Exploratory Analysis of Spatiotemporal Patterns in Meteorological Data

Rosa Virginia Encinas Quille[1], **Felipe Valencia de Almeida**[1],
Danielle Monteiro[1], **Pedro Luiz Pizzigatti Corrêa**[1]

[1]Escola Politécnica - Universidade de São Paulo

{encinas, fvalencia, pedro.correa}@usp.br

{danimonteirodba}@gmail.com

Abstract. *Meteorological studies of atmospheric variables impact different sectors of society, such as agriculture and health. Meteorological analysis is fundamental for atmospheric sciences, meteorological services, and international cooperation, while data quality and control are vital for environmental analysis and monitoring. This paper presents an open tool aimed at exploratory analysis of data from the Brazilian National Institute of Meteorology. A case study was carried out, involving the exploratory analysis of data from a measuring tower in the city of São Paulo. It was possible to develop a tool capable of working with data from several towers spread across Brazil, providing simplified plot views of the measured variables.*

Resumo. *Estudos meteorológicos sobre variáveis atmosféricas impactam diferentes setores da sociedade, como a agricultura e a saúde. A análise meteorológica é fundamental para as ciências atmosféricas, os serviços meteorológicos e a cooperação internacional, enquanto a qualidade e o controle dos dados são essenciais para a análise e o monitoramento ambiental. Este trabalho apresenta uma ferramenta aberta voltada para a análise exploratória de dados do Instituto Nacional de Meteorologia do Brasil. Foi realizado um estudo de caso envolvendo a análise exploratória de dados de uma torre de medição na cidade de São Paulo. Foi possível desenvolver uma ferramenta capaz de trabalhar com dados de diversas torres espalhadas pelo Brasil, oferecendo visualizações simplificadas das variáveis medidas.*

1. Introduction

Throughout history, humanity has maintained a continuous effort to understand and predict atmospheric phenomena [Anderson 2005, Daipha 2019]. Meteorology relies on a wide array of instruments that generate observations at various temporal resolutions [Defrise 1958]. Advances in measurement technologies have led to increasingly diverse and precise datasets. As a result, international standards, such as ISO/TC 146/SC 5, have been established to define technical terminology and ensure the consistency and integrity of meteorological data [Harrison 2014]. In this context of exponential data growth, the assurance of data quality has become a critical concern. Data Science plays a key role in addressing this challenge, enabling the extraction of knowledge and the generation of actionable insights from large volumes of meteorological observations.

By applying analytical and statistical techniques to meteorological time series, it is possible to identify patterns, trends, and significant weather events [Ravindra et al. 2019, Ceravolo et al. 2021]. These analyses contribute to improved climate understanding, more accurate forecasting models, and informed decision-making in response to environmental challenges. As data volumes continue to increase, there is growing demand for efficient exploratory tools that facilitate the initial understanding and structuring of time series data, especially in institutional contexts like the Brazilian National Institute of Meteorology (INMET), where such tools are still limited.

In recent years, the principles of Open Science and the availability of Open Data have gained increasing prominence in the scientific community. These practices promote transparency, reproducibility, and collaboration across disciplines. In meteorology, open access to observational data is particularly valuable, as it enables researchers to analyze and validate environmental information. The use of open-source tools to explore open meteorological datasets not only aligns with these principles but also fosters innovation and collective problem-solving in the face of climate and environmental risks.

Time series analysis is a dynamic and evolving research field, with a wide range of statistical and graphical techniques available for forecasting and anomaly detection [Hyndman and Athanasopoulos 2018]. Tools such as TimerSeacher 2 [Buono et al. 2005] and Visplore [Vuckovic and Schmidt 2020] demonstrate how visual and exploratory approaches can enhance the understanding of temporal data. Additionally, recent reviews [Choi et al. 2021] highlight the integration of deep learning methods for anomaly detection, and practical guides like [Tredennick et al. 2021] emphasize the importance of model selection in ecological time series.

Although these contributions cover various aspects of time series analysis, there remains a gap in applying Data Science specifically to the exploratory analysis of open meteorological datasets in the Brazilian context. This paper addresses this gap by presenting an open-source tool designed for the exploratory analysis of INMET data. To the best of the authors' knowledge, this is the first tool in the literature aimed at this specific purpose. A case study using data from a measurement tower located in the city of São Paulo is presented to demonstrate the tool's functionality and potential applications.

2. Material and Methods

Figure 1 presents a data science experiment method based on three steps: (i) Data acquisition from heterogeneous sources, (ii) Data ingestion and preprocessing into a unified repository, and (iii) Data analysis for extracting patterns, insights, and knowledge. Each step is described as follows.

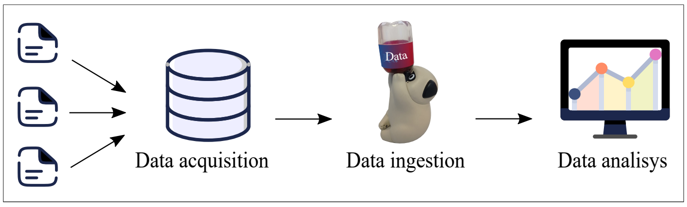
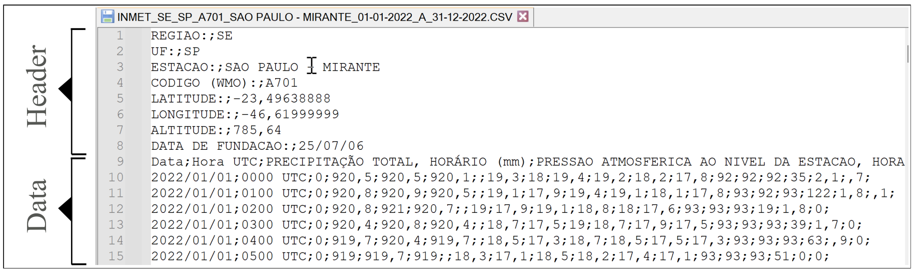


Figure 1. Overview of the proposed data science workflow, illustrating the three main stages.

2.1. Data acquisition

The first step is data acquisition. In this step, the researcher needs to find good data sources to locate the datasets that will be used in the experiment. Many government agencies usually have their own data portal to store their collected data. For example, satellite data can be found in the National Aeronautics and Space Administration (NASA) data portal. The data used in this experiment were acquired from the INMET Data Portal. It's possible to download time series data from many different measurement towers distributed in Brazil from 2 000 to 2 025. We obtained a total of 10 019 files distributed across 26 folders. Each folder contains datasets from 616 stations for its corresponding year, covering the period 2 000–2 025, with a total size of 6.87 GB.

Figure 2 shows an example of a data file downloaded from the INMET Data Portal. As can be seen, the first eight rows are always a header that presents information about the measurement tower, while the remaining rows are the data that can be analyzed. Data available from the measurement towers contains information about the measured timestamp, precipitation, temperature, humidity, solar radiation and wind speed within one hour between the measures. It is crucial to notice that the hour timestamp is in the UTC format. Since Brazil uses the UTC-3 time, we must consider that the hour information presented in the INMET data is always 3 hours ahead.



INMET_SE_SP_A701_SAO PAULO - MIRANTE_01-01-2022_A_31-12-2022.CSV

```
REGIAO:;SE
UF:;SP
ESTACAO:;SAO PAULO - MIRANTE
CODIGO (WMO):;A701
LATITUDE:;-23,49638888
LONGITUDE:;-46,61999999
ALTITUDE:;785,64
DATA DE FUNDACAO:;25/07/06
Data;Hora UTC;PRECIPITAÇÃO TOTAL, HORÁRIO (mm);PRESSAO ATMOSFERICA AO NIVEL DA ESTACAO, HORA
2022/01/01;0000 UTC;0;920,5;920,5;920,1;;19,3;18;19,4;19,2;18,2;17,8;92;92;92;35;2,1;,7;
2022/01/01;0100 UTC;0;920,8;920,9;920,5;;19,1;17,9;19,4;19,1;18,1;17,8;93;92;93;122;1,8;,1;
2022/01/01;0200 UTC;0;920,8;921;920,7;;19;17,9;19,1;18,8;18;17,6;93;93;93;19;1,8;0;
2022/01/01;0300 UTC;0;920,4;920,8;920,4;;18,7;17,5;19;18,7;17,9;17,5;93;93;93;39;1,7;0;
2022/01/01;0400 UTC;0;919,7;920,4;919,7;;18,5;17,3;18,7;18,5;17,5;17,3;93;93;93;63;,9;0;
2022/01/01;0500 UTC;0;919;919,7;919;;18,3;17,1;18,5;18,2;17,4;17,1;93;93;93;51;0;0;
```

Figure 2. INMET data file example

2.2. Data ingestion

The second step is the data ingestion. After collecting the data, it is necessary to preprocess it before doing the data analysis. Many different operations can be performed in this step, such as removing unnecessary data, filling missing values, and generating indirect parameters. The dataset was stored in a CSV file with 19 columns. In our workflow, the ingestion process was divided into two main tasks:

Data Selection and Station Metadata Extraction The raw dataset, originally stored in multiple CSV filesone for each year, was first filtered to include only the years 2 000–2 025 and ten selected stations. These was selected according to the problem to be analyzed.

Data Cleaning The data cleaning process was conducted systematically to prepare the INMET hourly dataset for subsequent analyses. The implemented code performed the following main steps: (1) Removal of unwanted columns; (2) **Standardization of dates and hours:** Dates presented initially in different formats and separators. They were normalized by replacing dots and slashes with hyphens. Parsing first attempted the ISO format (`YYYY-MM-DD`) and, when necessary, alternative formats prioritizing the Brazilian convention (day/month/year). The hour column (`Hour` or “Hora (UTC)”), which contained textual and mixed formats (e.g., “0”, “00:00”, “23 UTC”), was converted to integers in the range 0–23. Finally, a new column `datetime` was created by combining `Date` and `Hour`, ensuring that each record had a unique and consistent timestamp; (3) **Replacement of sentinel values:** INMET uses values such as `-9999` or `-9999.0` to indicate missing data. These were standardized to `NaN` to unify missing data marking; (4) **Removal of rows with complete measurement failure:** When precipitation (`Precip`) was missing, it was known that the entire row corresponded to a measurement failure. Such rows were removed. (5) **Specific treatment for solar radiation:** Missing values in the `Rad` (solar radiation) column were expected during nighttime and were replaced with zero. (6) **Filling remaining missing values in other variables:** For key meteorological variables such as atmospheric pressure (`Pstn`, `Pmax`, `Pmin`), temperature (`Temp`, `Tmax`, `Tmin`, `Dew`), relative humidity (`RH`, `RHmax`, `RHmin`) and wind (`WS`, `WSmax`, `WSdir`), interpolation was applied by averaging the previous value (forward fill) and the next value (backward fill). If missing values still remained (e.g., at the time series edges), they were filled with the column mean; (7) **Summary and saving:** After the process, quality metrics were computed such as the percentage of missing data before and after cleaning, the number of sentinel values replaced, and the amount of `NaN` values filled. The final dataset was saved in a `.csv` file containing 28 columns and 1 877 376 rows.

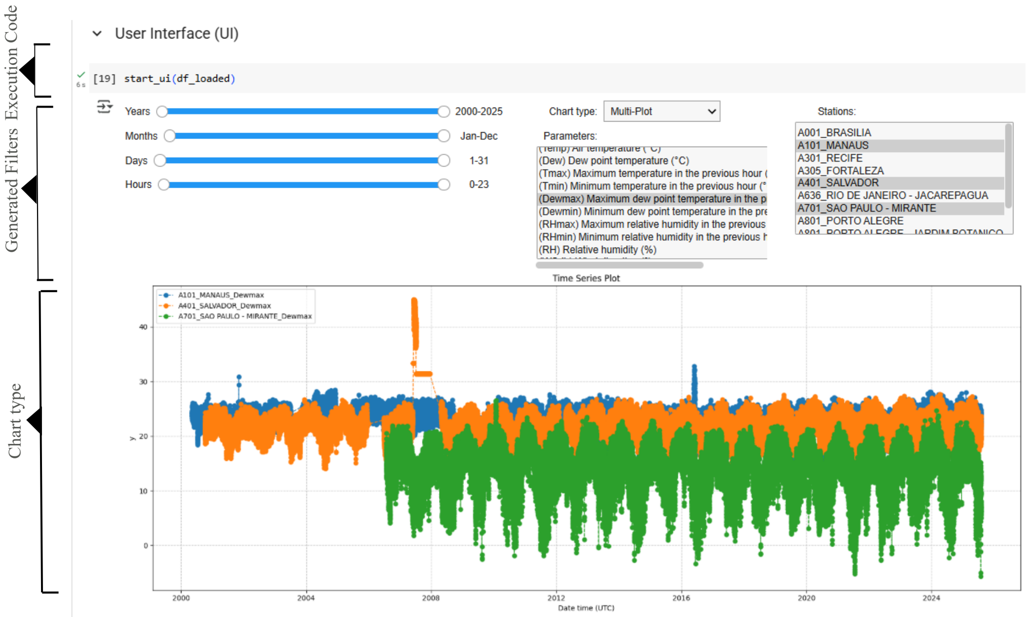
2.3. Data analisys

The exploratory analysis of the data was developed using filters that provide visual and intuitive interaction for the user. These filters offer flexibility in exploring the data interactively. Figure 3 shows the main interface of the tool. The filters can be manipulated through three types of interaction:

1. Sliders for selecting year, month, day, and hour.
2. Choice of chart type for data visualization. The types of plots considered are: (1) **Plot:** Visualizes the relationship between variables and identifies trends over time through line or scatter plots. (2) **Multi-Plot:** Compares and analyzes different datasets or variables simultaneously. (3) **Histogram:** This represents the distribution of a continuous variable by displaying the frequency or count of data points within predefined intervals or bins. (4) **Box-Plot:** Provides a concise summary of the dataset’s distribution, revealing the minimum, maximum, median, and quartiles, along with potential outliers. (5) **Pairplot:** Presents a comprehensive overview of variable interactions and correlations by displaying a matrix of scatter plots. (6) **Information Overview:** Provides three types of graphics on the same screen: Plot, box-plot, and histogram.
3. Selection of parameters (variables) to be analyzed. At least one parameter must be selected to explore the data graphically. You can also select the desired stations using the stations selection box on the right side.

3. Results and Discussion

The user can manipulate these filters through four-time sliders: year, month, day, and hour. Configuration is carried out automatically according to the imported dataset. For



User Interface (UI)

```
[19] start_ui(df_loaded)
```

Years 2000-2025

Months Jan-Dec

Days 1-31

Hours 0-23

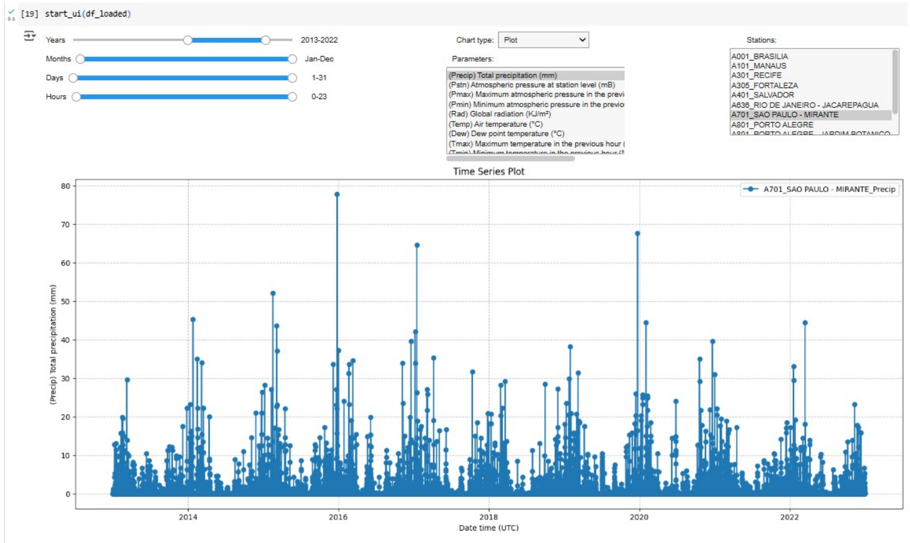
Chart type: Multi-Plot

Parameters:
(Temp) Air temperature (°C)
(Dew) Dew point temperature (°C)
(Tmax) Maximum temperature in the previous hour (
(Tmin) Minimum temperature in the previous hour (°
(Dewmax) Maximum dew point temperature in the p
(Dewmin) Minimum dew point temperature in the pre
(RHmax) Maximum relative humidity in the previous
(RHmin) Minimum relative humidity in the previous h
(RH) Relative humidity (%)

Stations:
A001_BRASILIA
A101_MANAUS
A301_RECIFE
A305_FORTALEZA
A401_SALVADOR
A636_RIO DE JANEIRO - JACAREPAGUA
A701_SAO PAULO - MIRANTE
A801_PORTO ALEGRE
A801_PORTO ALEGRE - JARDIM BOTANICO

Figure 3. User interface

example, Figure 4 shows the Plot of precipitation from 2013 to 2022. Since the measures are taken hourly, it is possible to see that most of the values are zero and there is a seasonal pattern, which is related to the seasons.



```
[19] start_ui(df_loaded)
```

Years 2013-2022

Months Jan-Dec

Days 1-31

Hours 0-23

Chart type: Plot

Parameters:
(Precip) Total precipitation (mm)
(Pstn) Atmospheric pressure at station level (mB)
(Pmax) Maximum atmospheric pressure in the previ
(Pmin) Minimum atmospheric pressure in the previo
(Rad) Global radiation (KJ/m²)
(Temp) Air temperature (°C)
(Dew) Dew point temperature (°C)
(Tmax) Maximum temperature in the previous hour (
(Tmin) Minimum temperature in the previous hour (°

Stations:
A001_BRASILIA
A101_MANAUS
A301_RECIFE
A305_FORTALEZA
A401_SALVADOR
A636_RIO DE JANEIRO - JACAREPAGUA
A701_SAO PAULO - MIRANTE
A801_PORTO ALEGRE
A801_PORTO ALEGRE - JARDIM BOTANICO

Figure 4. Precipitation Plot from 2013 to 2022

Note that the sliders allow selecting temporal subsets of data. For example, in Figure 5, we made a multiplot type graph for three temperature variables measured (mean, max, min) on December 25, 2022. Note that on that day, the temperature rises from 6 AM UTC-3 (9 AM UTC), reaching a peak of over 30 °C around 1 PM UTC-3. The filters' purpose is to select a subset of data to perform specific analyses for each situation.

Another type of plot that is available on the tool is the histogram. Figure 6 shows the histogram of the relative humidity parameter from 2013 to 2022, showing that most

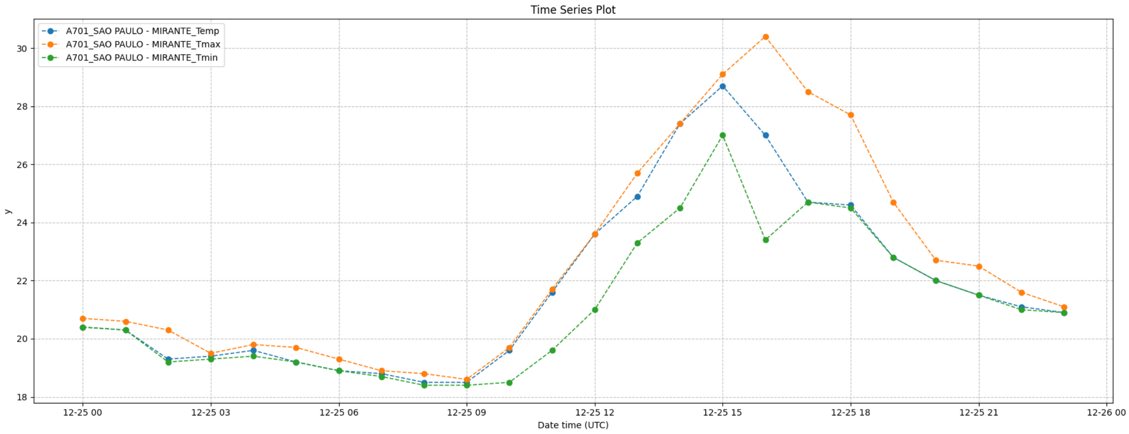


Figure 5. Multi plot for the three temperature variables (mean, max, min) on 25/12/22

measured values are located between 75% and 85%.

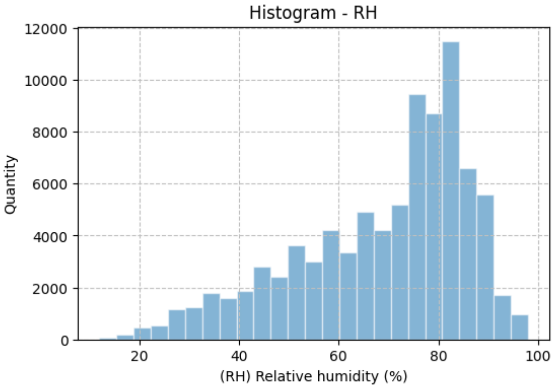


Figure 6. Relative Humidity histogram from 2013 to 2022.

Both the tool and the data used in this experiment are available at Zenodo for usage and reproducibility (DOI: `https://doi.org/10.5281/zenodo.16864784`).

The case study using the tool generated relevant plots, showcasing the tendencies of various variables in Sao Paulo from 2013 to 2022. These visual representations provide a clear and concise way to interpret complex data, aiding in identifying patterns, trends, and anomalies. By visually presenting the data, the tool enhances the effectiveness of exploratory analysis, enabling researchers to derive meaningful insights and make decisions based on the findings.

The climate forecast for ten years in São Paulo, according to the data generated by the São Paulo-Mirante station, is essential to understand climate trends and their impacts in the region—for example, the total precipitation variable presented in Figure 4. By analyzing historical data, it can be seen that the months from December to March generally have higher rainfall rates, indicating the arrival of the rainy season. Historical information plays an essential role in predicting future events, a factor of utmost importance in agriculture. By analyzing past data on rainfall patterns, farmers can effectively plan their planting and harvesting schedules, optimizing their agricultural practices based on higher or lower rainfall periods. This approach enables farmers to make informed decisions that ultimately lead to higher profits and more success in their agricultural endeavors. Currently, Agriculture 4.0 uses information systems and advanced technologies to collect meteorological data; this is reaffirmed in the work of [Zhai et al. 2020], stating that these analyzes can help with market demands, land uses, and help farmers in making decisions, as well as in the study carried out on the challenges detected in decision support systems in agriculture. In this article, we place exploratory analysis systems as a first step to carry out another type of analysis, such as using Artificial Intelligence techniques presented in the literature review by [Akkem et al. 2023].

Another example is the atmospheric pressure in Figure 7. We can identify the formation and displacement of high and low-pressure systems by analyzing the hourly values and previous maximum and minimum pressure records. For example, if there is a sharp drop in atmospheric pressure and previous lows are recorded, this could indicate the approach of a cold front, which is likely to result in changes in weather conditions, such as heavy rainfall and lower temperatures.

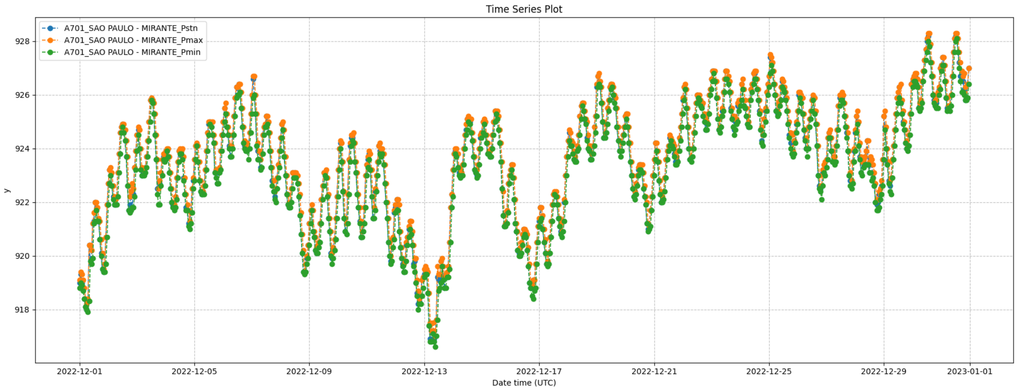


Figure 7. Multi Plot for the three Atmospheric pressure variables (mean, max, min) on 01/12/22 to 31/12/2022.

4. Conclusion

This work introduced an open-source tool for the exploratory analysis of meteorological data from INMET. The tool was validated using hourly data from ten selected INMET stations, chosen to represent different Brazilian regions and climate conditions. The stations used in the test include: Brasília, Salvador, Manaus, Porto Alegre, Cuiabá, Fortaleza, Curitiba, Recife, São Paulo – Mirante, and Rio de Janeiro – Jacarepaguá (ten stations). By applying it to a case study using data selected from the Mirante de Santana station in São Paulo (2013–2022), we demonstrated how the tool enables flexible and intuitive visualization of atmospheric variables such as temperature, precipitation, humidity, and atmospheric pressure. The ability to apply filters and dynamically generate plots allows researchers to detect seasonal trends, anomalies, and climatological patterns more effectively. Beyond its technical contributions, the tool reinforces the principles of Open Science and Open Data by promoting transparency, reproducibility, and collaborative research in environmental studies. For future work, we plan to expand the tool's applicability by integrating additional datasets and enabling statistical and predictive modeling capabilities.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

References

Akkem, Y., Biswas, S. K., and Varanasi, A. (2023). Smart farming using artificial intelligence: A review. *Engineering Applications of Artificial Intelligence*, 120:105899. https://doi.org/10.1016/j.engappai.2023.105899.

Anderson, K. (2005). *Predicting the weather: Victorians and the science of meteorology*. University of Chicago Press.

Buono, P., Aris, A., Plaisant, C., Khella, A., and Shneiderman, B. (2005). Interactive pattern search in time series. In Erbacher, R. F., Roberts, J. C., Grohn, M. T., and Borner, K., editors, *Visualization and Data Analysis 2005*, volume 5669, pages 175 – 186. International Society for Optics and Photonics, SPIE. https://doi.org/10.1117/12.587537.

Ceravolo, R., Coletta, G., Miraglia, G., and Palma, F. (2021). Statistical correlation between environmental time series and data from long-term monitoring of buildings. *Mechanical Systems and Signal Processing*, 152:107460. https://doi.org/10.1016/j.ymssp.2020.107460.

Choi, K., Yi, J., Park, C., and Yoon, S. (2021). Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines. *IEEE Access*, 9:120043–120065. https://doi.org/10.1109/ACCESS.2021.3107975.

Daipha, P. (2019). *Masters of uncertainty: Weather forecasters and the quest for ground truth*. University of Chicago Press.

Defrise, P. (1958). Petterssen, s.-introduction to meteorology. *Ciel et Terre, Vol. 74, p. 472*, 74:472.

Harrison, G. (2014). *Meteorological measurements and instrumentation*. John Wiley & Sons.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Ravindra, K., Rattan, P., Mor, S., and Aggarwal, A. N. (2019). Generalized additive models: Building evidence of air pollution, climate change and human health. *Environment International*, 132. https://doi.org/10.1016/j.envint.2019.104987.

Tredennick, A. T., Hooker, G., Ellner, S. P., and Adler, P. B. (2021). A practical guide to selecting models for exploration, inference, and prediction in ecology. *Ecology*, 102(6):e03336. https://doi.org/10.1002/ecy.3336.

Vuckovic, M. and Schmidt, J. (2020). Visual analytics approach to comprehensive meteorological time-series analysis. *Data*, 5(4). https://doi.org/10.3390/data5040094.

Zhai, Z., Martínez, J. F., Beltran, V., and Martínez, N. L. (2020). Decision support systems for agriculture 4.0: Survey and challenges. *Computers and Electronics in Agriculture*, 170:105256. https://doi.org/10.1016/j.compag.2020.105256.