Dengue Outbreak Climate Analysis

Lopamudra Biswal, 23b0049

November 27, 2025

Contents

1 Introduction

This analysis explores relationships between meteorological conditions and dengue outbreaks in New Delhi. Data is downloaded form ERA5-Land reanalysis (daily) and lag windows are created prior to each outbreak, environmental suitability indices are computed and analysis is done on time series basis.

2 Data

2.1 Datasets used

- ERA5-Land daily variables: 2m temperature (t2m), total precipitation (tp), 2m dewpoint temperature (d2m).
- Dengue outbreak dataset: new_delhi_dengue_data.csv (contains year, week_of_outbreak, Cases, latitude/longitude, etc.)
- MODIS LAI: Delhi_LAI_8day.csv (8-day MODIS LAI time series, later interpolated to daily)

2.2 Locations and temporal coverage

The ERA5 area bounding box used:

```
area = [28.8, 77.0, 28.4, 77.4]  % approximate New Delhi box
```

3 Preprocessing

3.1 ERA5 daily aggregation and units

Key conversions and derivations:

- Temperature: t2m (K) is converted to Celsius: T(°C) = t2m - 273.15.
- Dewpoint: d2m (K) to °C: Td(°C) = d2m - 273.15.
- Precipitation: tp (m) to mm: precip_mm = tp * 1000.
- Relative humidity (RH) estimated using the Magnus formula from T and Td.

3.2 Outbreak dates

Outbreaks in the dataset are converted from year + ISO week number to a Monday date using:

```
datetime.fromisocalendar(year, weeknum, 1)
```

Each outbreak's environmental window is defined for the 56 days prior to the outbreak date

4 Feature engineering

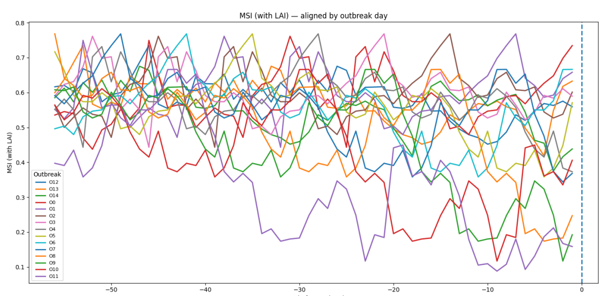
For each day in each outbreak window computed parameters are:

- 3-day rolling precipitation standard deviation: precip_roll3_std
- 7-day precipitation sum: precip_7sum
- A simple "stagnation" flag: precip_7sum $<$ median(precip_7sum)
- Suitability components:
  - temp_suit(T) = clip(1 - $((T - 28)/8)^2, 0, 1$)
  - rain_suit(p) = clip(p / 30, 0, 1)
  - rh_suit(r) = clip((r - 50) / 50, 0, 1)
  - lai_suit(l) = clip(l / 5, 0, 1) (LAI normalized to 0–5 typical MODIS range)
- MSI (original): MSI = 0.4*temp_suit + 0.3*rain_suit + 0.3*rh_suit
- MSI_new (with LAI): MSI_new = 0.4*temp_suit + 0.2*rh_suit + 0.2*rain_suit + 0.2*lai_suit

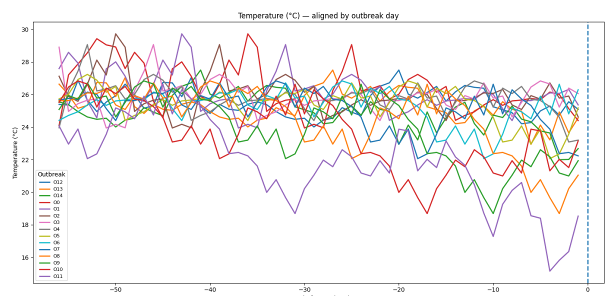
5 Visualizations

- Outbreak-specific time series: outbreak_plots/outbreak_{id}_{date}_var.png
- Summary plots with LAI: outbreak_plots_lai/outbreak_{id}_{date}_LAI.png
- Correlation matrix: /content/outbreak_analysis_outputs/plots_corr/correlation_matrix.png

Below are example figure placeholders embedded in the report. Replace the paths if your outputs differ.



(a) Example outbreak LAI–MSI time series



(b) Example outbreak temperature time series

Figure 1: Sample outbreak plots

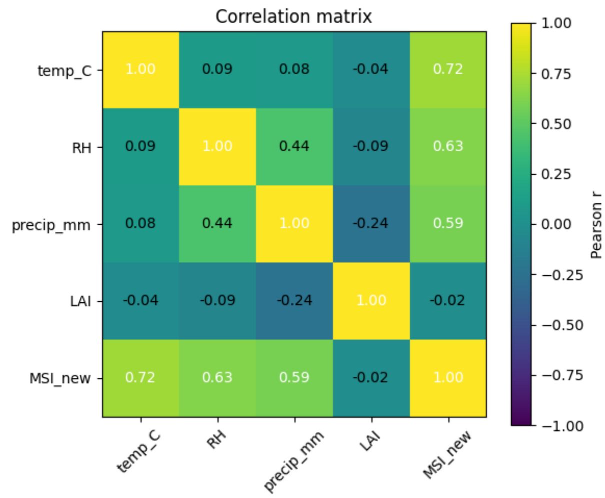


Figure 2: Feature correlation matrix

6 Key tables

A short variable reference:

Variable	Description
temp_C	2m temperature in °C
RH	Relative humidity (percent) estimated from T and Td
precip_mm	Daily precipitation in mm
LAI	Leaf Area Index (MODIS) interpolated to daily
MSI	Mosquito Suitability Index (original)
MSI_new	MSI including LAI weighting
outbreak_date	Monday date of outbreak week (ISO week)
outbreak_idx	outbreak identifier from source CSV

7 Causality Analysis Pipeline

To investigate whether environmental variables exhibit predictive or causal influence on dengue outbreak risk, we performed a full causality analysis using smoothed outbreak signals, Granger causality testing, and correlation structure evaluation. The entire pipeline is described below.

7.1 Data Preparation

The outbreak indicator was constructed as:

$$\text{outbreak\_flag}(t) = \begin{cases} 1, & \text{if an outbreak occurs on day } t, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Since daily outbreak values are sparse and discrete, a smoothed outbreak signal was required for time-series causality analysis. A centered 14-day moving average was applied:

$$\text{outbreak\_smooth}(t) = \frac{1}{15} \sum_{i=-7}^{+7} \text{outbreak\_flag}(t+i) \tag{2}$$

This smoothing produces a continuous signal that reflects outbreak intensity over time.

7.2 Granger Causality Testing

Granger causality determines whether past values of an environmental variable improve the prediction of the outbreak signal. For each variable X_t, the model tests:

$$X \Rightarrow \text{outbreak\_smooth}$$

using lags $1, 7, 14, 21, 30$.

The testing procedure:

```
result = grangercausalitytests(data, maxlag=30, verbose=False)
p = result[lag][0]["ssr_chi2test"][1]
```

If $p < 0.05$, the variable is considered Granger-causal at that lag.

Table 1: Granger Causality p-values for predicting *outbreak_smooth*

Variable	Lag 1	Lag 7	Lag 14	Lag 21	Lag 30
temp_C	0.98378	0.07781	**0.00218**	**0.00099**	**0.00000**
RH	0.62650	0.13610	0.13477	0.09479	0.11791
precip_mm	0.08155	**0.00248**	**0.01113**	**0.04942**	0.25634
LAI	0.36762	**0.01487**	**0.00010**	**0.00042**	**0.00000**
MSI_new	0.34407	0.07728	**0.01756**	**0.04540**	0.05542

7.3 Correlation Structure

To understand relationships among variables, a Pearson correlation matrix was generated:

$$\text{Corr}(X_i,\ X_j) = \frac{\text{cov}(X_i, X_j)}{\sigma_{X_i}\sigma_{X_j}}$$

The following environmental variables were included:

- Temperature (`temp_C`)
- Relative humidity (`RH`)
- Rainfall (`precip_mm`)
- Leaf Area Index (`LAI`)
- Mosquito Suitability Index (`MSI_new`)

The heatmap visualizes positive and negative linear dependencies which help interpret climatic interactions prior to causality tests.

8 LSTM-Based Outbreak Prediction Pipeline

We constructed a supervised learning framework to predict disease outbreaks using environmental and remote-sensing variables. The full pipeline consists of data preprocessing, window generation, sequence construction, model training, and evaluation.

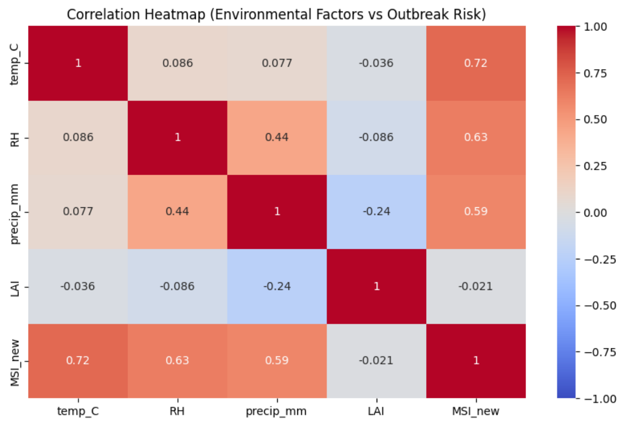


Figure 3: Correlation heatmap

8.1 Data Preprocessing

Daily data were loaded from the processed dataset and reindexed to a continuous timeline. Missing values were forward-filled. Known outbreak dates were encoded, and for each outbreak date d, a 56-day lookback window was extracted:

$$W_d = \{\, x(t) \mid t \in [d - 56,\, d] \,\}.$$

Each positive window was assigned a label $y = 1$. An equal number of non-outbreak windows were sampled and labeled $y = 0$, creating a balanced dataset.

8.2 Sequence Construction

For every window, only numerical variables were retained and standardized using a z-score transformation:

$$x' = \frac{x - \mu}{\sigma}.$$

Each standardized window forms a multivariate sequence of length 57, which is used as input to the LSTM:

$$X = \{x'(t), x'(t+1), \ldots, x'(t+56)\}.$$

8.3 LSTM Architecture

A two-layer LSTM network was used for binary classification. The final architecture is:

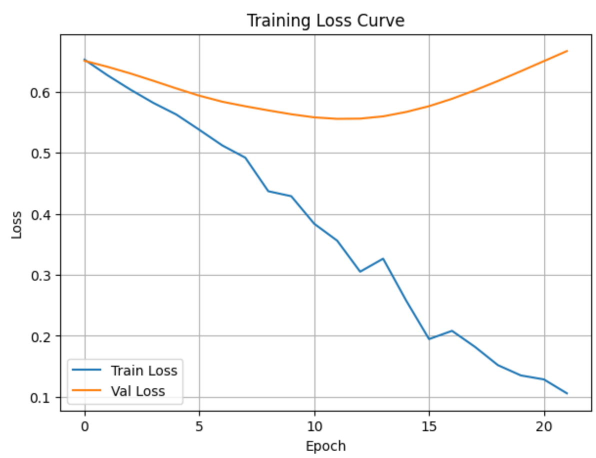
- LSTM(64) with return sequences
- Dropout(0.3)
- LSTM(32)
- Dropout(0.3)
- Dense(16, ReLU)
- Dense(1, Sigmoid)

The model was trained using the Adam optimizer and binary cross-entropy loss with early stopping to prevent overfitting.

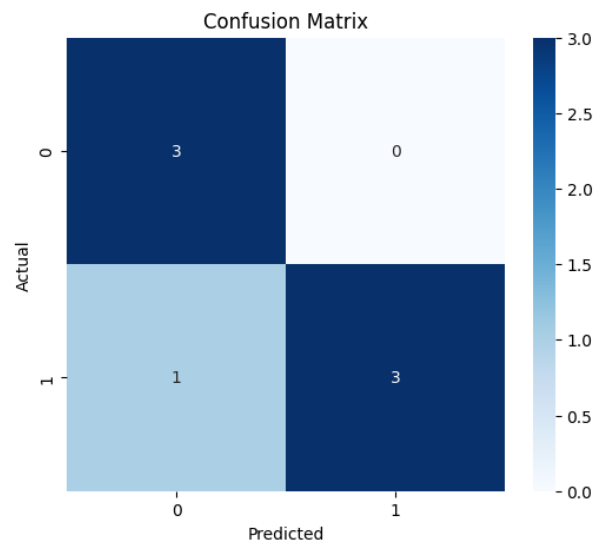
8.4 Model Evaluation

The dataset was split into 75% training and 25% testing. We evaluated the trained model using:

- training and validation loss curves,
- training and validation accuracy curves,
- confusion matrix,
- ROC curve and AUC score,
- precision, recall, and F1-score.



(a) Example outbreak LAI–MSI time series



(b) Example outbreak temperature time series

Figure 4: Sample outbreak plots

Table 2: Classification Report for LSTM Outbreak Prediction

Class	**Precision**	**Recall**	**F1-score**	**Support**
0 (No Outbreak)	0.75	1.00	0.86	3
1 (Outbreak)	1.00	0.75	0.86	4
Accuracy		0.86		7
Macro Avg	0.88	0.88	0.86	7
Weighted Avg	0.89	0.86	0.86	7