DEADLINE: 4 June 2025, 23:59 (11:59 p.m.).

1 Description of the problem

This project explores both short and long term climatological patterns using data from the Global Historical Climatology Network (GHCN) database. The GHCN database represents one of the most comprehensive and reliable sources of historical climate and weather data, making it particularly valuable for studying temporal patterns.

A fundamental distinction in atmospheric science lies between weather and climate analysis. Weather refers to short-term variations in temperature, precipitation, and other meteorological variables. These short-term fluctuations can be highly variable and are typically studied for day-to-day (or week-to-week) forecasting. Climate, in contrast, represents the long-term patterns and averages of these weather conditions, usually analyzed over periods of decades. We display in Figure 1 and Figure 2 two examples of climate and weather data, respectively.

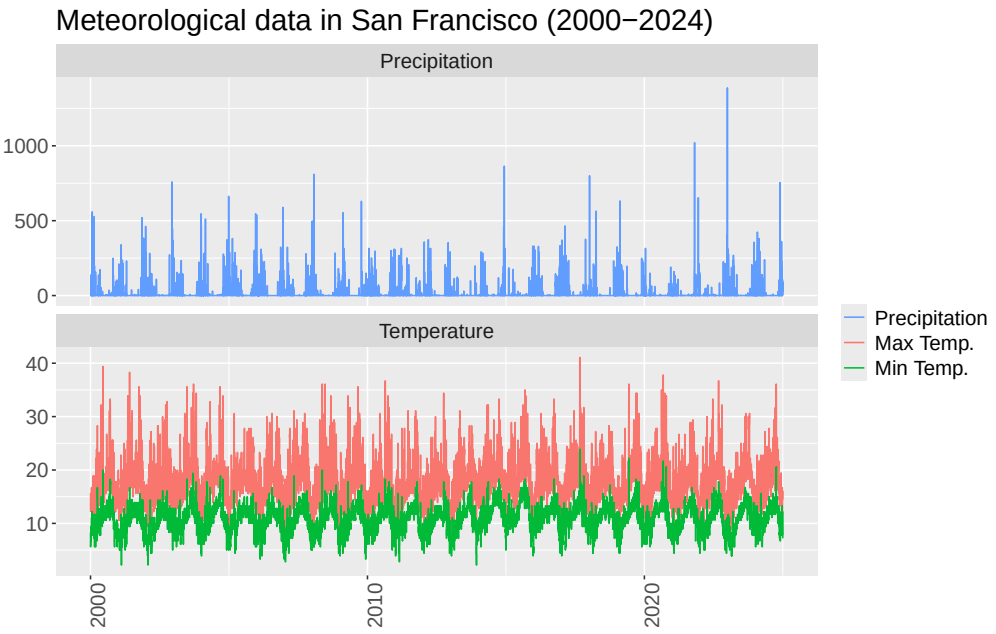

Figure 1: Daily precipitations (mm/10), minimum and maximum temperature (C°) in San Francisco (2000-2024).

Through this project you will investigate both phenomena exploiting different statistical methods you will learn during the course. For instance, some of the questions you will be asked to answer are: (a) Can we identify any long-term trends in temperature data? (b) Can we exploit recent data to forecast temperature of the following temporal steps?

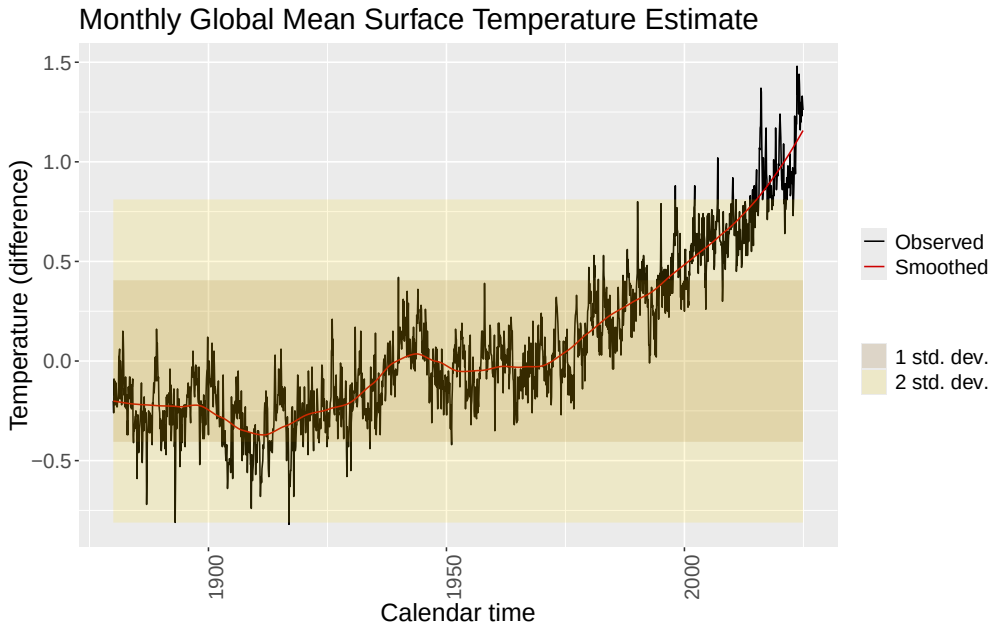

Figure 2: Global mean estimate for surface temperature from 1880 to 2024.

While for weather purposes you can directly work with raw data from the GHCN archive, in order to extract some long term information it is better to analyze aggregated data. In fact, some pre–processing steps stress macro–behaviors by smoothing out some short-term variations. In this direction, one of the most reliable and influential sources is GISTEMP [2] (GISS Surface Temperature Analysis) a product of the NASA Goddard Institute for Space Studies which merges some data gathered from GHCN meteorological stations with some ocean related data from ERSST stations.

2 Data Description

First, we will focus on the temperature data coming from GISTEMP [2]. The series provided represents changes in the global surface temperature over time. It is derived from a global network of land and ship weather stations, which track daily surface temperatures. Measurements are then adjusted for urban heat effects and other biases, aggregated monthly and averaged across stations. Lastly, since the quantity of interest is *the variation* of the global surface temperature, the final measurements are adjusted by subtracting the average global surface temperature over the period 1951-1980, which serves as a reference value.

The data set provides a reliable long-term record of temperature anomalies, offering valuable information on climate trends and variability.

The data is provided in a csv file named `gistemp.txt`. Each row refers to a calendar year (starting from 1880, up to 2024) and contains the following variables.

- 1st column: calendar year;
- 2nd to 13th: monthly temperature difference with respect to reference period;
- 14th to 15th: annual average of temperature difference taken as Jan-to-Dec (`J-D`) or Dec-to-Nov (`D-N`);
- 16th to 19th: seasonal average for winter (`DJF`), spring (`MAM`), summer (`JJA`) and autumn (`SON`).

Next, we turn our attention to the GHCN (Global Historical Climatology Network) dataset [1], which provides high-resolution daily climate observations from thousands of land-based weather stations around the world. The data is widely used for studies on local and regional climate patterns, extreme weather events, and short-term trends.

The dataset includes daily measurements of key meteorological variables such as temperature and precipitation. Each observation records minimum, maximum, and average daily temperatures, as well as the amount of precipitation, making it particularly valuable for fine-grained temporal analyses.

The data is provided in a `.txt` file named `ghcn.txt`. Each row in the dataset corresponds to a single daily observation from a specific weather station. The columns are as follows:

- 1st column: Station ID (a unique identifier for each weather station);
- 2nd column: Station name;
- 3rd to 5th columns: Geographic coordinates and elevation of the station (latitude, longitude, elevation in meters);
- 6th column: Date of observation (formatted as YYYY-MM-DD).

- 7th column: Minimum temperature of the day (TMIN), recorded in tenths of degrees Celsius;
- 8th column: Maximum temperature of the day (TMAX), recorded in tenths of degrees Celsius;
- 9th column: Average temperature of the day (TAVG), recorded in tenths of degrees Celsius;
- 10th column: Daily total precipitation (PRCP), recorded in tenths of millimeters.

Task #1. Data acquisition and exploration

Extract the data from the GISTEMP and GHCN datasets. Specifically, for daily data we will focus solely maximum and minimum temperature measurements from the San Francisco Downtown station. Describe suitably the two time series, with appropriate plots and comments. Perform a time series decomposition using appropriate tools and highlight relevant features (if present) for each component.

Task #2. GISTEMP: Do the data document global warming?

Consider the seasonality adjusted time-series from the GISTEMP data.

Step 1. Hidden Markov Models

Explore the use of Hidden Markov Models to identify the presence of long term temperature trends and/or change points inside the data. Comment on the results, highlighting advantages and limitations of the approach.

Step 2. Dynamic Linear Models

Explore the use of Dynamic Linear Models such as random walk plus noise or locally linear trend to address the presence of acceleration/deceleration in global warming. You can opt to use more structured models (if you choose a model with seasonality, you can use the full time series). Comment on the results, and compare them with the ones obtained using HMMs.

Task #3. GHCN: Weather preditcion.

The dataset includes data along space (several stations) and time. Typical aspects of interest are spatial interpolations (at a fixed time) or, in our case, DLM models for spatio-temporal data, considering temperature over time for multiple stations (say 2, for simplicity). You are welcome to explore this direction. However, to keep your workload lighter, here is our suggestion for your possible analysis, focusing only on one station, namely San Francisco Downtown.

Step 1. Data and question.

Extract the seasonality adjusted minimum and maximum daily temperatures from the San Francisco Downtown station. We want to obtain short term predictions for the minimum and maximum temperature and to investigate a potential common latent process that describes the weather in San Francisco.

N.B.: You need to divide columns 7, 8, and 9 (see Section 2) to obtain temperatures (TMIN, TMAX, TAVG) in Celsius degrees.

Step 2. Short term temperature prediction

Explore the use of Dynamic Linear Models to obtain short term temperature predictions. Evaluate the models based on their interpretability and the quality of their predictions.

Namely, consider the bivariate time series of minimum and maximum temperature, $Y_t = [T_{min,t}, T_{max,t}]^\top$ and explore the following models.

(a) Independent random walk plus noise models

$$\begin{aligned} \mathbf{Y}_t &= \boldsymbol{\theta}_t + \mathbf{v}_t, & \mathbf{v}_t &\sim \mathcal{N}(\mathbf{0}, \mathbf{V}) \\ \boldsymbol{\theta}_t &= \boldsymbol{\theta}_{t-1} + \mathbf{w}_t, & \mathbf{w}_t &\sim \mathcal{N}(\mathbf{0}, \mathbf{W}) \end{aligned}$$

where $\boldsymbol{\theta}_t = [\theta_{1,t}, \theta_{2,t}]^\top$ is the latent state at time t and $\mathbf{V}$ and $\mathbf{W}$ are variance-covariance diagonal matrices

$$\mathbf{W} = \begin{bmatrix} \sigma^2_{w,1} & 0 \\ 0 & \sigma^2_{w,2} \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} \sigma^2_{v,1} & 0 \\ 0 & \sigma^2_{v,2} \end{bmatrix}.$$

(b) "Seemingly unrelated" random walk plus noise models ($\mathbf{V}$ diagonal and $\mathbf{W}$ full)

$$\begin{aligned} \mathbf{Y}_t &= \boldsymbol{\theta}_t + \mathbf{v}_t, & \mathbf{v}_t &\sim \mathcal{N}(\mathbf{0}, \mathbf{V}) \\ \boldsymbol{\theta}_t &= \boldsymbol{\theta}_{t-1} + \mathbf{w}_t, & \mathbf{w}_t &\sim \mathcal{N}(\mathbf{0}, \mathbf{W}) \end{aligned}$$

where:

$$\mathbf{W} = \begin{bmatrix} \sigma^2_{w,11} & \sigma^2_{w,12} \\ \sigma^2_{w,21} & \sigma^2_{w,22} \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} \sigma^2_{v,11} & 0 \\ 0 & \sigma^2_{v,22} \end{bmatrix}$$

(c) Random walks plus noise driven by a "latent factor" (a common state process)

$$\mathbf{Y}_t = \mathbf{F}\boldsymbol{\theta}_t + \mathbf{v}_t, \quad \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$$
$$\boldsymbol{\theta_t} = \boldsymbol{\theta}_{t-1} + (0, w_t)^\top, \quad w_t \sim \mathcal{N}(0, \sigma_w^2)$$

where

$$\boldsymbol{\theta}_t = \begin{bmatrix} 1 \\ \xi_t \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} \alpha_1 & \beta \\ \alpha_2 & \frac{1}{\beta} \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} \sigma^2_{v,11} & 0 \\ 0 & \sigma^2_{v,22}. \end{bmatrix}$$

The parameters α_1, α_2 and β should be determined via MLE.

We recommend that you all do this analysis - which is the mandatory task for this final project. Then, if you like, you may explore other models (more stations, covariates, ...- as mentioned above) but these developments are <u>not</u> mandatory.

Guidelines

Remember that the final project is also a useful exercise of **presentation**. Below you find suggestions on how your analysis should be presented (they were posted, and are still available, on BBoard).

Submission

- Single .zip file with report and code to reproduce
- Report in pdf (if from rmarkdown: Knit to PDF. do not export HTML and then print)
- Code in .R or .rmd
- Name of zip file = group name

Length The PDF file must be ideally 8 pages long and absolutely no longer than 10 pages.

First page includes

- Group name
- Names of group components
- Scientific question you attempt to answer, and how (briefly)

Format Remember: you are supposed to send your code so the report should not include any!

- NO R console output: use tables
- NO R messages
- NO R code anywhere ever
- NO code chunks
- NO mention of the functions you use, and no explanation of your code
- Can the report be read 100% the same way if the code was not written in R? If yes, then good; if not, then make it independent of the code. The analyses and your interpretations are important, not the specifics of your code. Good code will lead to more elegant analyses, plots and overall presentation
- NO screenshots

Contents

- All models are written in formulas
- Notation is consistent
- Estimates for all unknowns are reported in tables/plots or discussed in the text and interpreted
- Model comparisons are meaningful

Plots, figures and tables

- All plots have short description/title/caption and are numbered
- All figures numbered sequentially
- All figures are mentioned in text, in the order in which they appear
- All plots have meaningful axis titles (if not redundant e.g. in the title)
- All plots are well positioned in the page (centered)
- All text in the plots is readable without zooming in
- No text is too big in the plot
- Plots are not "warped"
- No plot is pixelated or blurry or with jpeg artifacts
- All plots are useful for the purpose of answering the research question
- All plots are explained and interpreted (not just described passively)
- Report quantities with names meaningful to the application and not with generic ones (e.g., State1, State2, ecc)

General

- Spacing is used efficiently: no excessive white spaces
- Borders are normal, line spacing is standard, no other weirdness to fit everything within the page limit
- English: spelling mistakes? Too verbose? Concise enough? We're not the British Council but you don't want to be sloppy.
- Report does not look hastily made or sloppy

- Report looks professional
- Text is concise and to the point

Code: we will randomly pick some groups for a code check. Or, we may check the code when figures or values look funny (as it happens).

- Submitted code can be compiled/run without error generating all figures and tables in the report, with the same numbers
- Code is easily readable and it is possible for anybody to understand what is going on
- Variables are named to improve readability (i.e. avoid calling things "a1" "x9534", "asdfa", but rather use names such as "user_speed", "daily_price", "log_returns".
- The code would work with minor modification on different data

References

[1] Menne, M.J., I. Durre, B. Korzeniewski, S. McNeill, K. Thomas, X. Yin, S. Anthony, R. Ray, R.S. Vose, B.E.Gleason, and T.G. Houston, 2012: Global Historical Climatology Network - Daily (GHCN-Daily), Version 3.31. NOAA National Climatic Data Center. Dataset accessed 2025-02-03 at http://doi.org/10.7289/V5D21VHZ .

[2] GISTEMP Team, 2025: GISS Surface Temperature Analysis (GISTEMP), version 4. NASA Goddard Institute for Space Studies. Dataset accessed 2025-02-03 at https://data.giss.nasa.gov/gistemp/.

[3] Lenssen, N., G.A. Schmidt, M. Hendrickson, P. Jacobs, M. Menne, and R. Ruedy, 2024: A GISTEMPv4 observational uncertainty ensemble. J. Geophys. Res. Atmos., 129, no. 17, e2023JD040179, doi:10.1029/2023JD040179.