







Artigo

Refining Seasonal Precipitation Forecast in Brazil Using Simple Data-Driven Techniques and Climate Indices

Francisca Lanai Ribeiro Torres¹ , Cassia Akemi Castro Kuki¹ , Michelle Simões Reboita² ,
Luana Medeiros Marangon Lima³ , José Wanderley Marangon Lima⁴ ,
Anderson Rodrigo de Queiroz^{5,6} 

¹*Instituto de Sistemas Elétricos e Energia, Universidade Federal de Itajubá, Itajubá, MG, Brasil.*

²*Instituto de Recursos Naturais, Universidade Federal de Itajubá, Itajubá, MG, Brasil.*

³*Nicholas School of Environment, Duke University, Durham, NC, United States of America.*

⁴*Marangon Consulting and Engineering, Itajubá, MG, Brazil.*

⁵*Civil, Construction, and Environmental Engineering Department, North Carolina State University, Raleigh, NC, United States of America.*

⁶*Operations Research Graduate Program, North Carolina State University, Raleigh, NC, United States of America.*

Received: 18 November 2023 - Accepted: 3 May 2024

Abstract

Seasonal precipitation forecasts are essential for water resource management, agricultural activities, and the operational planning of hydropower systems. Any methodological advancement that enhances the accuracy of precipitation predictions will yield considerable societal benefits. In this context, this study proposes and evaluates two approaches for refining seasonal precipitation forecasts in Brazil, using simple data-based models, such as multiple linear regression (MLR) and nonlinear support vector machine (SVM). These models employ climate indices related to different teleconnection patterns that affect seasonal precipitation in Brazil, the unified gauge-based analysis of global daily precipitation from the Climate Prediction Center (CPC), and the precipitation forecasts from the Seasonal Forecast System 5 (SEAS5) as input variables. Both MLR and SVM models were validated from Jan-2017 to Dec-2020 using precipitation from the CPC as ground truth. The results suggest that, compared to SEAS5, MLR and SVM models enhance predictive accuracy and reduce bias in precipitation forecasts for the Southeast, Midwest, and North regions of Brazil during the austral summer. However, the performance of the models was found to be on par with the original predictions of SEAS5 in the Northeast and South regions, sectors of Brazil where the climate is significantly influenced by the El Niño-Southern Oscillation.

Keywords: data-driven models, SEAS5, seasonal precipitation forecast, teleconnection patterns, time series forecasting.

Aprimorando a Previsão de Precipitação Sazonal no Brasil Usando Técnicas Simples Baseadas em Dados e Índices Climáticos

Resumo

As previsões sazonais de precipitação são essenciais para a gestão dos recursos hídricos, a atividade agrícola e o planejamento operacional de sistemas hidroelétricos. Avanços metodológicos capazes de aprimorar a acurácia das previsões de precipitação geram benefícios significativos para a sociedade. Diante desse contexto, este estudo propõe e avalia duas abordagens para o refinamento das previsões sazonais de precipitação no Brasil, utilizando modelos matemáticos simples e baseados em dados, como a regressão linear múltipla (MLR) e a máquina de vetores de suporte não linear (SVM). Nas abordagens propostas, os modelos MLR e SVM são alimentados com índices climáticos relacionados a diferentes padrões de teleconexão que afetam a precipitação sazonal no Brasil, dados da análise de precipitação diária

global do Climate Prediction Center (CPC) e previsões de precipitação do Seasonal Forecast System 5 (SEAS5). Ambos os modelos MLR e SVM foram validados de janeiro de 2017 a dezembro de 2020, usando a precipitação do CPC como referência. Os resultados sugerem que, em comparação com o SEAS5, os modelos MLR e SVM melhoram a precisão e reduzem o viés nas previsões de precipitação para as regiões Sudeste, Centro-Oeste e Norte do Brasil durante o verão austral. No entanto, o desempenho dos modelos em consideração foi equivalente ao do SEAS5 nas regiões Nordeste e Sul, setores do Brasil onde o clima é significativamente influenciado pelo El Niño-Oscilação Sul.

Palavras-chave: modelos baseados em dados, SEAS5, previsão de precipitação sazonal, padrões de teleconexão, previsão de séries temporais.

1. Introduction

Weather and climate are natural factors that influence a variety of human activities, including agriculture, river transport, and renewable energy production (Smith, 1993; Schweighofer, 2014; Chavez *et al.*, 2015; Shannon and Motha, 2015; Sivakumar, 2015). In Brazil, the electrical matrix of the available sources for electricity generation is primarily hydraulic. Power generation is heavily reliant on streamflows, which are significantly affected by variations in precipitation (Fan *et al.*, 2014; Pontes *et al.*, 2013; Dias *et al.*, 2017). Therefore, accurate precipitation forecasting is crucial for hydropower generation, management of available water resources (Ali *et al.*, 2020; Carbone, 2005), agricultural activities (Lipper *et al.*, 2014; Ingram *et al.*, 2002), and flood and drought predictions (Brunner *et al.*, 2021). Fundamentally, precipitation forecasting can be classified as a weather or a climate forecasting problem depending on the forecast horizon considered. Weather forecasts aim to predict the spatio-temporal progression and the impacts of atmospheric systems, such as the resultant effects triggered by the incursion of a cold front. In contrast, climate forecasts focus on simulating the climate system over extended timeframes to provide a comprehensive outlook of climate variables (Toth and Buizza, 2019), which helps predict whether upcoming seasons will be rainier or drier, warmer or cooler compared to the climatological normals (i.e., long-term average values of climate variables).

Deviations from the usual seasonal patterns occur due to changes in the basic state of the atmosphere. These changes modulate the intensity and frequency of the atmospheric systems. For a clearer understanding, imagine anomalies in sea surface temperature (SST) causing disturbances in the atmosphere. When the basic state of the atmosphere is disturbed, it creates waves that propagate and influence the location, intensity, and frequency of atmospheric systems in distant regions, thereby affecting climate variables. The link between events occurring in a particular part of the globe and the subsequent changes they induce in remote regions' climates is termed teleconnection patterns (Liu and Alexander, 2007; Reboita *et al.*, 2021; Grimm and Dias, 1995). In the case of Brazil, several studies (Reboita *et al.*, 2021; Goddard *et al.*, 2001) indicate that SST anomalies strongly affect the climate of the North, north coast of the Northeast, and the South

regions of Brazil, while the remaining regions (Midwest and Southeast) show a weaker response to SST anomalies. Therefore, monitoring SST anomalies is essential for climate prediction in Brazil because they drive climate variability on a seasonal scale through teleconnection patterns (Nobre and Shukla, 1996).

Climate forecasts are typically generated using three methods (Goddard *et al.*, 2001): (a) meteorologists study current oceanic conditions, identify the active climatic drivers, and extrapolate these slow-changing forcings into the future to evaluate how they could potentially impact the climate; (b) applying numerical dynamical models, which are physically based models made up of equations that represent the physical processes found in nature. This modelling approach is employed to forecast the evolution of intricate dynamic systems over time, given certain initial conditions and assumptions, and (c) using data-driven models (also known as statistical models), that are designed employing explanatory variables and machine learning algorithms to forecast, for example, precipitation or surface air temperature. In this case, incorporating SST data from pertinent oceanic regions, such as the Central Pacific Ocean, is a common practice, given the significant influence of this explanatory variable over remote regions (Folland *et al.*, 2001; Diro *et al.*, 2008; Zeng *et al.*, 2011; Córdoba-Machado *et al.*, 2015; Choubin *et al.*, 2018; Pezzi *et al.*, 2000).

Despite the advanced methodologies employed to create climate forecasts, including both numerical and data-driven-based ones, it is important to recognize that the intricate nature of the climate system introduces a layer of complexity in the modeling process that cannot be ignored. Uncertainty in seasonal climate forecasts stems from the complex and chaotic nature of the climate system, influenced by numerous interacting factors, including atmospheric conditions, ocean currents, and human activities (Palmer, 2000). This complexity means that, even with advanced modeling and data acquisition techniques, forecasting seasonal precipitation remains a significant challenge. Among the many sources of uncertainty impacting these forecasts, two types receive greater attention: a) uncertainties of the initial conditions, associated with errors or inaccuracies in the input data, potentially leading to less reliable forecasts (Slingo and Palmer, 2011); and b) uncertainties related to the models' structure, due to limitations in modelling the climate system, such as the

imperfect representation of physical processes and interactions between the system's components (Palmer *et al.*, 2005). For a), the most common approach to representing uncertainties is the introduction of perturbations to the model's initial conditions, generating multiple scenarios (Tracton and Kalnay, 1993; Molteni *et al.*, 1996). This approach facilitates exploring the sensitivity of climate forecasts to different initial conditions, providing a method to quantify uncertainty and enhance understanding of the potential range of climate evolutions. In the case of b), ongoing research is focused on refining the representations of physical processes and interactions. In this context, refined climate models with diversified structures compose multimodel ensembles. This strategy, involving the exploration of outcomes from multiple models, aims to address the structural uncertainties of the forecasting process and assist in identifying more consistent and robust forecasts across different models (Hagedorn *et al.*, 2005).

1.1. Data-driven models in seasonal precipitation forecast

The advent of machine learning techniques has played a substantial role in the evolution of climate prediction over the years. Such algorithms can handle vast amounts of data and recognize complex and nonlinear patterns that are beyond human analysis or conventional computational models. They can learn from the data and improve their predictions over time, without explicit programming. In the specialized literature on data-driven models, many studies have explored the application of machine learning techniques to predict precipitation for months ahead. For example, Quan *et al.* (2006) analyzed the predictive performance and identified the sources of the forecasting skill of four atmospheric general circulation models (AGCMs) alongside a collection of forced linear regressions (using SST data from 1950 to 1999). The results indicate that the ability of AGCMs to forecast U.S. precipitation and surface air temperature largely depends on the linear atmospheric signal of the El Niño-Southern Oscillation (ENSO), a performance that is on par with the regression models. Regardless of this outcome, the authors believe we still need dynamic models to make enhanced seasonal forecasts, given that statistical methods, even those trained on 50 years of data, can experience notable performance fluctuations over decades. Later, and following a similar line of research, Diro *et al.* (2008) designed statistical models to forecast seasonal rainfall within eight homogeneous precipitation zones in Ethiopia. The authors used data from both tropical and extratropical SSTs around the world as predictors and implemented techniques such as multiple linear regression (MLR) and linear discriminant analysis. The statistical models exhibited enhanced accuracy during years of extreme precipitation, illustrating their effectiveness in predicting anomalously high and low rainfall levels.

Still in 2006, Coelho *et al.* introduced an approach aimed at enhancing seasonal rainfall predictions in South America through an integrated forecasting system. This system merges two distinct forecasting strategies: an empirical model (data-driven multivariate linear regression) using sea surface temperature anomalies from the Pacific and Atlantic Oceans, and a multi-model ensemble incorporating European models from 3 climate centers (Centre National de Recherches Météorologiques - CNRM, European Centre for Medium-Range Weather Forecasts - ECMWF, and United Kingdom Met Office - UKMO) that simulate both oceanic and atmospheric conditions. By applying Bayesian statistical techniques, this method refines the forecasts, particularly for the austral summer, yielding more reliable predictions. The findings indicate that this integrated approach significantly improves the accuracy of rainfall predictions across the Tropics and in specific regions of southern Brazil, Uruguay, Paraguay, and northern Argentina, especially during El Niño or La Niña events, although its accuracy diminishes in their absence.

In 2014, Badr *et al.* explored the use of artificial neural networks (ANN) to predict precipitation anomalies in Africa's Sahel region using SST and surface air temperature anomalies as predictors. The study emphasizes the superior accuracy of ANN algorithms compared to other statistical models, attributing their effectiveness to their ability to encapsulate the nonlinear influences that large-scale climate forcings exert on precipitation. Later, in 2016, Gerlitz *et al.* proposed a data-driven model based on the random forest algorithm to forecast seasonal precipitation anomalies in Central and Southern Asia. A correlation analysis study conducted to select predictors revealed a strong influence of ENSO on precipitation in both regions, with the central region additionally significantly impacted by the North Atlantic Oscillation (NAO) and East Atlantic (EA) patterns. The random forest model effectively forecasted wet conditions and moderate droughts in Central Asia; however, the prediction of severe dry spells proved to be a challenge yet to be overcome.

Dabernig *et al.* (2017) introduced a novel approach to improve weather prediction titled "Spatial Ensemble Post-Processing with Standardized Anomalies". This method improves the accuracy of weather predictions, specifically for temperature 2 meters above ground, by applying ensemble post-processing techniques across spatial domains instead of individual locations. By standardizing anomalies, this technique accounts for and removes seasonal and location-specific characteristics through climatological adjustments. This enables more precise forecasting in areas with limited observational data and often outperforms traditional methods. The approach uses non-homogeneous Gaussian regression models, modified to work with standardized anomalies, addressing challenges

related to spatial coherence and computational efficiency. The methodology's effectiveness is particularly highlighted in complex terrains, such as the Alps, where it significantly improves the forecast accuracy.

Xu *et al.* (2020) compared the performance of numerical dynamical and statistical models in precipitation forecasting (including linear regression, long short-term memory - LSTM neural networks, SVMs, and random forests, as well as the climate models from the North American Multi-Model Ensemble - NMME). The predictors considered in the data-driven modeling process included climate indices and wavelet-decomposed and non-decomposed historical precipitation data, along with mean, minimum, and maximum surface air temperatures. Wavelet-based models showed superior performance compared with other data-driven models and NMME. This result suggests the presence of nonlinear effects in precipitation that are revealed through wavelet decomposition, which in turn enable data-driven models to enhance their performances. Later, and following a similar line of research, Anochi *et al.* (2021) created and assessed a self-organized multilayer perceptron ANN for precipitation forecast in South America, benchmarking its performance against the Brazilian Global Atmospheric Model (BAM). The neural network model demonstrated superior performance over BAM in most regions, particularly reducing the forecast error from 8 mm to 2 mm in the central region during winter. However, larger errors were observed during the austral summer (rainy season). This was attributed to local processes and the abundant energy of this season, which pose a challenge for neural networks due to the spatiotemporal resolution limitation of the training data.

Gibson *et al.* (2021) presented a study on enhancing seasonal precipitation forecasts through the application of machine learning models trained on climate model outputs. Focusing on the Western United States, a region characterized by its challenging forecasting conditions due to low precipitation totals and high variability, the study leverages the potential of machine learning to interpret complex interactions among various sources of seasonal predictability, such as ENSO, tropical diabatic heating anomalies, and jet stream variability. The study's results demonstrate the skillfulness of machine learning models, such as Random Forest and Neural Networks, in forecasting seasonal precipitation with greater accuracy than traditional methods. Furthermore, the study advances interpretability in machine learning through variable importance analysis, revealing key predictors and offering insights into the physical processes driving seasonal precipitation. This research marks a step forward in seasonal forecasting, suggesting that machine learning models, especially when coupled with large climate model ensembles, can provide more accurate and interpretable predictions, thereby offering valuable tools for managing climate and weather risks.

Also in 2021, Wu *et al.* presented a novel hybrid model for predicting monthly precipitation in northeastern China. The authors emphasized that due to the nonlinear, stochastic, and highly complex nature of precipitation, accurate precipitation forecasting remains a major challenge. Models based on the autoregressive integrated moving average (ARIMA) and ANNs, which are commonly employed to forecast precipitation, have particular limitations. ARIMA lacks the capacity to emulate the nonlinear structure inherent to precipitation, whereas ANNs operate under the assumption of independence between the input and output variables. As a solution, the author proposed a combination of the aforementioned models with wavelet-based multiresolution analyses to form a more robust hybrid model. The results revealed that the proposed hybrid model outperformed ARIMA and LSTM in predicting the monthly precipitation. This enhanced performance may be ascribed to the integration of the strengths inherent in each model within the ensemble, thereby yielding forecasts that are more comprehensive and robust.

More recently, Pinheiro *et al.* (2023) conducted a study on the utilization of an ensemble of artificial neural networks (EANN) for short lead seasonal precipitation forecasting in Ceará, northeastern Brazil, for the February-April period. This research aimed to assess the forecasting skill of EANN using indices of low-frequency climate oscillations and to explore the integration of EANN with dynamical models into a hybrid multi-model ensemble (MME). Through leave-one-out cross-validation over four decades of data, the study compared the performance of EANN against traditional statistical models and dynamical models currently used in Ceará's seasonal forecasting system. The findings revealed that EANN outperformed in both deterministic and probabilistic forecasting skills, showcasing lower root mean squared error and ranked probability score across most regions of Ceará. Additionally, EANN demonstrated better calibration and resolution for predicting above-normal and below-normal rainfall categories compared to its counterparts. The integration of EANN with dynamical models into a hybrid MME resulted in improved reliability by reducing overconfidence in extreme forecasts. This study underscores the potential of EANN in enhancing seasonal rainfall forecasts in regions like Ceará, leveraging low-frequency climate oscillations, and highlights the advantages of adopting hybrid modeling approaches for refining forecast accuracy and reliability amidst the challenges posed by the region's climate variability and forecasting intricacies.

1.2. Purpose of this study

Especially for Brazil, the refinement of seasonal precipitation forecasts using data-driven models is a relevant research topic that holds national significance, given the

critical role of water resources in power generation and agricultural operations. Most literature is limited to specific sectors or does not detail the Brazilian geographical regions (Li *et al.*, 1996; Ward and Folland, 1991; Paz *et al.*, 2010; Anochi and Velho, 2016; Milléo and Almeida, 2021), and only some new comprehensive studies have considered the entire country (Anochi *et al.*, 2021; Monogo *et al.*, 2022). Our study fits precisely into this line of research. Our primary goal is to refine seasonal precipitation forecasting by using simple, yet effective, data-driven models and predictors with relevant information for climate forecasting. For this, we propose a new framework based on MLR and support vector machine (SVM) that explores a) climate indices related to teleconnection patterns affecting Brazil, b) the statistical relationship between future and past precipitation anomalies, and c) predictions from the Seasonal Forecasting System 5 (SEAS5), a numerical dynamical system developed by the European Centre for Medium-Range Weather Forecasts (ECMWF).

The MLR and SVM methods were chosen for the present study due to their straightforwardness in implementation, lack of need for large datasets compared to ANNs (i.e., ANNs need extensive datasets to work effectively, primarily due to their complexity and the requirement for extensive feature learning), and ability to provide satisfactory results, which are even equivalent to those obtained with sophisticated models according to studies developed in China (Xu *et al.*, 2020).

In fact, the use of machine learning and classical statistical methods in climate forecasting is not recent in Brazil. As mentioned earlier, several studies propose seasonal precipitation forecasting models for specific regions of the country, while only a few attempt to cover the entire territory. The reason behind this might be the substantial challenges posed by constructing data-driven models for a country as vast as Brazil. Brazil features a variety of climates and precipitation patterns, which can respond differently to large-scale phenomena such as ENSO. In this study, we developed data-driven models to address this challenge. For each grid cell within Brazil, considering both the annual season and forecast horizon month, we developed a machine learning model using climate indices and precipitation anomalies to estimate seasonal precipitation for the next seven months. The set of predictors effectively used by the models varies spatially, depending on the specific grid cell considered. Consequently, climate indices related to teleconnection patterns that affect only specific sectors of the country were employed in the models corresponding to those grid cells. In this way, our study distinguishes itself from previously proposed works not merely because of the use of MLR and SVM data-driven models, but because it jointly explores a broader range of climate indices, SEAS5 forecasts, as well as historical precipitation data in a systematic way, for each grid cell,

generating specialized forecasts for a country of continental dimensions while considering the unique characteristics of each area.

This paper is structured as follows. In Section 2, we provide an overview of the case study, where we detail the data sources, describe the rainfall forecasting models employed, and explain the implementation process of these models. Section 3 presents the results of the SEAS5 predictive performance evaluation for each Brazilian region, comparing it with the MLR and SVM models. Finally, in Section 4, we summarize the main conclusions drawn from our analysis and suggest potential avenues for future research.

2. Methods and Data

The framework outlined in this study is depicted in Fig. 1. It encapsulates the sequential procedures for refining seasonal precipitation forecasts in Brazil using machine learning methodologies. The process is delineated as follows: Study Area → 1. Inputs → 2. Input Selection → 3. Models → 4. Outputs. The following subsections provide details about the case study, datasets (climate indices, SEAS5 precipitation, and CPC precipitation), input selection methods, machine learning techniques, and evaluation metrics applied to assess seasonal precipitation forecasts from MLR, SVM, and SEAS5.

2.1. Study area

This study focuses on Brazil, a vast country situated in the eastern portion of South America. Extending from 34° S to 6° N in latitude and from 35° W to 74° W in longitude, its boundaries are outlined with a bold polygon in Fig. 2. Due to its vast area, Brazil is geopolitically divided into five regions (i.e., North, Midwest, Northeast, Southeast, and South regions) and several other sub-regions with different climates (Reboita *et al.*, 2010). Despite the variety of climates within individual geopolitical zones, each region is primarily distinguished by a dominant climate. According to Quadro *et al.* (1996), the North region (N) is characterized by an equatorial climate with high precipitation levels and a short dry period. The climatology of annual rainfall ranges between 1500 and 2500 mm in most sectors of this region, but values above 2500 mm are observed in the northwestern area. In the northern part of the South American continent, rainfall predominates during the austral autumn months. The heaviest rainfall begins in the austral spring in Central Brazil and then progresses northward. Consequently, the wettest season closer to the equator occurs in March-April-May, or even later (Grimm and Zilli, 2009; Grimm, 2011).

The Northeast region (NE), on the other hand, has a semi-arid climate in its central area. The rainy season (austral summer) is short-lived, and the climatological mean ranges from 200 to 800 mm in this sector, whereas a

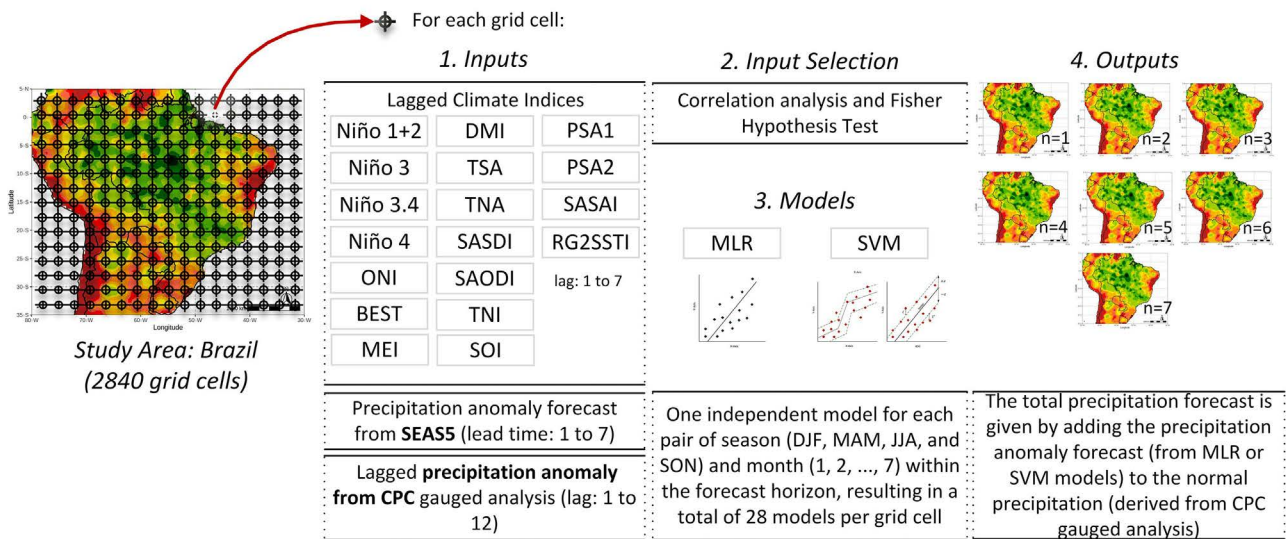


Figure 1 - Overview of the framework to improve seasonal precipitation forecasts using data-driven models.

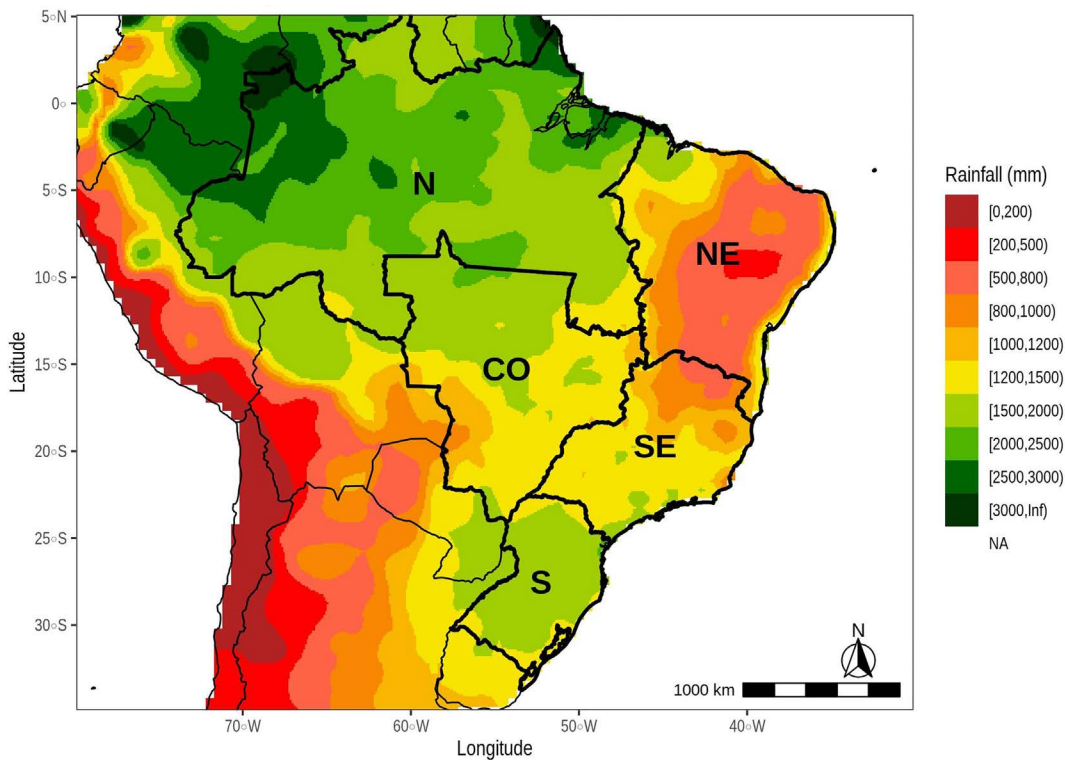


Figure 2 - The Average annual rainfall across Brazil was computed using the CPC unified gauge-based analysis of global daily precipitation (from January 1993 to December 2016). A bold polygon marks the boundaries of Brazilian territory. The acronyms N, CO, NE, SE, and S denote the North, Mid-west, Northeast, Southeast, and South regions of Brazil, respectively.

rainier climate is verified in the northwest area and along the east coast. The rainy season in the NE occurs from March to May when the Atlantic Intertropical Convergence Zone (ITCZ) (i.e., a non-static zone of increased convection, cloudiness, and precipitation encircling the Earth, near the equator) is at its southernmost position

(Grimm, 2011). Rainfall variability in this region is significantly influenced by anomalies in the Pacific Ocean SST. During the negative phase of ENSO (La Niña), the NE and parts of eastern N experience positive rainfall anomalies, while these anomalies reverse during the positive phase (El Niño). However, in the NE region, pre-

precipitation variability is not only linked to the Pacific Ocean but also to anomalies observed in the Atlantic Ocean. The anomalous SST gradient between the North Tropical Atlantic and the South Tropical Atlantic affects the position of the ITCZ, which in turn dictates the rainy season in the NE and exerts a significant influence on rainfall anomalies (Grimm and Zilli, 2009; Grimm, 2011).

Regarding the Midwest (CO) and Southeast (SE) regions, well-defined dry and rainy seasons are evident in their precipitation time series. Their climates are influenced by tropical and mid-latitude atmospheric systems, which result in annual rainfall levels between 1200 and 1500 mm, with a reduction in this range to the north of the Southeast region (500 to 1200 mm) and an increase in the north of the Midwest region (1500 to 2000 mm). The heaviest rainfall starts in the austral spring in Central Brazil and gradually extends to both the southern and northern Brazil. The next season, the austral summer, is the rainy season in the CO and SE due to a summer monsoon regime. The strongest variability occurs near the South Atlantic Convergence Zone (SACZ) (i.e., a zone analogous to the ITCZ, extending diagonally from the southeastern coast of Brazil to the central-southern Amazon), one of the most important features of the South American monsoon system (Grimm and Zilli, 2009; Grimm, 2011).

Lastly, the South (S) region is influenced by frontal systems throughout the year, which are the main drivers of rainfall in this area of higher latitudes. The climatology of annual rainfall ranges between 1500 and 2000 mm in almost the entire area, resulting in a uniform spatial distribution of precipitation. Spring marks the advent of the rainy season in many regions of Brazil, where the monsoon regime prevails, with precipitation concentrated in the hottest months of the year. Even in the S Region, which does not fit into the typical monsoon regime, spring brings a significant amount of precipitation. Much of this precipitation comes from the Mesoscale Convective Complexes (MCC) (i.e., large, organized, and long-lasting clusters of thunderstorms). These convective systems are frequent and contribute substantially to the total rainfall, especially during the transition between seasons. The most intense precipitation progresses from central Brazil towards southern South America from spring onwards, so that the wettest season in the S region peaks in January-February-March (JFM) (Grimm and Zilli, 2009; Grimm, 2011).

2.2. Data

2.2.1. CPC precipitation analysis

The CPC gauge-based analysis of global daily precipitation (Xie *et al.*, 2007; Chen *et al.*, 2008) is a gridded precipitation product. Daily accumulated precipitation data are gathered from approximately 30,000 weather stations (from 12:00 to 12:00 UTC of the next day) and com-

bined using interpolation techniques (Sun, 2018). The resulting gridded data cover the continental land area and are available at a resolution of 0.5° (NOAA, 2023). According to Silva *et al.* (2011), Carvalho *et al.* (2012), Almeida *et al.* (2018), and Torres *et al.* (2020), the CPC precipitation analysis has shown consistent and satisfactory accuracy in representing measured in situ precipitation. However, Hirata and Grimm (2018) noted that, compared to gauge station precipitation, CPC data usually underestimate the extreme rainfall in part of the South American monsoon core region (a sector within southeastern Brazil). According to the authors, this is a drawback that does not compromise the application of this dataset in validation analysis. For this reason, it is considered a suitable product for monitoring synoptic systems and climatic patterns at different time scales, such as the monsoon system, as well as for the validation of rainfall predictive models. In the present study, we use the daily CPC precipitation analysis to estimate the monthly accumulated rainfall from January 1992 to December 2020 across all grid cells within the study area depicted in Fig. 2. We then used this compiled rainfall data to both develop and validate the models proposed in our research.

2.2.2. SEAS5 precipitation forecast

SEAS5 (ECMWF, 2017; Johnson *et al.*, 2019) is a global and coupled general circulation system developed to predict the evolution of the ocean and atmosphere on a seasonal timescale. The system was developed by the ECMWF and supplanted the former System 4 (S4) on November 5, 2017 (Stockdale *et al.*, 2018). SEAS5 runs are initialized on the first day of the month at 00 UTC, generating seasonal precipitation forecasts with a spatial resolution of 36 km. Datasets are made available to the public at no cost, featuring a 1° spatial resolution and a forecast horizon of 215 days (approximately seven months). Since becoming operational, SEAS5 monthly generates an ensemble of 51 predictions. The first one of them is the control member, while the remaining 50 are created by perturbing the initial conditions of the atmosphere and sea surface temperature. Additionally, SEAS5 has produced a set of 25 monthly-based ensemble members via re-forecasting simulations (also known as hindcasts) from January 1993 to December 2016, a dataset applied for model calibration and anomaly computation. To reduce the extensive volume of data from SEAS5, we chose to work with the mean of the ensemble members rather than considering each member individually. The precipitation data from the hindcast (January 1993 to December 2016) and forecast (January 2017 to December 2020) simulations, available on the Copernicus data store (Copernicus, 2021), were downloaded, regridded to the same grid of the CPC precipitation analysis using bilinear interpolation, and then accumulated on a monthly basis.

2.2.3. Climate indices

SST anomalies are the primary drivers of climate variability. In order to predict medium-range climate variables, such as seasonal precipitation, a comprehensive understanding of the ocean-atmosphere interaction patterns that impact remote regions' climates is crucial (Liu and Alexander, 2007; Reboita *et al.*, 2021; Sacco, 2010). Among the existing teleconnection patterns, some have a notable impact on the Brazilian climate, including: the El Niño - Southern Oscillation, Tropical Atlantic SST Dipole,

Indian Ocean SST Dipole, Subtropical Atlantic SST Dipole, Pacific-South America Pattern, SST Anomalies on the south coast of Brazil and Uruguay, as well as the South Atlantic Anticyclone Variability (Souza and Reboita, 2021). These climate forcings are monitored through climate indices available online, with periodic updates provided by meteorological and climate centers. All climate indices used in this study (time series from June 1992 to December 2020), together with their corresponding data sources, are detailed in Table 1.

Table 1 - Climate indices and their data sources.

Index	Details	Link
Niño 1+2, 3, 3.4 and 4	The four climate indices are defined as the anomaly of the average SST in specific regions of the Tropical Pacific Ocean, based on the ERSST v5 data (Huang <i>et al.</i> , 2017), according to the following specifications: Niño 1+2 in the region (0° - 10° S, 90° W - 80° W), Niño 3 in the region (5° N - 5° S, 150° W - 90° W), Niño 3.4 in the region (5° N - 5° S, 170° W - 120° W), and Niño 4 in the region (5° N - 5° S, 160° E - 150° W) (Trenberth, 1998; Trenberth and Stepaniak, 2001).	https://psl.noaa.gov/data/correlation/nina1.anom.data https://psl.noaa.gov/data/correlation/nina3.anom.data https://psl.noaa.gov/data/correlation/nina3.4.anom.data https://psl.noaa.gov/data/correlation/nina4.anom.data
Oceanic Niño Index (ONI)	ONI (Glantz and Ramirez, 2020) is an index based on the three months moving average of the SST anomaly in the Niño 3.4 region, calculated using the ERSST v5 data (Huang <i>et al.</i> , 2017).	https://psl.noaa.gov/data/correlation/oni.data
Southern Oscillation Index (SOI)	SOI (Ropelewski and Jones, 1987) is an index defined as the difference between the standardized sea level pressure anomalies in Tahiti and Darwin. More information on the computation process can be found in (Shi, 2007).	https://psl.noaa.gov/data/correlation/soi.data
Multivariate ENSO Index (MEI)	MEI (Wolter, 1987; Wolter and Timlin, 1993) is the first component derived from the principal component analysis of several variables in the Pacific Ocean. These variables include anomalies of sea level pressure (SLP), sea surface temperature, zonal and meridional components of the surface wind, and outgoing longwave radiation (OLR) (30° S - 30° N, 100° E - 70° W).	https://psl.noaa.gov/enso/mei/data/meiv2.data
Bivariate ENSO Time Series (BEST)	BEST (Smith and Sardeshmukh, 2000) is calculated by combining the standardized time series of SST anomalies in the Niño 3.4 region with the SOI.	https://psl.noaa.gov/data/correlation/censo.data
Trans-Niño Index (TNI)	TNI (Trenberth and Stepaniak, 2001) is calculated by the difference between the standardized time series of SST anomalies in the Niño 1+2 and Niño 4 regions.	https://psl.noaa.gov/data/correlation/tni.data
Tropical Northern Atlantic (TNA) Index	TNA (Enfield <i>et al.</i> , 1999) is an index defined as the anomaly of average SST in the region (5.5° N - 23.5° N, 15° W - 57.5° W), based on HadISST and NOAA OI 1x1 product data.	https://psl.noaa.gov/data/correlation/tna.data
Tropical Southern Atlantic Index (TSA)	TSA (Enfield <i>et al.</i> , 1999) is an index defined as the anomaly of the average SST in the region (0° - 20° S, 10° E - 30° W), based on HadISST and NOAA OI 1x1 product.	https://psl.noaa.gov/data/correlation/tsa.data
Dipole Mode Index (DMI)	DMI (Saji <i>et al.</i> , 1999; Saji and Yamagata, 2003) is calculated as the difference between the anomalies of average SST in the western (50° E - 70° E, 10° S - 10° N) and southeast (90° E - 110° E, 10° S - 0°) regions of the Equatorial Indian Ocean, based on the HadISST data.	https://psl.noaa.gov/geos_wgsp/TimeSeries/Data/dmi.had.long.data
South Atlantic Ocean Dipole Index (SAODI)	SAODI (Nnamchi <i>et al.</i> , 2011) is defined as the standardized difference between the anomalies of average SST in the northeast (15° S - 0°, 10° E - 20° W) and southeast (25° S - 40° S, 10° W - 40° W) sectors of the South Atlantic Ocean, calculated with ERSST v5 data (Huang <i>et al.</i> , 2017).	https://meteorologia.unifei.edu.br/teleconexoes/indice?id=saodi
South Atlantic Subtropical Dipole Index (SASDI)	SASDI (Morioka <i>et al.</i> , 2011) is defined as the standardized difference between the anomalies of average SST in the northeast (15° S - 25° S, 0° - 20° W) and southwest (30° S - 40° S, 10° W - 30° W) sec-	https://meteorologia.unifei.edu.br/teleconexoes/indice?id=sasdi

(continued)

Table 1 - continued

Index	Details	Link
	tors of the South Atlantic Subtropical Ocean, calculated with ERSST v5 data (Huang <i>et al.</i> , 2017).	
Pacific South American 1 (PSA1)	PSA 1 (Mo and Higgins, 1998; Kidson, 1999; Mo, 2000) is the second component obtained with the application of the principal component analysis technique to seasonal anomalies of geopotential height at 700 hPa.	https://meteorologia.unifei.edu.br/teleconexoes/indice?id=psa1
Pacific South American 2 (PSA2)	PSA 2 (Mo and Higgins, 1998; Kidson, 1999; Mo, 2000) is the third component obtained with the application of the principal component analysis technique to seasonal anomalies of geopotential height at 700 hPa.	https://meteorologia.unifei.edu.br/teleconexoes/indice?id=psa2
Region 2 SST Index (RG2SSTI)	RG2SSTI (Reboita <i>et al.</i> , 2021; Reboita <i>et al.</i> , 2010; Jesus, 2021) is the anomaly of the average SST in the region (40° S - 30° S and 57° - 47° W) between southern Brazil and Uruguay, calculated with ERSST v5 data (Huang <i>et al.</i> , 2017).	https://meteorologia.unifei.edu.br/teleconexoes/indice?id=itsmrg2
South Atlantic Subtropical Anticyclone Index (SASAI)	SASAI (Reboita <i>et al.</i> , 2019) is the difference between the atmospheric pressure anomalies at mean sea level in the Southeast (25° S - 15° S, 50° W - 40° W) and South (37.5° S - 27.5° S, 60° W - 50° W) Brazilian regions using the ERA 5 reanalysis.	https://meteorologia.unifei.edu.br/teleconexoes/indice?id=iasas

2.3. Data driven models

2.3.1. Data structure and pre-processing

In this study, the data described in Section 2. 2. - including CPC precipitation analysis, SEAS5 precipitation forecasts, and the climate indices - compose the input datasets of the data-driven models developed here. However, the way these datasets are processed and used to feed the models is not uniform. Rather, it varies according to the data type and the specific month ahead for which we aim to predict precipitation, as detailed below.

The first input dataset consists of indices related to climate forcings from the Indian, Atlantic, and Pacific Oceans that affect the Brazilian climate remotely. During the forecasting process, climate indices for the upcoming months are unavailable, as their calculation requires observational data such as sea surface temperature (SST) and atmospheric pressure at sea level. For this reason, climate indices data from previous months, which are already part of the historical time series, have their lagged relationship with the precipitation anomaly explored. In this study, we use up to the last seven months from the indices' time series as inputs to data-driven models (see Fig. 3a). For instance, if the forecast horizon ranges from January to July 2016 (considering predictions made in January 2016), the data-driven model uses index data from June to December 2015 for the January 2016 prediction (lags 7 to 1), from July to December 2015 for the February 2016 prediction (lags 7 to 2), and so on. For the final month of the forecast horizon (July 2016), only the data from December 2015 (lag 7) is used.

Concerning the second dataset of predictors, Fig. 3b. illustrates the application of the results from the ECMWF dynamic system runs. The time series of differences between SEAS5 predictions and the CPC monthly normal

precipitation provides information on how much more or less rainfall the climate model is forecasting for the next seven months, taking as a reference the average rainfall over the last years. As shown in Fig. 3b., data-driven models explore this time series, taking into account the specific forthcoming month for which the precipitation is being forecasted.

The CPC precipitation anomalies comprise the third input dataset. Similar to the climate indices case, the lagged relationship between the predictor and the forecast variable is explored here. To achieve this, we use data from the last months of the CPC precipitation time series to calculate the precipitation anomalies of the last 12 months. Fig. 3c. provides an example of how the post-processed CPC data from January 2015 to December 2015 are used by the data-driven models to make predictions for the next seven months, from January 2016 to July 2016. Note that as the lead time increases from one to seven months, less short-term information becomes available to feed the data-driven models.

In summary, the data are processed in the following way to generate the predictors:

- Indices time series for each grid cell, lagged by 1, 2, 3, 4, 5, 6, and 7 months: (18 indices) x (7 lags) = 126 features;
- SEAS5 precipitation anomalies, predictions for 1, 2, 3, 4, 5, 6 and 7 months ahead: 7 features;
- CPC precipitation anomalies for each grid cell, lagged by 1, 2, 3, 4, 5, 6, through 12 months: 12 features.

2.3.2. Multiple linear regression

MLR (Hocking, 1976) is a classical data-driven model widely used in forecasting science. It has applications in predicting, for instance, seasonal precipitation (Xu *et al.*, 2020), hydrological droughts (Seibert *et al.*, 2017),

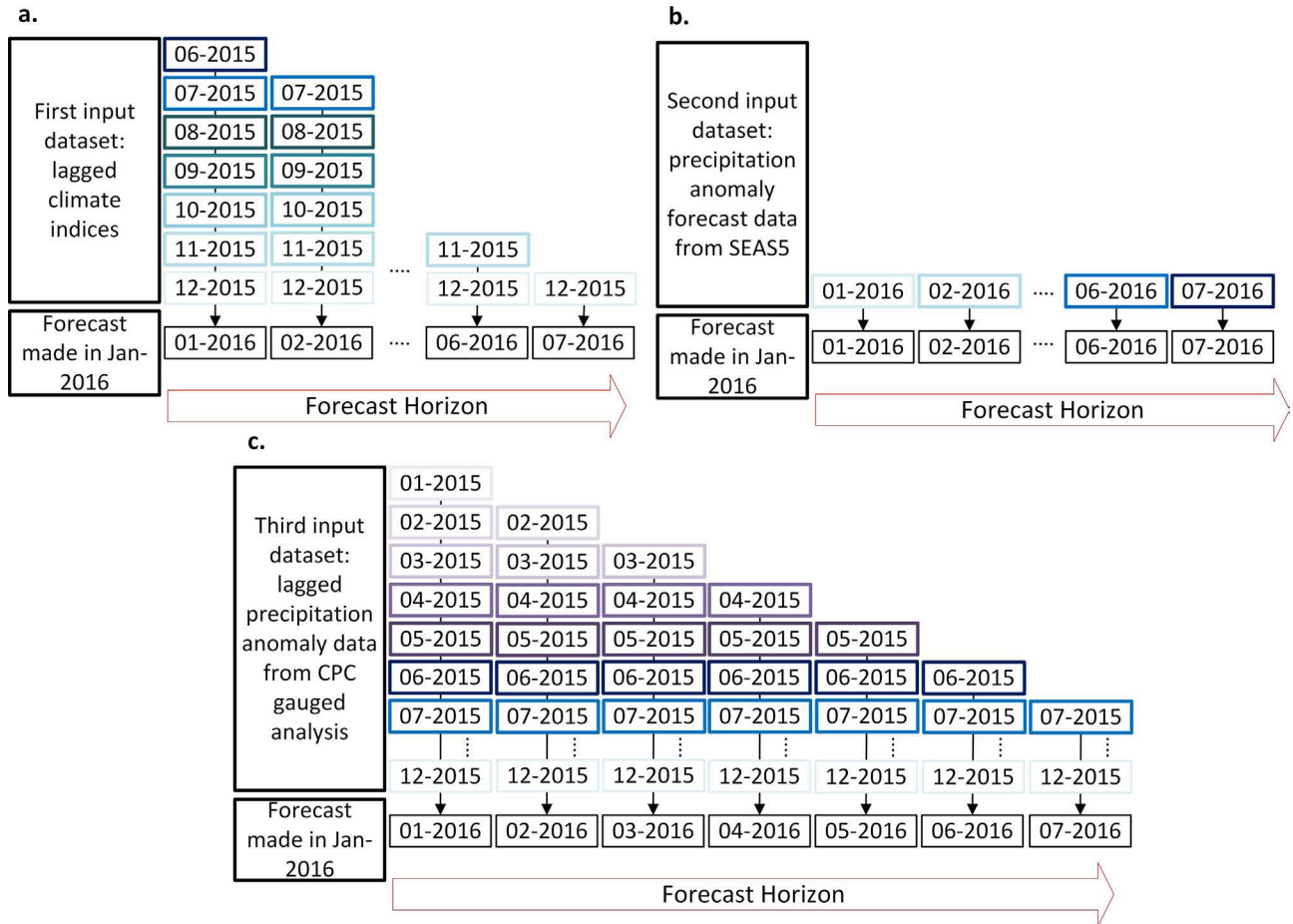


Figure 3 - The data-driven models take the following as inputs: (a) lagged climate indices (lags from 1 to 7), (b) SEAS5 precipitation anomalies (lead time of 7 months), and (c) lagged precipitation anomalies from CPC gauged analysis (lags from 1 to 12).

streamflow (Jozaghi *et al.*, 2021; Moradi *et al.*, 2020), and wind power generation (Gupta and Saraswat, 2020). In the MLR model, the dependent variable y is expressed as a function of a set of independent variables x_1, x_2, \dots, x_n , as shown in Eq. (1). Here, β_0 is the constant term, while $\beta_1, \beta_2, \dots, \beta_n$ are the linear coefficients associated with the predictors, and ε represents the residual. In this linear regression approach, the angular coefficients assign weights to the predictors, determining the influence of each independent variable on $y(t)$. These beta terms are computed through the least-squares method (Miller, 2017), which identifies the coefficients that minimize the sum of the squared residuals.

$$y(t) = \beta_0 + \beta_1 x_1(t) + \beta_2 x_2(t) + \dots + \beta_n x_n(t) + \varepsilon(t) \quad (1)$$

2.3.2.1. MLR input variable selection

Regarding the predictors included in the MLR models, the final set was fine-tuned for each grid cell within the study area, as depicted in Fig. 2. This optimization was performed using the stepwise method, a systematic approach where significant predictors are identified via

Fisher's Hypothesis Testing (F-test) (Pope and Webster, 1972). In this method, independent variables undergo rigorous scrutiny; they are sequentially added (in forward iterations) or removed (in backward iterations) based on their statistical significance to the regression model, configuring a bidirectional process of input selection. This ensures that the final model comprises only those predictors that provide meaningful and statistically relevant contributions. The objective is to strike a balance between model simplicity and predictive accuracy, eliminating any superfluous variables that do not enhance model performance.

A detailed description of the bidirectional stepwise regression algorithm is provided below:

- **Initialization:** Start with an empty model and choose the significance levels for entering and removing predictors from the model. These are typically denoted as alpha-to-enter and alpha-to-remove (e.g., $\alpha = 0.05$ in both cases);
- **Forward selection step (addition of predictors):** Rank the predictors according to certain criteria, such as the Pearson correlation between the set of predictors

and the forecast variable. Add the first predictor from the rank to the MLR model if the related p-value is less than the alpha-to-enter (the p-value is computed using the F-test). Repeat this step, considering the remaining predictors from the rank, until no more variables meet the criterion for entry;

- **Backward step (removal of predictors):** Remove the predictor that shows the least contribution to the model (often the one with the highest p-value) if its p-value is greater than alpha-to-remove. Repeat this step until no more variables meet the criterion for removal;
- **Iterate:** Alternate between forward and backward steps. At each step, reevaluate the model and the significance of each variable;
- **Stopping criterion:** The process stops when no predictors outside the model have a p-value lower than alpha-to-enter, and no predictors in the model have a p-value higher than alpha-to-remove. At this point, the model is considered optimized according to the step-wise criteria.

2.3.3. Support vector machine

SVM (Cortes and Vapnik, 1995; Vapnik, 2000) is a machine learning algorithm originally devised to solve data classification problems and later extended to address regression problems. It has been widely applied in forecasting applications. Examples include the prediction of seasonal precipitation (Xu *et al.*, 2020), short-term wind speed and power generation (Li *et al.*, 1996; Yang *et al.*, 2015; Wang *et al.*, 2018), streamflow (Rasouli *et al.*, 2012; Bhandari *et al.*, 2019), and urban flash floods (Yan *et al.*, 2018). The SVM method relies on determining the support vectors (i.e., data points influencing the regressive model), represented by w , and the hyperplane $y_i = w^t x + b$ (i.e., the regressive model) that maximizes the margin ε , as shown in Fig. 4a.

The SVM approach aims to produce a regression model that best fits the training dataset and minimizes the generalization error, using the outermost points from the dataset as a reference. Eq. (2) illustrates the SVM mathematical programming problem, where w is a vector perpendicular to the optimal hyperplane, b is a constant, and y_i and x_i represent the training data, respectively. The objective function in Eq. (2) seeks to maximize the margin depicted in Fig. 4a. (which is proportional to $1/w$), with constraints forming the frontier hyperplanes, where the errors remain below ε .

$$\min_{w,b} \frac{1}{2} \|w\|^2, \quad s.t. \begin{cases} y_i - w^t x_i - b \leq \varepsilon \\ w^t x_i + b - y_i \leq \varepsilon \end{cases} \quad (2)$$

The term w^2 in Eq. (2) is referred to as Ridge penalization. It is frequently employed in machine learning and statistical modeling for regularization purposes. Central to

the SVM methodology, its essence lies in penalizing weight magnitudes, guiding SVMs to simpler and more generalizable models with minimized weight vectors. This regularization technique is also essential in moderating the impact of correlated predictors by introducing a penalty based on the squared magnitude of coefficients, ensuring that models remain resilient to minor data fluctuations, and sidestepping the challenges of multicollinearity.

2.3.3.1. Handling non-linearities using kernel functions

Compared to MLR, SVM stands out for being suitable for nonlinear mathematical problems without requiring large datasets. When the training dataset is nonlinear, we can use a function Φ , as illustrated in Fig. 4b. and c., to map the data x_i from a low-dimensional space to a high-dimensional space, denoted as $\Phi(x_i)$. This transformation enlarges the feature space where regression can be more effectively performed. The new formulation of the SVM mathematical programming problem is provided in Eq. (3), where slack variables ξ_i and ξ_i^* (whose purpose is to avoid model overfitting) and a parameter C (which controls the relationship between error and margin) are also considered.

$$\min_{w,b,\xi_i,\xi_i^*} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N (\xi_i + \xi_i^*), \quad s.t. \begin{cases} y_i - w^t \Phi(x_i) - b \leq \varepsilon + \xi_i \\ w^t \Phi(x_i) + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \quad (3)$$

The new mathematical programming problem from Eq. (3) can be solved through its dual version (Mangasarian, 1994), a reformulated problem that was first proposed by Smola and Schölkopf (2004). The solution is detailed in Eq. (4), where $f(x)$ is the resulting SVM regression model, α and α_i are dual variables (i.e., Lagrange multipliers) and the product $\Phi(x_i)^t \Phi(x)$ is called kernel function $K(x_i, x)$, a term responsible for transforming the dataset implicitly (i.e., without detailing the exact mathematical form of the function Φ , but only the resulting expression $\Phi(x_i)^t \Phi(x)$).

$$f(x) = \sum_{i=1}^N N(\alpha_i - \alpha_i^*) K(x_i, x) + b \quad (4)$$

In the study discussed here, the sigmoidal kernel function $K(x_i, x) = \tanh(\gamma x^t x + c_0)$, with parameters γ and c_0 , was chosen after performing tests with linear, polynomial, and radial basis kernel functions. The optimal values of the SVM calibratable parameters, namely γ and c_0 , were determined through a heuristic process using the `tune.svm` function from the `e1071` package in R. The `tune.`

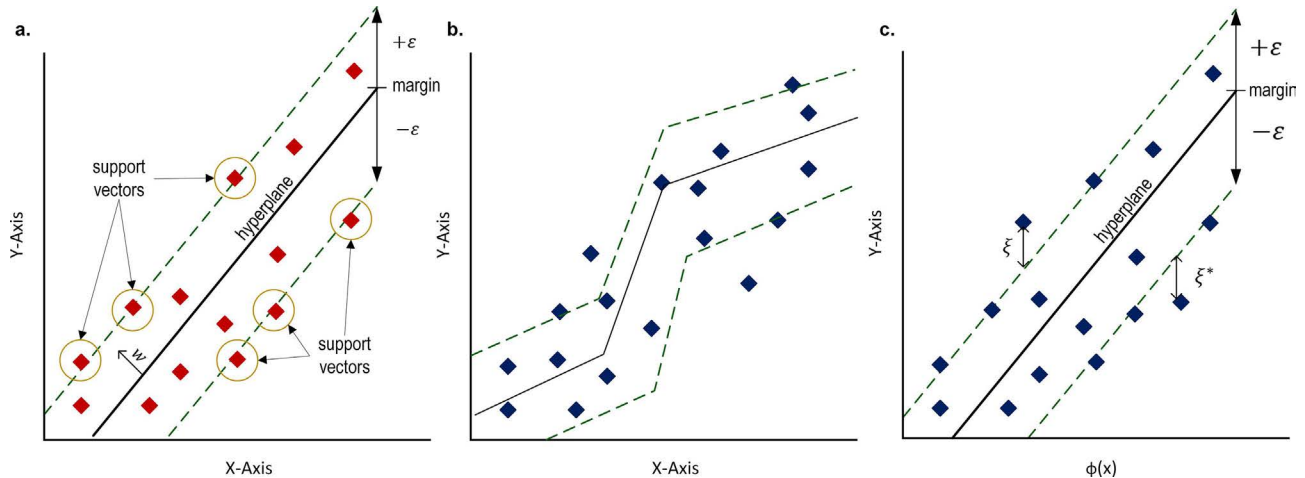


Figure 4 - SVM regression diagrams. In (a) we illustrate the basic concepts behind the SVM methodology (i.e., support vectors, optimal hyperplane, and margin of error). (b) and (c) exemplifies the use of kernel functions to remap the data x_i from a low- to a high-dimensional space $\Phi(x_i)$ where the regression process is easier to be accomplished.

svm function employs a 5-fold cross-validation strategy to fine-tune the parameters of SVM models, by segmenting data from 1993 to 2016 into five distinct parts. In each iteration, four parts are used for model training while the remaining one serves as validation, ensuring each segment is validated once. This iterative approach facilitates a comprehensive evaluation of the model's effectiveness and adaptability to different data subsets. Once this process is finished, the function provides the best-tuned model, making it ready for use in predictions.

2.3.3.2. SVM input variable selection

Correlation-based feature selection is a method used in machine learning to select the most relevant predictors for a model based on their correlation with the target variable. This method is particularly useful in reducing the dimensionality of a dataset and improving model performance. In our precipitation forecast problem, the final set of predictors effectively considered was optimized for each grid cell using as a selection criterion the correlation between the time series of each predictor and the CPC precipitation anomalies. Only predictors with correlations above 0.2 or below -0.2 make up the input set for SVMs. This range encompasses predictors with varying degrees of correlation strength, extending from weak to extremely strong, as classified in the work of [Evans \(1996\)](#). The rationale for adopting a relatively low threshold value stems from the observation that correlations between precipitation anomalies and climate indices typically do not achieve substantial magnitudes.

At first glance, this value may not appear highly selective. However, it grants the SVMs greater freedom to explore the set of predictors, focusing on the most promising ones, and reduces the dimension of the problem by discarding those with an absolute correlation value below

the threshold. In addition, it is important to remember that SVMs also have a regularization term in the objective function (Ridge penalization w^2) (see Section 2.3.3.), which minimizes the weights attributed to any other predictor that does not prove to be relevant to the model and helps in dealing with multicollinearity by avoiding over-reliance on any single feature.

2.3.4. Arrangement of models

In Section 2.3.1., we introduced three sets of predictors for seasonal precipitation forecasting: CPC precipitation anomalies, climate indices, and SEAS5 precipitation anomalies. However, it is important to note, as detailed in Sections 2.3.2 and 2.3.3, that not all the predictors listed in Section 2.3.1 are incorporated into the data-driven models. The exclusion of candidate predictors stems from the fact that not all of them can explain the variability in seasonal precipitation at the grid cells considered in this study. This limitation becomes particularly apparent with climate indices tracking teleconnection patterns that affect specific Brazilian regions. In such cases, not all grid cell precipitation anomalies are related to each climate index. Therefore, a pre-selection process is necessary to determine the most relevant predictors for the input dataset of each grid cell.

In the process of constructing the MLR-based models, the Fisher Hypothesis Test is applied to refine the final set of predictors by selecting those that have a statistically significant relationship with the predicted variable (i.e., only those that significantly enhance the quality of the regression model's fit are retained). Conversely, for SVMs, predictors are considered only if they have a correlation greater than 0.2 or less than -0.2 with the recorded precipitation anomalies in each grid cell. [Table 2](#) ranks the predictors detailed in Section 2.3.1. according to their

Table 2 - Ranking of key predictors used by MLR and SVM models for seasonal precipitation forecasting in Brazil (only for DJF months).

Data-driven model	Rank	SE [320 grid cells]		CO [542 grid cells]		S [207 grid cells]		NE [512 grid cells]		N [1259 grid cells]	
		Variable	%	Variable	%	Variable	%	Variable	%	Variable	%
MLR	1	CPC	54.7	CPC	47.5	CPC	34.2	CPC	29.0	CPC	31.1
MLR	2	SEAS5	12.5	SEAS5	12.4	SEAS5	14.4	PSA1	19.4	RG2SSTI	6.7
MLR	3	RG2SSTI	8.0	RG2SSTI	11.3	Niño 1+2	6.7	SEAS5	13.0	SEAS5	6.4
MLR	4	SAODI	6.1	TNA	5.5	RG2SSTI	5.4	SASDI	5.9	TNA	5.2
MLR	5	PSA1	5.8	PSA1	3.6	PSA2	4.1	RG2SSTI	5.2	ONI	5.0
MLR	6	TNA	3.2	Niño 1+2	2.9	Niño 3	4.0	Niño 3	4.8	Niño 3.4	4.5
MLR	7	PSA2	3.0	SASAI	2.4	MEI	3.9	TSA	3.3	Niño 4	4.1
SVM	1	CPC	67.3	CPC	63.7	CPC	47.4	CPC	52.2	CPC	37.4
SVM	2	SEAS5	13.5	SEAS5	12.7	SEAS5	9.5	SEAS5	10.4	SEAS5	7.5
SVM	3	PSA1	4.1	RG2SSTI	4.8	Niño 1+2	4.3	PSA1	5.9	Niño 4	4.6
SVM	4	TNA	3.4	TNA	3.2	Niño 3	4.2	TSA	3.2	Niño 3.4	4.2
SVM	5	RG2SSTI	3.0	PSA1	1.7	BEST	3.9	Niño 3	3.0	RG2SSTI	4.0
SVM	6	SAODI	2.6	SASAI	1.5	SOI	3.8	Niño 3.4	2.9	ONI	4.0
SVM	7	SASDI	1.0	Niño 1+2	1.5	Niño 3.4	3.8	SASDI	2.7	BEST	3.8

level of participation in data-driven models related to the five geographical regions of Brazil. The percentage specified in this context illustrates the contribution of each predictor to the set of models developed for the grid cells within SE, CO, S, NE, and N regions, with the table highlighting only the seven predictors most commonly used by the models. The results reveal that CPC and SEAS5 precipitation anomaly data are commonly exploited across all regions. In particular, CPC precipitation anomalies stand out due to their substantial percentage in the input dataset of both MLR and SVM models. This pattern indicates the existence of autocorrelation in the precipitation anomalies time series, a feature that is highly relevant to the data-driven models.

Regarding climate indices, there are noticeable regional variations in those that are commonly employed. In both the SE and CO regions, the indices RG2SSTI, TNA, and PSA1, derived from SST anomalies in the Atlantic Ocean and geopotential height anomalies at 700 hPa, are commonly employed by the models. In the case of the NE region, the models rely not only on PSA1 and Atlantic Ocean indices such as SASDI, TSA, and RG2SSTI, but also include a substantial number of indices associated with the ENSO phenomenon, specifically Niño 3 and Niño 3.4. This pattern is even more pronounced in the S region, where indices such as Niño 1+2, Niño 3, Niño 3.4, BEST, and SOI are notably prevalent. In the N region, the models similarly feature a selection of indices that includes ONI, Niño 3.4, Niño 4, and BEST.

The frequent integration of large-scale climate indices from both the Pacific and Atlantic Oceans underscores the essential role of the oceanic conditions in forecasting seasonal precipitation. In the Pacific, well-studied phe-

nomena like El Niño and La Niña exert a significant impact on rainfall patterns in the S, NE, and N regions, and are classified as the most influential teleconnection patterns affecting South America (Reboita *et al.*, 2021). Conversely, in the Atlantic, SST anomalies and oceanic dipoles influence the positioning of the ITCZ (Nnamchi *et al.*, 2011; Morioka *et al.*, 2011; Mo and Higgins, 1998), a key component that shapes the rainfall regime in the NE.

2.3.5. Framework of data-driven models

Figure 1 illustrates the key components of the seasonal rainfall forecasting framework presented in this paper. In this approach, data-driven models were constructed for the grid cells within the entire Brazilian territory, incorporating time series of climate indices and precipitation anomalies from the CPC and SEAS5 as predictors. The parameters of the MLR and SVM models were tuned using data from the period between January 1993 and December 2016. In order to take into account the unique characteristics of each annual season and the models' diminishing predictive accuracy as the lead time extends, individual models were developed for each month of the forecast horizon and for each meteorological season (MAM, JJA, SON, or DJF). This yielded a total of 28 data-driven models (four quarters \times seven months) per grid cell. As a result, the precipitation forecast for a specific grid cell is determined by adding the predicted precipitation anomaly (generated by the model created for a specific season and month ahead) to the precipitation climatology. The forecast for the next seven months is made by executing the aforementioned procedure for each month of the forecast horizon.

2.3.6. Validation and evaluation metrics

To validate the MLR and SVM data-driven models, we conducted a series of monthly backtest simulations from January 2017 to December 2020, consistently forecasting precipitation up to seven months ahead. We chose this particular time interval to align with the availability of real-time SEAS5 forecast data, thereby establishing a standardized timeframe that enables more straightforward comparisons with the European seasonal forecast system. While evaluating the models, precipitation forecast data from MLR, SVM, and SEAS5 were initially segregated based on the specific month (denoted as 'month n ') within the forecast horizon, and subsequently categorized by season (MAM, JJA, SON, and DJF). The segregation of data is justified because forecasts for the initial months of the time horizon are expected to be more accurate, and it is relevant to assess how this performance diminishes over time. Moreover, we aimed to examine the models' performance during the rainy season in the Southern Hemisphere, specifically during the austral summer (DJF). This objective led us to segregate the data by season, as illustrated by Fig. 5.

The datasets, organized by both the specific month within the 7-month forecast horizon (represented as n , where $n = 1, 2, \dots, 7$) and the season (MAM, JJA, SON, and DJF), were evaluated to understand the models' regional performance for bias and predictive accuracy. This assessment employed two specific metrics: Mean Absolute Error (MAE) and bias, along with a hypothesis test, as proposed by Diebold and Mariano (Diebold and Mariano, 1995). A detailed explanation of these metrics and the mathematical expressions used is provided in Table 3.

3. Results and Discussion

3.1. SEAS5 hindcast precipitation skill (1993-2016)

Figure 6 illustrates the spatiotemporal patterns of MAE for precipitation forecasts made by SEAS5 (from January 1993 to December 2016) for the austral summer (DJF quarter) and months 1, 3, and 7 of the forecast hor-

izon. The main goal of this set of maps is to reveal for which Brazilian regions the precipitation forecasts have high or low predictive accuracy.

The results reveal that regions with the largest MAE values are those where the SACZ affects the precipitation regime, from the Amazon Rainforest towards the southeastern Brazilian coast, encompassing N, CO, and SE. The SACZ is a meteorological system characterized by the emergence of a large band of clouds when stationary weather frontal systems, enduring for more than three days, interact with tropical convection over South America (Oliveira, 1986; Grimm *et al.*, 2021; Grimm, 2019). When active, the SACZ generates significant levels of precipitation, and for this reason, it is considered an essential meteorological system for the precipitation regime of the SE and CO regions during the rainy season. Regarding the S and NE regions, the precipitation forecasts are more accurate than those made for the rest of Brazil. This is especially true for the NE sector, where the lowest MAE values are observed. However, as the forecast horizon progresses from month 1 to 3 and then to 7, the absolute errors become more pronounced across Brazil. This trend indicates a marked decrease in SEAS5's predictive performance when the forecast is made further in advance, particularly in the SE region. To further illustrate this point, the corresponding average of the MAE values for each region of Brazil is shown in Table 4.

Figure 7 illustrates the spatiotemporal patterns of the bias for the SEAS5 precipitation forecasts. The primary aim of this second set of maps is to identify the regions of Brazil where SEAS5 underestimates or overestimates total precipitation, a feature that is crucial for understanding how model performance varies across Brazil during the rainy season. The results reveal that SEAS5 tends to overestimate precipitation in the S, SE, and parts of the CO and N regions of Brazil, especially in sectors affected by the SACZ, where the highest bias values are observed. Conversely, an opposite pattern is observed in the eastern N and northern CO sectors, where the precipitation forecasts tend to underestimate rainfall. In the case of the NE region

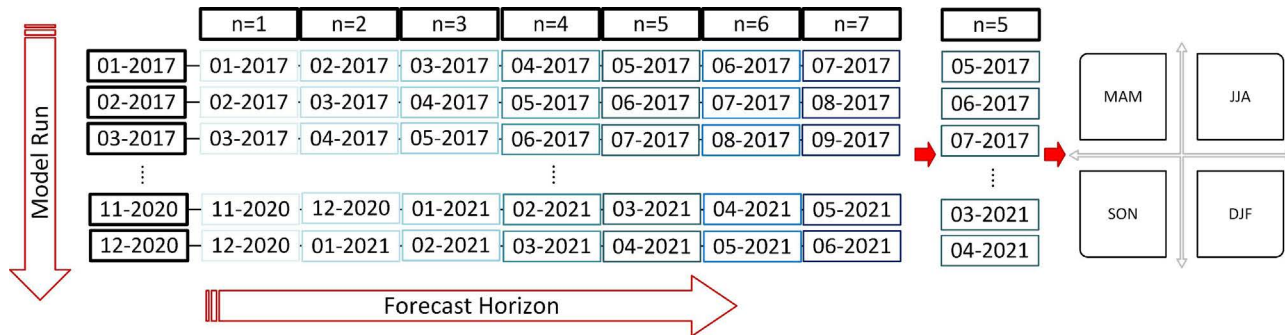
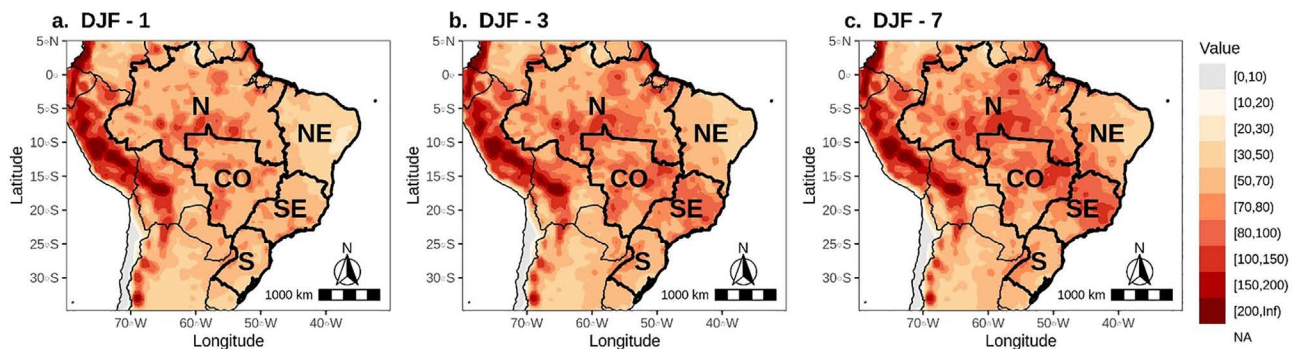


Figure 5 - Segregation of the seasonal rainfall forecasts according to the month of the forecast horizon ($n = 1, 2, \dots, 7$) and the season (MAM, JJA, SON, and DJF). As an illustrative example, the figure shows the segregation procedure applied to the dataset related to the fifth month ($n = 5$).

Table 3 - Evaluation metrics and hypothesis test.

Metric	Description	Equation
Mean Absolute Error (MAE)	The Mean Absolute Error (MAE), as described by Montgomery <i>et al.</i> (2008) , measures the accuracy of a model by determining the closeness between the predicted and observed values of a variable. This is achieved by averaging the absolute errors across a set of predictions. Since MAE is a metric that directly compares forecasts to observed data, an ideal value for this measure would be close to zero.	$MAE = \frac{1}{T} \sum_{t=1}^T x_{m,t} - y_t \quad (5)$ <p>where $x_{m,t}$ is the prediction made by model m for time t, and y_t is the measured value of the predicted variable for time t.</p>
Bias	Bias is a metric that measures the average difference between predicted and measured values of a variable (Pal, 2016). Since bias takes into account the resulting sign from the arithmetic operation, the metric can indicate whether the predicted variable is systematically overestimated (positive bias) or underestimated (negative bias) by the model. Ideally, the value of bias should be close to zero. However, caution must be exercised when analyzing this metric, as errors with opposite signs and similar magnitudes can cancel each other out, as illustrated in Eq. (6) .	$bias = \frac{1}{T} \sum_{t=1}^T (x_{m,t} - y_t) \quad (6)$ <p>where $x_{m,t}$ is the prediction made by model m for time t, and y_t is the measured value of the predicted variable for time t.</p>
Diebold-Mariano Test (DMT)	The Diebold-Mariano test (Diebold and Mariano, 1995) is a hypothesis test largely used to compare the predictive accuracy of two time series of forecasts A and B. The time series of absolute (or quadratic) error, $loss(e_{A,t})$ and $loss(e_{B,t})$, are compared along the time axis by the differential loss function $d_t = loss(e_{A,t}) - loss(e_{B,t})$. If the expectation of d_t is zero, A and B have the same level of predictive accuracy (null hypothesis H_0); otherwise, one of them is more accurate than the other one (alternative hypothesis H_a). The result of such a test is given by analyzing the p-value, a probability based on the cumulative distribution function of the DMT statistic, according to Eq. (7) . If p-value is lower (bigger) than the significance level α , we reject H_0 (accept H_0) and accept H_a (reject H_a).	$H_0 : E(d_t) = 0$ $H_a : E(d_t) \neq 0$ $DMT_{calc} = \frac{\bar{d}}{\sqrt{\frac{\gamma_0 + 2 \sum_{k=1}^{h-1} \gamma_k}{n}}} \quad (7)$ <p>$p\text{-value} = 2 \times [1 - CDF(DMT_{calc})]$</p> <p>where CDF is the cumulative distribution function of the DMT statistic; h is the forecast horizon; γ_k is the autocovariance of d at lag k; n is the total number of data points of each time series; and \bar{d} is the mean value of the differential loss function.</p>

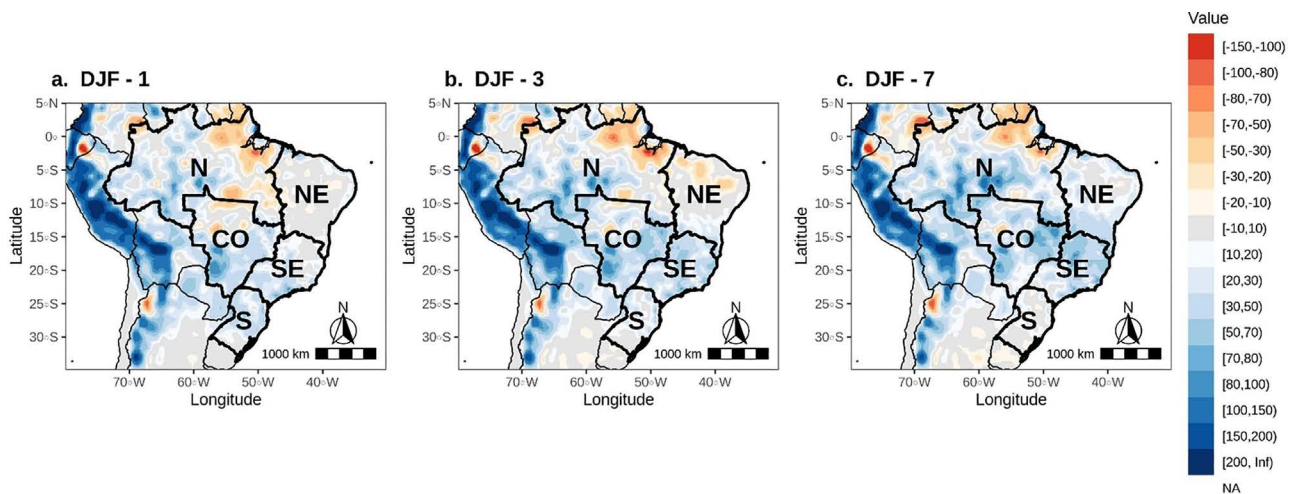
**Figure 6** - Spatiotemporal pattern of MAE [mm/month] for SEAS5 precipitation forecasts in the austral summer (DJF). The precipitation data from the hindcast simulations (January 1993 to December 2016) were accumulated and assessed according to the months of the forecast horizon (months 1, 3, and 7).

of Brazil, the bias sign alternates according to the month of the forecast horizon, and can be predominantly negative or positive, as shown in [Fig. 7b.](#) and [c.](#), respectively. [Table 4](#) not only supports these maps, but also indicates that the SE, CO, and N regions exhibit an almost continuous increase in positive bias as the month of the forecast time horizon advances from 1 to 7, particularly in the SE and CO regions where the highest mean bias values are found.

The results presented herein are similar to those reported by [Gubler *et al.* \(2020\)](#). The authors were the first to investigate the performance of the SEAS5 temperature and precipitation forecasts in South America using data from rain gauge stations as the ground truth. Although precipitation forecasting is a challenging task to perform for reasons such as the intermittent nature of precipitation, the strong dependence of rainfall on local factors, and the absence of proper equations for this variable ([Lopez,](#)

Table 4 - Average MAE and bias for SEAS5 precipitation forecasts (January 1993 to December 2016) in the rainy season (DJF), by region and month of the forecast horizon.

Region	Month of the forecast horizon													
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
	MAE [mm/month]							Bias [mm/month]						
S	58.79	62.71	62.71	63.01	61.36	62.26	62.55	24.70	11.94	13.97	12.75	5.30	4.34	4.17
SE	62.73	80.12	79.08	80.03	83.67	84.51	84.51	20.06	33.20	35.33	34.70	39.70	38.24	38.74
N	69.06	76.19	75.84	76.69	76.83	78.51	79.55	8.64	11.25	12.38	14.40	16.67	17.21	21.97
NE	43.93	54.86	54.48	57.10	58.37	59.54	61.79	3.65	1.05	-0.25	1.21	6.47	12.06	21.34
CO	69.35	77.62	77.42	76.70	78.81	79.96	78.87	22.55	32.74	34.30	32.91	36.50	36.51	36.55

**Figure 7** - Spatiotemporal pattern of bias [mm/month] for SEAS5 precipitation forecasts in the austral summer (DJF). The precipitation data from the hind-cast simulations (January 1993 to December 2016) were accumulated and assessed according to the months of the forecast horizon (months 1, 3, and 7).

2007), SEAS5 presents a high performance in the NE and S regions of Brazil, which are regions where ENSO has a strong influence on rainfall variability. The skill of SEAS5 forecasts in regions influenced by ENSO underlines the model's ability to represent one of the main teleconnection patterns affecting the seasonal precipitation in Brazil.

In the case of western Amazonia and other extratropical sectors in Brazil, disregarding the southern sector of the S region (Rio Grande do Sul), the performance of the SEAS5 precipitation forecasts is low according to Gubler *et al.* (2020), a result that was again validated in this study. The dynamical system performs poorly in the N, CO, and SE regions of Brazil during the rainy season. It produces high MAE values and tends to overpredict rainfall in these regions. Notable exceptions to this pattern are the eastern part of the N region and the northern area of the CO region, where the biases are negative.

Considering the aforementioned results and given the importance of seasonal rainfall predictions for society, it is clear that there is a pressing need for ongoing improvements to forecasting models. With this in mind, the following sections discuss the results of the framework

outlined in Fig. 1. This approach combines data-driven models with information from climate research centers to refine precipitation forecasts in Brazil.

3.2. DJF precipitation forecast skill of SEAS5, MLR, and SVM (2017-2020)

The present section provides a comprehensive overview of the validation results derived from the simulations conducted from January 2017 to December 2020, using MAE and bias as evaluative metrics. The primary objective is to evaluate the predictive accuracy of MLR and SVM data-driven models, comparing them with SEAS5 to identify which model yields better performance for the DJF quarter. A secondary aim is to examine the regional variations in the skills of these models.

Figure 8 presents the spatiotemporal pattern of MAE for predictions made by SEAS5 and the data-driven models (MLR and SVM). The set of maps shows that the NE and S regions have the lowest MAE values, with slight differences in the predictive performance of the models. Such outcomes suggest that in regions where seasonal rainfall variability heavily relies on ENSO, data-driven

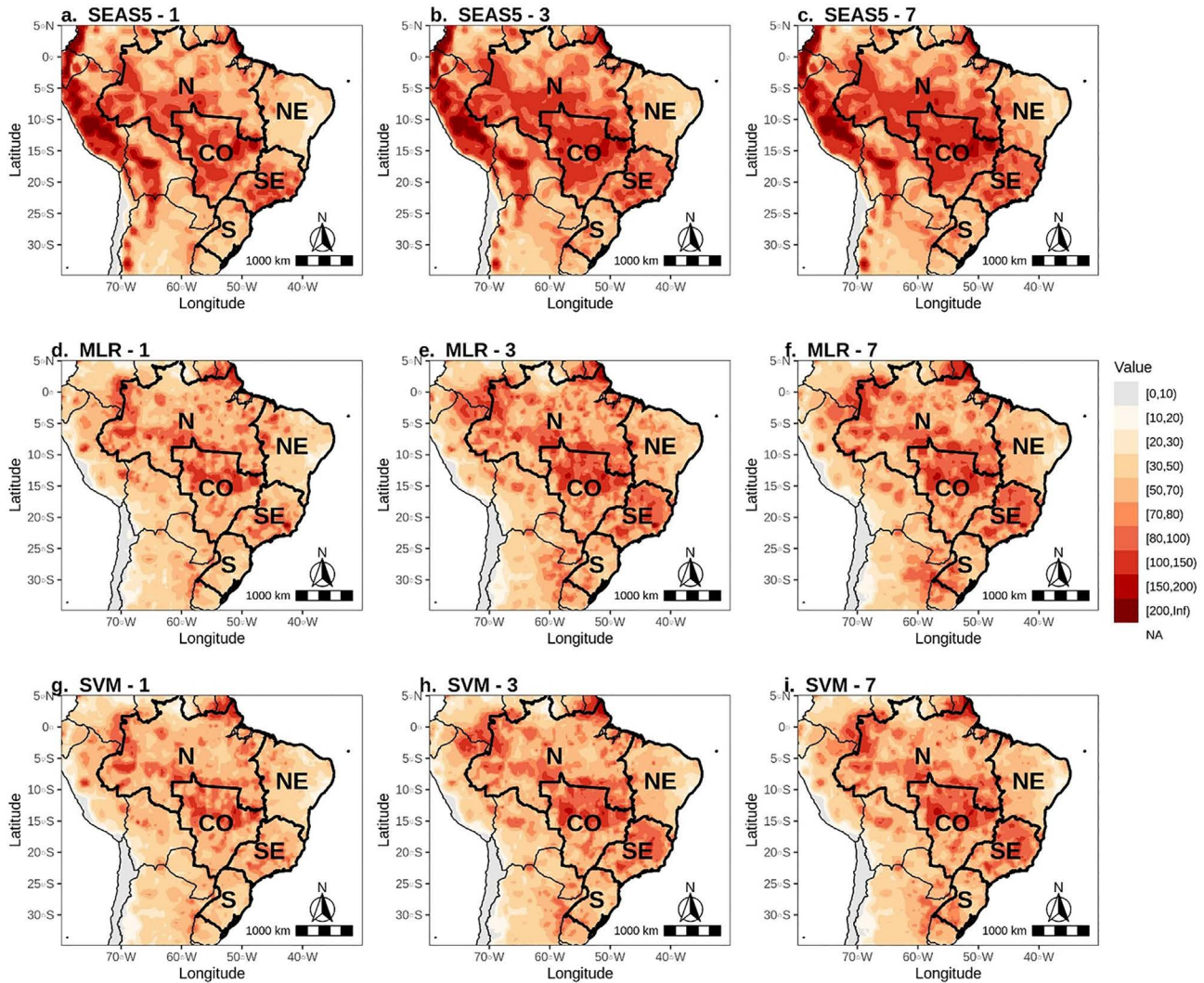


Figure 8 - Spatiotemporal pattern of MAE [mm/month] for SEAS5, MLR, and SVM precipitation forecasts in the austral summer (DJF). The precipitation data from the forecast simulations (January 2017 to December 2020) was accumulated and assessed according to the months of the forecast horizon (months 1, 3, and 7).

models do not significantly increase forecast accuracy compared to SEAS5. This can be attributed to SEAS5's ability to reproduce the teleconnection patterns that impact the NE and S regions during ENSO events. The model also exhibits a robust capacity to forecast precipitation in these areas, a finding previously confirmed by [Gubler *et al.* \(2020\)](#) and [Ferreira \(2021\)](#). This skill is thus reflected in the MLR and SVM data-driven models, as they employ ENSO-related climate indices and SEAS5 precipitation anomalies as predictors.

However, in the case of the SE, CO, and N regions of Brazil, there is a noticeable difference in the predictive performance between the data-driven models (MLR and SVM) and SEAS5. The ECMWF's dynamic system maintains the same spatial pattern of errors observed in the hindcast simulations, although it is accompanied by increased MAE values. A potential explanation for these

elevated MAE values could be that in the SE region, as well as in certain parts of the CO and N regions, surface-atmosphere interactions occurring from spring to summer exert a significant influence on the seasonal evolution of precipitation. These interactions may not be adequately captured by dynamical models. [Grimm *et al.* \(2007\)](#) uncovered a significant connection between the peak summer monsoon rainfall in Central-East Brazil and the preceding spring conditions, with a notable inverse correlation, especially during ENSO years. This relationship has been attributed to a proposed surface-atmosphere feedback mechanism that factors in spring soil moisture.

Lower rainfall in the spring leads to reduced soil moisture and higher surface temperatures by the end of the season, triggering anomalous convergence and cyclonic circulation over Southeast Brazil. This, in turn, increases the moisture flux from northern and central South America

into Central-East Brazil, contributing to heightened precipitation levels in this region. As shown in Fig. 8 (a., b., and c.) and Table 3 in the previous section, the largest increase in MAE is observed in the SE and CO regions during the early months of the forecast horizon. This may result from shifts in precipitation anomalies from spring to the subsequent summer, which are driven by regional surface-atmosphere interactions. Such interactions are inadequately represented in dynamical models, which tend to exhibit persistent anomalies from spring to summer. Data-driven models, which leverage observed precipitation from prior periods as predictors, may more effectively capture these interactions.

The escalation in forecast error observed from the first to the third month, compared to the moderate increase from the third to the seventh month, as depicted in Fig. 8 (d. through i.), suggests that certain regional processes introduce a significant error in December-February (DJF) forecasts based on initial spring conditions. However, this error does not significantly magnify with longer forecast lead times. This pattern, observed in both MLR and SVM models, highlights the critical impact of early-season conditions on short-term forecast accuracy without a proportional increase in error for forecasts extending further into the future. When it comes to data-driven models, despite their predictions showing higher error values, it is noteworthy that they consistently outperform SEAS5 across all forecast horizon months. This suggests that machine learning techniques, whether linear (such as MLR) or non-linear (such as SVM), have the potential to enhance the accuracy of seasonal precipitation forecasts in Brazil.

To further supplement the findings in Fig. 8, Fig. 9 displays the boxplot distributions of the MAE for the SEAS5, MLR, and SVM models. The boxplots depict the MAE distributions for each Brazilian region and month within the forecast horizon. Each boxplot illustrates the MAE distribution for grid cells within the Brazilian regions, whereas the seven sets of boxplots ($n = 1, 2, \dots, 6$, and 7) correspond to the respective months of the forecast horizon. The results show that the data-driven models consistently outperform SEAS5 in terms of average error across the N, SE, CO, and S regions for nearly all months of the forecast horizon, especially in the case of the CO region, where the disparity is most notable. When it comes to the S region, the difference in predictive performance between the models is less pronounced compared with the N, SE, and CO regions. However, in the NE region, the MLR model exhibits a higher average MAE and falls short of surpassing SEAS5 in terms of predictive accuracy. Considering the overall performance of these models across all regions and months, the SVM consistently yields more accurate precipitation forecasts. This outcome suggests a nonlinear relationship between the precipitation anomalies and the selected set of predictors, effectively captured by the SVM using a sigmoidal kernel function.

To complement the analysis previously presented, Fig. S1 in Supplementary Material displays the MAE calculated considering the precipitation anomaly forecasts from the SEAS5, MLR, and SVM models. Specifically for the SEAS5 model case, the anomalies were computed by subtracting the hindcast long-term mean of total precipitation from the forecasts of total precipitation. Regarding the spatial distribution of errors, maps in Fig. S1 exhibit a pattern that mirrors the one observed in the analysis of the MAE based on total precipitation data (see Fig. 8). Higher errors are observed in the SE, CO, and N regions of Brazil, while in the NE and S the error magnitude is diminished. The average MAE and bias values by region and model in Table S1 from Supplementary Material reinforce the results from the maps, and also provide new insights. When comparing models across regions, SVM forecasts are found to be more accurate in all regions and months of the forecast horizon, except for months 1, 2, and 3 in the S region, where SEAS5 outperforms the others.

According to specialized literature (Willmott, 1981; Wilks, 2011; Robeson and Willmott, 2023), MAE and other error metrics can be decomposed into systematic and unsystematic components. This decomposition allows researchers and analysts to better understand the nature of the errors generated by their models, facilitating more targeted improvements. The systematic component, denoted as MAE_s , represents the portion of the error that is consistent across different observations and reflects biases in the model's predictions, such as consistently overestimating or underestimating the actual values. Conversely, the unsystematic component, denoted as MAE_u , includes the random errors that remain after the systematic errors have been accounted for. These errors vary from one observation to another, illustrating the inherent uncertainty in the model's predictions that cannot be easily corrected through model adjustments.

The process of decomposing MAE into its systematic and unsystematic components involves statistical techniques such as regression analysis. Specifically, this decomposition employs the ordinary least squares (OLS) regression (\hat{x}) of the model predictions (x) on the observation (y). For the systematic component, the mathematical definition is given by $MAE_s = \frac{1}{T} \sum_{t=1}^T \frac{s_t}{s_t + u_t} |x_t - y_t|$, where $s_t = |x_t - y_t|$ is the difference between the OLS predictions and the actual observations. Conversely, the unsystematic component is given by $MAE_u = \frac{1}{T} \sum_{t=1}^T \frac{u_t}{s_t + u_t} |x_t - y_t|$. In this context, $u_t = |x_t - \hat{x}_t|$ quantifies the difference between the model's raw predictions and its OLS-adjusted predictions, at each time step t . Table S2 in Supplementary Material displays the outcomes of the MAE decomposition, conducted using precipitation anomaly data for each Brazilian region (SE, CO, S, NE, and N). Note that the values displayed in Table S2 are the systematic and unsystematic components of the MAE values from Table S1.

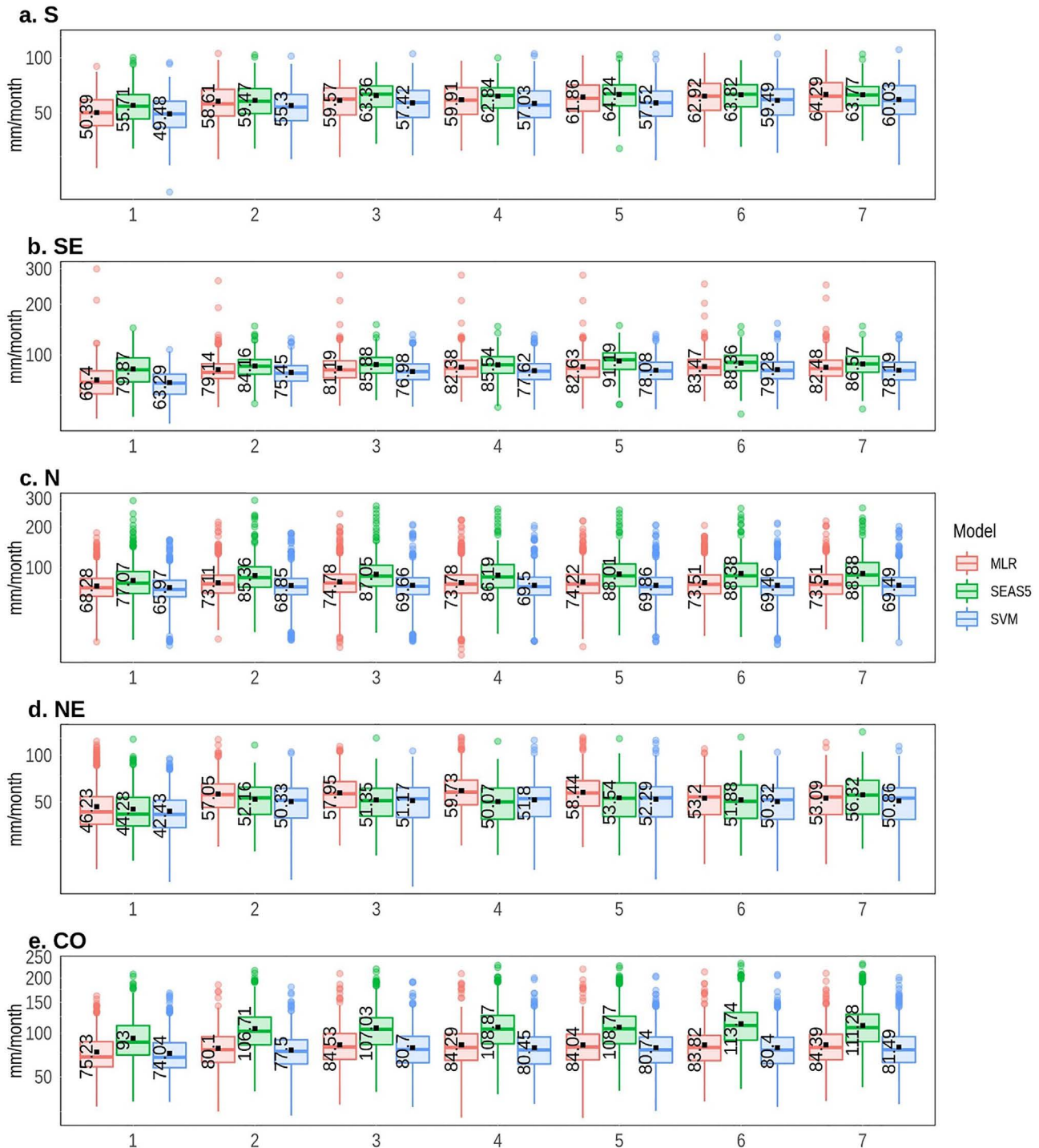


Figure 9 - Boxplot distributions of MAE [mm/month] for SEAS5, MLR, and SVM precipitation forecasts in the austral summer (DJF). Each boxplot encompasses the errors computed for the set of grid cells within the Brazilian regions and the seven sets ($n = 1, 2, \dots, 6$, and 7) correspond to the months of the forecast horizon.

These integral MAE values from Table S1 indicate that SVM forecasts exhibit higher accuracy across all regions when both systematic and unsystematic components are combined. Expanding on this point, Table S2 further reveals that, through the lens of MAE_s , the MLR model

stands out for its higher accuracy in S, SE, N, and CO regions (in the NE region SVM performs better), which means that the MLR is more proficient in reducing systematic errors. However, when it comes to unsystematic errors, MAE_u , the results change. The SVM model stands

out for providing the lowest values of MAE_u in the SE, NE, and CO regions, while the SEAS5 model outperforms in the S and N regions. This indicates that the performance of SVMs, as observed in Table S1, arises from their ability to reduce unsystematic errors.

Multiple studies have already confirmed the effectiveness of nonlinear models in addressing seasonal precipitation forecasting challenges (Xu *et al.*, 2020; Darji *et al.*, 2015; Choubin, 2016; Xu *et al.*, 2018; Fan *et al.* 2023). They excel in capturing intricate and nonlinear relationships between predictors and precipitation, thereby providing a more accurate representation of complex patterns and dynamics. The flexibility of nonlinear models allows for the incorporation of additional predictors, such as climate indices and soil moisture, enabling an adaptable and nuanced representation of the underlying processes

that drive precipitation variability. However, it is crucial to acknowledge that employing nonlinear models alone does not guarantee enhanced predictions. The selection of a suitable set of predictors that explain at least part of the precipitation variability within a region plays a critical role in the nonlinear modeling process. For this reason, in this research study, we selected climate indices that are associated with teleconnection patterns affecting rainfall in Brazil, and also explored the relationship between present and past precipitation anomalies.

The spatiotemporal pattern of bias for SEAS5, MLR, and SVM precipitation forecasts is depicted in Fig. 10. The set of maps shows that SEAS5 has an intense positive bias in the N, CO, and SE regions, indicating a systematic tendency to overestimate the precipitation during the rainy season. This finding agrees with the results of the hindcast

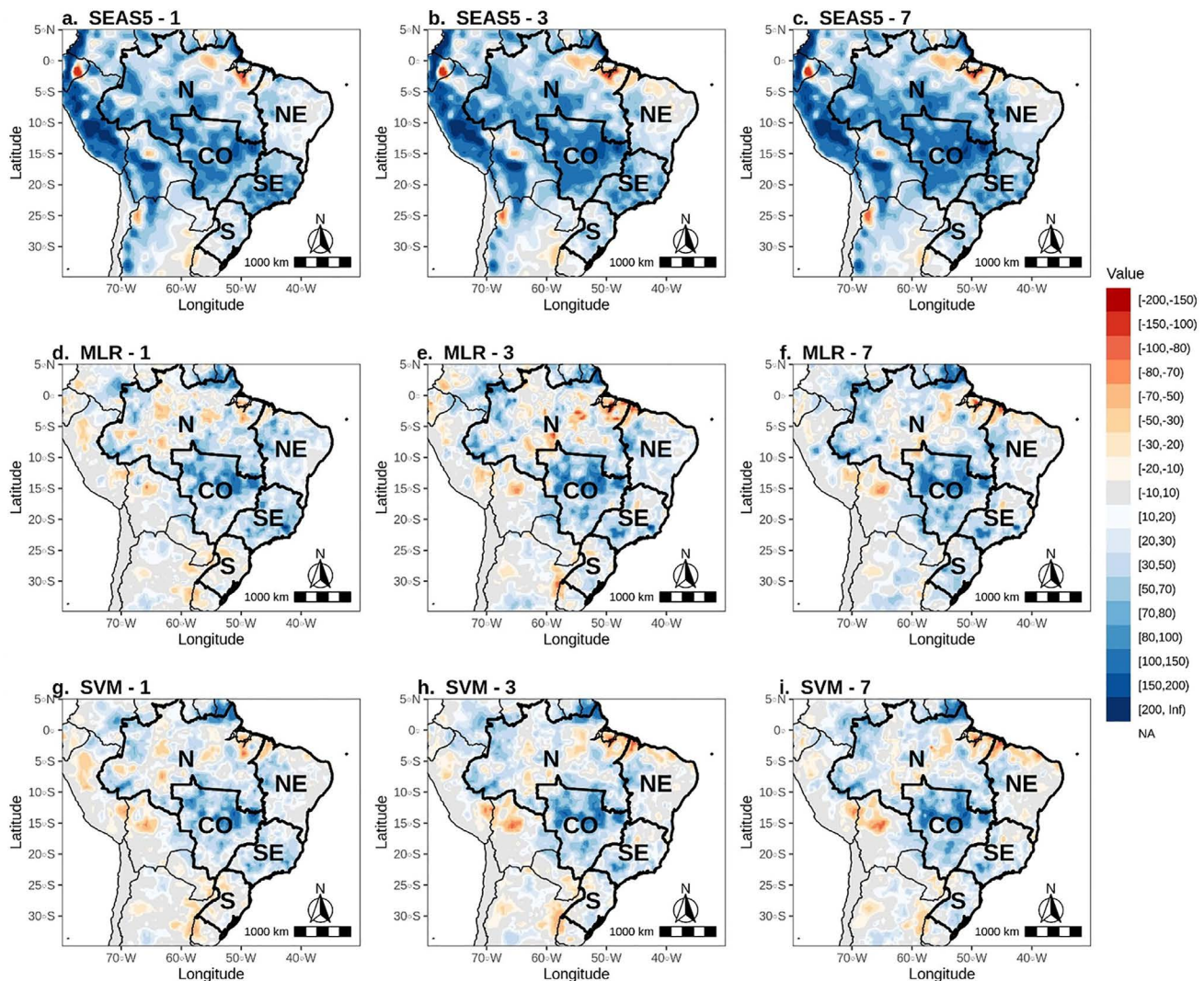


Figure 10 - Spatiotemporal pattern of bias [mm/month] for SEAS5, MLR, and SVM precipitation forecasts in the austral summer (DJF). The precipitation data from the forecast simulations (January 2017 to December 2020) was accumulated and assessed according to the months of the forecast horizon (months 1, 3, and 7).

assessment (January 1993 to December 2016) and is also supported by the evaluation conducted by [Ferreira \(2021\)](#). However, it is noteworthy that the positive bias intensifies during the forecast period (January 2017 to December 2020) in comparison to the hindcast period. The strengthened positive bias observed during the forecast period can be attributed to the nature of hindcast simulations. Hindcasts, according to [Risbey *et al.* \(2021\)](#), operate under perfect conditions compared to real-time forecasts. They enable a more comprehensive incorporation of initial condition data, along with the tuning and calibration of the model based on events that are part of the model's testing phase. These aspects mean that the skill observed in hindcasts may not accurately reflect the true skill of real-time forecasts. Hindcasts have some kind of "artificial skill," which refers to a skill that would not be attainable in a real-time forecast due to some aspect of the idealized nature of the hindcast. Thus, the higher skill of the SEAS5 hindcast dataset over the SEAS5 real-time forecast dataset is an expected result, given that the circumstances under which forecasts are conducted in a hindcast scenario are more advantageous.

In addition to the reduction in predictive accuracy between 2017 and 2020, the collection of maps from [Fig. 10](#) further shows that SEAS5 may increase the frequency or intensity of the SACZ during the DJF quarter. This may account for the high bias values detected in the N, CO, and SE regions. As previously mentioned, the influence of regional processes over central-eastern Brazil was established in studies by [Grimm and Zilli \(2009\)](#) and [Grimm *et al.* \(2007\)](#). Interactions between surface and atmosphere tied to late spring soil moisture have profound effects on the rainfall patterns during the austral summer. This leads to marked changes in the impact of teleconnection patterns, such as ENSO, in this region during the monsoon season. The authors also argue that dynamic models fall short in accurately reproducing this effect in precipitation forecasts, often indicating a continuity of rainfall anomalies from spring into summer. Unlike SEAS5, the MLR and SVM models generally show less pronounced positive biases, especially in the N region, where the bias is nearly neutral. When it comes to the NE and S regions, both the data-driven models and SEAS5 predominantly display positive biases, but less pronounced than those in the remaining regions of Brazil.

The boxplots in [Fig. 11](#) corroborate [Fig. 10](#) and reveal new information about the precipitation forecasts. The data-driven MLR and SVM models manage to reduce the average bias throughout Brazil, particularly in the SE, CO, and N regions, where the proposed models demonstrate a superior performance compared to SEAS5. For the S region, we also notice a trend of increasing positive bias throughout the forecast horizon, whereas a contrasting trend is evident in the NE region. Comparing the outcomes

of the SEAS5, MLR, and SVM models, it is clear that the SVM outperforms the other models by producing the least biased predictions for nearly all regions and months within the forecast horizon. Exceptions are observed in the N region, where the MLR model outperforms the SVM in some months.

3.3. General precipitation forecast skill of SEAS5, MLR and SVM models (2017-2020)

[Figure 12](#) shows the spatial pattern of the Diebold-Mariano test for the precipitation forecast of the SEAS5, MLR, and SVM models. This hypothesis testing process is designed to identify the regions of Brazil where predictions from a specific model are significantly more accurate than those from a competing model. The analysis takes into account a lead time of seven months ahead, with simulations initialized every month from January 2017 to December 2020.

Maps a. and b. show that both the MLR and SVM models produce more accurate precipitation forecasts than SEAS5 in the N, SE, and CO regions. This is particularly evident with the SVM model, which has more green grid cells in map b. These regions are strongly impacted by the SACZ during the austral summer. This meteorological system is responsible for much of the total precipitation observed over an extensive part of South America. For the other regions, it is worth noting that the MLR model typically delivers predictions that are less accurate than SEAS5, whereas the SVM model's precipitation forecasts are akin to those generated by the ECMWF dynamic system. Similarly, the S region in Brazil is a sector where none of the models particularly stands out. In fact, there are grid cells within the S region for which one or another model presents better predictive performance.

Regarding map c. from [Fig. 12](#), which compares the predictions of the two proposed data-driven models, we observe a distinct result. The SVM model generates significantly more accurate predictions in approximately half of the grid cells over Brazil, highlighted in green. For the remaining grid cells (colored in yellow), there is no significant difference in predictive performance between the models. Taking these observations into account, it becomes evident that the SVM model is especially suited for enhancing the accuracy of seasonal precipitation forecasts over Brazil, as it outperforms the competing models in a majority of the grid cells.

In [Fig. 13](#), the time series of the CPC precipitation analysis is presented alongside the one-month-ahead precipitation forecasts derived from MLR, SVM, and SEAS5. Each time series in the figure represents the average precipitation related to the grid cells within the N, NE, SE, CO, and S regions. The graphs indicate that the three models accurately represent the precipitation regime throughout Brazil. The N, NE, CO, and SE regions have clearly defined precipitation patterns, allowing for easy

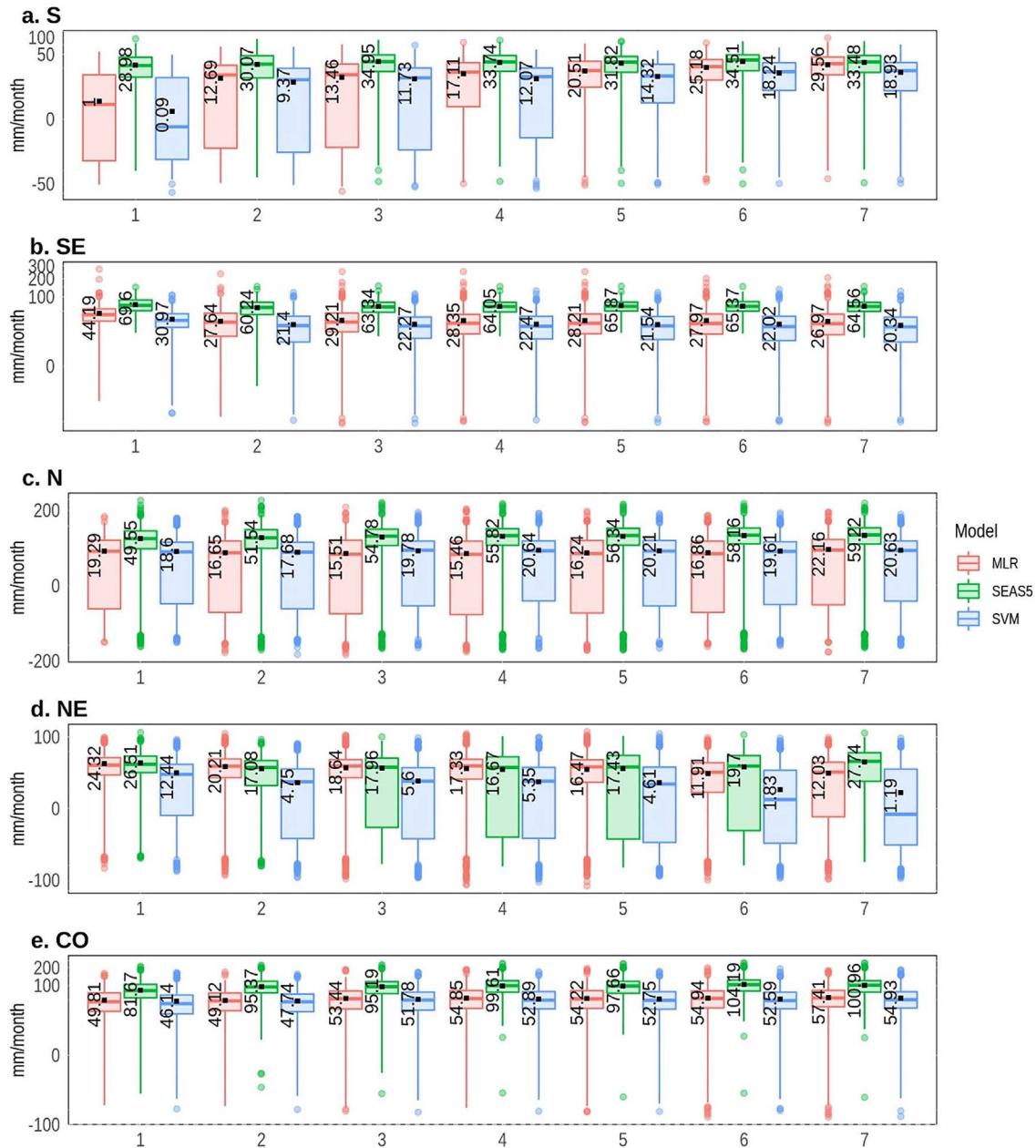


Figure 11 - Boxplot distribution of bias [mm/month] for SEAS5, MLR, and SVM precipitation forecasts in the austral summer (DJF). Each boxplot encompasses the errors calculated for the set of grid cells within the Brazilian regions, while the seven sets ($n = 1, 2, \dots, 6$, and 7) correspond to the respective months of the forecast horizon.

identification of the months corresponding to the rainy (from October to April) and dry (remaining months) periods in Brazil. In contrast, the S region exhibits a more uniform rainfall distribution throughout the year. The SEAS5 precipitation forecasts for the rainy period in the N, SE, and CO regions typically exceed the values recorded by the CPC. This result agrees with maps a., b., and c. in Fig. 10, where a strong positive bias can be observed. On the other hand, the precipitation time series from the MLR and SVM models are more adherent to the CPC precipita-

tion analysis, indicating a superior predictive performance compared to SEAS5. This trend continues into the dry period, where forecasts from all models generally align closely with the CPC data, with the exception of the SEAS5 forecasts for the N region, which tend to overestimate the precipitation. In the specific case of the S region, predicting monthly precipitation variability is a unique challenge, independent of the annual season. This difficulty highlights the complexity of forecasting in this area, stemming from the region's dependence on the

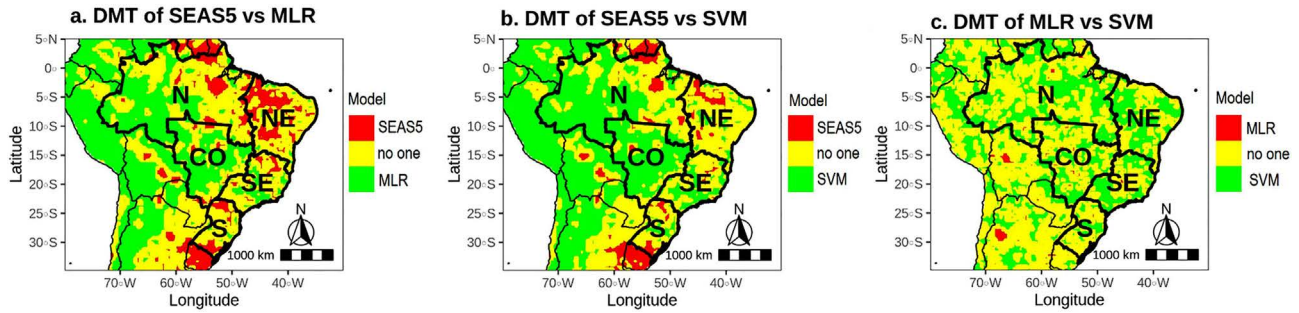


Figure 12 - Spatial pattern of results obtained through the Diebold-Mariano test considering a significance level α of 0.05, a forecast horizon of seven months ahead, and the monthly predictions made from January 2017 to December 2020. The set of maps depicts, in colors, the sectors of Brazil for which the predictions made by a specific model are significantly more accurate than those made by the competing model.

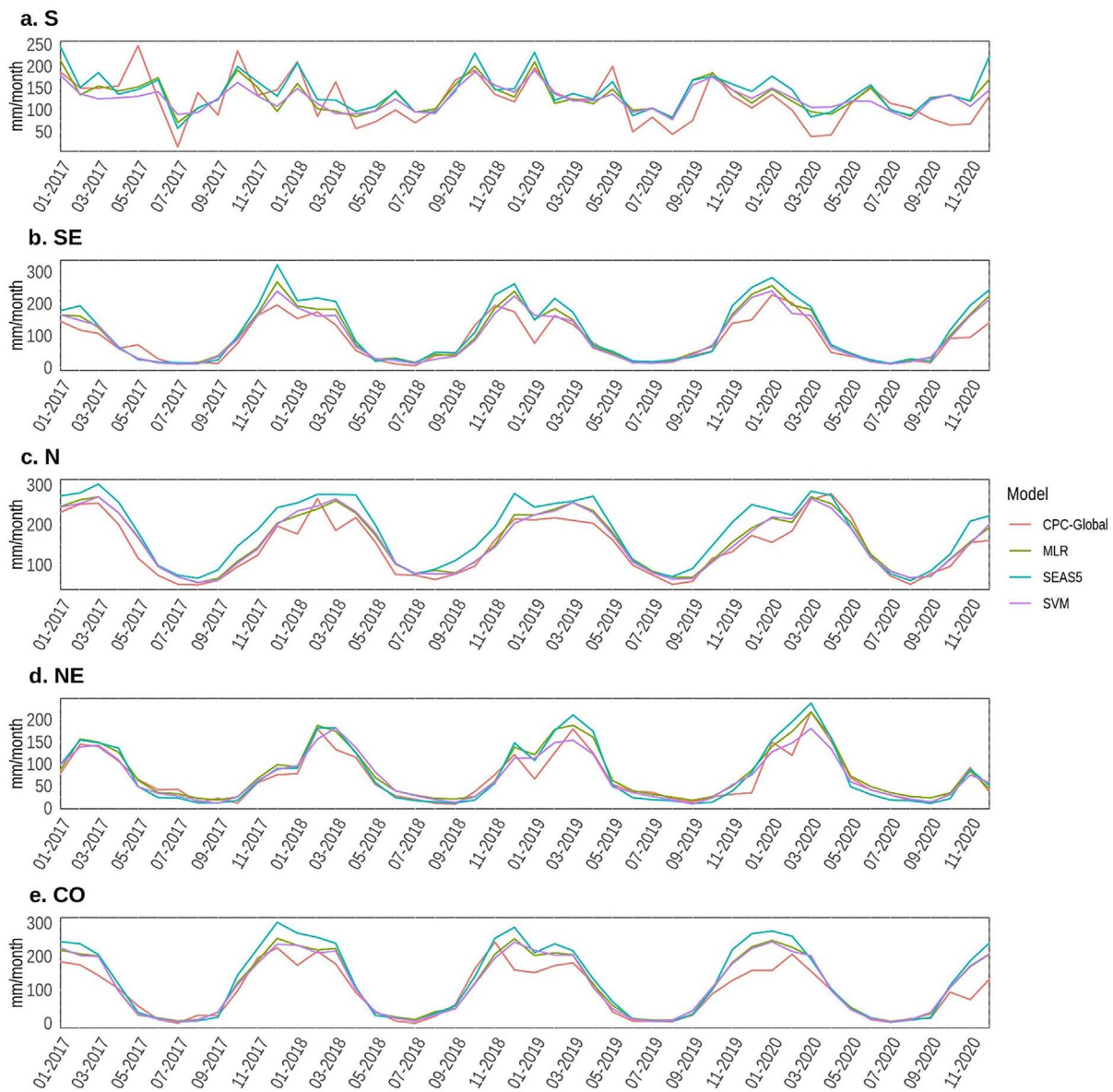


Figure 13 - Time series of precipitation forecasts for January 2017 to December 2020 from MLR, SVM, and SEAS5 models. The average precipitation for each Brazilian region is computed considering the results for the first month of the forecast horizon.

occurrence of high-frequency transients, as noted by [Reboita et al. \(2010\)](#).

The results demonstrate that, in general, data-driven models refine precipitation forecasts in Brazil, especially in the N, CO, and SE regions. Despite the advantage of having equations with physical interpretation, numerical dynamical models like SEAS5 are affected by uncertainties related to initial conditions, parameters, deficiency of model physics, and model structure. On the other hand, models based on MLR and SVM, which are also affected by several sources of uncertainty, have a crucial advantage over numerical dynamical models: the ability to identify relationships between inputs and outputs and use them with new input data to make predictions. This characteristic of data-driven models provides an even greater advantage when explored using a set of predictors whose relationships are physically sustained, such as in the case of climate indices related to teleconnection patterns affecting precipitation in Brazil ([Reboita et al., 2021](#)), and other important variables not yet considered in our study, such as soil moisture ([Grimm et al., 2007](#); [Grimm and Zilli, 2009](#)). Regarding the assessed models, the results reveal that the model based on SVM generates more accurate predictions than MLR, which suggests the existence of nonlinear relationships between the precipitation and the selected predictors. In fact, the linear model suffers from major disadvantages, as mentioned before by [Fan et al. \(2023\)](#), such as the inability to map nonlinear dependencies between predictors and the variable that is being predicted, which results in the loss of the nonlinear components of this relationship in the modeling process. Conversely, data-driven models based on nonlinear approaches, such as the SVM implemented here, do not suffer from such a deficiency and are consequently able to provide better predictions. Therefore, in future research, emphasis should be placed on investigating other nonlinear data-driven approaches, as well as optimizing the pre-processing of predictors to effectively explore the relationship between dependent and independent variables.

4. Conclusion

In this study, we present a MLR- and SVM-based framework designed to refine seasonal rainfall forecasts in Brazil, employing climate indices and precipitation anomalies from CPC and SEAS5 as predictors. Unlike dynamic models, in which equations encapsulate physical processes, models grounded in machine learning employ mathematical expressions that do not provide a direct interpretation of physical phenomena. Data-driven models extract information from a set of predictors, learn the relationship between them and the forecast variable, and then apply this learned relationship to generate forecasts using novel input data. Consequently, their ability to make a

prediction that meets physical expectations depends on both the correct choice of predictors and the utilization of a suitable mathematical structure to model the relationships.

Considering the aforementioned issues, the data-driven models developed in this study were designed to explore a base of predictors that varies spatially, depending on the relevance of the relationship with the precipitation anomalies observed for each grid cell in the study area. To illustrate, a model specifically tailored for a particular grid cell in the Northeast region of Brazil incorporates predictors capable of explaining at least part of the rainfall variability in that sector. It assimilates information intrinsic to climatic teleconnection patterns (e.g., the influence of ENSO on the precipitation regime of Northeast Brazil) as well as local processes, subsequently applying this acquired knowledge to generate predictions from a new set of input data.

The findings presented in Section 3 of this study demonstrate that the models introduced herein enhance the accuracy of precipitation forecasts for some grid cells. This enhancement is particularly notable in several Brazilian regions, including the North (N), Southeast (SE), and Central-West (CO). These regions, which are notably influenced by the SACZ, a primary atmospheric system that triggers precipitation during the austral summer, exhibit a persistent positive bias. This result suggests that the SEAS5, MLR, and SVM models may tend to increase the frequency or intensity of the SACZ. A potential explanation for this could be the regional surface-atmosphere interactions related to soil moisture in late spring, which influence the precipitation regime (including the SACZ) in central-eastern Brazil (SE and parts of CO) ([Grimm et al., 2007](#); [Grimm and Zilli, 2009](#)). This regional aspect may not be adequately represented in the models, notably in the case of SEAS5, which displays a distinct positive bias.

The performance of the models, compared to the other regions, particularly stands out in the NE and S regions. In the DJF quarter, the accuracy of precipitation forecasts made by SEAS5 and the data-driven models does not show a significant disparity. However, the SVM model outperforms SEAS5 in both cases. It is important to note that precipitation in these sectors is strongly associated with the ENSO phases ([Reboita et al., 2021](#)), a critical climate driver that impacts Brazil's precipitation regime. These findings suggest that SEAS5 effectively encapsulates the impact of these climate forcings on the precipitation forecasts. This characteristic is also reflected in the MLR and SVM models, as they use SEAS5 data and climate indices related to ENSO as predictors.

When the predictions from SEAS5 and the data-driven models are assessed and compared, we observe that SVM makes more accurate and less biased predictions than the remaining models for the austral summer. This result suggests the existence of nonlinear relationships

between the predictors and precipitation anomalies, which the MLR model cannot capture. Furthermore, upon conducting a more stringent statistical analysis using the Diebold-Mariano hypothesis test ($\alpha = 0.05$) across predictions made for all seasons, it is confirmed once again that SVM's precipitation forecasts are either significantly more accurate or on par with those generated by SEAS5 for the SE, CO, and N regions. However, it is critical to note that SVM's performance is also a function of the selected predictors, as they account for a substantial portion of the precipitation variability.

In future research, the primary goal is to refine precipitation forecasts for Brazil by exploring advanced techniques for ensemble prediction generation, input variable selection, as well as time-series decomposition (such as wavelet multiresolution analysis). A key focus will be on probabilistic forecasting to better represent the uncertainties in precipitation forecasts. One approach we are considering is Bayesian model averaging, which is a sophisticated approach to explore different data-driven models (including models beyond SVMs and MLR, like Ridge regression and random forest) based on their posteriori probability distributions.

Acknowledgements

The authors would like to thank the Energisa Company and the National Council for Scientific and Technological Development for their funding and support of the activities developed under the Research and Development Project "Applying Neural Networks to Forecast Inflow and Demand for Short-Term Electricity Price Formation," R&D PD -06585-RT-01.

References

- ALI, M.; DEO, R.C.; XIANG, Y.; LI, Y.; YASEEN, Z.M. Forecasting long-term precipitation for water resource management: A new multi-step data-intelligent modelling approach. **Hydrological Sciences Journal**, v. 65, n. 16, p. 2693-2708, 2020. [doi](#)
- ALMEIDA, V.A.; MARTON, E.; NUNES, A.M.B. Assessing the ability of three global reanalysis products to reproduce South American monsoon precipitation. **Atmosfera**, v. 31, n. 1, p. 1-10, 2018. [doi](#)
- ANOCHI, J.A.; VELHO, H.F.C. Climate precipitation prediction for South region by neural network self-configured. **Ciência e Natura**, v. 38, p. 98-104, 2016. [doi](#)
- ANOCHI, J.A.; ALMEIDA, V.A.; VELHO, H.F.C. Machine learning for climate precipitation prediction modeling over South America. **Remote Sensing**, v. 13, n. 13, p. 2468, 2021. [doi](#)
- BADR, H.S.; ZAITCHIK, B.F.; GUIKEMA, S.D. Application of statistical models to the prediction of seasonal rainfall anomalies over the Sahel. **Journal of Applied Meteorology and Climatology**, v. 53, n. 3, p. 614-636, 2014. [doi](#)
- BHANDARI, S.; THAKUR, B.; KALRA, A.; MILLER, W.P.; LAKSHMI, V.; *et al.* Streamflow forecasting using singular value decomposition and support vector machine for the upper Rio Grande river basin. **Journal of the American Water Resources Association**, v. 55, n. 3, p. 680-699, 2019. [doi](#)
- BRUNNER, M.I.; SLATER, L.; TALLAKSEN, L.M.; CLARK, M. Challenges in modeling and predicting floods and droughts: A review. **WIREs Water**, v. 8, n. 3, e1520, 2021. [doi](#)
- CARBONE, G.J.; DOW, K. Water resource management and drought forecasts in South Carolina. **Journal of the American Water Resources Association**, v. 41, n. 1, p. 145-155, 2005. [doi](#)
- CARVALHO, L.M.V.; JONES, C.; POSADAS, A.N.D.; QUIROZ, R.; BOOKHAGEN, B.; *et al.* Precipitation characteristics of the South American monsoon system derived from multiple datasets. **Journal of Climate**, v. 25, n. 13, p. 4600-4620, 2012. [doi](#)
- CHAVEZ, E.; CONWAY, G.; GHIL, M.; SADLER, M. An end-to-end assessment of extreme weather impacts on food security. **Nature Climate Change**, v. 5, n. 11, p. 997-1001, 2015. [doi](#)
- CHEN, M.; SHI, W.; XIE, P.; SILVA, V.B.S.; KOUSKY, V.E.; *et al.* Assessing objective techniques for gauge-based analyses of global daily precipitation. **Journal of Geophysical Research**, v. 113, n. D4, p. 1-13, 2008. [doi](#)
- CHOUBIN, B.; KHALIGHI-SIGAROODI, S.; MALEKIAN, A.; KIŞI, Ö. Multiple linear regression, multi-layer perceptron network and adaptive neuro-fuzzy inference system for forecasting precipitation based on large-scale climate signals. **Hydrological Sciences Journal**, v. 61, n. 6, p. 1001-1009, 2016. [doi](#)
- CHOUBIN, B.; ZEHTABIAN, G.; AZAREH, A.; RAFIEI-SARDOODI, E.; SAJEDI-HOSSEINI, F.; *et al.* Precipitation forecasting using classification and regression trees (CART) model: A comparative study of different approaches. **Environmental Earth Sciences**, v. 77, n. 8, 314, 2018. [doi](#)
- COPERNICUS. **Seasonal Forecast Daily and Subdaily Data on Single Levels**. Copernicus Climate Data Store, 2021. Available from <https://cds.climate.copernicus.eu/cdsapp#!/home>, accessed on February 17, 2023.
- COELHO, C.A.S.; STEPHENSON, D.B.; BALMASEDA, M.; DOBLAS-REYES, F.J.; VAN OLDENBORGH, G.J. Toward an Integrated Seasonal Forecasting System for South America. **Journal of Climate**, v. 19, n. 15, p. 3704-3721, 2006. [doi](#)
- CÓRDOBA-MACHADO, S.; PALOMINO-LEMUS, R.; GÁMIZ-FORTIS, S.R.; CASTRO-DÍEZ, Y.; ESTEBAN-PARRA, M.J. Influence of Tropical Pacific SST on seasonal precipitation in Colombia: Prediction using El Niño and El Niño Modoki. **Climate Dynamics**, v. 44, n. 5, p. 1293-1310, 2015. [doi](#)
- CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273-297, 1995. [doi](#)
- DABERNIG, M.; MAYR, G.J.; MESSNER, J.W.; ZEILEIS, A. Spatial ensemble post-processing with standardized anomalies. **Quarterly Journal of the Royal Meteorological Society**, v. 143, n. 703, p. 909-916, 2017. [doi](#)

- DARJI, M.P.; DABHI, V.K.; PRAJAPATI, H.B. Rainfall forecasting using neural network: a survey. In: **Proceedings International Conference on Advances in Computer Engineering and Applications**, Ghaziabad, p. 706-713, 2015. [doi](#)
- DIAS, T.L.; CATALDI, M.; FERREIRA, V.H. Application of neural network techniques and atmospheric modeling to prepare streamflow forecasts in the Rio Grande basin (MG). **Engenharia Sanitaria e Ambiental**, v. 22, n. 1, p. 169-178, 2017. [doi](#)
- DIRO, G.T.; BLACK, E.; GRIMES, D.I.F. Seasonal forecasting of Ethiopian spring rains. **Meteorological Applications**, v. 15, n. 1, p. 73-83, 2008. [doi](#)
- DIEBOLD, F.X.; MARIANO, R.S. Comparing predictive accuracy. **Journal of Business and Economic Statistics**, v. 13, n. 3, p. 253-263, 1995. [doi](#)
- ECMWF. **SEAS5 User Guide**. 2017. Available at https://www.ecmwf.int/sites/default/files/medialibrary/2017-10/System5_guide.pdf, accessed on Jan. 29, 2023.
- ENFIELD, D.B.; MESTAS-NUÑEZ, A.M.; MAYER, D.A.; CID-SERRANO, L. How ubiquitous is the dipole relationship in tropical Atlantic sea surface temperatures? **Journal of Geophysical Research: Oceans**, v. 104, n. C4, p. 7841-7848, 1999. [doi](#)
- EVANS, J.D. **Straightforward Statistics for the Behavioral Sciences**. Pacific Grove: Brooks/Cole Publishing, p. 145-147, 1996.
- FAN, F.M.; COLLISCHONN, W.; MELLER, A.; BOTELHO, L.C.M. Ensemble streamflow forecasting experiments in a tropical basin: The São Francisco River case study. **Journal of Hydrology**, v. 519, n. D, p. 2906-2919, 2014. [doi](#)
- FAN, Y.; KRASNOPOLSKY, V.; VAN DEN DOOL, H.; WU, C.; GOTTSCHALCK, J. Using artificial neural networks to improve CFS Week-3-4 precipitation and 2-m air temperature forecasts. **Weather and Forecasting**, v. 38, n. 5, p. 637-654, 2023. [doi](#)
- FERREIRA, G.W.S. **Validation of Seasonal Climate Predictions for South America: ECMWF-SEAS5 Global Model**. Dissertação de Mestrado, Instituto de Recursos Naturais, Universidade Federal de Itajubá, Itajubá, 2021.
- FOLLAND, C.K.; COLMAN, A.W.; ROWELL, D.P.; DAVEY, M.K. Predictability of northeast Brazil rainfall and real-time forecast skill, 1987-98. **Journal of Climate**, v. 14, n. 9, p. 1937-1958, 2001. [doi](#)
- GERLITZ, L.; VOROGUSHYN, S.; APEL, H.; GAFUROV, A.; UNGER-SHAYESTEH, K.; *et al.* A statistically based seasonal precipitation forecast model with automatic predictor selection and its application to central and south Asia. **Hydrology and Earth System Sciences**, v. 20, n. 11, p. 4605-4623, 2016. [doi](#)
- GIBSON, P.B.; CHAPMAN, W.E.; ALTINOK, A.; DELLE MONACHE, L.; DEFLORIO, M.J.; *et al.* Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts. **Communications Earth and Environment**, v. 2, n. 1, p. 1-13, 2021. [doi](#)
- GLANTZ, M.H.; RAMIREZ, I.J. Reviewing the Oceanic Niño Index (ONI) to enhance societal readiness for El Niño's impacts. **International Journal of Disaster Risk Science**, v. 11, n. 11, p. 394-403, 2020. [doi](#)
- GODDARD, L.; MASON, S.J.; ZEBIAK, S.E.; ROPELEWSKI, C.F.; BASHER, R.; *et al.* Current approaches to seasonal-to-interannual climate predictions. **International Journal of Climatology**, v. 21, n. 9, p. 1111-1152, 2001. [doi](#)
- GRIMM, A.M.; DIAS, P.L.S. Analysis of tropical-extratropical interactions with influence functions of a barotropic model. **Journal of the Atmospheric Sciences**, v. 52, n. 20, p. 3538-3555, 1995. [doi](#)
- GRIMM, A.M.; PAL, J.S.; GIORGI, F. Connection between spring conditions and peak summer monsoon rainfall in South America: Role of soil moisture, surface temperature, and topography in eastern Brazil. **Journal of Climate**, v. 20, n. 24, p. 5929-5945, 2007. [doi](#)
- GRIMM, A.M.; ZILLI, M.T. Interannual variability and seasonal evolution of summer monsoon rainfall in South America. **Journal of Climate**, v. 22, n. 9, p. 2257-2275, 2009. [doi](#)
- GRIMM, A.M. Interannual climate variability in South America: Impacts on seasonal precipitation, extreme events, and possible effects of climate change. **Stochastic Environmental Research and Risk Assessment**, v. 25, n. 4, p. 537-554, 2011. [doi](#)
- GRIMM, A.M. South American monsoon and its extremes. In: VENUGOPAL, V.; SUKHATME, J.; MURTUGUDDE, R.; ROCA, R. **Tropical Extremes: Natural Variability and Trends**. Amsterdam: Elsevier, p. 51-93, 2019. [doi](#)
- GRIMM, A.M.; DOMINGUEZ, F.; CAVALCANTI, I.F.A.; CAVAZOS, T.; GAN, M.A.; *et al.* South and North American monsoons: characteristics, life cycle, variability, modelling and prediction. In: **The Multiscale Global Monsoon System. Vol. 11**. Singapore: World Scientific Publishing Company, p. 49-66, 2021. [doi](#)
- GUBLER, S.; SEDLMEIER, K.; BHEND, J.; AVALOS, G.; COELHO, C.; *et al.* Assessment of ECMWF SEAS5 seasonal forecast performance over South America. **Weather and Forecasting**, v. 35, n. 2, p. 561-584, 2020. [doi](#)
- GUPTA, Y.; SARASWAT, A. Machine learning techniques for short-term forecasting of wind power generation. In: **Proceedings International Conference on Advanced Machine Learning Technologies and Applications**, Singapore, p. 439-448, 2020. [doi](#)
- HAGEDORN, R.; DOBLAS-REYES, F.J.; PALMER, T.N. The rationale behind the success of multi-model ensembles in seasonal forecasting - I. basic concept. **Tellus, Series A: Dynamic Meteorology and Oceanography**, v. 57, n. 3, p. 219-233, 2005. [doi](#)
- HIRATA, F.E.; GRIMM, A.M. Extended-range prediction of South Atlantic convergence zone rainfall with calibrated CFSv2 reforecast. **Climate Dynamics**, v. 50, n. 9, p. 3699-3710, 2018. [doi](#)
- HOCKING, R.R. A biometrics invited paper. The analysis and selection of variables in linear regression. **Biometrics**, v. 32, n. 1, p. 1-49, 1976. [doi](#)
- HUANG, B.; THORNE, P.; BANZON, V.; BOYER, T.; CHEPURIN, G.; *et al.* Extended reconstructed sea surface temperature, version 5 (ERSSTv5): Upgrades, validations, and intercomparisons. **Journal of Climate**, v. 30, n. 20, p. 8179-8205, 2017. [doi](#)
- INGRAM, K.T.; RONCOLI, M.C.; KIRSHEN, P.H. Opportunities and constraints for farmers of west Africa to use season-

- nal precipitation forecasts with Burkina Faso as a case study. **Agricultural Systems**, v. 74, n. 3, p. 331-349, 2002. doi
- JESUS, E.M.; ROCHA, R.P.; CRESPO, N.M.; REBOITA, M.S.; GOZZO, L.F. Multi-model climate projections of the main cyclogenesis hot-spots and associated winds over the eastern coast of South America. **Climate Dynamics**, v. 56, n. 1, p. 537-557, 2021. doi
- JOZAGHI, A.; SHEN, H.; GHAZVINIAN, M.; SEO, D.J.; ZHANG, Y.; *et al.* Multi-model streamflow prediction using conditional bias-penalized multiple linear regression. **Stochastic Environmental Research and Risk Assessment**, v. 35, n. 11, p. 2355-2373, 2021. doi
- JOHNSON, S.J.; STOCKDALE, T.N.; FERRANTI, L.; BALMASEDA, M.A.; MOLTENI, F.; *et al.* SEAS5: The new ECMWF seasonal forecast system. **Geoscientific Model Development**, v. 12, n. 3, p. 1087-1117, 2019. doi
- KIDSON, J.W. Principal modes of Southern Hemisphere low-frequency variability obtained from NCEP-NCAR reanalyses. **Journal of Climate**, v. 12, n. 9, p. 2808-2830, 1999. doi
- LI, W.G.; SÁ, L.D.A.; PRASAD, G.S.S.D.; NOWOSAD, A.G.; BOLZAN, M.J.A.; *et al.* Neural network adaptive wavelets for predictions of the Northeastern Brazil monthly rainfall anomalies time series. In: **Proceedings II Applications and Science of Artificial Neural Networks**, Orlando, p. 175-187, 1996. doi
- LIPPER, L.; THORNTON, P.; CAMPBELL, B.; BAEDEKER, T.; BRAIMOH, A.; *et al.* Climate-smart agriculture for food security. **Nature Climate Change**, v. 4, n. 12, p. 1068-1072, 2014. doi
- LIU, Z.; ALEXANDER, M. Atmospheric bridge, oceanic tunnel, and global climatic teleconnections. **Reviews of Geophysics**, v. 45, n. 2, p. 1-34, 2007. doi
- LOPEZ, P. Cloud and precipitation parameterizations in modeling and variational data assimilation: A review. **Journal of the Atmospheric Sciences**, v. 64, n. 11, p. 3766-3784, 2007. doi
- MANGASARIAN, O.L. **Nonlinear Programming**. New York: Society for Industrial and Applied Mathematics, 1994.
- MILLÉO, C.; ALMEIDA, R.C. Application of RBF artificial neural networks to precipitation and temperature forecasting in Paraná, Brazil. **Ciência e Natura**, v. 43, e40, 2021. doi
- MILLER, S.J. Chapter 24: The method of least squares. In: **The Probability Lifesaver: All the Tools You Need to Understand Chance**. Princeton: Princeton University Press, p. 625-635, 2017.
- MO, K.C.; HIGGINS, R.W. The Pacific-South American modes and tropical convection during the Southern Hemisphere winter. **Monthly Weather Review**, v. 126, n. 6, p. 1581-1596, 1998. doi
- MO, K.C. Relationships between low-frequency variability in the Southern Hemisphere and sea surface temperature anomalies. **Journal of Climate**, v. 13, n. 20, p. 3599-3610, 2000. doi
- MOLTENI, F.; BUIZZA, R.; PALMER, T.N.; PETROLIAGIS, T. The ECMWF ensemble prediction system: Methodology and validation. **Quarterly Journal of the Royal Meteorological Society**, v. 122, n. 529, 1996. doi
- MONEGO, V.S.; ANOCHI, J.A.; VELHO, H.F.C. South America seasonal precipitation prediction by gradient-boosting machine-learning approach. **Atmosphere**, v. 13, n. 2, 243, 2022. doi
- MONTGOMERY, D.C.; JENNINGS, C.L.; KULAHCI, M. **Introduction to Time Series Analysis and Forecasting**. 2. ed. Hoboken: John Wiley & Sons, 2008.
- MORIOKA, Y.; TOZUKA, T.; YAMAGATA, T. On the growth and decay of the subtropical dipole mode in the South Atlantic. **Journal of Climate**, v. 24, n. 21, p. 5538-5554, 2011. doi
- MORADI, A.M.; DARIANE, A.B.; YANG, G.; BLOCK, P. Long-range reservoir inflow forecasts using large-scale climate predictors. **International Journal of Climatology**, v. 40, n. 13, p. 5429-5450, 2020. doi
- NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION. **Gauge-Based Analysis of Global Daily Precipitation**. 2023. Available at https://ftp.cpc.ncep.noaa.gov/precip/CPC_UNI_PRCP/GAUGE_GLB/, accessed on January 29, 2023.
- NNAMCHI, H.C.; LI, J.; ANYADIKE, R.N.C. Does a dipole mode really exist in the South Atlantic Ocean? **Journal of Geophysical Research Atmospheres**, v. 116, n. D15, p. 1-15, 2011. doi
- NOBRE, P.; SHUKLA, J. Variations of sea surface temperature, wind stress, and rainfall over the Tropical Atlantic and South America. **Journal of Climate**, v. 9, n. 10, p. 2464-2479, 1996. doi
- OLIVEIRA, A.S. **Interactions Between Frontal Systems in South America and Convection in Amazonia**. Dissertação de Mestrado, Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 1986.
- PAL, R. **Predictive Modeling of Drug Sensitivity**. Amsterdam: Elsevier, 2016.
- PALMER, T.N. Predicting uncertainty in forecasts of weather and climate. **Reports on Progress in Physics**, v. 63, n. 2, p. 71-116, 2000. doi
- PALMER, T.N.; SHUTTS, G.J.; HAGEDORN, R.; DOBLASREYES, F.J.; JUNG, T.; *et al.* Representing model uncertainty in weather and climate prediction. **Annual Review of Earth and Planetary Sciences**, v. 33, 2005. doi
- PAZ, A.R.; UVO, C.B.; BRAVO, J.M.; COLLISCHONN, W.; ROCHA, H.R. Seasonal precipitation forecast based on artificial neural networks. In: **Computational Methods for Agricultural Research: Advances and Applications**. Hershey: IGI Global, p. 326-354, 2010. doi
- PEZZI, L.P.; UBARANA, V.; REPELLI, C. Performance and predictions of a regional statistical model for southern Brazil. **Revista Brasileira de Geofísica**, v. 18, n. 2, p. 128-144, 2000. doi
- PINHEIRO, E.; OUARDA, T.B.M.J. Short-lead seasonal precipitation forecast in northeastern Brazil using an ensemble of artificial neural networks. **Scientific Reports**, v. 13, n. 1, 2023. doi
- POPE, P.T.; WEBSTER, J.T. The use of an F-statistic in stepwise regression procedures. **Technometrics**, v. 14, n. 2, p. 327-340, 1972. doi
- PONTES, P.R.M.; FAN, F.M.; COLLISCHONN, W.; PAIVA, R.C.D. Sensitivity analysis of the Paraná River streamflow

- to potential precipitation change. In: **Proceedings XX Simpósio Brasileiro de Recursos Hídricos**, Bento Gonçalves, p. 1-8, 2013.
- QUADRO, M.F.L.; MACHADO, L.H.R.; CALBETE, S.; BATISTA, N.N.M.; OLIVEIRA, G.S. Precipitation and temperature climatology. **Climanálise - Bulletin of Climate Monitoring and Analysis**, edição especial comemorativa de 10 anos, art. 9, 1996. Available at <http://climanalise.cptec.inpe.br/~rcliman/boletim/cliesp10a/9.html>, accessed on February 25, 2023.
- QUAN, X.; HOERLING, M.; WHITAKER, J.; BATES, G.; XU, T. Diagnosing sources of U.S. seasonal forecast skill. **Journal of Climate**, v. 19, n. 13, p. 3279-3293, 2006. doi
- RASOULI, K.; HSIEH, W.W.; CANNON, A.J. Daily streamflow forecasting by machine learning methods with weather and climate inputs. **Journal of Hydrology**, v. 414-415, p. 284-293, 2012. doi
- REBOITA, M.S.; AMBRIZZI, T.; CRESPO, N.M.; DUTRA, L.M.M.; FERREIRA, G.W.S.; *et al.* Impacts of teleconnection patterns on South America climate. **Annals of the New York Academy of Sciences**, v. 1504, n. 1, p. 116-153, 2021. doi
- REBOITA, M.S.; GAN, M.A.; ROCHA, R.P.; AMBRIZZI, T. Precipitation regimes in South America: A literature review. **Revista Brasileira de Meteorologia**, v. 25, n. 2, p. 185-204, 2010. doi
- REBOITA, M.S.; ROCHA, R.P.; AMBRIZZI, T.; SUGAHARA, S. South Atlantic Ocean cyclogenesis climatology simulated by regional climate model (RegCM3). **Climate Dynamics**, v. 35, n. 7, p. 1331-1347, 2010. doi
- REBOITA, M.S.; AMBRIZZI, T.; SILVA, B.A.; PINHEIRO, R.F.; ROCHA, R.P. The South Atlantic subtropical anticyclone: present and future climate. **Frontiers in Earth Science**, v. 7, p. 1-15, 2019. doi
- RISBEY, J.S.; SQUIRE, D.T.; BLACK, A.S.; DELSOLE, T.; LEPORÉ, C.; *et al.* Standard assessments of climate forecast skill can be misleading. **Nature Communications**, v. 12, n. 1, 4346, 2021. doi
- ROBESON, S.M.; WILLMOTT, C.J. Decomposition of the mean absolute error (MAE) into systematic and unsystematic components. **PLoS ONE**, v. 18, n. 2, e0279774, 2023. doi
- ROPELEWSKI, C.F.; JONES, P.D. An Extension of the Tahiti-Darwin Southern Oscillation Index. **Monthly Weather Review**, v. 115, n. 9, p. 2161-2165, 1987. doi
- SAJI, N.; GOSWAMI, B.; VINAYACHANDRAN, P.; YAMAGATA, T. A dipole mode in the tropical Indian Ocean. **Nature**, v. 401, n. 6751, p. 360-363, 1999. doi
- SAJI, N.H.; YAMAGATA, T. Possible impacts of Indian Ocean Dipole mode events on global climate. **Climate Research**, v. 25, n. 2, p. 151-169, 2003. doi
- SACCO, M.A.L. **Atmospheric Teleconnections and Numerical Weather Forecast in South America**. Dissertação de Mestrado, Departamento de Ciências Atmosféricas, Universidade Federal de São Paulo, São Paulo, 106 p., 2010.
- SEIBERT, M.; MERZ, B.; APEL, H. Seasonal forecasting of hydrological drought in the Limpopo Basin: A comparison of statistical methods. **Hydrology and Earth System Sciences**, v. 21, n. 3, p. 1611-1629, 2017. doi
- SIVAKUMAR, M.V.K.; MOTHA, R.P. **Managing Weather and Climate Risks in Agriculture**. 1. ed. Berlin: Springer, 2007. doi
- SILVA, V.B.S.; KOUSKY, V.E.; HIGGINS, R.W. Daily precipitation statistics for South America: An intercomparison between NCEP reanalyses and observations. **Journal of Hydrometeorology**, v. 12, n. 1, p. 101-117, 2011. doi
- SLINGO, J.; PALMER, T. Uncertainty in weather and climate prediction. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, v. 369, n. 1956, p. 4751-4767, 2011. doi
- SMITH, K. The influence of weather and climate on recreation and tourism. **Weather**, v. 48, n. 12, p. 398-404, 1993. doi
- SCHWEIGHOFER, J. The impact of extreme weather and climate change on inland waterway transport. **Natural Hazards**, v. 72, n. 1, p. 23-40, 2014. doi
- SHANNON, H.D.; MOTHA, R.P. Managing weather and climate risks to agriculture in North America, Central America and the Caribbean. **Weather and Climate Extremes**, v. 10, p. 50-56, 2015. doi
- SUN, Q.; MIAO, C.; DUAN, Q.; ASHOURI, H.; SOROOSHIAN, S.; *et al.* A review of global precipitation data sets: Data sources, estimation, and intercomparisons. **Reviews of Geophysics**, v. 56, n. 1, p. 79-107, 2018. doi
- STOCKDALE, T.; JOHNSON, S.; FERRANTI, L.; BALMA-SEDA, M.; BRICEAG, S. ECMWF's new long-range forecasting system SEAS5. **ECMWF Newsletter 154 - Winter 2017/18**, p. 15-20, 2018. doi
- SHI, W. **Frequently Asked Questions Regarding CPC's Current Monthly Atmospheric and SST Index Values**. 2007. Available at <https://www.cpc.ncep.noaa.gov/data/indices/Readme.index.shtml#SOICALC>, accessed on February 13, 2023.
- SMITH, C.A.; SARDESHMUKH, P.D. The effect of ENSO on the intraseasonal variance of surface temperatures in winter. **International Journal of Climatology**, v. 20, n. 13, p. 1543-1557, 2000. doi
- SOUZA, C.A.; REBOITA, M.S. Tool for monitoring teleconnection patterns in South America. **Terrae Didactica**, v. 17, n. 0, e021009, 2021. doi
- SMOLA, A.J.; SCHÖLKOPF, B. A tutorial on support vector regression. **Statistics and Computing**, v. 14, n. 3, p. 199-222, 2004. doi
- TOTH, Z.; BUZZA, R. Chapter 2 - Weather Forecasting: What Sets the Forecast Skill Horizon? In: **Sub-Seasonal to Seasonal Prediction**. Elsevier, 2019. p. 17-45. doi
- TORRES, F.L.R.; KUKI, C.; VASCONCELLOS, B.; FREITAS, A.; SILVA, P.; *et al.* Validation of different precipitation databases of the Sapucaí and São Francisco River basins. **Revista Brasileira de Climatologia**, v. 27, p. 368-404, 2020. doi
- TRACTON, M.S.; KALNAY, E. Operational ensemble prediction at the National Meteorological Center: practical aspects. **Weather & Forecasting**, v. 8, n. 3, 1993. doi
- TRENBERTH, K.E. Development and forecasts of the 1997/98 El Niño: CLIVAR scientific issues. **CLIVAR Exchanges**, v. 3, p. 4-14, 1998.

- TRENBERTH, K.E.; STEPANIAK, D.P. Indices of El Niño evolution. **Journal of Climate**, v. 14, n. 8, p. 1697-1701, 2001. [doi](#)
- VAPNIK, V. **The Nature of Statistical Learning Theory**. 2. ed. New York: Springer, 2000.
- WANG, Y.; ZHOU, X.; LIANG, L.; ZHANG, M.; ZHANG, Q.; *et al.* Short-term wind speed forecast based on least squares support vector machine. **Journal of Information Processing Systems**, v. 14, n. 6, p. 1385-1397, 2018. [doi](#)
- WARD, M.N.; FOLLAND, C.K. Prediction of seasonal rainfall in the north Nordeste of Brazil using eigenvectors of sea-surface temperature. **International Journal of Climatology**, v. 11, n. 7, p. 711-743, 1991. [doi](#)
- WILKS, D.S. Chapter 7 - Forecast Verification. In: **International Geophysics: Statistical Methods in the Atmospheric Sciences**, v. 100, p. 301-394, 2011. [doi](#)
- WILLMOTT, C.J. On the validation of models. **Physical Geography**, v. 2, n. 2, p. 184-194, 1981. [doi](#)
- WOLTER, K. The Southern Oscillation in surface circulation and climate over the Tropical Atlantic, Eastern Pacific, and Indian Oceans as captured by cluster analysis. **Journal of Applied Meteorology and Climatology**, v. 26, n. 4, p. 540-558, 1987. [doi](#)
- WOLTER, K.; TIMLIN, M.S. Monitoring ENSO in COADS with a seasonally adjusted principal component index. In: **Proceedings 17 Climate Diagnostics Workshop**, Norman, p. 52-57, 1993.
- WU, X.; ZHOU, J.; YU, H.; LIU, D.; XIE, K.; *et al.* The Development of a hybrid wavelet-ARIMA-LSTM model for precipitation amounts and drought analysis. **Atmosphere**, v. 12, n. 1, p. 74, Jan. 2021. [doi](#)
- XIE, P.; CHEN, M.; YANG, S.; YATAGAI, A.; HAYASAKA, T.; *et al.* A Gauge-based analysis of daily precipitation over East Asia. **Journal of Hydrometeorology**, v. 8, n. 3, p. 607-626, 2007. [doi](#)
- XU, L.; CHEN, N.; ZHANG, X.; CHEN, Z. A data-driven multi-model ensemble for deterministic and probabilistic precipitation forecasting at seasonal scale. **Climate Dynamics**, v. 54, n. 7, p. 3355-3374, 2020. [doi](#)
- XU, L.; CHEN, N.; ZHANG, X. A comparison of large-scale climate signals and the North American Multi-Model Ensemble (NMME) for drought prediction in China. **Journal of Hydrology**, v. 557, p. 378-390, 2018. [doi](#)
- YAN, J.; JIN, J.; CHEN, F.; YU, G.; YIN, H.; *et al.* Urban flash flood forecast using support vector machine and numerical simulation. **Journal of Hydroinformatics**, v. 20, n. 1, p. 221-231, Jan. 2018. [doi](#)
- YANG, L.; HE, M.; ZHANG, J.; VITTAL, V. Support-vector-machine-enhanced Markov model for short-term wind power forecast. **IEEE Transactions on Sustainable Energy**, v. 6, n. 3, p. 791-799, 2015. [doi](#)
- ZENG, Z.; HSIEH, W.W.; SHABBAR, A.; BURROWS, W.R. Seasonal prediction of winter extreme precipitation over Canada by support vector regression. **Hydrology and Earth System Sciences**, v. 15, n. 1, p. 65-74, 2011. [doi](#)

Supplementary Material

Figure S1 - Spatiotemporal pattern of MAE [mm/month] for SEAS5, MLR, and SVM precipitation anomaly forecasts in the austral summer (DJF). The precipitation anomaly data from the forecast simulations (January 2017 to December 2020) were accumulated and assessed according to the months of the forecast horizon (months 1, 3, and 7).

Table S1 - Average MAE and bias for precipitation anomaly forecasts (January 1993 to December 2016) in the rainy season (DJF), by region and month of the forecast horizon. The smallest error values are colored in green.

Table S2 - Average MAE for precipitation anomaly forecasts (January 1993 to December 2016) in the rainy season (DJF), by region and month of the forecast horizon. The smallest error values are colored in green.



License information: This is an open-access article distributed under the terms of the Creative Commons Attribution License (type CC-BY), which permits unrestricted use, distribution and reproduction in any medium, provided the original article is properly cited.