UNIVERSITÉ DE MONTPELLIER

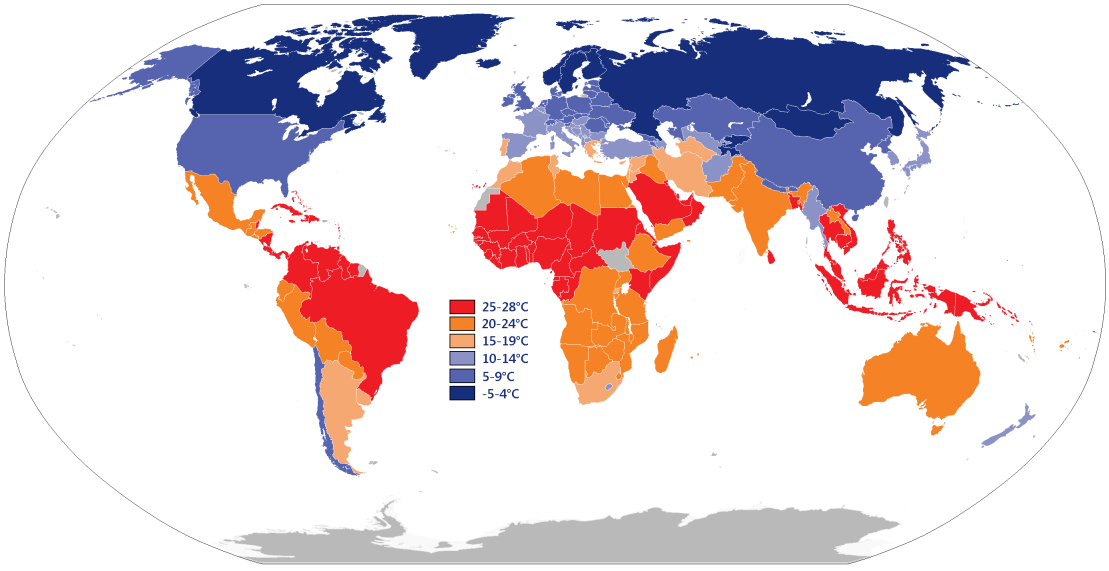
STATISTIQUE
SCIENCE DES DONNÉES
UNIVERSITÉ DE MONTPELLIER

Master 1 Statistics and Data Science

Project Report

Analysis of Meteorological Data: Comparative Analysis of a Climate Change Model and Historical Meteorological Observations

---



Thibault FERRETTI
Thomas LAUGIÉ

Supervised by: Mr. MÉTIVIER

May 28th, 2023

Contents

Acknowledgements

We wish to thank our advisor, David Métivier, for his guidance and assistance throughout the development of this project. His knowledge and patience have been greatly appreciated. We are grateful for the time and effort he has dedicated to helping us, which has contributed significantly to our project.

Introduction

The study of climate and its patterns has become increasingly important in recent years, with global warming being a major concern for the planet. Meteorological data can provide valuable insights into the behavior of climate and help us understand the long-term changes that may occur. Time series analysis is a powerful tool for exploring and analyzing meteorological data as it can help us identify patterns and trends over time.

In this project report, our main objective is to conduct in-depth analysis of meteorological data using advanced time series analysis techniques. We will adopt a comprehensive approach, starting with a thorough exploratory analysis to gain a better understanding of the structure and characteristics of the data. This will include visualizations of the time series, studies of variable distributions, as well as measures of correlation between different meteorological stations. This step will allow us to uncover general trends and detect any relationships.

Next, we will focus on identifying patterns and cycles present in the data. We will use techniques such as decomposing the time series into its seasonal, trend, and residual components. This will help us understand the seasonal patterns, long-term trends, and irregular variations present in the meteorological data.

After exploring patterns and cycles, we will move on to modeling. We will employ statistical models such as SARIMA (Seasonal Auto Regressive Integrated Moving Average) and SARIMAX (SARIMA with Exogenous variables) to capture seasonal structures, temporal dependencies, and potentially the influence of external variables on the meteorological data.

Finally, we will evaluate the performance of our models using metrics such as Root Mean Square Error (RMSE) and information criteria such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). This will help us assess the quality of our models and compare their respective performances.

The objectives of this project is to get a better understanding of weather patterns and trends from the available data. Secondly, we want to assess the potential impact of climate change by examining long-term variations in the time series.

In-depth analysis of meteorological data using advanced time series techniques can provide valuable insights for policymakers, researchers, and professionals in the environmental field. The findings from this report can contribute to a better understanding of climate trends.

1 Why Did We Choose Julia?

Julia is a high-level programming language specifically designed for numerical computation and data analysis. Here are a few reasons why we chose Julia for this project:

1.1 Advantages

- **Performance**: Julia is designed to be fast and is often compared to C in terms of performance.
- **Ease of Use**: Julia has a clear and intuitive syntax, making it easier to learn and use.
- **Compatibility**: Julia can call C, Python, or R code allowing for seamless integration with existing libraries.
- **Parallel and Distributed Computing**: Julia has built-in features for parallel and distributed computing.
- **Dynamic Typing**: Julia allows for greater flexibility in handling data due to its dynamic typing.

1.2 Disadvantages

- **Smaller Community**: Compared to Python or R, Julia has a smaller community.
- **Language Maturity**: Julia is a relatively new language and may present stability issues.
- **Fewer Available Libraries**: Julia does not have as many dedicated libraries as Python or R.
- **"Time To First Plot"**: Julia, as a compiled language, occasionally exhibits slower performance during the initial execution of code.

1.3 Julia vs Python and R

We chose Julia over Python or R for several reasons. Julia offers superior performance, ease of parallel computing, and a syntax closer to standard mathematical notation. Compared to R, Julia is faster, more versatile, and easier for parallel computing.

2 Exploratory Analysis

2.1 Dataset Exploration

The meteorological data used in this project is collected from several weather stations located across Europe, with a particular focus on stations situated in France.
The data includes measurements of temperature, humidity, precipitation, wind speed, and direction, among other variables.
In this report we deal only the temperature and precipitation variables. Our main focus being on the mean daily temperature. The data covers a period of several years, from 1900 to 2020, allowing us to investigate long-term patterns in the available data.
The data used in this project can be found here: https://www.ecad.eu/dailydata/predefinedseries.php

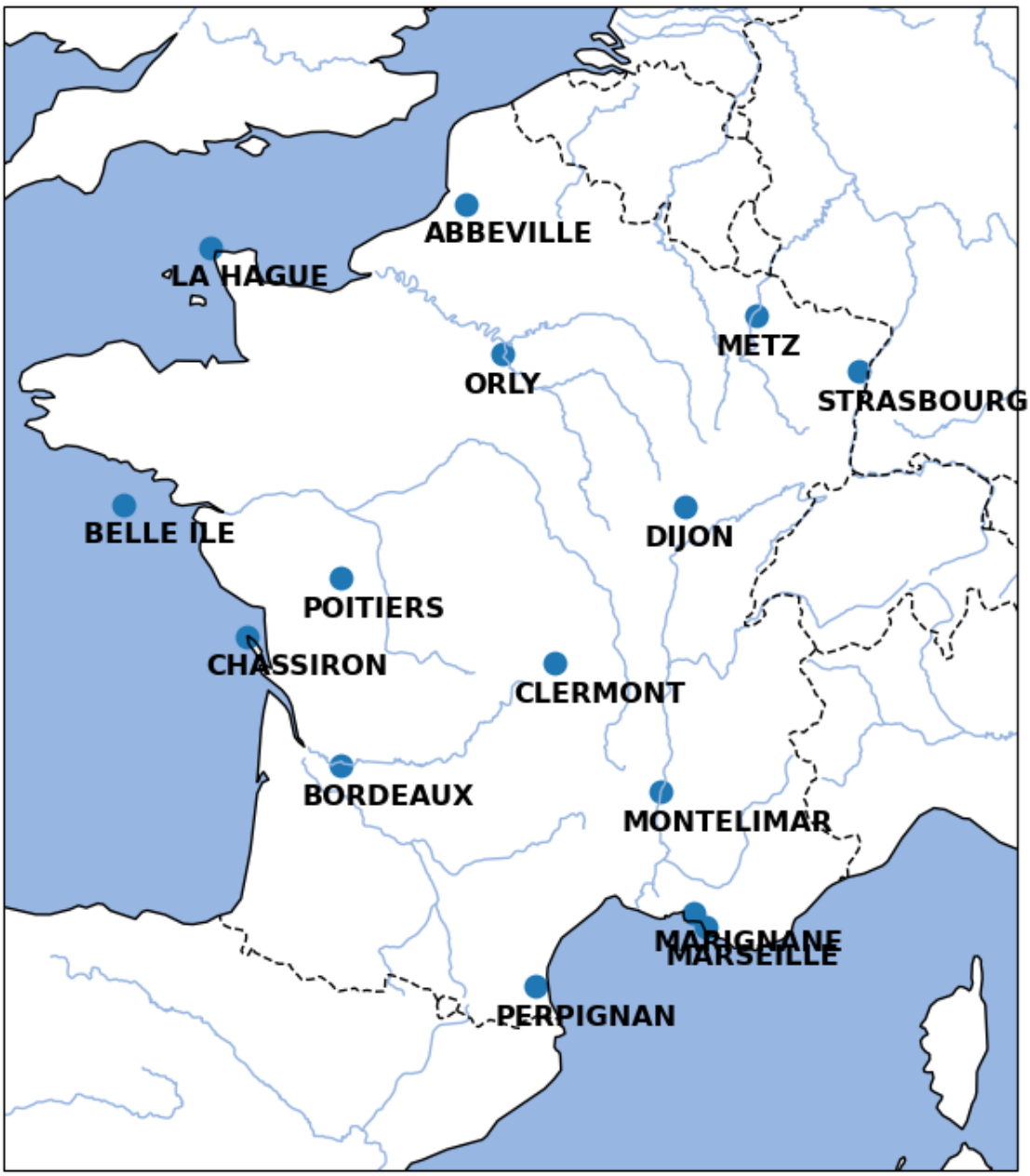


Figure 1: Location of the stations.

Here are the 15 stations that we will take for the study of the correlations. As for the study of trends, we will only consider the Orly station, similar analysis can be performed on the other stations.
We have chosen to focus on the Orly station for the study of trends due to its central location within France. This positioning allows it to be representative of a broad geographic area.

2.2 Correlation

Since we are dealing with multivariate Time Series across multiple locations. We can measure correlation in different ways.

2.2.1 Correlation heatmap

We will now turn our attention to a pairwise correlation analysis of temperatures across the various meteorological stations. This process involves evaluating the degree to which changes in temperature at one station correspond to changes at another.

By conducting this analysis, we aim to better understand the relationships and interactions among different geographical locations, which can offer valuable insights into regional climate patterns. Particularly, we will be focusing on how temperatures correlate between pairs of stations, providing us with a detailed understanding of the interconnectedness of these different areas.

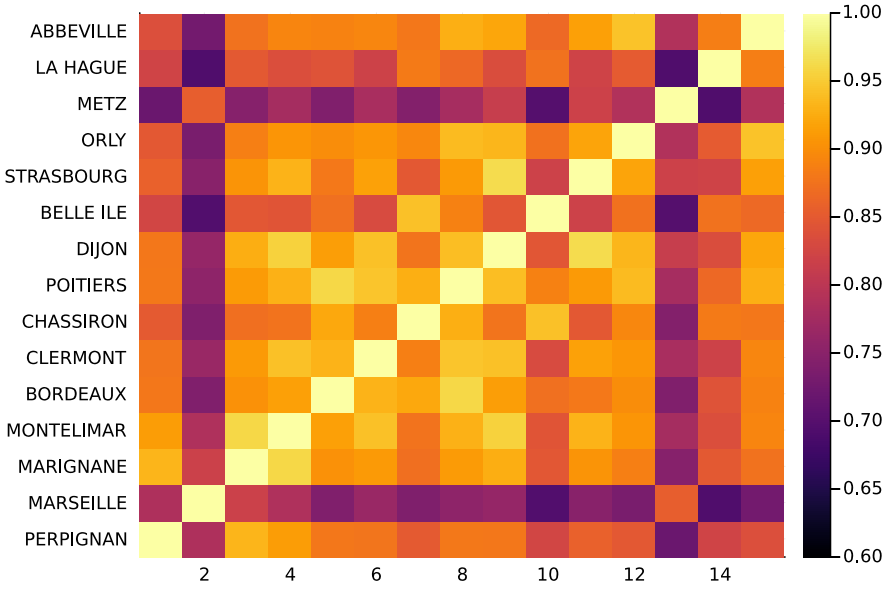


Figure 2: Pairwise correlation among different meteorological stations.

The stations are ordered from North to South. The first thing that is noticeable and expected is that the correlation are all positives. All stations being located in the same hemisphere a positive change in temperature at one station is likely to be seen on another station.

The least correlated stations appears to be Marseille and Metz, exhibiting a correlation of approximately 0.7 with other stations. This result might be attributed to the substantial geographical and climatic differences between the two cities: Marseille being located in the South by the Mediterranean sea and Metz situated in the North-Eastern part of the country.
One notable finding is the relatively low correlation between Marseille and Marignane. Being two cities that are really close, we expect them to have a high correlation. This can maybe be attributed to the altitudes of the stations, Marignane being at an altitude of 9 meters and Marseille at an altitude of 75 meters.

In contrast, most of the other station pairs demonstrate a higher level of correlation, around 0.8 to 0.9. This suggests that temperature changes in one location are likely to be accompanied by similar changes in the other, especially for geographically close stations. These findings underline the influence of regional climatic conditions on temperature correlation across different locations in France.

2.2.2 Correlation with respect to distance

Given our extensive dataset comprising of multiple stations, we aim at assessing the impact of distance on our analysis. Leveraging the latitude and longitude coordinates of each location, we will use the Haversine formula to accurately measure this effect.

Theorem 1 (Haversine Formula) *Let:*

- r *be the radius,*
- φ_1 *and* φ_2 *be the latitude of point 1 and point 2,*
- λ_1 *and* λ_2 *be the longitude of point 1 and point 2.*

Then the great-circle distance between point 1 and point 2 is equal to:

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos\varphi_1 \cdot \cos\varphi_2 \cdot \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

Here is the implementation of this function that we have done in Julia.

```
function haversine(ID1, ID2 = 11249)
    """
    Calculate the great circle distance in kilometers between two points
    on the earth (specified in decimal degrees).
    Compare with Orly by default.
    """
    # Convert decimal degrees to radians

    lon1, lat1 = get_lat_lon(ID1)
    lon2, lat2 = get_lat_lon(ID2)
    lon1, lat1, lon2, lat2 = deg2rad.([lon1, lat1, lon2, lat2])

    # Haversine formula
    dlon = lon2 - lon1
    dlat = lat2 - lat1
    a = sin(dlat / 2)^2 + cos(lat1) * cos(lat2) * sin(dlon / 2)^2
    c = 2 * asin(sqrt(a))
    r = 6371 # Radius of earth in kilometers.
    return c * r
end
```

Our reference point for this measurement will be station 11249 located in Orly, France.

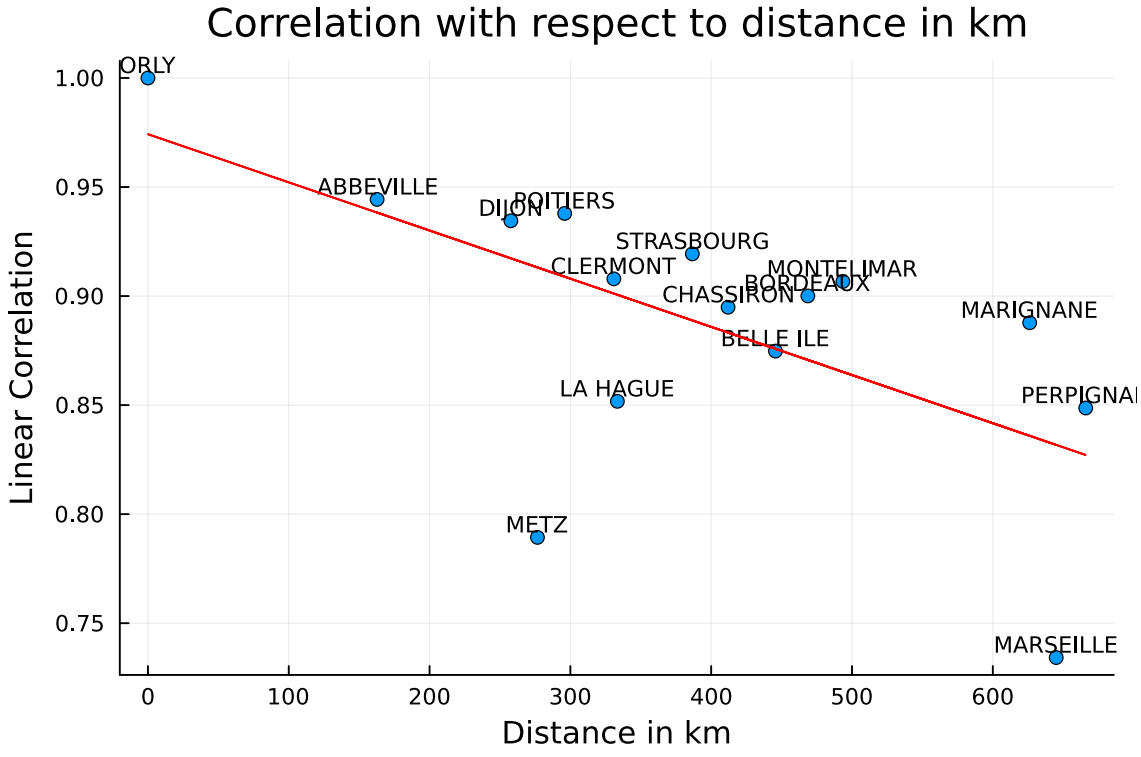


Figure 3: Correlation with respect to distance.

From the graphical representation, it is apparent that there is a linearly decreasing trend. This suggests that as the distance increases, the temperature demonstrates a tendency to evolve correspondingly but the decline does not follow a linear pattern but rather a power law distribution.

However, it is important to note that there are some outliers in the data. For instance, Marseille or Metz appears as an extreme point. This is possibly due to its coastal location, resulting in

warmer temperatures than would be expected based on distance alone. This instance underlines the complexity of temperature variations and how they can be influenced by local geographical factors beyond simple distance measurements. Such considerations are critical when interpreting the results of climate model forecasts.

In the interest of refining our model, we undertook an additional analysis by excluding the critical points from our dataset, as these can disproportionately influence our linear regression model. These outliers, as previously mentioned, might be influenced by specific local geographical factors.

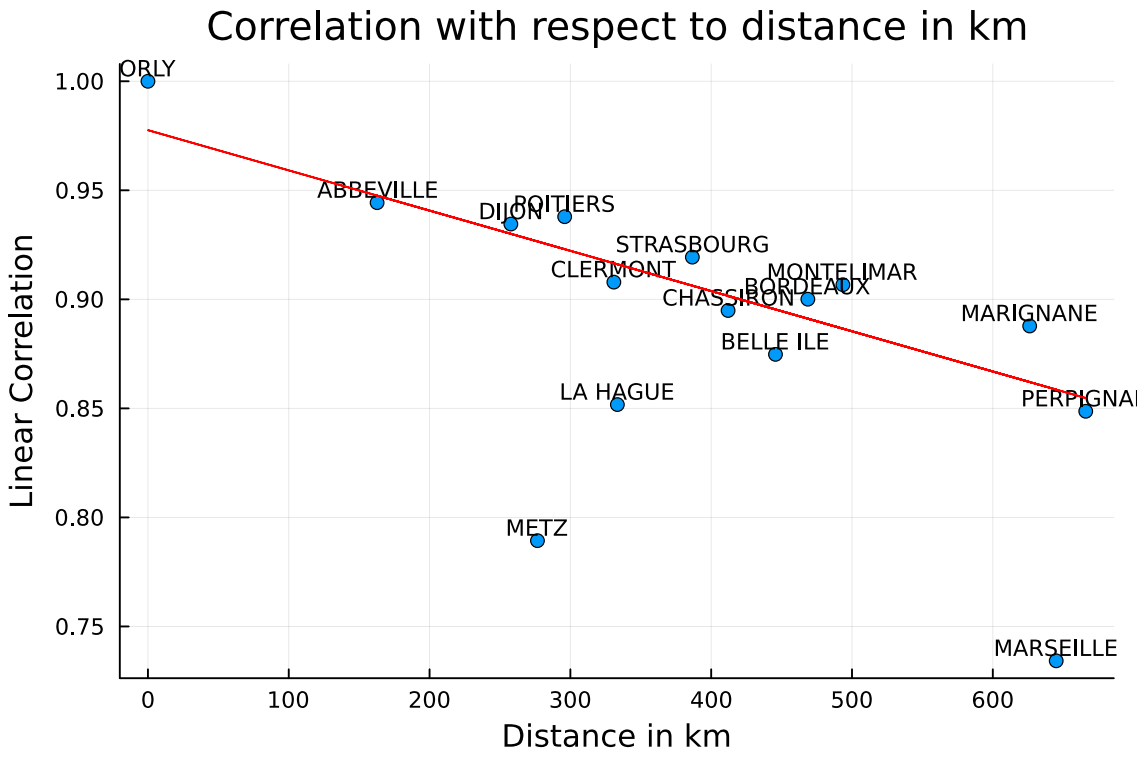


Figure 4: Adjusted correlation between temperature and distance

Upon re-evaluating the correlation excluding these critical points (the critical points were added on top of the graph but they do not influence the correlation), we observe that the points align more closely to the regression line. This indicates an improvement in our model, thereby offering a better fit to the data. This refined model, without the outliers, can provide a more accurate understanding of the general trend between temperature and geographical distance, excluding specific local anomalies.

2.2.3 Tail dependence heatmap

In addition to the standard correlation analysis, we also conducted a tail dependence correlation analysis between pairs of stations. This form of analysis focuses on the extreme values of the datasets, which are particularly relevant when considering phenomena such as heatwaves or cold spells.

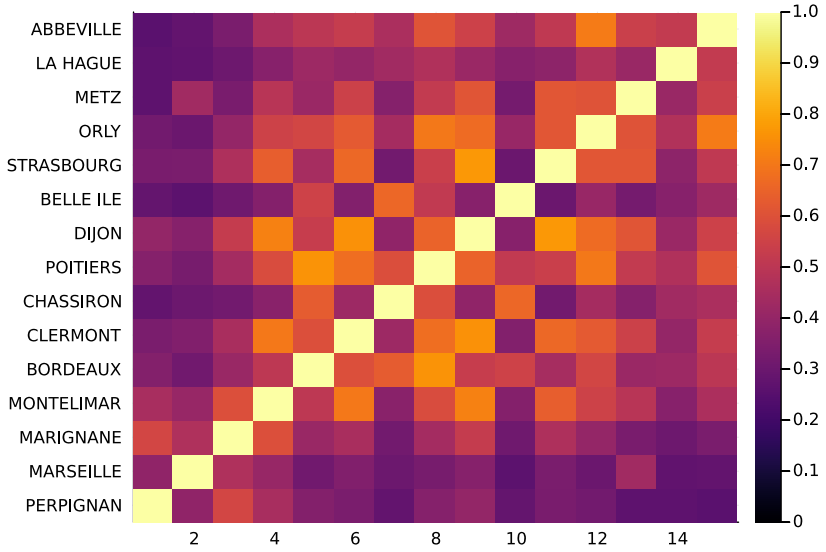


Figure 5: Tail Dependence Correlations Between Pairs of Stations

Interestingly, we found that there was virtually no tail dependence correlation for extreme values when the stations were somewhat distant from each other. This result is consistent with our expectations, as extreme temperature events in the south of France are not necessarily linked to similar events occurring in the north. These findings underscore the complex, localized nature of extreme weather events.

2.2.4 Tail dependence with respect to distance

Tail dependence is a statistical concept used to measure the probability that two random variables exceed a certain threshold together. It's a measure of the dependence between the extremes of the distributions of two variables.

It is used in extreme value theory to understand the joint behavior of two variables when their values are very large (or very small). In a more general context, it amounts to analyzing the dependence between two variables when they reach exceptionally high or low values, respectively in the upper or lower part of their distributions.

For instance, random variables that show no correlation can display tail dependence in extreme deviations.

Tail dependence can be quantified by tail dependence coefficients. For two random variables X and Y, these coefficients, usually denoted as L_T for the upper part of the distribution and L_B for the lower part, are defined as the limit of the probability that Y exceeds a high threshold, given that X has also exceeded this threshold.

For upper tail dependence :

$$L_T(u) = \lim_{t \to +\infty} P(Y > t | X > t) = \frac{P[X - F_X^{-1}(u) > 0, Y - F_Y^{-1}(u) > 0]}{1 - u}$$

For lower tail dependence :

$$L_B(u) = \lim_{t \to 0^+} P(Y \leq t | X \leq t) = \frac{P[X - F_X^{-1}(1-u) \leq 0, Y - F_Y^{-1}(1-u) \leq 0]}{u}$$

With :

- u in the interval (0,1)
- F_X and F_Y are the distribution functions of X and Y respectively
- F_X^{-1} and F_Y^{-1} are the corresponding quantile functions. If these measures are close to 1, it indicates a strong dependence between the extremes of X and Y. If they are close to 0, it indicates an independence of the extremes.

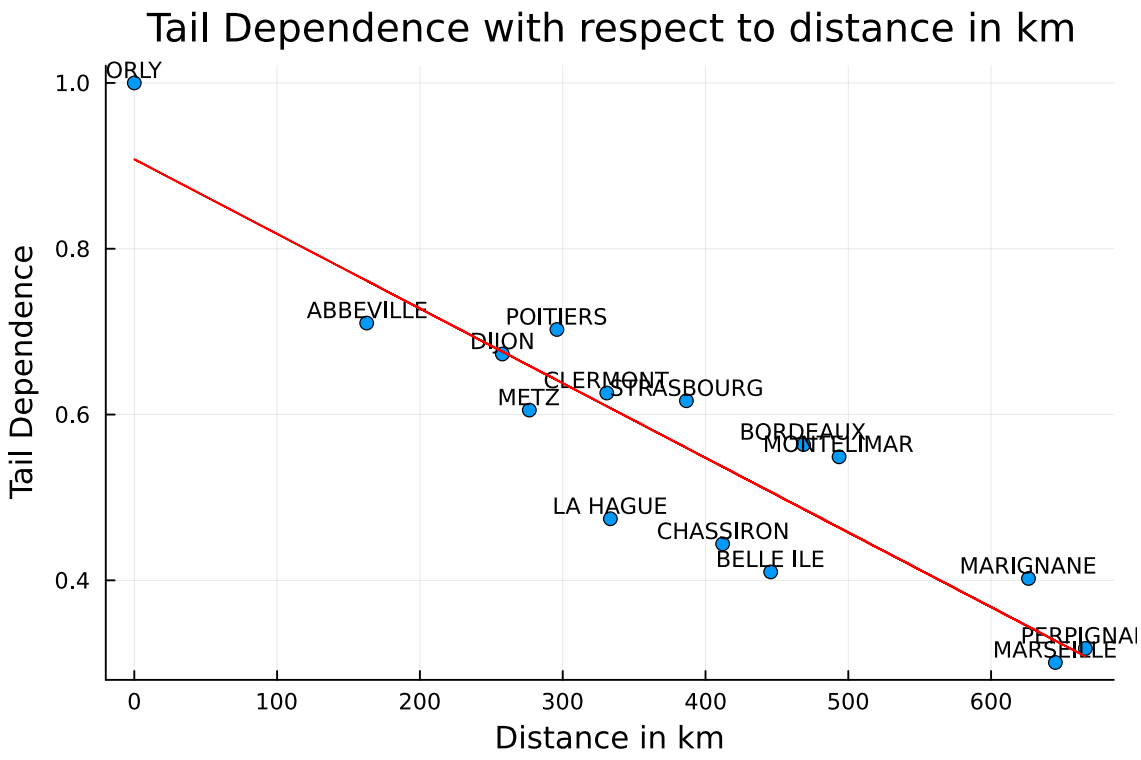


Figure 6: Tail dependence between temperature and distance

Based on our tail dependence analysis, it can be concluded that the model exhibits a strong fit for extreme values. This is an important observation as extreme weather conditions are crucial in climate studies, and their accurate modeling can provide vital insights for understanding and predicting significant climatic events.

Consequently, we observe that the tail dependence model is more consistent with extreme values, we don't have any outliers.

3 Trend and Model Selection

3.1 Time Series and trend

3.1.1 Analysis of Temperature

Our initial approach to analysis was to plot the complete time series of temperatures. However, given the vast amount of data collected over several decades from our chosen reference station, Orly, the resulting graph was difficult to interpret 7.

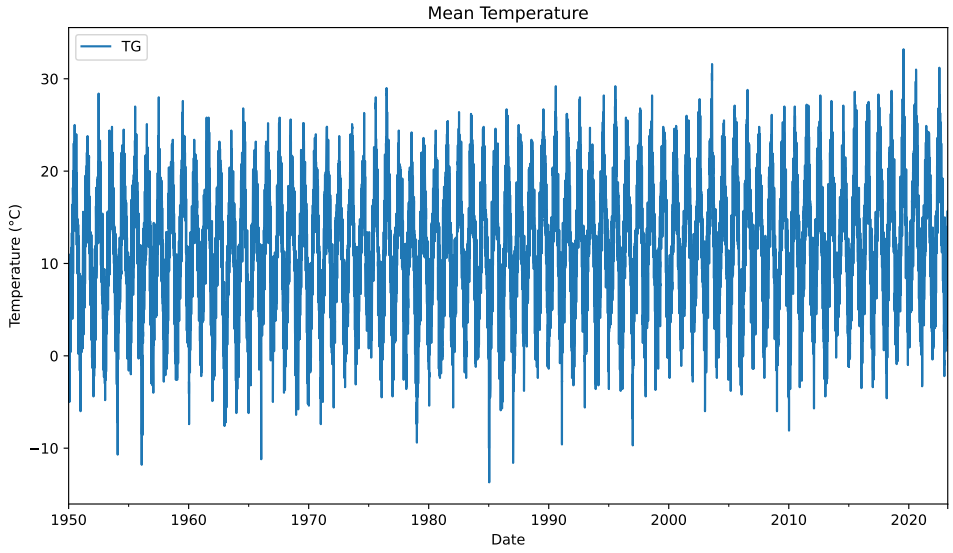


Figure 7: Evolution of average monthly temperature from 1950 to 2020.

As a result, we decided to break down the data by month for a more granular analysis.

Upon examining the average temperature of each individual month over the studied period, an overall trend of increase was observed. As shown in Figures 8, each month exhibited an increase in average temperature over time, suggesting a persistent and global warming. To determine the extent and certainty of this increase, a linear regression was drawn on each graph.

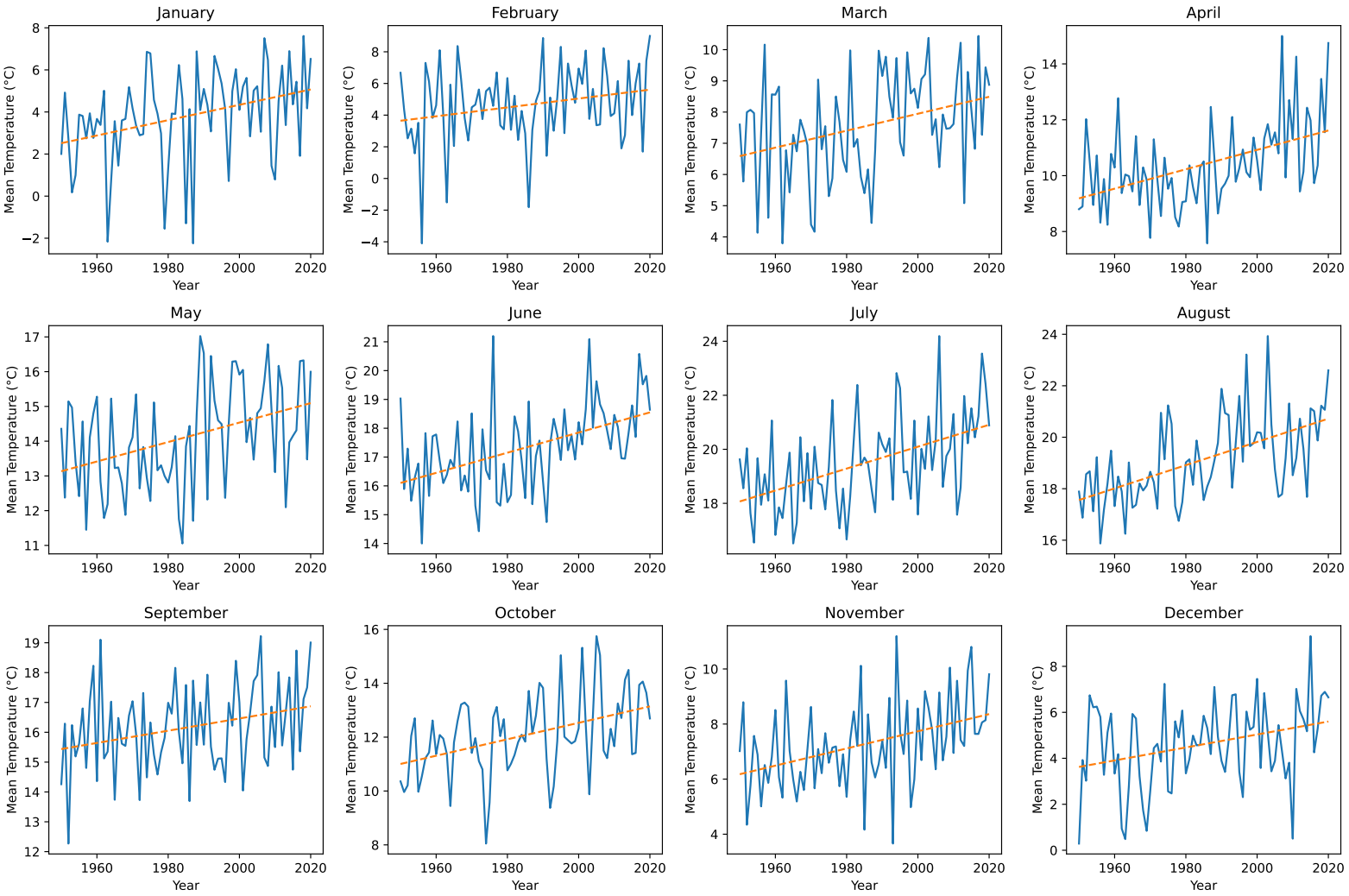


Figure 8: Evolution of average monthly temperature from 1950 to 2020 with linear regression.

To gain a broader understanding of the phenomenon, we also calculated the annual average temperature and plotted it over the entire study period. As illustrated in Figure 9, an upward trend is also observed at this scale, confirmed by the linear regression line.

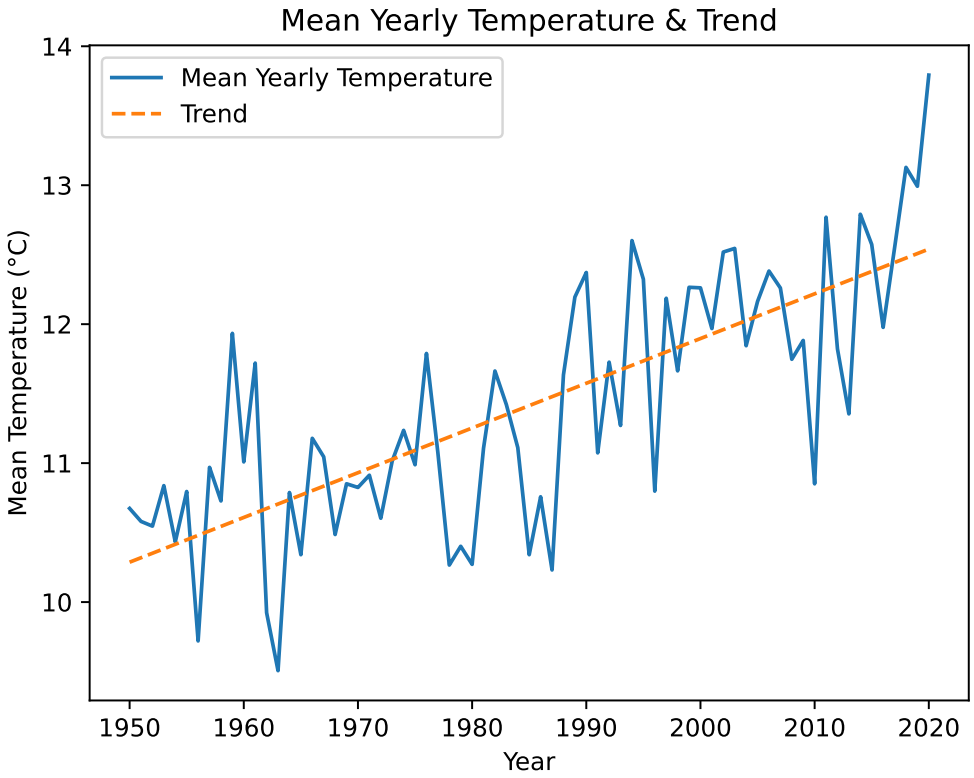


Figure 9: Evolution of average annual temperature from 1950 to 2020 with linear regression.

In addition to analyzing the overall and monthly trends, we also decompose the temperature time series into its constituent parts: trend, seasonality, and residuals. This decomposition allows us to study these components separately, and gain a deeper understanding of the underlying patterns in our data.

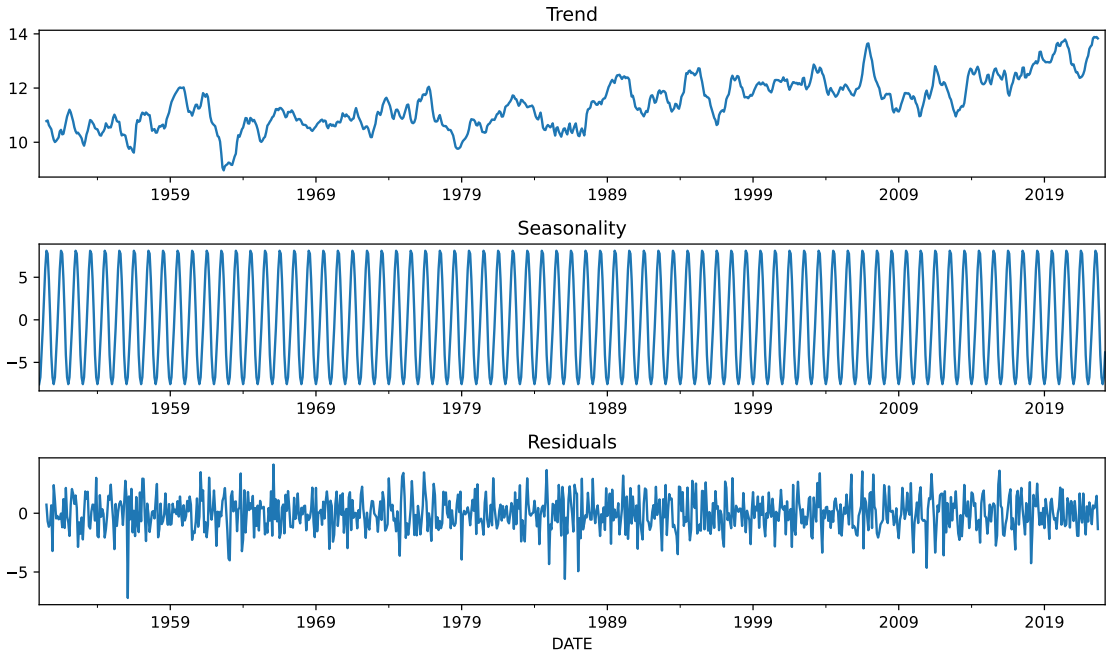


Figure 10: Decomposition of temperature time series into trend, seasonality, and residuals.

The 'trend' component in Figure 10 shows the long-term progression of temperatures, revealing an overall increase over the studied period.

The 'seasonality' component captures the regular, predictable changes in temperature associated with the changing seasons.

Lastly, the 'residuals' component captures the random, unpredictable fluctuations in temperature that are not explained by the trend or seasonality.

This could be due to a variety of factors, including weather events, changes in solar radiation, and other phenomena not captured by the model.

The upward trend aligns with the global narrative of climate change, demonstrating that the effects of a warming climate are tangible and observable in our data.

3.1.2 Analysis of Precipitation

Similar to temperature, we also performed an analysis on the time series of precipitation at Orly. Our first step was to plot the data in the form of a histogram to examine the distribution of rainfall. This histogram allowed us to visualize the distribution of daily precipitation amounts. To improve the visibility of the data, we applied a logarithmic scale to the histogram, as shown in Figure 11.

By using a logarithmic scale, we were able to better observe the distribution of precipitation values across different ranges. This scaling method enhances the representation of smaller values while still maintaining the relative proportions between larger values. It allows for a clearer visualization of the distribution, especially when dealing with a wide range of values, such as precipitation amounts.

As depicted in Figure 11, the majority of days exhibit little to no rainfall, which aligns with the common perception that rainfall is generally infrequent. However, we can also observe a smaller number of days with higher precipitation amounts, indicating occasional periods of significant rainfall.

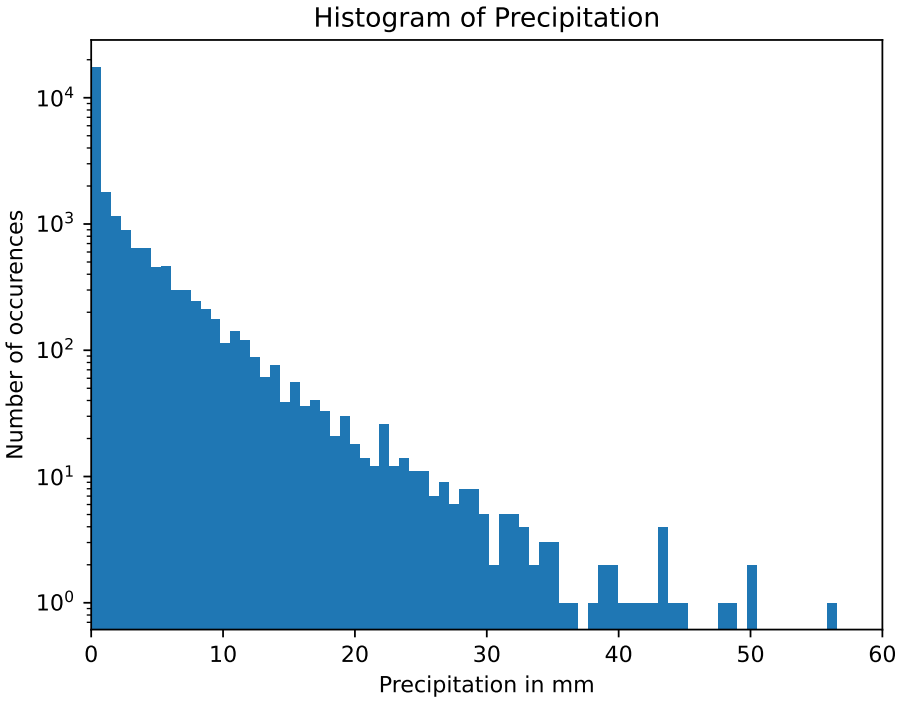


Figure 11: Histogram of daily precipitation from 1950 to 2020.

Next, we created a plot to examine whether precipitation levels have increased or decreased over the years. As shown in Figure 12, there is a slight upward trend in the annual average precipitation, suggesting that there may be an increase in precipitation over time. However, this trend is extremely mild and, given the variability of the data, may not be statistically significant.

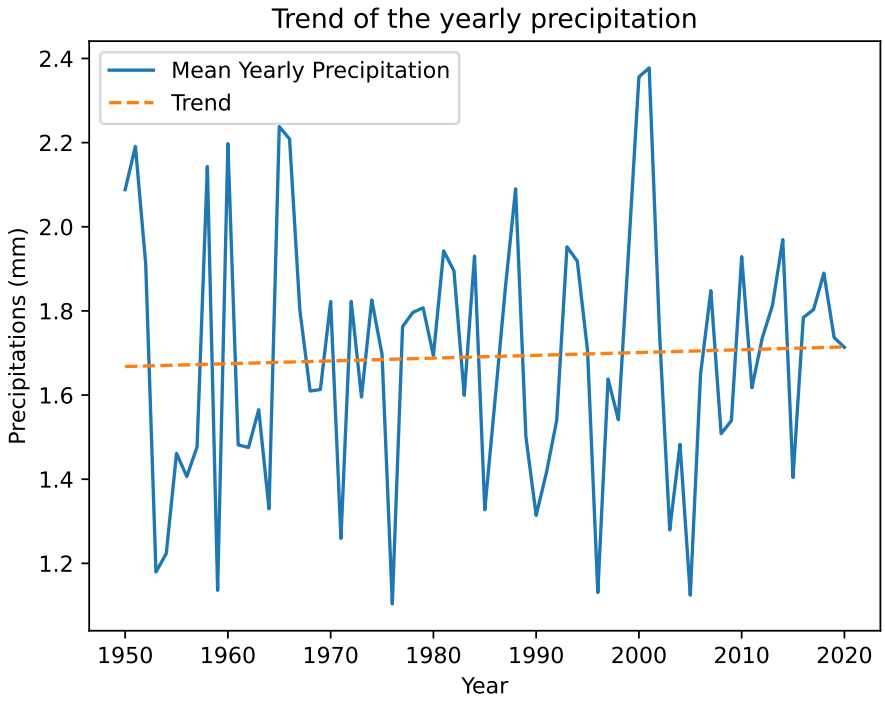


Figure 12: Evolution of average annual precipitation from 1950 to 2020.

3.1.3 Linear Regression

Linear regression is a statistical method for modeling the relationship between a dependent variable (response) and one or more independent variables (predictors). It aims to establish the best linear relationship between the variables, which can subsequently be used to predict the dependent variable values based on the independent variables.

- **Simple Linear Regression:** The simple linear regression model, involving only one independent variable, is defined by the equation:

$$y = a + bx \tag{1}$$

  where y is the dependent variable, a is the intercept (the value of y when x is zero), b is the slope (the rate of change of y with respect to changes in x), and x is the independent variable.

- **Multiple Linear Regression:** Multiple linear regression involves more than one independent variable and is defined by the equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{2}$$

  where $\beta_0, \beta_1, \ldots, \beta_k$ are the coefficients to be estimated, and $x_1, x_2, \ldots, x_k$ are the independent variables.

- **Least Squares:** The coefficients in linear regression models are estimated using a method called least squares. This method aims to find the line that minimizes the sum of the squared differences between the observed and the predicted values. The estimates of a and b are given by:

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \tag{4}$$

  where $\bar{x}$ and $\bar{y}$ are the means of x_i and y_i, and n is the number of observations.

3.2 SARIMA

The SARIMA model is a time series model that is used to analyze and forecast data with seasonal patterns. It is an extension of the ARIMA (AutoRegressive Integrated Moving Average) model, which is commonly used for non-seasonal time series data. The SARIMA model includes additional parameters to account for the seasonal component of the data.

The SARIMA model is specified by three sets of parameters: (p, d, q) for the non-seasonal component, (P, D, Q) for the seasonal component, and s for the length of the seasonal period. The non-seasonal component of the model is similar to the ARIMA model, while the seasonal component is based on the seasonal differences and seasonal lags of the time series data.

To estimate the parameters of the SARIMA model, various methods can be used, including maximum likelihood estimation and least squares. Once the model is fitted to the data, it can be used to generate forecasts for future time periods.

3.2.1 ARIMA Model

The **ARIMA** (AutoRegressive Integrated Moving Average) is a class of models that are used for analyzing and forecasting time series data. This methodology caters specifically to data showing evidence of non-stationarity and can be used to fit both seasonal and non-seasonal models.

The ARIMA model is specified by three order parameters: (p, d, q).

- 'p' is the order of the AutoRegressive part. This component captures the influence of the previous periods' values on the current period's value. In essence, it's the number of lag observations included in the model.
- 'd' is the order of differencing required to make the time series stationary. Differencing is a transformation used to eliminate the trend component of the time series or to remove the seasonal dependencies. It involves subtracting the observation in the current period with that in the previous period. If the first difference does not make a series stationary, a second difference might be needed and so on.
- 'q' is the order of the Moving Average part. This component accounts for the impact of the error term (the difference between the predicted and the actual values) from the previous periods. It's the number of lag error terms that should go into the model.

3.2.2 Autoregressive Processes AR(p)

Definition 3.1 (Autoregressive Process) *An autoregressive process of order p (AR(p)) is a second-order stationary process that is a solution to the equation* $X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t$, *where* Z_t *follows a standard normal distribution and* $\phi_1, \ldots, \phi_p$ *are real numbers.*

Properties:

- Causality of an AR(p) process An AR(p) process is said to be causal if there exists an absolutely summable sequence $(\psi_k)_{k\geq 0}$ such that $X_t = \sum_{k\geq 0} \psi_k Z_{t-k}$. That is, X_t is expressed in terms of Z_t and its past. With an appropriate change of white noise, it is always possible to obtain a causal process.
- Autocorrelation function Assuming that the process (X_t) is a causal and centered AR(p), its autocorrelation function is given by $\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \frac{E[X_t X_{t-h}]}{E[X_t^2]}$.
  The AR(p) process is defined by $X_t = \sum_{k=1}^{p} \phi_k X_{t-k} + Z_t$ for all $t \in Z$.
  In calculating its variance, we get $\gamma(0) = \frac{\sigma^2}{1-\sum_{k=1}^{p} \phi_k \rho(k)}$. For $h \geq 1$, the autocorrelation is defined by $\rho(h) - \sum_{k=1}^{p} \phi_k \rho(h-k) = 0, h \geq 1$.
- Characterization of autocorrelations By solving the recurrence equation, it can be shown that if the roots $r_1, \ldots, r_p$ of the polynomial $\phi(x) = 1 - \sum_{k=1}^{p} \phi_k x^k$ are distinct and of modulus greater than 1, then the solution is of the form:

$$\rho(h) = c_1 \left(\frac{1}{r_1}\right)^h + c_2 \left(\frac{1}{r_2}\right)^h + \ldots + c_p \left(\frac{1}{r_p}\right)^h$$

where $c_1, c_2, \ldots, c_p$ are real constants. If the roots $r_1, r_2, \ldots, r_p$ are all of modulus strictly greater than 1, the autocorrelation function decreases either exponentially (if the roots are real) or in damped cycles (if the roots are complex).

Autoregressive processes, and especially their autocorrelation function, play a crucial role in time series analysis. This is because they help describe the dependence structure of the data over time, which can be exploited to better understand the underlying processes, make predictions, and inform decision-making. By fitting an AR(p) model to a time series, one can capture patterns and trends that may be otherwise invisible.

3.2.3 Moving Average Process MA(q)

Definition 3.2 (Moving Average Process) *A moving average process of order q, denoted MA(q), is defined by the equation:*

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$$

where $Z_t \sim N(0, \sigma^2)$ is a sequence of white noise and $\theta_1, \ldots, \theta_q$ are real numbers.

Using the properties of linear processes and denoting $X_t = \sum_{k=0}^{q} \theta_k Z_{t-k}$ with $\theta_0 = 1$, the autocovariance function can be derived as:

$$\gamma_X(h) = \begin{cases} \sigma^2 \sum_{k=0}^{q-h} \theta_k \theta_{k+h} & \text{if } 0 \leq h \leq q \\ \sigma^2 \sum_{k=0}^{q+h} \theta_k \theta_{k-h} & \text{if } -q \leq h \leq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Properties:

- Autocorrelation of an MA(q) process

  The autocovariance function of an MA(q) process is given by:

$$\gamma(h) = \begin{cases} \sigma^2 \sum_{k=0}^{q-|h|} \theta_k \theta_{k+|h|} & \text{if } |h| \leq q \\ 0 & \text{if } |h| > q \end{cases}$$

  and thus $\gamma(0) = \sigma^2(1 + \sum_{k=1}^{q} \theta_k^2)$, and

$$\rho(h) = \begin{cases} \frac{\sum_{k=0}^{q-|h|} \theta_k \theta_{k+|h|}}{1+\sum_{k=1}^{q} \theta_k^2} & \text{if } |h| \leq q \\ 0 & \text{if } |h| > q \end{cases}$$

  The autocorrelation function of an MA(q) process is null beyond the order q.

Moving Average (MA) models are fundamental components in time series analysis. They capture the dependencies in the error terms of a univariate time series. Specifically, an MA model expresses the current value of the series as a linear combination of past errors. This is particularly useful in capturing sudden shocks or random disturbances that affect the system under study.

Moreover, MA models are employed to smooth out short-term fluctuations and highlight longer-term trends or cycles. They can also be used to forecast future values, where the assumption is that the error terms follow a specific distribution, typically a normal distribution.

However, it's important to note that an MA model assumes that the series is stationary—i.e., its properties do not change over time. This often requires pre-processing steps such as differencing or transformations to ensure the data meets the assumptions of the model.

3.2.4 How to find parameters for SARIMA with ACF and PACF

The determination of the parameters of a SARIMA (Seasonal AutoRegressive Integrated Moving Average) model for a time series can be done using two main plots: the ACF (Autocorrelation Function) and the PACF (Partial Autocorrelation Function).

The ACF represents the correlation between the time series and itself lagged by 'k' periods. The PACF, on the other hand, represents the partial correlation which removes the effect of intermediate lags.

To determine the parameters, we must observe the ACF and PACF plots:

- To determine the 'p' (AutoRegressive) parameter: we observe the PACF. 'p' is generally the value at which the PACF cuts the upper confidence band for the first time.
- To determine the 'q' (Moving Average) parameter: we observe the ACF. 'q' is generally the lag at which the ACF cuts the upper confidence band for the first time.
- To determine 'd' (Integrated), we must observe the number of differencing required to make the series stationary.
- For the seasonal parameters 'P' and 'Q', we must examine the seasonal peaks in the ACF and PACF.
- To determine 'D', we must observe the number of seasonal differencing required to make the series stationary.

These methods are heuristic, that is, they provide an initial estimate of the parameters. The model must then be adjusted and its quality tested with statistical methods to confirm or readjust these parameters.

In our specific study, we focus on the weather station located in Orly. We generate its ACF and PACF plots twice. The first time, the analysis is conducted directly on the raw time series, without any modification. The second time, we perform the same analysis on a differentiated version of the time series. This differentiation is conducted to remove any underlying seasonality and the trend present in the time series, thereby ensuring the stationarity required for ARIMA modeling.

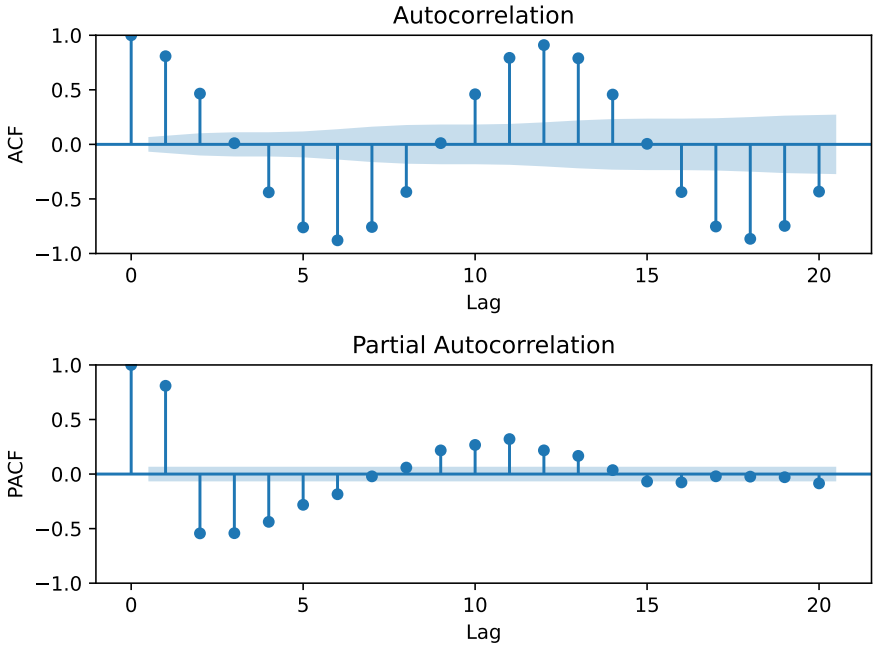


Figure 13: Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots before differencing

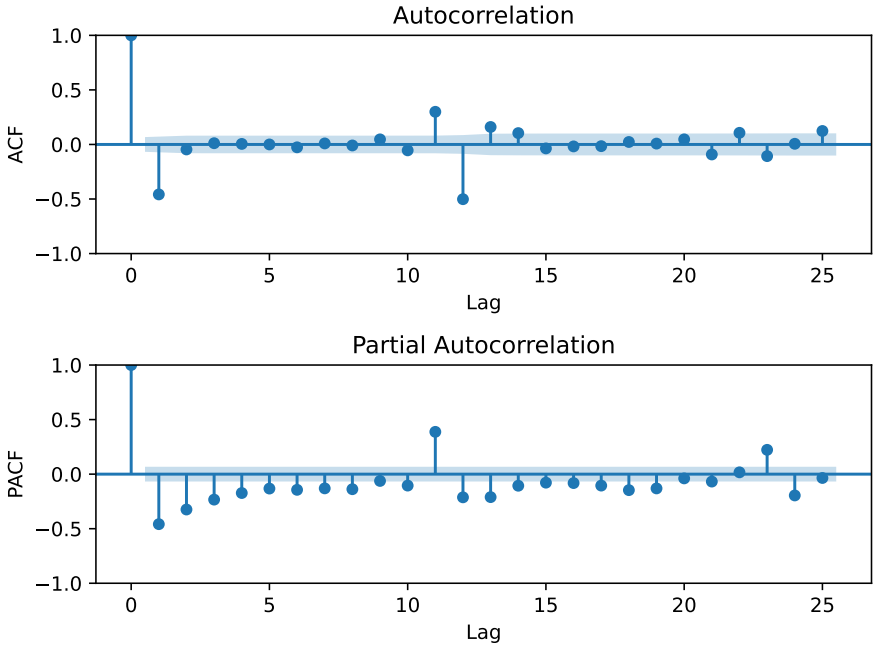


Figure 14: Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots after differencing

- Interpretation of the Plots
  The parameter d in an ARIMA model is the number of nonseasonal differences needed for the time series to become stationary. Observing the ACF and PACF plots before differencing, we can see that the autocorrelations are periodic. This suggests that the original series is non-stationary.

  After differencing (i.e., after transforming the series to a stationary one), the decay of the autocorrelations in the ACF and PACF plots is faster. This is the indication that the differencing was successful and the value of d in the ARIMA model would correspond to the order of differencing applied.

From our analysis, we can attempt to deduce the maximum order of (p, d, q) and (P, D, Q) which could provide us with

- The maximum order for 'p' in the autoregressive component of the model is expected to be 4.
- We can deduce from our analysis that the parameter 'd' for differentiation is likely to be 1.
- Furthermore, the maximum order for 'q' in the moving average part is projected to be 1.
- The maximum order for 'P' in the seasonal autoregressive component of the model is likely to be 1.
- We can deduce from our analysis that the parameter 'D' for seasonal differentiation is likely to be 1.
- The maximum order for 'Q' in the seasonal moving average part is anticipated to be 2.

This initial assessment allows us to have a model to test for performance and resemblance to reality. Subsequently, we plan to test other models identified through numerical methods.

3.2.5 Model Selection and Validation

The accuracy of statistical models such as SARIMA can be evaluated using various metrics. Among them, the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Root Mean Square Error (RMSE) are prominent. These metrics can be used in conjunction to assess and compare the accuracy of SARIMA models. However, it's important to note that each of these measures has its strengths and weaknesses, and none of them should be used independently to assess model accuracy.

- **AIC (Akaike Information Criterion):** The AIC is used for model selection. It seeks a balance between model complexity, indicated by the number of estimated parameters, and model fit. The AIC promotes parsimony by penalizing the inclusion of extra parameters, albeit less strictly than the BIC. The Akaike information criterion is defined as:

$$AIC(p, q) = \log(\hat{\sigma}^2) + \frac{2(p+q)}{n} \tag{5}$$

  where $\hat{\sigma}^2$ is the variance of the residuals, p and q are the orders of the autoregressive and moving average parts of the model, and n is the number of observations.

- **BIC (Bayesian Information Criterion):** Similar to the AIC, the BIC also aims to balance model complexity with model fit. However, the BIC penalizes the addition of extra parameters more heavily. Thus, when comparing models, the one with the lowest BIC is generally preferred. The Bayesian information criterion is defined as:

$$BIC(p, q) = \log(\hat{\sigma}^2) + \frac{(p+q)\log(n)}{n} \tag{6}$$

  where the variables are the same as in the AIC. These criteria are based on the principle of penalization. We seek a compromise between low variance and a small number of parameters to estimate: this is the principle of parsimony.

- **RMSE (Root Mean Square Error):** The RMSE is a measure of model accuracy. It is calculated as the square root of the average of the squared prediction errors, giving an indication of the magnitude of the model's prediction errors. A lower RMSE signifies a more accurate model. If y_i is the observed value and $\hat{y}_i$ is the predicted value, for n observations:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

3.2.6 Analysis of SARIMA Models

We will consider four different SARIMA models and compare their Akaike Information Criterion (AIC) values. The SARIMA models are characterized by their parameters $(p, d, q)(P, D, Q)_{12}$.

SARIMA Model Parameters	AIC
$SARIMA(1,1,1)(1,1,2)_{12}$	3119,22
$SARIMA(2,1,1)(1,1,2)_{12}$	3120,74
$SARIMA(3,1,1)(1,1,2)_{12}$	3122,57
$SARIMA(4,1,1)(1,1,2)_{12}$	3124,51

Table 1: Comparison of AIC for different SARIMA models

We observe that the model with the lowest AIC value is the $SARIMA(1,1,1)(1,1,2)_{12}$ model. Therefore, this is the model we will choose for our residual analysis.

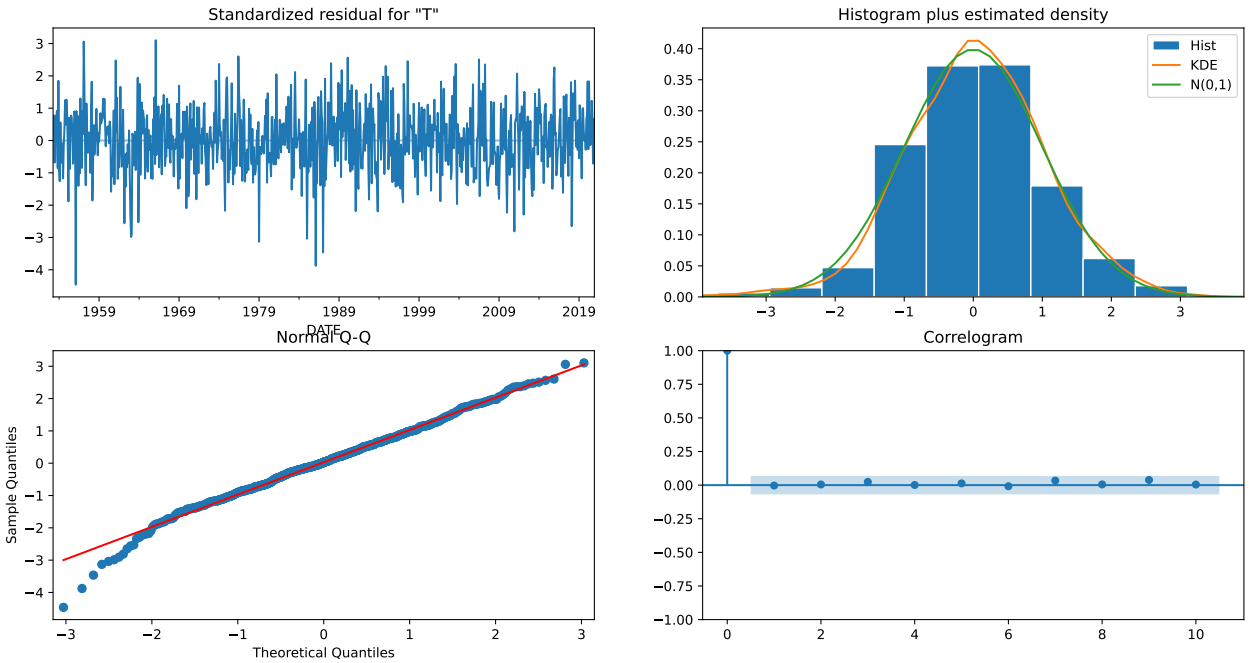


Figure 15: Residual analysis of the $SARIMA(1,1,1)(1,1,2)_{12}$ model

The residuals, crucial to the validation of this model, exhibit a normal distribution centered around zero. This observation is congruent with the theoretical expectations for residuals stemming from a SARIMA model, thereby providing a strong affirmation of our model's suitability and accuracy.

3.2.7 SARIMAX

The SARIMAX model is an extension of the SARIMA model, which allows for the inclusion of external variables in time series prediction. It comprises two types of variables: endogenous variables, which are the target variables, and exogenous variables, which are the external variables.

The ARIMA model, which the SARIMAX is an extension of, assuming that the time series is stationary. To include seasonality and exogenous variables in the model, we need to use the SARIMAX model.

In practice, to use the SARIMAX model, you first have to specify the underlying SARIMA model by determining the orders of differencing, autoregression, and moving average. Then, you need to determine the exogenous variables that will be included in the model. Finally, when you

are forecasting with the SARIMAX model, you need to specify the future values of the exogenous variables.

Having determined the optimal parameters for the SARIMA model to be $(1,1,1)(1,1,2)_{12}$based on the Akaike Information Criterion, we now fit a SARIMAX model using these parameters. The SARIMAX model extends the SARIMA model by incorporating exogenous variables, allowing for a more detailed representation of the data-generating process.

Figure 16 presents the results of fitting the SARIMAX model based on SARIMA$(1,1,1)(1,1,2)_{12}$

SARIMAX Results

Dep. Variable:	TG	No. Observations:	852
Model:	SARIMAX(1, 1, 1)x(1, 1, [1, 2], 12)	Log Likelihood	-1553.611
Date:	Sun, 28 May 2023	AIC	3119.221
Time:	14:17:30	BIC	3147.426
Sample:	01-31-1950	HQIC	3130.048
	- 12-31-2020		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1090	0.036	3.062	0.002	0.039	0.179
ma.L1	-0.9745	0.010	-95.178	0.000	-0.995	-0.954
ar.S.L12	-0.4524	0.258	-1.756	0.079	-0.957	0.053
ma.S.L12	-0.5561	85.427	-0.007	0.995	-167.990	166.877
ma.S.L24	-0.4439	37.943	-0.012	0.991	-74.812	73.924
sigma2	2.5399	216.990	0.012	0.991	-422.752	427.832

Figure 16: SARIMAX model based on SARIMA$(1,1,1)(1,1,2)_{12}$

4 Comparisons with current estimates

4.1 DRIAS

DRIAS is a platform initiated by the Ministry of Ecology, Energy, Sustainable Development, and the Sea for distributing climate scenarios for France http://www.drias-climat.fr/ It provides regional climate scenarios produced by various climate models to aid stakeholders in adapting to climate change. These data are instrumental in understanding past climate trends and predicting future changes.

In our case, we use DRIAS data as a climate model base based on the 15 meteorological stations seen earlier. This approach enables us to have a more comprehensive understanding of the climate in this specific region. We analyze the data starting from the oldest available 1951, up to 2005.

The model we employ is ALADIN, which serves as Météo-France's operational short-range numerical weather prediction model. It was developed specifically for use in regional forecasting. This model is used to generate forecasts over a particular area of interest, with a higher resolution than global models. This allows for a better representation of local climate characteristics, such as terrain or water distribution.

By using DRIAS data and the ALADIN model, we can thus carry out a detailed analysis of the climate across our chosen stations, comparing the model with the actual observations made by the meteorological stations. This approach allows us to check the model's accuracy and identify areas where improvements might be required.

4.2 Model Performance

To evaluate the performance of the DRIAS model, we compare it with historical data. We perform this comparison by carrying out correlation analyses among the 15 meteorological stations based on the DRIAS data and the historical data.

The following graph illustrates the comparison between the correlations obtained from the DRIAS model and those obtained from the historical data:

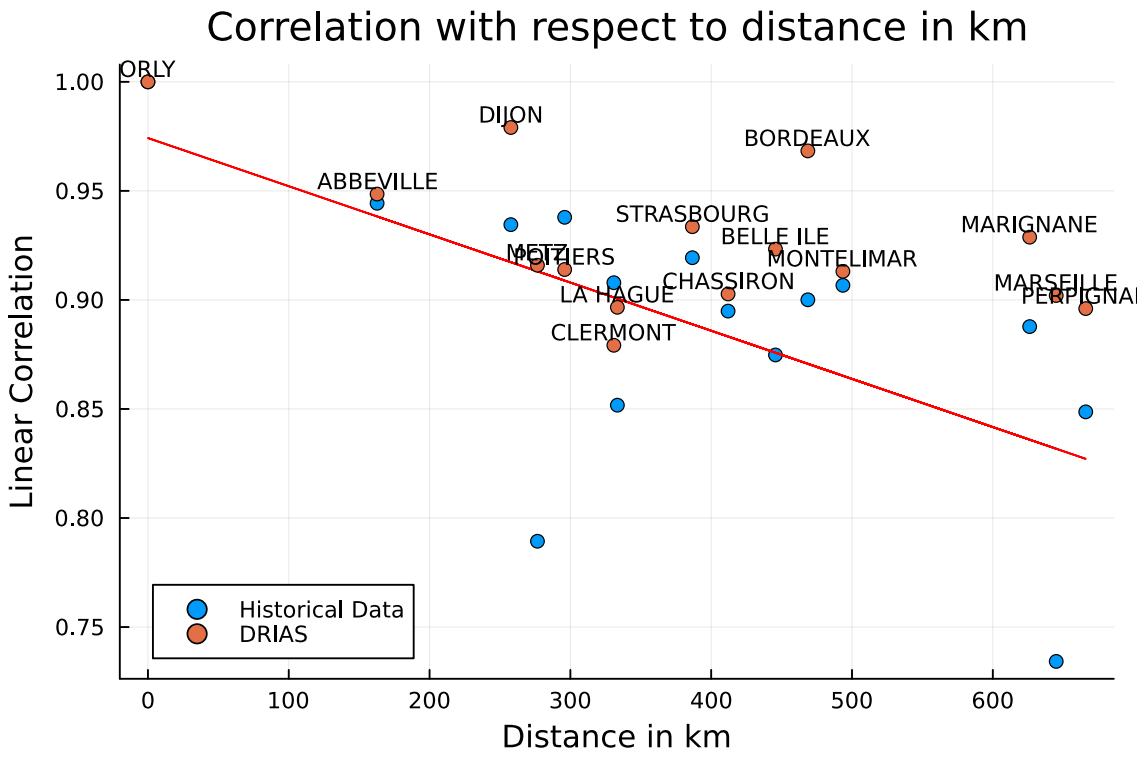


Figure 17: Comparison of correlations derived from DRIAS data and historical data.

Upon comparison, we find that the DRIAS model generally overestimates the correlations among the stations. This overestimation may be a result of the inherent assumptions and simplifications of the model, which do not fully capture the complex dynamics of the actual climate system.

In addition to the overall correlations, we also examine the tail dependence correlations between DRIAS data and historical data.

Here is a comparison of the tail dependence correlations obtained from the DRIAS data and the historical data:

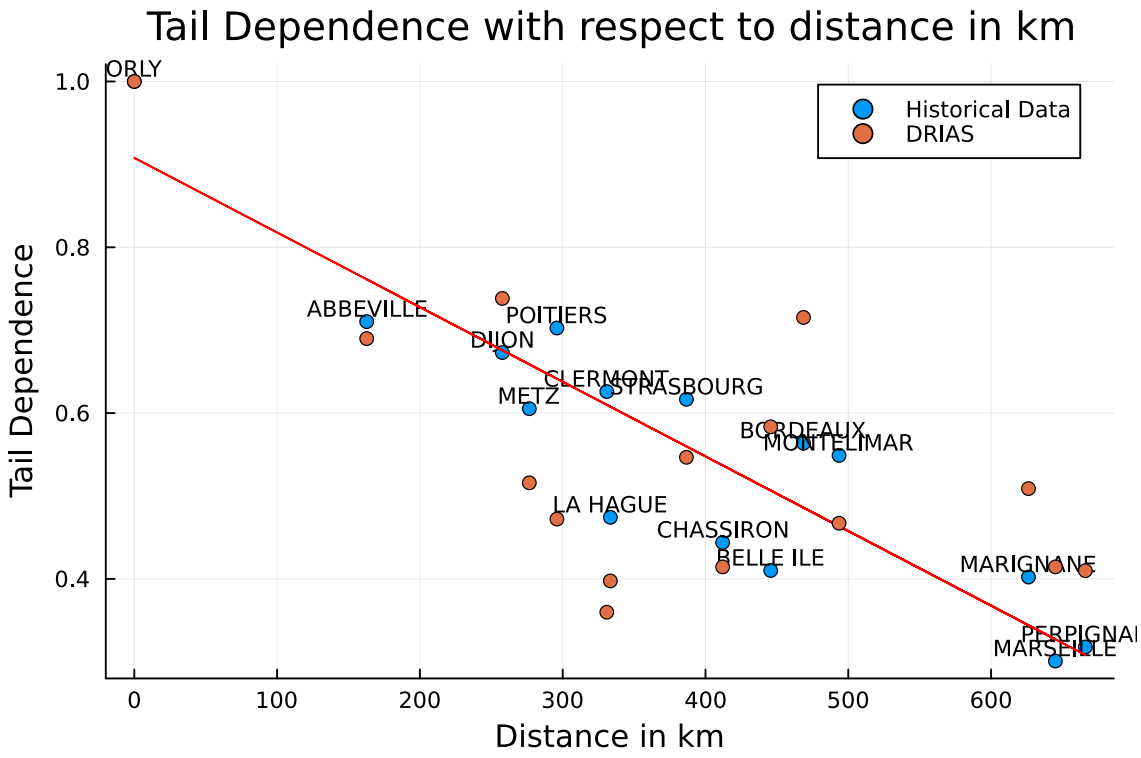


Figure 18: Comparison of tail dependence correlations derived from DRIAS data and historical data.

Interestingly, the DRIAS model slightly underestimates these tail dependence correlations compared to the historical data. This discrepancy might be due to the model's less precise representation of extreme climate phenomena.

These results emphasize the need for improving the model's performance in simulating extreme weather events, which are of paramount importance considering the projected increase in their frequency and intensity due to climate change.

4.3 RMSE

To further validate the reliability of the DRIAS model's forecasts, we evaluate its performance using the Root Mean Square Error (RMSE), a widely accepted metric for measuring predictive accuracy. In our case, the RMSE is computed to be 2.4. This value suggests that DRIAS's predictions, on average, deviate from the actual observed temperatures by approximately 2.4 degrees Celsius.

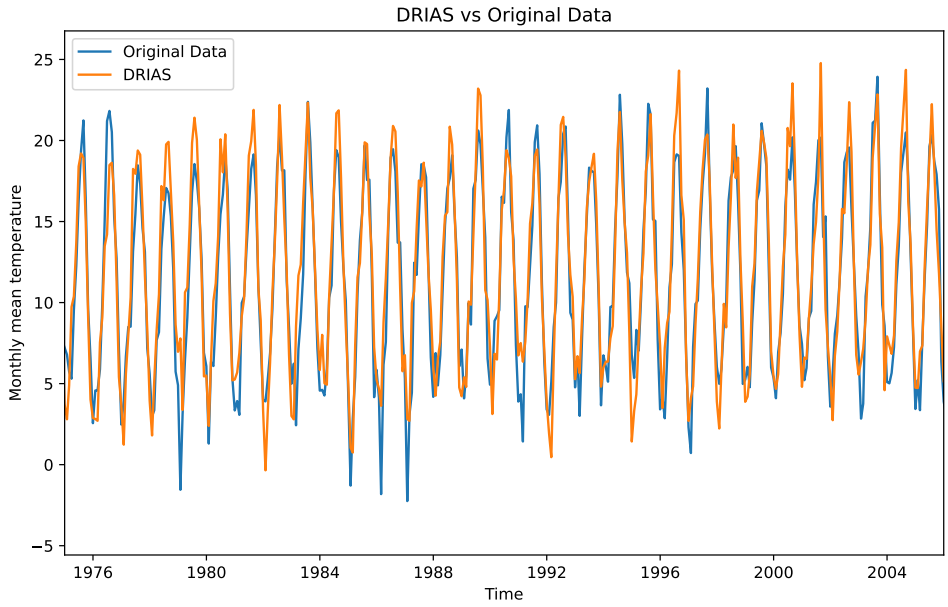


Figure 19: DRIAS vs original data

While an RMSE of 2.4 may initially appear to be a considerable margin of error, it is crucial to contextualize this figure within the broader scope of climate prediction. The task of accurately modeling and predicting temperatures is inherently challenging due to the myriad of factors contributing to climate variability.

In parallel, we also evaluate our SARIMA$(1,1,1)(1,1,2)_{12}$ model for its performance on the same historical data. The SARIMA model, in this case, yields a lower RMSE of 1.7, suggesting that it provides more accurate predictions compared to the DRIAS model when assessing the historical data.

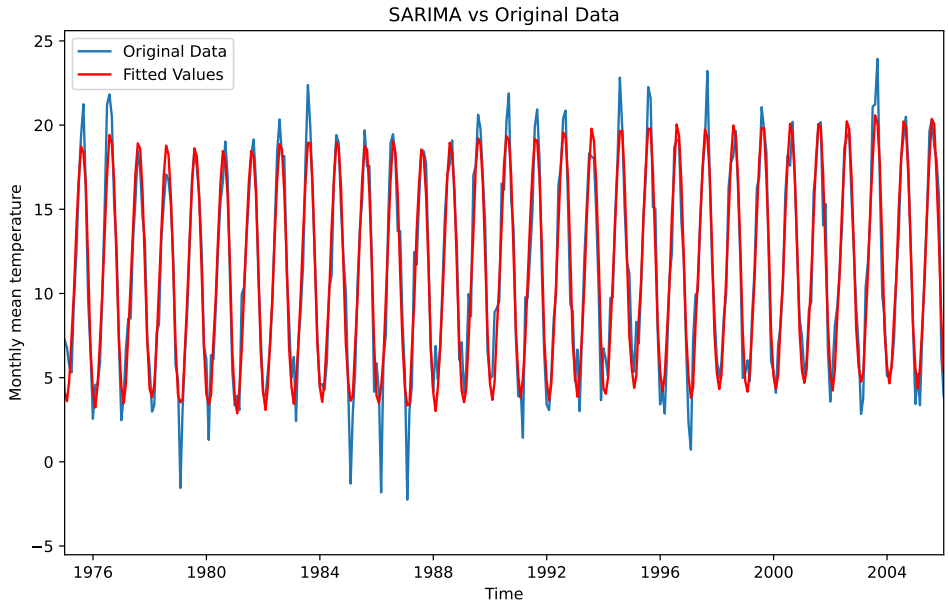


Figure 20: SARIMA vs original data

In summary, we observe that the DRIAS model, while generally less precise, excels in forecasting extreme temperature values. This ability makes it particularly valuable for predicting potential

climate anomalies and severe weather events.
On the other hand, the SARIMA model adheres more closely to historical values in its forecasts, thereby providing a more accurate picture of regular temperature patterns. However, it encounters difficulties when it comes to predicting extreme temperature events.

5 Conclusion

In conclusion, this project of analyzing meteorological data has provided critical insights into climate behavior and long-term trends. The use of time series analysis techniques, combined with the DRIAS climate projections, has created a valuable framework for understanding climate patterns and potential future trajectories.

The thorough exploration of the data, including detection of seasonal patterns, long-term trends, and irregular variations, has contributed significantly to our understanding of climate variations. Our models, particularly SARIMA and SARIMAX, have demonstrated their ability to capture these structures and generate accurate models.

Comparing our model with the DRIAS projections has been a beneficial exercise. While the DRIAS model excelled at predicting extreme values, it generally lacked precision. In contrast, the SARIMA model closely tracked historical data but struggled with extreme values. These findings underscore the importance of employing multiple models in climate studies to harness their unique strengths and counterbalance their weaknesses.

Performance evaluation, using metrics like the Root Mean Square Error (RMSE) and information criteria (AIC, BIC) for selection, has confirmed the efficacy of our models. The insights from this project can inform the planning of climate change adaptation policies, management of climate-related risks, and evidence-based policy-making.

However, we acknowledge certain limitations. The models used, while robust, are based on assumptions that may not fully encapsulate the complexity of climate systems. Additionally, our results, dependent on available data and chosen models, may vary across different geographic regions and time periods.

Future work could aim to refine these models, or even explore their integration, to improve both general and extreme value predictions. Moreover, including more variables impacting climate, such as greenhouse gas concentrations or solar radiation data, could create a more holistic model of climate change.

6 Bibliography

https://towardsdatascience.com/time-series-forecasting-with-arima-sarima-and-sarimax

http://www.sthda.com

https://www.drias-climat.fr/

https://www.ecad.eu/dailydata/predefinedseries.php

https://www.researchgate.net/figure/Mean-differences-black-line-and-their

Time Series Analysis Course, Master's degree in SSD at Monpellier University, by Elodie Brunel Piccinini.

https://www.researchgate.net/figure/Mean-differences-black-line-and-their-10-90-confidence-intervals-grey-shading_fig4_266670666