Title of Project 1

Module CA-EES-803, Fall 2022

Project Report

Garmani Thway

November 1, 2022

Contents

Abstract. In this paper, the hourly air temperature records collected at Kleine Feldberg Taunus from 1948 to 2020 are analysed. Shapiro-Wilk test on monthly air temperature distributions for the months of January, April, July, and October in the year 1985 shows that the underlying distribution is non-normal. However, using the same test, it was found that there is no support for rejecting the normality hypothesis for the monthly air temperature distributions at noon for the same months and the same years. Daily variation statistics were calculated for the same months and year, showing strong variations in January and October. Decadally averaged daily variations were also calculated and compared.

1 Introduction

Time series of air temperature provides insight into how regional climate changes over the studied time period. The average temperature over different time scales can be calculated to obtain temperature trends and can be used for comparison. For example, compared to average air temperatures in 1850-1900, recent air temperature trends have shown a warming of $1.1°C$ (IPCC, 2021). Furthermore, projections and predictions of different scenarios can be obtained by studying the impact of global warming, which are important in policy making. A similar analysis of air temperature time series (restricted to only one region and with more limitations) is done in this project.

In this paper, hourly air temperature records collected at Taunus Observatory in Kleine Feldberg is analysed. The aim of this paper is, firstly, to provide a demonstration of hypothesis testing and other methods in statistical analysis. For example, the underlying distribution of monthly air temperature distributions are tested for normality through Shapiro-Wilk test and graphical methods. Confidence intervals are also constructed using different methods and are compared. Secondly, daily variation statistics are computed on decadal timescales to provide a means to compare the average temperatures in the region across 7 decades.

2 Methodology

2.1 Air temperature measurements and data

The hourly air temperature data used in this project is provided by the Deutscher Wetterdienst. The data files contain local air temperature (in $°C$) and relative humidity (in %) values, and their timestamps, which are given in MEZ from 01.01.1948 to 01.11.1992, and in UTC after 01.11.1992. Air temperature, along with relative humidity, is measured by a Thermohygrograph from 01.01.1948 to 09.05.2006. Starting from 01.10.1992, a PT 100 resistance temperature detector was employed for electronic air temperature measurements.

The processing of the time series begins with locating missing values and replacing them with NaNs which are flagged as -999 if one of the two variables is correctly recorded. Otherwise, if both variables have missing values, they are not necessarily recorded. An hour counter since 1901 locates the records that were skipped and NaNs are inserted to make a consecutive data set.

2.2 Shapiro-Wilk Normality Test

In order to examine normality of the measured data, Shapiro-Wilk test (which is related to normal q-q plots) was applied. A probability plot with observations and expected values from a hypothesized distribution (normal distribution) tend to be linear when the hypothesis is true.(Shapiro, Wilk, 1965). The square of the slope of the probability plot regression line is compared with the total sum of squares to test normality. The Stats module from SciPy library provides a function (`scipy.stats.shapiro`) which calculates the test statistic used in Shapiro-Wilk test and the p-value, and is called in this project to perform the Shapiro-Wilk test.

2.3 Confidence intervals for true mean

A confidence interval for true mean μ has the form $[\bar{u} - h_\alpha, \bar{u} + h_\alpha]$, where u is the sample mean, and h_α the half-width for a certain significance level α (here, α is chosen to be 0.05).

2.3.1 Confidence interval using normal theory statistics

The half-width for a certain significance level h_α can be obtained by:

$$h_\alpha = z_\alpha \frac{\Delta u}{\sqrt{N}} \tag{1}$$

where z_α is the z-score, Δu the sample standard deviation, and N the sample size. The z-score (z_α) is the percent point function (inverse CDF) defined by:

$$z \;=\; \Phi^{-1}\left(\frac{1 + \mathrm{P}(\mu - z\sigma \leq U \leq \mu + z\sigma)}{2}\right)$$

In this project, since we know the probability (from α), the percent point function availble in the Stats module of Scipy library, `scipy.stats.norm.ppf` is used to obtain the z-score, and the confidence intervals are constructed from the half-width calculated by Equation 2.

2.3.2 Confidence interval using Student's t distribution

For a sample with small size, the Student's t distribution is used to construct confidence intervals. Here, the half-width is given by:

$$h_\alpha = t_\alpha \frac{\Delta u}{\sqrt{N}} \tag{2}$$

where t_α is, similarly as above, the percent point function and is determined by using `scipy.stats.t.ppf`.

2.3.3 Bootstrapping and confidence intervals

A bootstrap realization is the random resampling of the original data set with replacement. The distribution D of the difference $D = \bar{u} - \mu$ is approximated by $D^* = \bar{u^*} - \bar{u}$, where $\bar{u^*}$ is the mean of a bootstrap realization. This is repeated multiple times and the means $\bar{u^*}$ are sorted to obtain the extreme values used to construct the confidence interval.

3 Results

3.1 Test for data selection and extraction

Firstly, it was tested whether the hourly records were extracted correctly without gaps from the data set. The function `DataYearMonth(data,year,month)` extracts the specified data (timestrings, air temperature, or relative humidity) from the specified month and year and stores it in a matrix with 24 columns (rows representing days of the month and columns representing hour of the day). This was checked for the year = 2019 and month = 3. (See the JupyterNotebook included in the Digital Supplements section).

3.2 Monthly distributions of hourly air temperature in 1985

The year 1985 was chosen as a reference year and the monthly means and standard deviations for January, April, July, and October (month = 1,3,7,10) were calculated. Frequency histograms of the monthly distributions and probability q-q plots were produced Fig (1) for a graphical test for normality. Shapiro-Wilk test is then applied for a quantitative assessment of normality, and the confidence intervals are constructed from normal theory statistics, Student's t distribution and bootstrapping as described in Section 2.3. These results are summarized in Table (1)

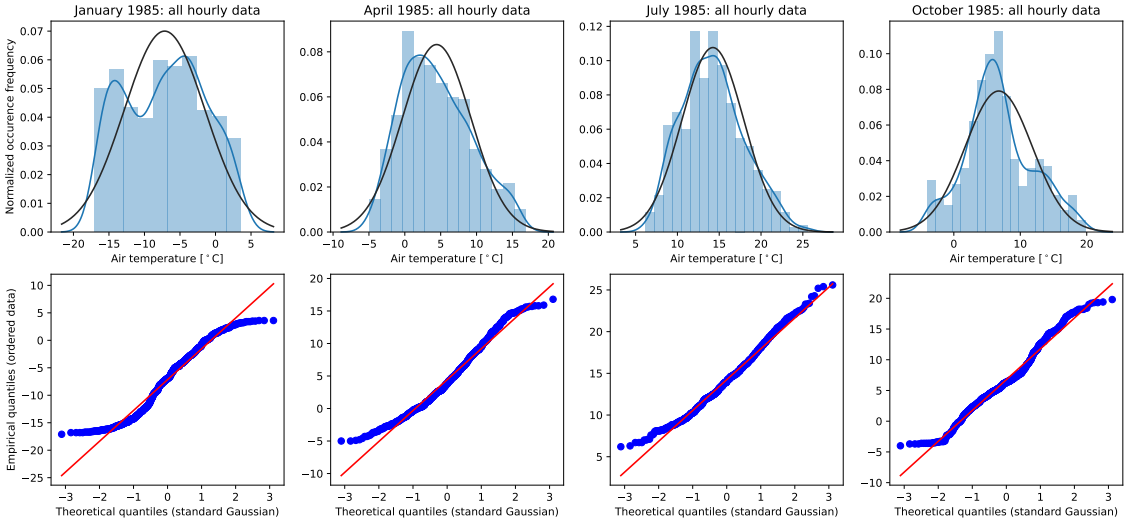

Figure 1: Monthly distributions shown as frequency histograms with a Gaussian fit (top panels) and corresponding normal q-q plots (bottom panels). Data provided by the DWD (Deutscher Wetterdienst, opendata.dwd.de)

3.3 Monthly distributions of hourly air temperature in the hour before noon in 1985

In the same year 1985, the air temperature records in the hour before noon of four months, January, April, July, and October are examined. The distributions and normal q-q plots were plotted and shown in Fig (2). Normality was tested and confidence intervals were constructed similarly as above, and the results are summarized in Table (2).

Month	January	April	July	October
Mean ($°C$)	-7.16	4.43	14.24	6.76
Standard deviation ($°C$)	± 5.70	± 4.80	± 3.71	± 5.05
p-value (Shapiro-Wilk)	1.68×10^{-13}	9.25×10^{-10}	1.75×10^{-6}	1.30×10^{-8}
Confidence interval (normal theory)	[-7.573,-6.754]	[4.084,4.785]	[13.974,14.507]	[6.395,7.121]
Confidence interval (Student's t)	[-7.574,-6.753]	[4.083,4.785]	[13.973,14.507]	[6.395,7.122]
Confidence interval (bootstrap)	[-7.595,-6.742]	[4.064,4.779]	[13.976,14.493]	[6.428,7.151]

Table 1: A summary of calculations, showing sample mean, standard deviation, p-value from Shapiro-Wilk test, and confidence intervals constructed from normal theory statistics, Student's t distribution, and bootstrapping for hourly air temperature records in the months of January, April, July, and October in the year 1985.

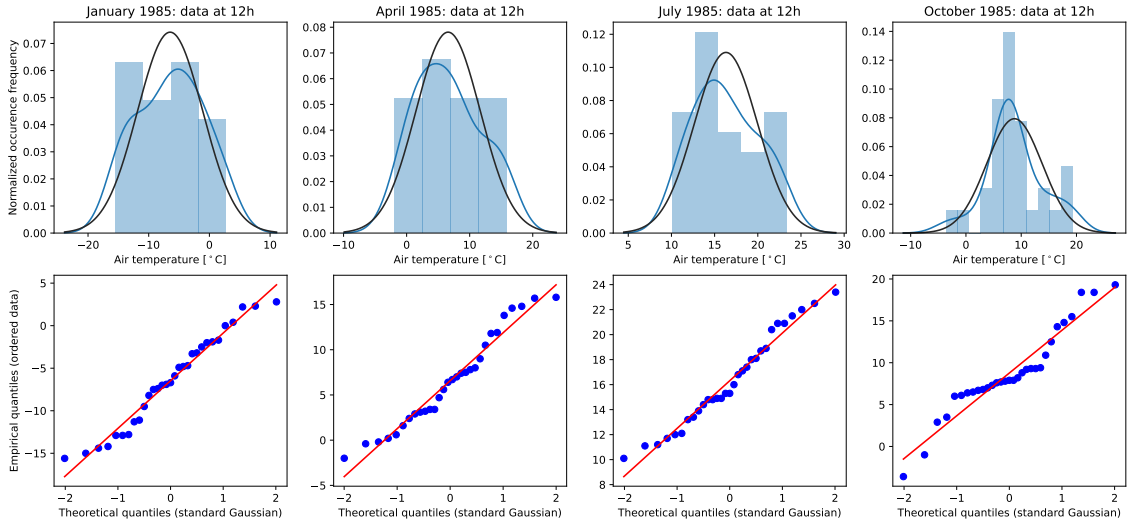

Figure 2: Monthly distributions shown as frequency histograms with a Gaussian fit (top panels) and corresponding normal q-q plots (bottom panels). Data provided by the DWD (Deutscher Wetterdienst, opendata.dwd.de)

3.4 Daily variation statistics of hourly air temperature

For the year 1985, and for the months of January, April, July and October, daily variation statistics was evaluated. The daily means and standard deviations of air temperature at each hour were calculated and plotted as shown in Fig (3)

Month	January	April	July	October
Mean ($°C$)	-6.47	6.57	16.31	8.77
Standard deviation ($°C$)	± 5.47	± 5.20	± 3.72	± 5.11
p-value (Shapiro-Wilk)	2.60×10^{-1}	1.82×10^{-1}	3.39×10^{-1}	5.11×10^{-2}
Confidence interval (normal theory)	[-8.364,-4.578]	[4.745,8.402]	[15.025,17.601]	[6.998,10.537]
Confidence interval (Student's t)	[-8.444,-4.498]	[4.666,8.481]	[14.971,17.655]	[6.924,10.611]
Confidence interval (bootstrap)	[-8.248,-4.816]	[4.747,8.330]	[15.068,17.503]	[6.948,10.506]

Table 2: A summary of calculations, showing sample mean, standard deviation, p-value from Shapiro-Wilk test, and confidence intervals constructed from normal theory statistics, Student's t distribution, and bootstrapping for air temperature records in the hour before noon in the months of January, April, July, and October in the year 1985.

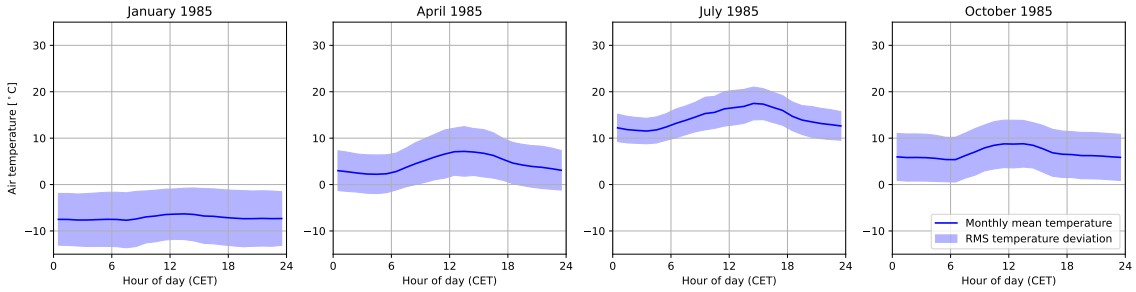

Figure 3: Daily variation statistics of hourly air temperature in the year 1985. Data provided by the DWD (Deutscher Wetterdienst, opendata.dwd.de)

3.5 Daily variation of hourly air temperature on a decadal timescale

For all months and for each decade starting from 1950-1959 to 2010-2019, the daily variations of hourly air temperature were compared. The daily variation statistics for each month and each year in the decade were calculated and averaged over the decade. Decadally averaged daily variations were plotted as shown in Fig (4)

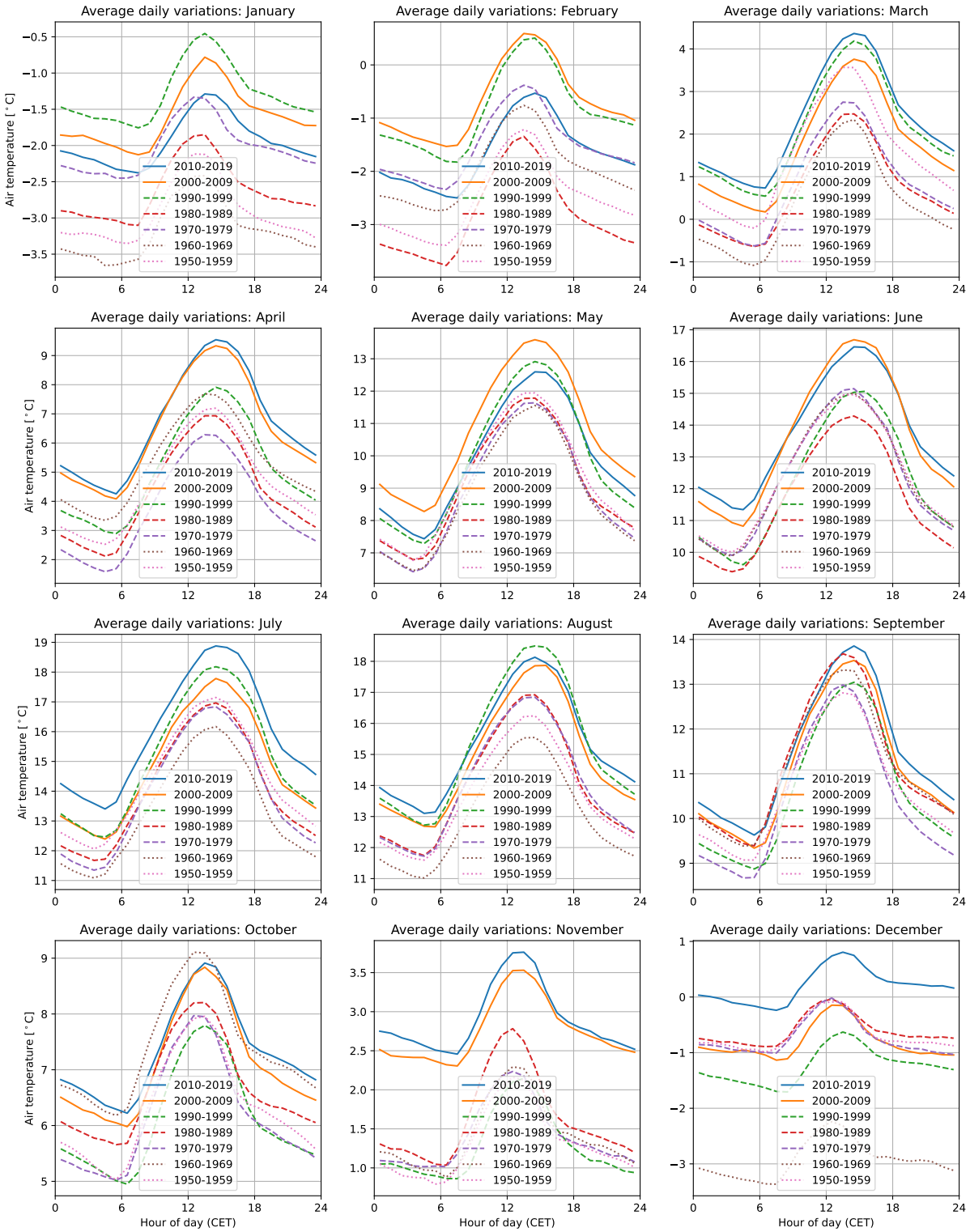

Figure 4: Daily variations of air temperature averaged over each decade for all months. Data provided by the DWD (Deutscher Wetterdienst, opendata.dwd.de)

4 Discussion

4.1 Monthly distributions of hourly air temperature in 1985

It can be seen that the Shapiro-Wilk test performed on the monthly distributions results in a probability value (see Table 1) that is much less than the significance level ($\alpha = 0.05$) for all four months (January, April, July, and October). This provides evidence for rejection of the null hypothesis (the underlying distribution is normal). In addition, from Fig 1, it can be seen that the q-q plots deviates from the regression line at the ends, suggesting skewness in the original data set. It can also be noted that the deviation is the strongest in the first plot (for the month of January), whose distribution seems bimodal. Both graphical and quantitative assessment for normality suggest that the underlying distribution is non-normal, assuming there is no fault in the data collection method (since there is no evidence against this).

In contrast, the results of the Shapiro-Wilk test performed on the monthly distributions of air temperature data at noon (for the same months and same year) show a probability value larger than the significance level ($\alpha = 0.05$) (Table 2). Therefore, there is no support for rejecting the normality hypothesis. Slight deviation from the regression line at the ends in the q-q plots suggests skewness in the data.

4.2 Daily variation statistics in the year 1985 and on decadal timescale

Figure 3 shows the daily air temperature variations in 1985 and from the figure, it can be seen that the months of January and October show strong daily variations while July shows the smallest daily variations. It can also be noted that the mean temperature stays somewhat constant throughout the day in January.

From the decadal averages, it can be seen that the last three decades show higher average temperatures than the previous decades, indicating warming of the climate. Compared to the 1990-1999 decade, winter months with the exception of December in the most recent decade show lower average temperatures whereas spring, summer and autumn months show higher average temperatures. It can also be noticed that average hourly temperatures throughout the day follow a similar trend over the decades in all months. However, with the current analysis, information on the temperature trends and whether they are statistically significant or not cannot be obtained.

5 Conclusions

In this paper, assessment for normality using Shapiro-Wilk test and graphical methods is demonstrated with two distinct cases, one where the normality hypothesis is rejected and one where there is no support for the normality hypothesis. In addition, daily variation

statistics was done for a single year and on a decadal time scale, allowing the comparison of decadally averaged daily air temperatures for different months. The results of this paper provides a glimpse of the evidence for warming climate but does not provide any quantitative measure of the change of air temperature over the studied time period as annual or seasonal trends were not examined.

This project can be further extended to include analysis of annual or seasonal trends which can be compared to general trends over the same time period in Western and Central Europe or on a global scale. The analysis could also include multiple stations within Germany to provide a description of temperature trends of the country. Furthermore, the relationship between air temperature variability and atmospheric circulation could also be studied through a correlation analysis with circulation indices.(See, for example, research by Feidas et al., 2004).

Digital supplements

The relevant data files, graphics, and Jupyter Notebook can be found at the course Teamwork page and are listed below.

- med-prj1-stats-jnb.ipynb
- dist_prob_02601_1985_all.pdf
- dist_prob_02601_1985_12h.pdf
- daily_variations_02601_1985.pdf
- daily_variations_decadal_02601.pdf
- stundenwerte_TU_02601_19480101_20201231_hist.zip

References

IPCC (2021) Climate Change 2021: The Physical Science Basis, Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change

Shapiro, S.S. Wilk, M.B. 1965 An analysis of variance test for normality

Feidas, H. Makrogiannis, T. Bora-Senta, E. 2004 Trend analysis of air temperature time series in Greece and their relationship with circulation using surface and satellite data: 1955–2001