là dữ liệu tại một bước thời gian cụ thể, vectơ này đã được định dạng thành dạng chuỗi các dữ liệu đầu vào (ma trận) với chiều dài chuỗi là 20. Mỗi bước thời gian tương ứng với một lần được quan sát. Trong bài báo này, dữ liệu dòng chảy được quan sát theo ngày. Điều này có nghĩa là mô hình sử dụng dữ liệu đầu vào là dữ liệu của 20 bước thời gian (tương ứng 20 ngày) được quan sát gần nhất để đưa ra các tính toán cho bước thời gian (hoặc ngày) tiếp theo. Thêm vào đó, nghiên cứu này là bài toán khôi phục lưu lượng dòng chảy trên sông và sử dụng dữ liệu từ hạ lưu để tính toán cho thượng lưu. Chính vì vậy, việc lựa chọn giá trị chiều dài chuỗi là 20 ngày cũng là để đảm bảo mô hình có thể học được đầy đủ quá trình xuất hiện một trận lũ từ lúc hình thành tới lúc kết thúc. Dữ liệu dòng chảy cần ước tính của trạm Lai Châu ở bước thời gian $t$ bất kỳ ($X_t^5$) sẽ được tính toán dựa trên công thức sau:

$$X_t^5 = f\begin{pmatrix} X_t^1, X_{t-1}^1, ..., X_{t-19}^1; \\ X_t^2, X_{t-1}^2, ..., X_{t-19}^2; \\ X_t^3, X_{t-1}^3, ..., X_{t-19}^3; \\ X_t^4, X_{t-1}^4, ..., X_{t-19}^4; \\ X_{t-1}^5, X_{t-2}^5, ..., X_{t-20}^5 \end{pmatrix} \quad (8)$$

Trong đó: $X_t^1$, $X_t^2$, $X_t^3$, $X_t^4$, và $X_t^5$ lần lượt là dữ liệu dòng chảy tại các trạm Nậm Mức, Tả Gia, Tạ Bú, Hòa Bình, và Lai Châu tại thời điểm $t$ bất kỳ.

Trong mô hình mạng nơ-ron, quá trình tối ưu hóa sẽ phụ thuộc vào hàm tối ưu hóa, thuật toán tối ưu hóa và hệ số học (learning rate). Hệ số học có liên quan chặt chẽ với thuật toán tối ưu hóa được lựa chọn. Trong bài báo này, thuật toán tối ưu hóa Adam (Kingma and Ba 2014) được lựa chọn và hệ số học mặc định là 0,001. Đây là thuật toán được sử dụng rộng rãi trong các bài toán học sâu (deep learning) vì hiệu quả của nó. Một thông số khác cũng được lựa chọn trong việc xây dựng mô hình đó là số lượng unit. Số lượng unit được hiểu như là số lượng nơ-ron trong mỗi tế bào GRU (GRU cell). Việc lựa chọn các giá trị này khác nhau với mục đích nhằm đánh giá ảnh hưởng

của số lượng unit đến hiệu suất mô hình. Số lượng unit trong mỗi tế bào GRU càng lớn thì độ phức tạp của mô hình càng tăng lên và thời gian để tính toán và cập nhập mỗi vòng lặp (epoch) sẽ tăng lên đáng kể. Ngoài ra, mô hình cũng được thiết lập để huấn luyện với số lần lặp tối đa là 100.000 lần.

Trong trường hợp kiểm định mô hình với tập dữ liệu độc lập năm 1984, để có thể đưa ra được chuỗi các giá trị dòng chảy được ước tính trong 1 năm, mô hình đã được thiết lập để xây dựng một chuỗi các vòng lặp tính toán liên tục. Ý tưởng cốt lõi của việc xây dựng vòng lặp tính toán là sử dụng giá trị được ước tính tại một một vòng lặp bất kỳ làm dữ liệu đầu vào của vòng lặp tiếp theo để tính toán và đưa ra các chuỗi giá trị theo yêu cầu.

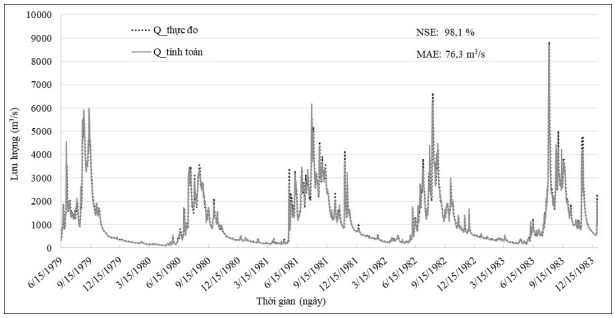
## 4. KẾT QUẢ NGHIÊN CỨU

### 4.1. Kết quả hiệu chỉnh mô hình

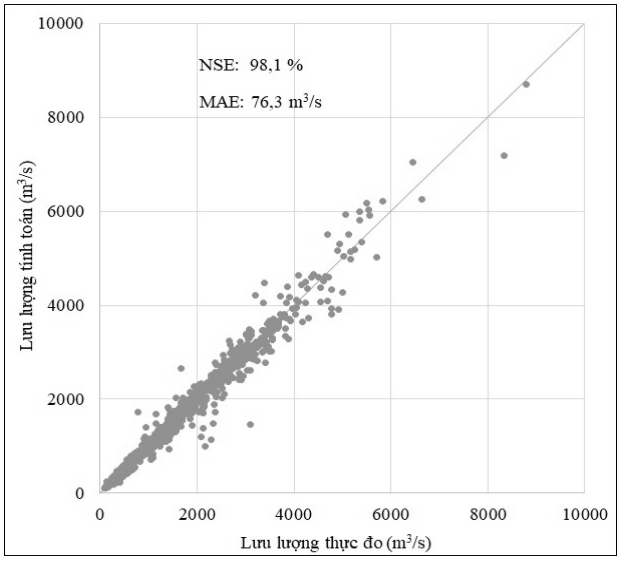
Mô hình đề xuất được huấn luyện và hiệu chỉnh với tập dữ liệu trong 23 năm từ 1961 đến 1983 tương ứng với tỉ lệ dữ liệu lần lượt là 80% và 20%. Kết quả hiệu chỉnh của mô hình được thể hiện ở Bảng 2. So sánh trực quan giữa giá trị thực đo và giá trị ước tính trong quá trình hiệu chỉnh được thể hiện trong Hình 3 và Hình 4.

**Bảng 2. Kết quả hiệu chỉnh mô hình GRU**

| Trường hợp | Chiều dài chuỗi | Số lượng unit | Hệ số học | Số lượng epoch | MAE ($m^3/s$) | RMSE ($m^3/s$) | NSE |
|---|---|---|---|---|---|---|---|
| TH1 | 20 | 20 | 0,001 | 9455 | 76,3 | 159,5 | 0,981 |
| TH2 | 20 | 30 | 0,001 | 8147 | 75,9 | 158,7 | 0,981 |
| TH3 | 20 | 50 | 0,001 | 5226 | 75,9 | 158,9 | 0,981 |




*Hình 3. So sánh giữa lưu lượng thực đo với tính toán trong quá trình hiệu chỉnh (TH1)*




*Hình 4. Tương quan giữa giá trị thực đo và tính toán trong quá trình hiệu chỉnh (TH1)*

Kết quả hiệu chỉnh mô hình cho thấy không có sự khác biệt giữa ba trường hợp được lựa chọn mặc số lượng unit trong mỗi tế bào GRU đã được thay đổi. Giá trị NSE trong cả ba trường hợp đều đạt 98,1% khi so sánh giữa lưu lượng được ước tính và lưu lượng thực đo. Các giá trị MAE và RMSE cũng cho thấy xu hướng tương tự như vậy, giá trị sai số giữa lưu lượng ước tính và thực đo trong cả ba trường hợp đều tương tự nhau, lần lượt là 76 $m^3/s$ và 160 $m^3/s$. Bảng 2 cũng cho thấy một xu hướng quan trọng khác, khi số lượng unit tăng lên thì số lượng epoch sẽ giảm xuống. Điều này có nghĩa là khi độ phức tạp của mô hình tăng lên hay thời gian tính toán cho mỗi vòng lặp tăng lên thì số lần tính toán (vòng lặp) để mô hình đạt được giá trị tối ưu sẽ giảm đi. Thời gian tính toán đối với mô hình mạng nơ-ron phụ thuộc vào cấu hình của thiết bị sử dụng.

Hình 3 và Hình 4 cho thấy rằng có sự tương quan chặt chẽ giữa giá trị lưu lượng được mô phỏng và giá trị được quan sát. Đối với trường hợp 1, hệ số tương quan NSE lên tới 98,1% trong khi trung bình sai số tuyệt đối chỉ là 73,3 $m^3/s$. Hình 3 cho thấy khả năng mô phỏng của mô hình trong trường hợp khôi phục dữ liệu dòng chảy vào

mùa lũ khi mà đỉnh lũ được tính toán xuất hiện cùng thời điểm với đỉnh lũ thực tế. Thêm vào đó, Hình 4 cũng cho thấy các giá trị được mô phỏng phù hợp với giá trị thực đo và sai số tuyệt đối trong trường hợp xuất hiện đỉnh lũ chỉ là 103,5 $m^3/s$ (so với giá trị đỉnh lũ thực đo là 8800 $m^3/s$), mức sai số tương ứng chỉ khoảng 1,2%.

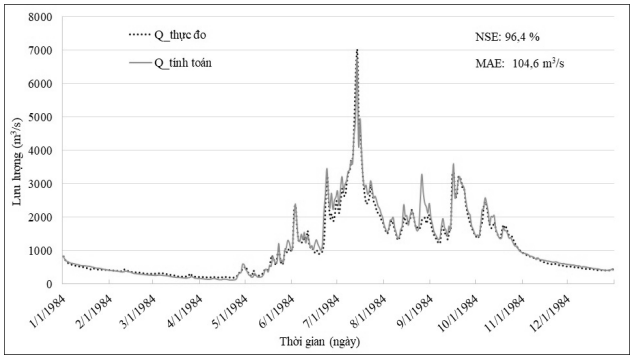
### 4.2. Kết quả kiểm định mô hình

Sau quá trình hiệu chỉnh, mô hình được kiểm định với tập dữ liệu độc lập năm 1984. Đây là tập dữ liệu chưa từng được sử dụng trước đó và mục đích kiểm định là để đánh giá khả năng tính toán của mô hình đề xuất. Kết quả kiểm định của mô hình GRU được thể hiện ở Bảng 3.

**Bảng 3. Kết quả kiểm định của mô hình GRU**

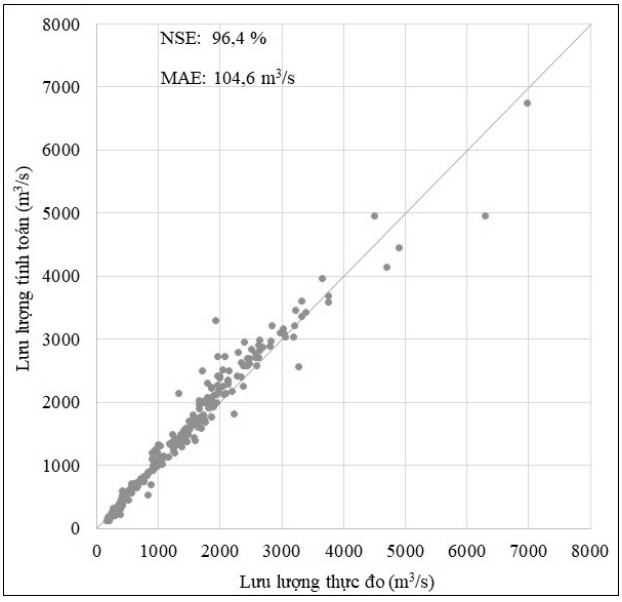
| Trường hợp | Chiều dài chuỗi | Số lượng unit | Hệ số học | Số lượng epoch | MAE ($m^3/s$) | RMSE ($m^3/s$) | NSE |
|---|---|---|---|---|---|---|---|
| TH1 | 20 | 20 | 0,001 | 9455 | 104,6 | 188,2 | 0964 |
| TH2 | 20 | 30 | 0,001 | 8147 | 154,5 | 228,8 | 0,947 |
| TH3 | 20 | 50 | 0,001 | 5226 | 126,6 | 212,7 | 0,954 |

Kết quả kiểm định cho thấy mô hình GRU vẫn đạt được kết quả rất ấn tượng. Mặc dù đã có một vài sự khác biệt nhỏ khi so sánh kết quả của ba trường hợp tính toán, nhưng có thể nói sự khác biệt này là không đáng kể khi mà độ chính xác (giá trị NSE) của mô hình vẫn đạt trên 95-96%. Trường hợp 1 (TH1) cho hiệu suất mô hình ổn định hơn cả so với 2 trường hợp còn lại. Giá trị sai số tương ứng MAE và RMSE trong quá trình kiểm định lần lượt là khoảng 105 $m^3/s$ và 190 $m^3/s$. Kết quả so sánh trực quan giữa giá trị thực đo và giá trị mô phỏng được thể hiện ở Hình 5 và Hình 6.




*Hình 5. So sánh giữa lưu lượng thực đo với tính toán trong quá trình kiểm định (TH1)*

Hình 5 so sánh tương quan giữa giá trị thực đo và tính toán trong trường hợp TH1. Có thể thấy rằng đỉnh đỉnh lũ dự báo vào đỉnh lũ thực đo xuất hiện cùng thời điểm vào khoảng giữa tháng 7. Sai số tuyệt đối trong trường hợp tính toán giá trị đỉnh lũ là 230,7 $m^3/s$ tương ứng với mức sai số tương đối khoảng 3,3%. Hình 6 biểu diễn các cặp dữ liệu được ghép đôi giữa giá trị thực đo và giá trị được tính toán. Các cặp dữ liệu này càng nằm gần đường chéo $45^0$ thì mô hình càng đạt hiệu suất. Có thể thấy rằng, các kết quả tính toán trong quá trình kiểm định có sự phù hợp cao với các giá trị thực đo. Các kết quả này khẳng định rằng mô hình đề xuất cho kết quả tính toán có độ chính xác cao và ổn định.




*Hình 6. Tương quan giữa giá trị thực đo và tính toán trong quá trình kiểm định (TH1)*

## 5. KẾT LUẬN

Trong bài báo này, tác giả đã xây dựng một mô

hình mạng GRU dựa trên mạng nơ-ron hồi quy với mục đích xây dựng lại dữ liệu dòng chảy ngày tại trạm Lai Châu trên sông Đà. Mặc dù chỉ sử dụng một lượng khiêm tốn dữ liệu, nhưng kết quả tính toán của mô hình đề xuất đã thể hiện sự phù hợp với dữ liệu thực đo. Các kết quả này đã được đánh giá một cách cẩn thận thông qua các quá trình huấn luyện, hiệu chỉnh và kiểm định. Cả ba trường hợp nghiên cứu đều đạt được hiệu suất xuất sắc gần như nhau khi mà các thông số của mô hình được thay đổi. Điều này cho thấy mô hình đề xuất đã thể hiện sự ổn định và cho hiệu suất cao.

Kết quả của nghiên cứu phụ thuộc vào các dữ liệu được thu thập. Trong nghiên cứu này, dữ liệu đầu vào là các giá trị lưu lượng thực đo tại các trạm thủy văn hạ lưu. Các dữ liệu về lượng mưa trong khu vực nghiên cứu cũng đã được quan tâm, tuy nhiên, việc đưa thêm các dữ liệu về lượng mưa không làm hiệu suất của mô hình tăng lên. Điều này có thể giải thích vì sự tương quan giữa dữ liệu về lượng mưa và lưu lượng tại trạm mục tiêu nhỏ hơn rất nhiều so với tự tương quan giữa lưu lượng và lưu lượng. Hơn nữa, trong mô hình mạng nơ-ron hồi quy, tương quan dữ liệu càng cao thì hiệu suất mô hình sẽ càng tốt (Le et al. 2019).

Mô hình mạng GRU hay mô hình mạng nơ-ron hồi quy đều là các mô hình dựa trên phương pháp định hướng dữ liệu. Phương pháp này có ưu điểm là đơn giản hơn so với các phương pháp dựa trên các mô hình vật lý vì không yêu cầu nhiều dữ liệu đầu vào như tình hình sử dụng đất hay diện tích bề mặt. Nghiên cứu này là bước đầu tiên trong việc xây dựng mô hình để tính toán và khôi phục lại dữ liệu dòng chảy tại trạm Lào Cai trên sông Hồng, nơi mà dữ liệu dòng chảy bị mất trong 15 năm từ 1979 đến 1994. Với kết quả nghiên cứu này, mô hình mạng nơ-ron hồi quy hoàn toàn có thể được áp dụng để ước tính và xây dựng lại các dữ liệu dòng chảy bị mất hoặc bị thiếu ở các trạm thủy văn ở hạ lưu hoặc thậm chí ở thượng lưu trên các lưu vực sông.

## TÀI LIỆU THAM KHẢO


Ben Aissia, M.-A., Chebana, F., and Ouarda, T. B. M. J. (2017). "*Multivariate missing data in hydrology – Review and applications*." Adv. Water Resour., 110, 299-309.

Cho, K., van Merrienboer, B., Gülçehre, Ç., Bougares, F., Schwenk, H., and Bengio, Y. (2014). "*Learning phrase representations using RNN encoder-decoder for statistical machine translation*." CoRR, abs/1406.1078.

Chung, J., Gülçehre, Ç., Cho, K., and Bengio, Y. (2014). "*Empirical evaluation of gated recurrent neural networks on sequence modeling*." CoRR, abs/1412.3555.

Gao, T., and Wang, H. (2017). "*Testing Backpropagation Neural Network Approach in Interpolating Missing Daily Precipitation*." Water, Air, & Soil Pollut., 228(10), 404.

Harvey, C. L., Dixon, H., and Hannaford, J. (2012). "*An appraisal of the performance of data-infilling methods for application to daily mean river flow records in the UK*." Hydrol. Res., 43(5), 618-636.

Kingma, D. P., and Ba, J. (2014). "*Adam: A method for stochastic optimization*." CoRR, abs/1412.6980.

Le, X. H., Ho, H. V., Lee, G., and Jung, S. (2019). "*Application of long short-term memory (LSTM) neural network for flood forecasting.*" Water, 11(7), 1387.

Sivapragasam, C., Muttil, N., Jeselia, M. C., and Visweshwaran, S. (2015). "*Infilling of Rainfall Information Using Genetic Programming.*" Aquatic Procedia, 4, 1016-1022.

Tencaliec, P., Favre, A.-C., Prieur, C., and Mathevet, T. (2015). "*Reconstruction of missing daily streamflow data using dynamic regression models.*" Water Resour. Res., 51(12), 9447-9463.

Woodhouse, C. A., Gray, S. T., and Meko, D. M. (2006). "*Updated streamflow reconstructions for the Upper Colorado River Basin.*" Water Resour. Res., 42(5).

**Abstract:**

**RECONSTRUCTION OF MISSING DAILY STREAMFLOW DATA USING RECURRENT NEURAL NETWORK**

*Streamflow data is one of the most important quantities in hydrology because of these data closely related to water resource management problems as well as flood forecasting problems. The lack of these data can lead to inadequate scientific analysis. Therefore, reconstruction of missing data is an important step to get more reliable and accurate information. The objective of this paper is to introduce an effective approach based on the recurrent neural network model to reconstructing missing daily discharge data. Lai Chau hydrological station, located upstream of the Da River basin, was selected as a case study. The findings of this study demonstrated that the recurrent neural network model yields reliable estimates for the problem of missing data. As a result, the RNN model can be applied to other hydrological stations upstream where the flow data is missing.*

**Keywords**: *GRU, RNN, missing data, data reconstruction, Da River.*

*Ngày nhận bài: 26/7/2019*

*Ngày chấp nhận đăng: 27/8/2019*