Environmental Data Analytics
Project 1 — Climate Data Analysis Using ERA5-Land Dataset

Anton Zaitsev, Othmane Mahfoud — University of Luxembourg

October 9, 2024

1 Data

1.1 Data Source

The dataset used in this analysis was downloaded from the ERA5-Land reanalysis dataset, which provides high-resolution climate data. The data was obtained through the Copernicus Climate Change Service (C3S) platform. We used **ERA5-Land monthly averaged data from 1950 to present**.

1.2 Data Download

The data for this analysis was originally downloaded for the entire planet. Afterwards, we selected only the region around **Esch-sur-Alzette** in southern Luxembourg for further analysis. Specifically, we used latitude **49.5°** and longitude **6.0°** values, representing the center of the grid cell, with the following boundaries:

- **Latitude**: From **49.55°N** to **49.45°S**
- **Longitude**: From **5.95°W** to **6.05°E**

This area includes southern Luxembourg and parts of neighboring France. The data spans the years from **1974 to 2023**.

1.3 Variables and Dataframe Creation

Five climate variables were selected from the dataset:

- **t2m**: Temperature at 2 meters above the surface
- **ssrd**: Surface solar radiation downwards
- **e**: Total evaporation
- **sp**: Surface pressure
- **tp**: Total precipitation

These variables were extracted using a grid cell corresponding to the aforementioned latitude and longitude values. The data was loaded using the **NetCDF4** library, and specified latitude and longitude indices were used to access the data values for the chosen location. Latitude **49.5°N** and longitude **6.0°E** values correspond to latitude index **405** and longitude index **60**.
After loading, the data was stored in a Python dictionary and then converted into a **pandas DataFrame**. Each variable was represented as a column, and a 'year' column was added to the dataframe. The dataframe is then saved as a *.csv* file.

1.4 Data Cleaning

The raw downloaded data is already clean, with no *NaN* values and all data properly formatted. The only modification made is converting e (total evaporation) from a negative to a positive scale.

1.5 Time Series Plots & Histograms

In order to better understand the data, we decided to visualize the following phenomena:

1. **Mean Values Over Time**: The yearly average of each variable was calculated, and these means were plotted to observe trends over time (see Figure 1).

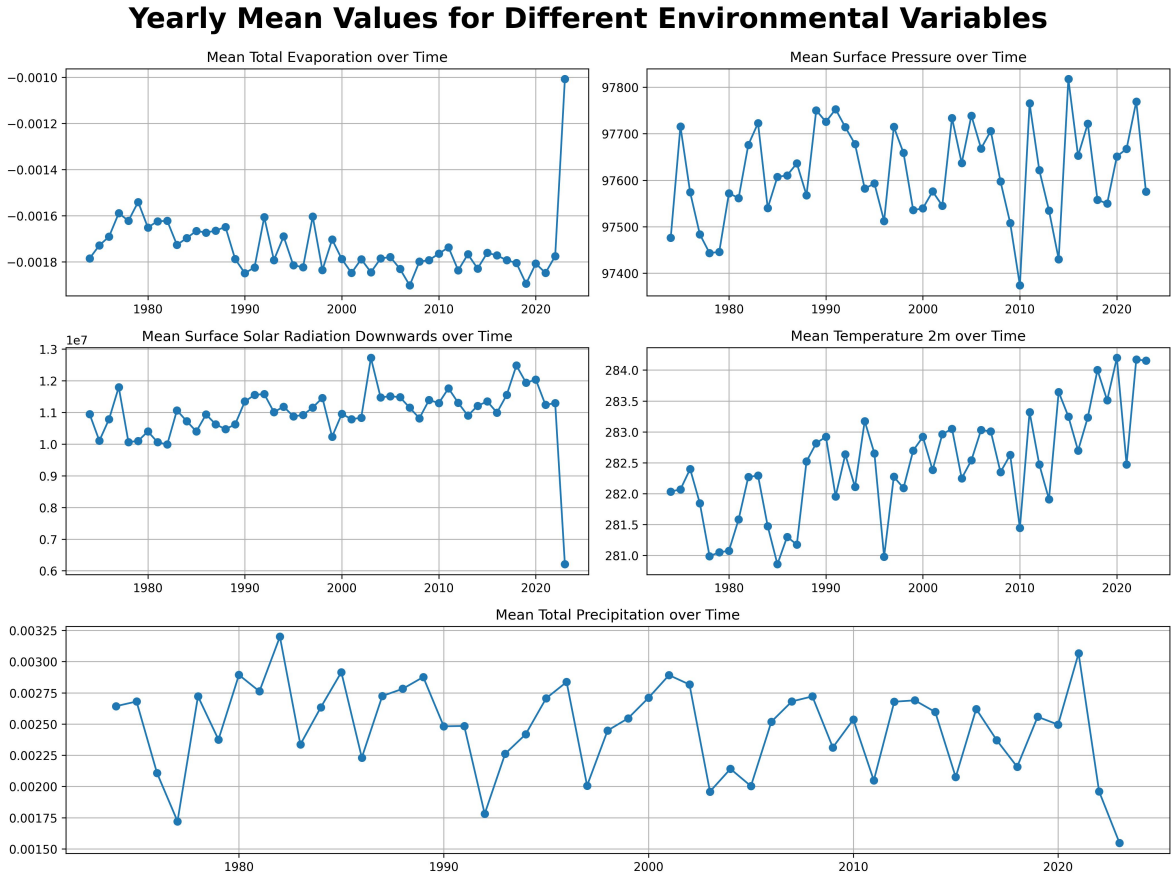

Figure 1: Mean Values Over Time

2. **Distribution Comparisons**: The distributions of each variable was calculated and plotted highlighting any potential anomalies (see Figure 2).

3. **Summary Statistics**: Summary statistics, such as mean, median, variance, maximum and minimum values, for each variable across different years were calculated, along with a rolling average for smoothing. (see Figure 3).

1.6 Anomaly in 2023

Upon analyzing the data, we observed an anomaly in the year **2023**, where:

- **Total evaporation (e)** showed significantly higher (i.e. lower values, since negative values indicate evaporation and positive values indicate condensation (ECMWF documentation)) compared to previous years.
- **Surface solar radiation downwards (ssrd)** and **total precipitation (tp)** showed a rapid decrease for 2023 relative to the previous years.

Even though the values for these variables for year 2023 lie in fairly probable range (see Figure 2), we ultimately decide to remove year 2023 data.

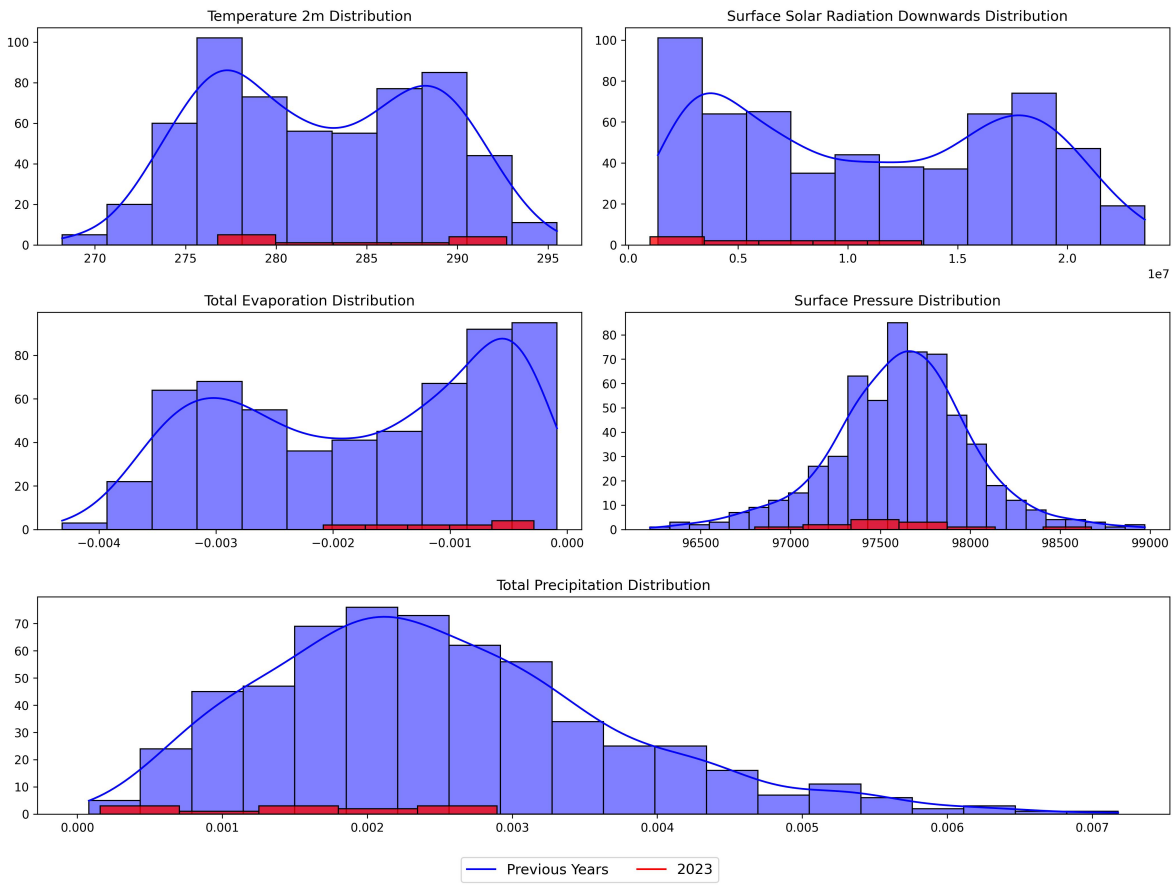

Figure 2: Distribution Comparisons

1.7 Summary Statistics

We compute summary statistics, such as mean, median, variance, minimum, and maximum values, for the chosen environmental variables to gain insights into their overall distribution and variability. These metrics allow us to understand the central tendencies (mean, median), the spread of data (variance), and the range (min, max) for each variable. While the table provides an aggregate snapshot of these statistics across the entire dataset, the Figure 3 shows how these summary statistics evolve over time.

	t2m	ssrd	e	sp	tp
Mean	282.422346	1.107641e+07	0.001751	97616.47083	0.002494
Median	282.128540	1.095859e+07	0.001585	97628.67250	0.002344
Variance	36.988568	4.239948e+13	0.000001	155287.66908	0.000001
Min	268.188780	1.345306e+06	0.000086	96218.16000	0.000079
Max	295.495480	2.353309e+07	0.004316	98971.14000	0.007179

Table 1: Summary statistics for the variables **t2m, ssrd, e, sp,** and **tp**.

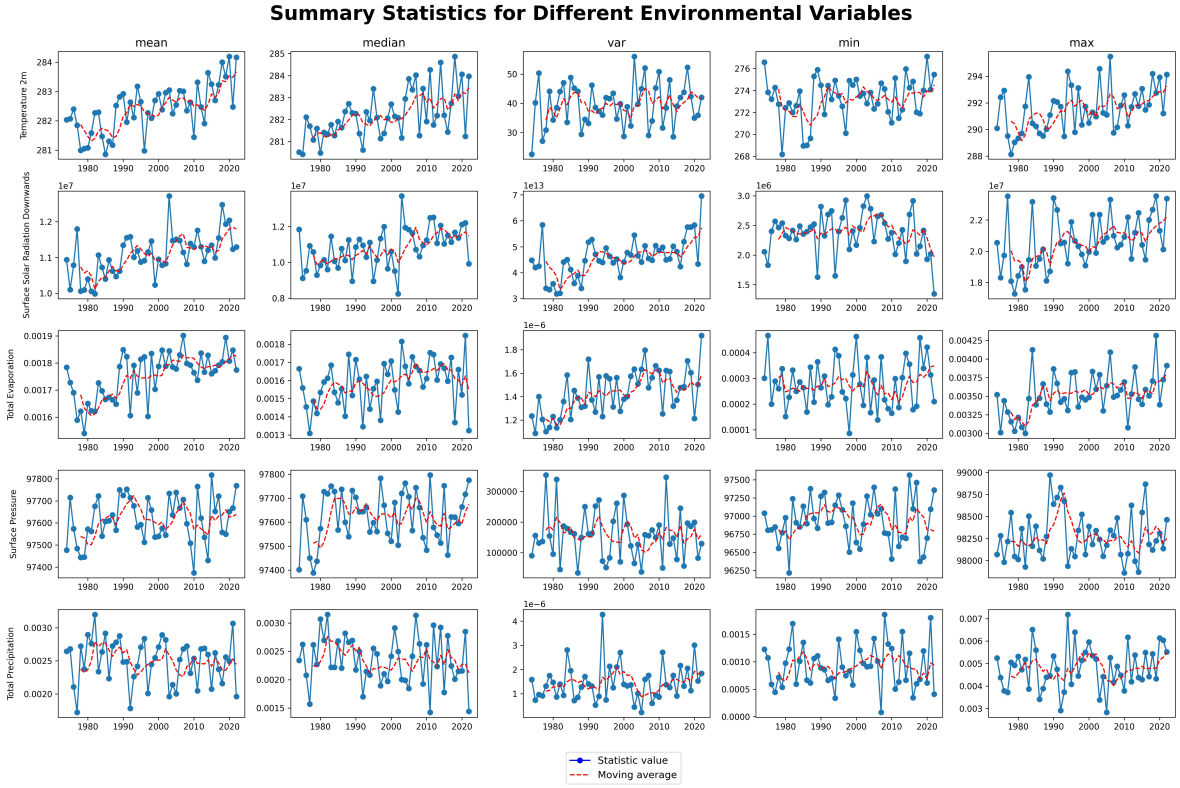

Figure 3: Summary Statistics & Rolling Average

2 Correlation Analysis

To perform correlation analysis we used the Pearson correlation coefficient, which can be used to summarize the strength of the linear relationship between two variables.

$$\text{Pearson} = \frac{\text{COV}(X, Y)}{\text{STD}(X) \cdot \text{STD}(Y)}$$

Note that the Pearson correlation coefficient is used for continuous variables.
From Table 2 and Figure 4 we have the following:

- Significant Correlations (p-value < 0.05):
  - **t2m** and **e**: p-value $= 0.0$
  - **t2m** and **ssrd**: p-value $= 0.0$
  - **t2m** and **tp**: p-value $= 0.012$
  - **t2m** and **sp**: p-value $= 0.027$
  - **ssrd** and **e**: p-value $= 0.0$
  - **ssrd** and **tp**: p-value $= 0.0$
  - **sp** and **tp**: p-value $= 0.0$
- Statistically Insignificant Correlations (p-value > 0.05):
  - **ssrd** and **sp**: p-value $= 0.634$
  - **e** and **sp**: p-value $= 0.735$
  - **e** and **tp**: p-value $= 0.103$

Most of these results are self-explanatory. For example, there is a positive correlation between temperature and surface solar radition, which is natural: radiant energy from the Sun transforms into thermal energy. We can summarize these findings as follows:

- Positive correlations:
  - **t2m** and **ssrd**: Solar radiation is a major driver of surface temperature. Increased solar radiation heats the surface, causing an increase in temperature.
  - **t2m** and **e**: Higher surface temperatures promote more evaporation, as warmer air has a higher capacity to hold water vapor. This can increase evaporation rates from bodies of water, soil, and vegetation.
  - **ssrd** and **e**: More solar radiation provides the energy needed for evaporation. As solar radiation increases, more energy is available to convert liquid water into vapor, leading to higher evaporation rates.
- Negative correlations:
  - **t2m** and **tp**: This inverse relationship shows that higher surface temperatures are be associated with lower precipitation, i.e. drier and hotter conditions reduce precipitation.
  - **ssrd** and **tp**: High solar radiation values typically occur during clear skies, which indicate drier conditions with less cloud cover and less precipitation.
  - **sp** and **tp**: Higher surface pressure is associated with anticyclonic conditions, which are typically dry and stable, leading to less precipitation.'
- No correaltions:
  - **t2m** and **sp**: The weak relationship between surface temperature and surface pressure indicates that these two variables do not directly influence each other in a significant way. Which is natural, since atmospheric pressure is caused by the force of gravity.

	t2m	**ssrd**	**e**	**sp**	**tp**
t2m	0.0	0.0	0.0	0.0269	0.0124
ssrd	0.0	0.0	0.0	0.6338	0.0
e	0.0	0.0	0.0	0.7346	0.1032
sp	0.0269	0.6338	0.7346	0.0	0.0
tp	0.0124	0.0	0.1032	0.0	0.0

Table 2: P-values for Pearson correlation coefficients between variables.

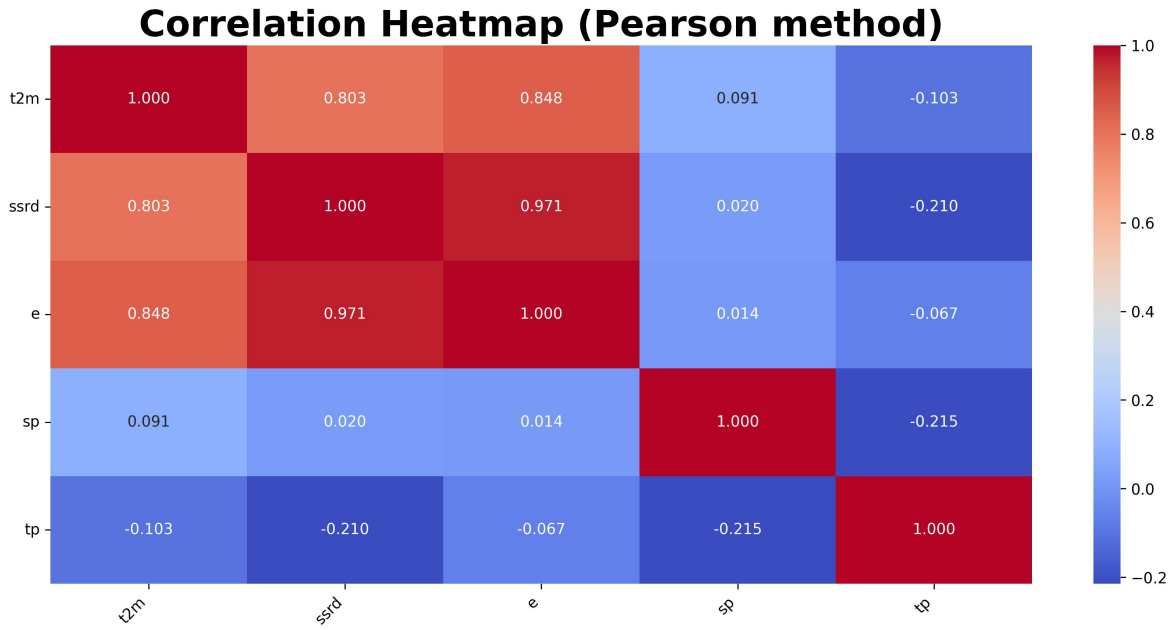

Figure 4: Correlation Coefficients between the Environmental Variables

3 Trend Analysis

In this part we will be analyzing the Temperature 2m over the years. To do so we will fit a regression model to our data using the **statsmodels** library and plot the trend line for our data with **matplotlib** for visualization. Finally we will analyze and interpret the observed results.

3.1 Model Summary

After preparing the data by selecting the **year** as our independent variable and **t2m** (Temperature 2m) as our dependent variable, we fit a linear model (OLS) to the data and get the results in Table 3.

Table 3: OLS Regression Results

Dep. Variable:	t2m			
Model:	OLS			
Method:	Least Squares			
Time:	20:21:39			
No. Observations:	588			
Df Residuals:	586			
Df Model:	1			
Covariance Type:	nonrobust			
	coef	**std err**	**t**	**P>t**
const	204.1045	35.318	5.779	0.000
year	0.0392	0.018	2.218	0.027
Omnibus:	298.055			
Durbin-Watson:	0.378			
Prob(Omnibus):	0.000			
Jarque-Bera (JB):	33.116			
Skew:	0.020			
Prob(JB):	6.44e-08			
Kurtosis:	1.838			
Cond. No.:	2.82e+05			

Standard Errors assume that the covariance matrix of the errors is correctly specified. The condition number is large, 2.82e+05. This might indicate that there are strong multicollinearity or other numerical problems.

3.2 Scatter Plot

Using **matplotlib** we plot the data for **t2m** by year using a scatter plot draw the regression line we fitted in the previous step. We obtain the plot in Figure 5.

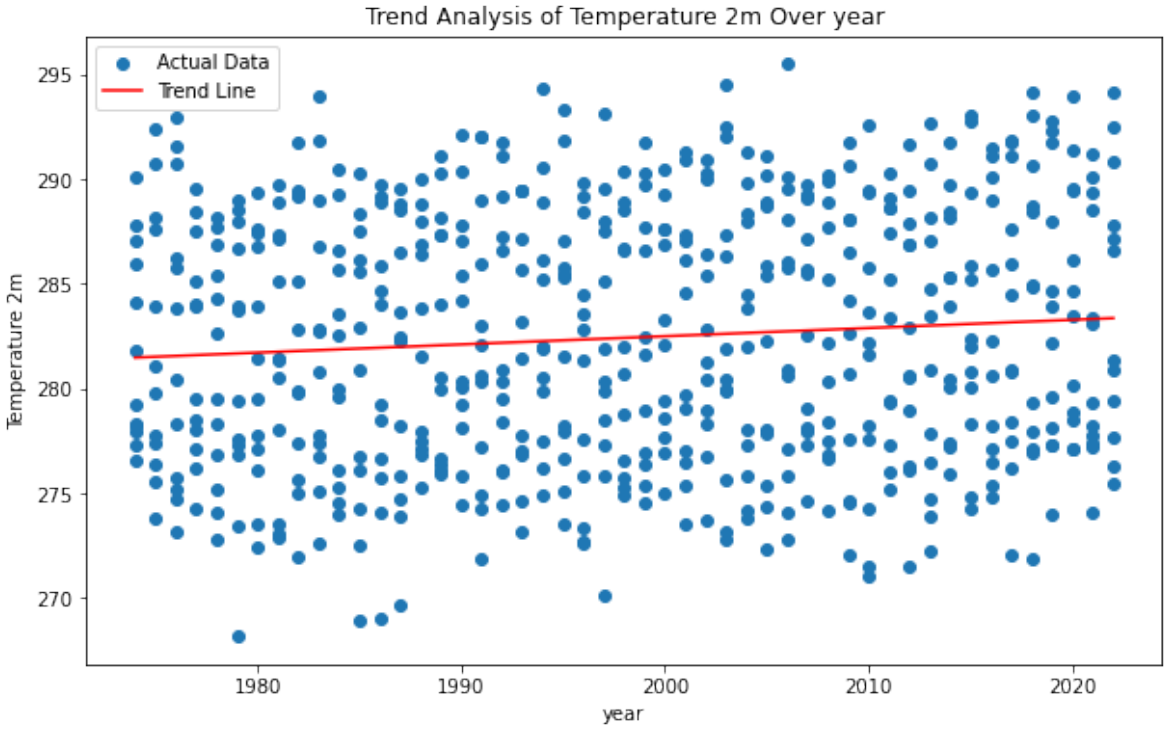

Figure 5: Trend analysis of Temperature 2m over Years

3.3 Interpretation

Looking at the regression model's summary and the trend line, we can derive the following conclusions:

1. Coefficient for year
   - At **0.0392**, this means that `t2m` increases by 0.0392 approximately every year
   - With the **p-value** for the year coefficient at **0.027** (< 0.05), this indicates a statistically significant temperature increase over time at the 5% significance level
2. R-squared and F-statistic
   - With a very low R-squared value of **0.008**, only 0.8% of the variability in temperature is explained by the year
   - F-statistic is **4.918** with a p-value of **0.0270**, which again shows that the model is statistically significant overall
3. Durbin-Watson and Prob(JB)
   - The Durbin-Watson statistic is **0.378**, which is quite low and indicates that there may be some autocorrelation in the residuals
   - The very low p-value of the Jarque-Bera test **(6.44e-08)** suggests that the residuals are not normally distributed
4. Plot
   - The scatter plot shows that the temperature varies a lot over the years, but the slight increase in temperature (slope of 0.0392) is still visually visible on the red line