FORECASTING AIR POLLUTION FOR PROACTIVE POLICY INTERVENTION

Samyak Jain

MOTIVATION

- India’s pollution crisis – over 2mn premature deaths in India can be attributed to long term exposure to high levels of air pollution.
- The worst states are Delhi and Haryana, where air quality can exceed 13 times the WHO guidelines.
- In response to these health concerns, in 2016, a short-term emergency response policy called **Graded Response Action Plan** (GRAP) was devised.
- GRAP stages (next page) are invoked reactively and not proactively, based on forecasts

Goal: **Predict PM2.5 concentration up to 14 days in advance**, so that policies GRAP can be invoked early to control severe pollution spikes and limit exposure, enhancing GRAP's effectiveness and limiting the health risks.

GRAP STAGES AND MEASURES

GRAP Stage	Air Quality Index (AQI) category	AQI range	Measures
I	Poor	201–300	-Dust control and road sweeping -Public transport promoted
II	Very Poor	301–400	-Intensify traffic management -Ban diesel generator
III	Severe	401–450	-Restrict entry of certain trucks -Close certain industrial plants
IV	Severe+ (Emergency)	>450	-Close schools and halt all construction -Odd even vehicle restrictions -Ban entry of trucks

Does GRAP grip at all? (Kattuman, Harvey, Singh)

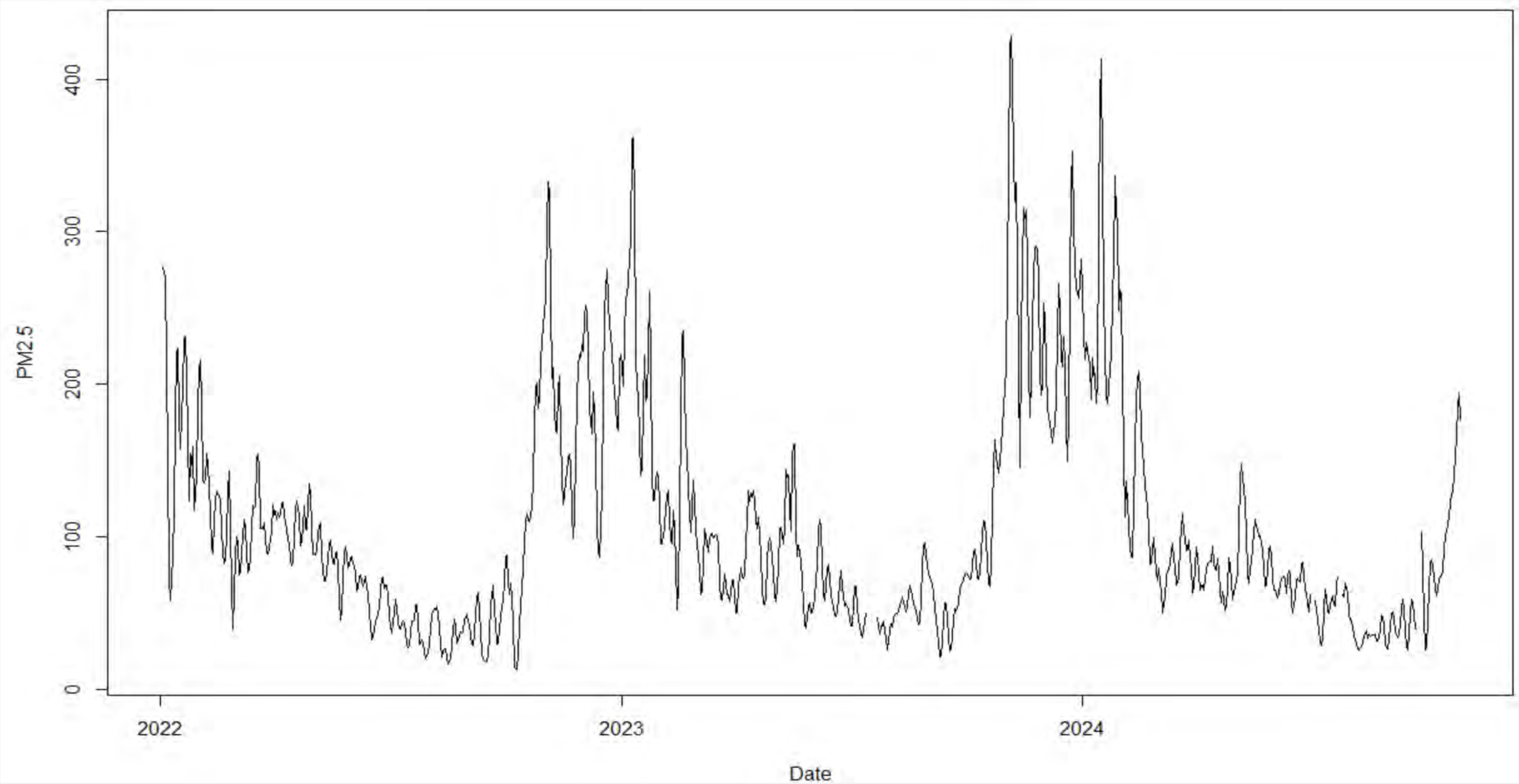
HOW IS POLLUTION MEASURED?

- The Air Quality Index takes into account several pollutant types such as
  - PM2.5 (particulate matter of diameter 2.5 micrometres or less)
  - PM10 (particulate matter of diameter 10 micrometres or less)
  - Ozone
  - Nitrogen Dioxide
  - Sulphur Dioxide
  - Carbon Monoxide

The most important of these is PM2.5 (measured in $\mu g/m^3$)

- Diameter <2.5μm
- Penetrates deeply into the lungs
- Linked to health problems such as asthma, heart disease, stroke

PM2.5 DATA FROM DELHI



- We have PM2.5 concentration levels from the RK Puram region in Delhi from Jan 2022 to Oct 2024.
- There is strong seasonality due to meteorological variables and agricultural practices.
- There are sharp spikes beginning late in the year (consistently around mid October), corresponding to periods of crop burning in neighbouring states

UNIVARIATE TIME SERIES MODELLING

Model	Description	Mean Square Error						
		Horizon 1	Horizon 2	Horizon 3	Horizon 4	Horizon 5	Horizon 6	Horizon 7
Delay (Baseline)	Use reading at time t as forecast	2,650	4,818	5,187	5,685	6,821	7,750	7,681
ETS	Exponential smoothing with trend/seasonality components	2,382	3,801	4,144	4,747	5,700	6,425	6,556
ARIMA	Autoregressive Integrated Moving Average model	2,258	3,883	4,381	4,921	5,574	6,130	6,271
SARIMA	Seasonal ARIMA, captures both non-seasonal and seasonal patterns	2,201	3,993	4,638	5,172	6,111	6,851	6,888
Harmonic Regression	Regression model with Fourier terms to capture seasonality	2,025	3,082	3,250	3,430	3,634	3,754	3,744
TBATS	Exponential smoothing with Box-Cox, ARMA errors, trend, and seasonal terms	2,415	4,103	4,362	4,746	5,422	5,923	6,004

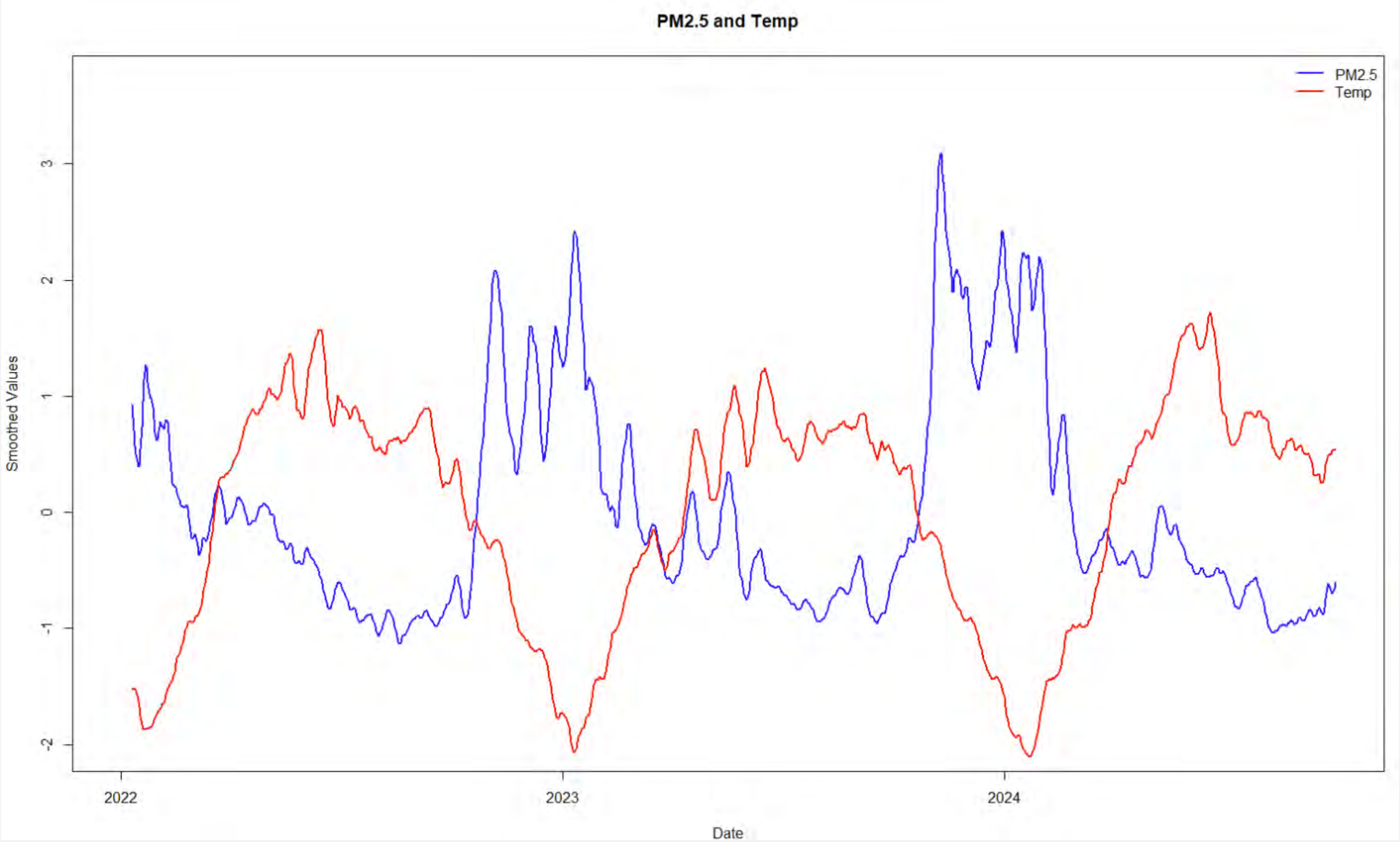
- Used increasingly more complex time series models to use previous readings of PM2.5 to predict readings of PM2.5 at different times in the future
- Harmonic Regression performs best, capturing the strong, regular seasonality well
- While TBATS is a more complex model, it tries to capture multiple seasonalities and overfits to the data

METEOROLOGICAL DATA

- **Temperature**

This is the temperature taken 2 meters above the surface and is a standard indicator of the temperature experienced by people.

- Temperature and PM2.5 have a complex relationship. In the short term, high temperatures can lead to increased PM2.5 due to photochemical reactions. In the long term, higher temperatures encourage more atmospheric mixing which cause a fall in PM2.5 concentration (this is key relation that this graph shows).

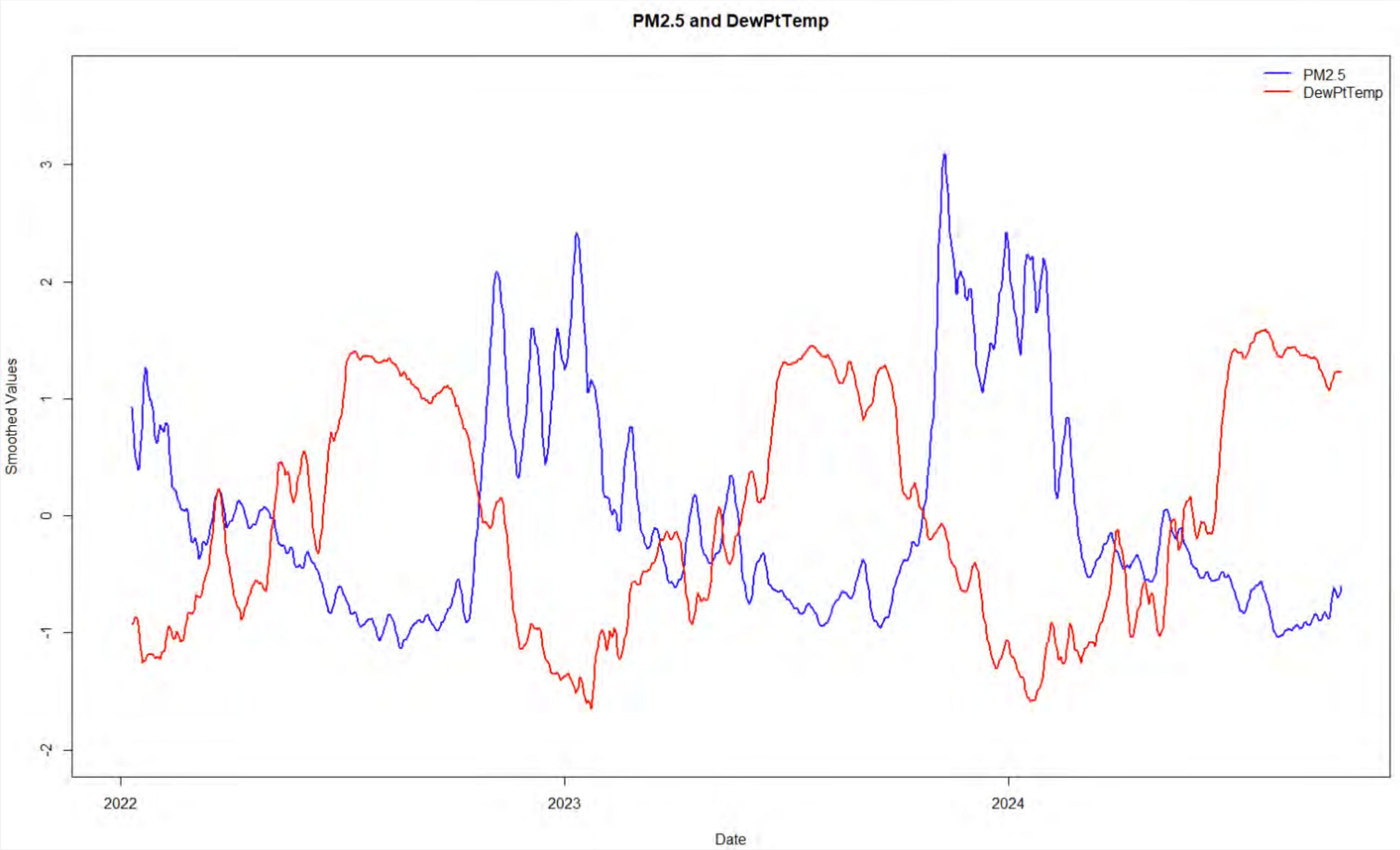


METEOROLOGICAL DATA

- **Dew Point Temperature**

The temperature to which the air must be cooled down to reach its moisture capacity.

- There are two opposing effects here. A high dew point temperature contributes to PM2.5 formation as water vapour condenses on gases forming larger particles. If the dew point is very high, then condensation can lower PM2.5.

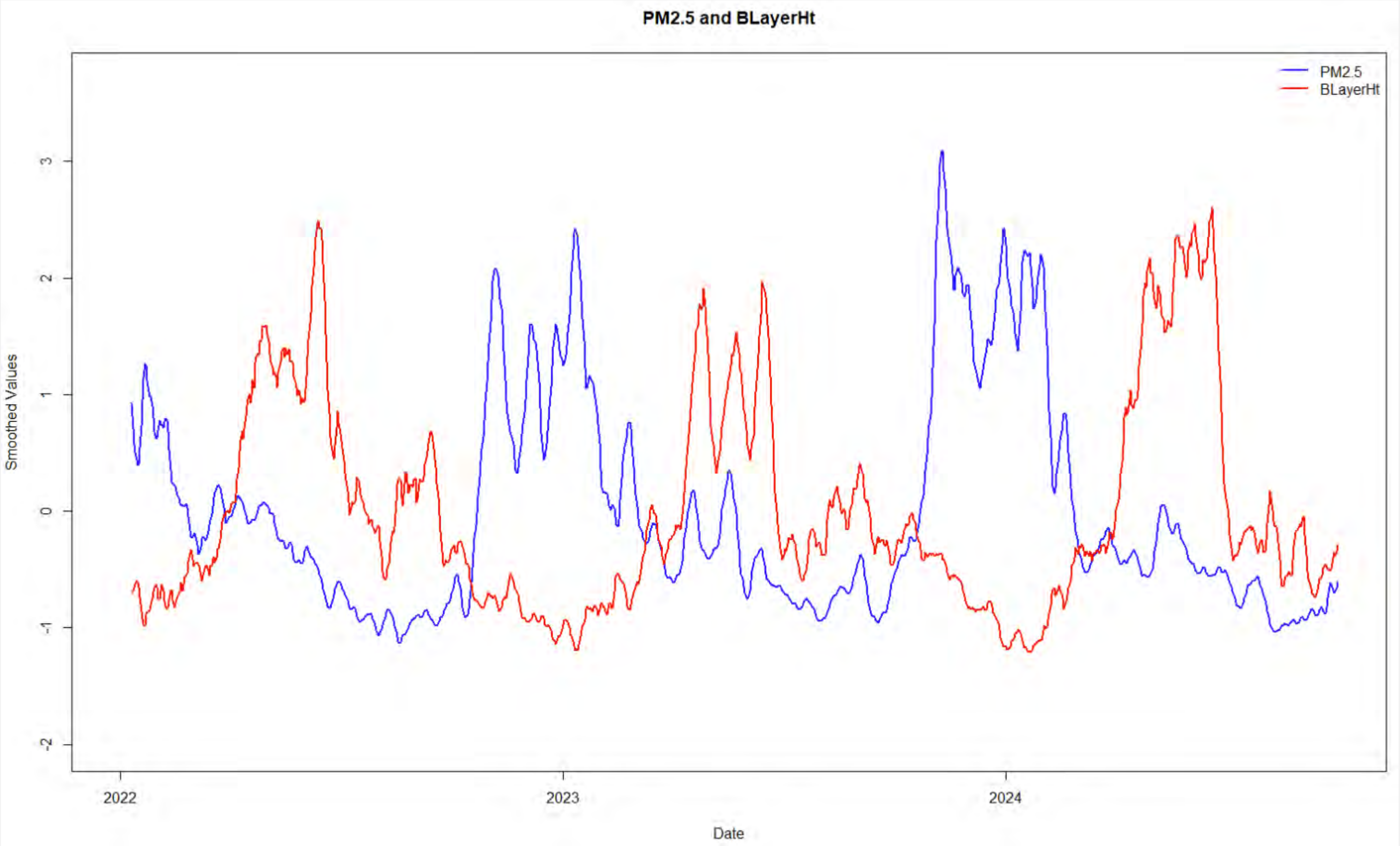


METEOROLOGICAL DATA

- **Boundary Layer height**

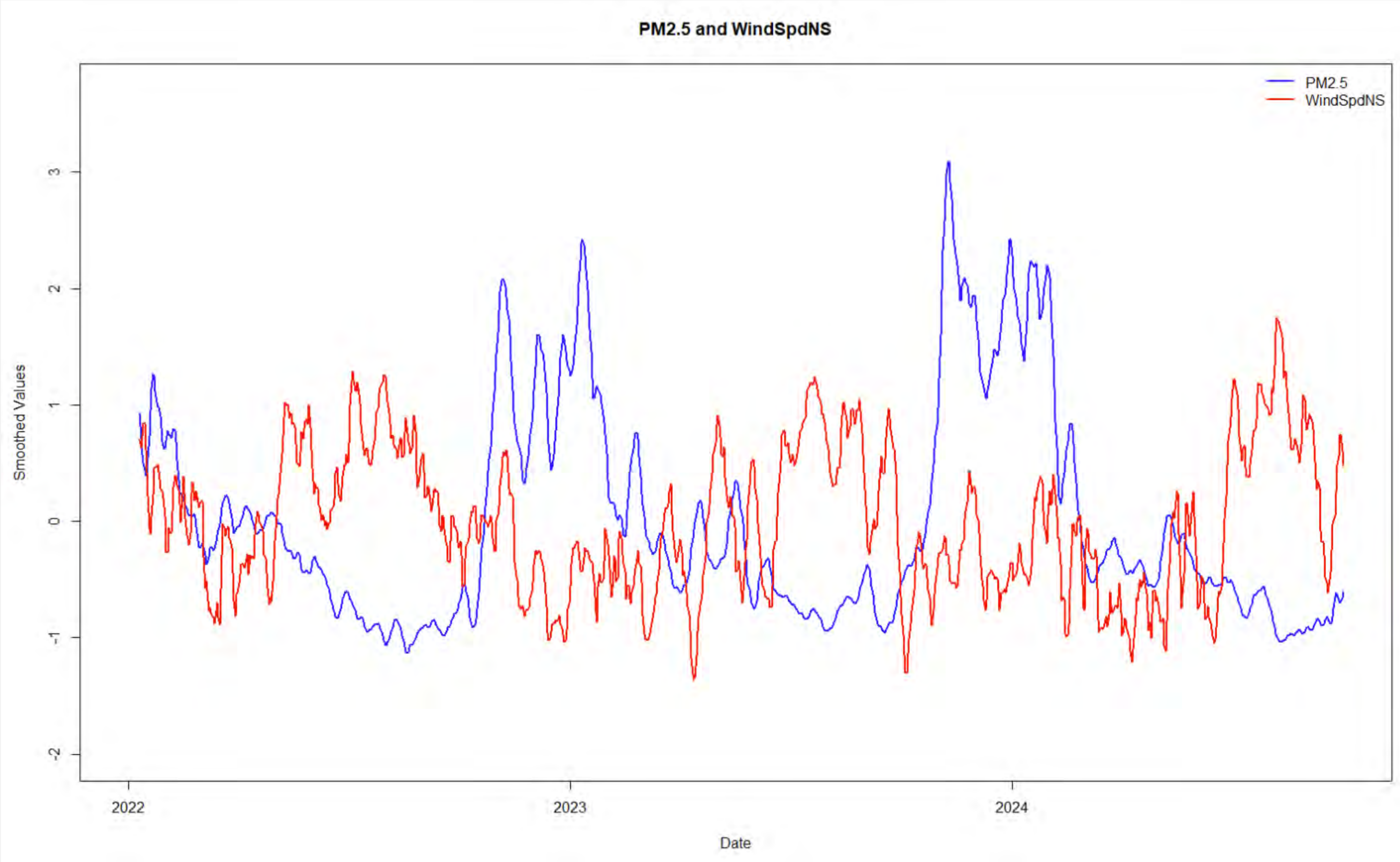
The planetary boundary layer/Troposphere is the lowest layer of the atmosphere.

- It is strongly negatively correlated with PM2.5, as a larger boundary layer height increases vertical mixing and pollutant dispersion.



METEOROLOGICAL DATA

- **Wind speed and Direction**
- Both North/South and East/West directions were included in the dataset.
- Generally, higher wind speeds disperse pollutants so lead to lower PM2.5.

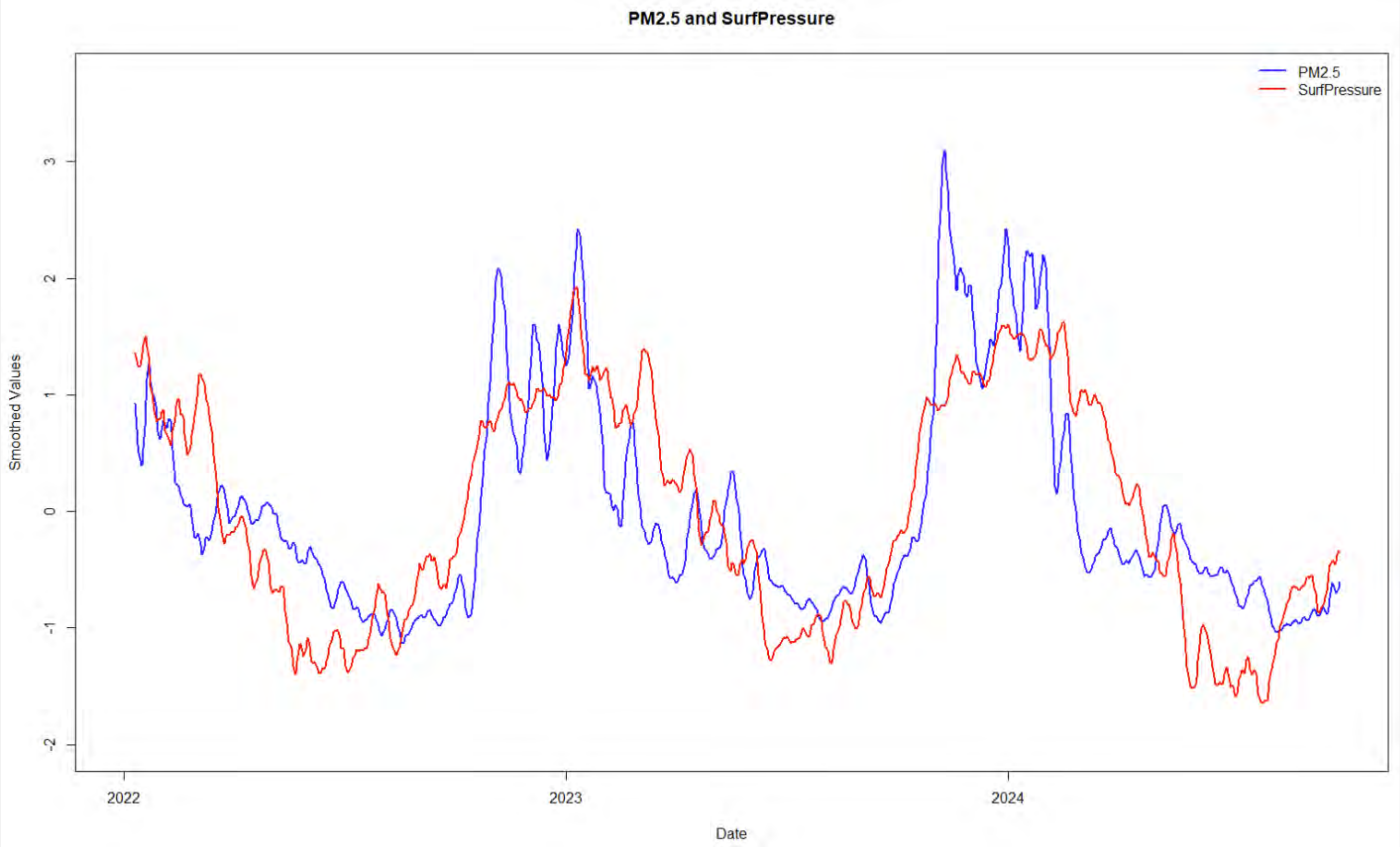


METEOROLOGICAL DATA

- **Surface Pressure**

This is the atmospheric pressure at a location on Earth's surface and is directly proportional to the total weight of the air directly above a specific area.

- We see it follows PM2.5. very well due to, high surface pressure leading to a stable atmosphere which allows pollutants to accumulate.

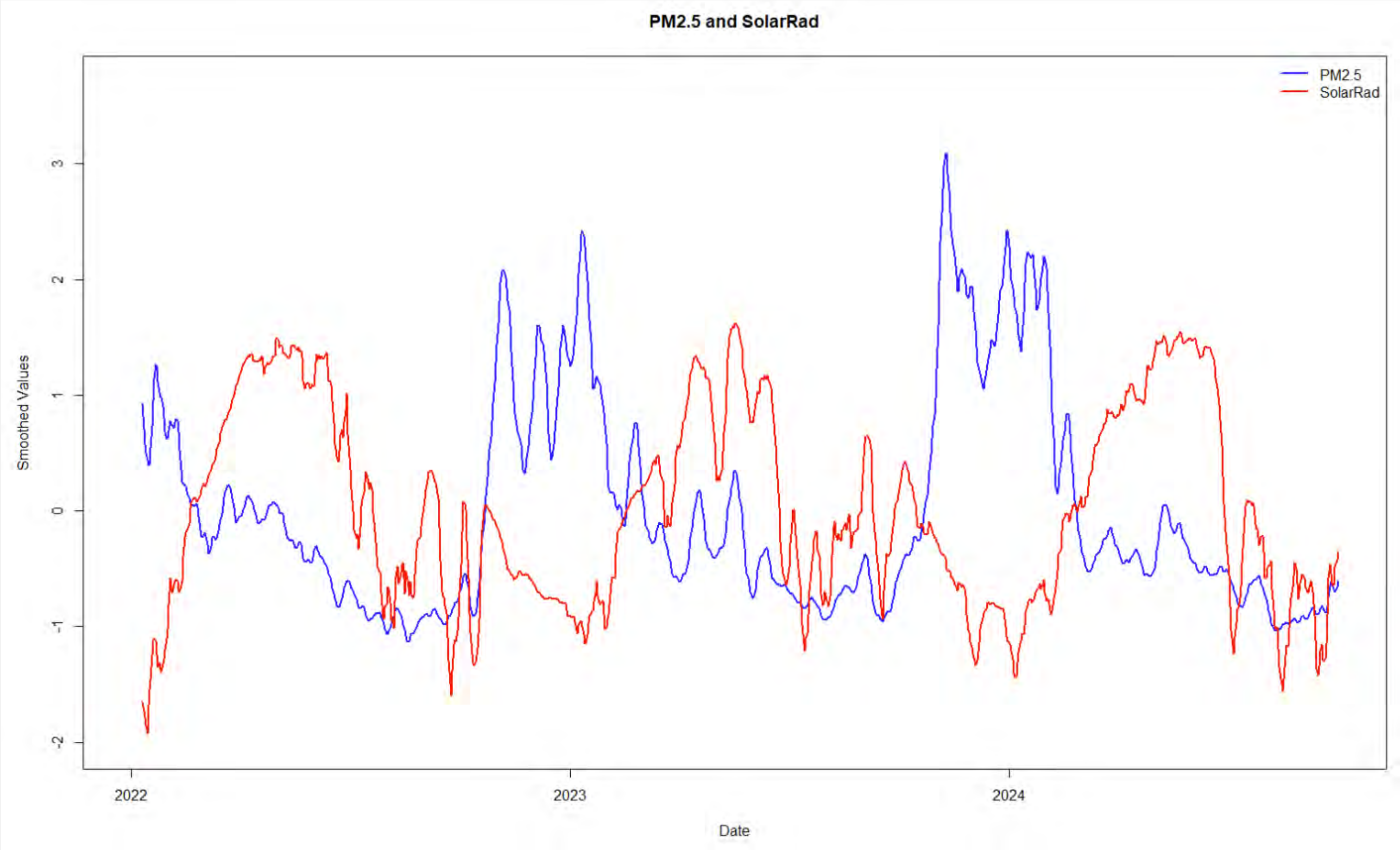


METEOROLOGICAL DATA

- **Solar Radiation**

Power per unit area received from the sun in the form of EM radiation.

- Solar radiation promotes photochemical reactions, which produces secondary PM2.5.
- Also, a high concentration causes dispersion of solar radiation, therefore during high PM2.5, lower solar radiation is observed.
- Indirectly affects PM2.5 through boundary layer height, temperature, humidity etc.



IDENTIFYING LAGGED STRUCTURE BETWEEN PM2.5 AND METEOROLOGICAL VARIABLES

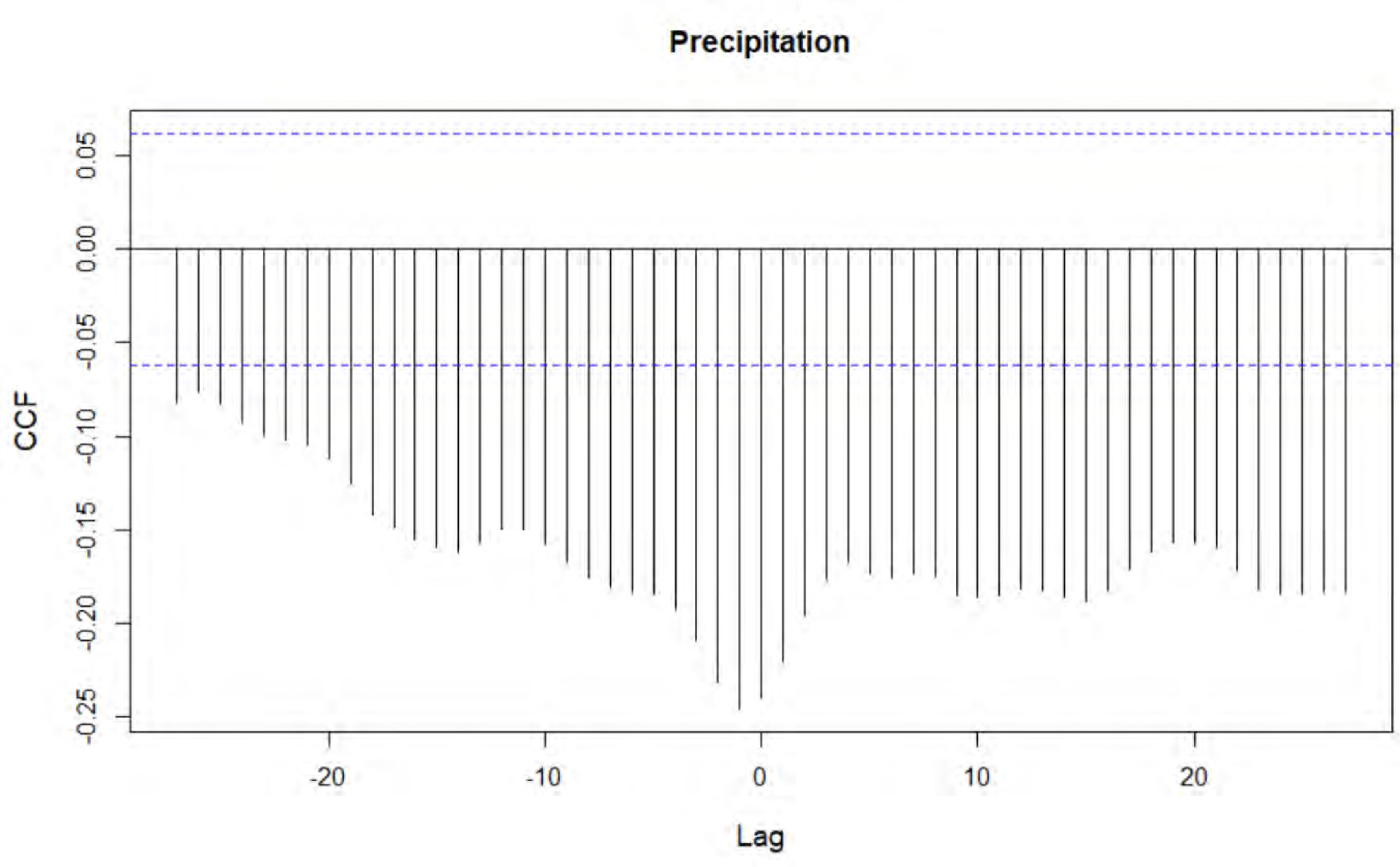
- Cross correlation functions measure the strength and direction of any linearity between two time series and different time lags.
- If x_t, y_t are two time series, then their cross-correlation function is

$$\text{CCF(d)} = \frac{Cov(x_{t-d}, y_t)}{\sqrt{Var(x_{t-d})Var(y_t)}} \approx \frac{\sum(x_{i-d}-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}}$$

where d is the lag.

Example with PM2.5 and precipitation

CROSS CORRELATION BETWEEN PRECIPITATION AND PM2.5



PRE-WHITENING TO REMOVE AUTOCORRELATION EFFECTS

- The idea is to find a linear, reversible transformation that can remove the confounding effect of seasonality on both PM2.5 and Precipitation.
- The general form of an ARMA(p,q) model

$$x_t = \sum_{i=1,\ldots,p} \phi_i x_{t-i} + \sum_{i=1,\ldots,q} \theta_i \epsilon_{t-i} + \epsilon_t$$

where $\epsilon_t \sim WN(\sigma^2)$

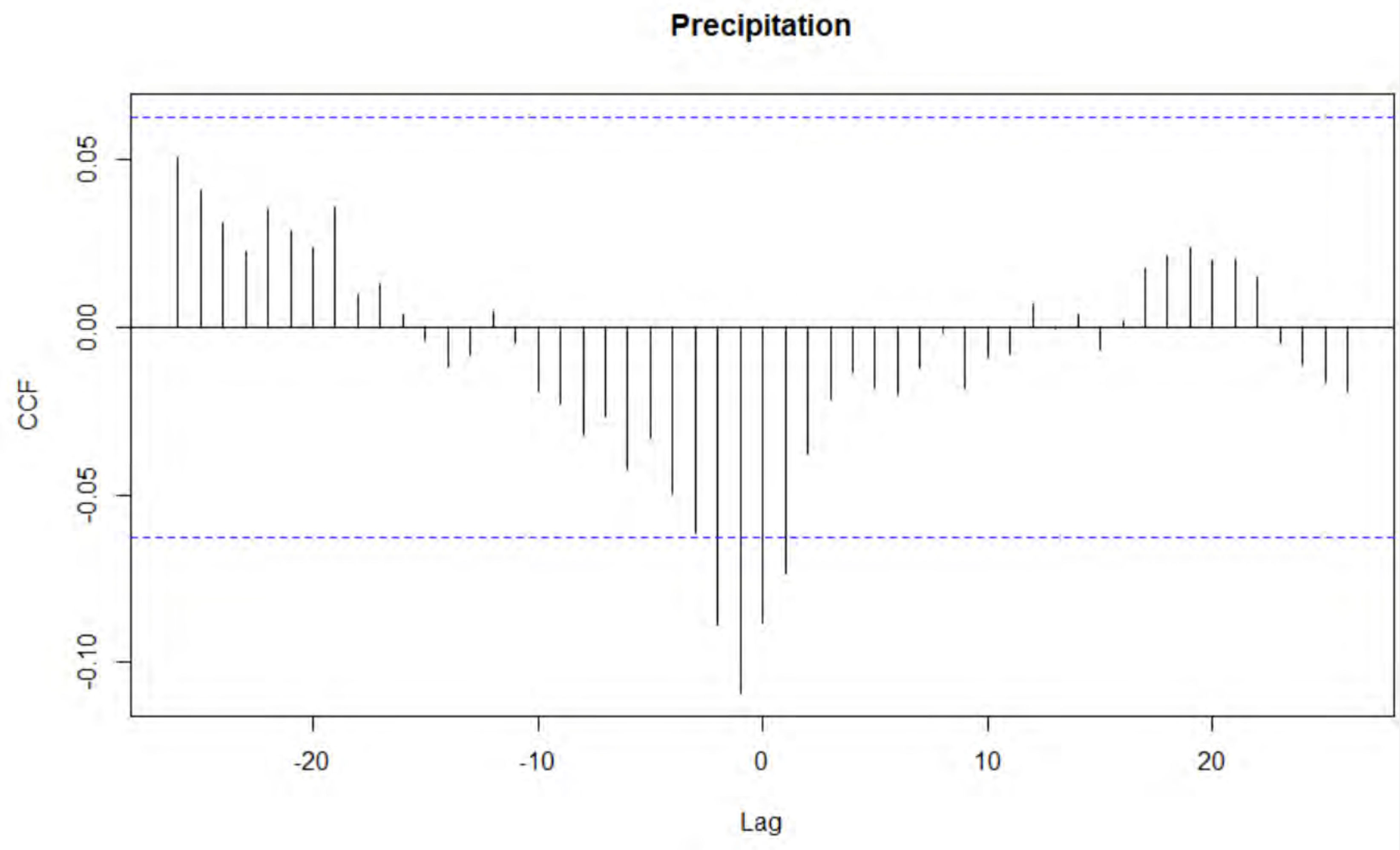
- In terms of the lag operator B, $Bx_t = x_{t-1}$, this becomes

$$\left(1 - \sum_{i=1,\ldots,p} \phi_i B^i\right) x_t = \left(1 + \sum_{i=1,\ldots,q} \theta_i B^i\right) \epsilon_t$$

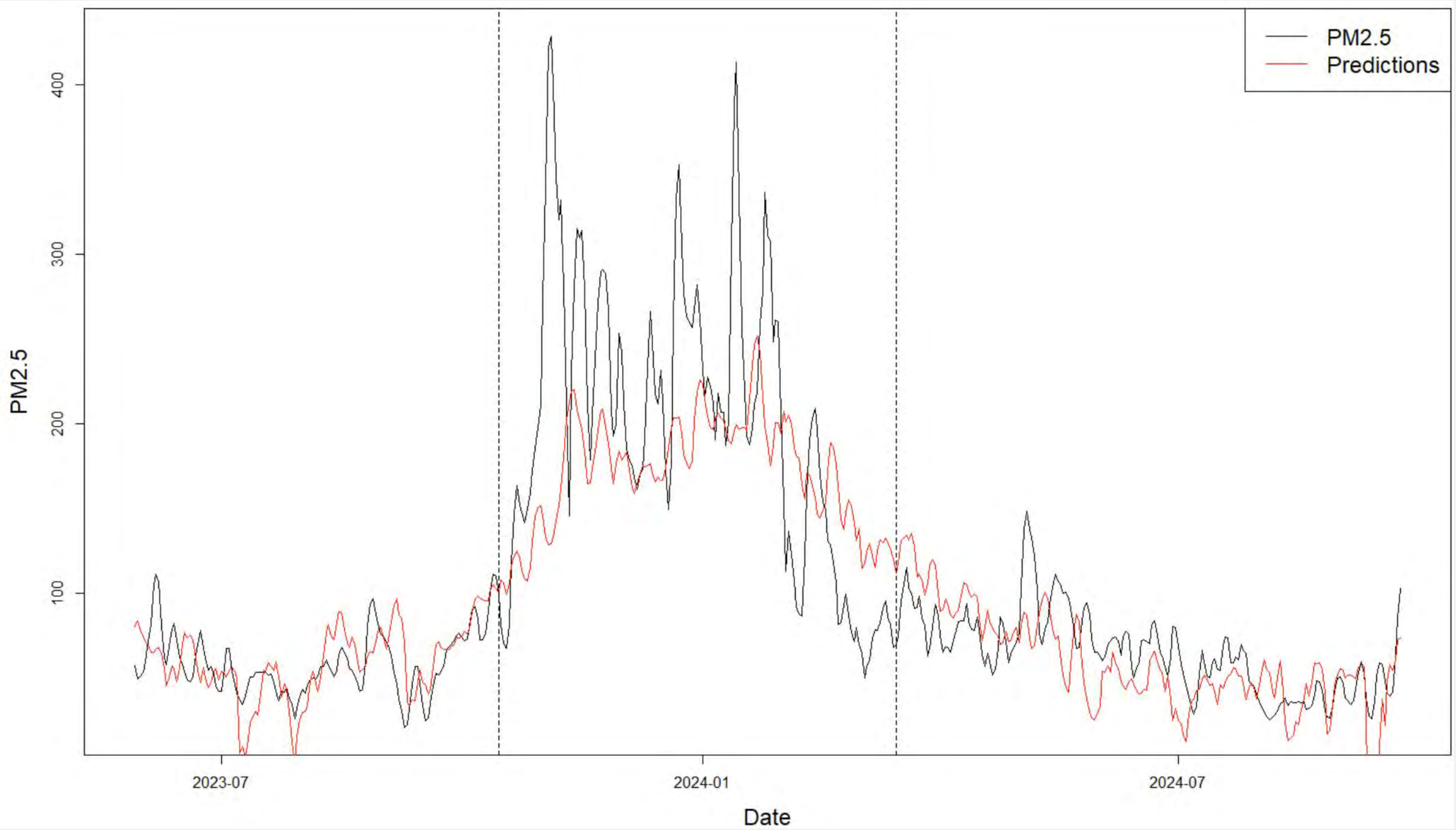
Rearranging $\left(1 + \sum_{i=1,\ldots,q} \theta_i B^i\right)^{-1} \left(1 - \sum_{i=1,\ldots,p} \phi_i B^i\right) x_t \sim WN(\sigma^2)$

CROSS CORRELATION FUNCTION AFTER PRE-WHITENING

- After pre-whitening, the cross-correlation function has a clearer lag structure
- We see significant correlations between PM2.5 and Precipitation lagged at days 1 and 2.
- We can interpret this as a direct cause of lower PM2.5 and include these lagged variables explicitly in our dataset.

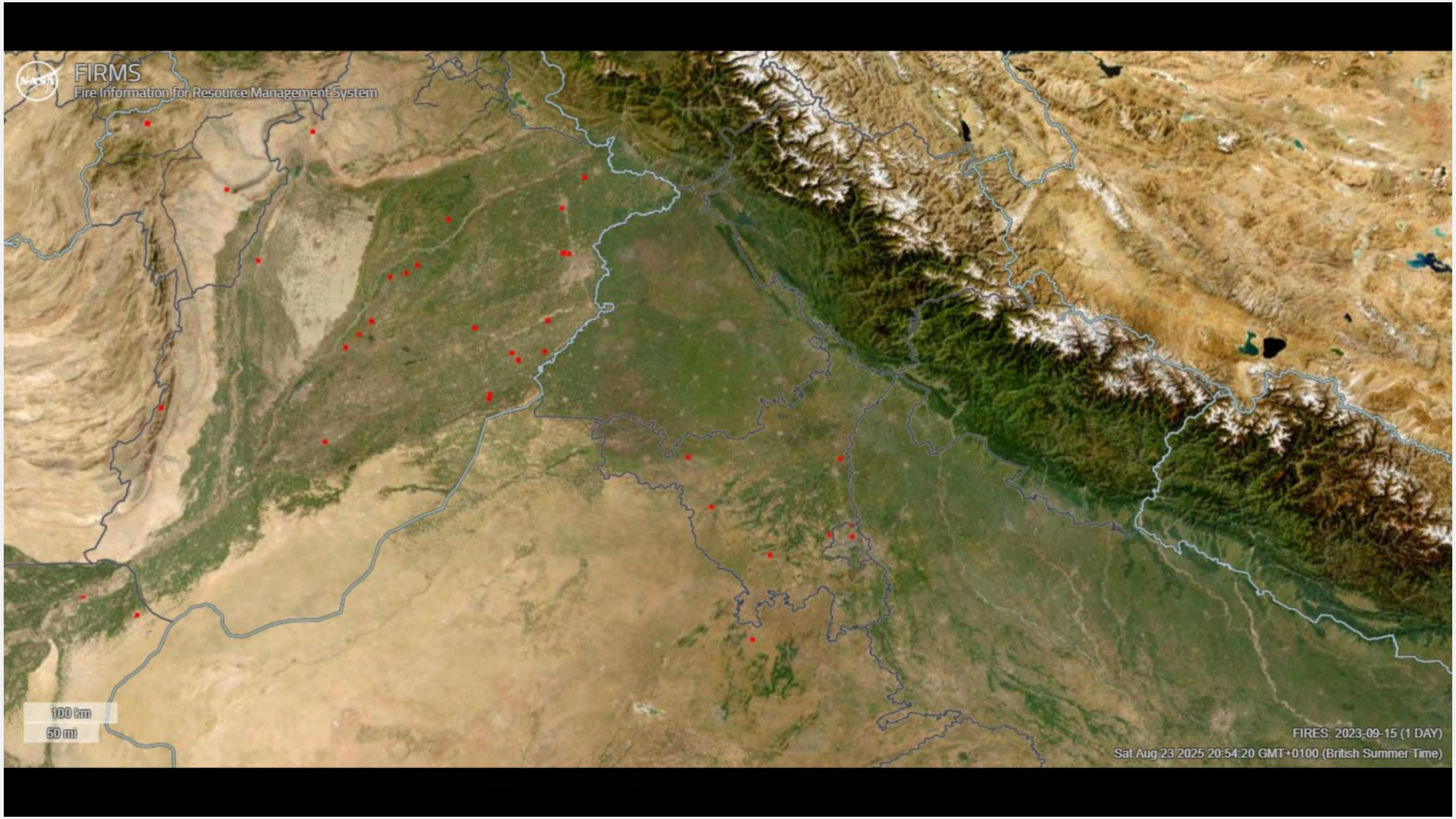


REGRESSING THE DATA ON PM2.5

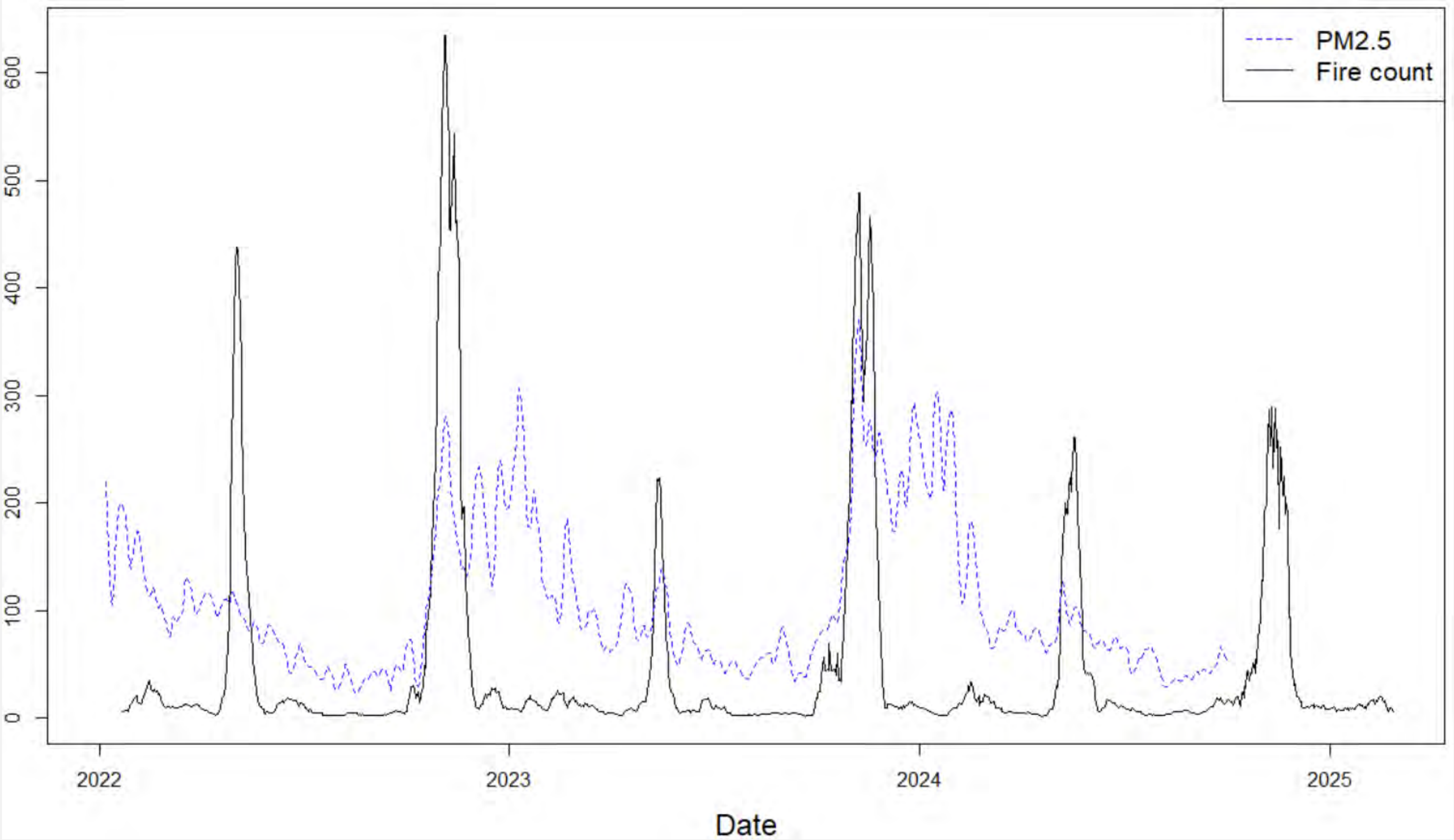


- In the non-winter period, the meteorological variables are effectively capturing the PM2.5 concentrations.
- The rapid spikes of PM2.5 however are not captured by this data, which suggests there are other relevant factors that have not been considered yet.

FIRE COUNT TIME LAPSE – HARYANA/DELHI



INCONSISTENCY IN FIRE COUNT DATA

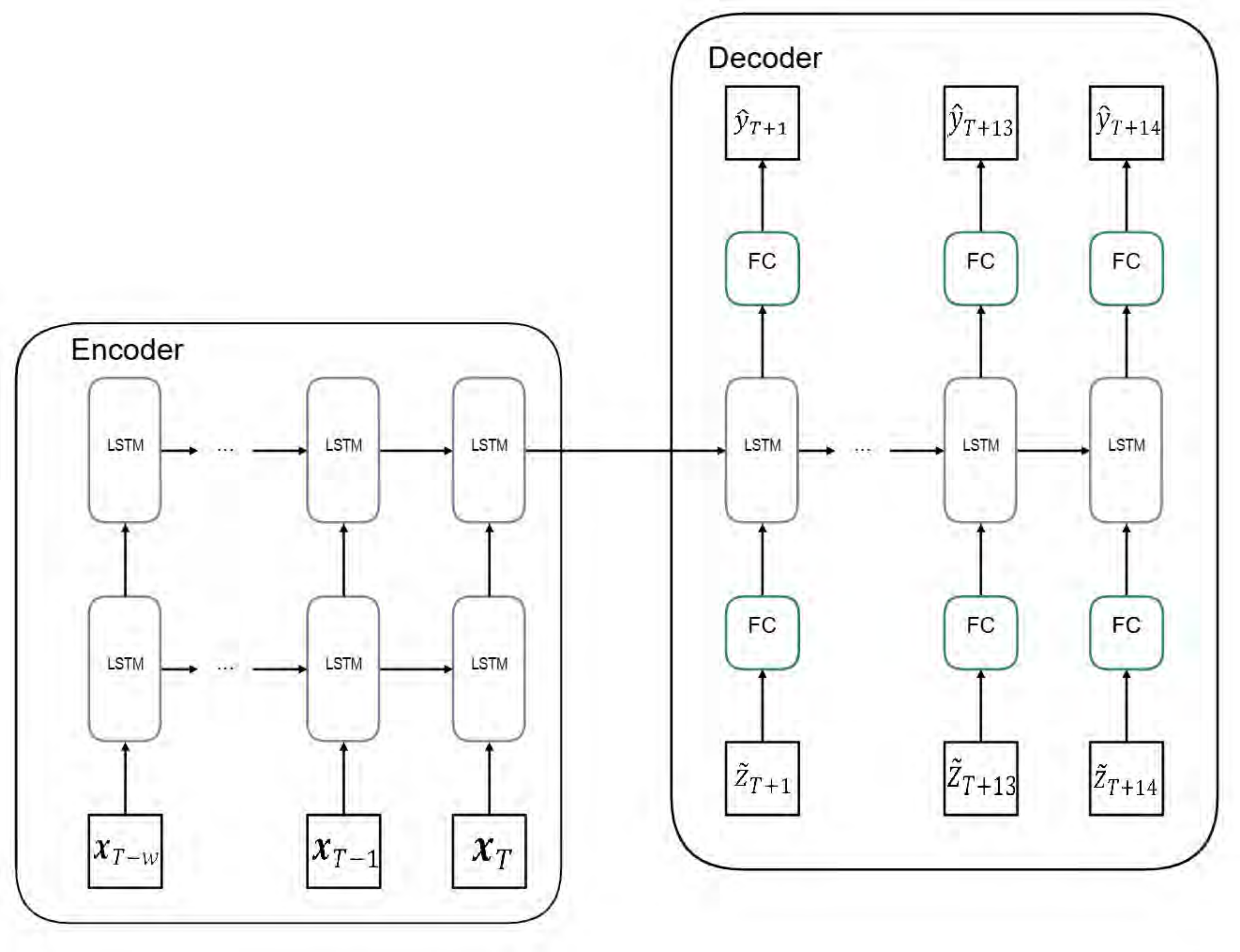


- The fire count is the number of fires detected by the NASA J1VIIRS satellite in the Haryana/Punjab region.
- The first spike in PM2.5 is driven by the rapid increase in the number of fires in this region.
- Notice the number of fires detected in the Winter stubble burning period are declining significantly, despite PM2.5 levels increasing.
- Predicting the initial rise will require a model to be able to be resilient to this misleading fire count data.

MODEL CHOICE – ENCODER DECODER LSTM

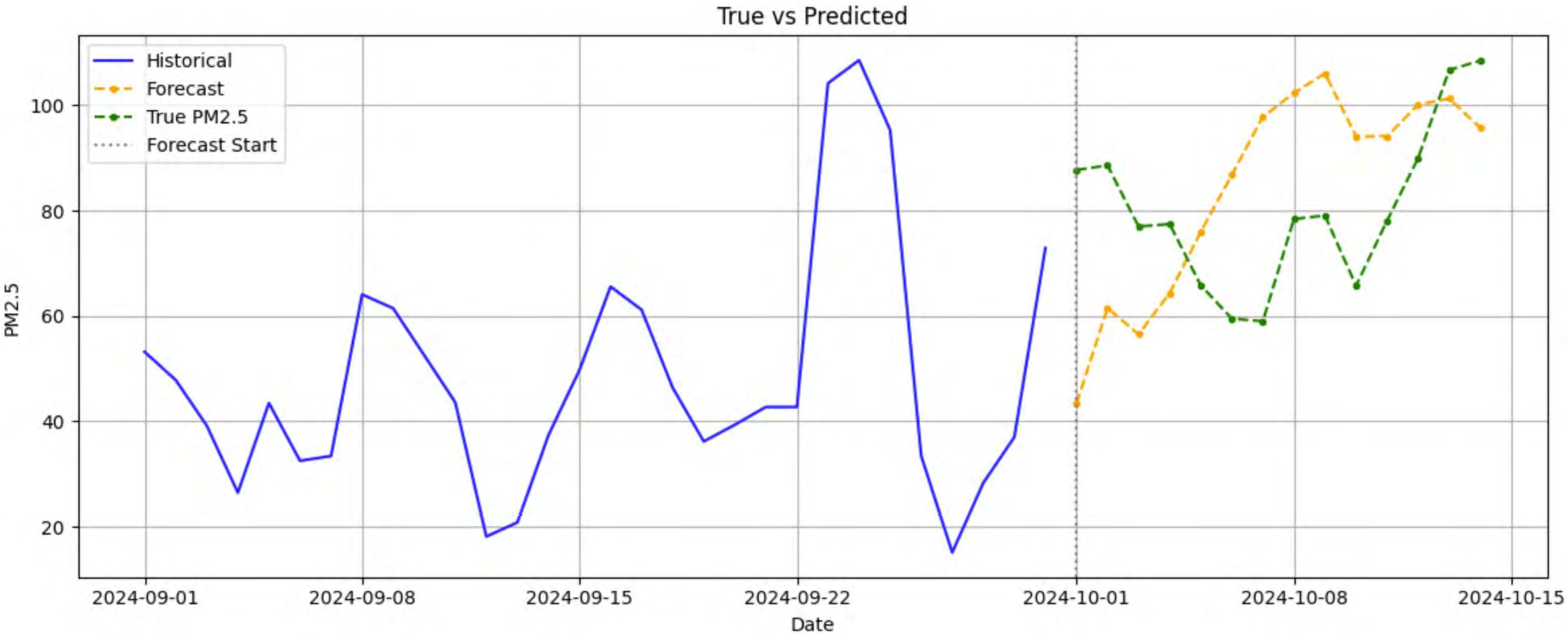
- $\boldsymbol{x}_t = (y_t, z_t, s_t)$
- y_t is the PM2.5 at time t
- z_t are the covariates for which we have future forecasts
- s_t are the covariates for which we don't have forecasts
- $\tilde{z}_t$ are the forecasts.

- The model is trained to minimise the mean square error loss between the predicted and true values for PM2.5.



Model Results

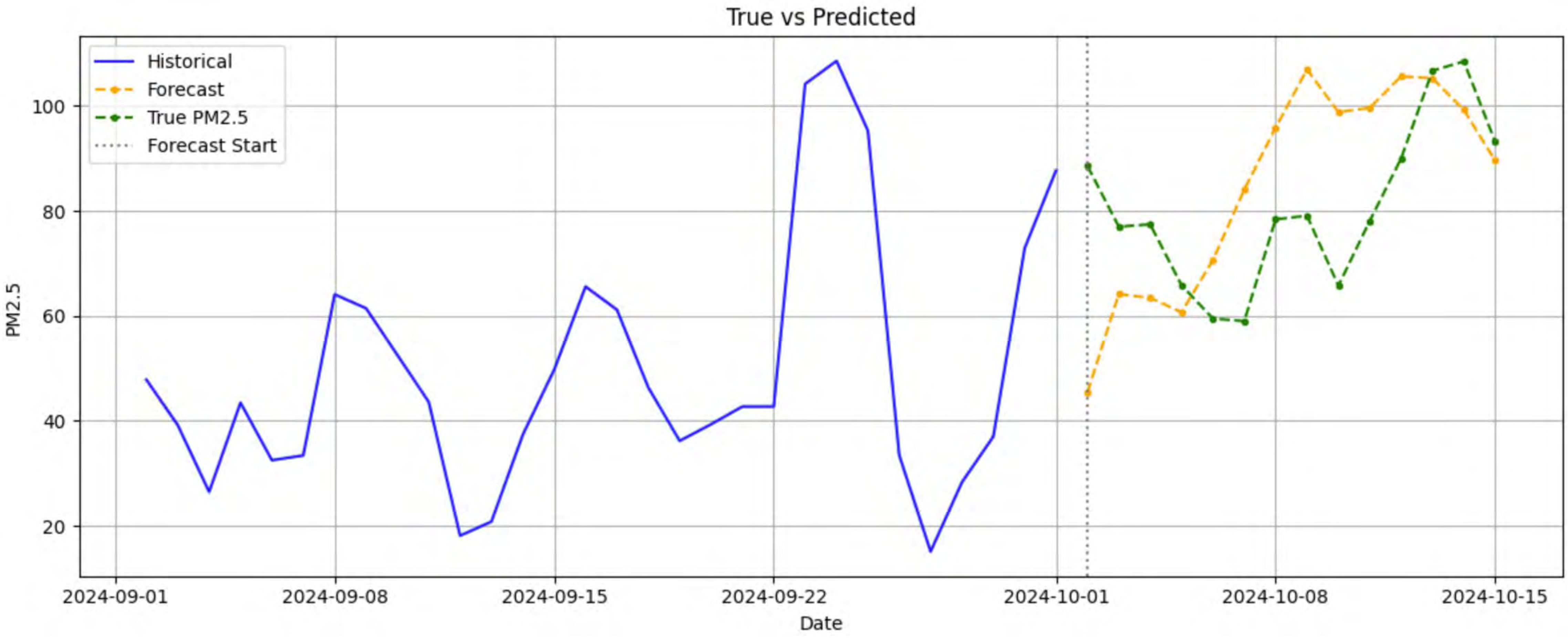
14 DAY ROLLING WINDOW FORECAST 1ST OCT 2024



Immediate horizons not necessarily more accurate than further out ones

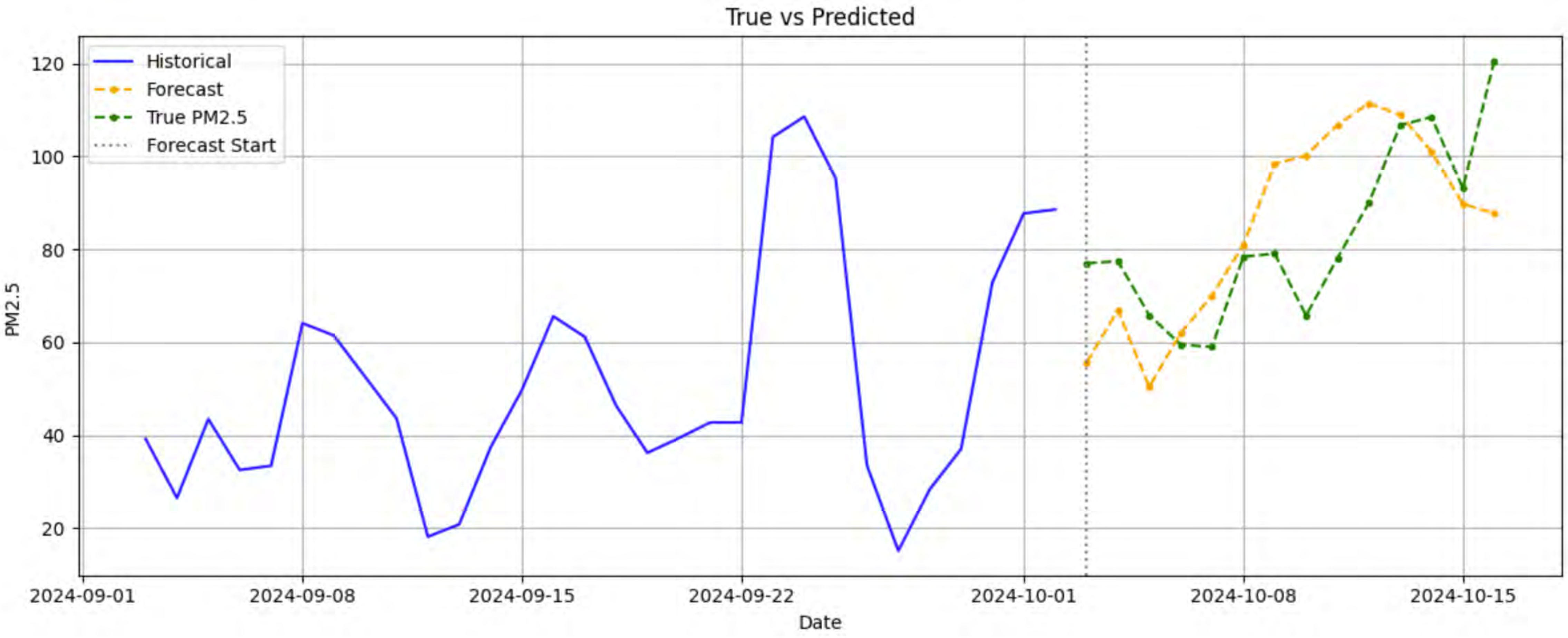
Model Results

14 DAY ROLLING WINDOW FORECAST 2ND OCT 2024



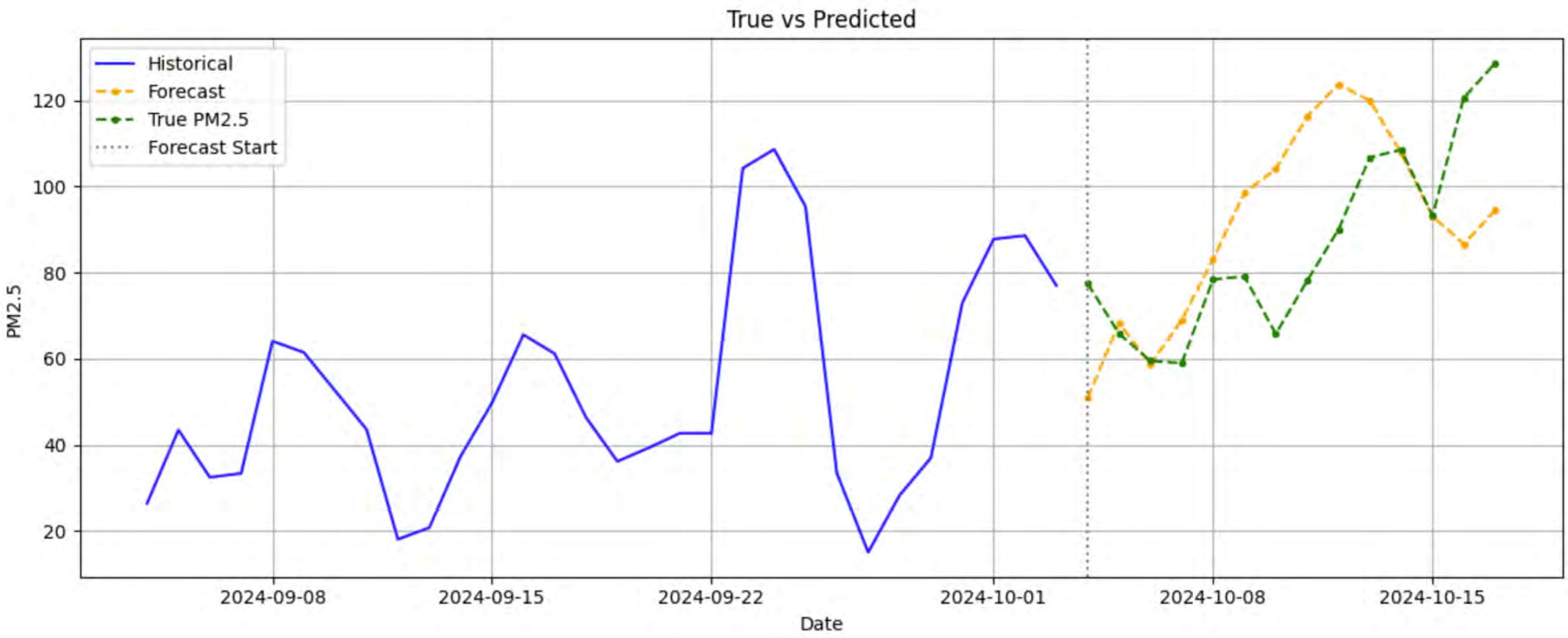
Model Results

14 DAY ROLLING WINDOW FORECAST 3RD OCT 2024



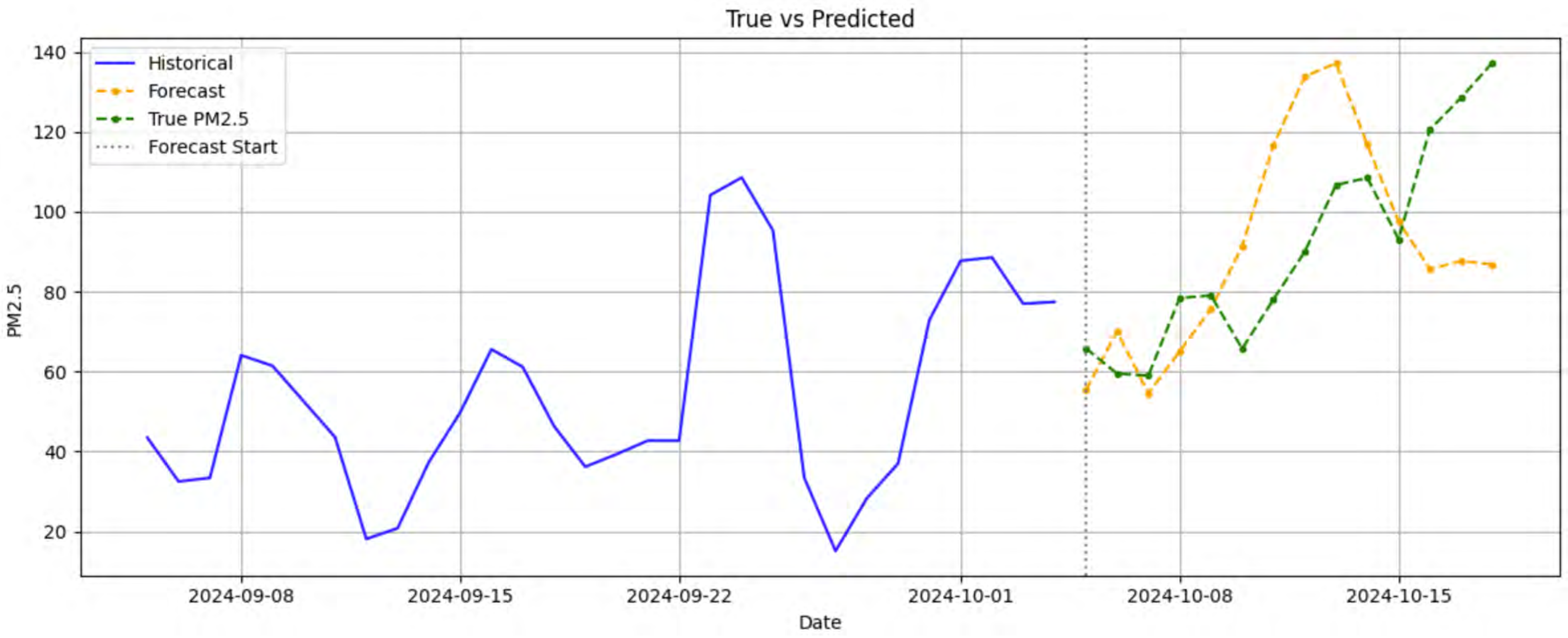
Model Results

14 DAY ROLLING WINDOW FORECAST 4TH OCT 2024



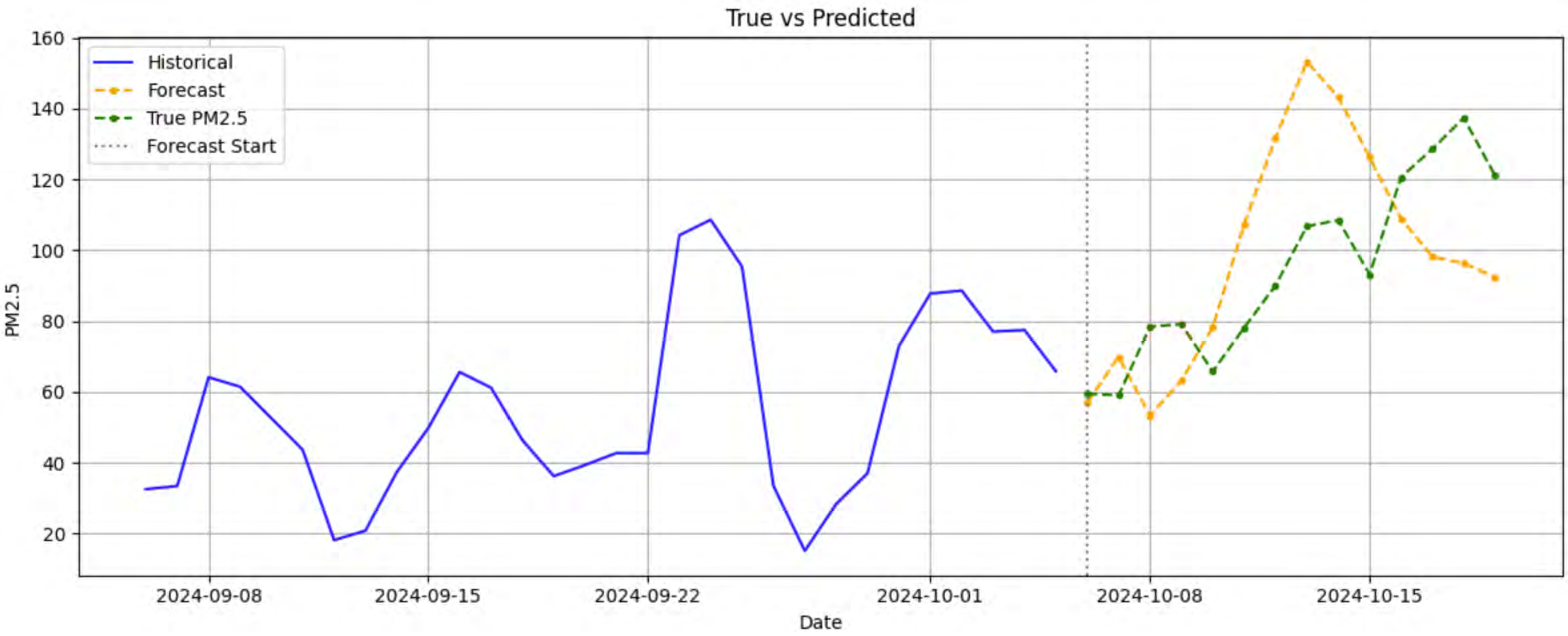
Model Results

14 DAY ROLLING WINDOW FORECAST 5TH OCT 2024

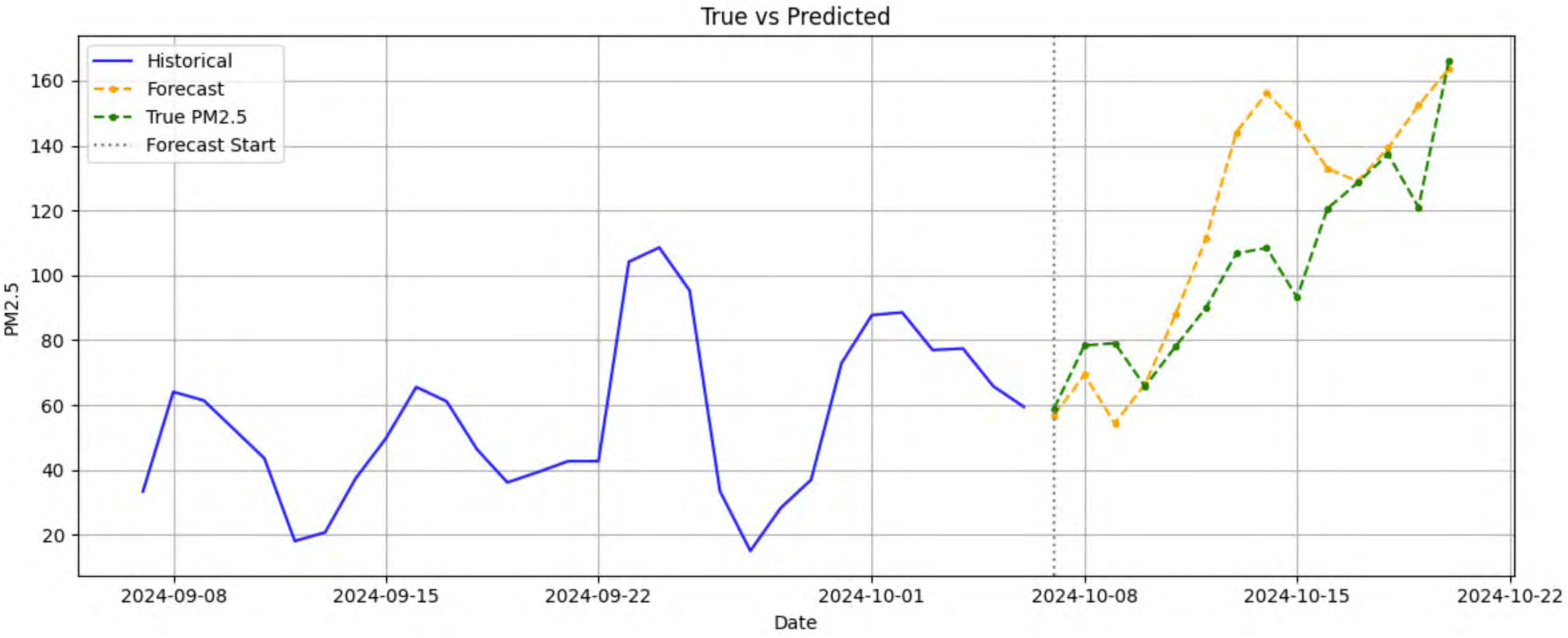


Consecutive 14 day horizons be drastically different, which needs to be addressed

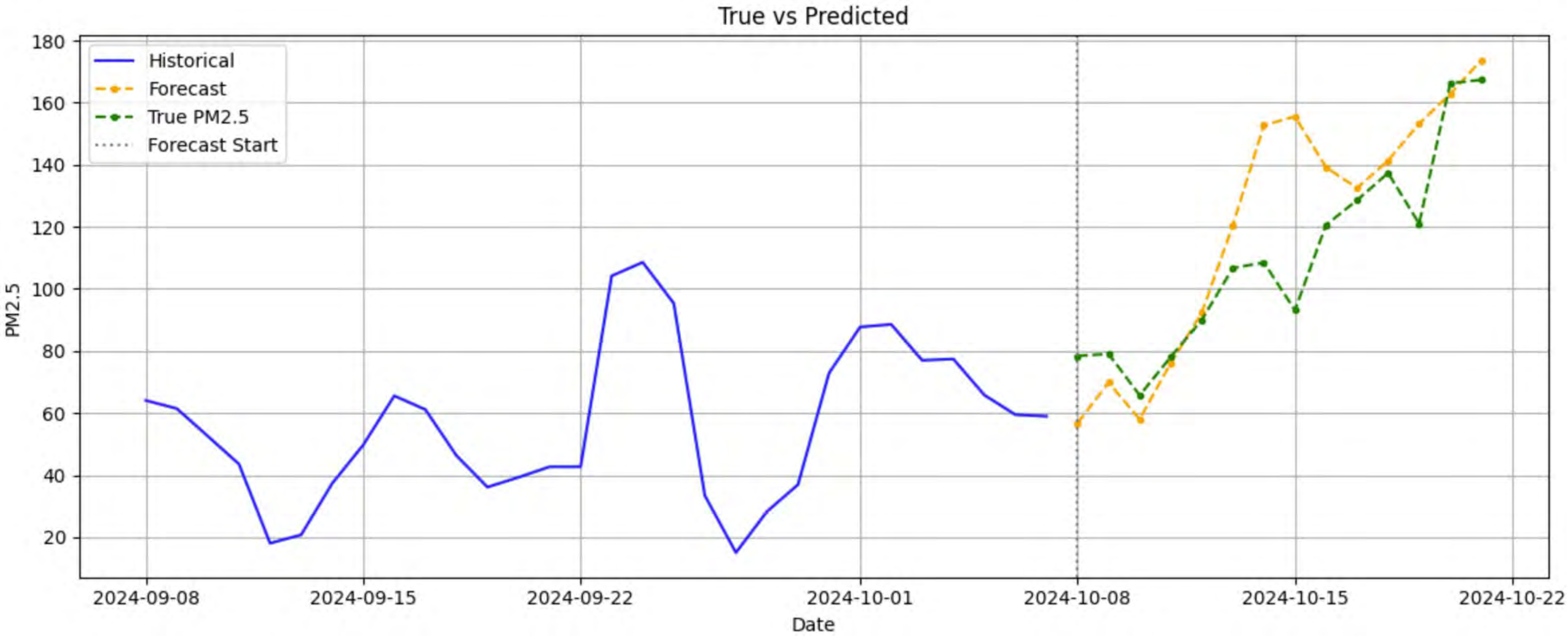
14 DAY ROLLING WINDOW FORECAST 6TH OCT 2024



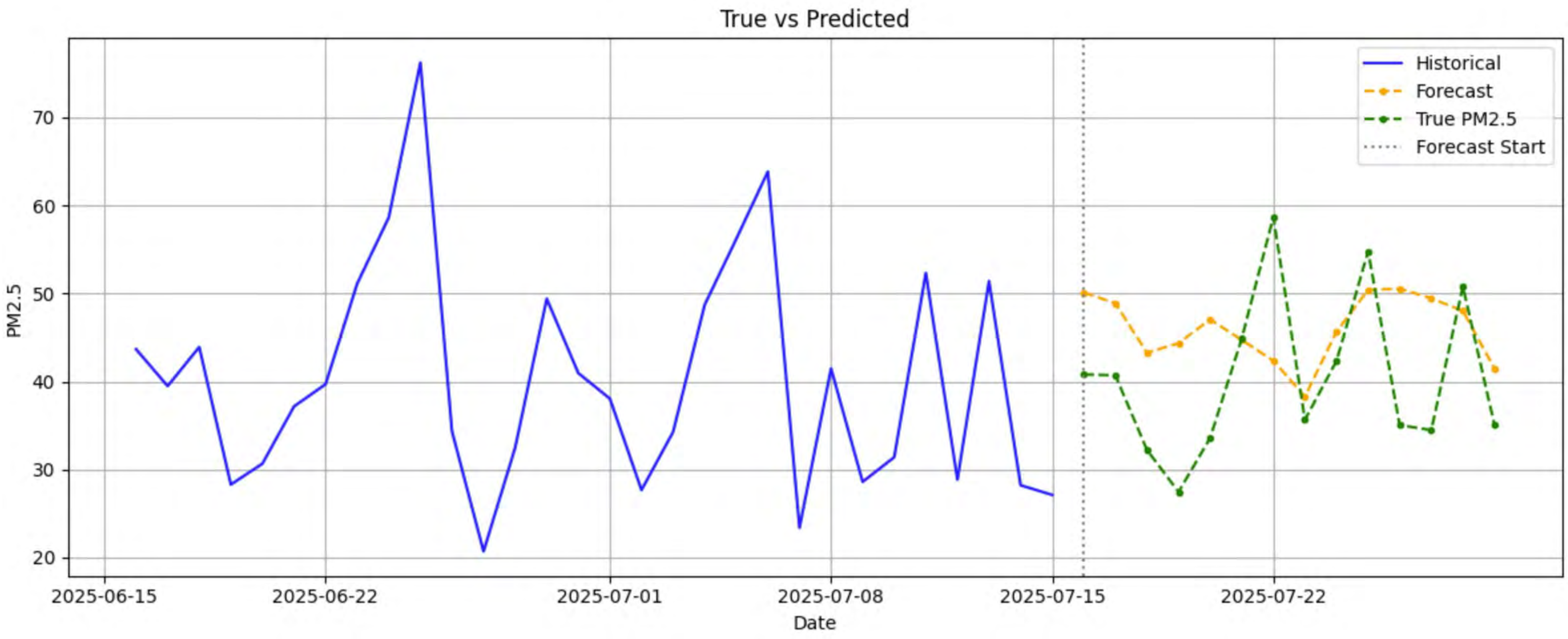
14 DAY ROLLING WINDOW FORECAST 7TH OCT 2024



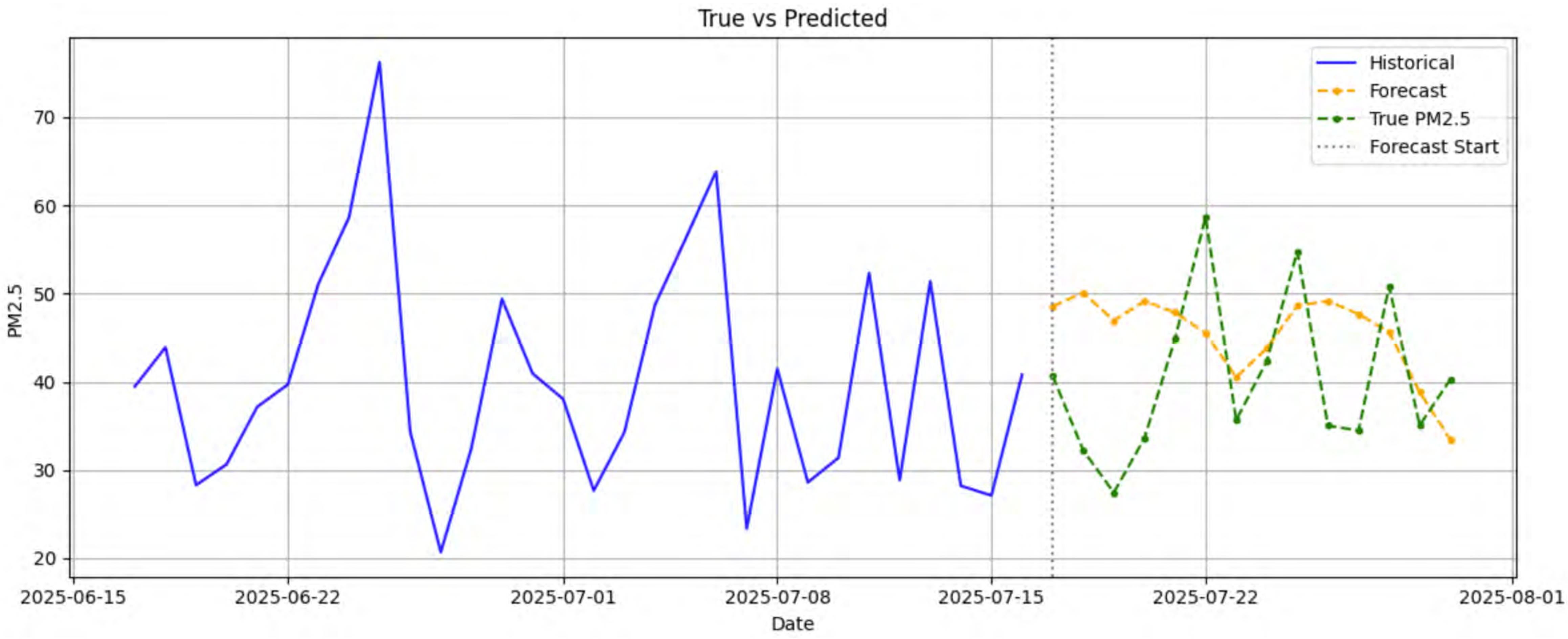
14 DAY ROLLING WINDOW FORECAST 8TH OCT 2024



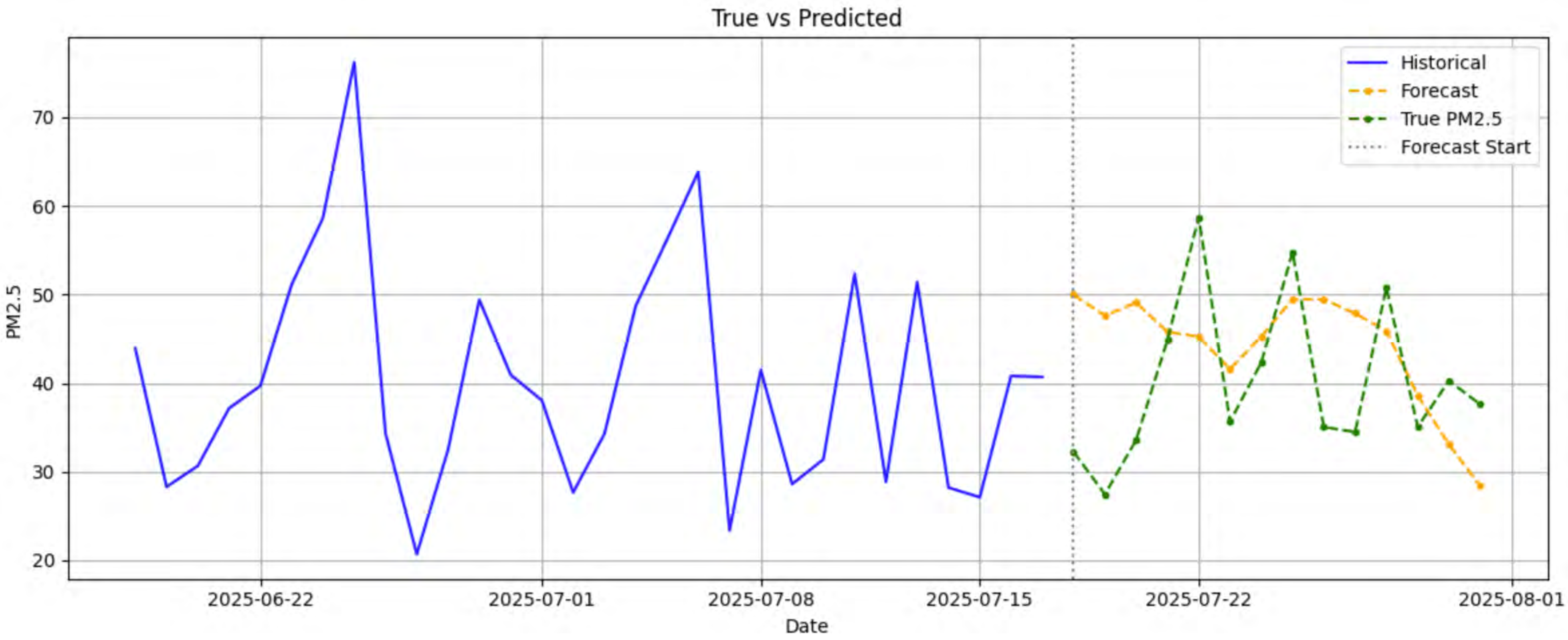
14 DAY ROLLING WINDOW FORECAST 16TH JULY 2025



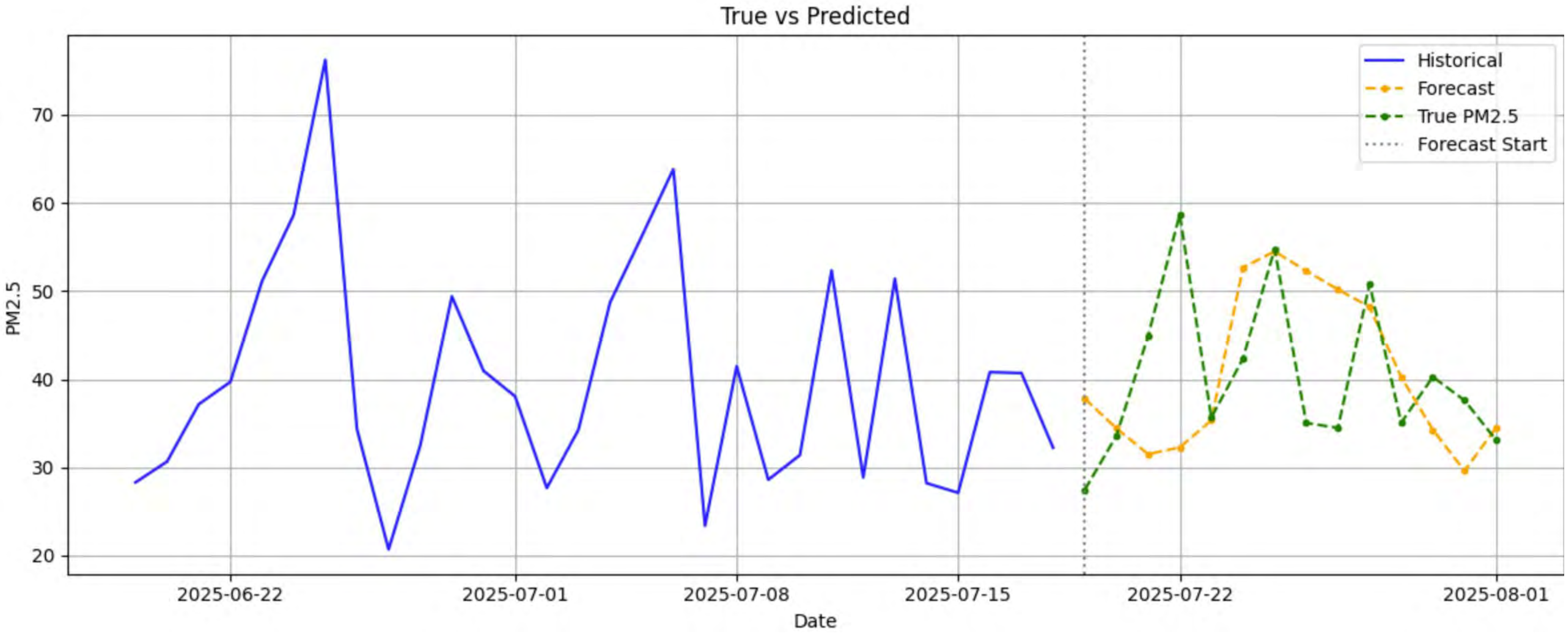
14 DAY ROLLING WINDOW FORECAST 17TH JULY 2025



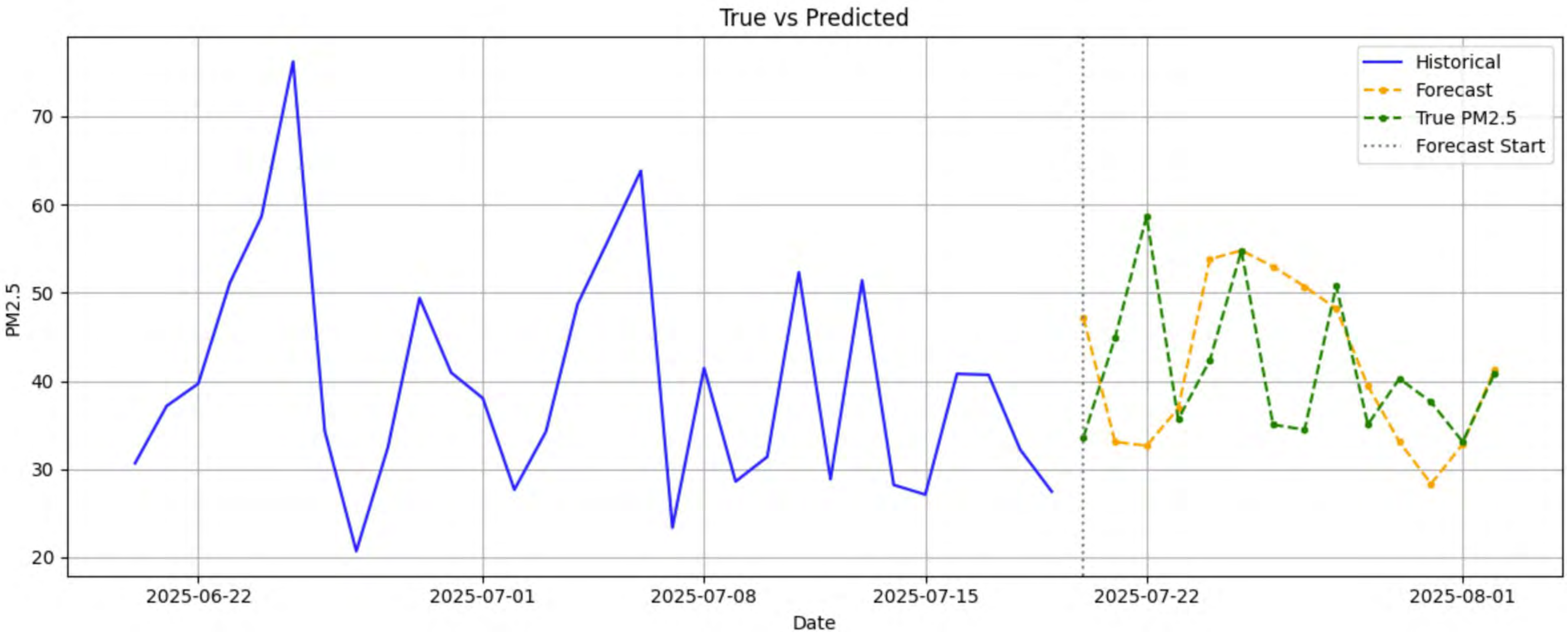
14 DAY ROLLING WINDOW FORECAST 18TH JULY 2025



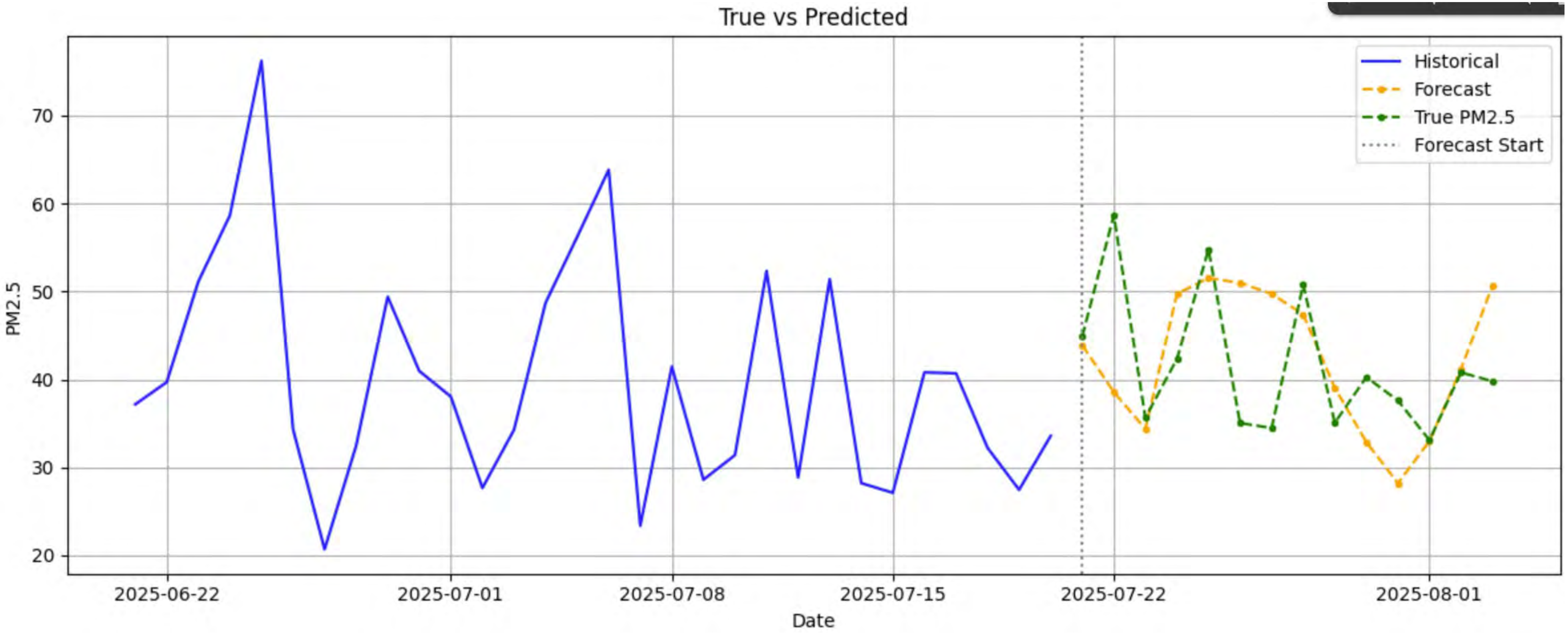
14 DAY ROLLING WINDOW FORECAST 19TH JULY 2025



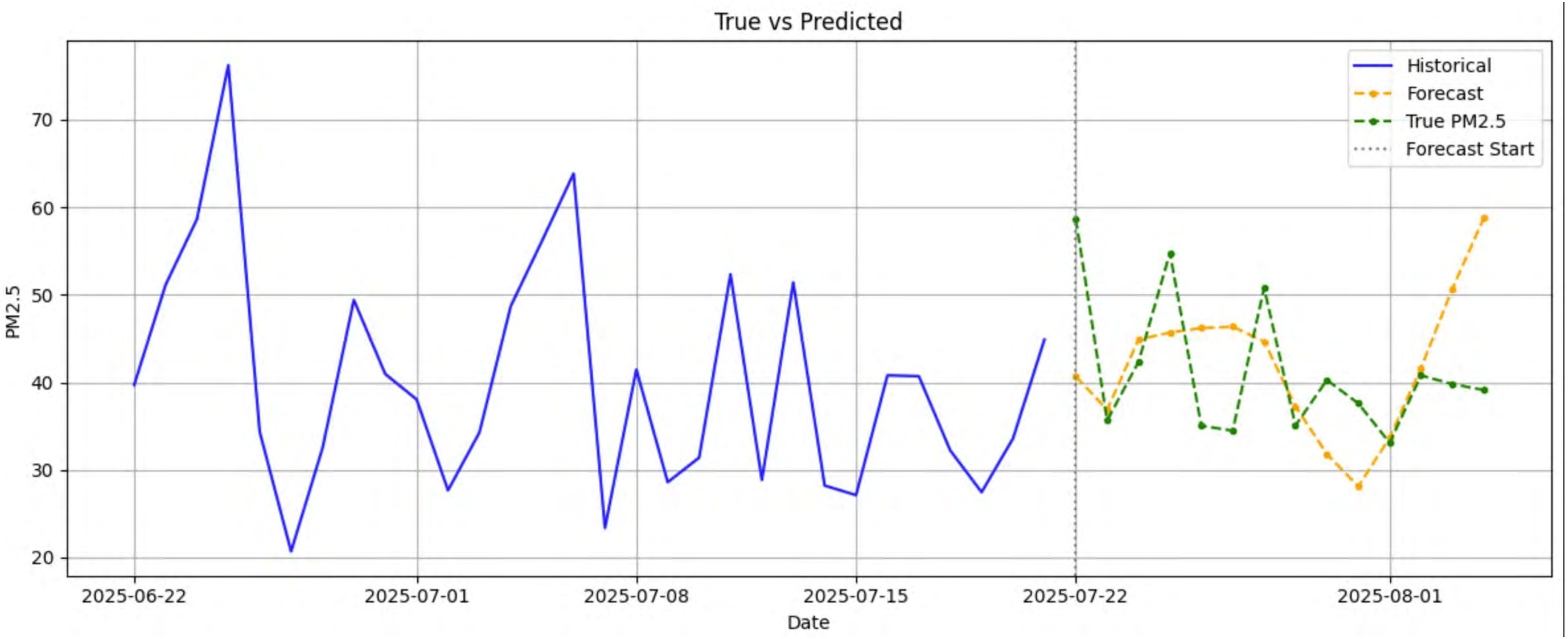
14 DAY ROLLING WINDOW FORECAST 20TH JULY 2025



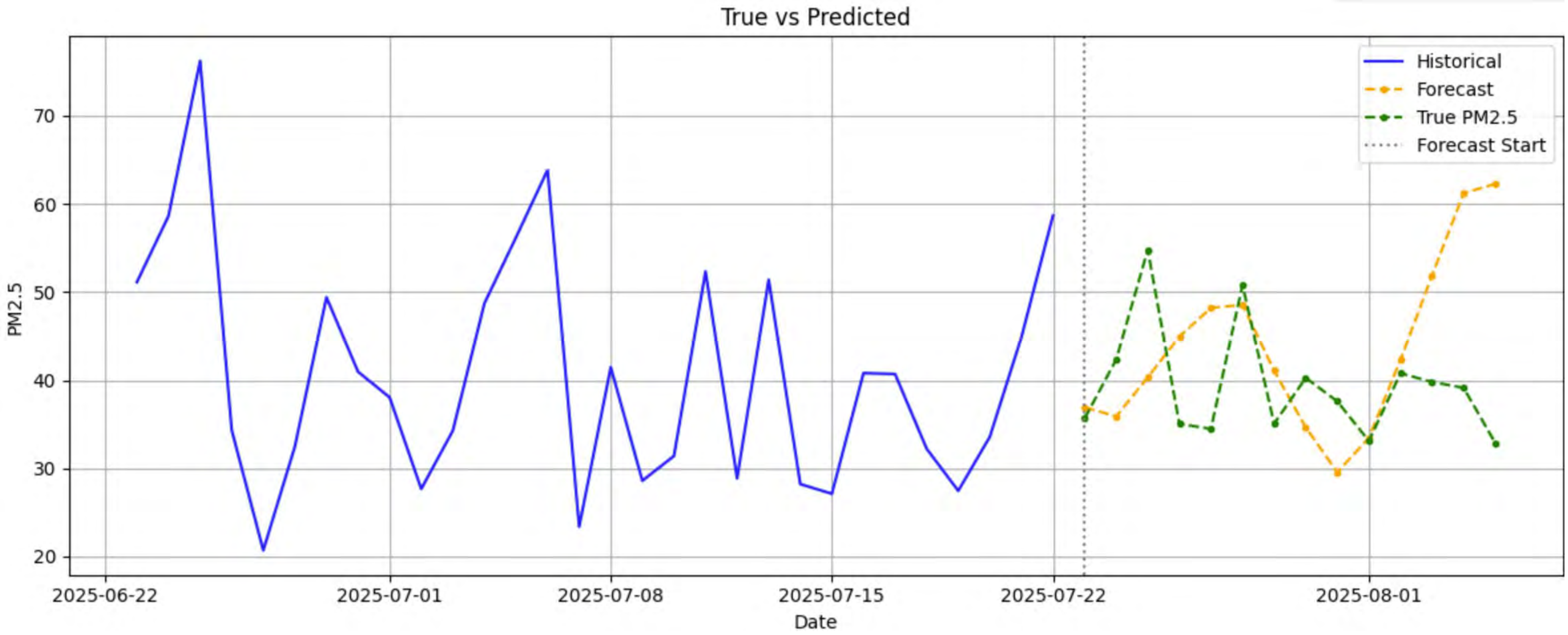
14 DAY ROLLING WINDOW FORECAST 21ST JULY 2025



14 DAY ROLLING WINDOW FORECAST 22ND JULY 2025



14 DAY ROLLING WINDOW FORECAST 23RD JULY 2025



Some larger error are present that need to be fixed

FUTURE/CURRENT DIRECTIONS

- Transfer learning
- Temporal Fusion Transformer (interpretable and designed to handle long term dependencies)
- Ensemble model – combine predictions from other models
- Find more explanatory variables e.g. including concentrations of other pollutants