Data Quality and Time Series Analysis Documentation

Introduction

Data quality management is a critical aspect of scientific research, particularly in surface modeling, where precision and accuracy are paramount. This document outlines the processes involved in managing data quality, validating time series data, and analyzing key variables from NASA's Land Information System (LIS).

Data Entry

Step 1: Initial Data Setup

1. Path Definition:
   - Define the path to the tar.gz file and the extraction directory.
   - Specify the directory containing the NetCDF files.
2. Loading Datasets:
   - Read and load the datasets into two primary data structures: ds_surface and ds_routing.
3. Dataset Overview: The ds_surface and ds_routing datasets contain essential variables related to surface modeling. Below are key variables included in the analysis:

Variable	Dimensions	Description	Unit
lat (Latitude)	(time, north_south, east_west)	Geographical latitude at each grid point	Degrees North
lon (Longitude)	(time, north_south, east_west)	Geographical longitude at each grid point	Degrees East
SoilMoist_tavg	(time, SoilMoist_profiles, north_south, east_west)	Average soil moisture content, providing depth profiles (4 layers defined by thicknesses of 0.1m, 0.3m, 0.6m, and 1.0m.)	m^3/m^3
TWS_tavg	(time, north_south, east_west)	Total terrestrial water storage	mm
Streamflow_tavg	(time, north_south, east_west)	Average streamflow in rivers/streams	m^3/s

Table 1: key variables included in the analysis

Data Validation

Step 2: Validating Time Series Data

Validating the accuracy of time series data involves methods to assess both data integrity and predictive model performance. The following techniques are employed:

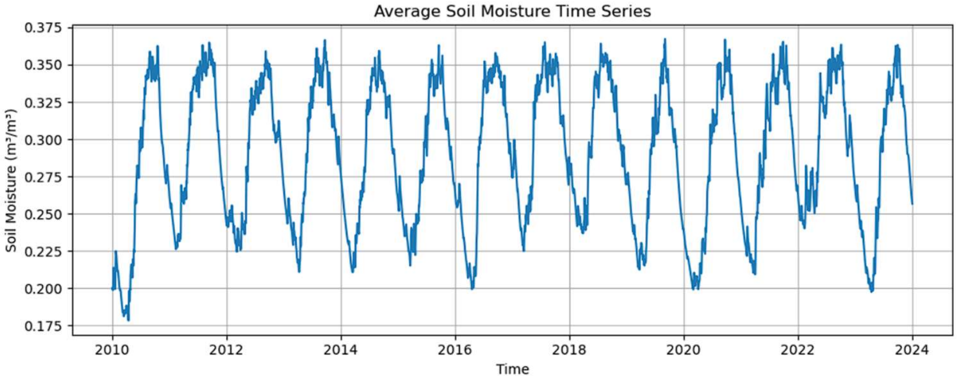
Methods of Validation:

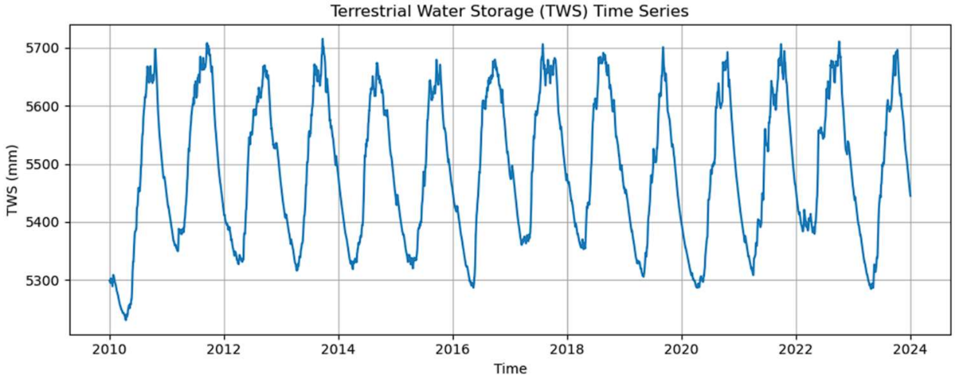
1. **Visual Inspection:**
   - Create line plots for the time series to identify patterns, trends, seasonality, and anomalies.
   - Key analyses include:
     - Trend Analysis: Analyze long-term trends across variables with averages across specified dimensions.
     - Decomposition: Break time series into its trend, seasonal, and residual components, allowing for anomaly detection.
2. **Trend Analysis**

   Describe the variables being analyzed (e.g., Soil Moisture, Terrestrial Water Storage, Streamflow).

   Output Generation:

   - Generate three distinct line plots for Soil Moisture to visualize changes from January 2010 to January 2024.





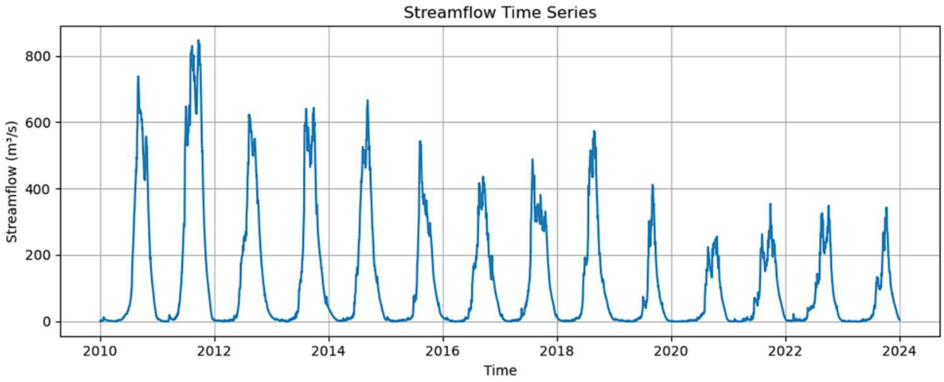


Figure 1: Display time series plot of Soil Moisture, Terrestrial Water Storage and Streamflow respectively

Observations:

- The line fluctuates, showing peaks and troughs that indicate variability in the Streamflow over the years.
- **Seasonal Trends**: The plot shows recurring patterns, such as higher streamflow during rainy seasons and lower during dry seasons.
- **Anomalies**: The significant deviations from the trend that diverge markedly from the established trend, often beyond what would be expected due to normal variability, which could indicate unusual events like floods or droughts.
- **Monitoring Systems**: In environmental monitoring, significant deviations in streamflow data might indicate flooding or drought conditions, prompting necessary actions.

3. Decomposition Analysis

- Sudden spikes or drops in the observed data may indicate unusual weather phenomena such as heavy rainfall or drought conditions.
- Analyze the trend to detect long-term shifts in soil moisture levels that may correlate with climate change.

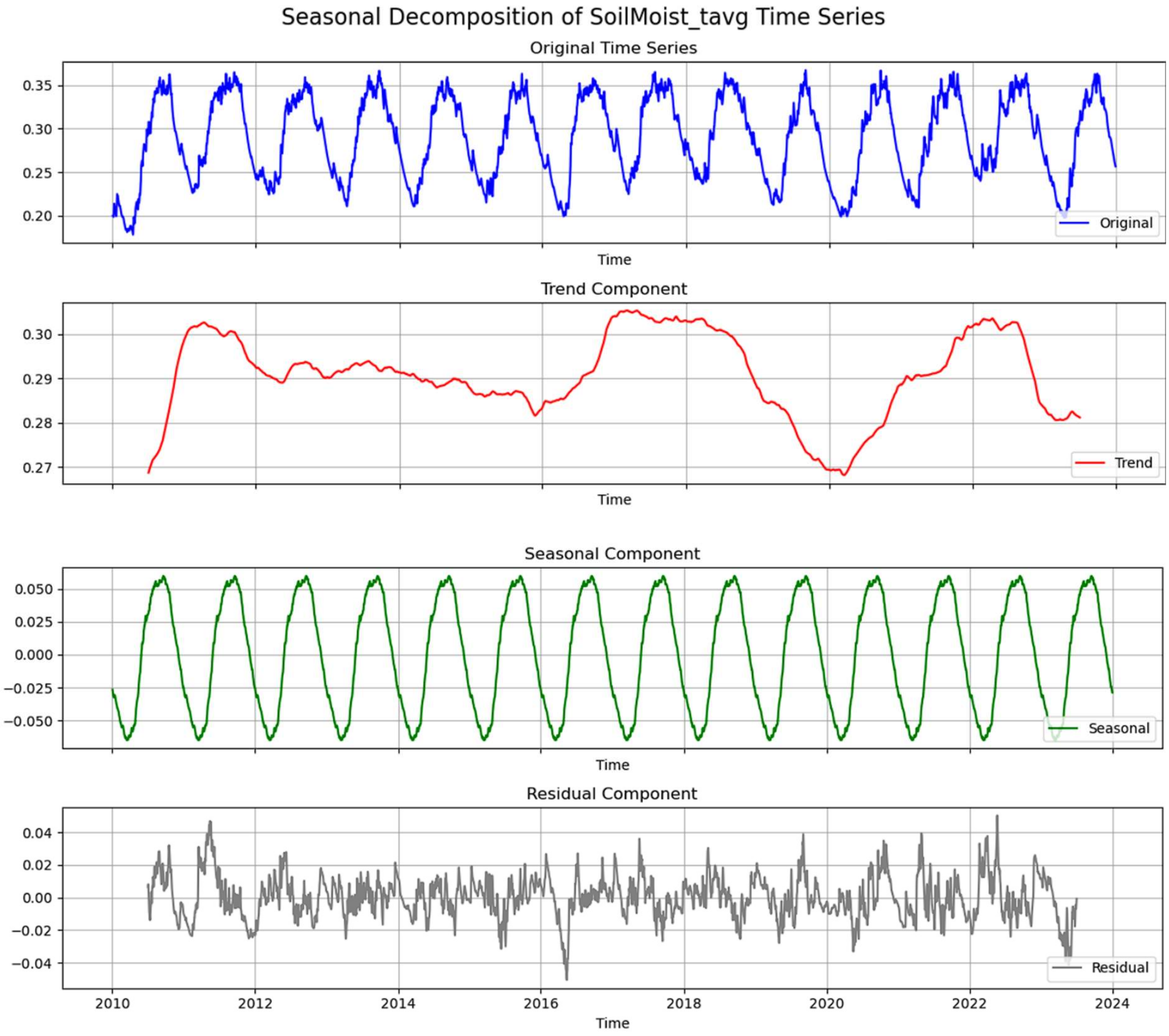


Figure 2: Display frequency of the seasonal decomposition of soil moisture.

Observations:

- **Trend Component**: Describe the overall pattern in the data over time.
  - The trend component shows a gradual increase in soil moisture levels over the years, indicating a long-term rise in soil moisture.
- **Seasonal Component**: Identify and describe any recurring cycles in the data.
  - The seasonal component reveals a clear annual cycle, with soil moisture peaking during the rainy season (June to September) and dipping during the dry season (December to February).
- **Residual Component**: Highlight any irregularities or noise in the data after removing the trend and seasonal effects.
  - The residual component shows some irregular fluctuations, which could be attributed to short-term weather events or measurement errors.

Analysis:

- The consistent seasonal trends suggest a stable climate pattern in the region. However, the anomalies in certain years highlight the need for improved water management strategies to cope with extreme weather events.

Conclusion:

- The analysis of soil moisture data from January 2010 to January 2024 reveals important seasonal trends and critical anomalies. It is recommended to enhance water conservation efforts and develop strategies to mitigate the impacts of droughts and floods.

Step 3: Anomaly Detection

1. **Outlier Identification:**

○ Use basic statistics (mean, median, standard deviation) alongside visualizations (box plots) to identify outliers.

Statistic	Soil Moisture	TWS	Streamflow
Mean	0.2894	5490.4771	120.7459
Median	0.2943	5487.1938	25.6520
Standard Deviation	0.0478	124.6536	176.3236
Q1 (25th percentile)	0.2476	5381.6465	2.9284
Q3 (75th percentile)	0.3354	5607.5884	184.1801
IQR	0.0878	225.9419	181.2517
Lower Bound	0.1160	5042.7336	-268.9491
Upper Bound	0.4671	5946.5012	456.0576
Number of Outliers	0	0	407

Table 2: Identifies and understands outliers in a time series based on basic statistics and IQR for each variable: Soil Moisture, Terrestrial Water Storage and Streamflow.

Interpretation:

- "The mean soil moisture (SoilMoist_tavg) is 0.2894, indicating the average moisture level across the study area."
- "Streamflow_tavg has a high number of outliers (407), suggesting significant variability in streamflow data."

- High-value outliers may indicate rare events meriting further examination, while measurement errors will be excluded to maintain data integrity.

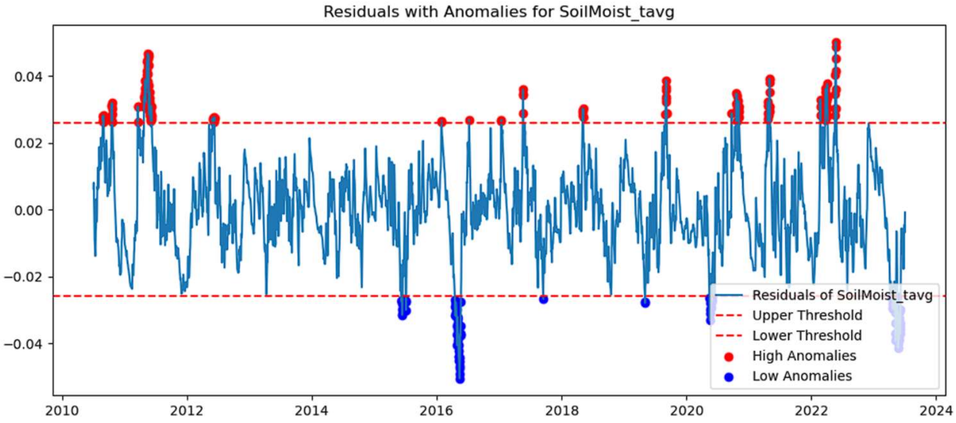


Figure 3: Display residuals with Anomalies of soil moisture.

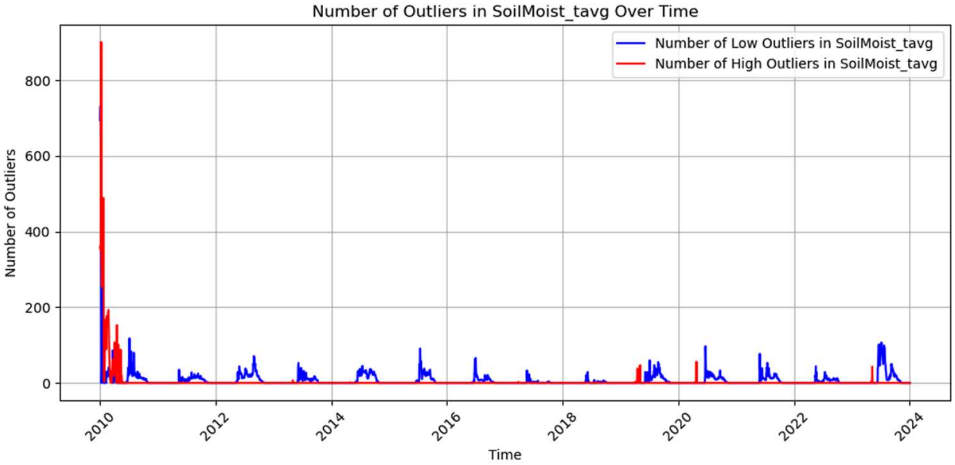


Figure 4: Display number of outliners in soil moisture over time.

Interpretation and Recommendations Soil Moisture Data Analysis

- **Outliers**: The presence of both low and high outliers suggests variability in soil moisture levels, which could be influenced by factors such as precipitation, evaporation, and soil properties.
- **Bounds**: The lower and upper bounds help in identifying the normal range of soil moisture values. Outliers outside these bounds need further investigation to understand their causes.

Conclusion

The analysis of Soil Moisture data reveals significant variability and the presence of outliers. Understanding these outliers and their causes is crucial for improving the accuracy of surface modeling. By regularly monitoring and analyzing these variables, we can gain valuable insights into environmental changes and their impacts.

Step 4: Analysis of Long-Term Trends

1. **Trend Calculation:**

   - Analyze spatial trends for variables, differentiating between positive and negative slopes through geographic mapping.

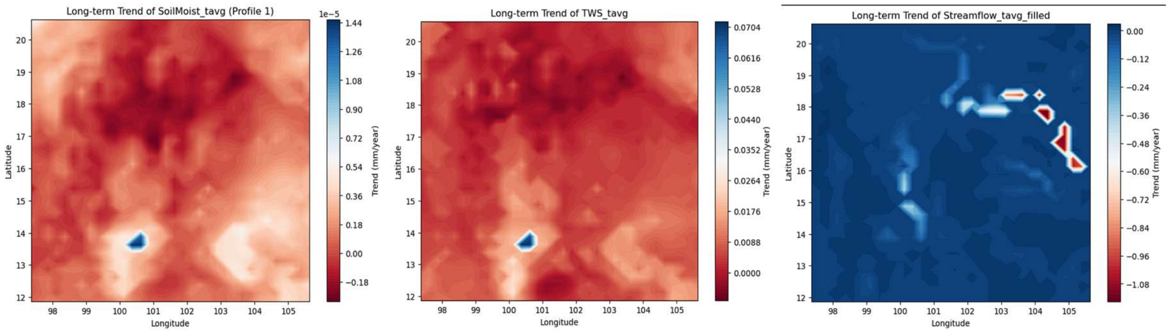


Figure 5: Display the trend map for Soil Moisture, Terrestrial Water Storage and Streamflow.

Result for soil moisture example:

- **Highest Positive Trend for soil moisture**:
  - **Location**: (7, 12)
  - **Trend**: 0.008962 mm/year
  - **Coordinates**: (Lat: 13.625, Lon: 100.375)
- **Highest Negative Trend for soil moisture**:
  - **Location**: (23, 9)
  - **Trend**: -0.001429 mm/year
  - **Coordinates**: (Lat: 17.625, Lon: 99.625)

Interpretation and Recommendations

- **Positive Trend**: The highest positive trend in soil moisture was observed at coordinates (13.625, 100.375), with an increase of 0.008962 mm/year. This indicates a gradual increase in soil moisture in this area, which could be beneficial for agriculture and vegetation.
- **Negative Trend**: The highest negative trend was noted at coordinates (17.625, 99.625), with a decrease of -0.001429 mm/year. This suggests a drying trend in this region, which may require water management interventions.

Step 5: Analysis and Statistical Tests Data

1. **Streamflow Disturbances and Statistical Testing**
   - **Statistical Testing**: In the analysis of streamflow data, statistical testing was conducted to identify significant differences in streamflow between two different periods. The following findings were derived from the t-test applied to the streamflow datasets:

- T-statistic: 134.95
- P-value: 0.0

- **Statistical Interpretation:**

  - **The t-test statistic** measures the size of the difference relative to the variation in the sample data. A t-statistic value of 134.95 indicates a substantial gap between the two periods being analyzed, suggesting that the changes observed in the dataset are far beyond what could be attributed to random chance.
  - **The p-value**, which quantifies the probability of observing the data or something more extreme given that the null hypothesis is true, was reported as 0.0. This extremely low p-value signifies strong evidence against the null hypothesis, which asserts that there is no significant difference in streamflow between the two periods.

- **Conclusion:**

  - Given the high t-statistic and the near-zero p-value, we confidently reject the null hypothesis. This outcome indicates a significant difference in streamflow between the evaluated periods, highlighting important changes in hydrological conditions that may be influenced by environmental factors, management practices, or climatic variations.
  - These findings are essential for understanding the dynamic behavior of water resources and provide a basis for further investigation into the causes behind the observed streamflow changes.

2. **Autocorrelation Analysis and Statistical Tests**

   Tools:

   - ACF and PACF: Used for detecting lag structure and seasonality.
     Statistical Tests: Perform tests such as the Augmented Dickey-Fuller (ADF) and Ljung-Box to validate assumptions of stationarity and check for autocorrelation in residuals.

Results of the Augmented Dickey-Fuller (ADF) Test:

- ADF Statistic: -6.955854304387079
- p-value: 9.43e-10 (very small)
- Critical Values:
  - 1%: -3.4316
  - 5%: -2.8621
  - 10%: -2.5671

Interpretation:

- **ADF Statistic**: The ADF statistic is -6.95, which is much lower (more negative) than all critical values at the 1%, 5%, and 10% significance levels.
- **p-value**: The p-value is close to zero (9.43e-10), which is much smaller than typical significance levels such as 0.05, 0.01, or 0.10.

Conclusion:

- Since the ADF statistic is significantly lower than the critical values, and the p-value is very small, we reject the null hypothesis of the ADF test. This indicates that the time series is stationary at all conventional significance levels (1%, 5%, and 10%).
- Given that the time series is stationary, difference is not necessary. The data is suitable for time series modeling without further transformation to remove non-stationarity, which should help improve ARIMA model fitting.

Overall Conclusion

Effective data quality management and time series analysis are essential for ensuring the reliability and accuracy of surface modeling research. By following a structured approach to data entry, validation, and analysis, researchers can identify and correct anomalies, understand long-term trends, and perform robust statistical tests. This process enhances the credibility of the findings derived from NASA's Land Information System (LIS) data.

Recommendations

1. Implement Rigorous Data Validation: Employ multiple methods such as visual inspection, trend analysis, and decomposition analysis to ensure the integrity of time series data.

2. Automate Anomaly Detection: Utilize advanced algorithms for outlier identification to streamline the detection of data anomalies and reduce manual effort.

3. Focus on Long-Term Trends: Regularly calculate and analyze long-term trends to understand underlying patterns and potential disturbances in streamflow data.

4. Conduct Comprehensive Statistical Tests: Perform autocorrelation analysis and other statistical tests to validate the consistency and reliability of the data.

5. Continuous Monitoring and Improvement: Establish a feedback loop for continuous monitoring and improvement of data quality management practices to adapt to new challenges and advancements in the field.