ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

BÁO CÁO BÀI TẬP LỚN

DỰ BÁO LƯỢNG MƯA SỬ DỤNG KẾT HỢP DỮ LIỆU KHÍ TƯỢNG VÀ DỮ LIỆU TRẠM VÀO CÁC MÔ HÌNH HỌC MÁY VÀ HỌC SÂU

Thành viên: Phạm Hoàng An Khánh - 23021597

Trần Gia Khánh - 23021599

Lê Nguyên Anh - 23021465

Vũ Nhật Tường Vân - 23021747

Môn học: Trí tuệ nhân tạo

Lớp học phần: INT3401E 4

Giảng viên hướng dẫn: PGS. TS Nguyễn Thị Nhật Thanh

ThS. Hoàng Gia Anh Đức

HÀ NỘI – 2025

Mục lục

1. Tóm tắt báo cáo

Báo cáo này tập trung vào việc dự báo lượng mưa trong khoảng thời gian ngắn tới, cụ thể là 6 giờ tới, dựa trên các thông số dữ liệu trạm AWS, dữ liệu khí tượng ERA5 trong khoảng thời gian tháng 4 và tháng 10 trong các năm 2019 và 2020. Việc dự đoán sử dụng phương pháp áp dụng các mô hình học máy, học sâu như XGBoost, Lasso, Ridge, LSTM. Việc đánh giá hiệu suất kết quả sẽ dựa trên các thông số MSE, MAE, R^2, Pearson.

2. Bài báo tham khảo

2.1. Tóm tắt bài báo tham khảo

2.1.1. Tổng quan

Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting.

Nghiên cứu này nhằm đánh giá và so sánh hiệu quả của các mô hình học máy hiện đại trong dự báo lượng mưa ngắn hạn. Nghiên cứu sử dụng dữ liệu thời tiết hàng giờ từ 5 thành phố ở Vương quốc Anh trong giai đoạn 2000–2020, bao gồm các đặc trưng như nhiệt độ, độ ẩm, áp suất, tốc độ gió và lượng mưa. Tác giả áp dụng chiến lược dự báo nhiều bước (multi-output forecasting), chuẩn hóa dữ liệu với MinMaxScaler và sử dụng cả mô hình học sâu lẫn ensemble truyền thống để so sánh độ chính xác, từ đó xác định các phương pháp phù hợp cho dự báo mưa theo chuỗi thời gian.

2.1.2. Mục tiêu của bài báo

- So sánh hiệu suất các mô hình học sâu (LSTM và các biến thể) với các mô hình học máy truyền thống (XGBoost, AutoML) trong bài toán hồi quy thời gian (time-series regression) dự báo lượng mưa
- Đưa ra phân tích dựa trên dữ liệu khí tượng thực tế tại 5 thành phố lớn ở Anh qua nhiều giai đoạn khác nhau
- Xác định mô hình phù hợp nhất để áp dụng cho dự báo lượng mưa trong thực tiễn.

2.2. Các mô hình đã được nghiên cứu

2.2.1. LSTM và các biến thể

Dựa trên kiến trúc mạng nơ-ron tái hồi (RNN), bài báo áp dụng 3 biến thể của mô hình LSTM như sau:

- LSTM thông thường: Là mô hình đơn giản, dễ triển khai.
- Stacked LSTM: Là mô hình học sâu giúp học được đặc trưng phức tạp, phù hợp với dữ liệu lớn.
- Bidirectional LSTM: Mô hình học thông tin từ cả chiều thời gian trước và sau, giúp tăng khả năng mô hình hoá bối cảnh.

2.2.2. XGBoost

Mô hình Boosting mạnh mẽ, được sử dụng như baseline truyền thống. Ưu điểm có tốc độ huấn luyện nhanh, hiệu quả cao trong nhiều bài toán dữ liệu dạng bảng.

2.2.3. AutoML (TPOT)

- Sử dụng genetic programming để tự động tìm kiến trúc mô hình tốt nhất (bao gồm Extra Trees, Gradient Boosting, SVR,...)
- Ưu điểm tối ưu toàn bộ quá trình tìm mô hình và tham số.

2.3. Kết quả của bài báo

- Stacked LSTM và LSTM thông thường cho chỉ số RMSE thấp nhất trên hầu hết các thành phố, cho thấy khả năng mô hình hoá chuỗi thời gian tốt hơn XGBoost
- XGBoost hoạt động ổn định nhưng không vượt qua được LSTM khi xét về chỉ số RMSE
- AutoML (TPOT) có kết quả khá tốt, đặc biệt ở các thành phố có biến động phức tạp.

Kết luận

- Mô hình Stacked-LSTM là lựa chọn mạnh mẽ nhất cho bài toán dự báo lượng mưa trong chuỗi thời gian
- XGBoost là baseline tốt nhưng học sâu cho kết quả chính xác hơn với bài toán chuỗi thời gian mưa ngắn hạn
- AutoML là lựa chọn khả thi nếu muốn tiết kiệm công sức xây dựng mô hình thủ công.

2.4. Nhóm đã ứng dụng thế nào vào dự án của mình

Dựa trên bài báo, nhóm kế thừa và áp dụng những kỹ thuật cốt lõi sau:

- **Chiến lược sử dụng sliding window để tạo chuỗi đầu vào**: Cho phép mô hình học được động lực học của thời tiết theo thời gian, ví dụ như xu hướng thay đổi của độ ẩm, tốc độ gió hoặc độ bất ổn khí quyển trước khi xảy ra mưa. Nhóm cũng kế thừa kỹ thuật kiểm tra tính liên tục của chuỗi (các bản ghi cách nhau đúng 1 giờ) nhằm loại bỏ các chuỗi bị thiếu giờ, từ đó đảm bảo độ chính xác và ổn định của dữ liệu đầu vào.
- **Xây dựng các mô hình XGBoost và Stack-LSTM**: Sau khi so sánh kết quả thu được từ các mô hình trong bài báo, nhóm chọn lọc những mô hình ưu tú nhất như Stack-LSTM và XGBoost cho thực nghiệm của mình và áp dụng những cải tiến của riêng nhóm để tạo kết quả dự đoán tốt hơn.

3. Dữ liệu

3.1. Mô tả dữ liệu đầu vào

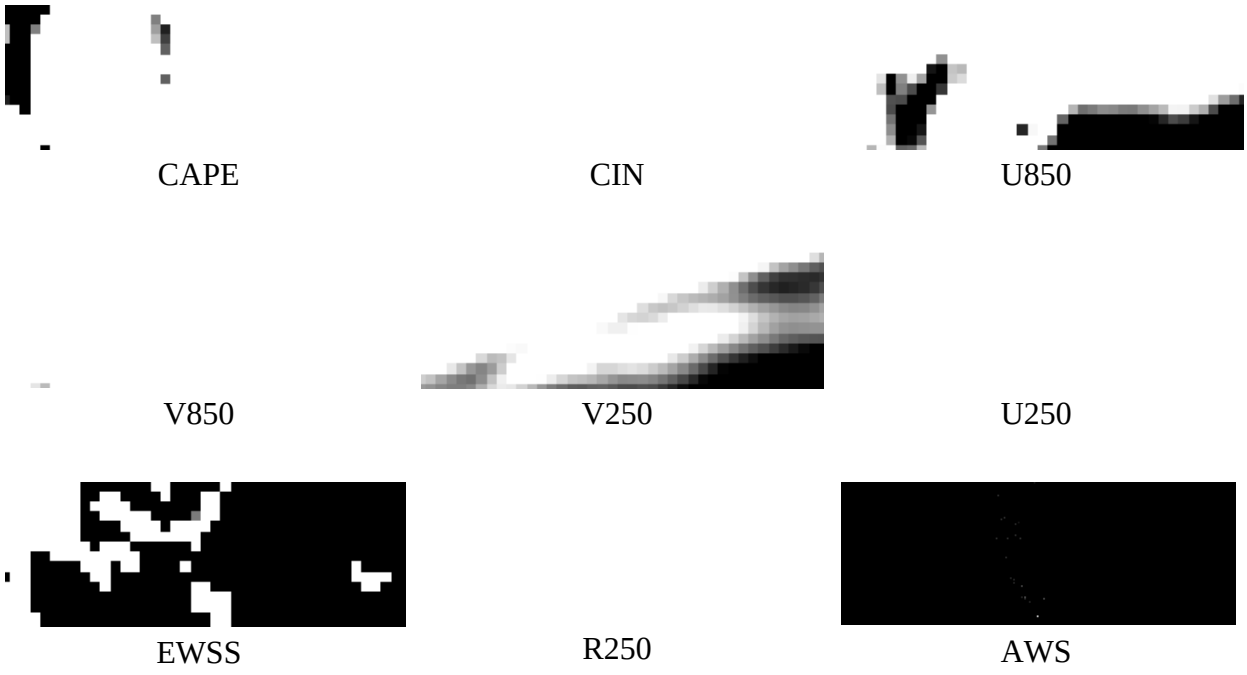
- Nguồn: dữ liệu: Các trạm đo thực tế đặt tại các khu vực quan sát
- Định dạng: GeoTIFF (.tif)
- Kích thước ảnh: 90x250
- Độ phân giải thời gian: 1 giờ
- Độ phân giải không gian: 4KM
- Khoảng thời gian: 4/2019, 10/2019, 4/2020, 10/2020
- Thông tin đo lường: Lượng mưa thực tế được ghi nhận theo thời gian tại từng vị trí trạm đo.

3.2. Xử lý dữ liệu đầu vào, biến đổi thành dữ liệu dạng bảng

3.2.1. Thu thập dữ liệu

- **Dữ liệu khí tượng (ERA5)**: Các biến khí tượng quan trọng như CAPE & CIN (khu vực có tiềm năng đối lưu mạnh), EWSS & KX (gió), R250 - R500 - R850 (biểu đồ khí tượng ở các tầng khí quyển khác nhau),... được sử dụng để mô hình hóa các yếu tố khí hậu có thể ảnh hưởng đến mưa.
- **Dữ liệu từ trạm AWS (Precipitation)**: Cung cấp các chỉ số về nhiệt độ, áp suất, độ ẩm và các chỉ số khí tượng khác tại các vị trí trạm quan trắc

- Những điểm ảnh giá trị >=0: Những vị trí đặt trạm và có bản ghi lượng mưa
- Những điểm trắng (-inf, nan hoặc -9999): Những vị trí không có dữ liệu.



Minh hoạ dữ liệu thô của các chỉ số khác nhau tại thời điểm 00:00:00 01/04/2020

3.2.2. Xử lý dữ liệu thô

- Xác định các tọa độ hợp lệ
- Duyệt qua từng file dữ liệu AWS theo năm, tháng, ngày, giờ
- Kiểm tra sự tồn tại của file GeoTIFF tương ứng
- Đọc dữ liệu từ file GeoTIFF (sử dụng thư viện rioxarray)
- Duyệt qua từng pixel trong ảnh: Mục tiêu là loại bỏ các toạ độ lỗi/thiếu (giá trị -inf) trong dữ liệu AWS
- Lưu danh sách tọa độ hợp lệ.
- Trích xuất dữ liệu từ file GeoTIFF
- Xác định đường dẫn file dữ liệu theo thời gian, vị trí và loại dữ liệu (AWS hay ERA5)
- Kiểm tra sự tồn tại của file tương ứng

- Mở file và trích xuất giá trị dữ liệu tại từng toạ độ hợp lệ
- Duyệt qua từng pixel: Mục tiêu là loại bỏ các toạ độ có giá trị -inf, inf và nan
- Tích hợp dữ liệu theo thứ tự thời gian vào danh sách.

• Tổng hợp dữ liệu thành bảng

- Duyệt qua danh sách tọa độ hợp lệ
- Gộp dữ liệu từ các thời gian 4/2019, 10/2019, 4/2020, 10/2020 vào 1 danh sách chung
- Chuyển thành DataFrame
- Lưu thành file CSV.

3.2.3. Dữ liệu đầu ra của bước

row	col	datetime	AWS	CAPE	CIN	EWSS	IE	ISOR	KX
0	114	2019-04-01 00:00:00	0.000000	446.125000	38.875000	-641.113000	-0.000014	0.588248	37.360200
0	114	2019-04-01 02:00:00	9.200000	519.750000	73.312900	-201.391000	-0.000037	0.588248	36.553200
0	114	2019-04-01 03:00:00	7.400000	447.500000	112.094000	-80.894500	-0.000043	0.588248	34.931900
0	114	2019-04-01 04:00:00	0.000000	421.500000	158.594000	-33.918000	-0.000043	0.588248	32.517400
0	114	2019-04-01 05:00:00	0.000000	435.625000	98.828200	11.640600	-0.000039	0.588248	31.718700

Minh hoạ dữ liệu đầu vào đã biến đổi thành dạng bảng

• 1 bảng dữ liệu gồm các cột sau

- row và column: Toạ độ tương ứng trong lưới ảnh
- datetime: Thời gian của dữ liệu
- Các cột AWS, CAPE, CIN,… : Các giá trị khí tượng từ ERA5 và lượng mưa trạm từ AWS tương ứng

• Thống kê dữ liệu:

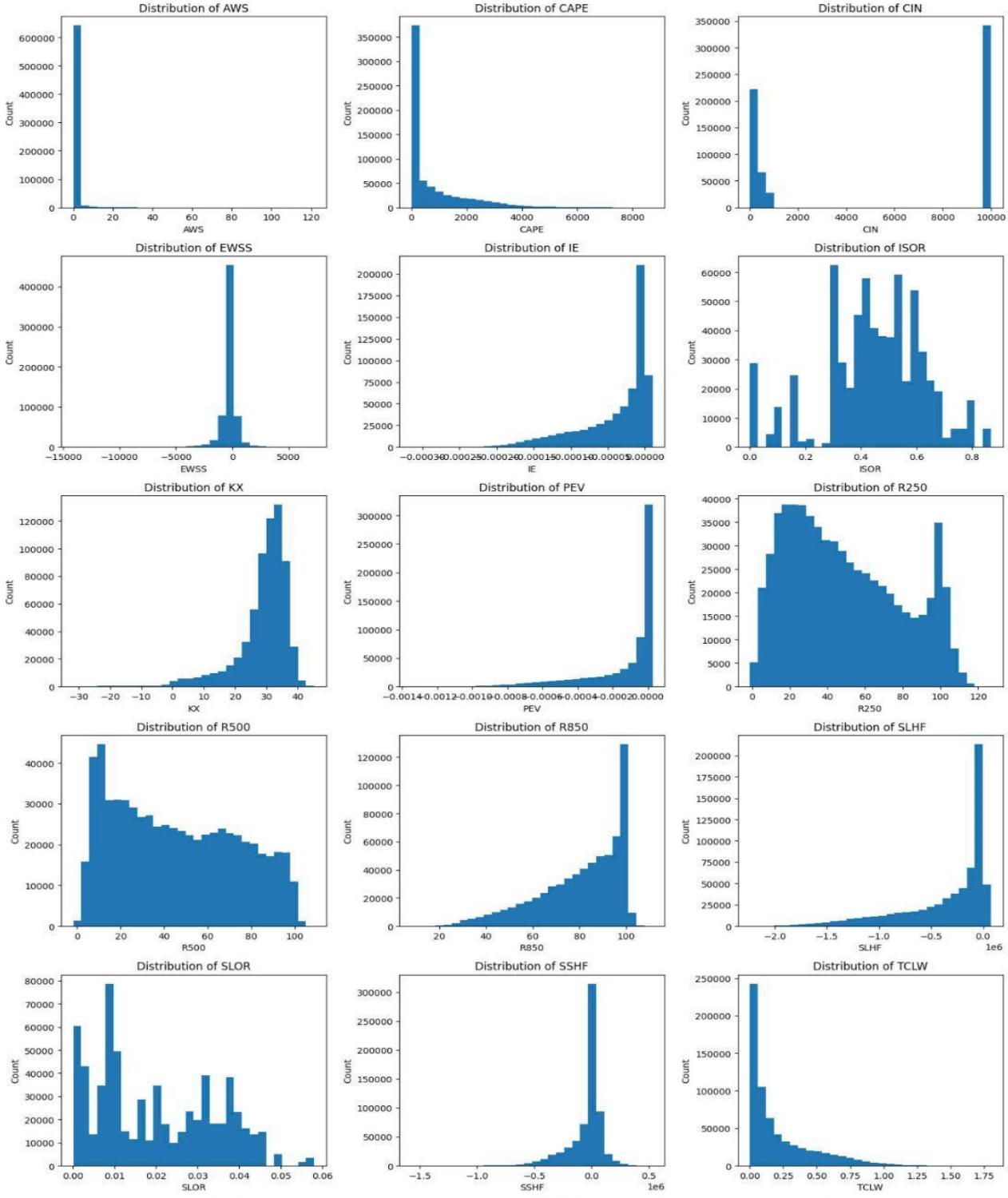
- Tổng số mẫu: 657681
- Số cột: 23

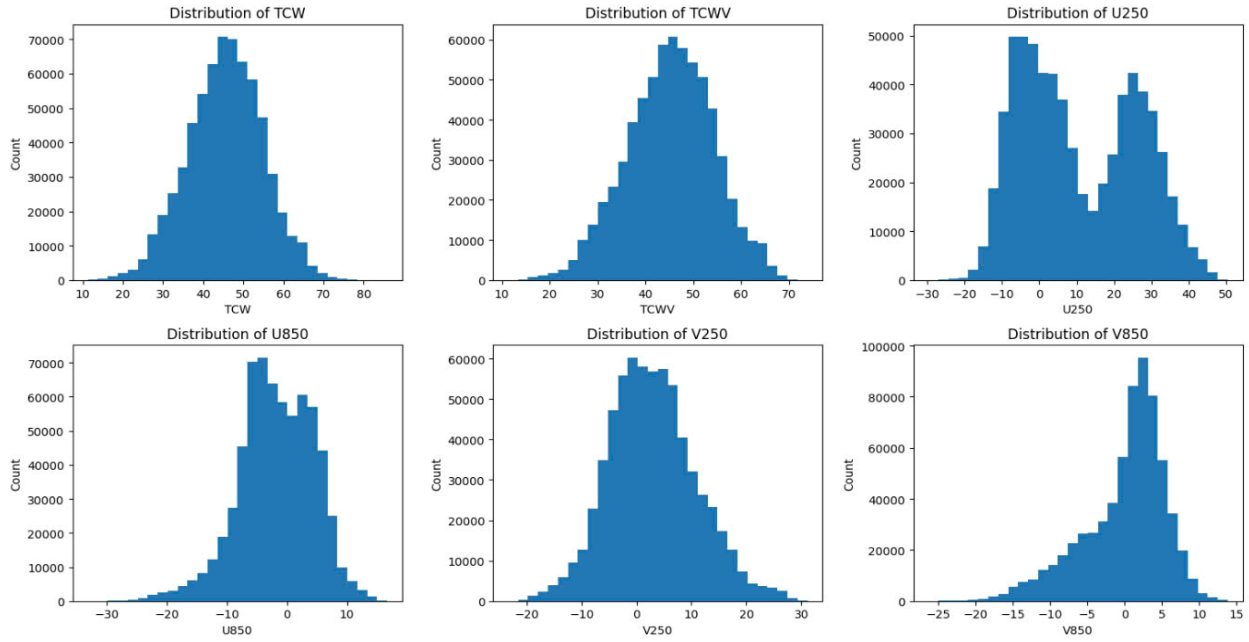
3.3. Phân tích sơ bộ

Trước khi xây dựng mô hình, nhóm thực hiện quan sát dữ liệu nhằm hiểu rõ cấu trúc và đặc điểm của tập dữ liệu. Các kỹ thuật trực quan hoá như biểu đồ phân phối, heatmap

tương quan và biểu đồ hộp (boxplot) được sử dụng để phát hiện xu hướng, giá trị ngoại lai và mối quan hệ giữa các biến.

Trước tiên nhóm sử dụng biểu đồ phân phối để kiểm tra độ lệch của các biến như sau:

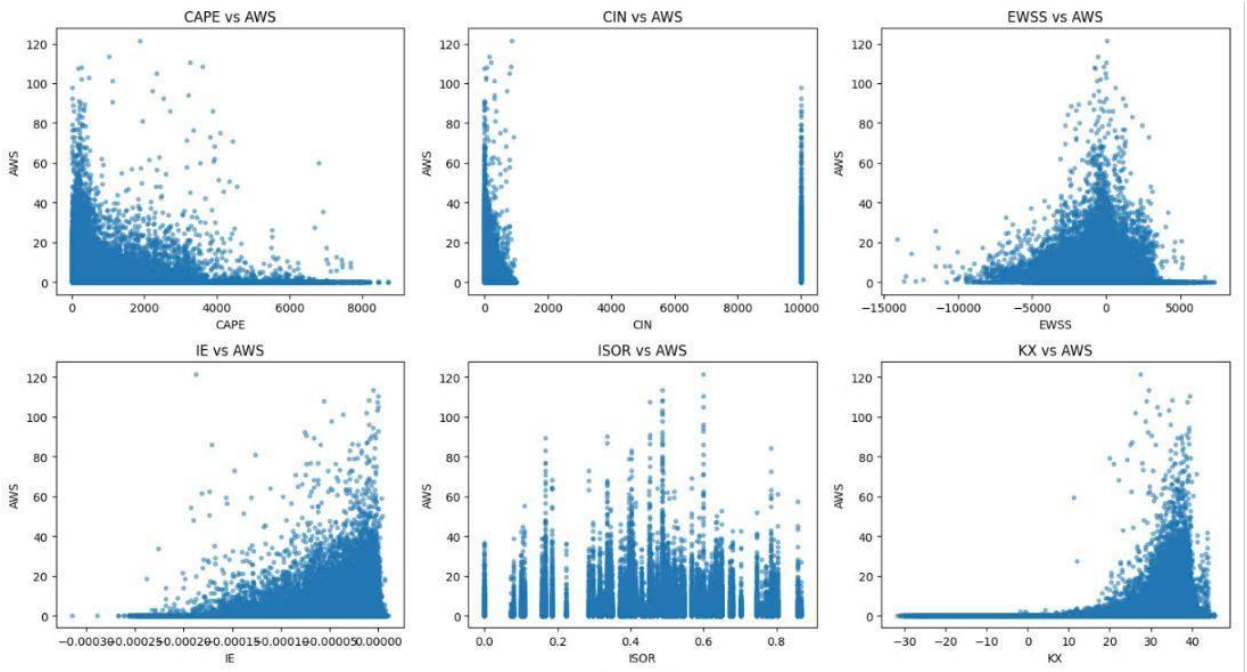


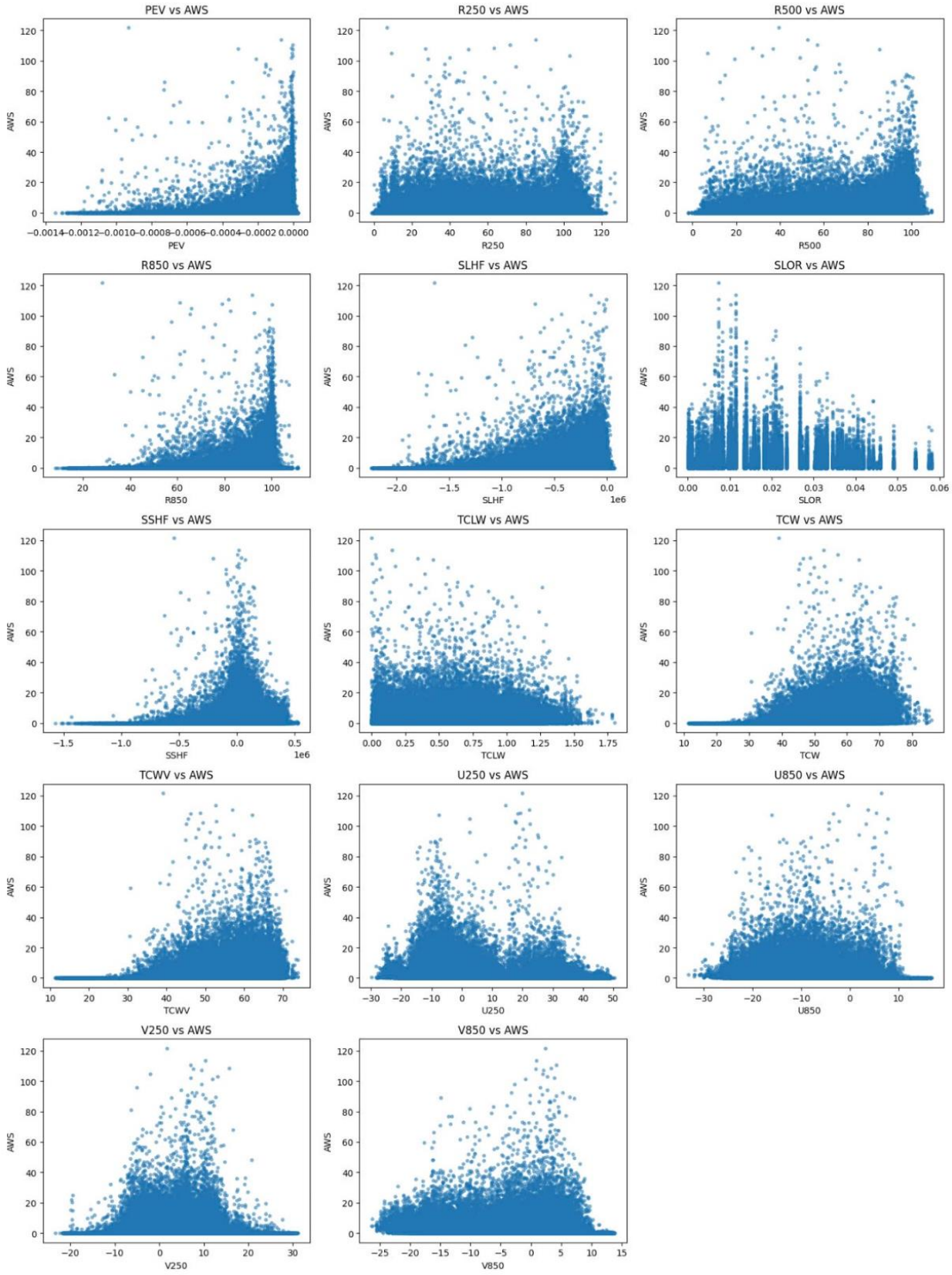


Phân bố của các chỉ số trong tập dữ liệu

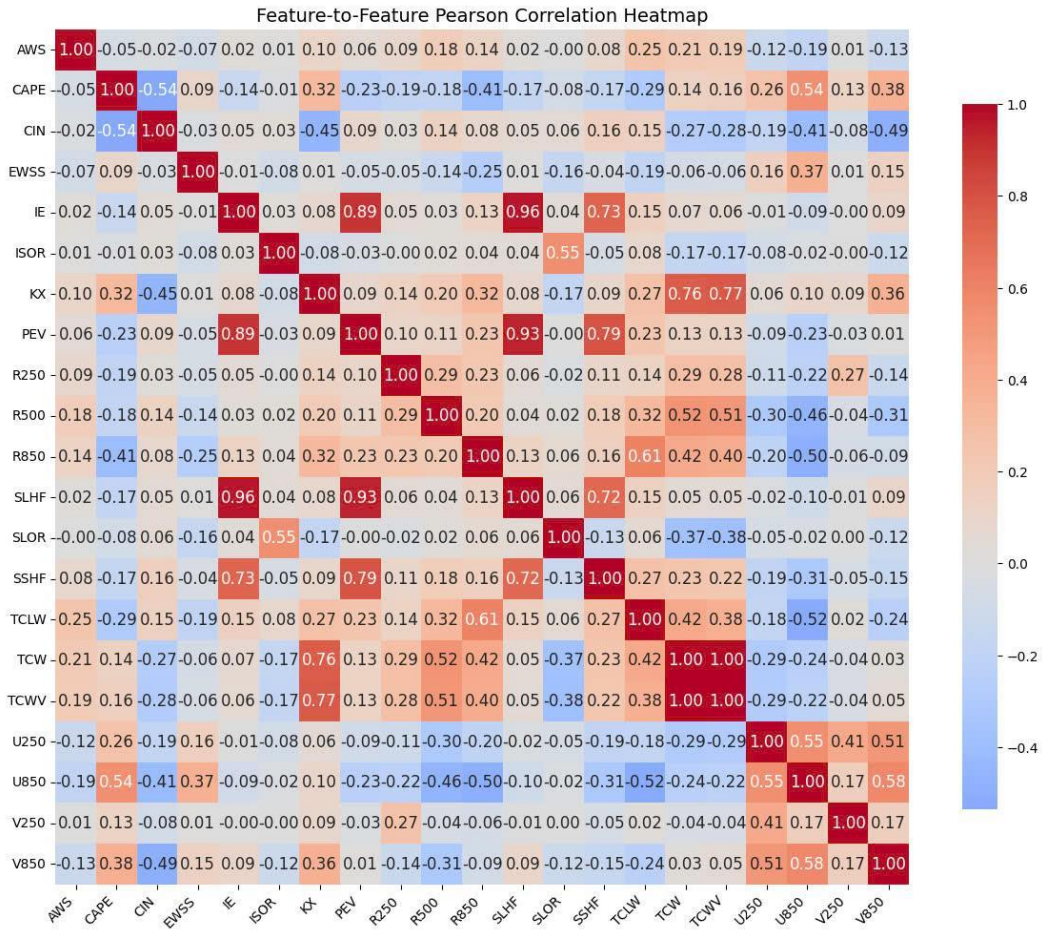
Có thể thấy một số biến như CIN có nhiều giá trị ngoại lai; biến mục tiêu/dữ liệu mưa có phân phối lệch phải, tập trung chủ yếu ở mức mưa nhỏ (<10mm). Điều này có thể gây mất cân bằng đầu ra, vì vậy khi chuẩn bị dữ liệu và xây dựng mô hình nhóm sẽ thực hiện loại bỏ giá trị ngoại lai và chuẩn hoá.

Tiếp theo, nhóm quan sát kĩ tương quan Pearson giữa biến mục tiêu AWS và các biến khác thông qua biểu đồ mật độ và heatmap như sau:





Biểu đồ tương quan của các chỉ số khí tượng với chỉ số lượng mưa AWS



Ma trận tương quan giữa các chỉ số

3.4. Trích rút đặc trưng

3.4.1. Tạo các đặc trưng mới

Từ cột datetime có nội dung dạng yyyy-mm-dd hh:mm:ss, nhóm tạo thêm 4 đặc trưng mới thông qua biến đổi các biến tuần hoàn hour_sin, hour_cos, doy_sin, doy_cos. Mục đích giúp mô hình hiểu chu kỳ giờ trong ngày và ngày trong năm.

3.4.2. Tiền xử lý dữ liệu

- **Loại bỏ giá trị ngoại lai**: Trước khi huấn luyện mô hình, dữ liệu được làm sạch bằng cách loại bỏ các mẫu có giá trị ngoại lai sử dụng **tiêu chuẩn thống kê z-score**. Cụ thể, các mẫu có giá trị lớn hơn |3σ| so với trung bình (tức z-score > 3 hoặc < -3) được coi là bất thường và bị loại bỏ.

Đây là phương pháp hiệu quả để xử lý ngoại lệ trong các đặc trưng có phân phối gần chuẩn, phù hợp với các mô hình như LSTM hoặc hồi quy tuyến tính khi các dữ liệu ngoại lai (dễ gặp ở lượng mưa) có thể gây ảnh hưởng mạnh đến kết quả huấn luyện.

- **Chuẩn hóa dữ liệu**: Toàn bộ dữ liệu huấn luyện được chuẩn hoá bằng **StandardScaler**, đưa các đặc trưng về phân phối chuẩn với trung bình = 0 và độ lệch chuẩn = 1.

Phương pháp này phù hợp với bộ dữ liệu đa biến và không đồng thang đo. Việc chuẩn hoá có vai trò đặc biệt quan trọng với mô hình LSTM, vốn yêu cầu đầu vào có phân phối đồng đều và ổn định để giảm thời gian hội tụ và tránh bùng nổ. Việc sử dụng StandardScaler giúp giữ tính nhất quán giữa các mô hình khi huấn luyện nhiều thuật toán khác nhau trong cùng pipeline.

Ngoài ra nhóm lưu ý việc chuẩn hóa chỉ thực hiện trên tập huấn luyện, sau đó áp dụng thống kê lên tập kiểm tra nhằm tránh hiện tượng rò rỉ dữ liệu từ tương lai về quá khứ.

- **Sinh chuỗi thời gian với sliding window**: Dữ liệu được chuyển đổi từ định dạng bảng sang chuỗi đầu vào bằng cách chia theo sliding window. Với mỗi chuỗi thời gian (theo từng trạm), thực hiện tạo ra:
    - Input: window_size = 1: 1 giờ gần nhất gồm các đặc trưng đầu vào
    - Output: horizon = 6 : dự đoán lượng mưa AWS liên tiếp cho 6 giờ tiếp theo.

Đồng thời nhóm kiểm tra rằng toàn bộ chuỗi có đúng khoảng cách 1 giờ giữa các bản ghi.

Sinh chuỗi thời gian là bước tiền xử lý quan trọng với mô hình LSTM, bốn yêu cầu đầu vào định dạng (samples, timesteps, features). Bước này giúp mô hình học được ngữ cảnh thời gian gần, phản ánh tính chất “tích lũy” hoặc “chuyển tiếp” của các biến khí tượng.

Sau khi thực hiện tiền xử lý, nhóm thu về dữ liệu như sau:

	row	col	datetime	AWS	CAPE	CIN	EWSS	IE	ISOR	KX	...	V250	V850
0	0	104	2020-10-16 00:00:00	1.4	-0.798054	1.164571	-0.195802	-0.000008	0.676296	0.716597	...	0.314106	-0.515163
1	0	104	2020-10-16 01:00:00	1.4	-0.792639	1.164571	-0.227186	-0.000014	0.676296	0.543833	...	0.217560	-0.643525
2	0	104	2020-10-16 02:00:00	0.0	-0.782601	1.164571	-0.306841	-0.000025	0.676296	0.403922	...	0.206267	-0.916225
3	0	104	2020-10-16 03:00:00	0.0	-0.770054	1.164571	-0.407982	-0.000034	0.676296	0.331123	...	0.294815	-1.219649
4	0	104	2020-10-16 04:00:00	0.0	-0.773224	1.164571	-0.338257	-0.000047	0.676296	0.260102	...	0.369559	-1.366925

Minh hoạ dữ liệu đầu vào đã tiền xử lý

3.4.3. Chọn lọc các biến cần thiết

Để có các đặc trưng phù hợp nhất, nhóm đưa dữ liệu vào mô hình Random Forest và sử dụng tính năng tính độ quan trọng của các biến (feature importances) của mô hình này. Sau khi quan sát các giá trị mô hình trả về, nhóm đưa ra lựa chọn như sau:

- Loại bỏ các cột row, column và datetime vì chúng không ảnh hưởng tới việc huấn luyện mô hình.
- Các biến AWS, CAPE, CIN, V850, KX, R250, V250, U250, U850, EWSS được giữ lại làm dữ liệu cho mô hình học máy. Trong đó AWS là biến mục tiêu, các biến còn lại là dữ liệu đầu vào.

Đặc trưng được chọn	**Mô tả**
CAPE	Chỉ số đo lường năng lượng có sẵn để hỗ trợ sự phát triển của các cơn bão đối lưu
CIN	Chỉ số thể hiện mức độ "cản trở" sự bốc lên của không khí

V850 & U850	Chỉ số đo lường thành phần gió ở các hướng khác nhau tại tầng 850 hPa, hỗ trợ phân tích sự vận chuyển độ ẩm và sự hội tụ không khí
U250	Chỉ số đo lường thành phần gió hướng Đông - Tây ở tầng 250 hPa, hỗ trợ đối lưu và hình thành mưa
V250	Thành phần gió hướng Bắc–Nam ở tầng cao (250 hPa), góp phần vào quá trình phân kỳ trên cao và hỗ trợ hình thành đối lưu
R250	Các chỉ số đo lường độ ẩm ở tầng khí quyển 250 hPa
KX	Chỉ số phát triển khí quyển dựa trên độ ẩm và nhiệt độ ở các tầng trung lưu
EWSS	Biến đánh giá sự thay đổi tốc độ gió theo chiều dọc

Bảng mô tả các đặc trưng được sử dụng

4. Phương pháp

4.1. Phân chia dữ liệu

- **Huấn luyện**: Dữ liệu từ 00:00:00 01/04/2019 đến 23:00:00 15/10/2020
- **Kiểm tra**: Dữ liệu từ 00:00:00 16/10/2020.

4.2. Lựa chọn mô hình

Dựa trên đặc trưng của bài toán, của dữ liệu, kết hợp kết quả của bài báo đã tham khảo, nhóm thử nghiệm trên 5 mô hình sau:

- **Lasso**: Lasso là mô hình thực hiện chọn lọc đặc trưng tự động bằng cách đưa hệ số một số biến về 0 (nhờ regularization L1), giúp giảm overfitting, dễ diễn giải. Lasso phù hợp với dữ liệu có số lượng đặc trưng nhiều và có nhiễu.
- **Ridge**: Ridge cũng là 1 mô hình chọn lọc đặc trưng bằng cách làm nhỏ hệ số (nhờ regularization L2) thay vì loại bỏ hẳn như Lasso, học tốt trong trường hợp có

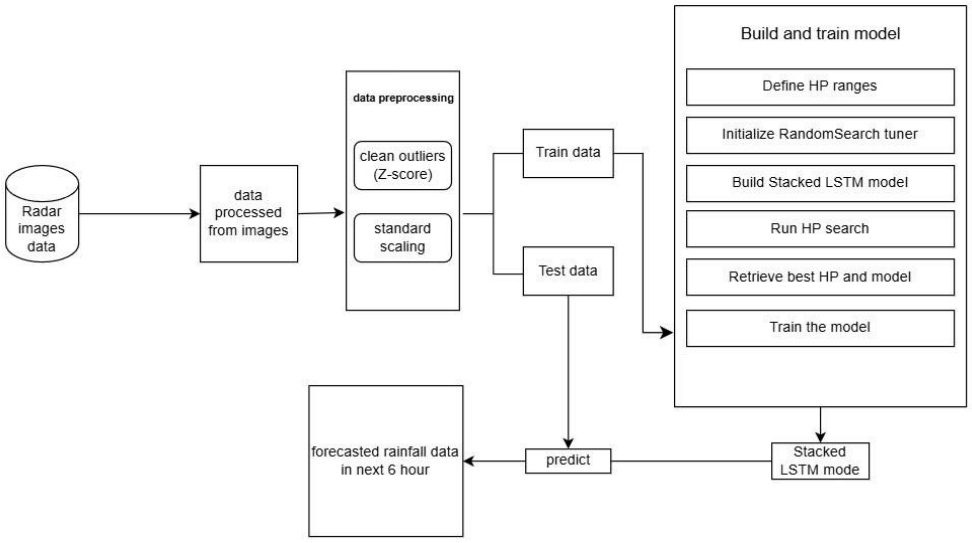
tương quan giữa các đặc trưng. Ridge phù hợp với dữ liệu có nhiều biến đa cộng tuyến.

- **Random Forest**: Random Forest là mô hình phi tuyến tính có khả năng tự động xử lý mối quan hệ phức tạp giữa các biến, ít nhạy cảm với ngoại lai và không yêu cầu nhiều tiền xử lý. Random Forest phù hợp với dữ liệu chứa nhiễu, nhiều tương tác.
- **XGBoost**: XGBoost là mô hình được sử dụng phổ biến trong nghiên cứu về dự báo khí tượng, xử lý tốt quan hệ phi tuyến giữa các biến và cho phép tinh chỉnh sâu. Mô hình này có hiệu suất cao, xử lý giá trị thiếu tốt, tránh overfitting nhờ regularization.
- **Stacked-LSTM**: Đây là mô hình mạng nơ-ron hồi tiếp sâu (RNN), phù hợp với xử lý dữ liệu dạng chuỗi thời gian liên tục nhờ khả năng ghi nhớ thông tin dài hạn.

Trong bài toán này, nhóm xây dựng Stacked-LSTM với cấu trúc như sau:

- LSTM lớp 1: học các đặc trưng thô từ dữ liệu đầu vào
- LSTM lớp 2: học các đặc trưng phức tạp hơn dựa trên đầu ra đã làm sạch qua Dropout và BatchNormalization của lớp trước
- Các lớp Dropout và BatchNormalization giúp mô hình học ổn định và giảm overfitting
    + Mỗi lớp Dropout ngăn mạng "nhớ" quá kĩ các đặc trưng huấn luyện
    + BatchNormalization giúp ổn định phân phối đầu vào cho lớp sau, làm cho các bước train không quá dao động do dropout hay learning_rate
- Output layer: Dense, linear activation giúp dự báo lượng mưa liên tục

Với 2 lớp LSTM, mô hình không chỉ học mối quan hệ temporal đơn giản mà còn có khả năng phát hiện các pattern phức tạp ở mức độ sâu (ví dụ: pattern ngắn hạn 1 giờ - 2 giờ - 3 giờ đến pattern dài hạn 1 giờ - 6 giờ).



Mô hình dự đoán của Stacked-LSTM

4.3. Tối ưu tham số

4.3.1. Phương pháp RandomizedSearchCV

Trong số các phương pháp tinh chỉnh tham số, nhóm lựa chọn RandomizedSearchCV vì đây là phương pháp cân bằng giữa hiệu quả tìm kiếm và chi phí tính toán, đặc biệt phù hợp với các mô hình có nhiều tham số hoặc thời gian huấn luyện dài như XGBoost và LSTM

- **Hiệu quả cao, tiết kiệm tài nguyên**: RandomizedSearchCV chỉ thử ngẫu nhiên 1 số tổ hợp thay vì thử mọi tổ hợp như GridSearchCV, giúp giảm mạnh chi phí tính toán và vẫn đảm bảo tìm được tham số tốt.
- **Phù hợp với mô hình phức tạp và dữ liệu lớn**: Các mô hình như Random Forest, XGBoost hoặc mạng nơ-ron như LSTM đều là các mô hình có nhiều siêu tham số, mỗi lần huấn luyện đều tốn thời gian, vì vậy RandomizedSearchCV là lựa chọn hợp lý.
- **Tính linh hoạt**: RandomizedSearchCV cho phép sampling từ phân bố liên tục (dropout, learning_rate,...) thay vì từ danh sách rời rạc.

4.3.2. Các siêu tham số được sử dụng và tinh chỉnh ở các mô hình dự đoán

Để đánh giá thực nghiệm của các mô hình, nhóm đã áp dụng 2 bộ tham số chính: Bộ siêu tham số mặc định và Bộ siêu tham số được tối ưu bằng RandomizedSearchCV

Mô hình	Bộ siêu tham số và giá trị được tối ưu bằng RandomizedSearchCV
Lasso/Ridge (chỉ áp dụng tham số mặc định)	alpha=1.0
Random Forest	'max_depth': [7, 8, 9, 10, 11, 12] 'n_estimators': [75, 100, 125, 150, 175, 200] 'min_samples_leaf': [1, 2, 3, 4, 5] 'bootstrap': [True]
XGBoost	'learning_rate': [0.01,0.03,0.05,0.07], 'max_depth': [14, 16, 18, 20, 22, 24, 26, 28, 30], 'n_estimators': [100, 125, 150, 175, 200], 'subsample': [0.6,0.7, 0.8, 0.9, 1.0], 'colsample_bytree': [0.6,0.7, 0.8,0.9, 1.0], 'gamma': [0, 0.1, 0.2, 0.3, 0.4], 'min_child_weight': [1, 2, 3, 4, 5, 6, 7],
Stacked-LSTM	units1=hp.Choice('units1', [64, 128, 256]), units2=hp.Choice('units2', [32, 64, 128]), dropout1=hp.Float('dropout1', 0.1, 0.5, step=0.1), dropout2=hp.Float('dropout2', 0.1, 0.5, step=0.1), lr=hp.Choice('lr', [1e-2, 1e-3, 1e-4])

Bảng siêu tham số

- Đối với 2 mô hình Lasso và Ridge, nhóm không tiến hành tinh chỉnh tham số để tập trung nguồn lực cho các mô hình phức tạp hơn.

- Đối với 2 mô hình Random Forest và XGBoost, nhóm áp dụng 2 lần tham số, với Bộ tham số mặc định nhằm thiết lập baseline và Bộ tham số được tối ưu để cải thiện độ chính xác, kiểm soát overfitting.
- Đối với mô hình Stacked-LSTM, nhóm áp dụng Bộ tham số được tối ưu ngay thông qua Keras Tuner (RandomSearch) với max_trials=5 để tìm tổ hợp tốt nhất với chi phí hợp lý. LSTM là mô hình nhạy cảm vì vậy tối ưu siêu tham số là điều cần thiết.

4.4. Đánh giá mô hình

- **Sai số bình phương trung bình (MSE - Mean Squared Error)**: Đo lường mức độ sai lệch trung bình giữa giá trị thực tế và giá trị dự báo. MSE càng nhỏ thì mô hình dự báo càng chính xác.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **Sai số tuyệt đối trung bình (MAE - Mean Absolute Error)**: Đo trung bình sai số tuyệt đối giữa giá trị thực và dự báo, không bị ảnh hưởng bởi outlier. MAE càng nhỏ thì mô hình càng tốt.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- **Hệ số xác định (R^2)**: Đánh giá mức độ mô hình giải thích phương sai của dữ liệu thực tế. Giá trị R^2 càng cao thì mô hình càng phù hợp.

$$R2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

- **Hệ số tương quan Pearson (Pearson's R)**: Đo lường mức độ tương quan tuyến tính giữa hai biến số. Giá trị R nằm trong khoảng từ -1 đến 1, trong đó R gần 1 thể hiện mối quan hệ tuyến tính dương mạnh, gần -1 là tuyến tính âm mạnh, và gần 0 là không có tương quan.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2 \sum_{i=1}^{n}(y_i - \bar{y}_i)^2}}$$

5. Thực nghiệm

5.1. Kết quả chỉ số đánh giá của các mô hình trên tập huấn luyện

Các mô hình dự đoán		MSE	MAE	R^2	Pearson
Lasso		0.3522	0.247	0.2155	0.5728
Ridge		0.3394	0.2594	0.2439	0.5711
Random Forest	Mặc định	0.3005	0.2339	0.3298	0.5791
	Đã tối ưu	0.2971	0.2297	0.3374	0.5862
XGBoost	Mặc định	0.3803	0.289	0.1508	0.4642
	Đã tối ưu	0.3486	0.2386	0.2239	0.5061
Stacked-LSTM	**Đã tối ưu**	**0.2967**	**0.2373**	**0.3385**	**0.6077**

Bảng so sánh kết quả giữa các mô hình

5.2. Nhận định và lựa chọn mô hình

Mô hình Lasso

- Điểm mạnh: Chỉ số đánh giá MSE và MAE ở mức ổn
- Điểm yếu: Chỉ số R^2 thấp, cho thấy mô hình chỉ giải thích được ~21% phương sai dữ liệu thực tế. Độ tương quan Pearson ở mức trung bình, cho thấy mô hình chưa nắm bắt tốt mối quan hệ giữa giá trị thực và dự đoán.

Mô hình Ridge

- Điểm mạnh: Hiệu suất tốt hơn Lasso về mọi chỉ số đánh giá
- Điểm yếu: Chỉ số R^2 còn thấp, thể hiện khả năng giải thích hạn chế.

Mô hình Random Forest

- Điểm mạnh: Chỉ số R^2 mặc định tốt nhất trong các mô hình phi tuyến tính sử dụng cây, độ tương quan Pearson cao, MAE thấp, cho thấy đây là mô hình ổn định

- Điểm yếu: So với mô hình Stacked-LSTM, mô hình này vẫn chưa đạt hiệu suất cao nhất trong các mô hình thử nghiệm.

Mô hình XGBoost

- Điểm mạnh: Sau khi tinh chỉnh đã cho thấy cải thiện rõ rệt với chỉ số R^2 và độ tương quan Pearson tăng
- Điểm yếu: Kết quả vẫn kém hơn Random Forest và Stacked-LSTM.

Mô hình Stacked-LSTM

- Điểm mạnh: Hiệu suất cao nhất trong các mô hình đơn lẻ
- Điểm yếu: Yêu cầu chuẩn hóa dữ liệu, định dạng theo chuỗi, huấn luyện lâu và khó kiểm soát.

Kết luận

- Stacked-LSTM được chọn làm mô hình chính thức do đạt hiệu suất dự báo tốt nhất trên toàn bộ tập chỉ số. Với cấu trúc nhiều lớp LSTM xếp chồng và khả năng khai thác đặc điểm chuỗi thời gian của dữ liệu, mô hình này giúp nhận diện các xu hướng khí tượng phức tạp và thay đổi nhanh trong thời gian thực.
- Random Forest (với Bộ tham số tối ưu) là mô hình có hiệu suất tốt thứ 2, vẫn cho thấy khả năng khai thác mối quan hệ phi tuyến giữa các đặc trưng đầu vào và lượng mưa. Đây cũng là lựa chọn tốt nếu ưu tiên tính ổn định, giải thích mô hình và thời gian huấn luyện ngắn.

6. Dự đoán và xây dựng bản đồ phân bố mưa

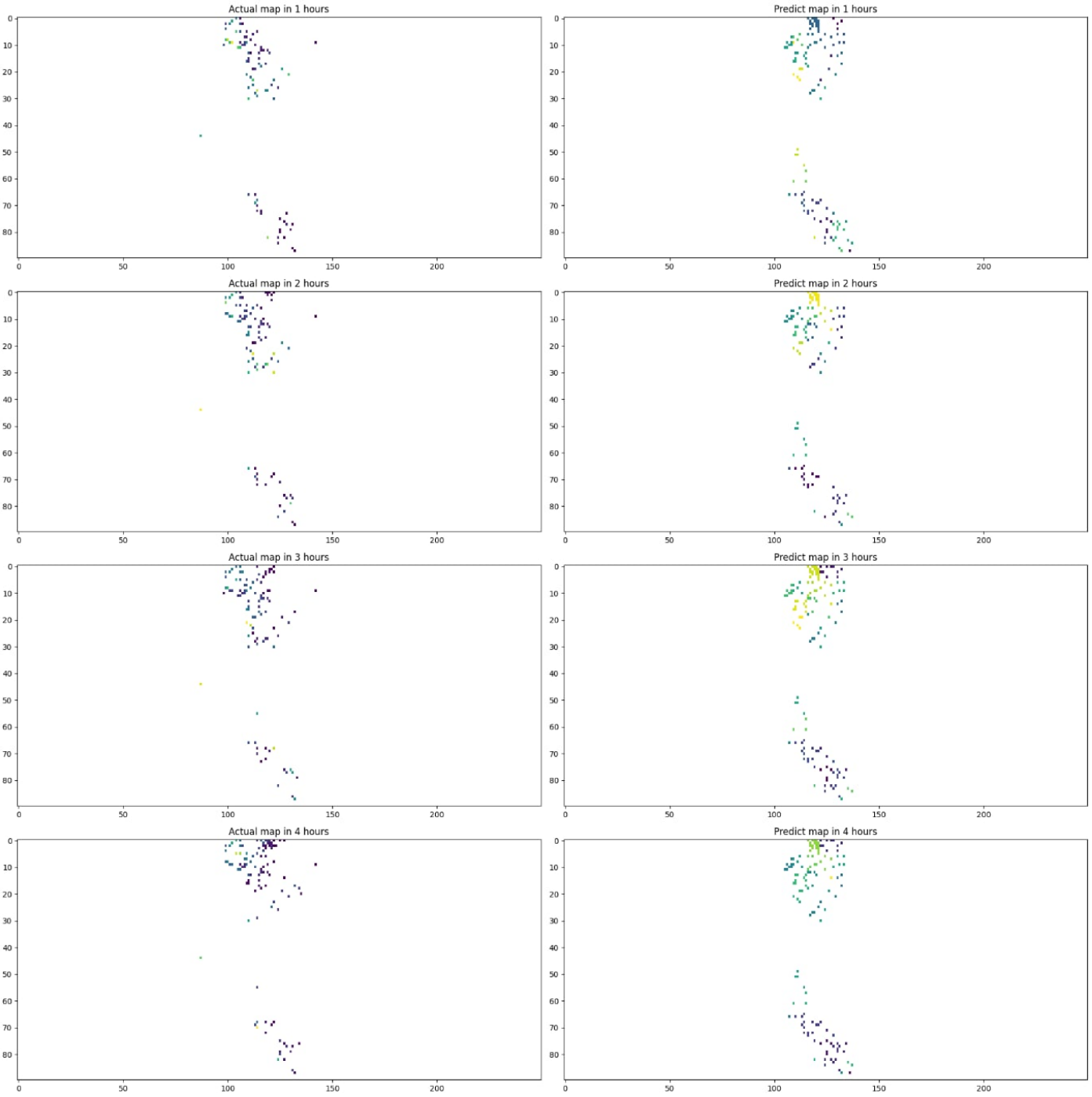
Nhằm đánh giá khả năng của mô hình trong việc dự báo phân bố không gian của lượng mưa theo thời gian, nhóm xây dựng bản đồ lượng mưa dự báo và thực tế tại các điểm trạm (grid) từ 1 đến 6 giờ tiếp theo.

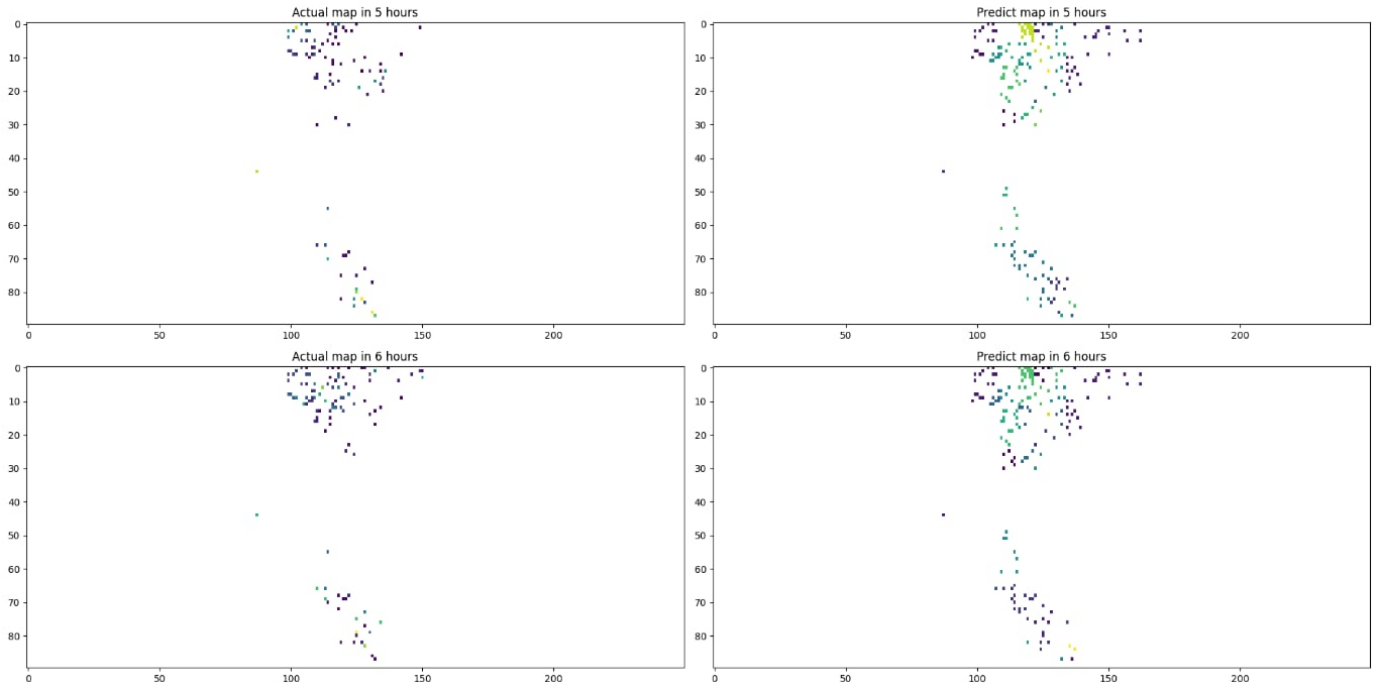
Để minh hoạ, nhóm lựa chọn mốc thời gian 15:00:00 29-10-2020 làm thời điểm dự báo, dữ liệu đầu vào là các giờ sau (tạo chuỗi sliding window).

Sau khi tạo danh sách các điểm trạm (lưới dự đoán) và chuẩn bị mảng chứa kết quả, nhóm thực hiện dự đoán theo từng điểm bằng cách trích xuất dữ liệu lịch sử, dự đoán

và ghi lại kết quả vào mảng. Nhóm cũng thực hiện so sánh với dữ liệu thực tế ứng với thời gian dự đoán.

Sau khi hoàn thành dự đoán, nhóm trực quan hoá kết quả như sau:





Mô phỏng dự đoán lượng mưa

Các trục X và Y đại diện cho tọa độ không gian 2 chiều. Mỗi điểm trên bản đồ đại diện cho 1 trạm AWS ở vị trí không gian cụ thể. Các màu sắc điểm đại diện cho các lượng mưa theo mức độ khác nhau: màu sắc càng sáng (ngả vàng) thể hiện lượng mưa càng lớn, còn màu sắc càng tối (ngả tím đậm) thể hiện lượng mưa thấp hoặc không có mưa.