

ĐỒ ÁN CHUYÊN NGÀNH 1

# DỰ BÁO THỜI TIẾT



# THÀNH VIÊN

1

TRÂN HÀ DUY

22IT054

2

HỒ ĐỨC ĐOAN

22IT065

3

ĐẶNG HỒNG  
NGUYÊN

22IT188

# NỘI DUNG

1

XỬ LÝ DỮ  
LIỆU

2

Quá trình  
huấn luyện

3

Kiểm Tra  
và Đánh  
Giá

# XỬ LÝ DỮ LIỆU

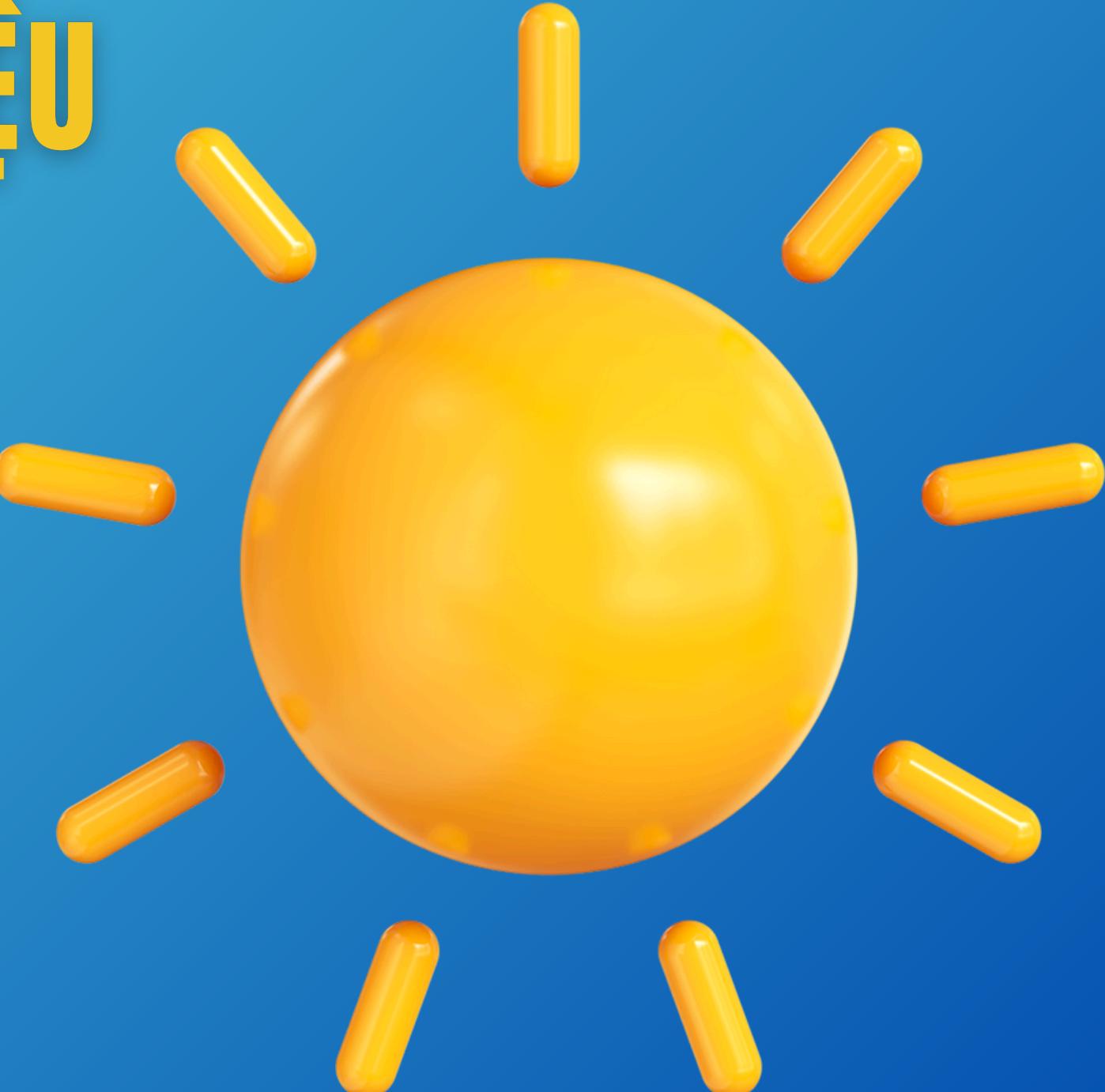


# 1. THU THẬP DỮ LIỆU

- Nguồn: NASA POWER API
- Phạm vi: Việt Nam ( $8^{\circ}$ - $25^{\circ}$  vĩ độ,  $102^{\circ}$ - $119^{\circ}$  kinh độ)
- Thời gian: 01/01/2022 - 31/07/2024
- Tham số: T2M, QV2M, PS, WS10M, PRECTOTCORR, CLRSKY\_SFC\_SW\_DWN

## 2. GỌI API & LƯU DỮ LIỆU

- Tối ưu:
  - 20 luồng song song
  - Retry 5 lần nếu lỗi, 0.5s
- Quản lý dữ liệu:
  - Tránh trùng lặp bằng biến `visited_coords`
  - Tự động retry 5 lần nếu lỗi khi gọi API (timeout, mất mạng,...)



### 3. GỘP & LÀM SẠCH DỮ LIỆU

- Quy trình:
  - Đọc toàn bộ các file trong thư mục `weather_chunks/` bằng glob
  - Gộp và xử lý từng chunk 25,000 dòng  
→ tối ưu bộ nhớ
  - Kết quả gộp vào `vietnam_weather_hourly.csv`
- Tiền xử lý:
  - Định dạng Datetime: `YYYY-MM-DD HH:MM:SS`
  - Loại dòng lỗi, NaT



## 4. CHIA & XỬ LÝ DỮ LIỆU

- Tính đặc trưng: giờ, ngày, tháng, mùa, sin/cos, lag, trend...
- Loại bỏ giá trị lồi (-999)
- Chia train/test theo thời gian (Tính điểm cắt:
  - $cutoff\_time = min\_date + timedelta(80\% \text{ thời gian})$
- Kiểm tra phân bố không gian cho thấy số điểm lưới đồng đều trên toàn quốc

## 5. LƯU TRỮ & TỐI ƯU

- File đầu ra:
  - `train_data.csv`: Dữ liệu huấn luyện
  - `test_data.csv`: Dữ liệu kiểm tra
- Quản lý file tạm:
  - File `temp_sorted.csv` được xóa sau khi hoàn tất bằng `remove_if_exists` để tiết kiệm dung lượng.



# 6. TỐI ƯU HÓA

- Tối ưu bộ nhớ:
  - Sử dụng chunksize=25000.
  - Gọi gc.collect() sau mỗi bước để giải phóng bộ nhớ không cần thiết.
  - Dùng kiểu dữ liệu nhẹ: float32, int8 thay vì float64
- Tối ưu tốc độ:
  - Dùng ThreadPoolExecutor với 20 luồng trong get\_weather\_data.py để lấy dữ liệu song song.



# 7. KẾT QUẢ CUỐI

- Dữ liệu sẵn sàng để huấn luyện LSTM:
  - Chuẩn hóa, đầy đủ, sạch
  - Liên tục theo thời gian
  - Phân bố không gian hợp lý
  - Định dạng chuẩn cho MinMaxScaler và TensorFlow Data Pipeline



# **QUÁ TRÌNH HUẤN LUYỆN**

# 1. GIỚI THIỆU TỔNG QUAN

Mục tiêu:

- Cải thiện độ chính xác dự báo thời tiết ngắn hạn (24 giờ tới).
- So sánh 3 mô hình: Old Model, New Model, Finetuned New Model.
- Đánh giá bằng MAE & RMSE, so với baseline (trung bình lịch sử).



## 2. TỔNG QUAN MÔ HÌNH & DỮ LIỆU

- Input: 48 hoặc 24 giờ gần nhất.
- Biến mục tiêu: T2M, PRECTOTCORR, PS, QV2M, WS10M  
CLRSKY\_SFC\_SW\_DWN.

```
model = Sequential([
    Input(shape=(timesteps, len(feature_cols))),
    LSTM(128, return_sequences=True),
    Dropout(0.2),
    LSTM(64, return_sequences=False),
    Dense(128, activation='relu'),
    Dropout(0.2),
    Dense(24 * len(target_cols)),
    tf.keras.layers.Reshape((24, len(target_cols)))
])
model.compile(optimizer=Adam(1e-3), loss=Huber(), metrics=['mae'])
```

## 2.1 OLD MODEL

**Đặc trưng đầu vào:**

- Vị trí: Latitude, Longitude
- Thời gian: hour, day, month, season
- Biến thời tiết: WS10M, QV2M, PS, PRECTOTCORR, T2M, CLRSKY\_SFC\_SW\_DWN
- Chu kỳ: sin\_hour, cos\_hour

**Nhược điểm:**

- Thiếu đặc trưng trễ (lag) và xu hướng (trend)
- MSE nhạy cảm với outliers (như lượng mưa)

## 2.2 NEW MODEL

Đặc trưng bổ sung:

- Trễ: T2M\_lag1, PRECTOTCORR\_lag1
- Xu hướng: T2M\_trend, PRECTOTCORR\_trend
- Phân biệt ngày - đêm: T2M\_day\_night

Tối ưu hóa:

- Optimizer: Adam ( $lr = 0.001$ , nhanh hơn)
- Callbacks: EarlyStopping (patience = 5), ModelCheckpoint, EpochTracker

## 2.3 FINETUNED NEW MODEL

**Hàm mất mát:**

- Huber loss có trọng số:
  - T2M: 1.5
  - PRECTOTCORR: 3.0
  - PS: 2.0

**Lý do cải tiến:**

- Bắt được xu hướng tốt hơn
- Huber loss giảm ảnh hưởng outliers

### 3. SO SÁNH CHI TIẾT BA MÔ HÌNH

Tiêu chí	Old Model	New Model	Finetuned New Model
Kiến trúc mô hình	Mạng LSTM cơ bản	Mạng LSTM với đặc trưng bổ sung	Mạng LSTM tinh chỉnh từ New Model
Đặc trưng đầu vào	<ul style="list-style-type: none"> <li>- Vị trí: Latitude, Longitude</li> <li>- Thời gian: hour, day, month, season</li> <li>- Biến thời tiết: WS10M, QV2M, PS, PRECTOTCORR, T2M, CLRSKY_SFC_SW_DWN</li> <li>- Chu kỳ: sin_hour, cos_hour</li> </ul>	<ul style="list-style-type: none"> <li>- Giảm timesteps từ 48 xuống 24</li> <li>- Bổ sung: T2M_lag1, PRECTOTCORR_lag1, T2M_trend, PRECTOTCORR_trend, T2M_day_night</li> </ul>	Như New Model
Hàm mất mát	Mean Squared Error (MSE)	Huber loss không trọng số	Huber loss với trọng số (T2M: 1.5, PRECTOTCORR: 3.0, PS: 2.0)
Optimizer	Adam, learning rate = 0.000070	Adam, learning rate = 0.001	Adam, learning rate = 0.00005
Ưu điểm	<ul style="list-style-type: none"> <li>- Tiết kiệm tài nguyên</li> <li>- Tự động điều chỉnh learning rate</li> </ul>	<ul style="list-style-type: none"> <li>- Tăng số mẫu huấn luyện</li> <li>- Huấn luyện nhanh hơn</li> </ul>	<ul style="list-style-type: none"> <li>- Đặc trưng phong phú</li> <li>- Hàm mất mát tối ưu</li> <li>- Ngăn overfitting</li> </ul>
Nhược điểm	<ul style="list-style-type: none"> <li>- Thiếu đặc trưng bổ sung</li> <li>- Hàm mất mát nhạy cảm với outliers</li> </ul>	<ul style="list-style-type: none"> <li>- Mất thông tin dài hạn</li> <li>- Hiệu suất giảm ở PRECTOTCORR</li> </ul>	<ul style="list-style-type: none"> <li>- Độ phức tạp tăng</li> <li>- Vẫn gặp khó khăn với PRECTOTCORR</li> </ul>

# KIỂM TRA VÀ ĐÁNH GIÁ

# **1. KIỂM TRA VỚI DỮ LIỆU THỰC**

**Chuẩn bị file dữ liệu thực tế khác với train và test (3/2025)**

## **Tạo đặc trưng bổ sung:**

- sin\_hour, cos\_hour: Chu kỳ giờ trong ngày.
- T2M\_lag1, PRECTOTCORR\_lag1: Giá trị trễ 1 giờ.
- T2M\_trend, PRECTOTCORR\_trend: Xu hướng dựa trên trung bình trượt 24 giờ.
- T2M\_day\_night: Phân biệt ngày (6-18h) và đêm.

**Chuẩn hóa: Sử dụng scaler\_X và scaler\_y từ file scaler.pkl.**

## **Tạo batch:**

- X\_batch: 48 giờ lịch sử (timesteps=48).
- y\_batch: 24 giờ thực tế (output\_steps=24).

## 2. KẾT QUẢ - OLD MODEL

Biến	MAE	RMSE	MAE (Baseline)	RMSE (Baseline)
PS	0.2837581	0.36288986	0.18182431	0.22983322
T2M	3.1665065	4.165608	2.7757084	3.1443548
QV2M	1.6127129	2.0228758	1.1080515	1.3801826
WS10M	1.5021013	1.8443718	1.1465452	1.4099838
PRECTOTCORR	3.6950228	5.345674	2.0389762	4.88422

So sánh với baseline:

- MAE và RMSE của Old Model cao hơn baseline ở tất cả các biến.
- Đặc biệt kém ở PRECTOTCORR (lượng mưa) và T2M (nhiệt độ).

Nhận xét:

- Mô hình không học tốt các mẫu thời gian phức tạp.
- Thiếu đặc trưng bổ sung và hàm măt mát không phù hợp với các giá trị ngoại lai (outliers).

## 2. KẾT QUẢ - NEW MODEL

Biến	MAE	RMSE	MAE (Baseline)	RMSE (Baseline)
PS	0.1838401	0.22654076	0.18182431	0.22983322
T2M	2.232472	2.4892335	2.7757084	3.1443548
QV2M	1.073823	1.3632295	1.1080515	1.3801826
WS10M	1.0297856	1.2771043	1.1465452	1.4099838
PRECTOTCORR	2.2042506	4.988393	2.0389762	4.88422

So sánh với baseline:

- MAE và RMSE thấp hơn baseline ở hầu hết các biến.
- PRECTOTCORR: MAE cải thiện, nhưng RMSE vẫn cao (4.9884 so với baseline 4.8842)

Nhận xét:

- Cải thiện đáng kể so với Old Model nhờ thêm đặc trưng và sử dụng Huber loss.
- Vẫn khó dự đoán chính xác các giá trị cực đại của PRECTOTCORR

## 2. KẾT QUẢ - FINETUNED NEW MODEL

Biến	MAE	RMSE	MAE (Baseline)	RMSE (Baseline)
PS	0.16728656	0.20782438	0.18182431	0.22983322
T2M	1.7294979	2.120125	2.7757084	3.1443548
QV2M	0.8721738	1.106748	1.1080515	1.3801826
WS10M	1.0704701	1.319345	1.1465452	1.4099838
PRECTOTCORR	2.9921608	5.202909	2.0389762	4.88422

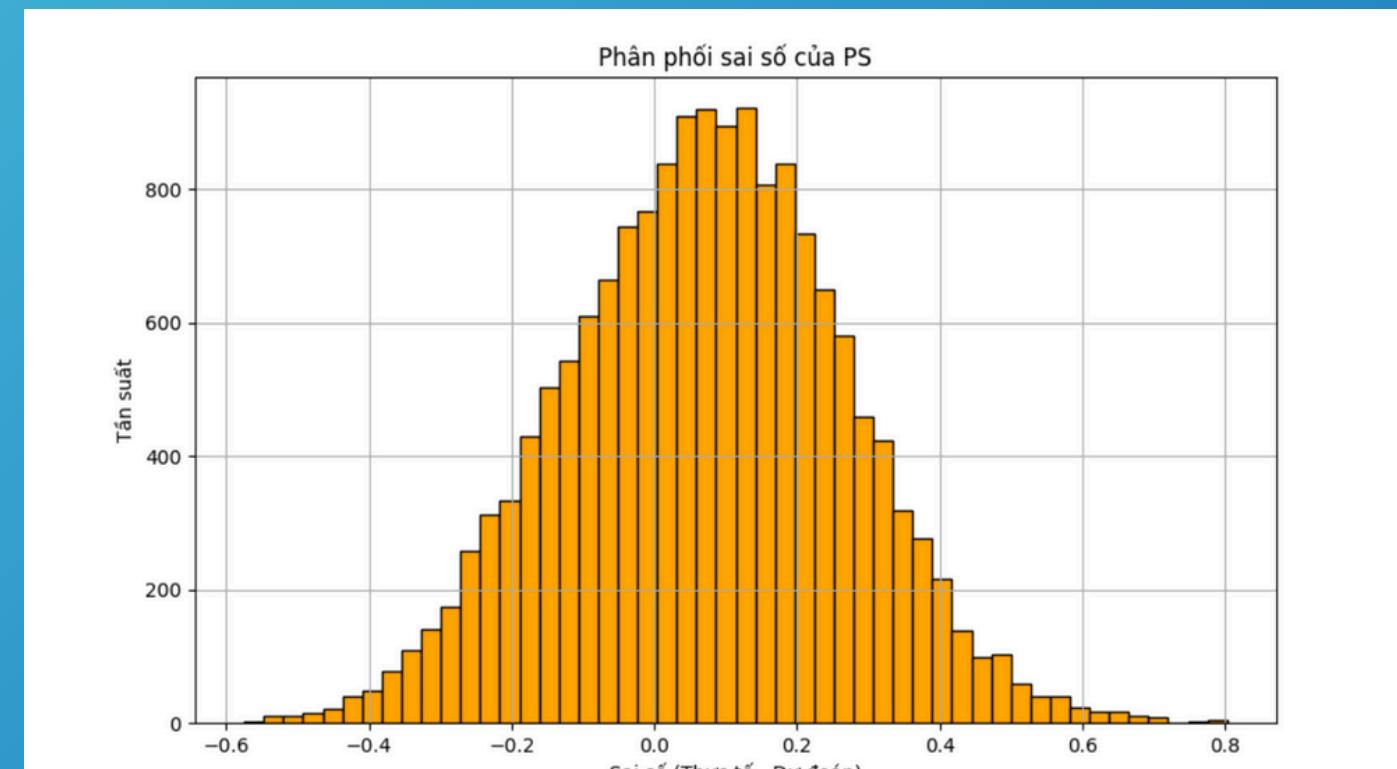
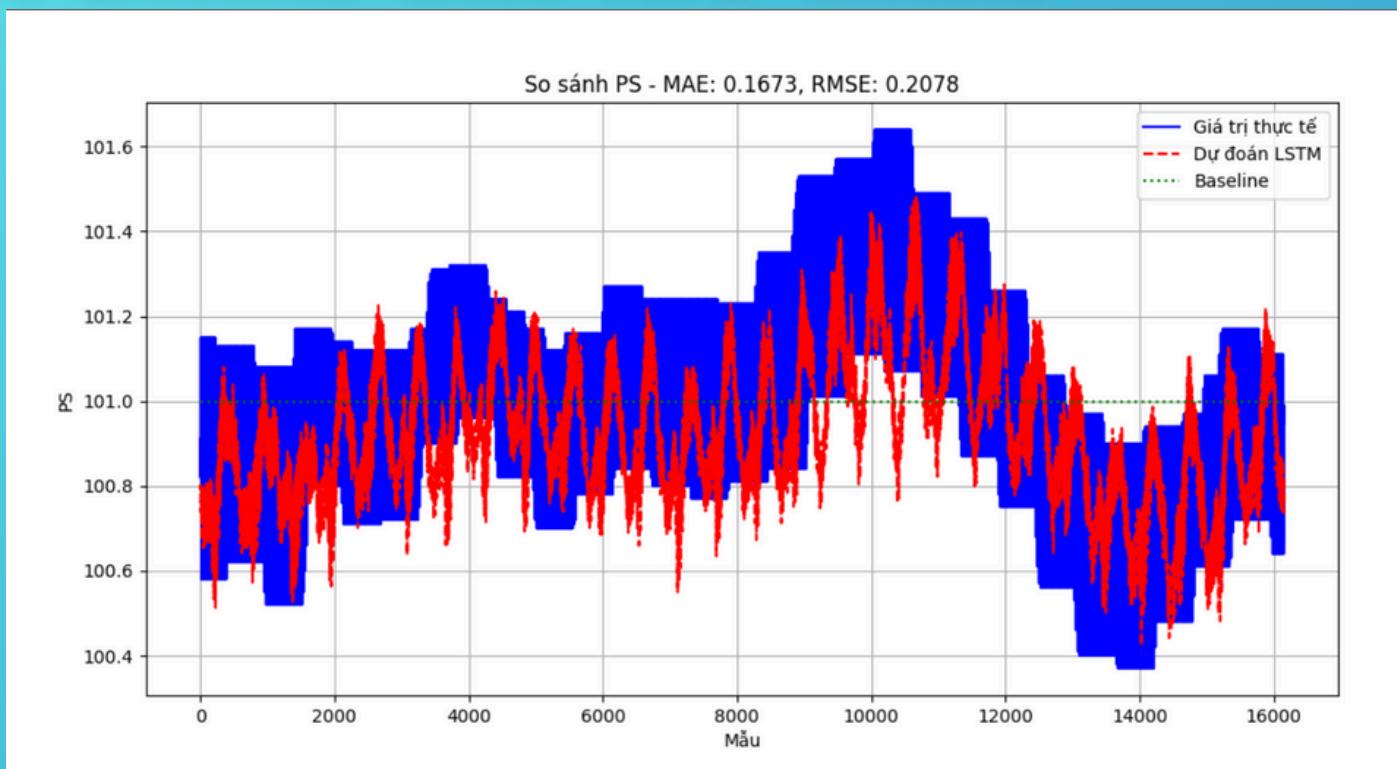
So sánh với baseline:

- MAE và RMSE thấp hơn baseline cho PS, T2M, QV2M, WS10M.
- PRECTOTCORR: MAE và RMSE cao hơn cả New Model và baseline.

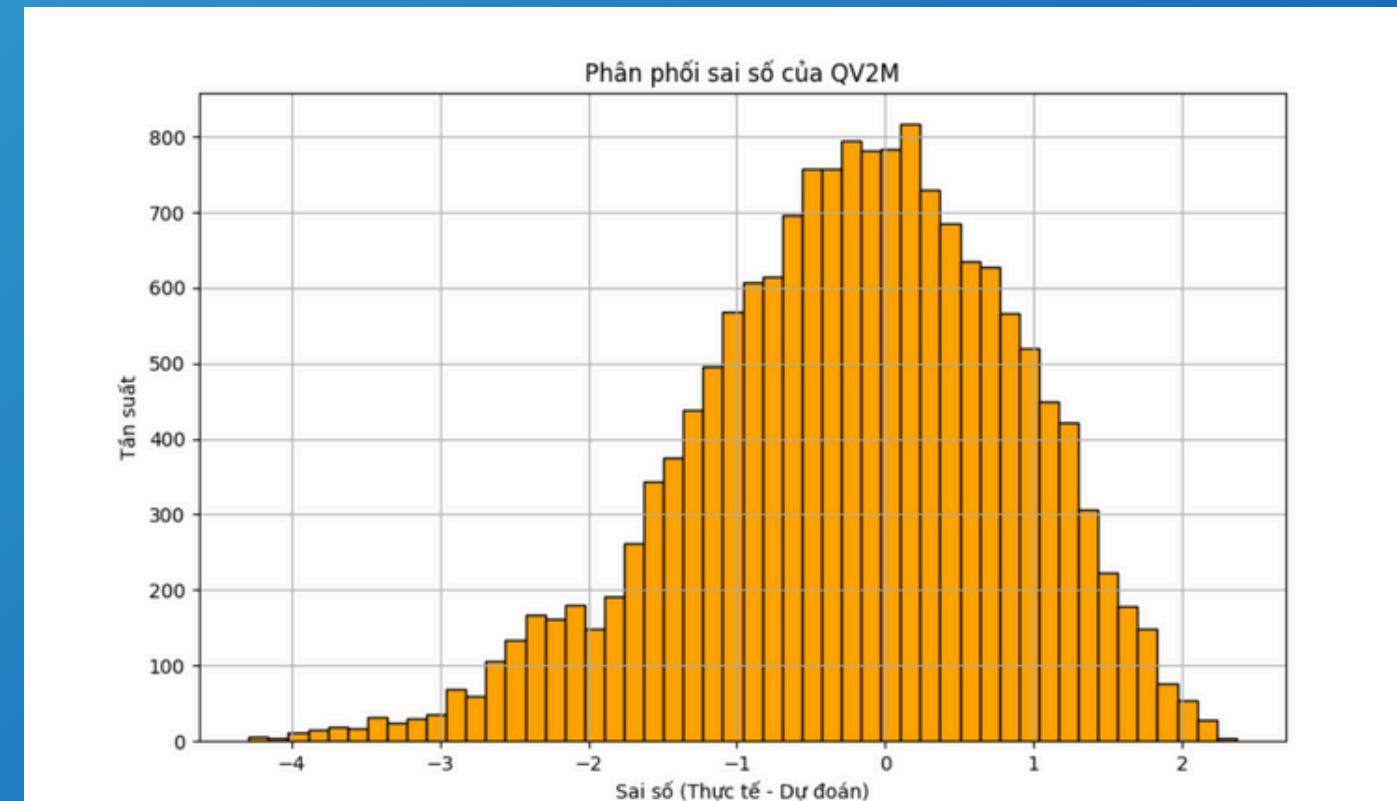
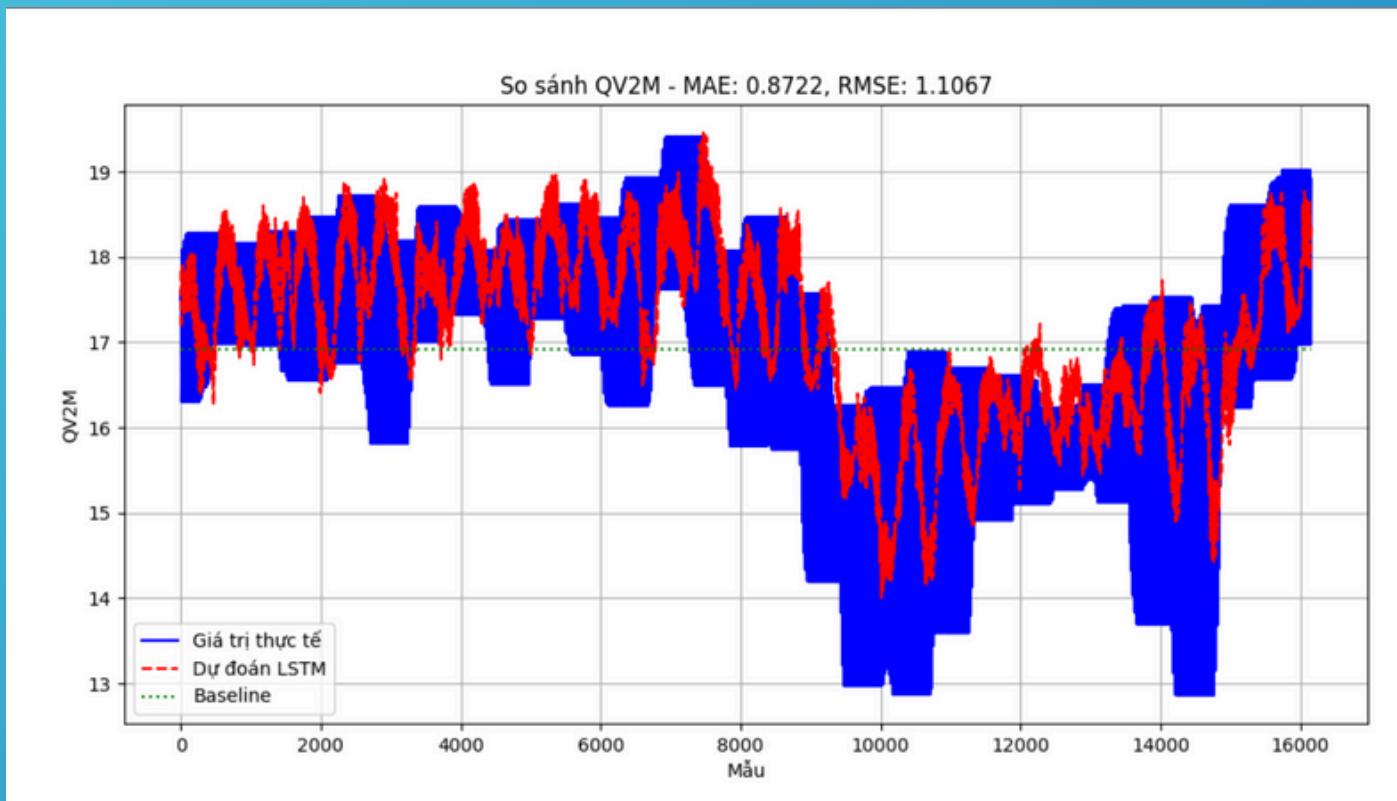
Nhận xét:

- Hiệu suất tốt hơn ở các biến ngắn hạn như T2M, PS.
- Kém hơn ở PRECTOTCORR do giảm số timesteps, làm mất thông tin dài hạn.

### 3. ĐÁNH GIÁ MÔ HÌNH

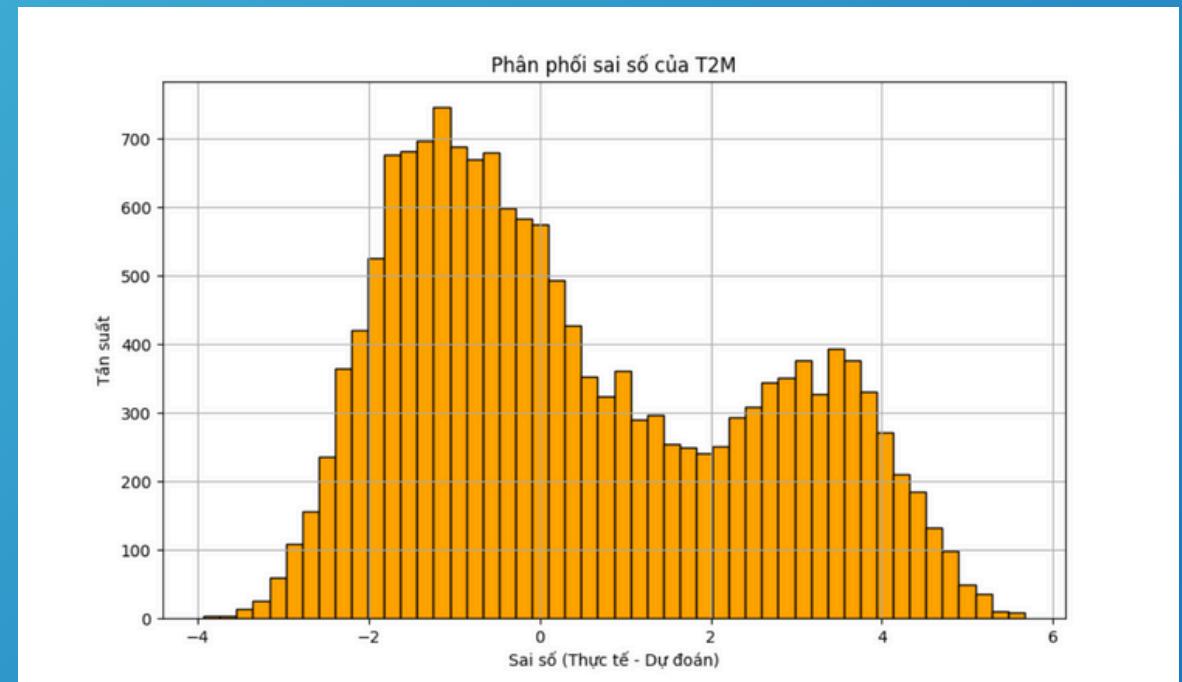
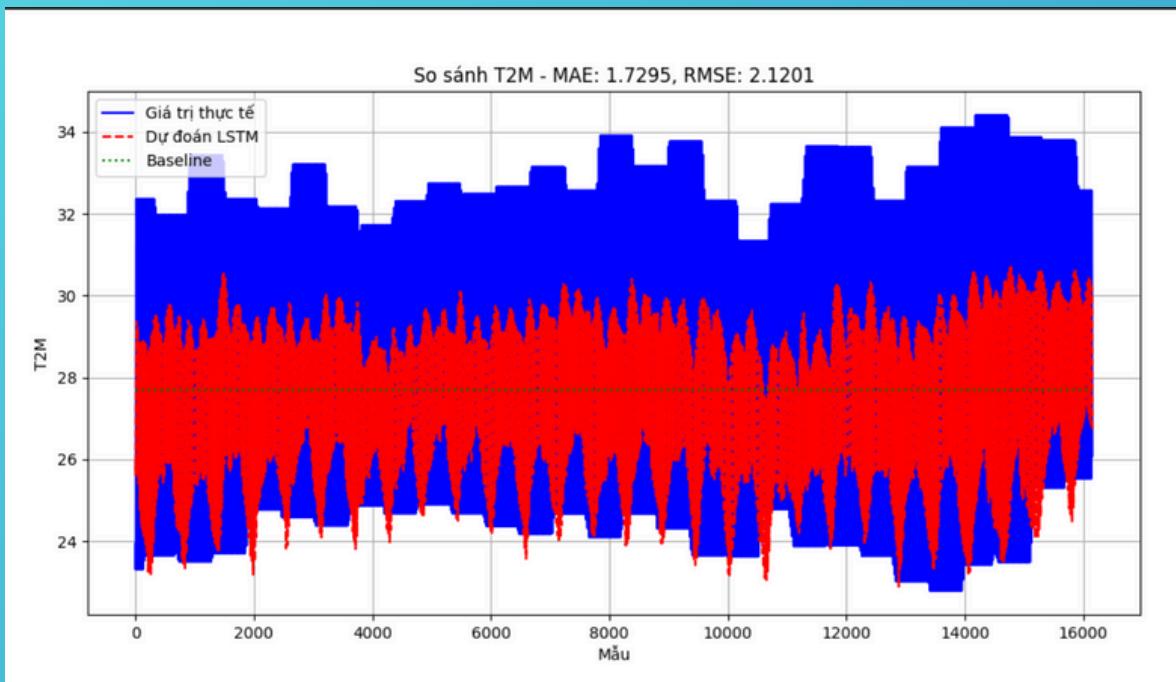


PS

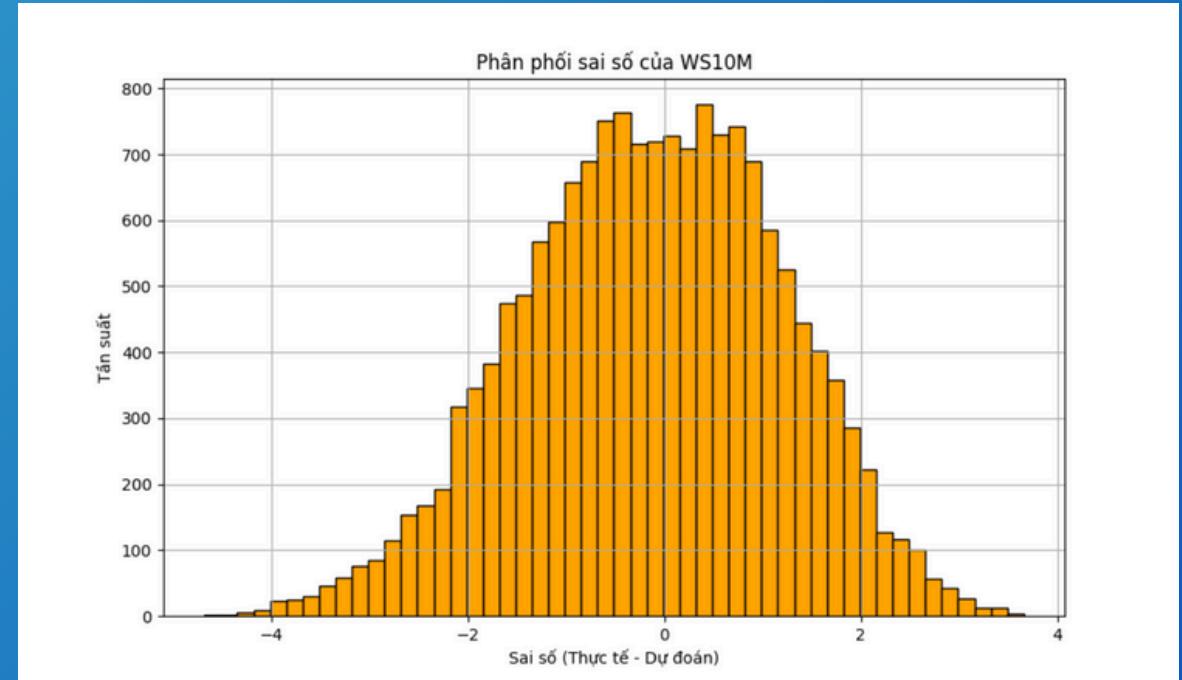
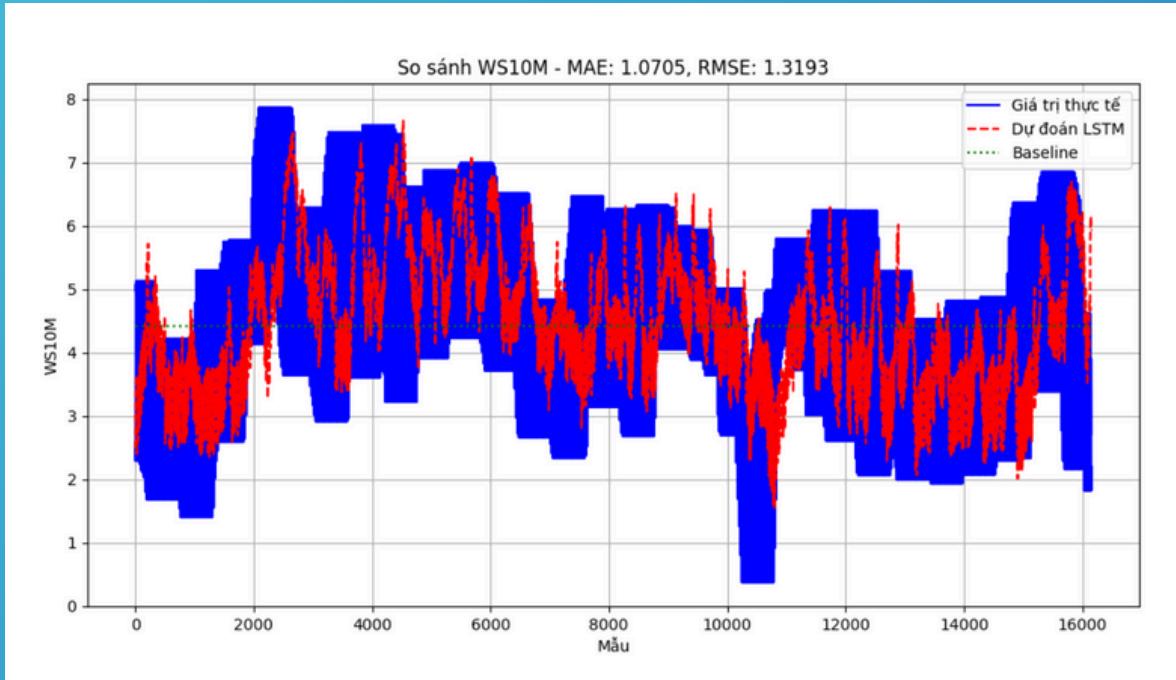


QV2M

### 3. ĐÁNH GIÁ MÔ HÌNH

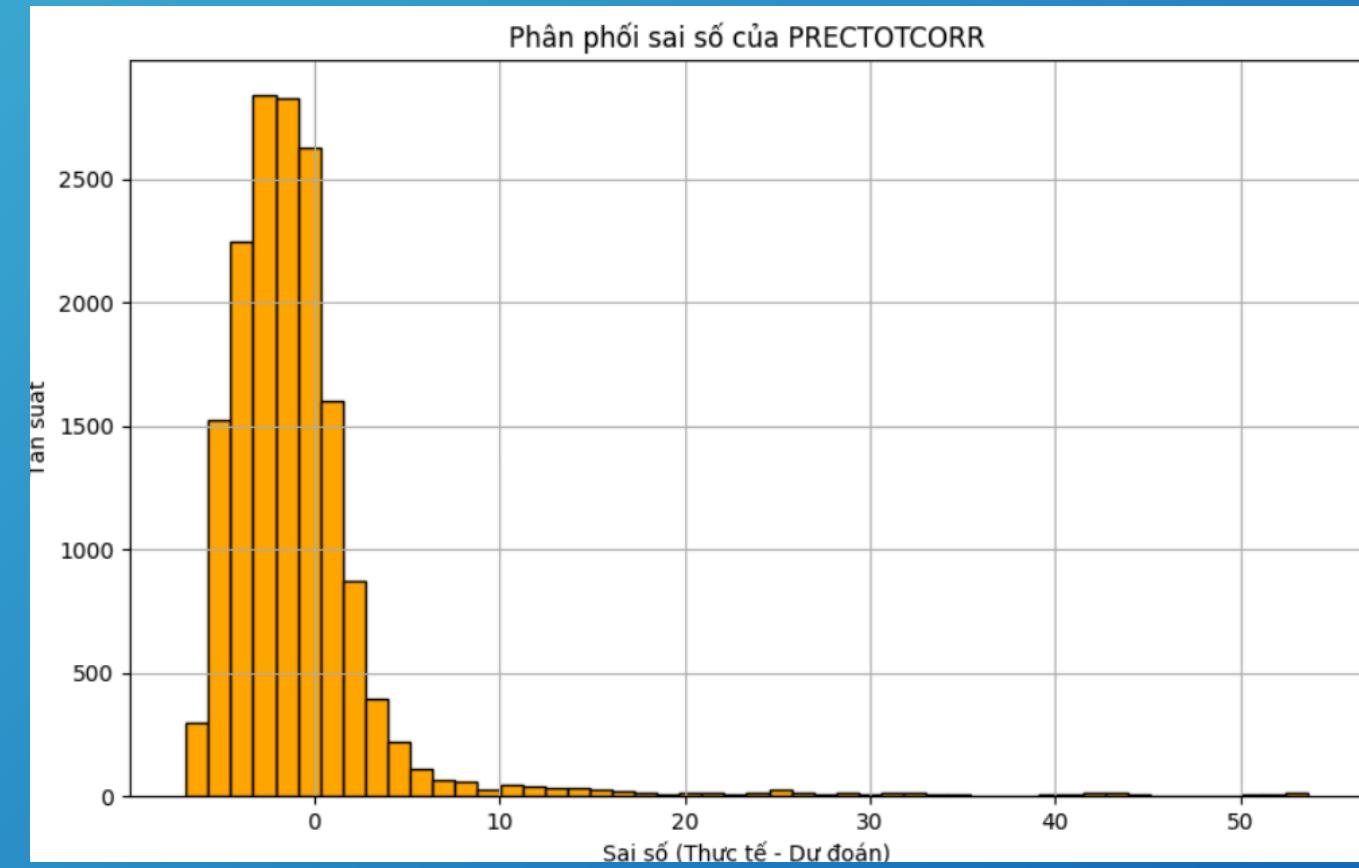
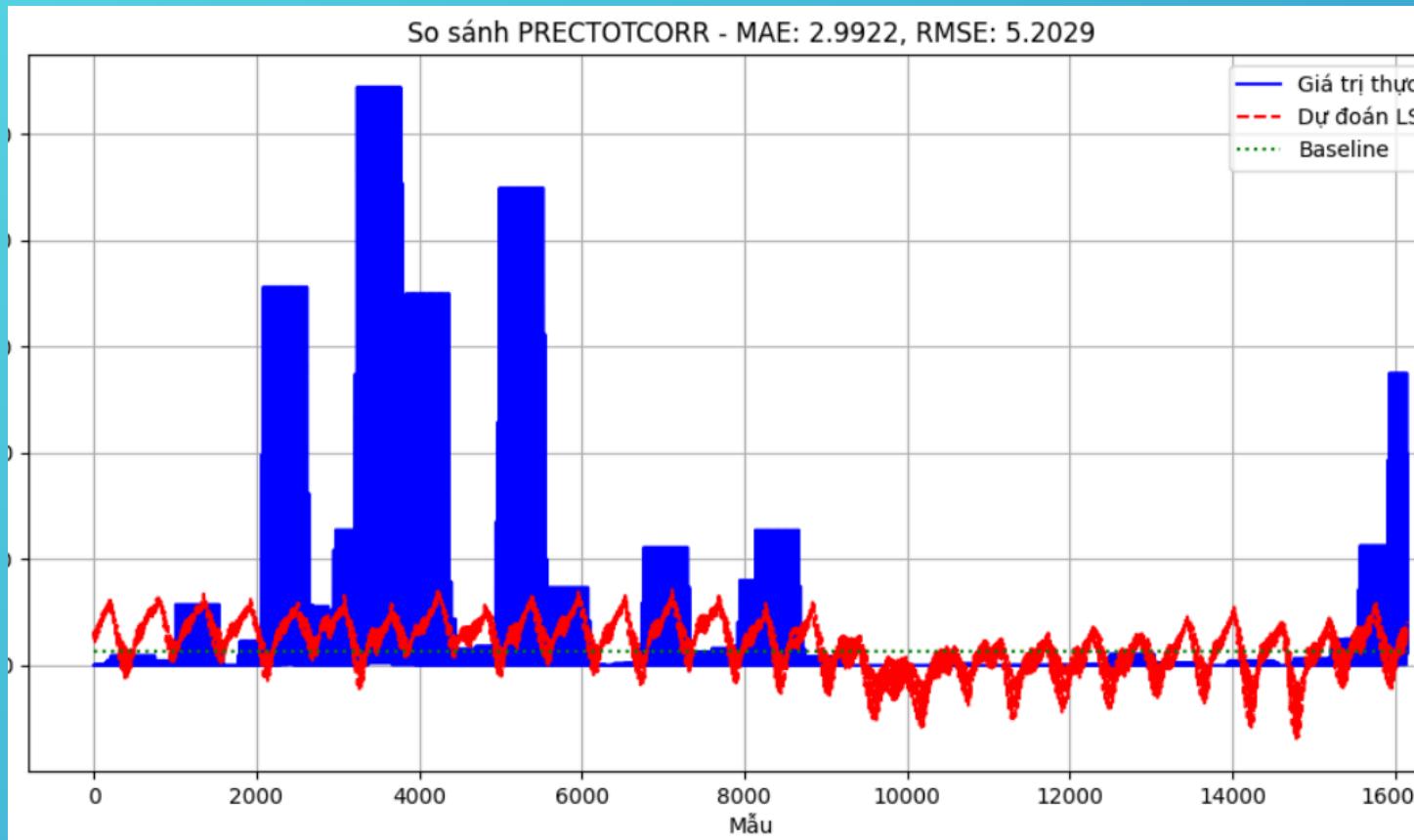


T2M



WS10M

### 3. ĐÁNH GIÁ MÔ HÌNH



PRECTOTCORR

### 3. KẾT LUẬN & HƯỚNG CẢI TIẾN

Tiến bộ rõ rệt từ Old → New → Finetuned

PRECTOTCORR vẫn là bài toán khó

Gợi ý cải tiến tiếp theo:

- Tối ưu timesteps: thử 36 giờ (giữa 24 và 48)
- Mô hình nâng cao: CNN + LSTM, hoặc Transformer
- Bổ sung dữ liệu: mở rộng dữ liệu khí tượng cho PRECTOTCORR
- Tinh chỉnh trọng số loss cho biến phức tạp



**THANK YOU  
FOR LISTENING!**

