Towards statistical-based prediction models for seasonal precipitation in the southeast of Brazil

Helder Arruda[1], Anita Drumond[1], Ewerton Oliveira[1], Julio Freitas[1,2], Rafael Rocha[1], Nikolas Carneiro[1], Renata Tedeschi[1], Ronnie Alves[1,3], Sergio Viademonte[1], Eduardo Carvalho[1,3]

[1]Instituto Tecnológico Vale, Belém PA, BRAZIL

[2]Federal University of Pará, Tucuruí PA, BRAZIL

[3]Federal University of Pará, Belém PA, BRAZIL

{helder.moreira,anita.drumond,ewerton.oliveira}@pq.itv.org

{julio.freitas,rafael.lima.rocha}@pq.itv.org

{nikolas.carneiro,renata.tedeschi,ronnie.alves}@itv.org

{sergio.viademonte,eduardo.costa.carvalho}@itv.org

Abstract. *This study evaluates precipitation forecasting at 12 monitoring points in southeastern Brazil using five statistical models (MLR, ARIMA, SARIMA, SARIMAX, and VARMAX). The forecast accuracy was assessed at two time points (M1 and M2) using RMSE. MLR performed best in the short term (M1) in half the locations (50%), while SARIMAX led in four (33%). In the long term (M2), VARMAX outperformed others in seven locations (58%), highlighting its strength in the capture of multivariate dynamics. The results underscore the value of statistical models for localized weather forecasting and infrastructure planning.*

Resumo. *Este estudo avalia a previsão de precipitação em 12 pontos de monitoramento no sudeste do Brasil, utilizando cinco modelos estatísticos (MLR, ARIMA, SARIMA, SARIMAX e VARMAX). A acurácia das previsões foi avaliada em dois momentos (M1 e M2), utilizando o RMSE. O modelo MLR apresentou o melhor desempenho no curto prazo (M1) em metade das localidades (50%), enquanto o SARIMAX foi superior em quatro delas (33%). No longo prazo (M2), o VARMAX superou os demais modelos em sete localidades (58%), evidenciando sua capacidade de capturar dinâmicas multivariadas. Os resultados destacam a importância dos modelos estatísticos para a previsão meteorológica local e o planejamento de infraestrutura.*

1. Introduction

Daily activities are intrinsically linked to meteorological conditions, influencing everything from routine commutes, such as traveling to work using public transportation, to more complex operations, such as scheduled maintenance in open-pit mining sites. Given the broad impact of atmospheric conditions on various social and economic spheres, weather forecasting has been the subject of numerous scientific studies

[Yu and Haskins 2021, Balaji et al. 2021]. Investigating the accuracy and methodologies employed in predicting precipitation events becomes highly relevant in this context.

Weather forecasting plays a strategic role in the mining sector, particularly in open-pit operations. Reliable meteorological information allows for better planning of maintenance activities, reduces loading and unloading times, and, most importantly, improves the safety of equipment operators and maintenance personnel. Anticipation of adverse weather events, such as thunderstorms and heavy rainfall, allows the preventive removal of workers from hazardous areas, thus contributing to the mitigation of accidents and operational continuity [Oliveira et al. 2023].

Among the various methods used for weather and climate forecasting, Multiple Linear Regression (MLR) stands out due to its simplicity and ability to model linear relationships between meteorological variables. Furthermore, traditional statistical techniques have prominence in the scientific literature [Kim et al. 2022], not only for their predictive modeling capabilities but also for providing a solid conceptual foundation for the application of more advanced approaches, such as machine learning algorithms. This methodological transition underscores the importance of statistics as a basis for developing more robust and adaptive predictive models.

Statistical models such as the Seasonal Autoregressive Integrated Moving Average (SARIMA) and its extension with exogenous variables, the Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors (SARIMAX), have proven effective in generating accurate forecasts across various domains. Their main advantage lies in their ability to capture recurring seasonal patterns, making them particularly suitable for regions or phenomena with well-defined seasonality. Under such conditions, these models tend to outperform approaches that disregard the seasonal structure of the data [Kim et al. 2019].

Historical data play a fundamental role in precipitation forecasting. They allow for the identification of recurring climatic patterns and the construction of more robust predictive models [Berrang-Ford et al. 2021, Garg and Mago 2021]. The availability of long and sufficiently variable time series enables the detection of seasonal and anomalous behaviors, thereby improving forecasts across different temporal scales. However, observational meteorological data often present gaps, inconsistencies, or measurement errors [Popp et al. 2020]. In this context, reanalysis datasets have gained prominence, offering continuous and spatially complete series, albeit partially based on estimates generated by numerical models and data assimilation systems [Hersbach et al. 2019]. Despite their limitations, reanalyses represent a valuable alternative for complementing or replacing missing observations, especially in regions with low station density.

Even with reanalysis data, several studies have explored their application in weather and climate forecasting. Weather forecasting focuses on short-term analysis, typically covering time scales from hours to a few days, aiming to describe the future state of the atmosphere based on current conditions [Hersbach et al. 2019]. In contrast, climate forecasting identifies variability patterns over longer periods, such as months, years, or decades. It is essential for understanding trends and climatic anomalies across different regions [Hayawi et al. 2025]. This methodological distinction is crucial for selecting appropriate models and data sources according to the temporal scope of the forecast.

In this context, the present study aims to forecast monthly precipitation, focusing on climate prediction in southeastern Brazil. The research seeks to address two central questions: (i) Is it possible to forecast precipitation using statistical methods and compare the results with those obtained through Multiple Linear Regression (MLR)?; and (ii) Which statistical methods perform best in the chosen region, which historically present greater complexity and uncertainty in meteorological forecasts?

This article is structured as follows: Section 2 presents a review of climatology in the studied region and statistical models; Section 3 describes the methodology for precipitation forecasting using statistical learning; Section 4 discusses the results obtained; and finally, Section 5 presents the conclusions and outlines future research directions.

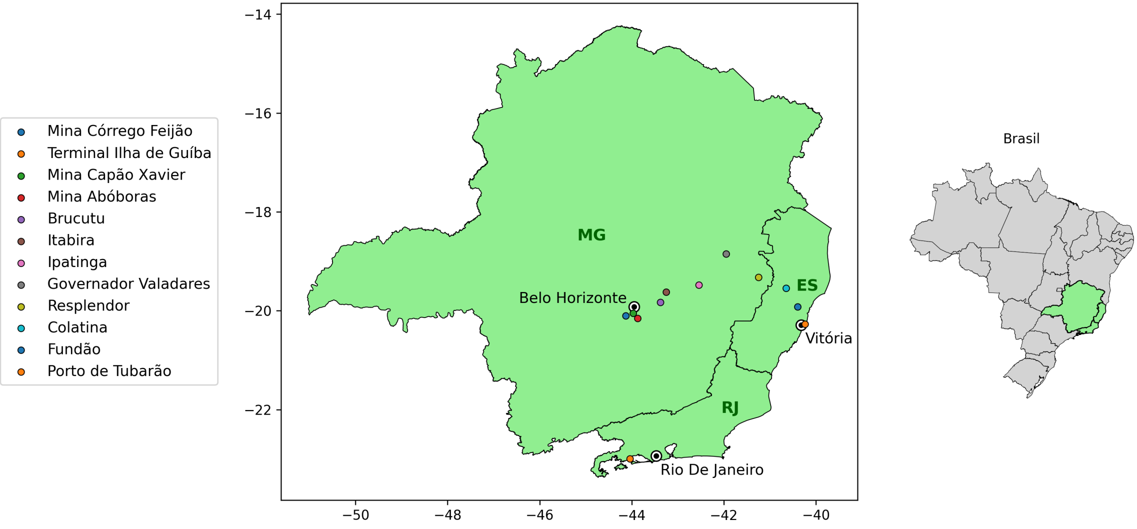
2. Climatology and statistical learning

This section provides a climatological overview of the Southeastern region of Brazil, emphasizing the meteorological characteristics that influence the operation of a major railway line used for transporting iron ore by a large mining company. The strategic importance of this logistical infrastructure, combined with the region's climatic variability, underscores the need for more accurate predictive studies. Additionally, this section introduces the fundamental concepts of statistical learning, outlining the algorithms employed in this study for precipitation forecasting.

2.1. Climatology in southeastern Brazil

The mining chain in Southeastern Brazil includes sites located near the Metropolitan Region of Belo Horizonte (MG). The southern mines are part of the so-called Southern Corridor, as shown in Figure 1, with production transported by rail to the Ilha de Guaíba terminal in the state of Rio de Janeiro. The northern mines belong to the Southeastern Corridor, with transportation carried out via the Vitória-Minas Railway (EFVM) to the Port of Tubarão, in the state of Espírito Santo [Silva Ferreira et al. 2021]. This logistics infrastructure is exposed to weather and climate variations, making precipitation forecasting a strategic element for planning and operations.

Figura 1. Locations of interest in the Southern/Southeast Corridor.

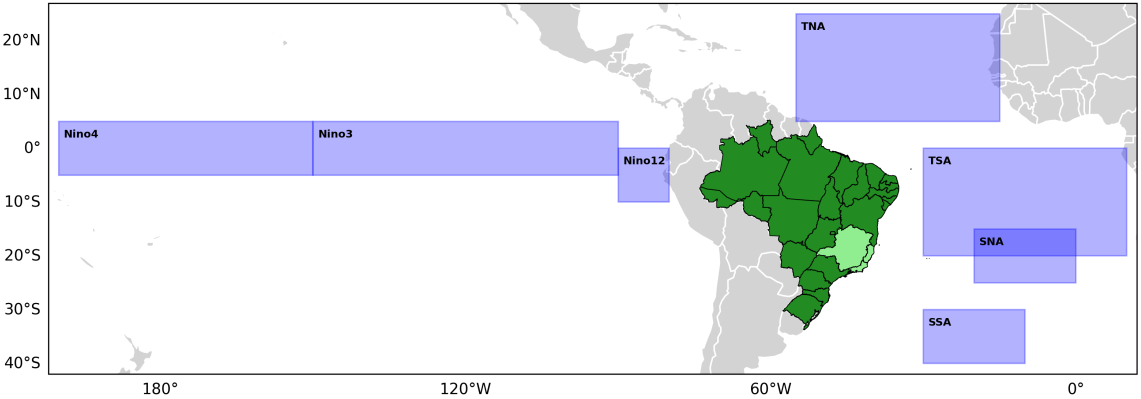


The Southeast region is home to many of the Brazilian population and the country's economic activities [Nunes et al. 2009]. It is located in a transitional area between tropical and subtropical climates, marked by climatic contrasts and influences from various geographic controls, such as proximity to the Atlantic Ocean and heterogeneous relief. In addition, changes in land use and increasing urbanisation promote anthropogenic changes in the regional climate [Nunes et al. 2009, Reboita et al. 2015].

Southeast Brazil is located under the influence of the South American Monsoon System (SAMS). This system contains rainy (October to March) and dry (April to September) seasons, which are well established. This region is impacted by atmospheric systems of different scales, from convection associated with local heating to transient frontal systems, easterly winds associated with the South Atlantic Subtropical High (SASH), and episodes of the South Atlantic Convergence Zone (SACZ). During the austral summer, these systems favour the occurrence of intense rainfall; in the austral winter, the weakening of convection and the displacement of the SASH over the continent reduce precipitation [Reboita et al. 2015].

In addition to seasonal variability, large-scale climate phenomena such as El Niño–Southern Oscillation (ENSO), Indian Ocean Dipole, South Atlantic Dipole, and Antarctic Oscillation influence the regional climate, modifying precipitation and temperature patterns [Carpenedo and da Silva 2022]. Events such as the 2002-2005 and 2014-2017 droughts and wet periods between 2007–2010 directly affected areas such as the Metropolitan Region of Belo Horizonte and the Rio Doce Valley [Petrucci et al. 2022]. The 2014 drought episode, for example, affected entire Southeastern Brazil, with socio-economic impacts and a water crisis [Coelho et al. 2016]. Some of these phenomena are shown in Figure 2.

Figura 2. Oceanic regions used in this work to predict rainfall.



Climate forecasting in the region is challenging. According to [Sampaio and Silva Dias 2015], the low predictability associated with numerical forecasting models is related to the great diversity and variability of meteorological systems operating in the region and the lower dependence on oceanic patterns. Given such complexity, techniques based on artificial intelligence emerge as promising for dealing with the region's climatic complexity, especially in forecasting transient meteorological systems, whose accurate representation is still limited within traditional models.

2.2. Statistical Learning for rainfall precipitation

Multiple Linear Regression (MLR) models the relationship between a continuous dependent variable and two or more independent variables, assuming a linear combination of predictors plus an error term [Kim et al. 2022]. Its strengths lie in simplicity and interpretability, with coefficients indicating the expected effect of each predictor while holding others constant. However, the method relies on key assumptions—such as the absence of multicollinearity, normally distributed residuals, and homoscedasticity — whose violation can compromise results and may necessitate model adjustments or alternatives.

The ARIMA model is a classical technique for univariate time series forecasting, combining autoregressive (AR), moving average (MA), and differencing (I) components to capture past dependencies and achieve stationarity [Yavuz 2025]. Effective modeling requires stationarity checks and careful selection of parameters (p, d, q). While ARIMA performs well for short-term forecasts in non-seasonal, single-variable series, it is less suitable for handling seasonality or multivariate data.

The SARIMA model extends ARIMA by adding seasonal components — captured through additional parameters (p, d, q, s), where s denotes the seasonal period—making it well-suited for time series with repeating patterns like monthly sales or annual temperatures [Yavuz 2025]. While effective for modeling complex temporal structures, SARIMA can be computationally intensive, and selecting appropriate parameters requires careful analysis. Still, it remains a widely used tool in meteorology, where seasonality is significant.

SARIMAX extends SARIMA by incorporating exogenous variables (external regressors that influence the time series but are not part of it) such as holidays, marketing actions, or economic indicators [Elshewey et al. 2022]. This addition enhances predictive accuracy in contexts where external factors play a significant role. However, it demands careful selection of relevant variables that must be available for the forecast horizon, adding complexity to model design and validation.

The VARMAX model generalizes ARMA/ARIMA to a multivariate context, allowing simultaneous modeling of multiple interdependent time series along with exogenous variables [Bowden and Clarke 2017]. Common in econometrics and finance, it captures complex dynamics among variables like inflation, interest rates, and GDP. While offering a comprehensive system view, VARMAX demands advanced expertise for proper implementation, particularly in ensuring stability and interpreting causal relationships.

The Student's t-test compares the means of two datasets to assess if their difference is statistically significant [Ruxton 2006]. In predictive modeling, the paired t-test suits comparisons like Root Mean Squared Error - RMSE (M1) vs. RMSE (M2) since values come from the same models under different conditions. It computes a t-statistic from paired differences and a p-value indicating the likelihood that the observed difference is due to chance. A p-value below a significance threshold (commonly 0.05) suggests that one method consistently outperforms the other. The test is favored for its simplicity and effectiveness, particularly with small to moderate samples [Ruxton 2006].

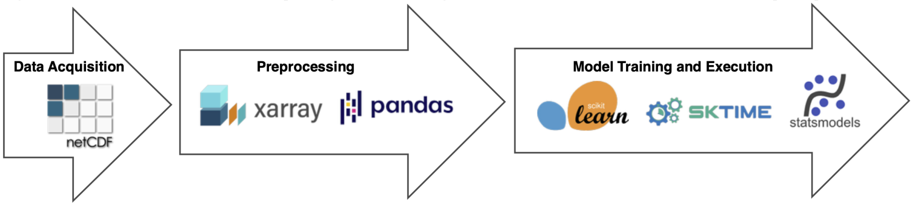
3. A method for predicting rainfall using statistical learning

For the precipitation forecast of the Southern/Southeastern Corridor, twelve interested locations were considered across the states of Minas Gerais (Córrego Feijão Mine - FEI, Capão Xavier Mine - CPX, Abóboras Mine - ABO, Brucutu - BRU, Itabira - ITA, Ipatinga - IPA, Governador Valadares - GOV, and Resplendor - RES), Espírito Santo (Colatina - COL, Fundão - FUN, and Tubarão Port - TUB), and Rio de Janeiro (Ilha de Guaíba Terminal - TIG) (Figure 1). Additionally, sea surface temperature from seven oceanic regions (Niño12, Niño3, Niño4, TNA, TSA, SNA, and SSA) (Figure 2) and zonal and meridional wind at the interested point were also taken into account.

3.1. Data manipulation

Data used in this study were obtained from the ERA5 database [Hersbach et al. 2020], provided by the European Centre for Medium-Range Weather Forecasts (ECMWF), in NetCDF (Network Common Data Form) format. The used variables include precipitation, sea surface temperature, and zonal and meridional wind (Data Acquisition) between 1940 and 2024. Data from 1940 to 2010 were used for model training, while between 2011 and 2024, they were used for testing and execution. In the preprocessing stage, the data were structured into time series corresponding to specific points of interest and oceanic regions, using the Xarray and Pandas libraries (Preprocessing). Finally, the model training and forecasting were conducted using the Scikit-Learn, SKtime, and StatsModels libraries (Model Training and Execution) (Figure 3).

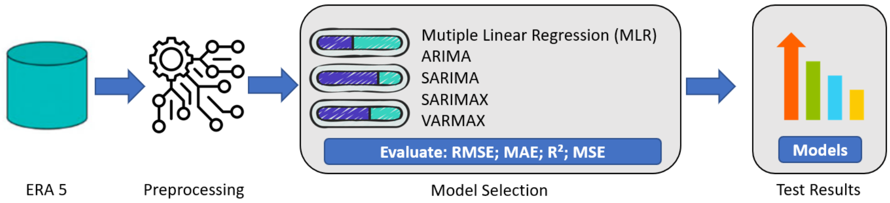
Figura 3. Data flux and programming structures used for data preparation.



3.2. Statistical learning algorithms

The MLR and SARIMAX models used two or three exogenous variables (sea surface temperature, zonal wind, and meridional wind) to predict precipitation, depending on the best correlation with the target point. The ARIMA, SARIMA, and VARMAX models used only precipitation data.

Figura 4. Methodology pipeline in predicting rainfall in Brazil's southeast.



The modeling process (Figure 4) begins with data preprocessing, which includes generating derived features and splitting the dataset into training and testing sets while preserving the temporal structure. The input data is sourced from ERA5, a state-of-the-art global atmospheric reanalysis produced by the ECMWF, which combines model data with observations to provide consistent and gap-free time series of multiple climate variables. Following this, a variety of models are considered for training, including regression techniques and time series models capable of capturing seasonal patterns and multivariate relationships. The goal is to explore different predictive strategies to identify the most suitable one for the data.

3.3. Evaluation methods

The R^2 is a statistical measure indicating the proportion of variance in the dependent variable explained by the regression model, with values ranging from 0 to 1—higher values suggesting greater explanatory power [Xu et al. 2022]. Despite its popularity, R^2 has limitations: it reflects trend alignment rather than absolute accuracy and can misleadingly increase with irrelevant variables. To address this, the adjusted R^2 is preferred in multivariable models, as it penalizes including non-contributory predictors [Xu et al. 2022].

The MAE measures the average absolute differences between predicted and observed values, offering an intuitive error metric expressed in the target variable's units [Chicco et al. 2021]. Its main strength lies in its robustness to outliers, as it avoids squaring errors, making it suitable when extreme deviations should not be overly penalized [Chicco et al. 2021]. However, since it ignores error direction, MAE is often used alongside other metrics for a more complete assessment of model performance.

The RMSE quantifies the average magnitude of prediction errors, emphasizing larger deviations by squaring them before averaging [Hodson 2022]. This makes it especially suitable for contexts where large errors are critical, such as control systems or demand forecasting. However, its sensitivity to outliers can also distort evaluations in data with high variability or noise, which is why RMSE is often analyzed alongside metrics like MAE for a more balanced assessment.

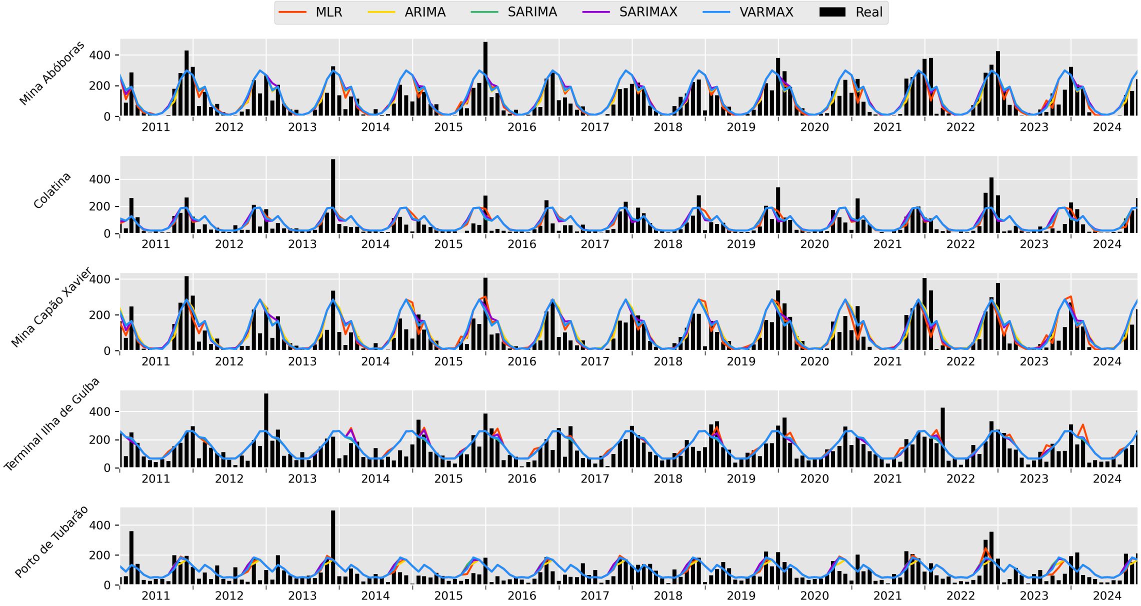
The avRMSE (Average RMSE) represents the mean of RMSE values across multiple runs, models, or data subsets, offering a more stable estimate of model performance in scenarios like cross-validation or multi-dataset testing [Bergmeir et al. 2018]. Average results mitigate the influence of run-specific variability, provided the individual RMSEs are computed under consistent conditions and scales. This makes avRMSE particularly valuable in predictive modeling experiments that prioritize performance consistency alongside accuracy [Vien et al. 2021].

4. Results

Figure 5 presents a time series of monthly precipitation from 2011 to 2024 for 5 points of interest in M1. The graph combines observed data and forecasts generated by statistical models. The observed data represent historical variability and seasonal precipitation patterns at the studied locations, while the forecasts correspond to values estimated by the models over the same period. The objective is to provide a visual comparison between observed data and model predictions, allowing an intuitive assessment of the performance of each method across different locations.

Table 1 presents the performance metrics for five statistical forecasting models — MLR, ARIMA, SARIMA, SARIMAX, and VARMAX — applied to monthly precipita-

Figura 5. Real data and statistical methods predictions



tion prediction across the 5 points of interest. The evaluation was conducted using two datasets, referred to as M1 and M2, representing different temporal partitions of the data from the given month and the following one. M1 and M2 correspond to data from current and subsequent month, respectively. For each location, models were assessed based on four metrics: MAE, MSE, RMSE, and R². RMSE was used as the primary metric for comparison due to its sensitivity to large errors and its interpretability in the same units as precipitation. The results are organized by location and model, allowing a detailed comparison of model performance both spatially (across different sites) and methodologically (across models).

The analysis of RMSE values across 12 geographical points at two time periods (M1 and M2) reveals clear differences in model performance over time. At M1, the MLR model performed best in 6 out of 12 points (IPA, GOV, RES, COL, FUN, and TUB). This suggests that, in the short term, linear relationships among variables were sufficient to generate reliable precipitation forecasts in many locations within the tropical region studied. Additionally, the SARIMAX model achieved the lowest RMSE in 4 points (FEI, CPX, BRU, and ITA), demonstrating the value of incorporating exogenous variables and seasonal patterns. ARIMA and VARMAX each had the best performance in only one point at this initial stage (TIG and ABO, respectively).

However, a marked shift occurs at M2. The VARMAX model emerged as the best performer in 7 of the 12 points (ABO, CPX, BRU, RES, COL, FUN, and TUB), indicating its superior capacity to capture complex, multivariate temporal dynamics in longer-term forecasts. SARIMA followed with 3 points (ITA, IPA, and GOV), showing consistent performance in regions where seasonal structure is dominant. ARIMA maintained relevance in 2 points (FEI and TIG), while MLR was no longer the best-performing model in any location.

This transition from the dominance of simpler models (like MLR) at M1 to the superiority of more sophisticated, multivariate models (especially VARMAX) at M2 high-

Tabela 1. Statistical learning metrics for 5 points of interest in the Southeastern Brazil, with lag 1 (M1) and lag 2 (M2) for rainfall prediction.

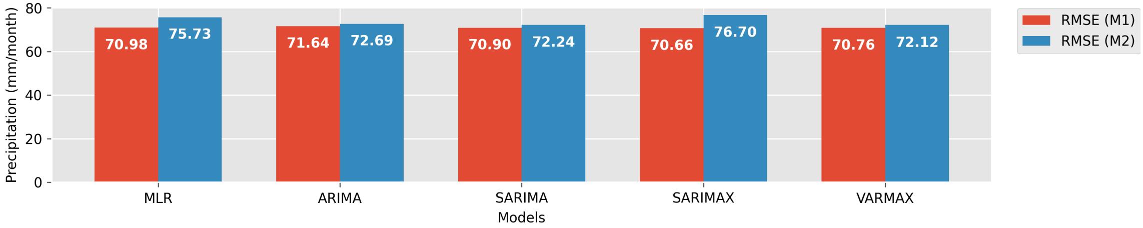
Point	Model	MAE (M1)	MSE (M1)	RMSE (M1)	R² (M1)	MAE (M2)	MSE (M2)	RMSE (M2)	R² (M2)
ABO	MLR	46.168	4634.337	68.076	0.595	48.990	5552.324	74.514	0.516
	ARIMA	45.959	4457.486	66.764	0.610	47.710	4767.089	69.044	0.585
	SARIMA	45.758	4412.438	66.426	0.614	47.671	4770.872	69.072	0.584
	SARIMAX	45.280	4432.171	66.575	0.612	49.559	5938.328	77.061	0.483
	VARMAX	45.609	4410.366	**66.411**	0.614	47.615	4745.104	**68.885**	0.587
COL	MLR	43.057	4277.743	**65.404**	0.434	43.962	4675.710	68.379	0.384
	ARIMA	43.495	4410.321	66.410	0.416	43.548	4468.262	66.845	0.411
	SARIMA	43.882	4480.667	66.938	0.407	43.796	4525.571	67.272	0.403
	SARIMAX	44.048	4500.540	67.086	0.405	45.388	4884.161	69.887	0.356
	VARMAX	43.487	4413.912	66.437	0.416	43.529	4464.227	**66.815**	0.411
CPX	MLR	45.610	4869.403	69.781	0.505	48.700	5550.513	74.502	0.437
	ARIMA	46.192	4512.365	67.174	0.541	46.526	4768.178	69.052	0.516
	SARIMA	45.384	4420.873	66.490	0.551	46.828	4719.249	68.697	0.521
	SARIMAX	44.737	4384.542	**66.216**	0.554	48.995	5807.255	76.205	0.411
	VARMAX	45.148	4413.755	66.436	0.551	46.782	4708.993	**68.622**	0.522
TIG	MLR	47.885	4570.106	67.603	0.445	47.884	4733.708	68.802	0.429
	ARIMA	45.529	4243.109	**65.139**	0.485	47.355	4641.584	**68.129**	0.440
	SARIMA	45.400	4255.245	65.232	0.484	47.278	4652.730	68.211	0.438
	SARIMAX	45.904	4276.414	65.394	0.481	47.496	4668.750	68.328	0.436
	VARMAX	45.566	4247.397	65.172	0.484	47.395	4644.255	68.149	0.439
TUB	MLR	44.767	3712.877	**60.933**	0.310	45.613	4182.193	64.670	0.228
	ARIMA	47.115	3975.492	63.052	0.261	46.080	3995.343	63.209	0.263
	SARIMA	46.853	4021.573	63.416	0.253	46.551	4051.429	63.651	0.253
	SARIMAX	46.578	3987.451	63.146	0.259	47.479	4202.498	64.827	0.225
	VARMAX	46.026	3969.912	63.007	0.262	45.425	3958.054	**62.913**	0.270

lights the importance of selecting models based on forecast horizon and climate complexity. In tropical regions—where precipitation is influenced by nonlinear and interdependent processes—models like VARMAX that account for multiple variables and their interactions tend to offer better long-term performance.

Moreover, spatial variability plays an important role in forecasting performance. Stations such as TUB, COL, and ABO consistently showed low RMSEs in both M1 and M2, suggesting they are more predictable. Conversely, stations like ITA, IPA, and GOV consistently exhibited high RMSEs, pointing to greater forecast uncertainty—possibly linked to local geographic or atmospheric variability. These differences emphasize that model performance is not only tied to the method used but also to the intrinsic characteristics of each location.

Figure 6 chart compares average RMSE values between two lags, M1 and M2, across five forecasting models: MLR, ARIMA, SARIMA, SARIMAX, and VARMAX. Each pair of bars represents the RMSE performance of a model under the two configurations.

Figura 6. Root Mean Square Error comparison between M1 and M2.



From the visualization, it is evident that M2 consistently yields higher RMSE values than M1 across all models, indicating a generally lower predictive accuracy. This pattern suggests that the M1 configuration may be more effective in minimizing prediction errors. The difference is particularly notable in models like SARIMAX and Linear Regression, where the gap between M1 and M2 is more pronounced.

Based on the statistical analysis performed using the paired Student's t-test, a statistically significant difference was identified between the mean RMSE (Root Mean Square Error) values of models M1 and M2. The test yielded a t-value of -2.8040 and a p-value of 0.0486, indicating that, at a 5% significance level, the performance of the models in terms of mean squared error is not equivalent. This suggests that the observed differences between models M1 and M2 are not due to random variation, highlighting the importance of considering this difference when selecting the most appropriate model.

5. Conclusion and future works

This study used reanalysis data to address the relevance of precipitation forecasting at 12 collection points. This data type is justified by its completeness and spatial-temporal availability, even though it represents a modeled estimate of observational reality. Currently, a MLR model is employed to forecast precipitation with two temporal lags, referred to as M1 and M2.

The main objective was to compare this model's performance against other statistical methods, as described in Section 2.2. In this context, two guiding questions were investigated: (i) the feasibility of comparing the models, with the results summarized in Table 1, and (ii) the identification of the best-performing algorithms among those tested, considering the 12 points distributed along a railway network.

(i) Can precipitation be forecasted using statistical methods, and can the results be compared with multiple linear regression (MLR)? Yes. The study demonstrates that several statistical models — including ARIMA, SARIMA, SARIMAX, and VARMAX — can effectively forecast precipitation and, in many cases, outperform MLR. While MLR showed strong performance at the initial time point (M1), its effectiveness declined in the subsequent period (M2), when it was not the best model in any location.

(ii) Which statistical methods perform best in the studied region, which historically presents greater complexity and uncertainty in meteorological forecasts? The results indicate that VARMAX is the most robust model for precipitation forecasting in the studied region, especially in longer-term scenarios. SARIMA also demonstrated good performance, particularly in areas with strong seasonal patterns. Simpler models like MLR may still be helpful in short-term applications, but are limited when addressing the nonlinear and dynamic nature of precipitation systems.

In summary, the study reinforces the importance of using advanced multivariate time series models, such as VARMAX, to improve forecast accuracy in regions characterized by high variability and forecasting uncertainty.

For future work, the inclusion of machine learning models such as decision trees and neural networks is planned, as well as variation in the hyperparameters of the methods using cross-validation. Additionally, the exploration of other exogenous variables capable of aiding the precipitation forecasting process will be considered.

Referências

Balaji, T. K., Annavarapu, C. S. R., and Bablani, A. (2021). Machine learning algorithms for social media analysis: A survey. *Computer Science Review*, 40:100395.

Bergmeir, C., Hyndman, R. J., and Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83.

Berrang-Ford, L., Sietsma, A. J., Callaghan, M., Minx, J. C., Scheelbeek, P. F., Haddaway, N. R., Haines, A., and Dangour, A. D. (2021). Systematic mapping of global research on climate and health: a machine learning review. *The Lancet Planetary Health*, 5(8):e514–e525.

Bowden, R. S. and Clarke, B. R. (2017). Using multivariate time series methods to estimate location and climate change effects on temperature readings employed in electricity demand simulation. *Australian & New Zealand Journal of Statistics*, 59(4):413–431.

Carpenedo, C. B. and da Silva, C. B. (2022). Influência de teleconexões na precipitação pluvial do cerrado brasileiro. *Revista Brasileira de Climatologia*, 30(18):26–46.

Chicco, D., Warrens, M. J., and Jurman, G. (2021). The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *Peerj computer science*, 7:e623.

Coelho, C. A. S., de Oliveira, C. P., Ambrizzi, T., Reboita, M. S., Carpenedo, C. B., Campos, J. L. P. S., Tomaziello, A. C. N., Pampuch, L. A., Custódio, M. d. S., Dutra, L. M. M., Da Rocha, R. P., and Rehbein, A. (2016). The 2014 southeast brazil austral summer drought: regional scale mechanisms and teleconnections. *Climate Dynamics*, 46:3737–3752.

Elshewey, A. M., Shams, M. Y., Elhady, A. M., Shohieb, S. M., Abdelhamid, A. A., Ibrahim, A., and Tarek, Z. (2022). A novel wd-sarimax model for temperature forecasting using daily delhi climate dataset. *Sustainability*, 15(1):757.

Garg, A. and Mago, V. (2021). Role of machine learning in medical research: A survey. *Computer science review*, 40:100370.

Hayawi, K., Shahriar, S., and Hacid, H. (2025). Climate data imputation and quality improvement using satellite data. *Journal of Data Science and Intelligent Systems*, 3(2):87–97.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Dee, D., Horányi, A., Nicolas, J., Peubey, C., Radu, R., Rozum, I., et al. (2019). The era5 global atmospheric reanalysis at ecmwf as a comprehensive dataset for climate data homogenization, climate variability, trends and extremes. In *Geophysical Research Abstracts*, volume 21.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al. (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049.

Hodson, T. O. (2022). Root mean square error (rmse) or mean absolute error (mae): When to use them or not. *Geoscientific Model Development Discussions*, 2022:1–10.

Kim, S.-J., Bae, S.-J., and Jang, M.-W. (2022). Linear regression machine learning algorithms for estimating reference evapotranspiration using limited climate data. *Sustainability*, 14(18):11674.

Kim, T., Shin, J.-Y., Kim, H., Kim, S., and Heo, J.-H. (2019). The use of large-scale climate indices in monthly reservoir inflow forecasting and its application on time series and artificial intelligence models. *Water*, 11(2):374.

Nunes, L. H., Vicente, A. K., and Candido, D. H. (2009). Clima da região sudeste do brasil. In *Tempo e clima no Brasil*, pages 243–258. Oficina de Textos, Sao Paulo.

Oliveira, E. C. L. d., Nogueira Neto, A. V., Santos, A. P. P. d., da Costa, C. P. W., Freitas, J. C. G. d., Souza-Filho, P. W. M., Rocha, R. d. L., Alves, R. C., Franco, V. d. S., Carvalho, E. C. d., et al. (2023). Precipitation forecasting: from geophysical aspects to machine learning applications. *Frontiers in Climate*, 5:1250201.

Petrucci, E., Oliveira, L. A., and Silva, R. C. (2022). Secas pluviométricas no estado de minas gerais, de 1980 a 2017. *Raega*, 54:129–153.

Popp, T., Hegglin, M. I., Hollmann, R., Ardhuin, F., Bartsch, A., Bastos, A., Bennett, V., Boutin, J., Brockmann, C., Buchwitz, M., et al. (2020). Consistency of satellite climate data records for earth system monitoring. *Bulletin of the American Meteorological Society*, 101(11):E1948–E1971.

Reboita, M. S., Rodrigues, M., Silva, L. F., and Alves, M. A. (2015). Aspectos climáticos do estado de minas gerais. *Revista brasileira de Climatologia*, 17.

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test. *Behavioral Ecology*, 17(4):688–690.

Sampaio, G. and Silva Dias, P. L. (2015). Evolução dos modelos climáticos e de previsão de tempo e clima. *Revista USP*.

Silva Ferreira, D. B., Kuhn, P. A. F., Silva, F. d. O., Costa, C. P. W., Tedeschi, R. G., and Santos, A. P. P. (2021). Sistema de previsões meteorológicas para corredores sul-sudeste da vale. Technical report, Instituto Tecnológico Vale.

Vien, B. S., Wong, L. D. Z., Kuen, T., Rose, L. R. F., and Chiu, W. K. (2021). A Machine Learning Approach for Anaerobic Reactor Performance Prediction Using Long Short-Term Memory Recurrent Neural Network. In *8th Asia Pacific Workshop on Structural Health Monitoring*, pages 61–70.

Xu, X., Du, H., and Lian, Z. (2022). Discussion on regression analysis with small determination coefficient in human-environment researches. *Indoor air*, 32(10):e13117.

Yavuz, V. S. (2025). Forecasting monthly rainfall and temperature patterns in van province, türkiye, using arima and sarima models: a long-term climate analysis. *Journal of Water and Climate Change*, 16(2):800–818.

Yu, N. and Haskins, T. (2021). Bagging machine learning algorithms: A generic computing framework based on machine-learning methods for regional rainfall forecasting in upstate new york. In *Informatics*, volume 8, page 47. MDPI.