Climate Anomaly Detection in Sri Lankan Weather Data Using Unsupervised Machine Learning

*Data Mining Individual Project - Final Report

Deeghayu Adhikari
Department of Computer Science and Engineering
University of Moratuwa
Moratuwa, Sri Lanka
nethmina.25@cse.mrt.ac.lk

***Abstract*—With increasing concerns about global climate change, the ability to identify localized, extreme weather events from vast datasets has become critical for risk assessment, adaptation strategies, and effective policy-making. This study presents a comprehensive data mining approach to detect climate anomalies using a granular Sri Lankan weather dataset spanning from 2010 to 2023. Following a meticulous data pre-processing and feature engineering phase, which included the creation of time-aware features like rolling averages and lagged variables, an unsupervised Isolation Forest model was deployed. The model analyzed 14 distinct meteorological variables across 30 cities to isolate statistically significant deviations from normative weather patterns. The algorithm successfully identified 20,065 anomalous daily records, constituting approximately 13.6% of the total data. Key findings reveal a distinct geographical and temporal clustering of these anomalies, with cities in the central highlands (Hatton, Badulla, Kandy) being disproportionately affected. The years 2014, 2010, and 2016 exhibited the highest frequency of such events. A detailed characterization of these anomalies indicates they are primarily defined by extreme precipitation, higher wind speeds, and greater temperature variance, findings that strongly align with anticipated indicators of regional climate change.**

***Index Terms*—Anomaly Detection, Isolation Forest, Climate Change, Time Series Analysis, Unsupervised Learning, Sri Lanka Weather**

I. Introduction & Problem Statement

The tangible impacts of global climate change are most acutely observed at a regional level through an increased frequency and intensity of anomalous weather events. For an island nation like Sri Lanka, characterized by a diverse geography that ranges from low-lying coastal plains to mountainous central highlands, vulnerability to such events is particularly pronounced and multifaceted. The nation's economy is heavily reliant on climate-sensitive sectors such as agriculture (tea, rubber, rice) and tourism, both of which can be devastated by unpredictable and extreme weather.

While traditional climatology often focuses on analyzing long-term trends in average temperature or precipitation, this approach can obscure the short-term, high-impact events that pose the most immediate threat. A data mining approach, in contrast, offers the capability to automatically sift through massive temporal datasets to pinpoint specific instances of deviation from the norm, thereby providing granular insights into climate instability.

This project addresses the challenge of identifying these non-obvious, multi-variate anomalies within a large-scale, high-dimensional temporal weather dataset. The primary research question is: **Can an unsupervised anomaly detection algorithm effectively identify and characterize statistically significant deviations in daily weather patterns across Sri Lanka, and do these detected anomalies align with expected climate change indicators such as increased extremity and variability?**

The key objectives of this study are:

1) To pre-process and enrich a large-scale weather dataset with time-aware features suitable for temporal anomaly detection.
2) To implement and tune an efficient, unsupervised machine learning model (Isolation Forest) capable of handling high-dimensional data without prior labeling.
3) To identify and analyze the spatial and temporal distribution of detected anomalies across Sri Lanka.
4) To characterize the meteorological profile of these anomalies to understand their nature and potential implications.

The beneficiaries of this analysis are diverse, including climatologists seeking data-driven evidence of local climate shifts, governmental and non-governmental bodies responsible for disaster preparedness and management, and agricultural stakeholders whose livelihoods depend on predictable weather patterns.

II. Related Work & Background

The application of data mining and machine learning to climatology is a rapidly evolving field, moving beyond simple statistical thresholds.

Traditionally, methods for identifying outliers in time-series data relied on statistical models. For instance, techniques based on Z-scores or Mahalanobis distance are effective for detecting deviations in uni-variate or low-dimensional data, assuming a Gaussian distribution, as outlined by Barnett and Lewis [1]. However, these methods often falter when faced with the high dimensionality and complex, non-linear interdependencies present in modern climate datasets.

Chandola et al. provide a foundational survey of anomaly detection techniques, categorizing methods suitable for various data types, including sequential and temporal data [2]. This work highlights the critical shift towards machine learning models that can learn complex patterns without strict statistical assumptions. Within this domain, unsupervised models are particularly valuable, as ”ground truth” labels for climatic anomalies are rare and context-dependent.

The Isolation Forest algorithm, proposed by Liu et al. [3], represents a significant advance in this area. Unlike distance-based (e.g., k-NN) or density-based (e.g., DBSCAN) methods, it does not rely on calculating distances, making it highly efficient for large, high-dimensional datasets. Its core principle—that anomalies are ”few and different” and thus easier to isolate—is particularly well-suited for meteorological data, where variables like precipitation are highly skewed.

Recent studies have successfully applied similar machine learning paradigms to climate science. Autoencoders, a type of neural network, have been used to detect abnormal sea-surface temperature patterns indicative of El Niño events [4]. Elsewhere, clustering algorithms have been used to identify distinct weather states, with anomalies being points that do not fit well into any cluster [5]. This project builds upon such work by applying the robust and scalable Isolation Forest technique to a comprehensive, multi-feature regional dataset, aiming to identify generalized climate instability rather than a single predefined phenomenon like a heatwave or cyclone. The focus is on finding days that are holistically unusual across a suite of weather variables, a task for which Isolation Forest is ideally designed.

III. Data Description & Pre-processing

A. Data Source and Attributes

The study utilizes the `SriLanka_Weather_Dataset.csv`, a comprehensive dataset containing 147,480 daily records from 30 distinct cities across Sri Lanka. The data spans a 13.5-year period from January 1, 2010, to June 17, 2023. The dataset includes 24 attributes, which can be categorized as shown in Table I.

TABLE I
Key Dataset Attributes and Descriptions

Category	Attributes
Temporal	`time`, `sunrise`, `sunset`
Temperature	`temperature_2m_(max/min/mean)`, `apparent_temperature_(max/min/mean)`
Precipitation	`precipitation_sum`, `rain_sum`, `snowfall_sum`, `precipitation_hours`
Wind	`windspeed_10m_max`, `windgusts_10m_max`, `winddirection_10m_dominant`
Other	`weathercode`, `shortwave_radiation_sum`
Geospatial	`latitude`, `longitude`, `elevation`, `city`

B. Data Pre-processing

The pre-processing pipeline was foundational to ensure data quality and suitability for time-series modeling:

1) **Missing Value Analysis:** A primary check of the dataset revealed no missing values in any of the 24 columns. This obviated the need for imputation techniques, which could have introduced artificial data points and skewed the analysis, thus preserving the integrity of the original data.
2) **Data Type Conversion:** The `time`, `sunrise`, and `sunset` columns, initially stored as object types, were converted to `datetime` objects using `pandas.to_datetime`. This transformation was crucial for enabling time-series operations, feature engineering, and correct temporal sorting and plotting.
3) **Data Structuring:** The DataFrame was sorted by `city` and then by `time` to ensure that all subsequent time-aware operations (e.g., rolling averages, lagged features) were calculated correctly within the context of each city’s unique timeline.

C. Exploratory Data Analysis (EDA)

Initial EDA was conducted to understand the underlying structure and characteristics of the data. Key findings from this phase guided the feature engineering and model selection process.

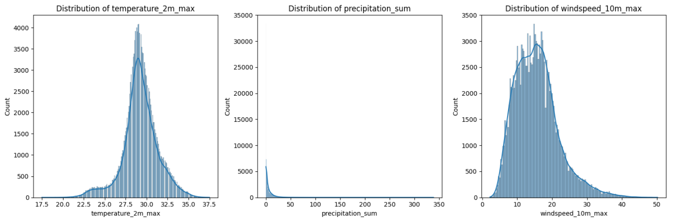

Fig. 1. Distribution of key numerical weather variables. Temperature is near-normally distributed, while precipitation and wind speed are highly right-skewed, indicating rare but extreme events.

As seen in Fig. 1, the distributions of core weather variables were revealing:

- **Temperature Variables:** Metrics like `temperature_2m_max` and `temperature_2m_mean` exhibited approximately normal distributions, which is typical for climatic temperature data centered around a seasonal mean.
- **Precipitation and Wind:** In sharp contrast, `precipitation_sum` and `windspeed_10m_max` were highly right-skewed. This indicates that most days have little to no rain and moderate winds, but a small number of days experience extreme levels of precipitation and high wind speeds. These rare events are prime candidates for being anomalies.
- **Time-Series Trends:** Time-series plots for mean temperature in representative cities like Colombo, Kandy, and Galle showed clear seasonal cycles, providing a visual

baseline for what constitutes ”normal” behavior over an annual cycle.

IV. Methodology

A. Feature Engineering

To enhance the model’s ability to detect temporal anomalies, raw data was augmented with engineered features designed to capture time-dependent patterns. These features were calculated on a per-city basis to respect geographical differences.

- `rolling_avg_temp`: A 7-day rolling average of `temperature_2m_mean`. This feature helps smooth out daily noise and establishes a short-term local baseline. A deviation from this rolling average is more significant than a deviation from the global mean.
- `month` & `day_of_week`: These cyclical features were extracted from the `time` column to allow the model to learn seasonal and weekly patterns. For example, a high temperature in January might be normal, while the same temperature in July could be an anomaly.
- `lagged_temp`: The mean temperature from the preceding day (*t-1*). This feature was created to capture temporal auto-correlation, as weather on any given day is often strongly influenced by the conditions of the previous day.

B. Anomaly Detection Algorithm: Isolation Forest

The Isolation Forest algorithm was selected as the primary tool for anomaly detection. This choice was driven by several key advantages it offers for this specific problem context.

- **Justification:**
  1) *Efficiency*: Unlike distance-based methods, its time complexity is linear with the number of data points, making it highly scalable for large datasets like the one used here.
  2) *High-Dimensionality Handling*: It performs well in high-dimensional spaces, a known challenge for many other algorithms (the ”curse of dimensionality”).
  3) *No Distributional Assumptions*: It does not assume any particular distribution (e.g., Gaussian) for the data, making it robust for handling the mixed and skewed distributions observed during EDA.
- **Working Principle:** The algorithm operates by building an ensemble of ”isolation trees” (iTrees). For each tree, data is recursively partitioned by randomly selecting a feature and then randomly selecting a split value for that feature. The core idea is that anomalous points are ”few and different,” so they are more susceptible to isolation and will thus be found closer to the root of the tree (i.e., they will have a shorter average path length across all trees).
- **Implementation:** The model was implemented using the `scikit-learn` library. It was trained on a set of 14 carefully selected numerical features, including all temperature metrics, precipitation, and wind variables. Geospatial columns and engineered features derived from the target variable were excluded from the model input to prevent data leakage and bias. The model’s `contamination` parameter, which specifies the expected proportion of outliers, was set to `'auto'`. This allows the algorithm to determine a suitable anomaly threshold based on the original methodology proposed by Liu et al. [3], making it more data-driven than a manually set value. The model outputs a prediction of -1 for anomalies and 1 for inliers for each data point.

V. Experiments & Results

The trained Isolation Forest model analyzed all 147,480 data points and successfully identified **20,065 records as anomalies**, representing approximately **13.6%** of the entire dataset. A deeper dive into these anomalies reveals distinct patterns.

A. Spatial and Temporal Distribution of Anomalies

The analysis showed that anomalies are not uniformly distributed across time or space. Specific locations and years are clear hotspots.

- **Spatial Distribution:** The central highlands of Sri Lanka emerged as a region with a significantly higher concentration of anomalies. Table II lists the top five cities by anomaly count, showing that Hatton, a high-elevation city known for its tea plantations, accounts for nearly a quarter of all detected anomalies.
- **Temporal Distribution:** The anomalies also clustered in specific years. As shown in Table III, 2014, 2010, and 2016 recorded the highest number of anomalous days. This temporal clustering suggests that large-scale climate phenomena, such as strong El Niño or La Niña cycles, may have driven these periods of instability.

TABLE II
Anomalies by City (Top 5)

City	Anomaly Count	% of Total Anomalies
Hatton	4,913	24.5%
Badulla	2,528	12.6%
Kandy	1,847	9.2%
Nuwara Eliya	1,770	8.8%
Bandarawela	1,326	6.6%

TABLE III
Anomalies by Year (Top 5)

Year	Anomaly Count
2014	1,719
2010	1,674
2016	1,618
2011	1,600
2021	1,595

B. Characterization of Anomalies

To understand what defines an ”anomalous” day, a comparative statistical analysis was performed between the anomalous subset and the full dataset. The results, summarized in Table IV, reveal a distinct meteorological profile for anomalies.

Key takeaways from this comparison are:

- **Temperature:** Anomalous days have a notably lower mean temperature but a significantly higher standard deviation. This suggests anomalies include not just uniform hot/cold days, but days with extreme temperature fluctuations. The gap between real and apparent temperature is also smaller on average for anomalies, potentially due to high winds.
- **Precipitation:** Anomalous days are much wetter, with an average `precipitation_sum` nearly three times that of a normal day.
- **Wind:** Wind and gust speeds are substantially higher during anomalies, with averages increasing by 45% and 42% respectively. This points to storms and other severe weather systems.

The most frequent weather codes during anomalies were Moderate Rain (Code 63), Light Drizzle (Code 51), and surprisingly, Clear Sky (Code 1). The prevalence of ”Clear Sky” anomalies suggests that some anomalies are defined not by precipitation but by extreme temperature events (e.g., heatwaves or unusual cold snaps) on otherwise clear days.

C. Visualization of Detected Anomalies

To provide an intuitive understanding of the results, time-series plots were created for the cities with the highest number of anomalies. Fig. 2 shows the mean temperature for Hatton, with detected anomalies marked in red. The visualization clearly shows that anomalies are not random noise but are specific points that deviate sharply from the established seasonal pattern. Many red points correspond to dramatic dips or peaks in temperature, confirming their unusual nature.

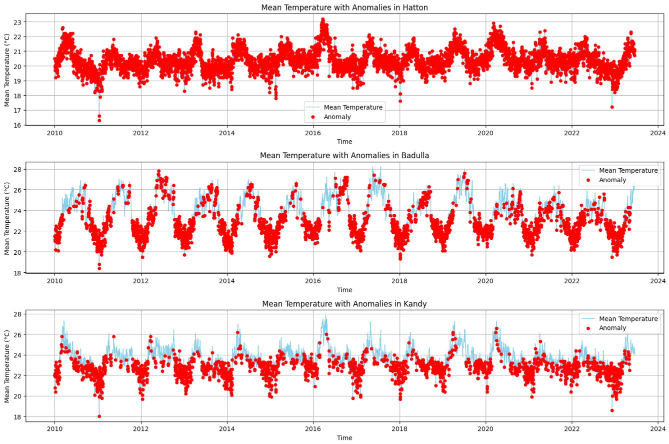

Fig. 2. Time series of mean daily temperature for the top 3 cities with the most anomalies (Hatton, Badulla, Kandy). Normal days are plotted in blue, while days identified as anomalies by the Isolation Forest are highlighted with red dots.

VI. Discussion

The results of this study provide compelling, data-driven evidence of significant weather instability in specific Sri Lankan regions. The model successfully moved beyond simple thresholding to identify complex, multi-variate events that represent true deviations from climatic norms.

The concentration of anomalies in Hatton, Badulla, and Nuwara Eliya—all high-elevation areas in the central highlands—strongly suggests these mountainous regions are more susceptible to extreme weather fluctuations. This may be due to orographic effects amplifying precipitation and wind, making these areas particularly vulnerable to events like landslides and flash floods. This finding has direct implications for regional disaster management and agricultural planning, especially for the tea industry which dominates this region.

The years with high anomaly counts (2014, 2010, 2016) warrant further climatological study. These periods could be cross-referenced with records of large-scale climate drivers, such as the El Niño-Southern Oscillation (ENSO) or the Indian Ocean Dipole (IOD), to investigate potential teleconnections that amplify local weather extremes. For example, 2016 was a strong El Niño year, which correlates well with the high anomaly count found by the model.

A key insight is the multi-faceted nature of the anomalies themselves. They are not merely small deviations; they are characterized by compound extremes—often high precipitation combined with high wind speeds and significant temperature variance. This aligns with scientific consensus on the effects of climate change, which projects not just a warming trend, but an increase in the frequency and intensity of extreme, multi-hazard events. The presence of ”clear sky” anomalies is also notable, indicating that the model successfully captured extreme heat or cold events not associated with storms, which could represent heatwaves or unusual drops in temperature.

A. Limitations

This study, while comprehensive, has several limitations:

1) **Algorithm Sensitivity:** The performance of the Isolation Forest model is sensitive to the feature set chosen and the `contamination` parameter. While `'auto'` provides a reasonable baseline, a more nuanced threshold could potentially be set with domain expert input, which might refine the set of detected anomalies.
2) **No Causal Inference:** This analysis identifies statistical anomalies but does not inherently prove a causal link to anthropogenic climate change. Such a conclusion would require more sophisticated attribution studies that are beyond the scope of this data mining project. The findings are strong indicators, but not definitive proof.
3) **Data Granularity:** The dataset provides daily summaries. Intra-day data (e.g., hourly) could reveal more nuanced anomalies, such as sudden, intense bursts of rain (cloudbursts) that might be averaged out in a daily summary.

TABLE IV
COMPARISON OF DESCRIPTIVE STATISTICS: FULL DATASET VS. ANOMALY SUBSET

	Full Dataset (Inliers + Outliers)				**Anomaly Subset Only**			
Feature	**Mean**	**Std Dev**	**Min**	**Max**	**Mean**	**Std Dev**	**Min**	**Max**
`temperature_2m_mean` (°C)	26.16	2.66	16.7	32.3	24.44	4.29	16.7	32.1
`apparent_temp_mean` (°C)	31.79	4.90	16.0	44.1	28.52	8.16	16.0	43.6
`precipitation_sum` (mm)	6.04	13.04	0.0	210.9	17.33	24.93	0.0	210.9
`windspeed_10m_max` (km/h)	21.08	7.97	5.1	74.5	30.63	10.98	5.9	74.5
`windgusts_10m_max` (km/h)	39.86	13.56	9.7	123.1	56.63	17.51	12.2	123.1

B. Ethical Considerations

The data used in this study consists of publicly available meteorological records, posing no personal privacy concerns. However, the interpretation and communication of the findings carry significant ethical weight. It is crucial to frame these results as potential indicators of climate instability and risk rather than as definitive predictions of future disasters. Care must be taken to avoid causing undue public alarm. The most responsible application of these findings is to share them with relevant stakeholders—climatologists, disaster management agencies, and agricultural bodies—to inform their domain-specific models and preparedness plans.

VII. CONCLUSION & FUTURE WORK

This project successfully designed and executed a data mining pipeline to detect and analyze climate anomalies in a large-scale Sri Lankan weather dataset. By leveraging a robust unsupervised algorithm, Isolation Forest, the study moved beyond simple trend analysis to identify 20,065 statistically significant anomalous weather days. The analysis revealed clear spatial and temporal patterns, with Sri Lanka's central highlands and specific years like 2014 and 2016 emerging as hotspots of climate instability. The characterized anomalies, marked by extremes in precipitation, wind, and temperature variance, serve as tangible indicators of a changing and more volatile regional climate.

Future work could build upon this foundation in several promising directions:

1) **Anomaly Severity Scoring:** Instead of a binary classification (anomaly/inlier), future work could implement a scoring system based on the Isolation Forest's path length or by measuring the deviation from a seasonal baseline. This would allow for the categorization of anomalies by severity (e.g., mild, moderate, extreme), providing more actionable information for risk assessment.
2) **Comparative Modeling:** Apply and compare other anomaly detection methods, such as Seasonal-Trend decomposition using Loess (STL) or deep learning models like LSTM Autoencoders. A comparative analysis could validate the findings of this study and potentially discover different types of anomalies (e.g., long-term deviations vs. short-term shocks).
3) **Root Cause Analysis:** Employ interpretable machine learning models (e.g., SHAP or LIME) on top of the anomaly detection results. This would help determine which specific features (e.g., wind gusts, apparent temperature) contribute most to a data point being flagged as an anomaly, offering deeper insights into the physical drivers of these events.
4) **External Data Correlation:** Integrate the detected anomalies with external datasets, such as those tracking climate indices (ENSO, IOD), agricultural crop yields, or reports of natural disasters. This would facilitate a more robust investigation into the causes and, more importantly, the real-world consequences of these identified climate anomalies.

REFERENCES

[1] V. Barnett and T. Lewis, *Outliers in Statistical Data*. Wiley, 1994.

[2] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[3] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*, pp. 413-422, 2008.

[4] A. G. Muñoz et al., "The Crystal Ball of the Climate: A survey on Machine Learning applications to seasonal climate forecasting," *Artificial Intelligence Review*, vol. 54, pp. 5319–5342, 2021.

[5] D. J. Gagne II et al., "Interpretable deep learning for convective storm annotation and forecasting," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 317–324, 2019.