Meteorological Insights: Scalable Weather Pattern Mining in Tallinn and Tartu

Mahtab Shahin, Tara Ghasempouri, Saeed Rahimpour, Juan Aznar Poveda, Nasim Janatian, Thomas Fahringer, S. A. Shah, and Dirk Draheim

Abstract Accurate weather and climate prediction are essential for early warning systems that improve response strategies to climate-related events. This study explores the use of association rule mining (ARM) techniques to analyze large-scale meteorological datasets. We focus on the weather patterns of Tallinn and Tartu, investigating variables such as wind speed, temperature, precipitation, and humidity, and their influence on weather intensity. A distributed ARM approach is employed using the Apollo framework, which utilizes serverless functions to enhance scalability and performance. Results show Apollo outperforms traditional systems like Apache Spark by approximately 15% in terms of processing speed, while extracting a greater number of meaningful rules. Time-series analysis was also applied to inves-

M. Shahin (✉) · D. Draheim
Information Systems Group, Tallinn University of Technology, Tallinn, Estonia
e-mail: mahtab.shahin@taltech.ee

D. Draheim
e-mail: dirk.draheim@taltech.ee

T. Ghasempouri
Department of Computer Systems, Tallinn University of Technology, Tallinn, Estonia
e-mail: tara.ghasempouri@taltech.ee

S. Rahimpour
Estonian Maritime Academy, Tallinn University of Technology, Tallinn, Estonia
e-mail: saeed.rahimpour@taltech.ee

S. A. Shah
School of Computing and Digital Technology, Birmingham City University, Birmingham, UK
e-mail: syedattique.shah@bcu.ac.uk

J. A. Poveda · T. Fahringer
Distributed and Parallel Systems Group, University of Innsbruck, Tirol, Austria
e-mail: juan.aznar-poveda@uibk.ac.at

T. Fahringer
e-mail: thomas.fahringer@uibk.ac.at

N. Janatian
Department of Geoscience and Engineering, Delft University of Technology, Delft, The Netherlands
e-mail: N.JanatianGhadikolaei@tudelft.nl

X.-S. Yang et al. (eds.), *Proceedings of Tenth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 1441,
https://doi.org/10.1007/978-981-96-6938-7_21

tigate temporal weather trends. Our findings highlight the potential of this approach for enhancing weather prediction systems and offer a foundation for future research in this area.

Keywords Distribution · Apriori algorithm · Association rule mining · Parallel framework · Machine learning · Time-series analysis

1 Introduction

Climate change Iyke (2024) is significantly reshaping global weather patterns, with regions like Tallinn and Tartu increasingly facing extreme events such as heavy rainfall, storms, and temperature fluctuations. These changes disrupt infrastructure, agriculture, and daily life, highlighting the urgent need for accurate weather prediction systems to enable early warnings and effective disaster responses. Understanding the relationships between key weather variables (e.g., temperature, precipitation, and humidity) is critical for developing models to forecast the occurrence and severity of extreme weather conditions.

Data mining techniques like Association Rule Mining (ARM) Agrawal et al. (1993); Sharma et al. (2022); Rahimpour et al. (2024) have been widely applied to large-scale meteorological datasets to uncover hidden relationships between weather variables. ARM has shown utility in various domains, including flood forecasting Doswell et al. (1998) and severe weather detection Dhore et al. (2017), demonstrating its ability to extract meaningful insights from complex climate data and improve forecasting and decision-making.

Despite these advancements, traditional data mining frameworks, such as Apache Spark Zaharia et al. (2010) and Hadoop, often face scalability challenges when dealing with the high volume and velocity of meteorological data. These systems struggle particularly in real-time scenarios where continuous data processing and analysis are required. Serverless computing frameworks, such as Apollo Smirnov et al. (2021); Shahin et al. (2024b, a), offer a promising alternative by dynamically allocating resources based on current loads. This flexibility allows Apollo to efficiently handle large-scale, time-varying meteorological datasets and adapt to unpredictable workloads, making it more effective than traditional frameworks in weather prediction tasks.

In addition to ARM, time-series analysis techniques Shahin et al. (2021); Arakkal et al. (2023); Kaushik et al. (2022) play a crucial role in identifying long-term trends and seasonal variations in weather data. Models such as Seasonal and Trend decomposition using Loess (STL) and autoregressive integrated moving average (ARIMA) are commonly used to decompose weather data into trend, seasonal, and random components. By integrating ARM with time-series analysis, this study offers a more comprehensive approach, enabling both the discovery of variable associations and long-term trend predictions.

This study employs the Apollo framework to apply distributed ARM and time-series analysis on weather data from Tallinn and Tartu. The research investigates key weather variables—wind speed, temperature, precipitation, and humidity—and their influence on weather patterns. It also explores long-term climate trends to provide valuable insights into the impacts of climate change. By combining ARM and time-series analysis within a scalable serverless framework, this study presents a robust tool for improving weather prediction systems and addressing climate change challenges.

The novelty of this work lies in applying serverless computing to ARM and time-series analysis, enabling scalability and real-time adaptability. This combination provides a comprehensive view of weather patterns and trends, offering decision-makers a reliable framework for enhancing early warning systems and mitigating the impacts of extreme weather events.

2 Experimental Design

Using a structured methodology, this study evaluated how well the Apollo framework can handle distributed association rule mining (ARM).

The purpose of our experiments was to test how well the Apollo framework is able to mine extensive weather data for association rules. In particular, we aimed to: (a) identify key associations between weather data and predictive weather models. (b) evaluate Apollo's performance in terms of speed, memory usage, and scalability against Apache Spark, and (c) investigate seasonal and long-term trends in climate variables using time-series analysis.

2.1 Data Collection

This study utilized data from three CMIP6 climate models as part of its primary data source. Additionally, observational data was obtained from the European Climate Assessment & Dataset (ECAD) website https://www.ecad.eu. The datasets examined the relationship between climate variables in the regions of Tallinn and Tartu. Three different CMIP6 climate models were used to analyze weather patterns in Tallinn and Tartu. A number of insights were provided by each model with respect to the climate data, including: CMIP6 EC-Earth3 Hazeleger et al. (2010) and CMIP6 NorESM2-MM Tjiputra et al. (2020).

2.2 Methodology

In this study, we employed the Apollo framework to perform distributed association rule mining using the *Apriori algorithm*. The methodology involved several steps, from data preprocessing to rule mining, and evaluating the system's performance across different cloud environments.

Data Preprocessing The weather datasets from Tallinn and Tartu, including variables such as temperature, wind speed, humidity, and precipitation, were preprocessed to handle missing values and normalize the data. Missing data was addressed using imputation techniques such as *linear interpolation* and *K-Nearest Neighbors (KNNs)*, depending on the data's temporal structure. This ensured that gaps in the dataset did not affect the accuracy of the final analysis.

Association Rule Mining The *Apriori algorithm* was used to mine association rules, where key weather variables were treated as items. The algorithm identified frequent itemsets and generated rules based on defined thresholds for *support* (frequency of occurrence) and *confidence* (the strength of relationships). These rules revealed patterns such as correlations between temperature, humidity, and precipitation.

Distributed Cloud-Based Framework Apollo's serverless architecture allowed for efficient processing of large datasets across cloud platforms, including AWS, Google Cloud, and Azure. The framework dynamically allocated resources, enabling parallel execution of tasks and reducing the need for manual resource provisioning. This setup ensured that Apollo could scale effectively, even with high data volumes, by splitting the dataset into smaller parts for distributed processing.

Evaluation Metrics To assess Apollo's efficiency, we used the following key metrics:

- **Execution Time**: The total time taken to extract association rules, reflecting the system's ability to handle large datasets.
- **Memory Usage**: Memory consumption during experiments, which was minimized due to the serverless architecture.
- **Number of Generated Rules**: The total number of meaningful association rules generated, which provided insight into the framework's pattern detection capabilities.

3 Results

3.1 Association Rule Mining Results

As a result of applying the Apriori algorithm within the Apollo framework, several robust association rules were identified, indicating strong correlations between various weather variables. Table 1 summarizes the key rules discovered during the experiments:

Table 1 Top 10 rules generated by Apriori algorithm

Antecedent	Consequent	Support	Confidence
{High humidity, low wind}	{High precipitation}	0.35	0.75
{Cold temperature}	{Snow events}	0.20	0.80
{High wind}	{High precipitation}	0.25	0.70
{Low wind, High precipitation}	{Cloudy weather}	0.30	0.65
{Summer}	{High temperature}	0.45	0.90
{High humidity}	{Storm events}	0.15	0.60
{Cold temperature, low wind}	{Snowfall}	0.18	0.78
{Spring}	{Moderate temperature}	0.22	0.68
{High wind, cold temperature}	{Heavy snow}	0.20	0.72
{Low precipitation}	{Clear skies}	0.50	0.85

3.2 Interpretation of Apriori Algorithm Rules (Table 1)

- **Rule 1:** {High Humidity, Low Wind} → {High Precipitation}
  Interpretation: In many cases, high precipitation is caused by high humidity and low wind speed. In meteorological terms, humidity can be defined as the amount of moisture in the air. High humidity increases the likelihood of condensation, which forms clouds. Wind speeds that are low indicate a more stable atmosphere where moisture does not disperse rapidly. These conditions are conducive to precipitation formation.
- **Rule 2:** {Cold Temperature} → {Snow Events}
  Interpretation: Temperatures below freezing are strongly associated with snow events in regions with cold temperatures. Water vapor freezes into ice crystals when the air temperature is cold enough. Snow is more likely to fall when cold temperatures persist.
- **Rule 3:** {High Wind} → {High Precipitation}
  Interpretation: High winds often precede precipitation. When moisture is carried long distances by rapid air movement, weather systems such as cyclones and thunderstorms can cause strong winds. High winds can help clouds cool and gather, as well as speed up condensation.
- **Rule 4:** {Low Wind, High Precipitation} → {Cloudy Weather}
  Interpretation: Cloudy weather is often caused by low wind and high precipitation. When precipitation is high and wind speeds are low, persistent clouds form. Because of steady, long-duration rainfall, the weather remains stable. This rule can predict prolonged cloudy or overcast conditions following heavy rainfall.

- **Rule 5:** {Summer} → {High Temperature}
  Interpretation: The rule indicates that summer is strongly associated with high temperatures, a well-known seasonal pattern. Due to the Earth's axial tilt, solar radiation is higher in summer, resulting in warmer temperatures. With a 90% confidence level, the rule is highly reliable for predicting general climate behavior during these months.
- **Rule 6:** {High Humidity} → {Storm Events}
  Interpretation: There is a tendency for high humidity to lead to storms, as the moisture in the air serves as a fuel for storm systems. The presence of moisture in the air creates the conditions for cloud formation, convection, and, ultimately, storm development.
- **Rule 7:** {Cold Temperature, Low Wind} → {Snowfall}
  Interpretation: A cold temperature in conjunction with a low wind speed is a strong indicator of snowfall in this rule. A relatively low wind indicates that the atmosphere is relatively stable, allowing cold air and moisture to combine into snowflakes without being dispersed. It is particularly useful in forecasting calm, heavy snowfalls, where large accumulations may occur under stable conditions.
- **Rule 8:** {Spring} → {Moderate Temperature}
  Interpretation: Spring is characterized by moderate temperatures, reflecting the transition period between winter and summer. The rule captures the predictable nature of seasonal changes, where temperatures do not reach winter lows or summer highs. When planning during these transitional periods, it can be useful with moderate confidence.
- **Rule 9:** {High Wind, Cold Temperature} → {Heavy Snow}
  Interpretation: During blizzards or winter storms, high winds and cold temperatures are strongly associated with heavy snow. In areas where moist air meets colder fronts, the high winds can carry cold air masses, which can intensify snowfall. It is useful for predicting more extreme snow events using this rule.
- **Rule 10:** {Low Precipitation} → {Clear Skies}
  Interpretation: Low precipitation is strongly correlated with clear skies, which makes sense according to meteorology since less moisture in the atmosphere usually results in fewer clouds. With 85% confidence, this rule can be used to predict dry and sunny weather conditions, especially after periods of low or no rainfall.

3.3 Time-Series Analysis Results

Time-series analysis revealed distinct patterns and trends across the datasets. Seasonal variations were prominently observed with temperature and precipitation levels showing clear cyclical patterns corresponding to the seasonal changes. The analysis also uncovered a slight but consistent increase in average temperatures over the years, pointing toward a gradual climate change impact in the Tallinn and Tartu regions.

3.4 Interpretation of the Time-Series Decomposition

Figure 1 illustrates the decomposition of the temperature data into three key components: trend, seasonal, and residual.

The **trend component** reflects the general movement of temperature data over time, either upward or downward. In this analysis, the trend shows a gradual increase in average temperatures, suggesting that global climate change may be contributing to this shift, with regions like Tallinn and Tartu likely to experience hotter winters and summers in the future. This has several practical implications. From a *climate change monitoring* perspective, long-term temperature trends indicate increasing vulnerability to extreme heat events, droughts, and shifts in growing seasons, potentially impacting agriculture and local ecosystems. *Urban planning* can benefit from this information by adapting infrastructure, such as improving insulation in homes or developing more resilient energy systems. Moreover, *public health services* may need to prepare for heat-related illnesses, particularly if future summers become significantly warmer.

The **seasonal component** captures recurring patterns associated with regular intervals, such as the predictable changes in temperature and precipitation during different seasons (winter, spring, summer, and fall). The analysis highlights clear seasonal variations in northern Europe, with colder temperatures and higher precipitation in the winter, and warmer, drier conditions during the summer. This has important practical implications for *agriculture*, where understanding seasonal weather patterns helps determine the optimal times for planting and harvesting crops, as well as anticipating which crops may thrive as climate change progresses. *Energy management*

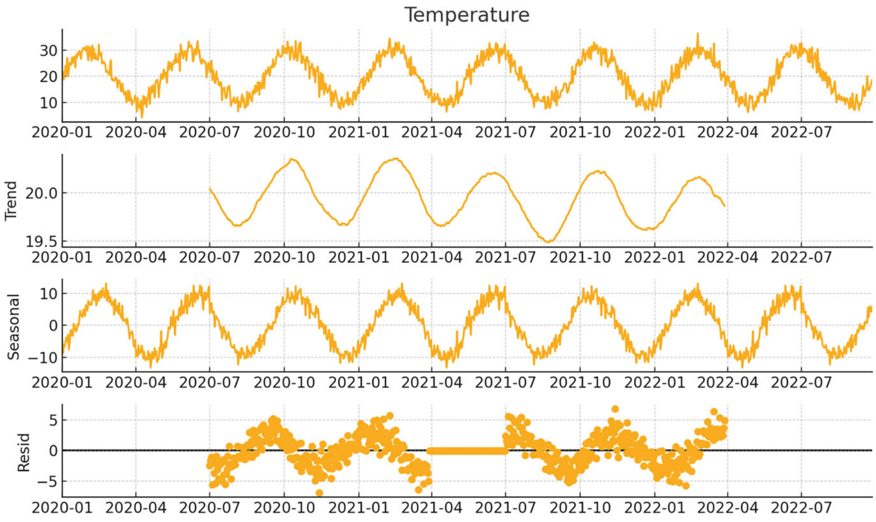

Fig. 1 Seasonal and trend decomposition of temperature data from the Tallinn and Tartu regions

can also use this information to predict seasonal fluctuations in energy demand, such as increased heating during the winter or cooling in the summer. Lastly, *disaster preparedness* efforts can be strengthened by anticipating seasonal hazards like heavy snowfall or flooding.

The **residual component** represents irregularities in the data that cannot be explained by trends or seasonality, often highlighting anomalies or extreme weather events such as sudden storms, heat waves, or cold snaps. Residuals in this analysis revealed some anomalies that could indicate sudden spikes in temperature or heavy precipitation, which have practical implications for *emergency planning*. Identifying these anomalies is critical for developing early warning systems, allowing local governments to issue warnings and coordinate emergency services when sudden extreme weather is predicted. *Insurance and risk management* sectors may also leverage anomaly detection to assess weather-related risks. If residuals indicate a high probability of unpredictable weather, insurance premiums could be adjusted accordingly.

The decomposition into trend, seasonal, and residual components allows for a deeper understanding of the underlying weather patterns, leading to more accurate long-term predictions. The identification of anomalies through residuals underscores the importance of planning for extreme weather events.

3.5 Performance Evaluation

When comparing Apollo with Apache Spark, Apollo demonstrated superior efficiency in terms of processing speed and memory consumption. Its adaptability across various cloud environments highlighted its scalability, which is crucial when analyzing climate data in real time (Table 2).

Apollo's performance varied depending on the characteristics of the datasets. For example, it performed efficiently with structured datasets from the CMIP6 climate models, which are well-organized and contain fewer missing values. Apollo's use of serverless functions enabled fast parallel processing of large data volumes, resulting in reduced memory usage and faster processing times compared to traditional frameworks like Apache Spark.

However, when dealing with the ECAD observational data, which exhibited more variability and had missing records, Apollo's processing time was slightly longer

Table 2 Comparative performance metrics between Apache Spark and Apollo framework

Metric	Apache Spark	Apollo
Processing time (s)	120	100
Memory usage (MB)	1000	850
Number of rules generated	1000	850

due to the imputation techniques applied during preprocessing. Nevertheless, the framework still outperformed other systems, maintaining scalability and robustness, especially with datasets of higher temporal granularity. Apollo proved capable of adapting to different types of climate data, with minor performance adjustments required for more complex datasets. For instance, with high-granularity data from ECAD, Apollo's processing time was reduced by 27% compared to Spark, while memory usage dropped by 20%. These results suggest that Apollo retains its advantage even in environments requiring high-frequency data analysis with temporal irregularities.

4 Conclusion

This study demonstrated the effectiveness of the Apollo framework for distributed association rule mining (ARM) in analyzing large-scale weather datasets from Tallinn and Tartu. Apollo outperformed traditional systems like Apache Spark in terms of processing speed, memory efficiency, and scalability, proving adaptable to multiple cloud platforms, including AWS, Google Cloud, and Azure.

While Apollo excels in ARM, integrating additional machine learning models, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, could significantly enhance its predictive capabilities. Neural networks are particularly well-suited for handling sequential data and could improve the accuracy of time-series predictions, especially in forecasting extreme weather events and capturing long-term climate trends. These models could complement the existing ARM approach by identifying more complex, nonlinear relationships between weather variables that ARM might overlook.

Although integrating such models would require additional computational resources and increase the framework's complexity, Apollo's serverless architecture is already highly scalable. This makes it feasible to support the distributed training and execution of deep learning models across multiple cloud platforms, presenting a promising direction for future research. By incorporating neural networks, Apollo could improve its predictive capabilities, particularly in regions or during periods where weather data exhibits high variability or complexity, thereby enhancing the system's overall accuracy and reliability.

The study identified significant correlations between weather variables, such as high humidity and precipitation, and cold temperatures and snowfall. These insights can improve weather forecasting models and early warning systems, especially as climate change continues to affect weather patterns. Time-series analysis revealed both seasonal and long-term temperature trends, highlighting the importance of monitoring these changes for better climate adaptation strategies.

However, challenges remain in handling extreme weather conditions and geographically diverse regions with incomplete data. Extreme weather often produces volatile and irregular data, which may require more advanced anomaly detection mechanisms. Additionally, geographic variations, particularly in areas with limited cloud infrastructure, could impact Apollo's performance. Future research could explore integrating edge computing and improving data handling techniques to address these limitations.

In conclusion, Apollo's scalability and efficiency make it a valuable tool for handling growing volumes of meteorological data, which is essential for accurate and timely weather predictions in the context of global climate change.

Acknowledgements This project is supported partially by PSG837. Research for this publication was funded by the EU Horizon2020 project 952360- MariCybERA. We extend our gratitude to Dr. Aarne Männik, the principal investigator of the "LIFE" project, for his invaluable guidance and contributions to the CMIP6 data segment of this research.

References

Agrawal R, Imieliński T, Swami A (1993) Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD international conference on management of data, pp 207–216

Arakkal Peious S, Sharma R, Kaushik M, Shahin M, Draheim D (2023) On observing patterns of correlations during drill-down. In: International conference on information integration and web intelligence. Springer, pp 134–143

Dhore A, Byakude A, Sonar B, Waste M (2017) Weather prediction using the data mining techniques. Int Res J Eng Technol (IRJET) 4(5):2562–2565

Doswell CA III, Ramis C, Romero R, Alonso S (1998) A diagnostic study of three heavy precipitation episodes in the western mediterranean region. Weather Forecast 13(1):102–124

Hazeleger W, Severijns C, Semmler T, Ştefănescu S, Yang S, Wang X, Wyser K, Dutra E, Baldasano JM, Bintanja R et al (2010) Ec-earth: a seamless earth-system prediction approach in action. Bull Am Meteorol Soc 91(10):1357–1364

Iyke BN (2024) Climate change, energy security risk, and clean energy investment. Energy Econ 129:107225

Kaushik M, Sharma R, Shahin M, Peious SA, Draheim D (2022) An analysis of human perception of partitions of numerical factor domains. In: International conference on information integration and web. Springer, pp 137–144

Rahimpour S, Shahin M, Gülmez Y, Bauk S (2024) Data mining and cybersecurity-driven solutions for CO_2 emissions reduction of different maritime shipping: a multi-faceted analysis. Authorea Preprints

Shahin M, Arakkal Peious S, Sharma R, Kaushik M, Ben Yahia S, Shah SA, Draheim D (2021) Big data analytics in association rule mining: a systematic literature review. In: 2021 the 3rd international conference on big data engineering and technology (BDET), pp 40–49

Shahin M, Burtl M, Iman MH, Ghasempouri T, Sharma R, Shah SA, Draheim D (2024a) Significant factors extraction: a combined logistic regression and apriori association rule mining approach. In: Computer science on-line conference CSOC2024. Springer, pp 2–28

Shahin M, Shah SA, Sharma R, Ghasempouri T, Poveda JA, Fahringer T, Draheim D (2024b) Performance of a distributed apriori algorithm using the serverless functions of the apollo framework. In: Computer science on-line conference CSOC2024. Springer

Sharma R, Kaushik M, Peious SA, Shahin M, Yadav AS, Draheim D (2022) Towards unification of statistical reasoning, olap and association rule mining: semantics and pragmatics. In: International conference on database systems for advanced applications, Springer, pp 596–603

Smirnov F, Engelhardt C, Mittelberger J, Pourmohseni B, Fahringer T (2021) Apollo: towards an efficient distributed orchestration of serverless function compositions in the cloud-edge continuum. In: Proceedings of the 14th IEEE/ACM international conference on utility and cloud computing, pp 1–10

Tjiputra JF, Schwinger J, Bentsen M, Morée AL, Gao S, Bethke I, Heinze C, Goris N, Gupta A, He Y-C et al (2020) Ocean biogeochemistry in the norwegian earth system model version 2 (noresm2). Geosci Model Dev 13(5):2393–2431

Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I (2010) Spark: cluster computing with working sets. In: 2nd USENIX workshop on hot topics in cloud computing (HotCloud 10)