

Nâng cao khả năng dự báo dòng chảy của mô hình học máy - Hiện đại hóa công tác dự báo thủy văn tại Việt Nam

Để tài nghiên cứu đã xuất sắc giành giải Nhất sinh viên nghiên cứu khoa học cấp ĐHQGHN và tham gia báo cáo tại Đại hội Thế giới lần thứ 41 của Hiệp hội Kỹ thuật và Nghiên cứu Thủy văn Quốc tế, đồng thời được công bố trên 3 tạp chí khoa học uy tín (trong đó, 2 bài báo tiếng Anh thuộc danh mục Q1 ISI và Scopus). Sự ghi nhận này đã khẳng định giá trị ứng dụng thực tiễn cao, khả năng góp phần vào công tác quản lý tài nguyên nước và vận hành hồ chứa trong điều kiện thực tế của nghiên cứu.

NHỮNG TRĂN TRỞ VỚI NGÀNH THỦY VĂN TẠI VIỆT NAM

Tại Việt Nam, lưu vực sông Đà giữ vai trò đặc biệt quan trọng với hệ thống năm hồ chứa lớn gồm Lai Châu, Bản Chát, Huội Quảng, Sơn

La và Hòa Bình. Đây là các công trình thủy điện chiến lược, đóng góp hơn 40% sản lượng điện toàn quốc, góp phần đảm bảo an ninh năng lượng và điều tiết nước trong mùa khô. Vì vậy, dự báo dòng chảy đầu vào cho hệ thống này (tại hồ Lai Châu, hồ Bản Chát và trạm thủy văn Nậm Giàng) là nhiệm vụ then chốt trong quản lý và vận hành hồ chứa.

Tuy nhiên, lưu vực sông Đà là lưu vực quốc tế điển hình, với gần 50% diện tích thượng nguồn thuộc lãnh thổ Trung Quốc dẫn đến thiếu hụt nguồn dữ liệu quan trắc mặt đất. Thiếu dữ liệu, nhiễu trong quan trắc và ảnh hưởng phức tạp từ hệ thống hồ chứa thượng nguồn khiến việc mô hình hóa và dự báo dòng chảy trở nên đặc biệt khó khăn.

DỰ BÁO DÒNG CHẢY TRÊN LƯU VỰC SÔNG ĐÀ - KHU VỰC GIỮ VAI TRÒ TRỌNG YẾU TRONG AN NINH NĂNG LƯỢNG VÀ QUẢN LÝ TÀI NGUYÊN NƯỚC QUỐC GIA, ĐANG LÀ NHIỆM VỤ TỒN TẠI NHIỀU THÁCH THỨC LỚN TRONG ĐIỀU KIỆN THIẾU DỮ LIỆU. ĐỂ GIẢI QUYẾT BÀI TOÁN NÀY, NHÓM NGHIÊN CỨU GỒM 3 SINH VIÊN: LÊ HỮU MINH QUÂN, NGUYỄN THỊ QUẾ CHI, NGUYỄN THỊ PHƯƠNG ANH ĐẾN TỪ TRƯỜNG ĐH KHOA HỌC TỰ NHIÊN, ĐHQGHN ĐÃ ĐỀ XUẤT PHƯƠNG PHÁP CẢI THIỆN KHẢ NĂNG DỰ BÁO DÒNG CHẢY VỚI MÔ HÌNH MÁY HỌC TIẾN TIẾN, SỬ DỤNG DỮ LIỆU DỰ BÁO ĐA ĐIỂM, ĐA QUY MÔ, VỚI ĐỘ CHÍNH XÁC CAO, ĐỂ DÀNG ỨNG DỤNG VÀO VIỆC QUẢN LÝ HỆ THỐNG HỒ CHỨA.

 BÌNH DIỆU



Các mô hình vật lý truyền thống (HEC-HMS, MIKE, SOBEK, VIC...) tuy mô phỏng được quá trình tự nhiên nhưng đòi hỏi dữ liệu phức tạp, thời gian tính toán dài, hạn chế khả năng ứng dụng thời gian thực. Trong khi đó, các mô hình học máy và học sâu lại nhạy cảm với nhiễu và chưa khai thác đầy đủ đặc trưng mùa vụ của dòng chảy.

Trong bối cảnh biến đổi khí hậu ngày càng khó lường, cùng nhu cầu quy hoạch tổng hợp tài nguyên nước, điều tiết vận hành hồ chứa và cảnh báo thiên tai, việc phát triển mô hình dự báo dòng chảy chính xác, thích ứng với điều kiện thiếu dữ liệu là yêu cầu cấp thiết.

Trên cơ sở đó, nhóm nghiên cứu hướng tới hai mục tiêu chính: (1) Đánh giá khả năng nâng cao hiệu quả dự báo dòng chảy; (2) Đề xuất mô hình dự báo tối ưu cho các điểm quan trắc thượng lưu trên lưu vực sông Đà, phục vụ cho công tác vận hành hồ chứa và quản lý tài nguyên nước trong điều kiện thực tế.

Đề tài vì thế mang tính cấp thiết cả về khoa học lẫn thực tiễn, góp phần hiện đại hóa công tác dự báo thủy văn bằng trí tuệ nhân tạo, đồng thời cung cấp giải pháp khả thi để nâng cao độ chính xác dự báo dòng chảy cho một trong những con sông trọng yếu của Việt Nam.

ĐI TÌM GIẢI PHÁP HIỆN ĐẠI HÓA PHƯƠNG PHÁP DỰ BÁO THỦY VĂN

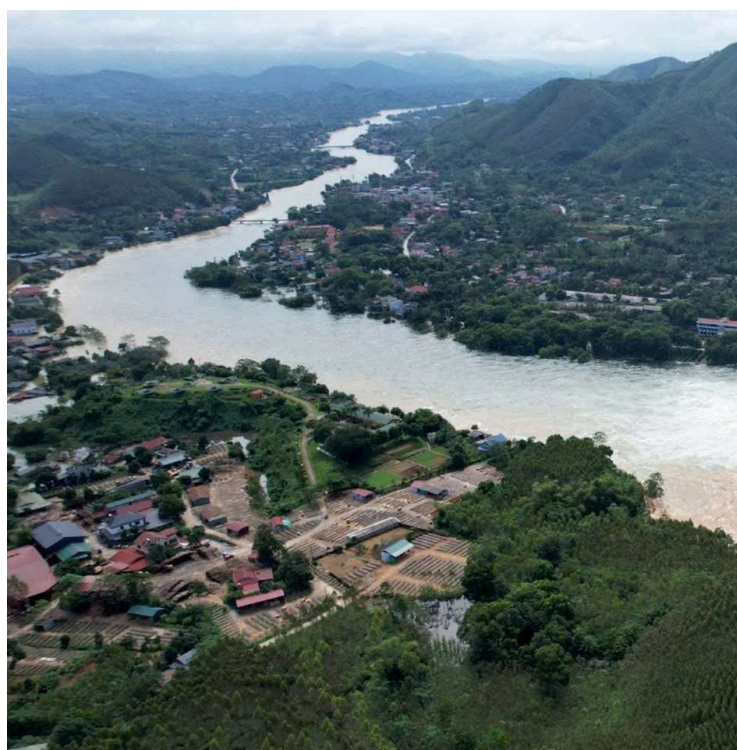
Trên thế giới, dự báo dòng chảy bằng mô hình học máy hiện là phương pháp hiện đại, có tính ứng

dụng cao, có khả năng đưa ra dự báo ngắn hạn và khoảng tin cậy chính xác hơn. Đây cũng chính là phương pháp mà nhóm nghiên cứu lựa chọn nhằm ứng dụng những thành tựu của trí tuệ nhân tạo để hiện đại hóa công tác dự báo thủy văn tại Việt Nam.

Nhóm nghiên cứu đã lựa chọn hướng tiếp cận riêng để tối ưu hiệu quả dựa của mô hình máy học, dựa trên bối cảnh thực tế tại lưu vực Sông Đà. Lê Hữu Minh Quân – đại diện nhóm nghiên cứu, chia sẻ: “Dựa trên đặc điểm dòng chảy ở lưu vực sông Đà, nhóm lựa chọn hướng tiếp cận bằng mô hình tự tương quan dòng chảy, tức là chỉ sử dụng chính chuỗi dòng chảy quan trắc tại trạm trong quá khứ để dự báo tương lai. Cách làm này phù hợp với điều kiện hạn chế dữ liệu, đồng thời đảm bảo tính độc lập và ổn định của mô hình trong thực tiễn vận hành”.

Cụ thể, nhóm nghiên cứu tập trung vào việc khử nhiễu dữ liệu và tăng cường nhận diện đặc trưng dòng chảy theo mùa để nâng cao độ chính xác cho mô hình dự báo. Hướng tiếp cận này được triển khai thông qua việc xử lý và đánh giá trên đa điểm, đa quy mô và theo nhiều bước dự báo, nhằm nâng cao khả năng dự báo của mô hình.

Trước hết, để giúp mô hình học máy tập trung vào tín hiệu dòng chảy chính xác thay vì bị lẫn bởi nhiễu, nhóm nghiên cứu đã áp dụng các kỹ thuật lọc nâng cao. Nghiên cứu sử dụng hai phương pháp lọc phổ biến trong xử lý tín hiệu là



DWT và FFT. Đây là hai công cụ hiệu quả trong việc tách nhiễu ra khỏi dữ liệu mà vẫn giữ lại được các đặc điểm thủy văn quan trọng (như đỉnh lũ, lưu lượng cơ bản).

Cùng với đó, nhóm triển khai kỹ thuật xử lý dữ liệu với phương pháp phân tách chuỗi thời gian dòng chảy thành ba thành phần cơ bản là: Xu hướng (Trend) - Thay đổi dài hạn của dòng chảy; Mùa vụ (Season) - Chu kỳ biến đổi lặp lại theo mùa; Phần dư ngẫu nhiên (Residual) - Hầu hết các thành phần nhiễu.

Sau khi tách, nhiễu trong Phần dư ngẫu nhiên sẽ được loại bỏ bằng phương pháp DWT hoặc FFT, sau đó ba thành phần này sẽ được tổng hợp lại thành một chuỗi dữ liệu mới đã được làm sạch, sẵn sàng đưa vào huấn luyện mô hình.

Đặc biệt, do dòng chảy sông Đà có tính mùa vụ rất rõ rệt, thể hiện qua sự khác biệt lớn giữa mùa lũ và mùa cạn, nhóm nghiên cứu xác định rằng việc nâng cao khả năng nhận diện đặc trưng dòng chảy theo mùa là yếu tố then chốt quyết định độ chính xác của mô hình dự báo.

Để mô hình học máy không chỉ dự báo dựa trên các giá trị dòng chảy trước đó, mà còn hiểu được bối cảnh thời gian đang xảy ra, nghiên cứu đã áp dụng phương pháp One-Hot Encoding. Phương pháp này chuyển đổi các thông tin thời gian mang tính chu kỳ (như tháng hoặc mùa) thành định dạng số mà mô hình học máy dễ dàng xử lý. Việc bổ sung

các đặc trưng thời gian đã được mã hóa vào dữ liệu đầu vào giúp mô hình tăng cường khả năng nhận diện các quy luật thay đổi của dòng chảy theo từng giai đoạn trong năm, từ đó đưa ra dự báo chính xác hơn cho từng mùa.

Hướng tiếp cận với tham vọng về quy mô và chất lượng này yêu cầu nhóm phải làm việc với khối lượng dữ liệu rất lớn. Theo chia sẻ từ đại diện nhóm nghiên cứu, bộ dữ liệu dòng chảy phục vụ mô hình tự tương quan cho lưu vực sông Đà kéo dài hàng chục năm, được chia thành nhiều kịch bản dự báo khác nhau theo từng điểm dự báo, thời đoạn và nhiều mô hình máy học khác nhau như LSTM, RF, LGBM, ANN, ARIMA, SVR,...

“Trong giai đoạn đầu, khi chưa có máy chủ mạnh, nhóm phải chạy toàn bộ mô hình trên laptop cá nhân. Mỗi kịch bản huấn luyện có thể kéo dài từ vài chục tiếng đến vài ngày, và đã có thời điểm máy tính của tất cả các bạn trong nhóm đều phải hoạt động liên tục suốt một tuần không tắt chỉ để hoàn tất quá trình huấn luyện. Khối lượng chỉ số cần tính toán cũng không hề nhỏ”, Minh Quân, trưởng nhóm cho hay.

Để khắc phục khó khăn này và đảm bảo tính minh bạch, độ tin cậy của kết quả, ở giai đoạn tổng hợp cuối cùng, nhóm đã chạy lại toàn bộ mô hình trên server có cấu hình cao, giúp mọi mô hình được đánh giá trong cùng một điều kiện tính toán. Song song, nhóm cũng tối ưu hóa thuật toán và quy trình huấn luyện, nhằm rút ngắn thời gian xử lý và nâng cao hiệu quả tính toán trong các lần chạy sau.

Dựa trên kết quả phân tích sau khi đã xử lý dữ liệu đầu vào, nhóm đã đánh giá được về khả năng nâng cao hiệu quả dự báo dòng chảy, từ đó đề xuất thành công mô hình dự báo tối ưu cho các điểm quan trắc thượng lưu trên lưu vực sông Đà, phục vụ cho công tác vận hành hồ chứa và quản lý tài nguyên nước trong điều kiện thực tế.

BƯỚC TIẾN TỪ NGHIÊN CỨU ĐẾN ỨNG DỤNG THỰC TIỄN

Với mô hình dự báo tối ưu cho các điểm quan trắc thượng lưu trên lưu vực sông Đà, nhóm đã lựa chọn được những mô hình dự báo dòng chảy trung bình ngày và trung bình 10 ngày tốt nhất và có thể dự báo trước xa nhất. Các phương pháp xử lý dữ liệu như khử nhiễu (DWT, FFT, RWT, RFFT) và bổ sung





đặc trưng thời gian (One Hot Encoder) đều góp phần cải thiện đáng kể hiệu quả dự báo lưu lượng dòng chảy. Cụ thể, các mô hình LGBM_DWT, RF_DWT, LSTM_RWT và LSTM_DWT cho phép dự báo trước 7 ngày với dòng chảy trung bình ngày và trước từ 50 đến 70 ngày với dòng chảy trung bình 10 ngày - vượt trội hơn hẳn so với các mô hình truyền thống, đáp ứng các yêu cầu thực tiễn phục vụ cho quản lý tài nguyên nước và vận hành hệ thống các hồ chứa lớn trên lưu vực sông Đà.

Đây là bước tiến lớn, không chỉ về mặt học thuật mà còn ở khả năng ứng dụng vào dự báo nghiệp vụ thực tế. Các mô hình được xây dựng có thể vận hành độc lập, chỉ cần dữ liệu quan trắc sẵn có trong nước, mà không cần phụ thuộc vào dữ liệu xuyên biên giới - yếu tố thường gây khó khăn cho các lưu vực quốc tế như sông Đà.

Hiện tại, nhóm đang từng bước triển

khai kết quả nghiên cứu vào hệ thống dự báo nghiệp vụ thực tế, phục vụ điều hành liên hồ chứa, phát điện và cảnh báo lũ, hạn trên lưu vực sông Đà. Theo Minh Quân: “Đây là bước tiến quan trọng hướng tới hiện đại hóa công tác dự báo thủy văn bằng trí tuệ nhân tạo ở Việt Nam, đồng thời góp phần giải quyết bài toán dự báo dòng chảy trong điều kiện thiếu dữ liệu, một thách thức lớn của ngành thủy văn hiện nay”.

Đề tài không chỉ dừng lại ở phạm vi nghiên cứu mà còn tạo nền tảng cho chuyển đổi số trong lĩnh vực dự báo khí tượng, thủy văn. Việc áp dụng mô hình học máy giúp: Tăng tốc độ xử lý và khả năng cập nhật dự báo theo thời gian thực; Giảm chi phí và nguồn lực vận hành, nhờ mô hình dữ liệu tự động học và cập nhật; Mở rộng khả năng ứng dụng sang các lưu vực khác có đặc điểm tương đồng, như

sông Mã, sông Cà hoặc hệ thống sông miền Trung - nơi dữ liệu đo đạc còn hạn chế.

Kết quả của nghiên cứu không chỉ mang ý nghĩa học thuật, mà còn có tính ứng dụng cao trong công tác dự báo, vận hành hồ chứa và quản lý tài nguyên nước trong điều kiện thực tế. Quan trọng hơn, kết quả nghiên cứu khẳng định giải pháp trí tuệ nhân tạo hoàn toàn có thể trở thành công cụ hiệu quả để hỗ trợ đưa ra dự báo trong quản lý tài nguyên nước, góp phần đảm bảo an ninh nước, an ninh lương thực và an ninh năng lượng cho khu vực miền Bắc nói riêng và Việt Nam nói chung.