Modeling of Earth System Data

Jacobs University Bremen, Course CA-EES-803, Fall 2022

— Project description —

1 Statistical Concepts in Earth Data Modeling

The topics addressed in the first course chapter *Statistical Concepts in Earth Data Modeling* include density estimation, sample statistics, standard errors, confidence intervals, normality tests, statistical hypothesis testing, and the bootstrap approach to error estimation. In a series of Jupyter notebooks, the methods are presented and demonstrated using geophysical time series and random variates drawn from model distributions.

In this project, temperature records from the *CDC (Climate Data Center)* of the *DWD (Deutscher Wetterdienst)* are used to demonstrate the statistical techniques presented in the first course chapter. Monthly distributions of hourly air temperature time series are studied, average daily variations are constructed, and sample statistics are computed using methods for normally distributed data as well as bootstrap techniques.

1.1 Preparation

Consult the presentation file of the course chapter *Statistical Concepts in Earth Data Modeling* to review the underlying theoretical concepts. Go through the computational exercises to recall the numerical methodology and the syntax of relevant Python functions.

Consult the CDC website and familiarize with its *terms of use*:

- `https://www.dwd.de/EN/climate_environment/cdc/cdc_node_en.html`,
- `https://www.dwd.de/EN/service/copyright/copyright_node.html`.
- `https://opendata.dwd.de/`.

The data for this project are hourly temperature measurements available from the directory

- `https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/hourly/air_temperature/historical/`.

Unzip the archive of the station assigned to you, read the documentation, and inspect the data.

On the course teamwork space you find the Jupyter notebook `med-prj1-stats-jnb.ipynb` with further instructions, function templates, and scripting examples.

The functions

- `TimeStr2Int()`,
- `TimeInt2Str()`,
- `HourSince1901()`,
- `TimeStrFromHS1901()`,

are provided in the Jupyter notebook `med-prj1-stats-jnb.ipynb`, to facilitate the processing of the DWD time series. Familiarize with the functions and check if they work as intended.

1.2 Data selection and extraction

As outlined in the Jupyter notebook `med-prj1-stats-jnb.ipynb`, hourly time series for a numerical `month` (range 1–12) in a `year` (four-digit integer) are extracted from a one-dimensional array `data` by calling the function

- `DataYearMonth(data,year,month)`.

Familiarize with the functions and check if it works as intended.

1.3 Monthly distributions of hourly air temperature measurements

As a reference year select 1985, and collect all hourly air temperature measurements at your observing station for the months January, April, July, and October in a data vector (one-dimensional data array, hint: apply `np.ravel()` to the output of `DataYearMonth()`).

- Construct histograms and kernel density estimators (KDEs), and plot them together with a fitted normal probability density function (PDF), most conveniently by means of the function `distplot()` from the Python module `seaborn`.
- Produce normal probability plots using the function `probplot()` from the Python module `scipy.stats`.
- As a quantitative normality check perform the Shapiro-Wilk test. Compare the resulting probability value with standard significance levels ($\alpha = 0.1, 0.05, 0.001$).
- Construct confidence intervals for the (true) mean air temperature in the respective month using (1) normal theory statistics, (2) Student's t distribution, and (3) the bootstrap approach to error estimation.

See Figure 1 for the empirical distributions and probability plots based on the Bremen data (station ID 691).

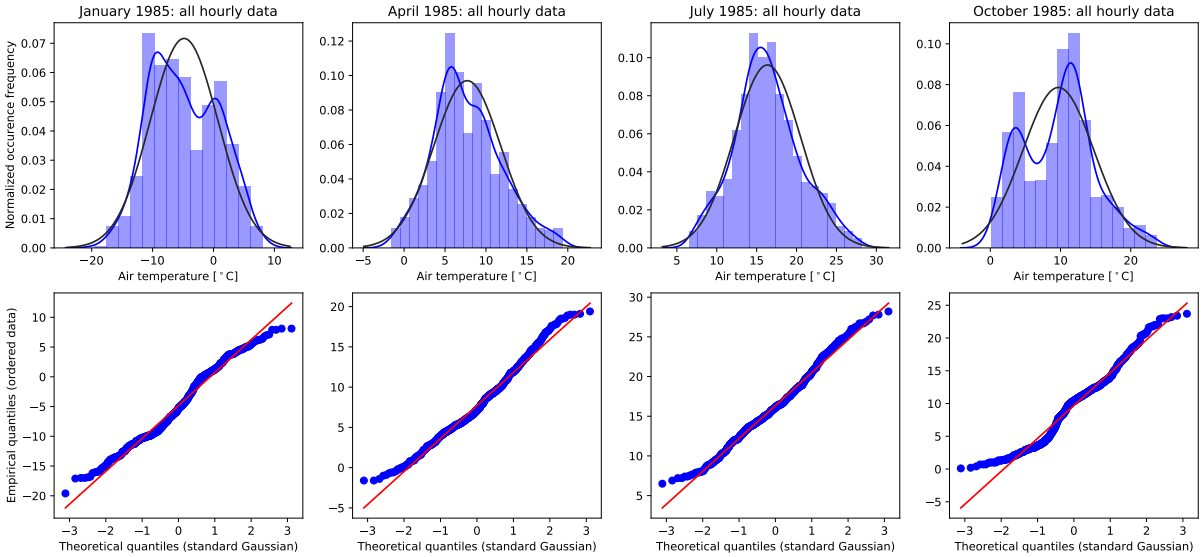

Figure 1: Empirical distributions (top panels) and probability plots (bottom panels) for all hourly air temperature measurements in Bremen. Data provided by the DWD (Deutscher Wetterdienst, opendata.dwd.de).

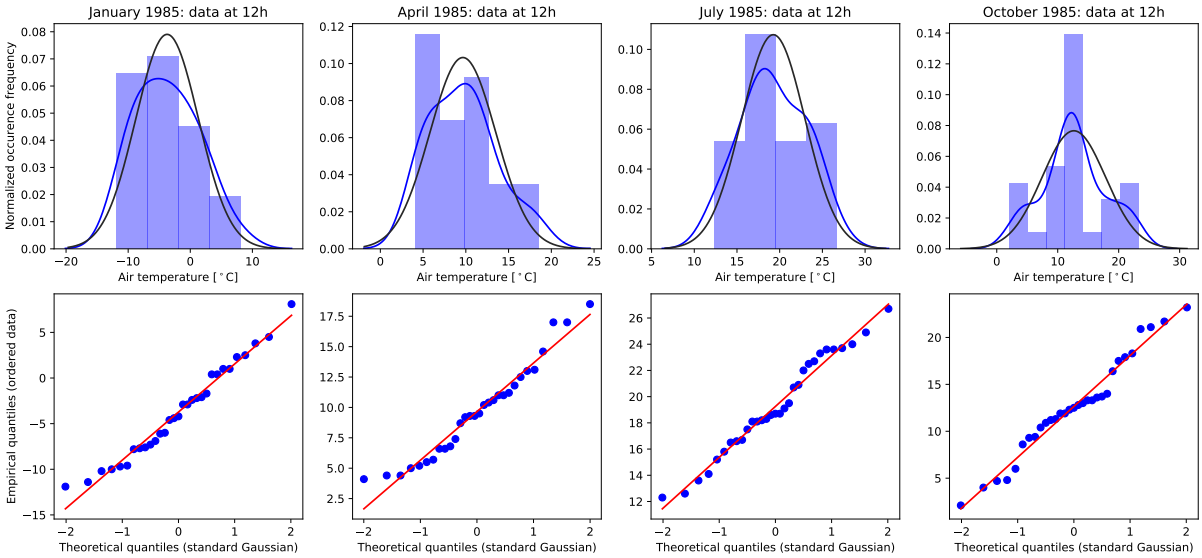

Figure 2: Empirical distributions (top panels) and probability plots (bottom panels) for hourly air temperature measurements in Bremen at noon (CET). Data provided by the DWD (Deutscher Wetterdienst, opendata.dwd.de).

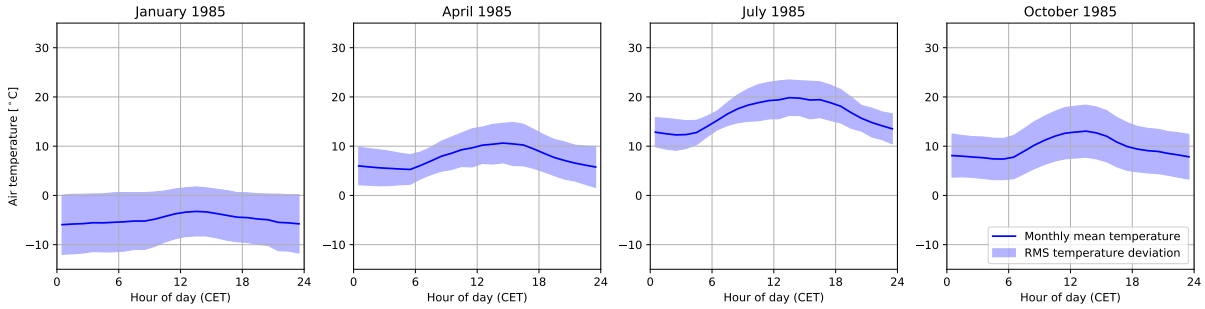

Figure 3: Monthly averaged daily variations of hourly air temperature measurements in Bremen. Data provided by the DWD (Deutscher Wetterdienst, opendata.dwd.de).

1.4 Monthly distributions of air temperatures at noon

Instead of considering all hourly air temperature measurements from the months January, April, July, and October in the reference year 1985, restrict your analysis to measurements taken in the hour before noon, and attributed to the time 12:00 CET. Repeat the analysis described above, namely,

- plot histograms, KDEs, and fitted normal PDFs, e.g., using `seaborn.distplot()`;
- produce normal probability plots by means of `scipy.stats.probplot()`;
- compute the probability value of the Shapiro-Wilk test;
- construct confidence intervals for the (true) mean air temperature at noon in the respective month using (1) normal theory statistics, (2) Student's t distribution, and (3) the bootstrap approach to error estimation.

Figure 2 shows the diagram for Bremen (station ID 691).

1.5 Daily variation statistics of hourly air temperature data

The shape of the data matrix returned by the function `DataYearMonth()` facilitates the computation of daily variation statistics by applying operations along the first dimensions (`axis=0`). For the months January, April, July, and October in the reference year 1985, compute the mean $\overline{T}$ and the standard deviation ΔT separately for each daily hour (time t). Plot the daily variation of the mean $\overline{T}(t)$ on top of a filled region between $\overline{T}(t) - \Delta T(t)$ and $\overline{T}(t) + \Delta T(t)$ using the function `fill_between()` from the Python module `matplotlib.pyplot`.

The results for Bremen (station ID 691) are shown in Figure 3.

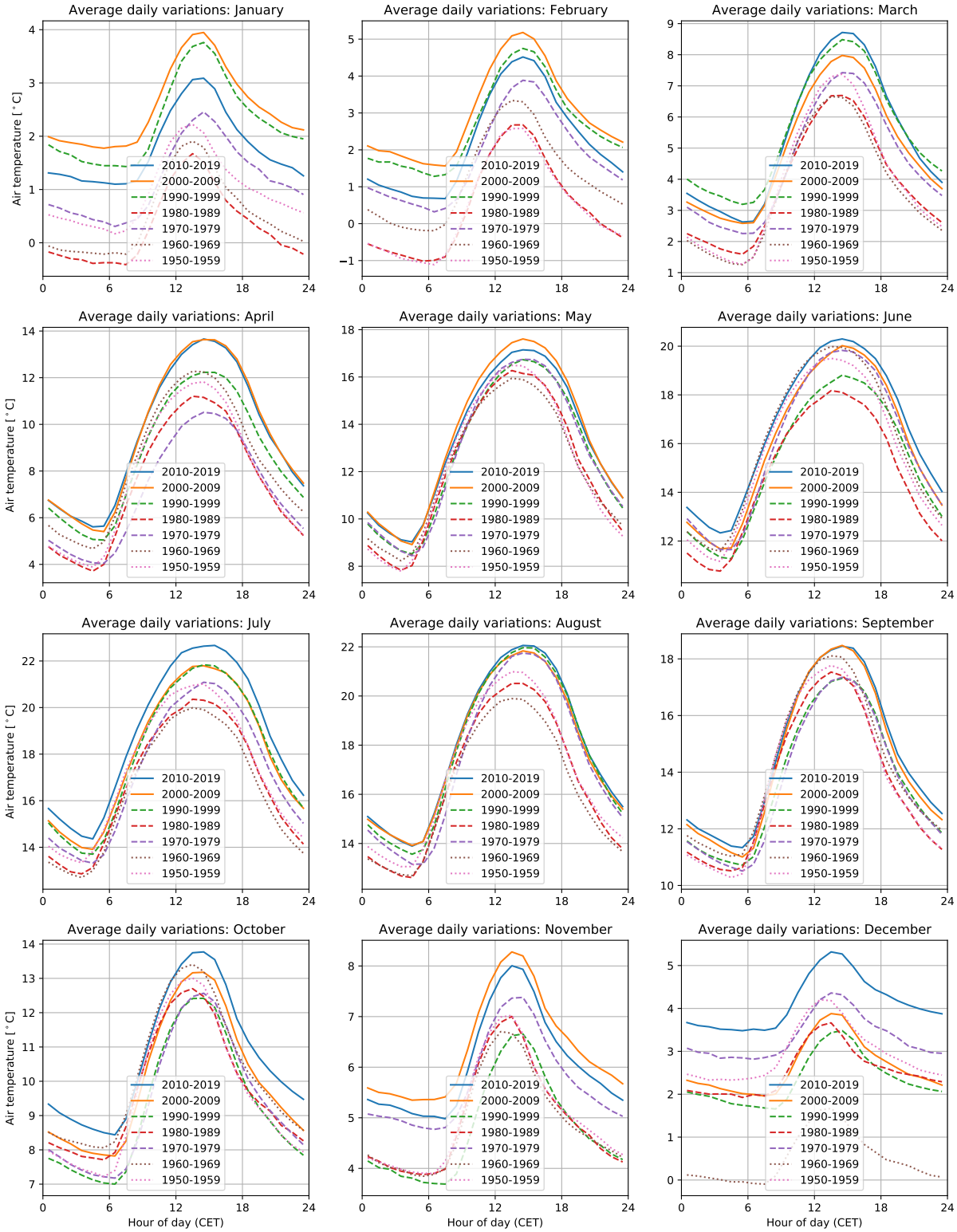

Figure 4: Decadal averages of daily air temperature variations in Bremen for all months. Data provided by the DWD (Deutscher Wetterdienst, opendata.dwd.de).

1.6 Decadal variation of monthly temperature statistics

For all months and each decade from 1950–1959 to 2010–2019, determine the average daily variations of air temperature at your observing station. For a particular month plot all seven decadally averaged curves into one panel, and combine all plots in a (4×3) diagram. Observe trends and patterns. Think about possible hypotheses and how they could be statistically tested.

Figure 4 shows the diagram for Bremen.

1.7 Project report and digital supplements

Scientific reports typically show the following structural elements.

- *Introduction*: background, motivation, objectives.
- *Methodology*: briefly describe key theoretical concepts and computational tools.
- *Results*: comments on implementation and procedure, description of graphics.
- *Discussion*: discuss the results in the light of the objectives.
- *Conclusions*: suggestions for further work, outlook.
- *References*

Digital supplements (Jupyter notebook, graphics, data files) are to be uploaded to your personal folder on the course teamwork space. Organize the teamwork wiki and add comments.