Analyzing Weather Temperature Trends

Prerak Patel

***Abstract*—Analyzing temperature trends for a specific region provides researchers deep insights into the impact of global warming and helps develop strategies for its mitigation. It also helps in planning for farming and managing water resources [1]. This report focuses on daily and monthly data with a combined approach of parametric and non-parametric tests, along with deep visualizations, to identify the underlying trends in the data. Preprocessing is performed along with Exploratory Data Analysis (EDA), followed by a Linear Regression Model, the Mann-Kendall test (MK), the Pettitts test, and the Theil-Sen slope estimator.MK test showed a significant trend in daily and monthly temperature data.**

***Index Terms*—IEEE, Trend, Linear Regression,Mann-Kendall test, Modified-Mann-Kendall test, Theil-Sen estimator LaTeX, paper, template.**

I. Introduction

TEMPERATURE is one of the most important variable in climate change and also studying it helps us in getting knowledge of effects of global warming on a particular region. This study focuses on analyzing trends in temperature data with help of statistical tests.

A trend can be defined as an upward or downward movement in data. Parametric and Non-Parametric tests are compared here and used to find the underlying trends in the data.

Jan 30, 2025

II. Approach/Methods/Materials

The dataset was downloaded from Kaggle, which consists of 25 years of weather data for a particular place and could be found here [2]. The approach is simple: preprocess the data, perform exploratory data analysis, and then apply tests to the data. The language used for the coding is Python. [3] The libraries used are Sklearn, Pymannkendall, scipy, stats models, numpy, and pandas. [4] The technology used is the Jupyter Notebook.

The tests performed are as follows:

- Linear Regression model
- Mann-Kendall test
- Theil-Sen slope estimator
- Pettitts test

A. Exploratory data analysis

We collected the hourly data and converted it to daily and monthly data by resampling it. Then, we plotted the raw data and some random chunks over time using Matplotlib to visualize how the data looks over time.

The first step after preprocessing is plotting the moving averages of daily temperature data with a window of 7 days, 30 days, and 90 days. This method was applied to distinguish between short-term and long-term trends in the data.

B. Linear Regression

Linear Regression model is used to find a best-fit straight line for the data approximation. [5] It is a simple parametric approach used in statistics, trend analysis, and machine learning that does assume any specific data distribution. The equation of linear regression is given by:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

Where:

- y: Dependent variable
- β_0: Intercept
- β_1: Slope of the straight line
- x: Independent variable
- ϵ: error in the data

The slope of the line indicates the trend in data: increasing if positive, no trend if zero, and decreasing if negative.

C. Mann-Kendall test

It is a statistical and non parametric method which is used to identify monotonic trend in data [6]. Its useful in trend analysis as it does not need the data to be normally distributed and its robust to abrupt changes in data.

Let the time series be $x_1, x_2, x_3, ...x_n$ where n is the number of data points. The Mann-Kendall statistic is given by [7]:

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} sign(x_j - x_k) \tag{2}$$

$$\text{sign}(x_j - x_k) = \begin{cases} +1 & \text{if } (x_j - x_k) > 0, \\ 0 & \text{if } (x_j - x_k) = 0, \\ -1 & \text{if } (x_j - x_k) < 0. \end{cases} \tag{3}$$

Take the difference of each data point with all other data point and then sum the signs of the difference.

- Null Hypothesis: There is no trend in data
- Alternative Hypothesis: There is trend in data

The variance V(S) of the data can be calculated as:

$$V(S) = \frac{n(n-1)(2n+5)}{18} \tag{4}$$

where n is the number of data points

The normalized test statistic Z_s is given by:

$$Z_s = \begin{cases} \frac{S-1}{\sqrt{V(S)}} & \text{if } S > 0, \\ 0 & \text{if } S = 0, \\ \frac{S+1}{\sqrt{V(S)}} & \text{if } S < 0. \end{cases} \tag{5}$$

Department of Electrical and Computer Engineering, Florida Institute of Technology, Florida,FL,32901,USA
Professor Anthony Smith

The final results could be interpreted based on this:

- Decreasing Trend: $Z < 0$ and p-value<significance level
- No-trend: p-value $>$ significance level
- Increasing Trend: $Z > 0$ and p-value<significance level

D. Theil-Sen slope estimator

Theil-Sen slope estimator is a simple and robust method for finding trends in data. It works by finding the median slope of all possible data pairs. It is significantly less sensitive to outliers and could be performed in the following steps:

- Find the slope using all possible data pairs and calculate its median slope.
- Then calculate y-intercept using $y_i - m * x_i$ where m is the slope

E. Pettitts test

Its a well-known test to find the change point detection in the data. A non-parametric test which means no assumption regarding the distribution of the data [8].

- Null Hypothesis: No change point in data
- Alternative Hypothesis: change point detected.

Steps to perform pettitts test are as follows [9]:

$$\text{Test Statistics } U_t = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} sign(x_j - x_k) \tag{6}$$

$$\text{sign}(x_j - x_k) = \begin{cases} -1 & \text{if } (x_j - x_k) > 0, \\ 0 & \text{if } (x_j - x_k) = 0, \\ +1 & \text{if } (x_j - x_k) < 0. \end{cases} \tag{7}$$

The test statistic K_t is given by $Max|U_t|$

III. Experiments/Results

A. Exploratory data analysis

The data is first converted to daily from hourly, and then both plots are plotted for 24 years. It is hard to visualize the data this way, so plotting random data intervals was necessary to get deeper insights. The plots are as shown below:

As a part of exploratory data analysis, moving averages over 7 days, 30 days, and 90 days are plotted below. A part of the moving average data plot is shown in Fig 5.The plotly library makes the plots more interactive, as it could be more concise as seen in Fig 4.

B. Linear regression

The result of the linear regression showed increasing trend in the data. The result are as below:

	coef	standard error	table value	P-value
const	-105.4141	23.429	-4.499	$< 0.05*$
x1	0.0002	3.19e-05	4.937	$< 0.05*$

TABLE I: Regression Results

The slope and intercept for daily data without smoothing is 0.000157 °C/day and intercept is -105.4141°C .This would be around 0.57°C over a decade. The plot of linear regression is as follows:

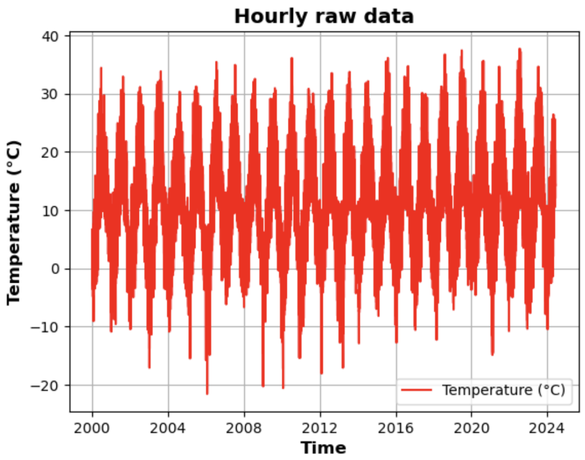

Fig. 1: Raw data hourly plot

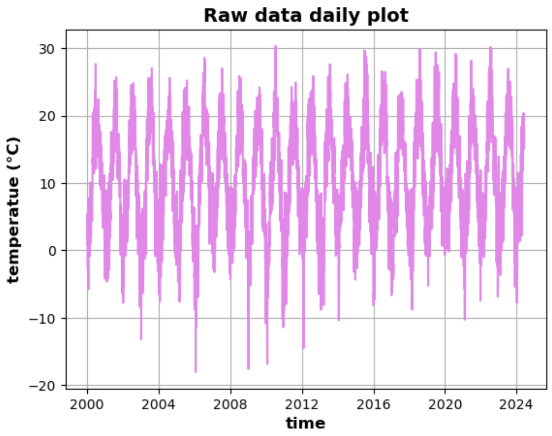

Fig. 2: Raw data daily plot

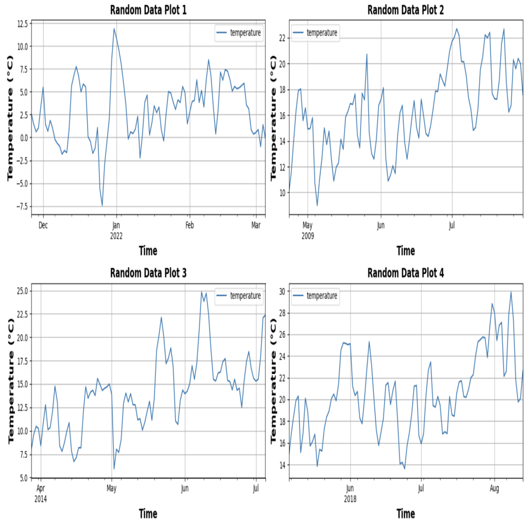

Fig. 3: Random data chunk plot

C. Mann-Kendall test

The results revealed an increasing trend in the overall data. It can be seen in table II

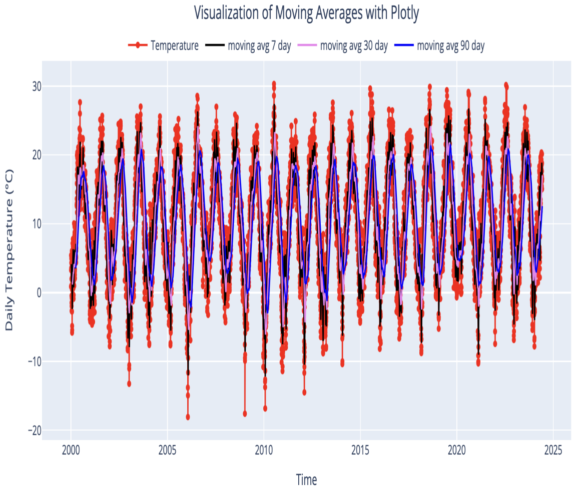

Fig. 4: Moving Average whole data plot

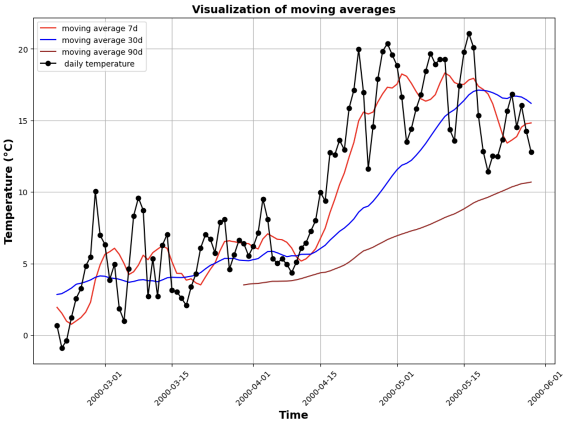

Fig. 5: Moving Average data chunk plot

S	Variance of S	z	p-value	Trend
1227056	78951996522.6	4.36	1.25e-5	increasing

TABLE II: Mann-Kendall Result

Seasonal Mann-Kendall test is also performed on the data to find the underlying trends in each month's data. The month of June showed a significant trend in the data. The results are as follows:

D. Theil-Sen slope estimator

The slope of the trend line found by Theil-Sen slope estimator is 0.00017°C/day and the intercept is -116.16°C.

This test was also performed for each month data and the results are as shown in table IV:

E. Pettitts Test

The result of pettitts test detects change point in data in month of June the results are in table V

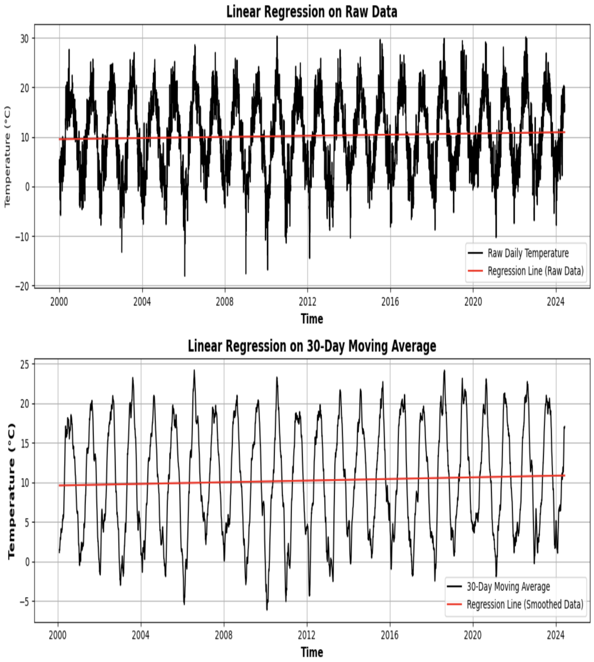

Fig. 6: Linear Regression with 30-day average smoothing and without smoothing

Index	S	Z	P-value	Trend	Month
0	52	1.19	0.23	No Trend	1
1	54	1.23	0.21	No Trend	2
2	56	1.28	0.19	No Trend	3
3	-8	-0.16	0.87	No Trend	4
4	-22	-0.49	0.62	No Trend	5
5	101	2.33	0.01	Trend Detected	6
6	60	1.46	0.14	No Trend	7
7	50	1.21	0.22	No Trend	8
8	66	1.61	0.10	No Trend	9
9	42	1.01	0.30	No Trend	10
10	36	0.86	0.38	No Trend	11
11	54	1.31	0.18	No Trend	12

TABLE III: Seasonal Mann-kendall test Results

Month	Slope	Intercept
1	0.086148	-172.141050
2	0.105088	-209.657906
3	0.058868	-113.514687
4	-0.008237	25.491772
5	-0.022492	59.634493
6	0.104608	-192.438929
7	0.052850	-86.373490
8	0.067085	-115.513306
9	0.065494	-116.429367
10	0.075719	-141.259882
11	0.039486	-73.808633
12	0.083264	-164.919141

TABLE IV: Monthly Theil-Sen Slope Estimates

IV. Discussion

The analysis of 25 years of temperature data revealed a warming pattern all over the region over the long term.

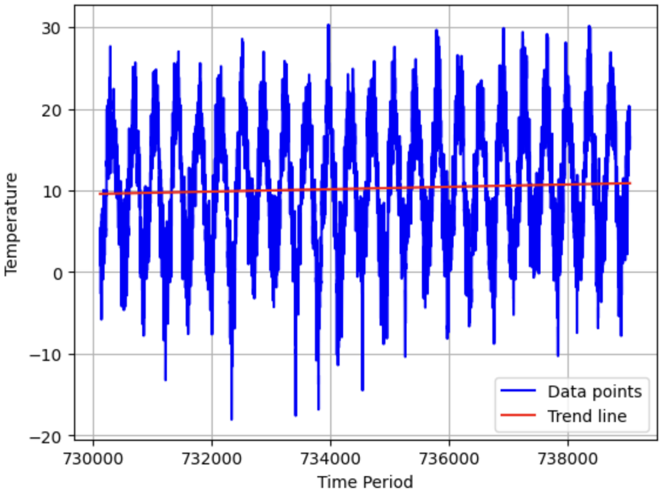

Fig. 7: Theil-Sen slope estimator result plot

Months/Seasons	**KT**	**P-value**	**Change point (Year)**
January	66.0	0.24	No
February	72.0	0.17	No
March	68.0	0.23	No
April	44.0	0.70	No
May	40.0	0.79	No
June	110.0	0.01	2016-06-30 00:00:00
July	63.0	0.24	No
August	71.0	0.14	No
September	62.0	0.26	No
October	53.0	0.42	No
November	44.0	0.64	No
December	83.0	0.05	No

TABLE V: Kendall Tau (KT) and Change Points by Month/Season

The result of linear regression is a positive slope, indicating the same increase over the long term. However, the low coefficient of variation suggests that this test is not a good fit for the data. The Mann-Kendall and Theil-Sen slope estimator test suggested an increase in overall temperature data. The seasonal Mann-Kendall showed a positive trend in June.

The Pititts test showed a significant change point over time in June 2016. No other significant change point was found in the data.

V. Conclusion

This study analyzed long-term trends using combination of both parametric and non-parametric test. The key findings are as follows:

- A increasing trend is detected in linear regression, the Mann-Kendall test and Theil-Sen slope estimation.
- Seasonal Mann-Kendall test suggested increasing trend in month of June.
- Pettitts test revealed a change point in year 2016.

The increase in trend is probably due to global warming, which highlights the importance of trend analysis for this particular region. Future work can focus on including more climate features such as precipitation, rainfall, and humidity.

References

[1] V. HT, A. Ghosh, S. Ojha, P. Poddar, and P. Basak, "Trend analysis of rainfall and detection of change point in terai zone of west bengal," *International Journal of Environment and Climate Change*, vol. 14, no. 1, pp. 603–613, 2024.

[2] ParthDande, "Weather dataset," 2023, accessed: 2023-10-15. [Online]. Available: https://www.kaggle.com/datasets/parthdande/timeseries-weather-dataset?select=Weather_dataset.csv

[3] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 2011, vol. 12.

[4] W. McKinney, "Data structures for statistical computing in python," *Proceedings of the 9th Python in Science Conference*, pp. 51–56, 2010.

[5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009.

[6] U. Alhaji, A. Yusuf, C. Edet, C. O. Oche, and E. Agbo, "Trend analysis of temperature in gombe state using mann kendall trend test," *J. Sci. Res. Rep*, vol. 20, no. 3, pp. 1–9, 2018.

[7] Y. S. Güçlü, "Improved visualization for trend analysis by comparing with classical mann-kendall test and ita," *Journal of Hydrology*, vol. 584, p. 124674, 2020.

[8] A. N. Pettitt, "A non-parametric approach to the change-point problem," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 2, pp. 126–135, 1979. [Online]. Available: http://www.jstor.org/stable/2346729

[9] V. HT, A. Ghosh, S. Ojha, P. Poddar, and P. Basak, "Trend analysis of rainfall and detection of change point in terai zone of west bengal," *International Journal of Environment and Climate Change*, vol. 14, no. 1, pp. 603–613, 2024.