CAUSAL DISCOVERY REPORT ON DWD

TECHNICAL REPORT

Causal Copilot

October 28, 2024

ABSTRACT

This report investigates the causal relationships among environmental and meteorological variables, namely altitude, temperature, precipitation, longitude, and sunshine hours, utilizing a data-driven approach to enhance our understanding of climate dynamics. The methodology involves a comprehensive causal discovery process aided by a large language model (LLM) to select suitable algorithmsPC, GES, and NOTEARSbased on statistical characteristics of the data. Preprocessing steps were performed to clean the data and manage missing values, followed by the selection of specific hyperparameters for optimal causal inference. The results indicate complex interdependencies, revealing that altitude significantly impacts both temperature and precipitation, while longitude plays a critical role in influencing temperature and sunshine hours. The findings contribute to the field by providing a refined causal graph that elucidates the interactions among these climatic factors, supported by both statistical evidence and expert knowledge, thus laying the groundwork for future investigations into climate-related phenomena.

Keywords Causal Discovery, Large Language Model, PC, Dwd

1 Introduction

In this report, we aim to explore the causal relationships among a set of environmental and meteorological variables, namely altitude, temperature, precipitation, longitude, and sunshine hours. Each of these variables plays a critical role in understanding climate dynamics and ecological interactions. For instance, altitude is known to influence both temperature and precipitation patterns, while temperature can significantly affect evaporation rates and thus precipitation outcomes. Additionally, the geographical positioning denoted by longitude may contribute to local climate variations, impacting temperature distributions. Sunshine hours further modulate temperature and precipitation, suggesting potential interconnectedness among these factors. Utilizing statistical methods for causal discovery, we will investigate the underlying mechanisms that govern these relationships, drawing on insights from climatology, ecology, and agricultural science to enhance our understanding of how these variables interact and influence one another.

2 Background Knowledge

2.1 Detailed Explanation about the Variables

- **Altitude**: This variable represents the height of a location above sea level, usually measured in meters. Altitude can have a significant effect on various environmental factors, including temperature, precipitation, and biodiversity.
- **Temperature**: This variable indicates the degree of heat present in the atmosphere, typically measured in degrees Celsius (ŕC). Temperature is a crucial factor in determining weather patterns and ecological conditions.
- **Precipitation**: This variable refers to the amount of water, in the form of rain, snow, sleet, or hail, that falls to the earth's surface over a defined period. It is measured in millimeters (mm). Precipitation is vital for understanding water availability in ecosystems and agricultural settings.

- **Longitude**: This variable specifies the east-west position of a point on the Earth's surface, measured in degrees. Longitude is important for geographical orientation and can be used in analyses that account for geographical variations in climate and weather patterns.
- **Sunshine hours**: This variable measures the amount of time (usually in hours) that sunlight is received at a specific location during a given period. Sunshine hours influence temperature and can also affect vegetation and agricultural yields.

2.2 Possible Causal Relations among these Variables

- **Altitude** → **Temperature**: Typically, as altitude increases, the temperature tends to decrease due to the thinning atmosphere at higher elevations, resulting in a cooler climate.
- **Temperature** → **Precipitation**: Higher temperatures can lead to increased evaporation rates, causing more moisture in the atmosphere. This can result in increased precipitation as warmer air can hold more water vapor, leading to clouds and eventual rainfall or snowfall.
- **Altitude** → **Precipitation**: Higher altitudes can enhance orographic lifting, where moist air is forced to rise over mountains, cooling and condensing to produce greater precipitation in these elevated areas.
- **Longitude** → **Temperature**: The longitudinal position can influence climate due to the different thermal properties of land and water bodies, as well as variations in predictable weather patterns, which alter temperature profiles across different regions.
- **Sunshine Hours** → **Temperature**: An increase in sunshine hours generally leads to higher temperatures, as more solar radiation contributes to heating the earth's surface, influencing local temperature measurements throughout various times of the day and year.
- **Sunshine Hours** → **Precipitation**: Areas characterized by higher sunshine hours often have lower precipitation levels, especially in arid or semi-arid climates, while regions with fewer sunshine hours may experience higher instances of precipitation due to increased cloud cover and moisture retention.
- **Temperature** → **Sunshine Hours**: Higher temperatures can lead to fewer clouds and clearer skies, resulting in increased sunshine hours in a location, thereby reinforcing a feedback loop where temperature and sunshine influence each other.
- **Longitude** → **Precipitation**: Longitude can also determine precipitation patterns as it affects the proximity to oceanic influences, which provide moisture necessary for precipitation, indicating a complex interaction with regional climatic changes.

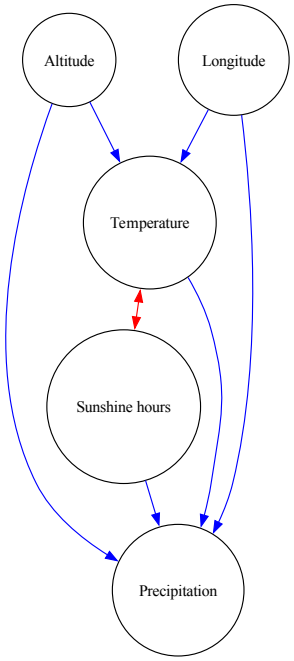


Figure 1: Possible Causal Relation Graph

3 Dataset Descriptions and EDA

The following is a preview of our original dataset.

Table 1: Dataset Preview

Altitude	Temperature	Precipitation	Longitude	Sunshine hours
205.0	9.7	828.3	6.08	1552.0
46.0	8.2	791.5	10.20	1443.3
794.0	6.4	1057.8	9.00	1096.5
325.0	8.1	803.0	13.08	1571.7
500.0	6.2	1252.2	10.43	1368.0

3.1 Data Properties

We employ several statistical methods to identify data properties.

The shape of the data, data types, and missing values are assessed directly from the dataframe. Linearity is evaluated using Ramseys RESET test, followed by the Benjamini & Yekutieli procedure for multiple test correction. Gaussian noise is assessed through the Shapiro-Wilk test, also applying the Benjamini & Yekutieli procedure for multiple test correction. Time-Series and Heterogeneity are derived from user queries.

Properties of the dataset we analyzed are listed below.

Table 2: Data Properties

Shape (n x d)	Data Type	Missing Value	Linearity	Gaussian Errors	Time-Series	Heterogeneity
(349, 5)	Continuous	False	False	False	False	False

3.2 Distribution Analysis

The following figure shows distributions of different variables. The orange dash line represents the mean, and the black line represents the median. Variables are categorized into three types according to their distribution characteristics.

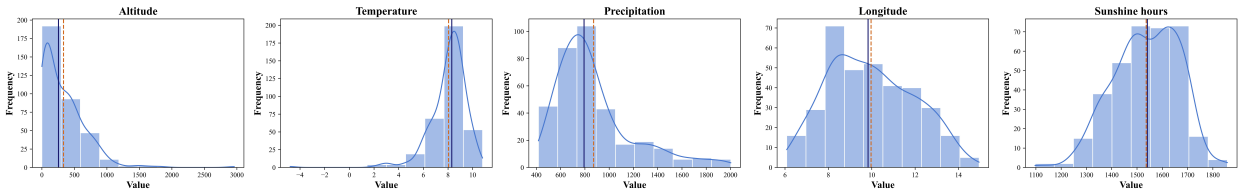


Figure 2: Distribution Plots of Variables

- Slight left skew distributed variables: None
- Slight right skew distributed variables: Altitude, Precipitation, Sunshine hours
- Symmetric distributed variables: Temperature, Longitude

3.3 Correlation Analysis

In this analysis, we will categorize the correlation statistics of features in the dataset into three distinct categories: Strong correlations ($r > 0.8$), Moderate correlations ($0.5 < r < 0.8$), and Weak correlations ($r < 0.5$).

- Strong Correlated Variables: Temperature and Altitude
- Moderate Correlated Variables: Precipitation and Altitude, Precipitation and Temperature
- Weak Correlated Variables: None

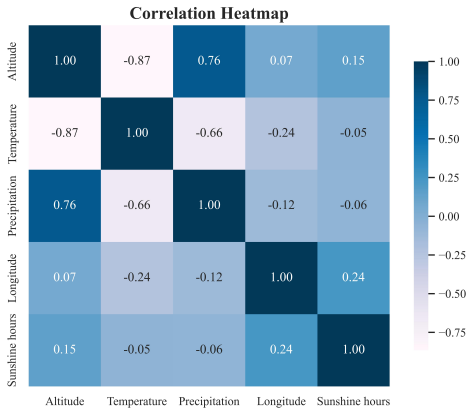


Figure 3: Correlation Heatmap of Variables

4 Discovery Procedure

In this section, we provide a detailed description of the causal discovery process implemented by Causal Copilot. We also provide the chosen algorithms and hyperparameters, along with the justifications for these selections.

4.1 Data Preprocessing

In this initial step, we preprocessed the data and examined its statistical characteristics. This involved cleaning the data, handling missing values, and performing exploratory data analysis to understand distributions and relationships between variables.

4.2 Algorithm Selection assisted with LLM

Following data preprocessing, we employed a large language model (LLM) to assist in selecting appropriate algorithms for causal discovery based on the statistical characteristics of the dataset and relevant background knowledge. The top three chosen algorithms, listed in order of suitability, are as follows:

- **PC**:
  - **Description**: The PC algorithm is a constraint-based method that learns the structure of a causal graph from data by testing conditional independencies between variables. It constructs a directed acyclic graph (DAG) representing the causal relationships.
  - **Justification**: Given the dataset's large sample size and the absence of missing values, the PC algorithm is efficient for quickly identifying the causal structure in the data while accommodating the non-linear relationships and continuous nature of the variables.
- **GES**:
  - **Description**: Greedy Equivalence Search (GES) is a score-based causal discovery algorithm that identifies the optimal causal structure by navigating the space of equivalence classes of Directed Acyclic Graphs (DAGs).
  - **Justification**: GES is well-suited for large datasets and can handle non-Gaussian distributions, which fits the dataset's characteristics. Its ability to efficiently explore complex search spaces makes it a strong candidate for the dataset.
- **NOTEARS**:
  - **Description**: NOTEARS transforms the combinatorial problem of learning Directed Acyclic Graphs (DAGs) into a continuous optimization problem, making it scalable for large datasets.
  - **Justification**: NOTEARS is suitable for this dataset, particularly due to its adaptability to high-dimensional data and its capability to handle non-linear relationships, aligning with the observed relationships in the dataset.

4.3 Hyperparameter Values Proposal assisted with LLM

Once the algorithms were selected, the LLM aided in proposing hyperparameters for the [ALGO] algorithm, which are specified below:

- **alpha**:
  - **Value**: 0.01
  - **Explanation**: Given the large sample size of 349, a lower significance level (0.01) is appropriate to reduce the risk of false positives in identifying true causal relationships. This is especially important given that the dataset is not heterogeneous and the relationships may not be linear.
- **indep_test**:
  - **Value**: fisherz
  - **Explanation**: Fisher's Z test is suitable here as the data is continuous. While typically it assumes linearity and Gaussian distribution, the PC algorithm can still effectively utilize Fisher's Z in the context of large sample sizes like 349, despite the noted non-linearities.
- **uc_rule**:
  - **Value**: 0
  - **Explanation**: Using the default value of 0 is appropriate as it adheres to the standard PC algorithm approach, which is well-suited for the characteristics of this dataset.
- **uc_priority**:
  - **Value**: 2

– **Explanation**: Prioritizing stronger colliders can enhance the robustness of causal inference. A value of 2 offers a balanced approach to resolve conflicts while being aware of potential relationships that may exist among variables.

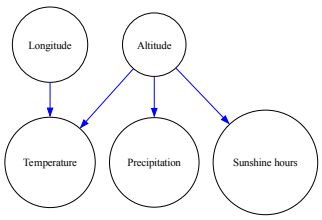
4.4 Graph Tuning with LLM Suggestion

In the final step, we performed graph tuning with suggestions provided by the LLM. We utilize LLM to help us determine the direction of undirected edges according to its knowledge repository. By integrating insights from the LLM to refine the causal graph, we can achieve improvements in graph's accuracy and robustness.

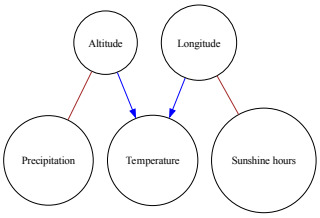
- **Altitude** $\rightarrow$ **Precipitation**: Higher altitude often leads to orographic precipitation, where moist air rises over mountains, cools, and condenses, resulting in increased precipitation.
- **Sunshine hours** $\rightarrow$ **Longitude**: Longitude influences geographical and climatic conditions, which can determine the amount of sunshine hours received at a specific location.

This structured approach ensures a comprehensive and methodical analysis of the causal relationships within the dataset.

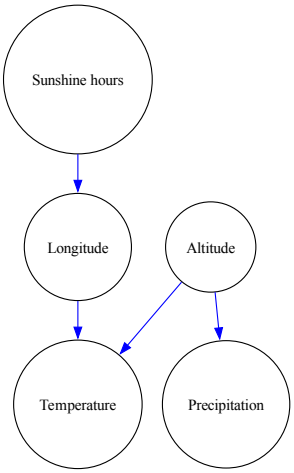
5 Results Summary



(a) True Graph



(b) Initial Graph



(c) Revised Graph

Figure 4: Graphs Comparision of PC

The above are result graphs produced by our algorithm. The initial graph is the graph in the first attempt, and the revised graph is the one pruned with LLM suggestion.

The causal relationships among the variables reveal a complex interplay influenced by geographic and atmospheric factors. Altitude impacts Temperature, as higher elevations typically result in lower temperatures due to thinner air and reduced atmospheric pressure. Additionally, altitude also affects Precipitation, as mountains can induce orographic lift, leading to increased rainfall in higher regions. Interestingly, Precipitation appears to have a feedback effect on Altitude, suggesting that the accumulation of precipitation may contribute to geological processes that shape landforms over time. Furthermore, Longitude affects Temperature, likely due to variations in climate zones across different longitudes, which influence average temperatures. Longitude is also linked to Sunshine hours, as geographical position determines the amount of sunlight received throughout the year. There is a reciprocal influence between Sunshine hours and Longitude, indicating that changes in sunlight exposure may also reflect variations in longitudinal positioning. Overall, these connections highlight the dynamic interactions among altitude, temperature, precipitation, longitude, and sunshine hours in shaping climatic and environmental patterns.

5.1 Graph Reliability Analysis

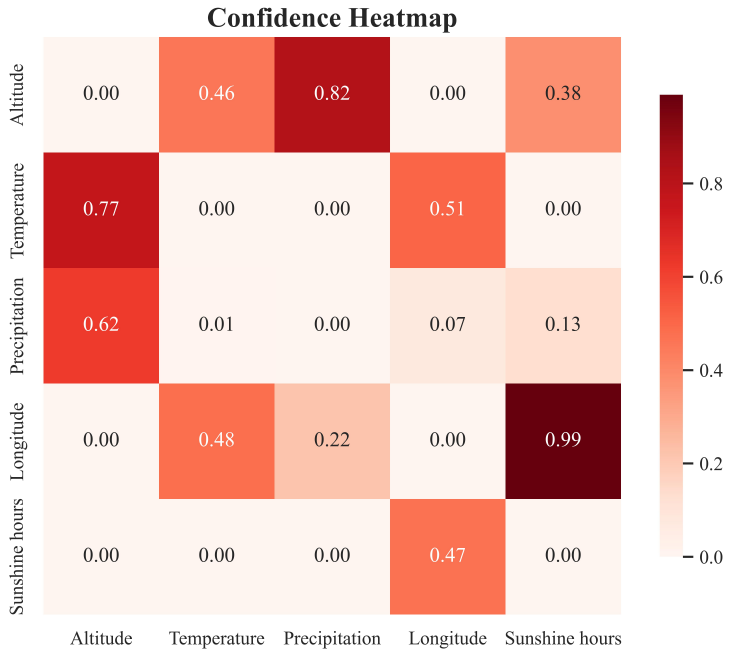


Figure 5: Reliability Graph

Based on the confidence probability heatmap and background knowledge, we can analyze the reliability of our graph.

From the statistics perspective, we have high confidence to believe that the edge between Longitude and Sunshine hours exists (bootstrap probability of 0.99), and the edge between Altitude and Temperature exists (bootstrap probability of 0.46) does not hold much confidence. Additionally, we observe that the edges Altitude $\rightarrow$ Precipitation (0.82) and Precipitation $\rightarrow$ Altitude (0.62) show reasonably high bootstrap probabilities, suggesting a strong mutual influence. The edge between Longitude and Temperature (0.48) also lacks sufficient confidence, while the edge Sunshine hours $\rightarrow$ Longitude (0.47) indicates a low likelihood of a causal relationship.

However, based on expert knowledge, we know that it is widely understood that higher altitudes typically result in lower temperatures, supporting the idea that altitude influences temperature despite the low bootstrap probability. Furthermore, orographic precipitation due to altitude reinforces the edge between Altitude and Precipitation. The relationship between Sunshine hours and Precipitation can also be rationalized based on geographical and climatological patterns, where locations with higher sunshine hours typically correlate with lower precipitation, reinforcing the idea of a causal connection.

Therefore, while certain edges, particularly those involving Longitude and Temperature, show low statistical confidence, expert knowledge supports the existence of relationships primarily between Altitude, Temperature, and Precipitation. As such, the result of this causal graph can be considered partially reliable, needing careful interpretation and possibly further analysis to validate uncertainty in less confident edges.