*Federal University of Paraíba - UFPB*

*Technology Center*

**Postgraduate Program in Civil and Environmental Engineering**

**– PhD Thesis –**

# IMPROVEMENT OF PRECIPITATION ESTIMATION AT MONTHLY AND DAILY SCALES FOR BRAZIL BASED ON REMOTE SENSING PRODUCT AND MACHINE LEARNING TECHNIQUES

*by*

***Emerson da Silva Freitas***

*PhD thesis defended at Federal University of Paraíba for attaining the degree in Doctor of Civil and Environmental Engineering with emphasis on Water Resources*

**João Pessoa – Paraíba** **September 2024**

*Federal University of Paraíba - UFPB*

*Technology Center*

**Postgraduate Program in Civil and environmental Engenieering**

**– PhD Thesis –**

# IMPROVEMENT OF PRECIPITATION ESTIMATION AT MONTHLY AND DAILY SCALES FOR BRAZIL BASED ON REMOTE SENSING PRODUCT AND MACHINE LEARNING TECHNIQUES

PhD Thesis submitted to the Postgraduate Program in Civil and Environmental Engineering at the Federal University of Paraíba, as part of requisites for attaining the title of Doctor.

**Emerson da Silva Freitas**

**Advisor: Prof. Dr. Cristiano das Neves Almeida**

**Co-advisor: Prof. Dr. Victor Hugo Rabelo Coelho**

**João Pessoa – Paraíba** **September 2024**

**Catalogação na publicação**
**Seção de Catalogação e Classificação**

F866i Freitas, Emerson da Silva.
Improvement of precipitation estimation at monthly and daily scales for brazil based on remote sensing product and machine learning techniques / Emerson da Silva Freitas. - João Pessoa, 2024.
85 f. : il.

Orientação: Cristiano das Neves Almeida.
Coorientação: Victor Hugo Rabelo Coelho.
Tese (Doutorado) - UFPB/CT.

1. Aprendizado de máquina. 2. Precipitação - Ciclo hidrológico. 3. Dados de reanálise. 4. K-Nearest Neighbours. 5. Sensoriamento remoto. I. Almeida, Cristiano das Neves. II. Coelho, Victor Hugo Rabelo. III. Título.

UFPB/BC CDU 004.891(043)

Elaborado por WALQUELINE DA SILVA ARAUJO - CRB-15/514

PPGECAM
Programa de Pós-Graduação
Engenharia Civil e Ambiental
UFPB Universidade Federal da Paraíba

**Mestrado e Doutorado**

# IMPROVEMENT OF PRECIPITATION ESTIMATION AT MONTHLY AND DAILY SCALES FOR BRAZIL BASED ON REMOTE SENSING PRODUCT AND MACHINE LEARNING TECHNIQUES

**EMERSON DA SILVA FREITAS**

**Tese aprovada em 27/09/2024**
**Período Letivo: 2024.2**

gov.br Documento assinado digitalmente
CRISTIANO DAS NEVES ALMEIDA
Data: 13/03/2025 10:36:53-0300
Verifique em https://validar.iti.gov.br

**Prof. Dr. Cristiano das Neves Almeida – UFPB**
**Orientador**

gov.br Documento assinado digitalmente
VICTOR HUGO RABELO COELHO
Data: 13/03/2025 14:53:58-0300
Verifique em https://validar.iti.gov.br

**Prof. Dr. Victor Hugo Rabelo Coelho – UFPB**
**Coorientador**

gov.br Documento assinado digitalmente
DAVI DE CARVALHO DINIZ MELO
Data: 13/03/2025 10:56:58-0300
Verifique em https://validar.iti.gov.br

**Prof. Dr. Davi de Carvalho Diniz Melo – UFPB**
**Examinador Interno**

gov.br Documento assinado digitalmente
GUILLAUME FRANCIS BERTRAND
Data: 19/03/2025 15:42:37-0300
Verifique em https://validar.iti.gov.br

**Prof. Dr. Guillaume Francis Bertrand – UFPB**
**Examinador Interno**

gov.br Documento assinado digitalmente
SAULO AIRES DE SOUZA
Data: 20/03/2025 19:17:46-0300
Verifique em https://validar.iti.gov.br

**Dr. Saulo Aires de Souza – ANA**
**Examinador Externo**

**Drª. Ana Paula Martins do Amaral Cunha – CEMADEN**
**Examinador Externo**

gov.br Documento assinado digitalmente
ANA PAULA MARTINS DO AMARAL CUNHA
Data: 13/03/2025 11:49:40-0300
Verifique em https://validar.iti.gov.br

**João Pessoa/PB - 2024**

And to those who are part of my life.

I dedicate!

## ACKNOWLEDGEMENTS

I want to take this opportunity to thank everyone who helped me, directly or indirectly, to complete this thesis, because this work required a lot of effort, courage, determination and hardship, and I would not have been able to achieve this goal if I had been alone in this life's endeavour.

First, I would like to thank God for all the opportunities and experiences that contributed to my learning, especially the realisation of this dream.

I want to thank my mother, Gerlande da Silva Freitas, for her affection and devotion, for always supporting me, for the moral principles that made me who I am today, and above all for everything she did to help me reach this moment.

To my wife, Juliene Diniz Pereira, for all the love, understanding and support she received, even when I was away, which were necessary to seek the knowledge needed to complete this work.

To Professor Dr Cristiano das Neves Almeida, for all the teachings, support and encouragement I received over all these years, which were fundamental to the completion of this thesis and also fundamental to my professional and personal development. You have been like a father to me.

To Professor Dr Victor Hugo Rabelo Coelho for the wisdom you have imparted and your constant availability. Without you, this thesis would not have been possible, you were like a big brother to me.

To all the professors of the PPPGECAM, who have imparted their knowledge with great skill.

To my lab colleagues, Filipe Carvalho, Marcela Meira, Geraldo Moura and Abner Lins, for all the help I have received.

To everyone who said I would not be able to do it, this has become fuel for this journey.

To CAPES (Coordination for the Improvement of Higher Education Personnel) for granting the scholarship.

## ABSTRACT

Precipitation is one of the main components of the hydrological cycle and its accurate quantification is essential to provide information for understanding and predicting physical processes. Occurrence observations based on ground-based devices (manual and automatic rain gauges) are highly accurate but have limited spatial coverage. On the other hand, remote sensing products cover large areas but with lower precision. In this context, this study aims to evaluate machine learning models to create a product with better occurrence estimation, with lower latency than other products and without directly relying on field data. The methodology consists of choosing the best machine learning model (classification and regression) and applying it to satellite-based remote sensing data (IMERG Early Run product) and reanalysis-based variables (MERRA-2). The method was applied throughout the Brazilian territory, on monthly and daily scales, which presents a wide variety of supply regimes. This methodology first resulted in the development of an adjusted IMERG product at the monthly scale (IMERG-BraMaL) and later an improved product at the daily scale with a multiple machine learning technique (IMERG-BraMMaL). Compared to the original IMERG products (Early Run and Final Run) and global estimation products (MSWEP, CHIRPS and PERSIANN-CDR), IMERG-BraMaL improved the analyses evaluated between terrestrial and satellite data in almost all analyses. For example, the KGE (Kling-Gupta Efficiency) went from lower values (0.70, 0.82, 0.09, 0.60 and 0.81 for IMERG Early, IMERG Final, PERSIANN, MSWEP and CHIRPS, respectively) to values above 0.86 in IMERG-BraMal at the monthly scale. On a daily scale, IMERG BraMMAL proved to be more efficient, presenting better results, with a CC of 0.79 compared to 0.68 for IMERG BraMaL. The main conclusions of the study were: (i) much faster availability to end users; (ii) no dependence on any field data, allowing its application in areas where rainfall data are not available or are of low quality; (iii) no correlation of errors with local characteristics; and (iv) much improved estimates in regions of Brazil where, historically, satellite-based products often underestimate the observed data.



**KEYWORDS:** machine learning, precipitation, re-analysis data, k-nearest neighbours, remote sensing.

## RESUMO

A precipitação é um dos principais componentes do ciclo hidrológico e sua quantificação precisa é essencial para fornecer informações para a compreensão e previsão de processos físicos. As observações de ocorrência baseadas em dispositivos terrestres (pluviômetros manuais e automáticos) são altamente precisas, mas têm cobertura espacial limitada. Por outro lado, os produtos de sensoriamento remoto cobrem grandes áreas, mas com menor precisão. Neste contexto, este estudo tem como objetivo avaliar modelos de aprendizado de máquina para criar um produto com melhor estimativa de ocorrência, com menor latência que outros produtos e sem depender diretamente de dados de campo. A metodologia consiste em escolher o melhor modelo de aprendizado de máquina (classificação e regressão) e aplicá-lo a dados de sensoriamento remoto baseados em satélite (produto IMERG Early Run) e variáveis baseadas em reanálise (MERRA-2). O método foi aplicado em todo o território brasileiro, em escalas mensais e diárias, que apresenta uma grande variedade de regimes de abastecimento. Esta metodologia primeiramente resultou no desenvolvimento de um produto IMERG ajustado na escala mensal (IMERG-BraMaL) e posteriormente um produto melhorado na escala diária com uma técnina de múltiplos machine leraning (IMERG-BraMMaL). Comparado aos produtos originais do IMERG (Early Run e Final Run) e produtos de estimativas globais (MSWEP, CHIRPS e PERSIANN-CDR), o IMERG-BraMaL melhorou as análises avaliadas entre dados terrestres e de satélite em quase todas as análises. Por exemplo, o KGE (Eficiência Kling-Gupta) passou de valores mais baixos (0.70, 0.82, 0.09, 0.60 e 0.81 para IMERG Early, IMERG Final, PERSIANN, MSWEP e CHIRPS, respectivamente) para valores acima de 0.86 no IMERG-BraMal na escala mensal. Na escala diária, o IMERG BraMMAL se mostrou mais eficiente, apresentando melhores resultados, com CC de 0,79 comparado a 0,68 do IMERG BraMaL. As principais conclusões do estudo foram: (i) disponibilidade muito mais rápida para os usuários finais; (ii) não dependência de quaisquer dados de campo, permitindo sua aplicação em áreas onde os dados pluviométricos não estão disponíveis ou são de baixa qualidade; (iii) a não relação dos erros com as características locais; e (iv) estimativas muito melhoradas em regiões do Brasil onde, historicamente, os produtos baseados em satélites frequentemente subestimam os dados observados.



**PALAVRAS-CHAVES:** Aprendizado de máquina, precipitação, dados de reanálise, k-nearest neighbours, sensoriamento remoto.

**SUMMARY**

## LIST OF FIGURES

## LIST OF TABLES

# 1 INTRODUCTION

Rainfall is one of the main components of the hydrological cycle, and its accurate spatio-temporal quantification is essential to provide basic information for a large range of hydrometeorological processes and socio-economic activities (Breugem et al., 2020; Y. Du & Xie, 2020; Markonis et al., 2019). The temporal resolution of the rainfall data is an important characteristic for many hydro-meteorological applications. For instance, analyses of extreme events and flash floods require high-resolution data (e.g. sub-daily) (Llauca et al., 2021; Sadeghi et al., 2020), while low-resolution data (daily onwards) can be adequate for trend and climate change analysis (Bonnema et al., 2016).

Although crucial for many applications, achieving accurate global precipitation estimates (i.e. without gaps and large biases) is still challenging because some monitoring instruments, retrieval methods, and numerical models still present remarkable uncertainties (Sun et al., 2018; Wehbe et al., 2020). Also, dense precipitation measurements still remain a costly and laborious procedure, especially for fine temporal scales (i.e., sub-hourly) (Blenkinsop et al., 2018; Freitas et al., 2020).

Rain gauges are considered to be the most accurate method for obtaining precipitation data (L. Xu et al., 2020). However, the global availability of rain gauges is still sparse and uneven (Becker et al., 2013; Harris et al., 2014; Schneider et al., 2016; Zhang et al., 2021) and has geographic discontinuity and coverage limitation (Raj et al., 2022). The lack of rain gauges is even more prominent in developing and sizeable countries, like Brazil. When the rainfall data based on rain gauges are available in these countries, the users often face problems, such as: i) difficulty of access; ii) gaps and a lack of information; and iii) absence of quality control and standardisation, making the identification of problems difficult (e.g. recording the same amount of precipitation over longer periods) (Blenkinsop et al., 2018).

To overcome these inherent problems with rain gauges and facilitate access to data, precipitation estimated from orbiting remote sensors emerged in the 1960s (Kidd & Huffman, 2011). The development of remote sensing techniques has provided an unprecedented opportunity to create products that estimate spatial precipitation continuously on a global scale over recent decades (Zhang et al., 2021). These products are mainly obtained through algorithms that combine measurements from orbiting remote sensors operating in the microwave and infrared bands of the electromagnetic spectrum (Kidd & Levizzani, 2011).

The use of orbiting remote sensors has increased the amount of regional and global

studies, enabling, among others, a better understanding of precipitation characteristics (Chen et al., 2020; Freitas et al., 2020; Gupta et al., 2020; Kidd & Huffman, 2011; Markonis et al., 2019) and applications in hydrometeorological studies and monitoring (Belabid et al., 2019; Llauca et al., 2021; Munier et al., 2014; Pellet et al., 2019). Several satellite-based products are currently providing sub-daily precipitation information (Llauca et al., 2021; Ramos Filho et al., 2021), with different characteristics, in terms of resolutions (spatial and temporal), spatial coverage, latency etc (Beck, Van Dijk, et al., 2017).

Although internationally recognised, the satellite-based precipitation data still present some associated inconsistencies, such as: i) temporal and spatial resolutions that may not be sufficient for use in some hydrological modelling (Behrangi and Wen, 2017); ii) difficulties in estimating light rain and detecting precipitation on snow and ice-covered surfaces (Kidd & Levizzani, 2011); iii) errors due to variability and uncertainties introduced by orographic effects (Bhuiyan et al., 2018; Bhuiyan et al., 2019; Derin & Yilmaz, 2014; Houze, 2012; Mei et al., 2014); iv) failures in estimating convective rainfall (Gadelha et al., 2019); v) inconsistencies in identifying extreme precipitation events that trigger hydrological disasters (Brunetti et al., 2018; Brunetti et al., 2021; Ramos Filho et al., 2021); and v) errors in identifying precipitation events and their properties (e.g. intensity and duration) (Freitas et al., 2020).

Worldwide studies have pointed out that the inconsistencies in satellite precipitation estimates have been reduced since the advent of the Global Precipitation Measurement (GPM) mission (e.g. Bhuiyan et al., 2017; Gadelha et al., 2019; Li et al., 2017; Oliveira et al., 2016; Prakash et al., 2016; Satgé et al., 2019; Silva Lelis et al., 2018; Tan et al., 2018), due to the new dual-frequency precipitation radar (DPR). The GPM mission makes available three IMERG (Integrated Multi-satellitE Retrievals for GPM) products: Early Run, Late Run, and Final Run. The Early and Late Run products are near-real-time products, which are available to end-users 4 and 12 hours after observation, respectively. The Final Run product is available 3.5 months after observation, and, subsequently, monthly gauge observations from the Global Precipitation Climatology (GPCC) data, and other ancillary data, are incorporated to improve satellite estimations. The Final Run product shows improved performance in estimating precipitation when compared to the near-real-time products (Jiang et al., 2021; Ramadhan et al., 2022; Sungmin et al., 2017; Wang et al., 2017; Zhou et al., 2021). However, despite the progressive reduction of the inconsistencies, it is recognised by the scientific community that some disagreements between the data observed by rain gauges and those estimated by the satellite-based products still prevail, leaving them open to further

improvements (Li et al., 2017; Ma et al., 2020; Ning et al., 2017; Tang et al., 2016; R. Xu et al., 2017).

In this perspective, some techniques have been used in recent years, such as: (i) the fusion of remote sensing estimations with ground-based rainfall data (Beck et al., 2017; Beck et al., 2019; Bhuiyan et al., 2017; Bhuiyan et al., 2018; Chen et al., 2022; Rafieeinasab et al., 2015; Rozante et al., 2018; Wang et al., 2020), and (ii) a "Bottom-up" approach that uses satellite-based soil moisture observations to infer or agree on land-based soil moisture observations, such as the SM2RAIN algorithm (e.g. Brocca et al., 2013, 2016, 2019; Pellarin et al., 2013; Wanders et al., 2015). Despite efforts to improve the satellite-based precipitation data, some issues persist, including underestimations of extreme precipitation events and false event estimates, which were reported in both bottom-up and fusion techniques (Brocca et al., 2019). Specific to bottom-up techniques, large-scale in situ soil moisture data are also required (Brocca et al., 2019) but dense observed networks are not available in many regions. Additionally, these models have low accuracy when there is dense vegetation coverage and complex terrain (Brocca et al., 2013). In addition, the fusion-based products need ground-based precipitation data but, as mentioned above, the monitoring networks are not dense in some regions (Schneider et al., 2016; Zhang et al., 2021), which can affect the final quality of the product.

In recent years, artificial intelligence (AI), such as machine learning techniques, has been widely used. According to Zhang et al. (2021), machine learning algorithms have advantages over other methods because they: (i) can effectively handle complex and nonlinear relationships between input and output data; (ii) do not contain any rigid assumptions; (iii) are highly flexible in terms of incorporating various types of explanatory variables; (iv) can combine field observations with multiple remote sensing products; and (v) are easy for implementation.

Overall, the following machine learning algorithms have been used to enhance satellite-based precipitation products: i) Quantile Regression Forests (QRF) (Tyralis et al., 2023; Y. Yang & Luo, 2014); ii) Support Vector Machine (SVM) (Kumar et al., 2019; Sehad et al., 2017) ; iii) Random Forests (RF) (Assiri & Qureshi, 2022; Bhuiyan et al., 2020a; Kumar et al., 2019; Sengoz et al., 2023; Wolfensberger et al., 2021); iv) Artificial Neural Networks (RNA) (Bhuiyan et al., 2020a; Sadeghi et al., 2020; Sengoz et al., 2023; Wehbe et al., 2020); v) Gradient boosting (GB) (Sengoz et al., 2023; Tyralis et al., 2023; R. Wang et al., 2023); and vi) Linear regression (LRi) (Sengoz et al., 2023; C. Wang et al., 2021).

To our knowledge, the studies using machine learning on satellite-based precipitation

products focused on the accuracy of the machine learning algorithms, comparing their results with the original estimation of the product to confirm the improvements (Bhuiyan et al., 2017, 2019, 2020a; Zhang et al., 2021). However, these new precipitation products are not available to the scientific community. Additionally, some of these studies use specific regional data for the studied region, such as ground radar data (H. Chen et al., 2020; R. Wang et al., 2023; Wehbe et al., 2020) and gauge observations (Bhuiyan et al., 2020a; Papacharalampous et al., 2023c; Yu et al., 2023; Zhang et al., 2021), which may limit their applicability in other regions. Moreover, these studies were conducted at local scales (mostly at the river basin scale), without the evaluation of some available machine learning models (e.g. K-Nearest Neighbours, KNN). Only the study by Ma et al. (2020) generated a new precipitation product for Asia (namely, aIMERG) applying a new calibration approach to the IMERG product that included 57,835 ground-based observations from rain gauges, i.e. limiting its applicability to other regions with a lack, or uneven distribution, of rain gauges.

### 1.1. Hypotheses

Following the contextualisation above, this PhD thesis is based on the hypothesis that machine learning algorithms can significantly enhance the accuracy of daily and monthly satellite-based precipitation, without dependence on local or regional in-situ data.

### 1.2. Main aims

This study aims to optimise monthly and daily satellite-based precipitation estimates, based on machine learning techniques and reanalysis data.

### 1.3. Specific aims

- To develop a machine learning model to obtain a satellite-based precipitation product with reduced latency and higher accuracy compared to the existing alternatives;
- To create a machine learning model to obtain a satellite-based precipitation product without dependency on any field data for its calibration, allowing its application in areas where rain gauge data are unavailable or present low quality;
- To build a machine learning model to obtain a satellite-based precipitation product whose errors are unrelated to local features (e.g. climate, precipitation regimes, topography);
- To identify the best machine learning models for classification and regression

of precipitation data;

- To investigate the application of individual and combined machine learning techniques (classification/regression and stacking) to enhance daily rainfall estimates based on remote sensing data, aiming for more precise identification of precipitation events;
- To compare the products calibrated on monthly and daily scales to determine the more accurate for improving the precipitation estimate;
- To compare the monthly and daily precipitation data, obtained from the proposed products, with other global satellite-based precipitation products.

## 1.4. Thesis Structure

The thesis was divided into six major items in the following order: 1) Introduction, presenting a contextualisation and justification that motivated the development of this thesis; 2) Literature review, covering conceptual and basic topics essential to better understanding the study; 3) Study area, presenting the characteristics of the studied area and justifying its selection; 4) An improved gridded monthly rainfall product; 5) Evaluation of single and combined machine learning models to improve daily rainfall estimations; and 6) Conclusions and recommendations, highlighting the main results found and proposing further studies.

# 2 LITERATURE REVIEW

## 2.1. Precipitation

Precipitation is a fundamental component of the climate system and global water cycle, whose observation and measurement are crucial for human well-being (Kidd & Huffman, 2011). Excessive and insufficient rainfall threatens lives, properties, and agriculture. Precipitation also holds significant economic relevance, playing a central role in water resource management, agribusiness (Kidd et al., 2009; Thornes et al., 2010). From an environmental point of view, precipitation serves as a transformative agent in the surrounding landscape, influencing both the sustenance of natural vegetation and erosive processes. Simultaneously, it plays a role in the dispersion of atmospheric pollution and the transport of nutrients and pollutants (Michaelides et al., 2009).

Due to the importance of precipitation, its physical characteristics need to be frequently and systematically measured at a fine spatiotemporal resolution. Historically, quantitative precipitation measurements have been limited to a relatively short period in recent history (Kidd & Huffman, 2011). Although precipitation data has been available since the mid-1850s, early records exhibit variations in terms of accessibility, completeness, and consistency. Also, the availability of data at shorter intervals (i.e., daily or sub-daily) remains scarce (Blenkinsop et al., 2018; Freitas et al., 2020; Jones et al., 2002; New et al., 2001).

Over the years, studies on precipitation have been continuously capitalised on by technological advancements to obtain more accurate measurements and fill gaps in our knowledge and understanding of the processes influencing precipitation. The first measurements were conducted using manual rain gauges, i.e. simple collectors placed on the earth's surface, whose reading is done by an observer daily. Over time, these instruments were gradually improved to more advanced versions (e.g. automatic rain gauges). The accumulation of data over several years has driven numerous hydrometeorological and climatological studies, initially locally and regionally and, subsequently, on global scale. Recently, technologically sophisticated devices, installed on the Earth's surface (e.g. ground-based radars, disdrometers) or aboard space platforms (e.g. microwave and infrared sensors), have expanded our spatial and temporal understanding of precipitation.

### 2.1.1. Ground-based measurements

Manual rain gauges (Figure 1) consist of a cylindrical tube, opened at the top, which

collects rainwater. The amount of precipitation is then measured by the height of the water accumulated inside this tube. This approach provides a direct measure of the amount of rainfall, making the rain gauge a valuable tool for providing accurate and reliable data on precipitation in a specific location (Garcez & Alvarez, 1988; Pinto et al., 1976).

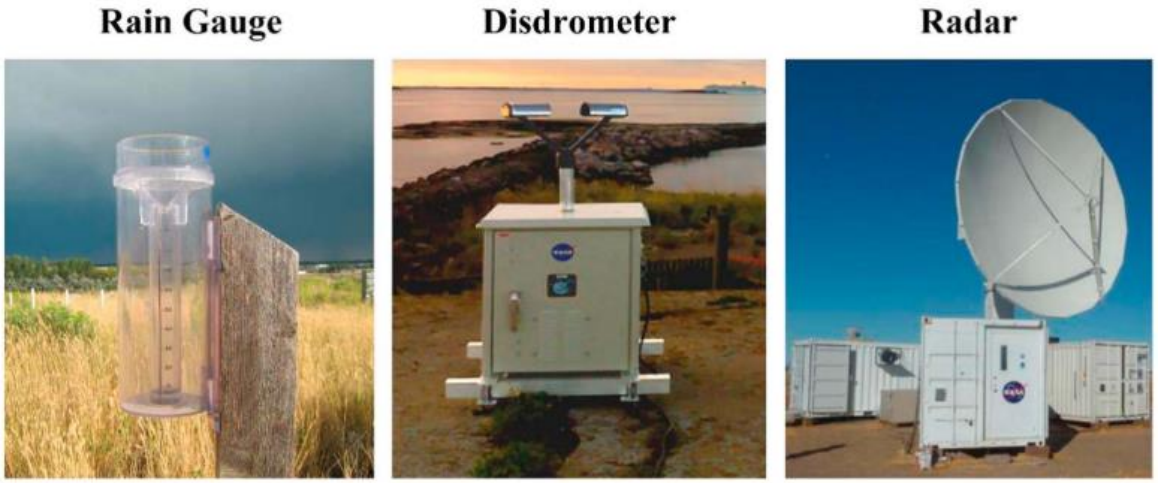


Figure 1 – Three instruments for making ground observations of precipitation. (Courtesy: Sun et al. 2018)

The use of rain gauges is widespread throughout the world due to their simplicity and effectiveness. However, rain gauges provide a point-scale measurement, which poorly represents the spatial variability of precipitation, ranging from a few metres to several kilometres. The spatial distribution of these instruments can affect the accuracy of precipitation estimates over large areas, as highlighted by Harris et al. (2014). In regions with limited rain gauge density, the spatial representation of precipitation can be compromised, challenging hydrological and meteorological modelling and forecasting.

To overcome some of these challenges, automatic rain gauges, which use electronic sensors to measure precipitation continuously and at shorter time intervals, were developed (Blenkinsop et al., 2018). This type of device offers a more dynamic approach to data collection, allowing for a better understanding of the temporal variability of precipitation.

Although less dynamic, manual rain gauges are the most frequent equipment to provide accurate precipitation data from the ground (Kidd 2001). Sub-daily precipitation data is an even more difficult task because sparsely covers the global landmass (Hegerl et al., 2015.; Lewis et al., 2019). The number of in-situ sub-daily precipitation records is even lower in tropical regions, probably due to the higher costs of implementing rain gauges capable of measuring sub-daily events, compared to those that measure on a daily time scale (Freitas et al., 2020; Kidd et al., 2017).

Thus, new projects emerged intending to create a reliable precipitation database. The INTENSE project (Intelligent use of climate models to adapt to non-stationary hydrological extremes) is the first major international effort to focus on global extreme sub-daily precipitation, enabling substantial advances in quantifying observed historical changes. However, the INTENSE project identified, from a data collection initiative, the lowest availability of sub-daily precipitation data in countries in Africa and Latin America (Blenkinsop et al. 2018). In Brazil, CEMADEN created a sub-daily monitoring network with approximately 3,400 automatic rain gauges distributed over the country. This network was created to support the prediction and development of warning systems for precipitation-related disasters, including floods and landslides, so most rain gauges are within the cities.

### 2.1.2. Ground-based radar data

Ground-based radars (Figure 1) can be an alternative to rain gauges, providing real-time measurements with high temporal and spatial resolution. However, the spatial coverage of ground-based radars is limited only by land and is also affected by a lack of accessibility due to their high cost (Varma, 2018).

The radar system consists of a transmitter producing electromagnetic microwaves that are backscattered by particles in the atmosphere and then converted into a measure of rainfall intensity (Kidd & Huffman, 2011). However, the accuracy of ground-based radar measurements is often low, especially for extreme precipitation magnitudes (Marra & Morin, 2015), since intensity is indirectly derived from radar-measured reflectivity (i.e. subject to multiple sources of error) rather than being a direct measurement (Ochoa-Rodriguez et al., 2019; Pellarin et al., 2013).

## 2.2. Remote sensing data

### 2.2.1. Satellite-based precipitation data

Sensors onboard satellites are currently the only instruments to provide global homogeneous precipitation measurements. Nowadays, many promising satellite-based precipitation products are available for many applications, followed by the advancement in the number of satellite sensors and imaging technology. Such products provide valuable distributed information on sub-daily precipitation data (Llauca et al., 2021; Yuan et al., 2019).

The sensors onboard satellites, used to estimate precipitation, can be classified into three categories: Visible/InfraRed sensors, passive Microwave, and active Microwave sensors (Michaelides et al., 2009; Prigent, 2010). Corresponding methods used to derive

precipitation from electromagnetic interactions with clouds and the atmosphere have been developed, including the Visible/InfraRed-based methods, active and passive Microwave techniques, and merged Visible/InfraRed and Microwave approaches (Kidd & Levizzani, 2011).

The GPM IMERG is produced using passive microwave techniques, a more direct method of measuring precipitation than Visible/Infrared techniques, as microwave lengths can penetrate clouds and detect precipitation-sized particles. The most critical disadvantage of passive microwave precipitation estimation is its low spatiotemporal coverage (Sun et al., 2018).

As mentioned previously, rain gauges provide accurate measurements of precipitation at single points, but they are sparsely distributed across the globe and can be affected by sampling error (Habib et al., 2001; Kidd et al., 2017). On the other hand, satellite observations have homogeneous spatial coverage, but random errors and biases linked to the algorithms can be detected (Chen et al., 2021; Kidd et al., 2017). Therefore, merging different sources of information to overcome these errors and biases, by combining the individual advantages of specific methods, is currently the subject of several studies (Figure 2).

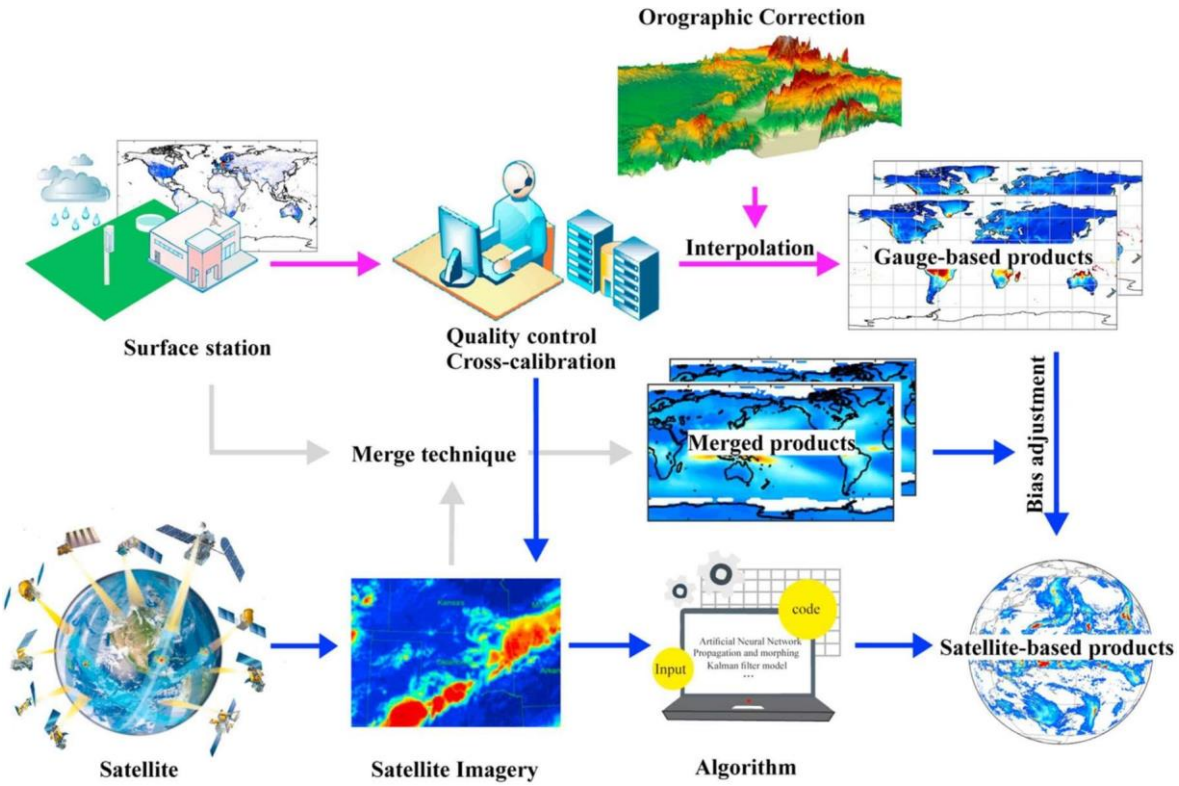


**Figure 2** – Flowchart for the precipitation products. (Courtesy: Sun et al. 2018).

### 2.2.2. Reanalysis datasets

Reanalysis systems fuse irregular observations and models spanning many physical and dynamic processes to generate an estimate of the state of the system through a uniform grid, with spatial homogeneity, temporal continuity, and a multidimensional hierarchy (Sun et al., 2018). Many essential climate variables, resulting from reanalysis systems, can be obtained after short periods. A reanalysis system includes a historical prediction model and a data assimilation routine.

Reanalysis generates a large variety of atmospheric, sea-state, and land surface parameters across a uniform grid with spatial homogeneity over long periods through data assimilation, a process that relies on both observations and model-based forecasts to estimate conditions (Dee et al., 2014; Sun et al., 2018).

Successive generations of reanalysis products produced by various organisations have improved their quality, with improved models, input data, and assimilation methods. As examples, we can mention the two NCEP/NCAR reanalysis systems (NCEP1 and NCEP2), the two reanalysis systems of the European Center for Medium-Range Weather Forecasts (ECMWF) (ERA-40 and ERA-Interim), the Century Reanalysis (20CRv2), the Modern-Era Retrospective Analysis for Research and Application (MERRA) system, the NCEP Climate Forest System Reanalysis (CFSR) system, and the Japanese 55-Year Reanalysis (JRA-55).

Version 2 of the MERRA system (MERRA-2) is a global atmospheric reanalysis product produced by the Goddard Earth Observing System Model, Version 5 (GEOS-5), at NASA's Goddard Space Flight Center (GSFC), which combines satellite observations, sounding data, and other observational data to create a three-dimensional, temporally consistent representation of the atmospheric state (Gelaro et al., 2017).

## 2.3. Machine Learning

### 2.3.1. Definition

Over the past few decades, machine learning has played a crucial role in advancing artificial intelligence, providing machines with the ability to learn from data and improve their performance over time (Das et al., 2022; Papacharalampous et al., 2023a; Zhang et al., 2021). This definition highlights the autonomy of learning, emphasising the importance of systems adaptively acquiring knowledge.

One of the pioneers in this field was Arthur Samuel, who first introduced the term machine learning (Samuel, 1959). In his work, Samuel described machine learning as the field of study that allows computers to learn from data without being explicitly programmed.

This fundamental insight still influences the contemporary understanding of machine learning. More recently, Bishop (2006) explored fundamental machine-learning techniques, highlighting the importance of pattern recognition and probabilistic inference.

There is still confusion regarding the terms artificial intelligence, machine learning, and deep learning; they are often used as synonyms, but there are differences. These concepts represent different layers of complexity and automation in a machine's ability to learn and make decisions. The field of artificial intelligence seeks to develop systems capable of performing tasks that normally require human intelligence (McCarthy et al., 2006). Machine learning is a subcategory of artificial intelligence that focuses on systems that can learn from data (Mitchell, 1977). Deep learning is a specific machine learning approach that uses deep neural networks to model and solve complex problems (Hinton et al., 2006), and has played a significant role in expanding the boundaries of machine learning (Goodfellow et al., 2016).

### 2.3.2. Brief history

The history of machine learning began in 1950, when Alan Turing, considered the father of computing and artificial intelligence, published the article "Computing Machinery and Intelligence". In his work, Turing introduced the famous test that assessed the ability of a machine to exhibit behaviour indistinguishable from that of humans. In 1956, artificial intelligence was formally established as a field of research during the Dartmouth Conference, whose term was coined by John McCarthy.

However, in the 1970s and 1980s, the field faced challenges and technological limitations, resulting in the so-called "Artificial Intelligence Winter." Funding for artificial intelligence research has declined due to unmet expectations and technical difficulties.

The resurgence began in the 1980s and 1990s, with advances in machine learning algorithms, including SVM (Support Vector Machines), proposed by Vladimir Vapnik, and neural networks. The popularity of machine learning continued to grow, but it was from the 2000s onwards that these techniques became more widespread.

The advent of deep learning marked a turning point in the machine learning field, with contributions like "ImageNet Classification with Deep Convolutional Neural Networks" (Krizhevsky et al., 2017), demonstrating the effectiveness of deep neural networks. This period also witnessed the rise of big data, providing massive data sets to train complex models.

### 2.3.3. Machine learning model types

Machine learning models can be categorised into three main types, based on distinct

approaches: supervised, unsupervised, and reinforcement. These categories reflect different learning paradigms, with specific applications and challenges.

Supervised machine learning models are algorithms trained by a set of predictor data, i.e. a set of data in which the desired outputs are already known. The main objective of supervised models is to learn the relationship between inputs and outputs from the given training examples (James et al., 2013). These models are fundamental for regression and classification works, in which choosing the right algorithm for a specific problem and the use of training data are fundamental (Hastie et al., 2009). In supervised machine learning models, the algorithm is fed with input and output pairs and trained to learn a function that maps these inputs to the desired outputs. After training, the model can make predictions or decisions when presented with new unlabeled data.

Unsupervised machine learning models are approaches where the algorithm is trained on a dataset without predictions, where the desired outputs are not provided. The main goal of unsupervised models is to discover intrinsic insights in data, such as natural groupings, relationships, or distributions. These models are often used in situations where the nature of the task is not known in advance or when data labelling is difficult or expensive (James et al., 2013). The most common unsupervised models are the k-means algorithm (MacQueen, 1967), which is used to group data into clusters, identifying intrinsic patterns and structures without relying on labels; and Principal Component Analysis (PCA) (Hotelling, 1933), which reduces the dimensionality of the data while keeping most of the variance.

Reinforcement models in machine learning refer to an approach in which an agent learns to make sequential decisions through interactions. The model is not based on input and output pairs but rather on feedback in terms of rewards or penalties, adjusting its strategies to maximise the rewards (James et al., 2013). Reinforcement models have their roots in the field of artificial intelligence and have stood out in practical applications, such as games and robotics, in which the idea of reinforcement learning was introduced, where an agent learns to make optimised decisions by interacting with a dynamic environment (Sutton & Barto, 2018).

#### 2.3.3.1. Classification models

Classification machine learning models are algorithms that learn to assign labels or categories to data instances based on their characteristics. The main task is to map inputs to pre-determined categories, known as classes. These models are fundamental in situations aiming to make categorical predictions or determine to which class an instance belongs

(James et al., 2013).

The main classification models are:

i) Logistic Regression (RL): a supervised ML algorithm used for binary classification problems (Belyadi & Haghighat, 2021). Logistic regression essentially uses a nonlinear logarithmic odds ratio logistic function to model a binary output variable (Tolles & Meurer, 2016). Logistic regression normally has a classification between 0 and 1, not requiring a linear relationship between input and output variables.

ii) Support Vector Machines (SVM): their characteristic is to find the hyperplane that best separates the classes in the feature space, which can be linear or non-linear (Vapnik et al., 1995). SVM can model highly nonlinear processes without knowledge of the statistical distributions of classes. Another important property is its good generalisation performance even in the case of high-dimensional data and a small training set. Several works have shown the superiority of SVM classifiers over traditional statistical and neural classifiers (Sehad et al., 2017).

iii) Decision Tree (DT): tree structure that makes decisions based on conditions in attributes, which works by dividing the data set based on characteristics, forming a set of decision rules (Quinlan, 1986). It consists of inner nodes representing the structures of the branches, representing the verdict given by the algorithm, and each leaf node representing an outcome. The decision node, which is used to make a decision, has various branches, while the leaf node is the output of decision nodes and has no further branches (Bansal et al., 2022).

iv) Random Forest: an ensemble machine learning approach that aggregates the results of multiple decision tree models (Breiman, 2001). This model creates several trees and aggregates their results to reduce overfitting (increase variance) and improve generalisation.

v) Gradient boosting (GB): consists of several decision trees that are built sequentially, where each tree is trained to correct the errors made by the previous ones, allowing iterative improvement of the model's performance using relatively few trees (Friedman, 2001).

vi) Stochastic Gradient Descent (SGD): a widely used in training machine learning model due to its efficiency (Amari, 1993), whose approach is to update the model parameters with each training example in a stochastic way, making it o suitable for large data sets (Meng et al., 2019; Mu et al., 2017).

vii) K-Nearest Neighbour (KNN): is a non-parametric supervised learning technique used

in classification and regression problems, which classifies instances based on the class of the majority of the k nearest neighbours in the feature space (Altman, 1992). This is characterised by its simplicity and flexibility, as it does not require the assumption of specific data distributions. KNN calibration is crucial in the appropriate choice of the value of k, the number of neighbours considered, which directly affects the sensitivity of the model to specific patterns and is essential to avoid overfitting or underfitting, ensuring the robustness and effectiveness of the algorithm in the task at hand (Rajagopalan & Lall, 1999).

viii) Artificial Neural Network (ANN): The algorithm is an information processing paradigm inspired by biological neural networks (Gardner & Dorling, 1998). The basic elements of ANN are neurons (or units), which are interconnected by weighted links. In each unit, the output is calculated by a transfer (or active) function of the weighted sum of the inputs. It has a three-layer structure (i.e. input, hidden, and output layers), which is one of the widely used forms of ANN algorithms.

#### 2.3.3.2. Regression models

In statistics and machine learning, regression models are techniques that aim to model the relationship between a dependent variable and one or more independent variables (predictors). Regression is widely used to predict or understand how a change in the independent variables may affect the dependent variable (James et al., 2013).

Some regression models have already been mentioned in the previous subitem of this literature review, as many of them can be used for both functions. As regression models, we can mention:

i) K-Nearest Neighbour (KNN): is a non-parametric supervised learning technique used in classification and regression problems (Altman, 1992). Regression KNN is an extension of classification KNN for problems where the dependent variable is continuous. Instead of assigning a class to the instance based on the majority of k nearest neighbours, regression KNN calculates an average of the dependent variable values of the nearest instances (James et al., 2013).

ii) Linear regression (LRi): is a statistical technique for modelling the relationship between a dependent variable and one or more independent variables (Yang & Chen, 2023). The essence of this method lies in the search for a line of best fit that minimises the sum of the squares of the differences between the predicted and observed values.

### 2.3.4. Overfitting e Underffiting

The concepts of overfitting and underfitting are related to the model's ability to generalise from training data to new data but in different contexts. Overfitting (Figure 3) occurs when a model overfits the training data, capturing specific patterns, even noises, which may not be representative of the true relationship between the input and output data, resulting in a model that does not generalise well for new data (James et al., 2013). When overfitting, complex models can get lost in high-dimensional spaces, memorising instead of learning patterns (BISHOP, 2006). Underfitting occurs when a model is too simple to capture the complexity of the training data. This results in a lack of adaptation to the standards, making the model ineffective to generalise, even on training data (James et al., 2013). Underfitting is easier to identify, as the model does not present good results even for the training data. Good fit is ideal, as the model can adjust to the complexity of the training data and can generalize well to the test data.

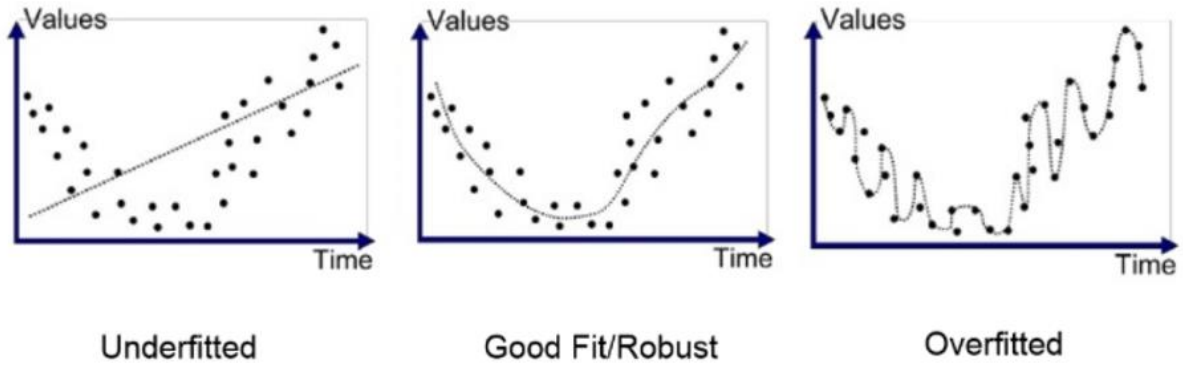


**Figure 3** – Overfitting, good fit, and underfitting in machine learning models (Raghav, 2022).

## 3 STUDY AREA

Brazil is a continental-sized country covering approximately 8.5 M km² between the latitudes 5°16′N - 33°45′S and longitudes 34°47′W – 73°59′W (Figure 4). The Brazilian territory is divided into five official geographic regions: South (S), Southeast (SE), Central-West (CW), North-East (NE), and North (N). Due to its large territorial extension, Brazil covers different climatic zones and precipitation patterns. According to Alvares et al. (2013), the Brazilian territory encompasses twelve of Köppen's climate types, divided into three main zones (Tropical, Semi-arid, and Humid Subtropical) (Figure 4), with a mean annual air temperature of approximately 10-26 °C. The annual rainfall in Brazil is characterised by high spatial variability, with values ranging from 380 mm (semi-arid climate in the NE) to 4,000 mm (tropical forest in the N). According to the mean monthly precipitation distribution, the Brazilian territory can be divided into five regions with homogeneous characteristics (Rozante et al., 2018) (Table 1).

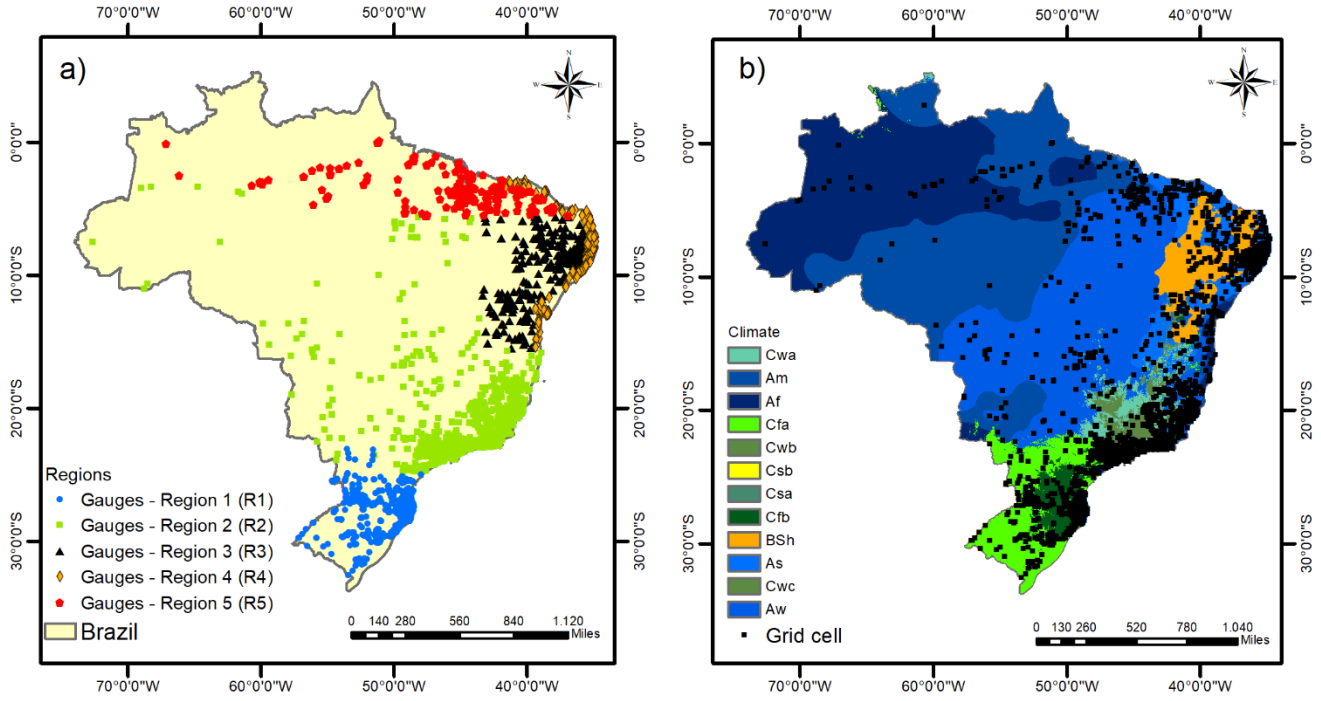


**Figure** 4 – Study area showing (a) the rain gauges used in the study, identified according to the five homogeneous regions defined by Rozante et al. (2018), in terms of precipitation, and (b) grid cells obtained from rain gauges to match the IMERG Early Run 0.1°.

**Table 1 –** Aspects of the five regions with homogeneous mean monthly rainfall characteristics in Brazil according to Rozante et al. (2018).

| Region | Number of grid cells | Köppen classification | Climatic features |
|---|---|---|---|
| R1 | 559 | C | This group is influenced by the Pre-frontal Storm Line (PSL) and Cold Fronts (FC) climatic systems (Reboita et al., 2010). The monthly and annual precipitation averages are around 120 and 180 mm, respectively. The mean annual air temperature is around 22º C. |
| R2 | 1530 | C | The main climatic systems influencing this group are the South Atlantic Subtropical Anticyclone (SASA) and the South Atlantic Convergence Zone (ZCAS) climatic systems (Reboita et al., 2010). The annual average precipitation is around 1,500 mm, mostly from November to February. The mean annual air temperature is around 22º C. |
| R3 | 299 | Bsh | The main climatic systems influencing this group are the South Atlantic Subtropical Anticyclone (SASA) and the South Atlantic Convergence Zone (ZCAS) (Reboita et al., 2010). The annual average precipitation is around 820 mm, mainly from December to May. The monthly mean air temperature ranges between 18 and 24 °C throughout the year. |
| R4 | 322 | AS | This group is influenced by the atmospheric systems of Instability Lines (LI) and the Southeast Trade Winds (SETW) (Reboita et al., 2010). This mean annual precipitation is around 1600 mm. The mean annual air temperature is above 23 °C. |
| R5 | 211 | Am and Aw | The main atmospheric systems acting in this group are the Intertropical Convergence Zone (ITCZ), the Instability Lines (LI), the Mesoscale Convective Complex (MCC), the South Atlantic Convergence Zone, the Upper-Level Cyclonic Vortices (ULCV), and the Easterly Wave Disturbances (Lemos et al., 2023). The mean annual air temperature is around 26 °C. The mean annual precipitation reaches 2600 mm, mainly from January to April. |

# 4 AN IMPROVED GRIDDED MONTHLY RAINFALL PRODUCT

## 4.1. Contextualisation

In this chapter, we focus on the development of a more accurate monthly precipitation product for Brazil, the so-called IMERG BraMaL, which uses IMERG Early Run and a re-analysis database as the input of a machine learning approach based on a regression model, without dependence on local or regional data. The general and transferable strategy of this new product would also be relevant to account for: (i) reducing the latency period for a satellite-based product with improved precipitation estimates (e.g. compared with the IMERG Final Run, which is currently available 3.5 months after observation); (ii) correcting well-known errors and biases in satellite-based precipitation products at monthly scale in some regions of Brazil (Freitas et al., 2020; Gadelha et al., 2019; Rozante et al., 2010); and (iii) not relying on observed data, i.e. being general enough to be applied in other larger regions with diverse climates, terrains, and precipitation regimes.

## 4.2. Materials and methods

The methodology proposed in this study considers estimated data from satellite-based (rainfall) and re-analysis (meteorological variables) products, obtained between 2014 and 2021, as inputs to create the IMERG BraMaL product. Observed rainfall data (rain gauges) obtained for the same period were used to calibrate (70%) and validate (30%) the proposed regression model. The methodological steps detailed below, include: (a) the description of the observed and estimated dataset; (b) the prediction model used; and (c) the statistical metrics to evaluate the performance of the proposed product (Figure 5).

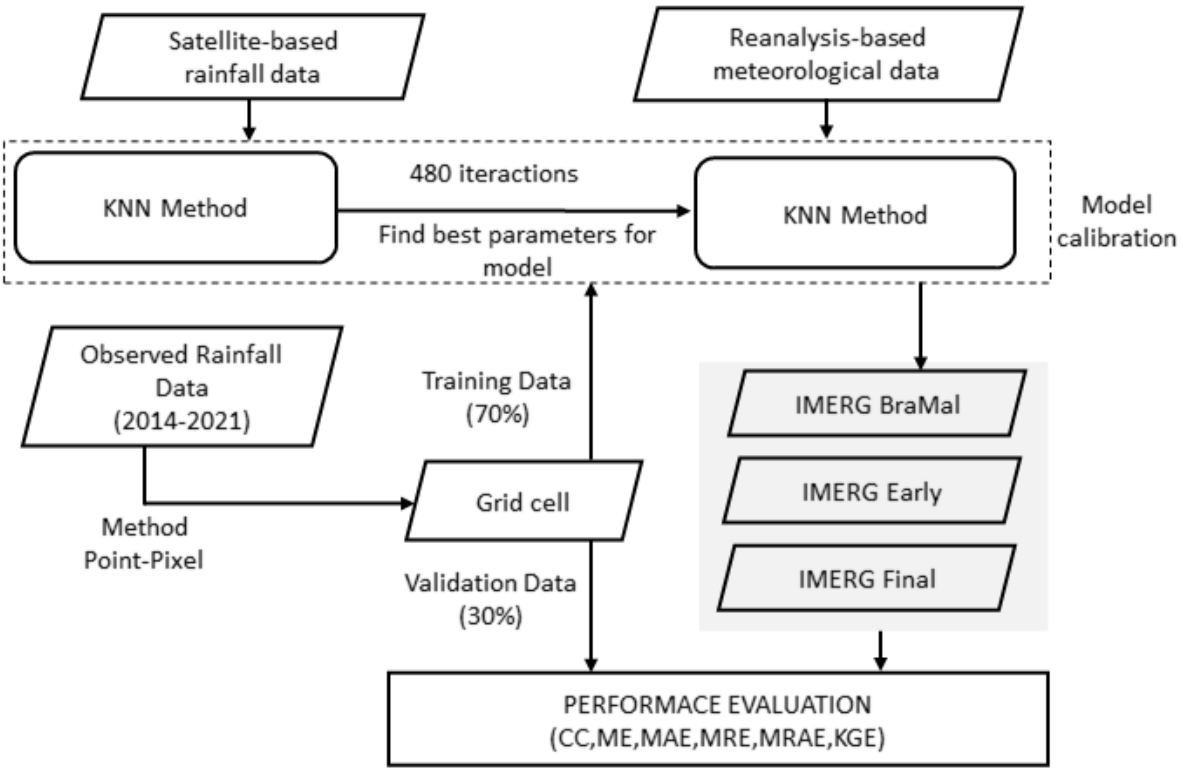


**Figure 5** – Schematic representation of the prediction process to create the IMERG BraMaL product.

### 4.2.1. Observed and estimated dataset

#### 4.2.1.1. Observed rainfall data

The rainfall data used in this study were obtained from automated rain gauges operated by the Brazilian Centre for Monitoring and Early Warnings of Natural Disasters (CEMADEN, an acronym in Portuguese). The ground-based rainfall observation network of CEMADEN is made up of tipping bucket gauges with a 10 min temporal resolution, when it rains, and 60 min, when there is no rain. In this study, we aggregated this high temporal resolution dataset to the monthly scale. Currently, CEMADEN operates more than 4,000 rain gauges distributed throughout Brazil, whose data are made available in UTC (Coordinate Universal Time) without quality control.

This study used rainfall data from 3,039 rain gauges distributed throughout the Brazilian territory, containing at least one calendar year of complete data from 1 January 2014 to 31 December 2021. Considering this monitoring period, the following number of rain gauges per monitoring year were available: 832 (1 year), 652 (2 years), 571 (3 years), 411 (4 years), 301 (5 years), 206 (6 years), and 66 (7 years). The selection of these rain gauges resulted from a strict quality control procedure, following the steps used by Freitas et al. (2020), which included: (i) the analysis of the amount of data recorded by the gauges,

considering unsuitable rain gauges with more than 60-days of missing data in each analysed year; (ii) the comparison of these stations selected in the first step with their five closest neighbours, based on a visual analysis with a double-blind test and taking into account the monthly and instantaneous (10 min) precipitation data; and (iii) the checking of the range of values and changes over subsequent measurements to identify constant or null rainfall records that probably indicate gauge clogging.

#### 4.2.1.2. Rainfall satellite-based data

The Global Precipitation Measurement (GPM) mission launched its Core Observatory in 2014, to succeed the Tropical Rainfall Measuring Mission (TRMM), which began to provide rainfall and snowfall information globally with better temporal (half-hour) and spatial (0.1° × 0.1°) resolutions, via the Integrated Multi-satellitE Retrievals for GPM (IMERG) products (Skofronick-Jackson et al., 2018a, 2018b, 2017). IMERG obtains information about precipitation in the latitude band 90° North and 90° South, with temporal coverage from 1 June 2000 to the present.

This study used Version-06B of IMERG Early Run (i.e. with a latency of 4 hours to the end-users) as a baseline satellite rainfall product, to calibrate the proposed IMERG BraMaL. The gauge-calibrated IMERG Final Run (i.e. with a latency of 3.5 months), which performed better compared to the other two near-real-time IMERG products (Jiang et al., 2022; Ramadhan et al., 2022; Zhou et al., 2021; Wang et al., 2017; Sungmin et al., 2017), was used in addition to IMERG Early Run for comparisons with the estimations of the IMERG BraMaL product. Additionally, the IMERG BraMaL product was compared with three consolidated global satellite-based products: i) Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks - Climate Data Record (PERSIANN-CDR), whose algorithm is calibrated by an artificial neural network using data from satellites to produce a product with 0.25º x 0.25º spatial resolution (Ashouri et al., 2015); ii) Climate Hazards Group InfraRed Precipitation with Station (CHIRPS), which incorporates interpolation techniques, satellite information of precipitation estimates, and station data to provides a product with grid cells of 0.05º x 0.05º (Funk et al., 2015); and iii) Multi-Source Weighted-Ensemble Precipitation MSWEP, which is derived from the optimal fusion of a series of measurement, satellite, and reanalysis estimates to provides a product with 0.1º x 0.1º spatial resolution (Beck et al., 2019).

#### 4.2.1.3. Reanalysis meteorological data

The re-analysis meteorological dataset obtained from the Modern-Era Retrospective analysis for Research and Applications, version 2 (MERRA-2), was used as input data to calibrate the proposed IMERG BraMaL product. MERRA-2 is the latest atmospheric re-analysis of the modern satellite era, produced by NASA's Global Modelling and Assimilation Office (GMAO), which uses the Goddard Earth Observing System (GEOS) model and analysis scheme to provide a viable ongoing climate analysis (Gelaro et al., 2017). Some recent studies verified the reliability of the MERRA-2 dataset, which presented a good overall agreement with the observed data (Huang et al., 2022; Guo et al., 2021; Zhang et al., 2020), as shown by Reichle et al. (2017) for runoff, rainfall, and soil moisture.

MERRA-2 provides global information with monthly temporal resolution from 1 January 1980 to the present, using a cubed-sphere horizontal discretisation at an approximate resolution of 0.500° × 0.625° and 72 hybrid-eta levels from the surface to 0.01 hPa (Gelaro et al., 2017). The latency of the MERRA-2 product is approximately three weeks after the end of each month, which will constrain the latency of IMERG BraMaL to 3 weeks after the event. We considered for each grid cell 53 MERRA-2 monthly variables with a particular connection to precipitation (i.e. its formation and impacts on nature) (Table 2). These variables can be divided into three large groups of data: single-level variables, Earth surface forced variables, and radiation diagnostic variables.

**Table 2** – Variables of MERRA-2, IMERG and CEMADEN used for model calibration

| Type | Data | Type | Spatial Resolution | Temporal Resolution | Period | Latency |
|---|---|---|---|---|---|---|
| Observed data | CEMADEN | Gauges | Gauge | 10 min | 2014 to present | 1 day |
| Satellite data | IMERG Early | precipitation estimate | 0.1° x 0.1° | Monthly | 2000 to present | 12 hours |
| | IMERG Final | precipitation estimate | 0.1° x 0.1° | Monthly | 2000 to present | 3.5 months |

| | | | | | | |
|---|---|---|---|---|---|---|
| Reanalysis meteorological data | MERRA- 2 | surface pressure; specific humidity*; air temperature*; total precipitable ice, liquid and vapor water; surface skin temperature; eastward wind**; northward wind**; accretion loss of cloud water to rain; convective source of cloud ice and water; convective production of rain water; vertically integrated water vapor tendency due to analysis, chemistry and dynamics; vertically integrated water vapor tendency due to moist processes, physics and turbulence; evaporation from turbulence, loss of cloud water and loss of precipitation water; eastward flux of atmospheric liquid water and water vapor; northward flux of atmospheric liquid water and water vapor; liquid water convective precipitation and large scale precipitation; cloud area fraction for high, low and middle clouds; total cloud area fraction; in cloud optical thickness of low, middle and all clouds; cloud top pressure and temperature; height***; specific humidity****; air temperature****; eastward and northward wind****; air temperature*****; | 0.5° x 0.65° | Monthly | 1980 to present | 3 weeks |

* 2 e 10 meters
** 2, 10 e 50 meters
*** 250, 500, 850 and 1000 hpa
****250, 500 and 850 hpa
*****500 and 850 hpa

### 4.2.2. Model calibration to reduce the errors in rainfall estimates

The model used the IMERG Early Run and MERRA-2 products for the calibration parameters through the k-nearest neighbours (KNN) algorithm, a non-parametric supervised machine learning technique (Altman, 1992). The model runs on a monthly basis to minimise the magnitude of the mean squared residual errors, producing the monthly IMERG BraMaL rainfall product. KNN is an algorithm that identifies K samples in the training dataset (whose independent variables are similar to the target values) and uses the average of these K samples to perform classifications or regressions (Alizadeh & Nikoo, 2018).

We used 55 variables for the model calibration, 54 inputs (53 from MERRA-2 and 1 from IMERG Early Run) and 1 output (monthly rain gauge observations). The observed rainfall data was only used for the calibration of the IMERG BraMaL product, which enables its application in regions with sparse or unavailable rain gauge networks after this step.

The observed point-scale rainfall was converted into grid cells to match the 0.1° × 0.1° IMERG Early Run grid. For grid cells containing more than one rain gauge, the average of the observed rainfall data was considered. In total, 1846 grid cells were used as input data, subdivided into training (70%) and test (30%) datasets to calibrate and validate the model, respectively, as performed by Zhang et al. (2021). In contrast to Zhang et al. (2021), however, we considered this subdivision randomly in time (monthly) and in space (indirectly), i.e. the same grid cell can be used for training and testing simultaneously but for different periods (months and years). The data model was developed using the Scikit-learn library: Machine Learning in Python (Pedregosa et al., 2011). Before the calibration, all input data were standardised and normally distributed using the following equation:

$$z=(x-u)/s \quad (1)$$

where z is the standardised value, x is the value to be standardised, and u and s are the mean and the standard deviation of the data, respectively. This process prevents data of different magnitudes from unequally influencing the determination of neighbours and distances calculated by the model.

To avoid overfitting, we attempted to extract the better calibration parameters for the model by testing 480 variations of the parameters, which returned the best performance with K=13 neighbours, the Euclidean distance, and weight points by the inverse of their distances for each variable. After calibration, the model scores were 0.98 and 0.81 for the training and testing data, respectively. The calibrated model was then evaluated.

### 4.2.3. Performace evaluation metrics

Four statistical metrics were used to compare the estimations of the IMERG BraMal product with the ground-based rainfall observations (i.e. the validation data) and the original IMERG products (Early and Final Run), which includes the Error, the relative error (RE), the mean absolute error (MAE), the mean relative absolute error (MRAE), the Pearson's coefficient correlation (CC), and the Kling-Gupta Efficiency (KGE):

$$Error = E_i - O_i \quad (2)$$

$$RE = \frac{100 \sum_{i=1}^{n} E_i - O_i}{O_i} \quad (3)$$

$$MAE = \frac{\sum_{i=1}^{n} |E_i - O_i|}{n} \quad (4)$$

$$MRAE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{O_i - E_i}{O_i} \right| \quad (5)$$

$$CC = \frac{\sum_{i=1}^{n} (O_i - \bar{O})(E_i - \bar{E})}{\sqrt{\sum_{i=1}^{n} (O_i - \bar{O})^2} \cdot \sqrt{\sum_{i=1}^{n} (E_i - \bar{E})^2}} \quad (6)$$

$$KGE = 1 - \sqrt{(1 - CC)^2 + (1 - Beta)^2 + (1 - Gama)^2} \quad (7)$$

$$Beta = \frac{\bar{E}}{\bar{O}} \quad (8)$$

$$Gama = \frac{\sigma E}{\sigma O} \quad (9)$$

where O is the rain gauge-observed rainfall data, $\bar{O}$ is the mean rain gauge-observed rainfall data, E is the rainfall data estimated by the products, $\bar{E}$ is the mean rainfall data estimated by the products, σ is the standard deviation, i is the time step, and n is the total number of compared pairs. MRE and MRAE measure the accuracy of the IMERG BraMaL product, with values close to zero indicating smaller errors. The CC ranges from -1 to +1, where extreme values represent total negative and positive linear correlations, respectively. The KGE values range from -∞ to 1, with desirable values close to 1 and negative values representing worse performances.

The evaluation procedures were performed considering the following four perspectives: (i) an overall national-scale analysis, considering all monthly rainfall values for the studied period; (ii) a seasonal analysis, considering the rainfall data month by month throughout the analysed time series; (iii) a spatial analysis of the errors; and (iv) a group analysis based on regions with the homogeneous mean monthly rainfall characteristics defined by Rozante et al. (2018).

## 4.3. Results and discussion

### 4.3.1. National-scale comparison of products with field-based data

Figure 6 shows the scatter plots considering all 1846 grid cells distributed throughout the country. Overall, the Early Run product presented the largest dispersion (CC = 0.69), followed by the Final Run product (CC = 0.86). This significantly lower dispersion observed in Final Run can be attributed to the product calibration incorporating observed rainfall data

(Sungmin et al., 2017). Both original IMERG products tended to underestimate the observed data, as shown by the greater number of points concentrated below the line of equality. Similar underestimations were also identified for the monthly analyses of the IMERG Final and Early Run products in the studies carried out by Wang et al. (2017) and Zhou et al. (2021) in China. The IMERG Early and Final run products exhibited MAE and MRAE values ranging from 52.21 mm (Early Run) to 33.80 mm (Final Run) and from 1.14 (Early Run) to 1.04 (Final Run), respectively. Previous studies also found similar differences between the monthly errors from the IMERG Final Run and Early Run products, compared to gauge observations (Chen et al., 2022; Wang et al., 2021; Guo et al., 2019).

Overall, the IMERG BraMaL product performed better than the two original IMERG estimates, with the lowest errors (MAE = 29.1 mm and MRAE = 0.86) and a higher positive correlation (CC = 0.91). The KGE scores confirm that the BraMaL product (KGE = 0.86) provides a better estimation of precipitation than the Final Run (KGE = 0.82) and Early Run (KGE = 0.70) products, i.e. presenting the same variability magnitude of the rain gauge measurements and a better agreement with the observed data. Bhuiyan et al. (2020) showed that the modified IMERG product using ML techniques for a river basin in Bangladesh, presented a smoother reduction of MAE for all quartiles evaluated (from 3.75-3.25 mm) when compared to the IMERG Late Run product, i.e. 64% lower than the improvement for IMERG BraMaL. Zhang et al. (2021) revealed an improvement of the best-modified product created by DML techniques on the Chinese mainland, with CC increasing from 0.64 to 0.78 and KGE from 0.54 to 0.71, compared to IMERG Early Run, i.e. lower than that reached with the BraMaL product. Overall, it was observed that the BraMaL product presented better performance than other studies in our literature review.

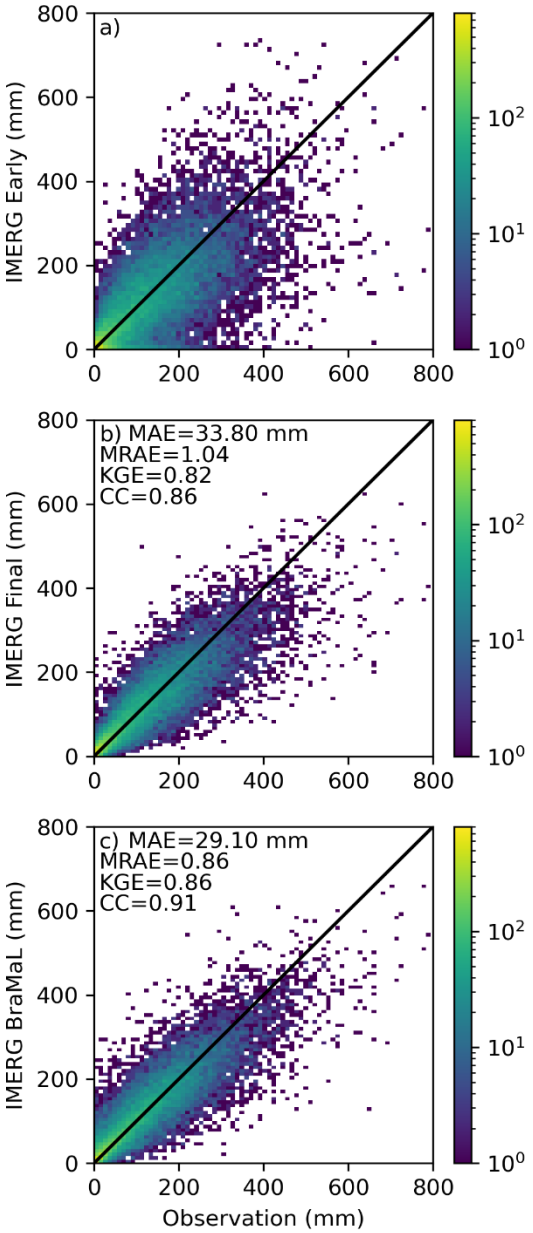


**Figure 6** – Scatter plots of the grid cell precipitation values vs (a) IMERG Early Run, (b) IMERG Final Run and (c) IMERG BraMaL, considering the national-scale analysis. The colours represent the number of events, from lower (violet) to higher (yellow).

Figure 7 shows the MAE and MRAE values for different monthly rainfall intervals. Overall, the IMERG BraMaL product exhibited lower relative and absolute errors than IMERG Early Run for almost all rainfall intervals, especially between 0-550 mm. The MAE values of the BraMaL product were similar to IMERG Final Run for monthly rainfall intervals lower than 200 mm and lower than the same gauge-calibrated product for overall monthly rainfall intervals between 200-550 mm. The values of MRAE for the IMERG BraMaL product are lower than the IMERG Final Run for monthly rainfall intervals between 50-550 mm, showing that the error's magnitude of BraMaL is lower for most rainfall

intervals. For some monthly rainfall intervals (e.g. > 550 mm), the MAE and MRAE values of the two original IMERG products tend to be close to BraMaL errors.

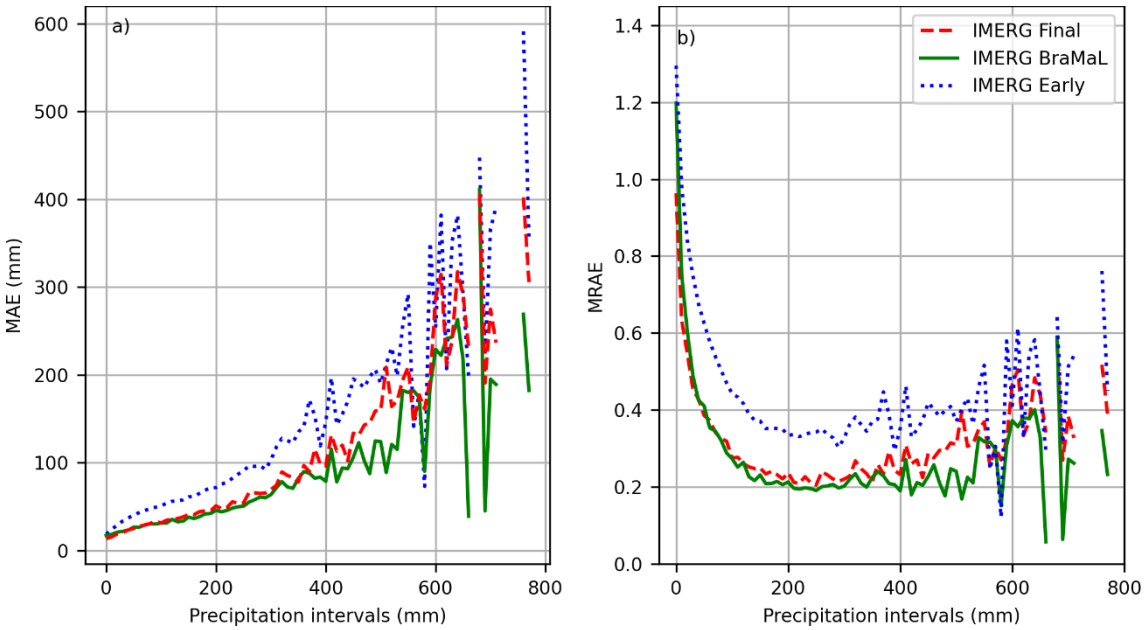


**Figure 7 -** (a) MAE and (b) MRAE values of the IMERG Early, IMERG Final and IMERG BraMaL products for different monthly precipitation intervals.

### 4.3.2. Seasonal influence on product quality

In this analysis, the IMERG-based products (i.e., Early Run, Final Run, and BraMal) were evaluated against the observed rainfall data month by month throughout the analysed time series (2014-2021). Figure 8 shows the scatter plots for each rainfall product against the observed data for each month of the year. Overall, it is possible to observe a greater density area of points (i.e. in yellow) concentrated close to the line of equality for all products, which is more explicit from February to May for IMERG BraMaL. The IMERG Early Run product presented more dispersed points for all months, clearly exhibiting a greater tendency to underestimate the gauge data, especially from June to August, where monthly observed precipitation close to 500 mm was not detected by the satellite-based estimations. Overall, the Final Run product improved the Early Run estimates, reducing the data dispersion, mainly between June and August.

Figure 9 shows the values of the four statistical metrics for each product, exhibited month by month. The MAE values reveal that the IMERG BraMaL product presented lower errors than the two original IMERG products for all months, especially the Early Run product. Overall, the MAE values of the IMERG Early and Final Run products were 23 and 5 mm larger than the IMERG BraMaL product, with values of MAE for the Early Run reaching up to 60.9 mm in March. It is worth highlighting that, in August and September,

the MAE for the IMERG BraMaL product presented values below 20 mm. For the MRAE metric, the IMERG BraMaL product presented the same superiority as in MAE, especially in June and July, which reached MRAE values around 160 and 260%, respectively. However, these higher relative values occur in months with low-magnitude precipitations, i.e. not interfering with the application or use of the BraMaL product. The best CC and KGE values were observed for IMERG BraMaL in almost all months, but mainly from June to August, with CC ranging from 0.50 to 0.90 and KGE varying between 0.15 and 0.82. For these three months, the two original products presented the worst performances, especially the Early Run, with KGE and CC lower than 0.2 and 0.5, respectively. Although higher for almost all months, the CC of IMERG BraMaL presented values similar to the IMERG Final Run product. However, the KGE values confirm the efficiency of the IMERG BraMaL product monthly, with more substantial differences for the IMERG Final compared to the CC.

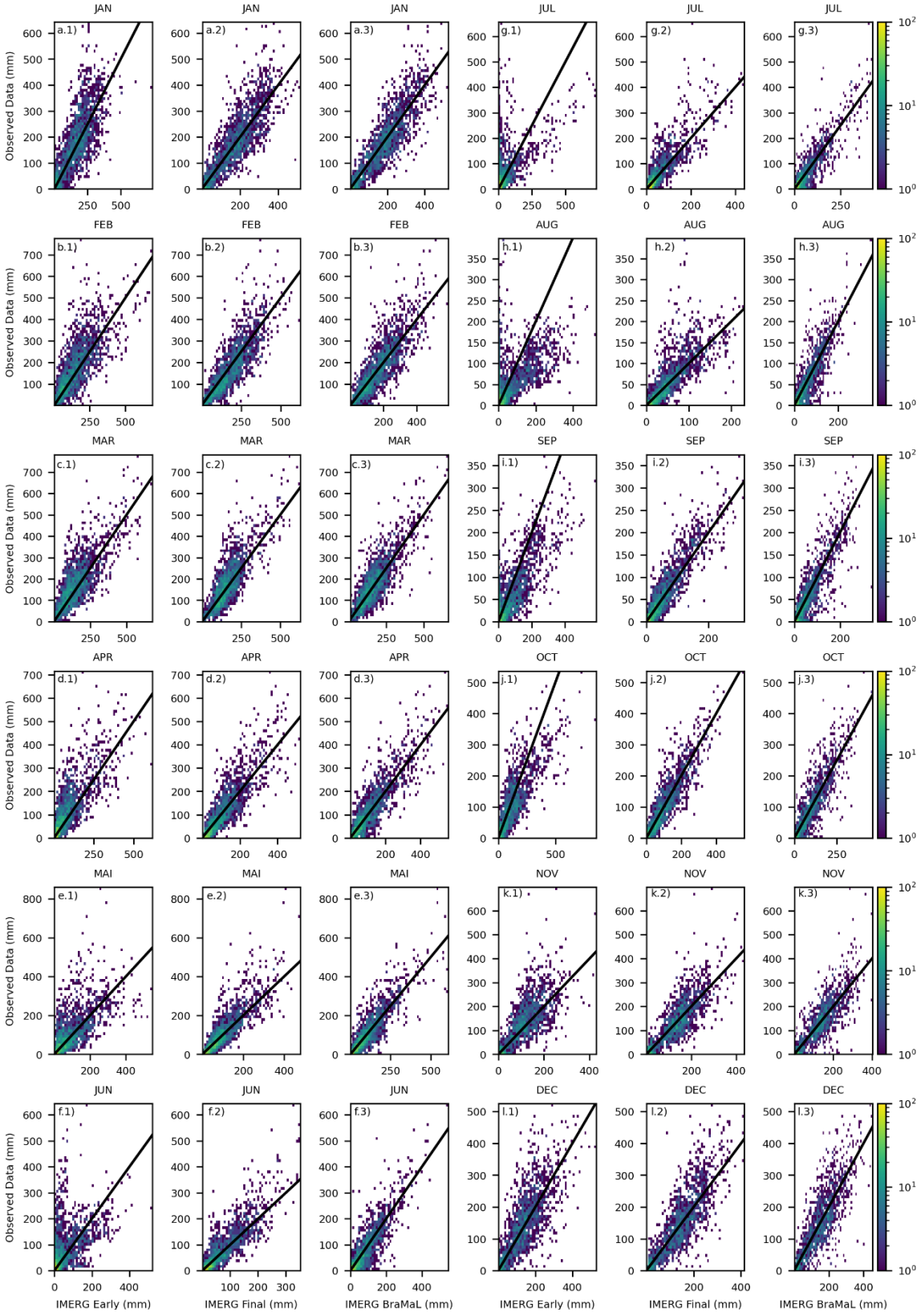


**Figure 8** – Scatter plots of the grid cell precipitation values vs variations of IMERG Early Run (first and fourth columns), IMERG Final Run (second and fifth columns), and IMERG BraMaL (third and sixth columns) estimates per month. The colours represent the number of events, from lower (violet) to higher (yellow).

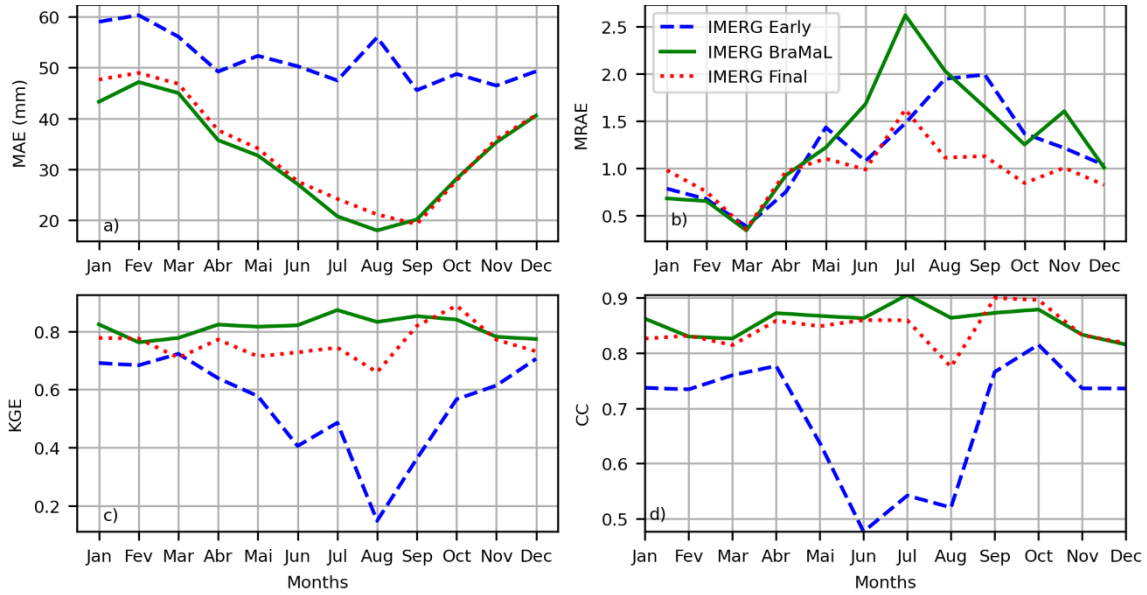


**Figure 9** – Monthly variations of (a) MAE, (b) MRAE, (c) KGE, and (d) CC for the IMERG Early Run, IMERG Final Run, and IMERG BraMaL products.

Figure 10 shows the MAE and MRAE values month by month, considering different intervals of monthly precipitation. Overall, it is possible to observe an increase in the MAE and MRAE values as the monthly rainfall accumulates. Once again, the absolute and relative errors of IMERG Early Run were considerably greater than the two other products, especially from June to November. The IMERG BraMaL product presented errors slightly lower than IMERG Final Run, except in two months for the MAE (October and November) and MRAE (January and February), where the errors were similar. The remarkably better performance of IMERG BraMaL means that the IMERG near-real-time product was well-corrected with the atmospheric parameters used in the calibration to produce the proposed product.

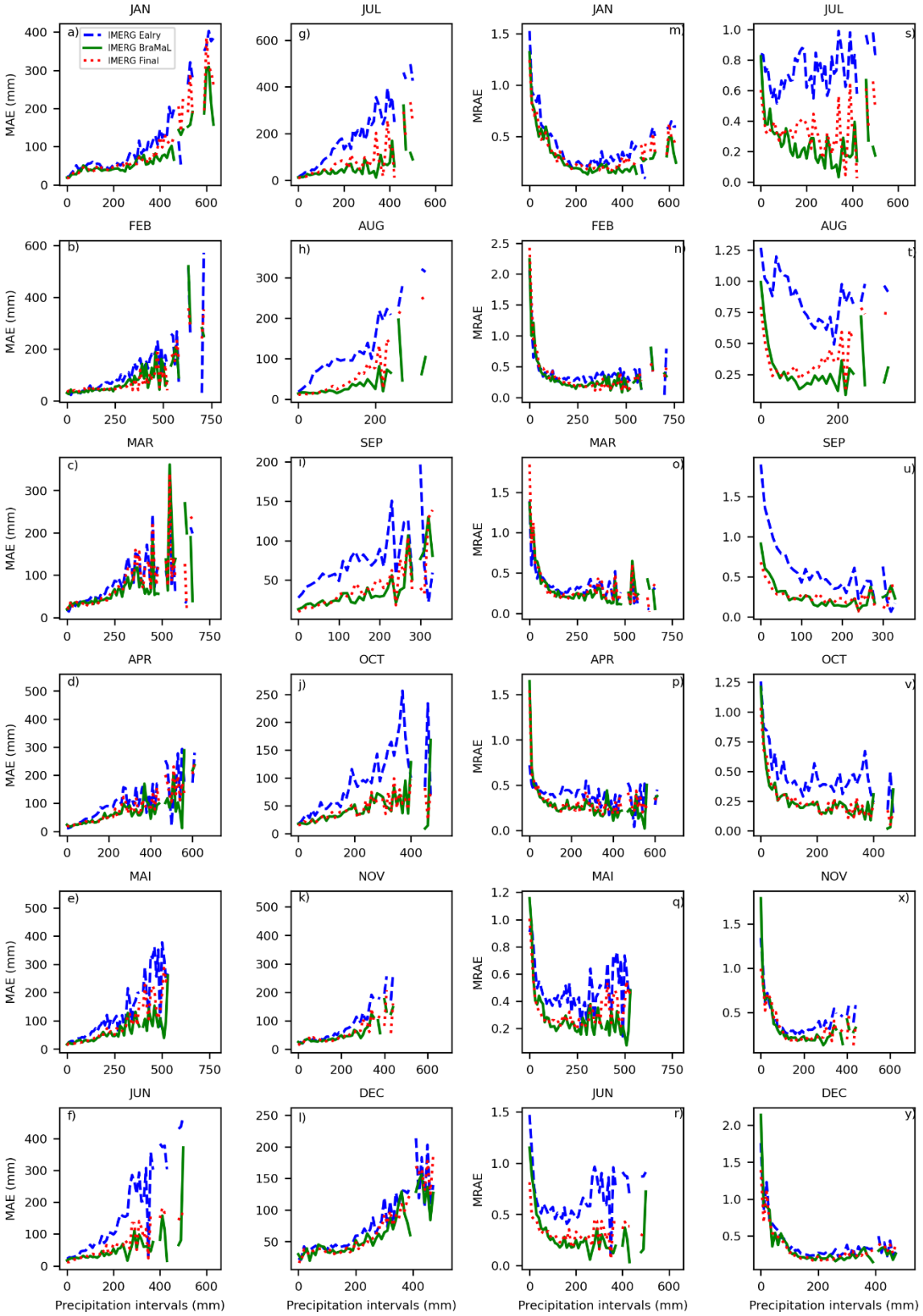


**Figure 10** – Monthly variations of the IMERG Early Run, IMERG Final Run, and IMERG BraMaL products for different intervals of precipitation: MAE from (a) January to December (l) and MRAE from (m) January to (y) December.

### 4.3.3. Spatial variability of product quality

This analysis evaluates the spatial distribution of the monthly MRAE of IMERG BraMaL over Brazil. Figure 11 shows that the REs are not concentrated in a specific region and are mostly between -50 and 50%. For instance, the studies by Rozante et al. (2018), Gadelha et al. (2019), and Freitas et al. (2020) identified remarkable variations in the performance of the IMERG Final Run product as a function of the analysed region in Brazil, with higher errors and biases in the N and CW regions, as well as along the Atlantic coast of the NE region. This indicates that the proposed model was able to correct the estimates for the IMERG BraMaL product across the whole of Brazil, a large country with diverse rainfall regimes and climates.

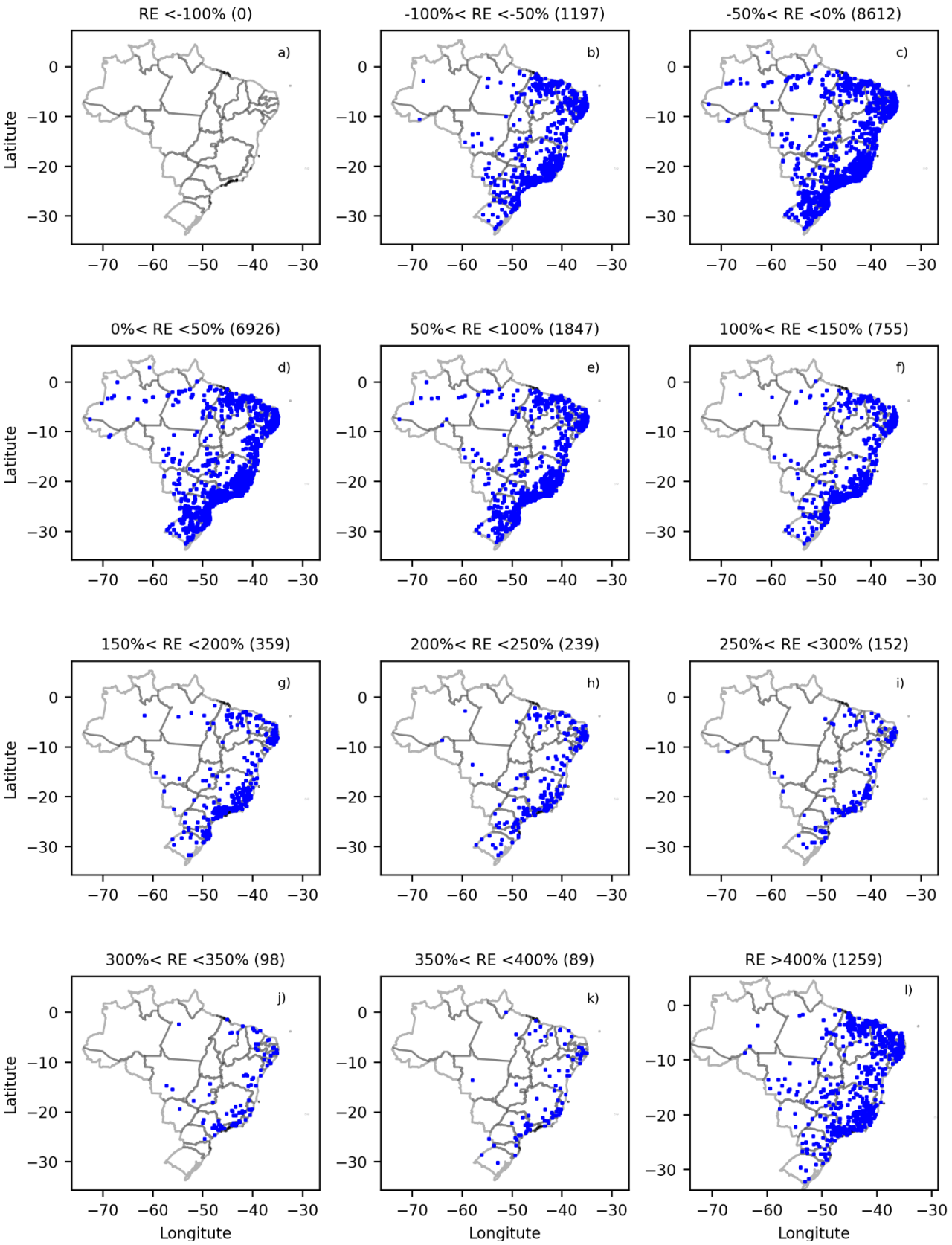


**Figure 11** – Spatial distribution of RE for the IMERG BraMaL product by interval classes: (a) RE < -100%, (b) -100% < RE < -50%, (c) 50% < RE < 0%, (d) 0% < RE < 50%, (e) 50%< RE < 100%, (f) 100% < RE < 150%, (g) 150% < RE < 200%, (h) 200% < RE < 250%, (i) 250% < RE < 300%, (j) 300% < RE < 350%, (k) 350% < RE < 400%, and (l) RE >400%. The number inside the parentheses identifies the quantity of grid cells.

Figure 12 shows that approximately 60% of the monthly precipitation values of IMERG BraMaL exhibit Errors smaller than 25 mm. When considering Errors smaller than 50 mm, this percentage rises to approximately 83% of the values. Regarding the RE, approximately 60% of the data present under-estimation or over-estimation errors, up to 25% of the monthly precipitation value. By expanding the analysis to an RE of up to 50%, approximately 80% of the data is in this range. This indicates that around 65-80% of the estimated precipitation could replace the field data without a notable change in the observed values, considering the RE of 25 and 50%, respectively. Only 6% of the data present RE exceeding 300%. However, as mentioned above, when evaluating the Errors of these events, a small difference in the precipitation amount is observed. These high RE occur in various parts of the territory and are mostly associated with low monthly precipitations, with an average of 30 mm and a standard deviation of 39 mm. It is important to emphasise that these estimates do not compromise the overall quality of the product due to the low magnitude of the values.

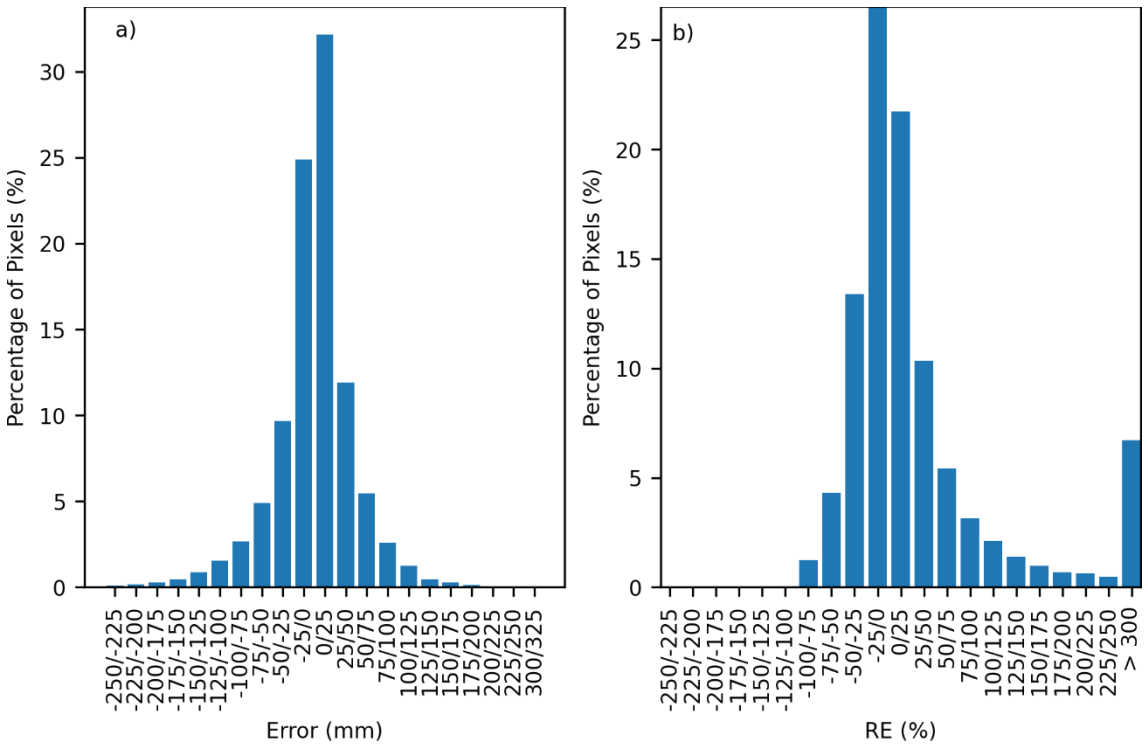


**Figure 12** – Histogram of the (a) Errors (mm) and (b) Relative Errors (%) of precipitation estimated by the IMERG BraMaL product.

### 4.3.4. Regional scale analysis of the product quality

In order to compare the performance of the IMERG BraMaL and the two original IMERG, we evaluated separately these products in the five regions with homogeneous monthly rainfall characteristics defined by Rozante et al. (2018). Figure 13 shows a large dispersion of data for the estimates of the IMERG Early Run product in almost all regions, except for R5, which had the least number of grid cells analysed (i.e. 1,782). A reduction of this dispersion is observed for IMERG Final Run, followed by an improvement in agreement with the observed data for the BraMaL product. Overall, the regions R1 and R2 (with 4,838 and 10,581 grid cells analysed, respectively) present higher densities of points concentrated along the 1:1 line, compared to other regions. On the other hand, the region R4 (with 2,248 grid cells analysed) exhibits several underestimations of the two original IMERG products, which other authors (Rozante et al., 2018; Gadelha et al., 2019; Freitas et al., 2020) attributed to the influence of convective clouds that occur along the Northeast coast of Brazil during the rainy period. These underestimations were mainly adjusted by IMERG BraMaL, demonstrating the robustness of the proposed product.

The variations of four metrics for the homogeneous regions are shown in Figure 14. A similar order of magnitude of MRAE was observed for all products. On the other hand, the MAE of IMERG Early Run presented higher values (from 76 to 43 mm) than the IMERG Final Run (from 43 to 26 mm) and IMERG BraMaL (from 41 to 29 mm) products, which were similar to each other, except for region R4. For region R4, IMERG BraMaL presented MAE of 37 mm, while the IMERG Final Run product exhibited MAE of 43 mm, indicating a better overall performance of the proposed product. The other two performance metrics confirm that the region R4 showed improvement of BraMaL (KGE = 0.80 and CC = 0.86) compared to the IMERG Early (KGE = 0.19 and CC = 0.56) and Final Run (KGE = 0.57 and CC = 0.86). This improvement in performance in R4 was mainly observed during the rainy season, as shown in the seasonality analysis. In contrast to the study performed by Bhuiyan et al. (2018) in the Iberian Peninsula, with a study period spanning 11 years, the results observed in the region R4 show that the machine learning-based model can successfully correct the precipitation data with higher values.

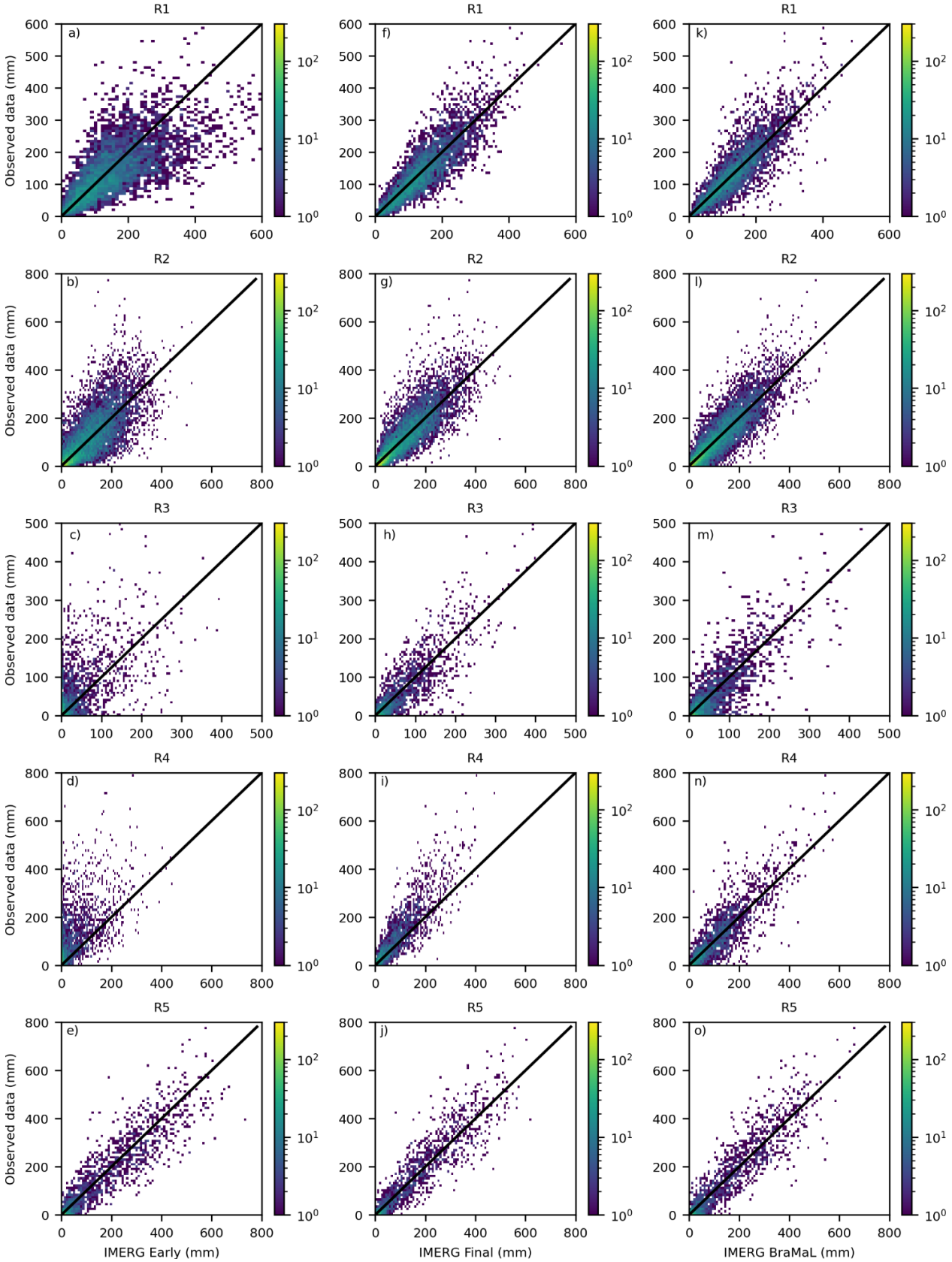


**Figure 13** – Scatter plots of the monthly precipitation observed values vs IMERG Early Run (first column), IMERG Final Run (second column), and IMERG BraMaL (third column) by homogeneous regions (from the top to bottom: R1 to R5). The colours represent the number of events, from lower (violet) to higher (yellow).

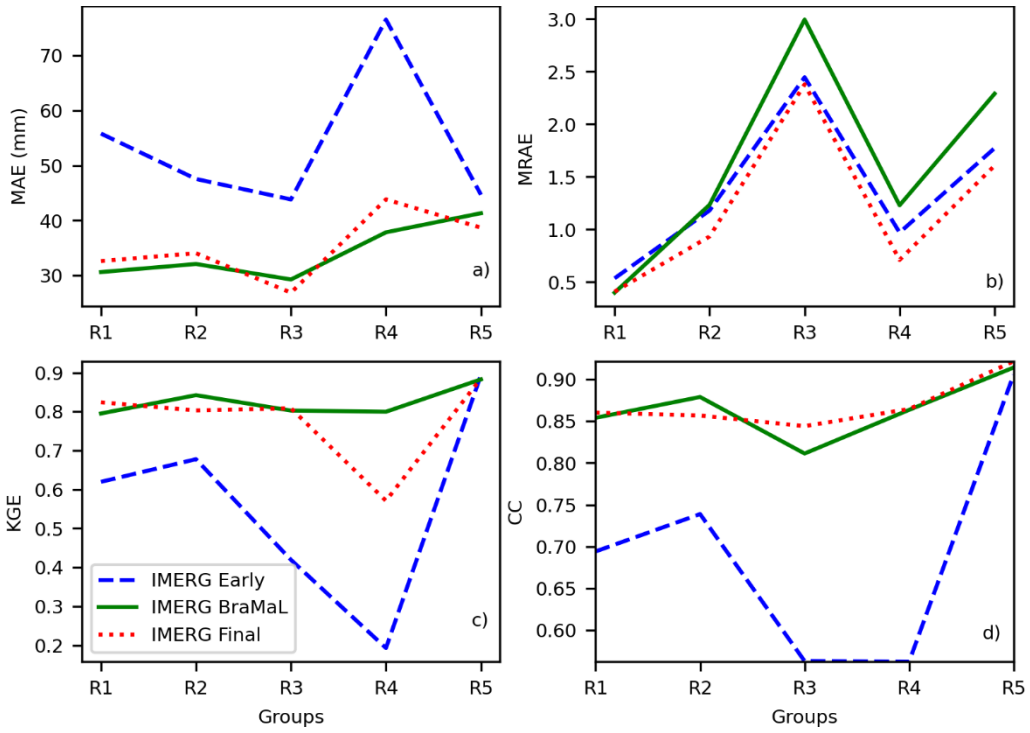


**Figure 14** – Statistical indexes (a) MAE, (b) MRAE, (c) KGE, and (d) CC for the IMERG Early Run, IMERG Final Run and IMERG BraMaL per analysed region.

### 4.3.5. Comparison of IMERG BraMaL estimates with other global precipitation products

The comparison of monthly precipitation estimates by IMERG BraMaL with the CHIRPS, PERDIANN-CDR, and MSWEP products for all 1846 evaluated gris cells show an overall lower dispersion of the data for the proposed product (CC = 0.91), especially when confronted with PERSIANN-CDR (CC = 0.68) (Figure 15). The PERSIANN-CDR and MSWEP (CC = 0.78) products presented an overall trend to underestimate the observed data, while CHIRPS (CC = 0.86) exhibited a better agreement with the rain gauges, with a higher density of points along the 1:1 line. The KGE scores attest that IMERG BraMaL (KGE = 0.86) presents a better rainfall estimation when compared to PERSIANN-CDR (KGE = 0.09), MSWEP (KGE = 0.60), and CHIRPS (KGE = 0.81). A similar lower agreement between PERSIANN-CDR and the observed data was also identified by Ramos et al. (2020) when evaluating the ability of 14 satellite-based precipitation products to characterise extreme events that trigger floods.

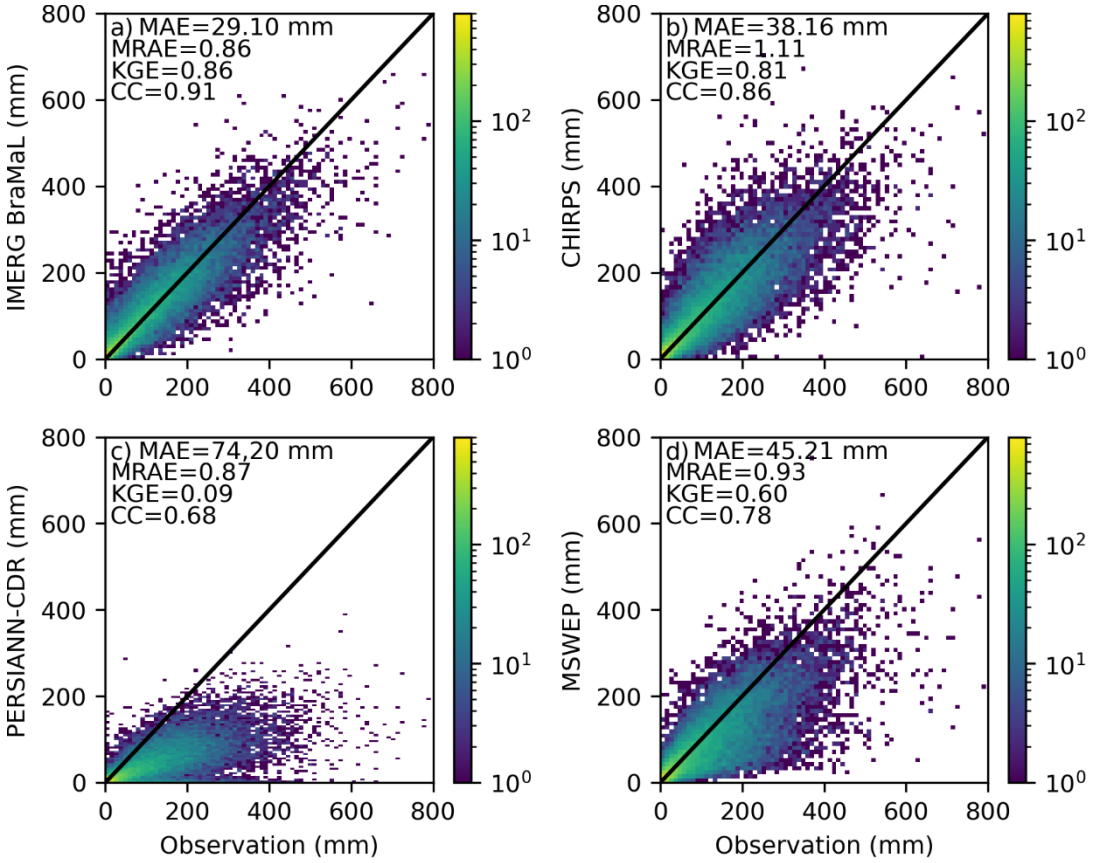


**Figure 15** – Scatter plots of the monthly precipitation observed values vs (a) IMERG BraMaL, (b) CHIRPS, (c) PERSIANN-CDR, and (d) MSWEP. The colours represent the number of events, from lower (violet) to higher (yellow).

Regarding the errors, IMERG BraMaL presented lower MAE (29.1 mm) and MRAE (0.86) in relation to CHIRPS (MAE = 38.16 mm and MRAE = 1.11), PERSIANN-CDR (MAE = 74.20 mm and MRAE = 0.87), and MSWEP (MAE = 45.21 mm and MRAE = 0.93). Such performance metrics confirm that IMERG BraMaL better estimates the monthly rainfall when compared to the main global satellite-based precipitation products. For instance, the evaluation of 10 satellite-based products carried out by Wati et al. (2021) in Indonesia found lower values of error-based metrics for CHIRPS, PERSIAN-CDR, and MSWEP, highlighting the ability of IMERG BraMaL to estimate more accurately the monthly rainfall.

# 5 EVALUATION OF SINGLE AND COMBINED MACHINE LEARNING MODELS TO IMPROVE DAILY RAINFALL ESTIMATIONS

## 5.1. Contextualisation

Machine learning algorithms proved to be a good tool for improving satellite-based precipitation, as shown in Chapter 4 and some other studies (e.g. Sengoz et al., 2023; Tyralis et al., 2023). However, some previous studies also identified that the combination of machine learning models can further increase its performance. Currently, there are three different procedures available in the literature to combine machine learning algorithms for improving satellite-based precipitation data, which include: i) a combination between classification and regression algorithms (Zhang et al., 2021); ii) a combination and stacking of regression algorithms (Montero-Manso et al., 2020; Papacharalampous et al., 2023c); and iii) the use of hybrid algorithms (Di Nunno et al., 2022).

Although presenting satisfactory results, the studies combining machine learning algorithms still present some limitations, as they: i) still rely on the use of specific regional ground-based data for the studied region, such as rain gauge observations (Bhuiyan et al., 2019, 2020b; Papacharalampous et al., 2023a, 2023b); ii) use only the mean and the median of the model outputs to obtain the final precipitation estimations (i.e. without considering that each model contributes with different weights) (Papacharalampous et al., 2023c); iii) use predicted data in a model together with their corresponding true values as inputs for other models (i.e. propagating the errors for training steps) (Papacharalampous et al., 2023c); and iv) consider a pre-defined and limited combination of classification and regression algorithms (i.e. without evaluations of more appropriate algorithms) (Zhang et al., 2021).

In this chapter, we focus on improving the IMERG BraMaL product to produce more accurate daily precipitation estimations for Brazil, also based on using the IMERG Early Run and re-analysis database as input and without dependence on ground-based local or regional data. The general and transferable strategy of this chapte also accounts for i) evaluating the performance of single and combined machine learning algorithms to improve these precipitation estimates, ii) testing various combinations of machine learning models (regression, classification, and the combining regression-classification), and iii) considering different weights for each model used in the combination.

## 5.2. Materials and methods

### 5.2.1. Observed and estimated dataset

This study used precipitation data from 3,039 rain gauges operated by CEMADEN

throughout Brazil, covering the period from 01/01/2014 to 31/12/2022. The selection of such rain gauges involved a rigorous quality control process, ensuring the reliability of this data (Freitas et al., 2020), as previously detailed in section 4.2.1.1. Similarly to the monthly IMERG BraMaL product, Version-06B of the IMERG Early Run was also used as a baseline satellite precipitation product to calibrate the proposed daily IMERG BraMaL product, while Version-06B of the IMERG Final Run product was used for comparisons with the estimations of the IMERG BraMaL product (see item 4.2.1.2). Furthermore, the same 53 meteorological variables from the MERRA-2 reanalysis product were used to provide input data for calibration of the daily IMERG BraMaL, similar to the monthly basis product (see item 4.2.1.3).

### 5.2.2. Calibration of machine learning models

Nine classification and six regression machine learning models were trained and evaluated. Classification models: Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN), Artificial Neural Network (RNA), Stochastic Gradient Descent (SGD), Gradient Boosting (GB), and Xtreme Gradient Boosting (XGB). Regression models: Linear Regression (LRi), Decision Tree (DT), K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), Stochastic Gradient Descent (SGD), and Gradient Boosting (GB).

We used 61 variables for the model calibration, 60 inputs (57 from MERRA-2 and 3 from IMERG Early Run) and 1 output (daily rain gauge observations). Similarly to the monthly IMERG BraMaL, the observed rainfall data was only used for the model's calibration, enabling its application in regions with sparse or unavailable rain gauge networks after this step. In addition, the observed point-scale rainfall was also converted into grid cells to match the 0.1° × 0.1° IMERG Early Run grid for the calibration of the daily-based models. The average of the observed rainfall data was also considered for grid cells containing more than one rain gauge. For all models, 1,846 grid cells were used as input data, subdivided into training (70%) and test (30%) datasets to calibrate and validate the model, respectively (Zhang et al., 2021; Bansal et al., 2022; Nunno et al., 2022) (**Erro! Fonte de referência não encontrada.**). In contrast to Zhang et al. (2021), however, we considered this subdivision randomly in time (monthly) and in space (indirectly), i.e. the same grid cell can be used for training and testing simultaneously but for different periods (months and years).

Similarly to the monthly IMERG BraMaL, the data model was developed using the

Scikit-learn library: Machine Learning in Python (Pedregosa et al., 2011), and all input data were standardised and normally distributed before the calibration (see item 4.2.2). Three different methods were considered to define the best daily precipitation estimation product: i) single machine learning (SML), based on the calibration of a single regression model, ii) dual machine learning (DML), combining a classification model and a regression model, and iii) multiple machine learning (MML), which combines many regression models.

#### 5.2.2.1. Calibration of the simple machine learning (SML) method

For the SML method, six regression models (i.e. RLi, GB, DT, RNA, KNN, and SGD) were directly considered for the training data (Figure 16). After model calibration, the performance of the models was evaluated using statistical metrics applied to the test data. The simple regression model with the best daily precipitation estimates was selected and used to be further evaluated and compared to double and multiple machine learning techniques.

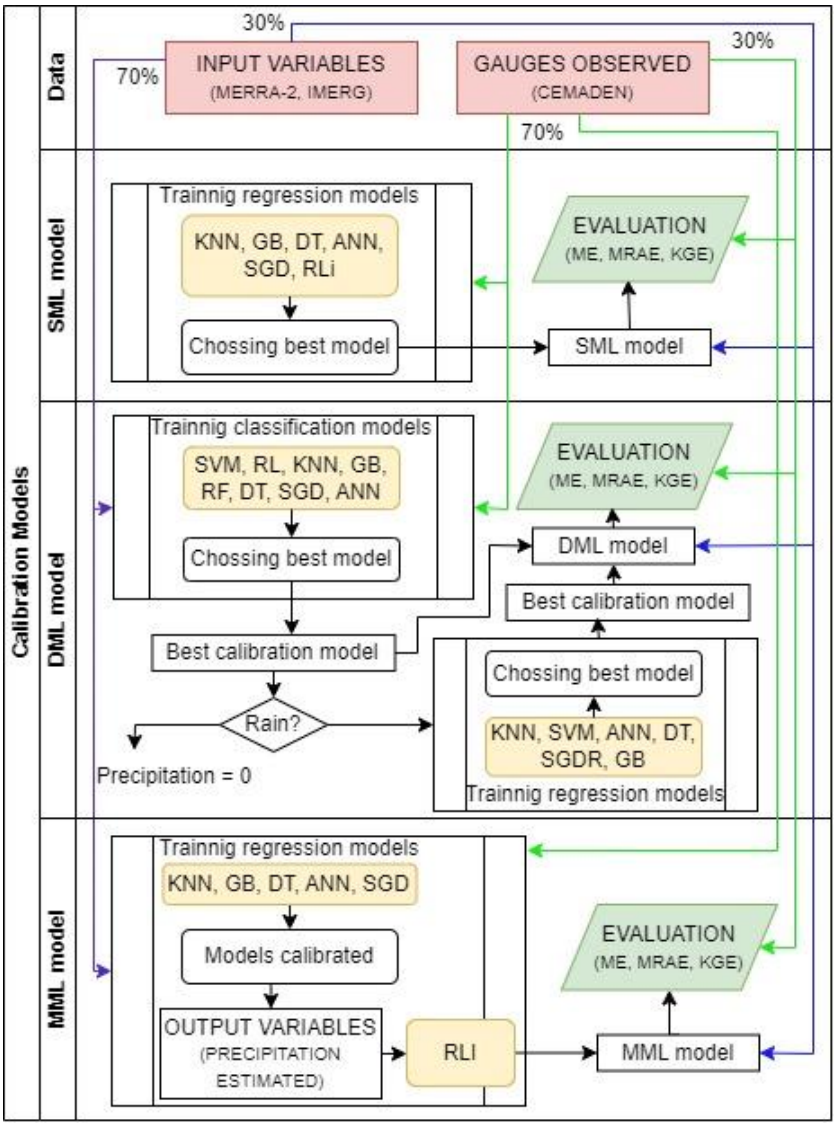


**Figure 16** – Schematic representation of training and testing the simple, double, and multiple machine learning models.

### 5.2.2.2. Calibration of the double machine learning (DML) method

For DML models, calibrations were first performed for the classification models (Figure 16), which assigned a value of 1 for observed precipitation greater than 5 mm/day and a 0 for those below this threshold. This first calibration aimed to allow the identification of significant daily rain by the classification models, as from the analysis of the CEMADEN dataset used in this study, precipitation higher than 5 mm/day comprises at least 80% of annual rainfall. Eight classification models (i.e. SVM, RL, KNN, GB, RF, DT, SGD, and ANN) were tested for this first calibration. The most efficient classification model was selected based on a performance evaluation to separate rain and non-rain events. The rain events were then calibrated by seven regression models (i.e. SVM, KNN, GB, DT, ANN, SGD, and RLi), whose performances for daily precipitation estimates were evaluated using

statistical metrics.

#### 5.2.2.3. Calibration of the multiple machine learning (MML) method

Similarly to DML, the MML models were calibrated into two steps (Figure 16). Firstly, five regression models (i.e. KNN, GB, DT, SGD, and ANN) were trained using both the input estimated variables (i.e. MERRA-2 and IMERG dataset) and the gauge-based precipitation data (i.e. CEMADEN). The precipitation estimated for the five models was then linearly combined and used as input data for a stacking regression model, considering the particularities and weights of each model.

### 5.2.3. Performance evaluation of the models and methods

Four statistical metrics divided into three main groups were used to assess the quality of the ML-based models and methods. The first group aimed to analyse the efficiency of the classification models and methods in detecting days with precipitation events (i.e. classifying into rainy and non-rainy days), including F1-Score, Precision, Critical Success Index (CSI), Accuracy, and Area under the Receiver Operating Characteristic Curve (AUC-ROC):

$$\mathrm{F1-score} = \frac{2\mathrm{x}(\mathrm{Precision\ x\ Recall})}{\mathrm{Precision} + \mathrm{Recall}} \quad (10)$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (11)$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (12)$$

$$\mathrm{CSI} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \quad (13)$$

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} \quad (14)$$

where TP and TN represent the true positives and negatives, and FP and FN represent the false positives and negatives, respectively. F1-score is an evaluation metric that combines precision and recall to assess imbalanced classes. Precision is evaluate the amount of positives correctly evaluated. The Critical Success Index (CSI) is a useful metric for measuring the fraction of positive events correctly detected out of all FP and FN events. Accuracy measures the proportion of correctly classified events among the total events, providing an overall effectiveness. The AUC-ROC represents the area under the ROC (Receiver Operating Characteristics) curve, which shows the relationship between the FP and TP rates, and measures the discriminative power of a model at different classification

thresholds. The values of F1-score, Precision, CSI, and Accuracy range from -1 to 1, while for AUC-ROC. For all metrics, the desirable values are close to 1.

We used a second group of metrics to analyse the errors and agreements of the daily precipitation estimates adjusted by the regression models for the rainy days in comparison to the observed data, including the Mean Error (ME), Mean Absolute Error (MAE), Mean Relative Error (MRE), Mean Relative Absolute Error (MRAE), Correlation coefficient (CC) and Kling-Gupta Efficiency Index (KGE):

$$ME = \frac{1}{n}\sum_{i=1}^{n} O_i - E_i \quad (15)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |O_i - E_i| \quad (16)$$

$$MRE = \frac{1}{n}\sum_{i=1}^{n} \frac{O_i - E_i}{O_i} \quad (17)$$

$$MRAE = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{O_i - E_i}{O_i}\right| \quad (18)$$

$$CC = \frac{\sum_{i=1}^{n}(O_i - \bar{O})(E_i - \bar{E})}{\sqrt{\sum_{i=1}^{n}(O_i - \bar{O})^2}\sqrt{\sum_{i=1}^{n}(E_i - \bar{E})^2}} \quad (19)$$

$$KGE = 1 - \sqrt{(CC - 1)^2 + (Beta - 1)^2 + (Gama - 1)^2} \quad (22)$$

$$Gama = \frac{\sigma E}{\sigma O} \quad (20)$$

$$Beta = \frac{\bar{E}}{\bar{O}} \quad (21)$$

where O is the value observed by the rain gauges (mm), $\bar{O}$ is the mean gauged values (mm), E is the value estimated by the models (mm), $\bar{E}$ is the mean of estimated values (mm), and $\sigma$ is the standard deviation. Values close to zero indicate smaller errors in the estimations for the error-based metrics, while values close to 1 indicate better agreement between estimated and observed data.

### 5.3. Results

#### 5.3.1. Evaluation of SML models

Figure 17 shows the overall mean values of the metrics used to evaluate the performance of the six regression models used in the SML method to estimate daily precipitation. Overall, the KNN model presented the best performance, with lower errors (MAE = 2.47 mm and MRAE = 2.25) and better agreement (KGE = 0.70), followed by GB

(MAE = 2.73 mm and MRAE = 2.30) and ANN (KGE = 0.64) regarding the errors and agreement metrics, respectively. The mean values of MRAE and ME for these three models by precipitation intervals (i.e. 1-5, 5-10, 10-15, 15-25, 25-50, 50-75, 75-350 mm/day) show that KNN exhibited better performance for precipitation intervals from 15-50 to 50-350 mm/day, which represent only 29% of the events but correspond to 70% of depth (Figure S1).

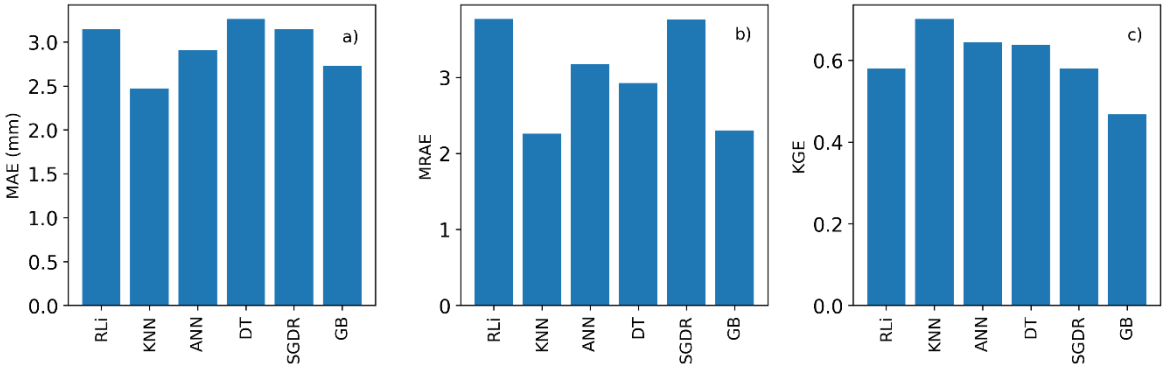


**Figure 17** - (a) MAE (mm), (b) MRAE, and (c) KGE values of the single machine learning (SML) models for estimating precipitation.

### 5.3.2. Evaluation of DML Models

Figure 18 shows the overall mean values of the efficiency metrics used to evaluate the quality of the eight classification models in separating rain and non-rain events. The models SGDC, SVM, and GB performed better for the F1-score (0.61, 0.60, and 0.60, respectively) and AUC-ROC (0.73, 0.73, and 0.73, respectively) metrics (Fig. 19ab). These high values of the metrics F1-score and AUC-ROC indicate that these models can minimise false positive and negative precipitation events. Overall, all models exhibited high Accuracy, varying between 0.82 (i.e. DT) and 0.89 (i.e. SGDC), probably influenced by the imbalance between the rain and no-rain classes since the number of days without precipitation is commonly higher along the year in most regions (Fig. 19d). The models SVM, SGDC, and GB also presented good performances for CSI (i.e. 0.48, 0.44, and 0.42, respectively) (Fig. 19e). These results indicate the ability of these three models to identify true positives and true negatives. On the other hand, an underperformance of most efficiency metrics (i.e. F1-score, AUC-ROC, and CSI) was identified especially for RF and KNN. The analysis of the efficiency metrics of these three models by precipitation intervals (i.e 5-10, 10-15, 15-25, 25-50, 50-75, 75-350 mm/day) shows that GB presented a slight superiority of ~7.5%, on average, when compared with SVM and SGDC for precipitation intervals from 1-5 to 5-15 mm/day (i.e. representing 70% of precipitation events) (Figure S2). Because classifications tend to improve when precipitation intervals increase (i.e. correctly classifying rain events

with lower volumes is more challenging), models with low rates of false positives for smaller intervals and without a high number of false negatives for larger intervals are shown to be more appropriate. Therefore, GB was considered the best classification model for identifying rain and non-rain events.

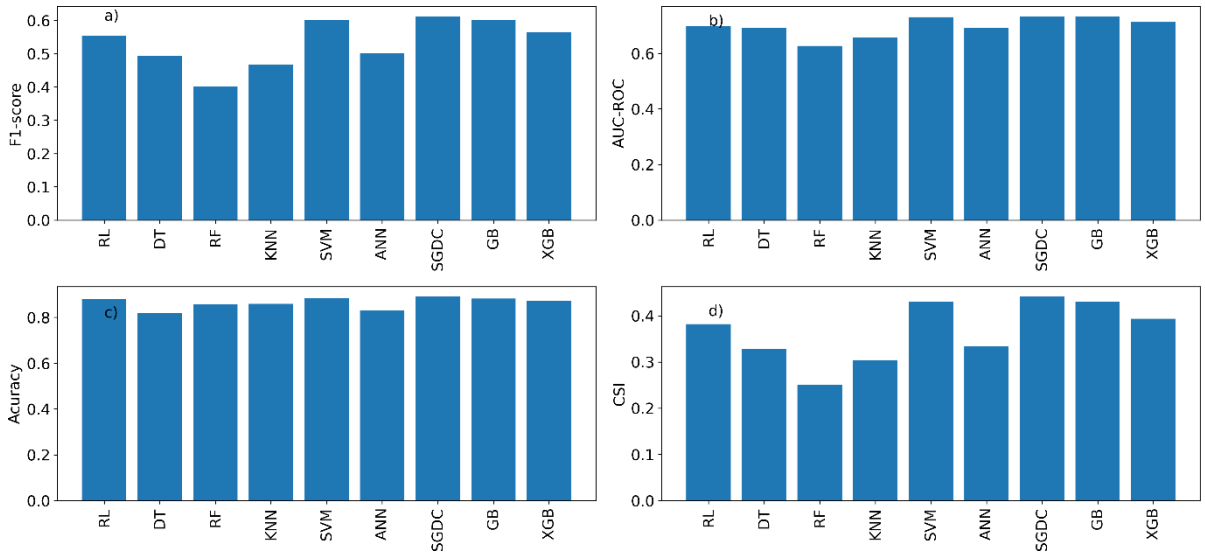


**Figure 18** – Performance of the double machine learning (DML) models RL, DR, RF, KNN, SVM, ANN, SGD, GB and XGB to estimate rain and non-rain events based on the statistical metrics (a) F1-score, (b) AUC-ROC, (c) Accuracy, and (d) CSI.

Figure 19 shows the overall mean values of the statistical indices used to evaluate the quality of the six regression models for estimating daily precipitation from the DML model. KNN presented an overall best performance, with lower errors (MAE = 11.0 and MRAE = 17.2 mm) and higher agreement (KGE = 0.36), followed by SGDR (MAE = 11.40 mm, MRAE = 1.63, and KGE = 0.28) and GB (MAE = 11.51 mm, MRAE = 1.39, and KGE = 0.22). Compared to the KNN model calibrated by the SML method, the performance of KNN using the DML method seems to be worse in classifying rain, with errors about 345% higher. However, it is worth mentioning that these index values cannot be individually compared since a regression procedure considering all data was used for the SML models, conversely to the DML models, where this regression only considered the data classified as rain. Furthermore, this stage aims to show the best models of each method instead of comparing methods. By precipitation intervals (i.e. 5-10, 10-15, 15-25, 25-50, 50-75, 75-350 mm/day), KNN presented lower MAE (from 15 to 350 mm) and MRAE (from 25 to 350) for ranges upper to 10-15 mm, which represent 52% of the events and 83% of the depths (Figure S3). Therefore, following the same criteria for choosing the SML model, KNN was selected for the DML regression.

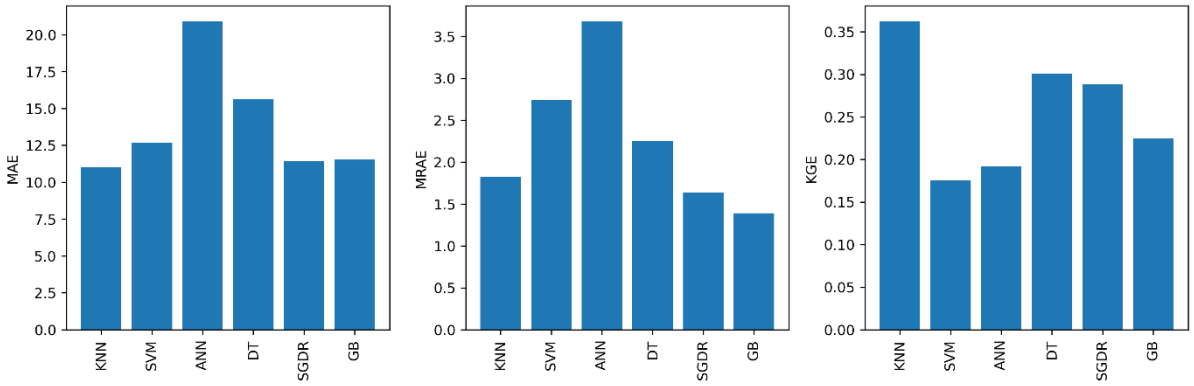


**Figure 19** - (a) MAE, (b) MRAE and (c) KGE statistical metrics of the double machine learning (DML) models to estimate precipitation.

### 5.3.3. National-scale comparison of the IMERG BraMaL products with field-based data

Figure 20 shows the scatterplots with the metrics comparing the similarity between the daily precipitation observed by the rain gauges and estimated by the products IMERG Final Run and the three versions of IMERG BraMaL (i.e. based on SML, DML, and MML models), considering the 1,846 grid cells nationwide. The lowest agreements were noticed for IMERG Final Run (CC = 0.68 and KGE = 0.66) and IMERG BraMaL based on the DML model (CC = 0.68 and KGE = 0.50), with both tending to underestimate the observed data, as shown by the greater number of points concentrated below the line of equality. Conversely, the IMERG BraMaL based on the MML model contains a cloud of points more concentrated close to the line of equality, with CC = 0.77 and KGE = 0.70. Overall, the three versions of IMERG BraMaL exhibited lower values of MAE (ranging from 5.27 to 6.49 mm) compared to IMERG Final Run (6.83 mm), with the version considering the MML model performing better. Considering the relative errors, IMERG Final Run presented a lower MRAE (i.e. 1.14) compared to IMERG BraMaL (ranging from 2.25 for MML and 2.59 for DML). These lower absolute and relative errors of IMERG BraMaL considering SML and MML models were noticed for almost all precipitation intervals, especially between 0 and 170 mm, with a slightly lower MRAE of

IMERG BraMaL SML for the ranges between 10 and 50 mm (Figure S4).

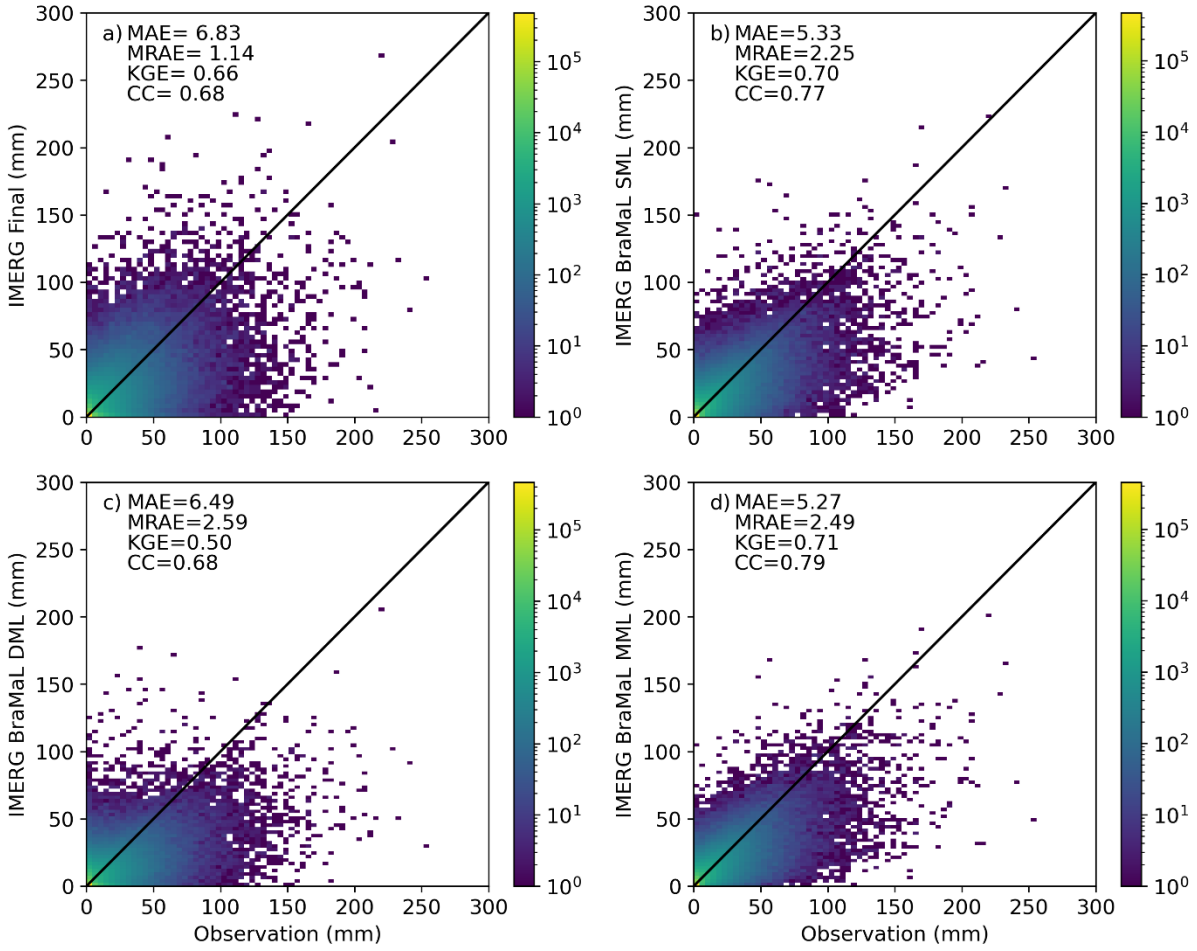


**Figure 20** – Scatter plots of the grid cell values vs (a) IMERG Final Run, (b) IMERG BraMaL SML, (c) IMERG BraMaL DML e (d) IMERG BraMaL MML, considering the national-scale analysis. The colours represent the number of events, from lower (violet) to higher (yellow).

### 5.3.4. Seasonal analysis of the IMERG BraMaL products

**Figure *21*** shows the scatterplots for each precipitation product against the observed data for each month of the year. From January to April, the products show similar behaviour, with a higher density of points (yellow represented) closer to the line of equality and an overall trend to underestimate mainly the observed data > 50 mm/day (Fig. 21 a.1- e.4). Visually, the IMERG BraMaL considering the SML and MML models presented a considerable improvement from June to September, with a lower dispersion of data and, consequently, with the cloud of points more adjusted to the line of equality compared with the IMERG Final Run product (Fig. 21 f.1 - i.4 ). The higher KGE and lower errors (MAE and MRAE) presented during this 4-month window for the IMERG BraMaL based on SML (0.84, 2.80 mm, and 1.41, respectively) and MML (0.86, 2.78 mm, and 1.58, respectively) models can confirm such enhancement (Figure 22), which mainly occurred for precipitation

intervals higher than 50 mm (Fig. S5). For instance, the IMERG Final Run product exhibited KGE = 0.57, MAE = 4.95 mm, and MRAE = 1.60 during the same period. From October to December, the products return to similar behaviour, with slight superiority of the IMERG BraMaL product considering SML and MML models in terms of MAE and KGE. It is noticeable that the DML-based product showed the worst KGE performance for all the months analysed.

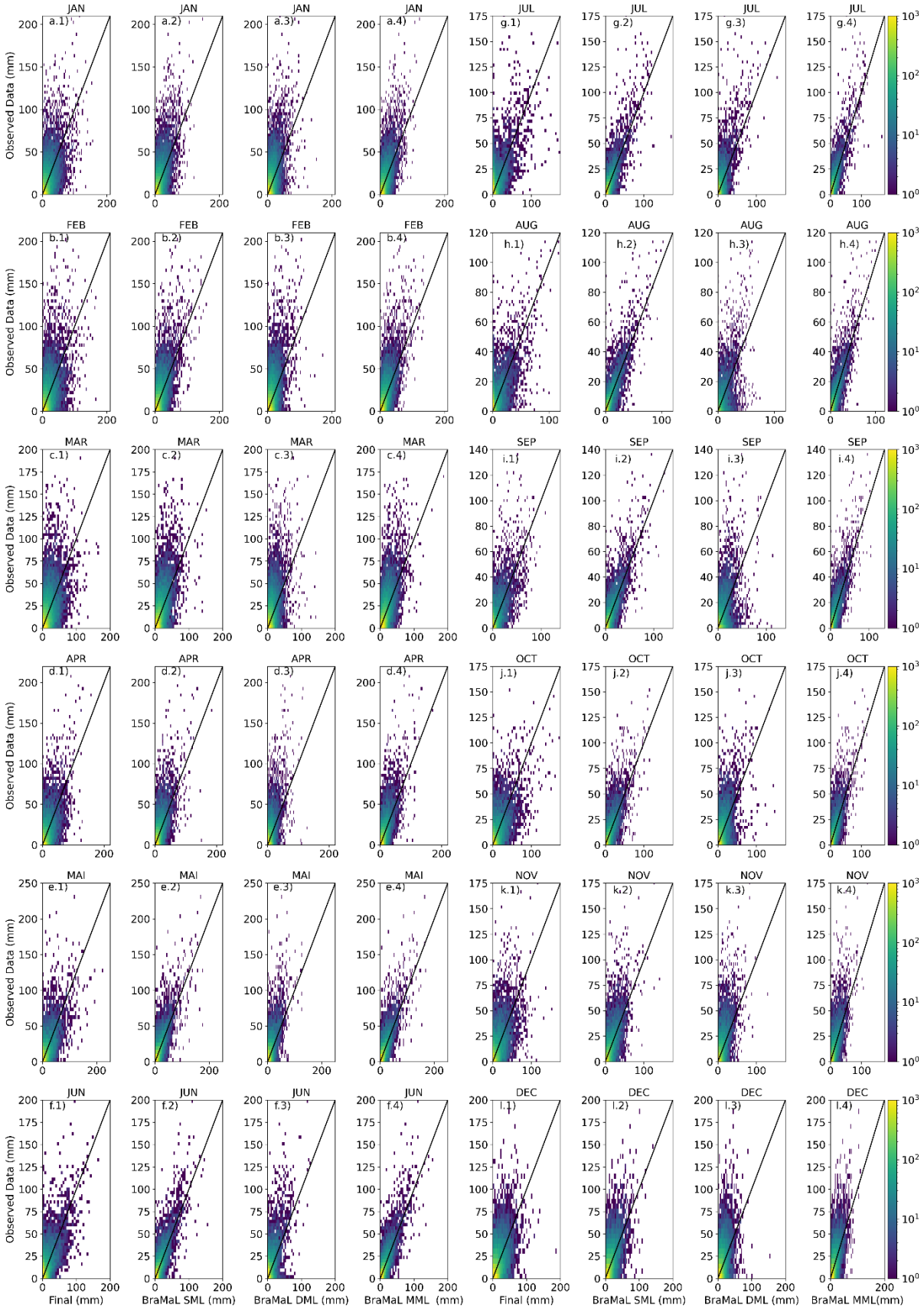


**Figure 21** – Scatter plots of the grid cell values vs variations of IMERG Final Run (first and fifth columns), IMERG BraMaL SML (second and sixth columns), IMERG BraMaL DML (third and seventh columns), and IMERG BraMaL MML (fourth and eighth columns) estimates per month.

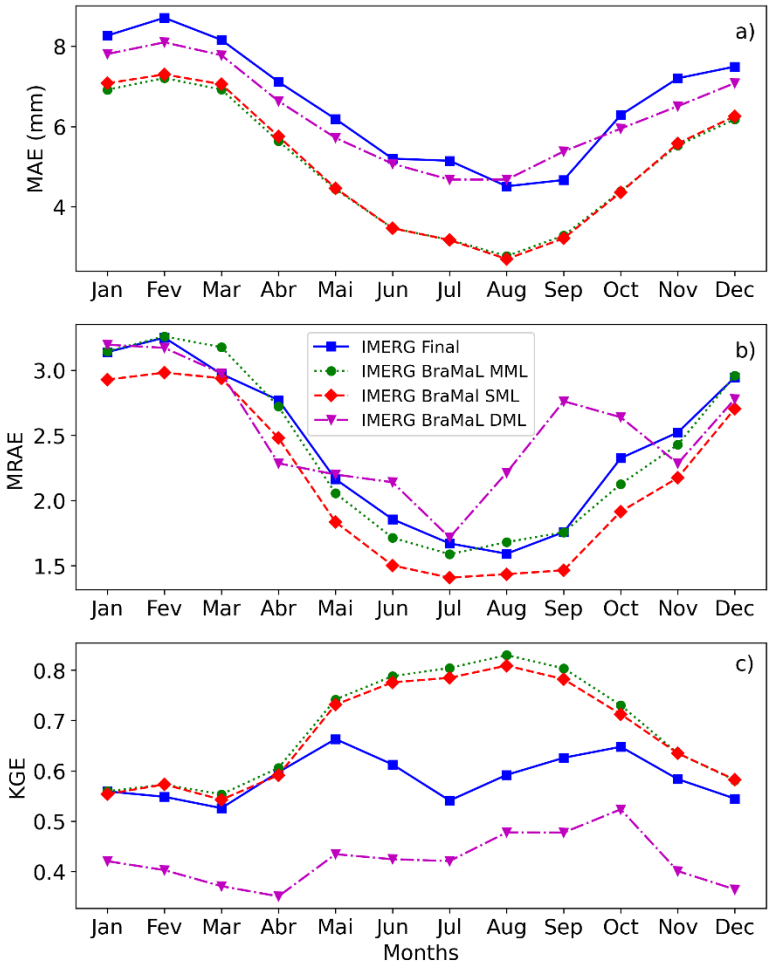


**Figure 22** – Monthly variations of (a) MAE, (b) MRAE, and (c) KGE for the IMERG Final Run and the IMERG BraMaL based on SML, DML, and MML models.

### 5.3.5. Regional-scale analysis of the IMERG BraMaL product

The performance evaluation of the IMERG BraMaL products by regions in Brazil with homogeneous monthly rainfall characteristics shows a greater dispersion of data for the estimations of the IMERG Final Run and IMERG BraMaL DML-based products from R1 to R4, with KGE ranging from 0.30 to 0.43 (R4) and from 0.52 to 0.70 (R1), respectively (Figure 23 and Figure 24). A reduction in the dispersion and the underestimation propensity of data can be observed for IMERG BraMaL based on SML and MML models between R1 and R4 when compared with the IMERG Final Run product, with the KGE values going from 0.70 to 0.74 (MML) and 0.76 (SML) in R1, from 0.53 to 0.61 (SML) and 0.60 (MML) in R2, from 0.50 to 0.57 (SML) and 0.60 (MML) in R3, and from 0.57 to 0.64 (SML) and 0.65 (MML) in R4. The error metric MAE confirms this better performance of the IMERG BraMaL product based on SML and MML models, with lower values in R1 (4.57 and 4.63 mm), R2 (5.74 and 5.67 mm), R3 (4.10 and 3.98 mm), and R4 (4.40 and 4.32 mm) compared with IMERG Final Run (6.68, 7.26, 5.06 and 5.95 mm, respectively). Conversely, such well-

defined improvement of the IMERG BraMaL product considering SML and MML models cannot be identified by the MRAE values.

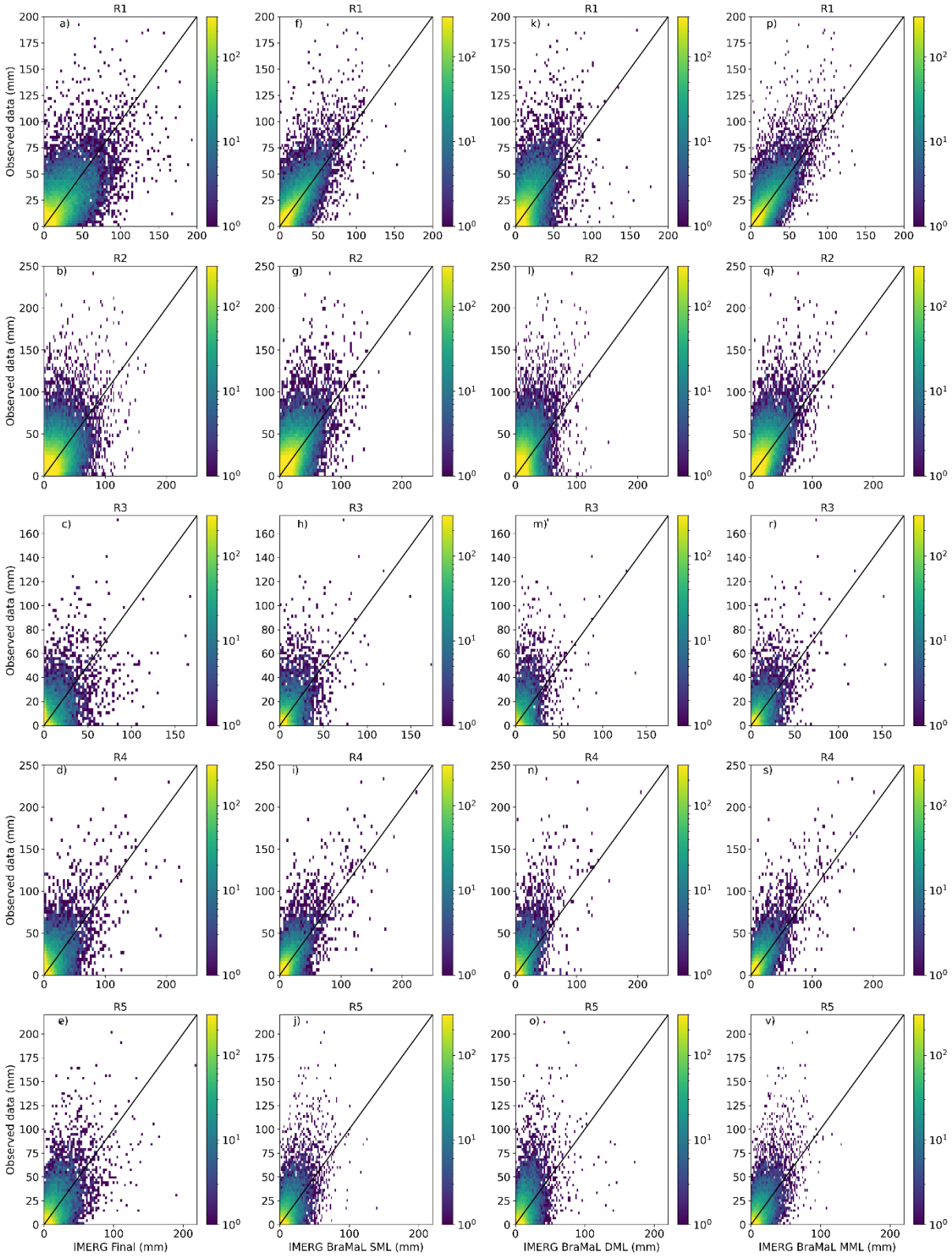


**Figure 23** – Scatter plots of the monthly observed values vs IMERG Final (first column), IMERG BraMaL SML (second column), and IMERG BraMaL DML (third column) and IMERG BraMaL MML(fourth column) by homogeneous regions (from the top to bottom: R1 to R5).

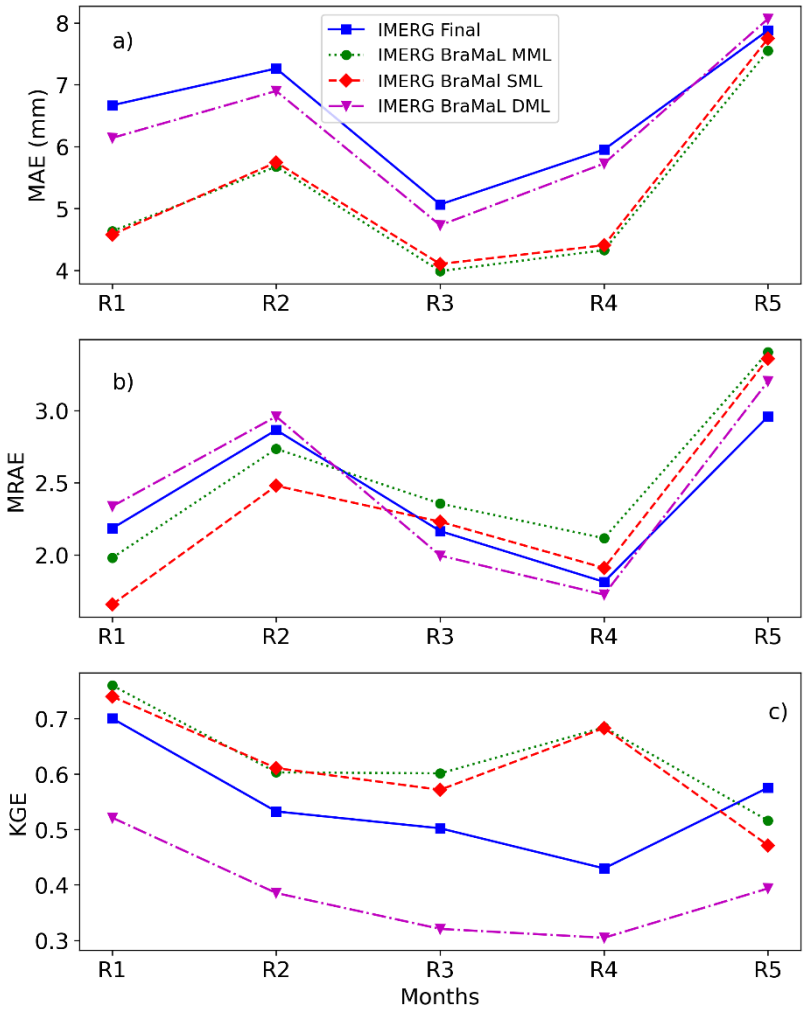


**Figure 24** – Statistical indexes (a) MAE, (b) MRAE, and (c) KGE for the IMERG Final Run and the IMERG BraMaL based on SML, DML, and MML models per homogenous monthly rainfall characteristics in Brazil.

### 5.3.6. Comparison of daily IMERG BraMMaL estimates with other global precipitation products

As shown in the national, regional, and seasonal scale analyses, IMERG BraMaL based on SML and MML models presented an overall better performance in estimating daily precipitation, with a slight superiority of the MML-based product, which considers a linear combination of models that presented good results in the SML-based analyses, including KNN, ANN, and GB. Therefore, the MML model can be considered the best choice for the daily precipitation estimates of this new product, called IMERG BraMMaL (Intercalibrated Merged Retrievals for GPM in Brazil with Multiple Machine Learning). The comparison of these daily estimates by IMERG BraMMaL with the CHIRPS, PERDIANN-CDR, and MSWEP products for all 1846 evaluated grid cells show an overall lower dispersion of the data for the proposed product (KGE = 0.70), especially when confronted with PERSIAN-

CDR (KGE = 0.05) (Figure 25). After IMERG BraMMaL, the product that presented the best agreement was MSWEP, with KGE = 0.54. This superiority of IMERG BraMMaL in estimating daily precipitation can be attested by the lower values of MAE and MRAE compared with CHIRPS (MAE = 0.67 mm and MRAE = 4.42), PERSIANN-CDR (MAE = 7.76 mm and MRAE = 2.02), and MSWEP (MAE = 6.22 mm and MRAE = 3.05). The lower MRAE value of PERSIAN-CDR was observed because this product greatly underestimates the observed data, directly impacting the relative error.

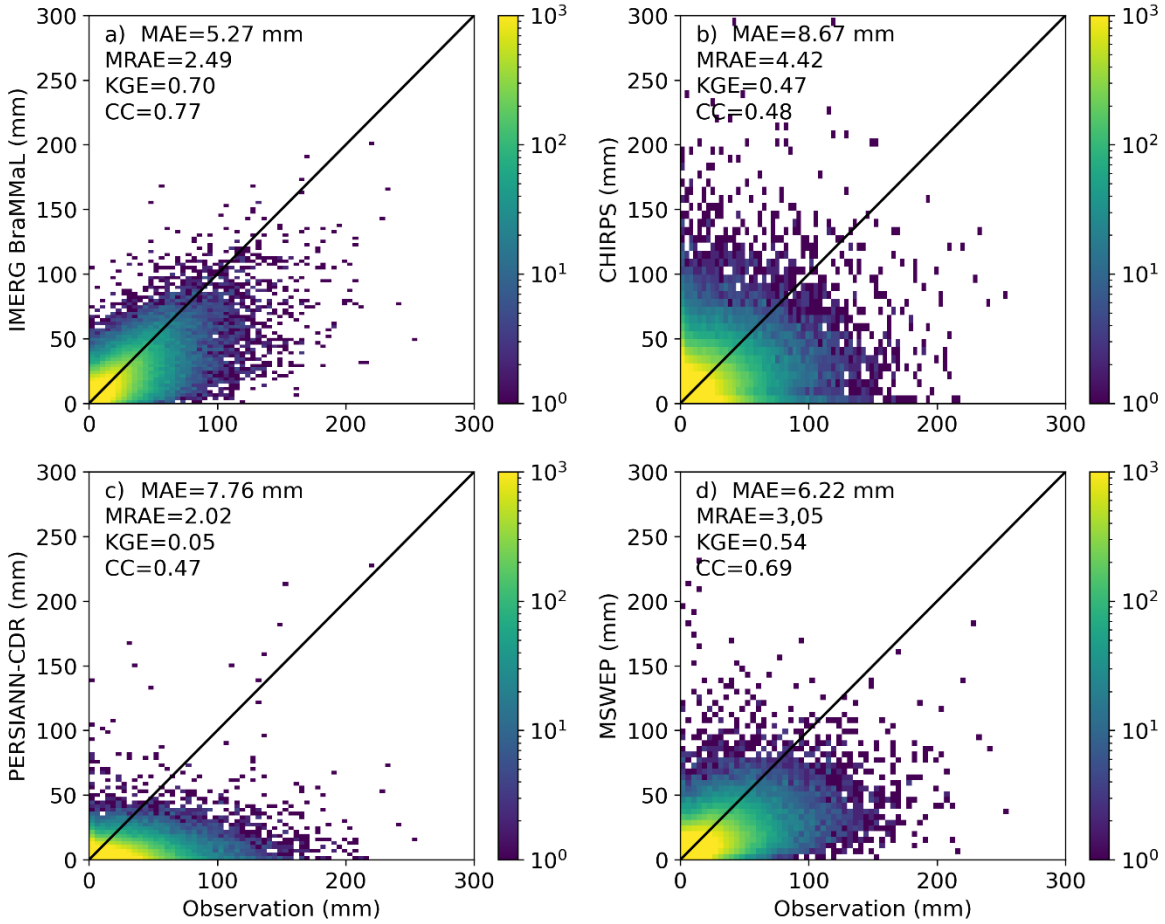


Figure 25- Scatter plots of the daily precipitation observed values vs (a) IMERG BraMMaL, (b) CHIRPS, (c) PERSIANN-CDR, and (d) MSWEP. The colours represent the number of events, from lower (violet) to higher (yellow).

### 5.3.7. Data evaluation of the monthly accumulated IMERG BraMMaL estimates

Daily precipitation estimations of the IMERG BraMMaL product were accumulated monthly and compared with estimations of the monthly IMERG BraMaL product based on the KNN regression model developed by Freitas et al. (2024) (Figure 26), which already presented better estimations compared with other global precipitation products, including IMERG Final Run and the three products previous analysed (i.e. CHIRPS, PERSIANN-

CDR, and MSWEP). The IMERG BraMMaL product visually presents a much smaller data dispersion, with a solid yellow blotch with more than 300 pixels concentrated exactly above the line of equality. Compared with the early monthly IMERG BraMaL, the improved product also presented statistically better performance, with CC and KGE rising from 0.91 to 0.97 and 0.86 to 0.95, respectively. This better performance of KGE was identified for the IMERG BraMMaL product over the months (Fig. S6). Similarly, a better performance was also observed for the error analyses, with values of MAE and MRAE for the IMERG BraMMaL product reducing from 29.10 to 13.33 mm and 0.86 to 0.28, respectively. Such lower MAE and MRAE values were noticed for the improved daily IMERG BraMMaL product over the months, especially from January to March and May to September, respectively (Figure S6).

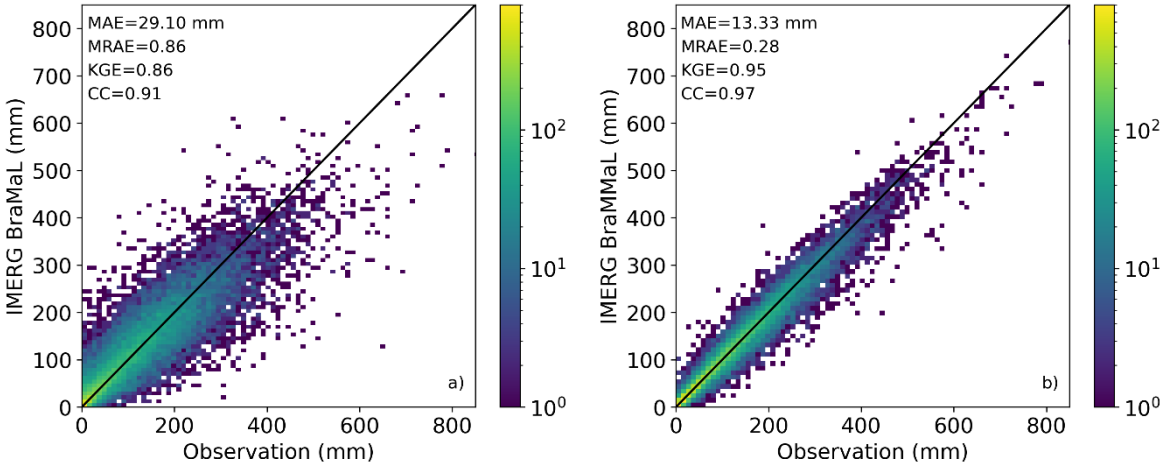


**Figure 26** – Scatterplots of observed monthly precipitation values in the grid cell vs the values of the (a) monthly IMERG BraMaL product based on the KNN regression model and proposed by Freitas et al. (2024), and (b) IMERG BraMMaL product based on the MML model and monthly accumulated.

## 5.4. Discussion

### 5.4.1. Performance of machine learning models as tools to improve precipitation estimates

The selection of machine learning models to classify rainfall events and improve precipitation estimates requires the consideration of their performance. The findings of this study were consistent with other works worldwide (e.g. Papacharalampous et al., 2023b; Hengel et al., 2018; Das et al. 2022), which found that the model GB performed better than other models (e.g. SVM and SGDC) for classifying events into rain and non-rain. Similar to what was pointed out by Hengel et al. (2018), we believe that such improved performance comes from the efficiency of GB in dealing with imbalanced classes of data, as the internal

decision trees are iteratively adjusted to focus on misclassified instances. Also, such decision trees of GB can extrapolate better beyond the training set range when compared with other models. Conversely, SVM and SGDC seek to find the hyperplane that best separates the classes in the data space and uses part of the data to adjust and update the calibration parameters, respectively. Although less accurate for detecting rain and non-rain when compared with GB, SVM and SGD also presented good performances as pointed out by Yu et al. (2023), Bansal et al. (2022), and (Sattari et al., 2023).

Our results also indicated that KNN was the best regression model regarding performance to improve the satellite-based precipitation estimates, whether individually or combined with classification models, since their predictions are based directly on the training data, being effective for nearest neighbours that present close relationships (Liu et al., 2020). This result corroborates the findings by Freitas et al. (2024), which applied KNN to enhance the IMERG Early Run estimates at a monthly scale. This performance of KNN was closely followed by GB, which also performed well in the study by Papacharalampous et al. (2023b) in the entire contiguous United States for merging satellite and earth-observed precipitation data at a daily scale. Our findings diverge from the study by Bansal et al. (2022), which points to SVM, after a literature review and research, as rendering the best results to predictive analytics in real-time applications related to a multidisciplinary sphere.

Previous studies identified substantial differences between the daily satellite-based estimates from the IMERG Final Run product and the rain gauge observations worldwide (e.g. Ramadhan et al., 2022; Sungmin et al., 2017; Yu et al., 2021), including in Brazil (e.g. Freitas et al., 2020; Gadelha et al., 2019). Also, prior studies using distinct machine learning models to improve precipitation estimates observed different performances. For instance, the studies by Bhuiyan et al. (2019) and Bhuiyan et al. (2020) respectively identified the RF and ANN models as the best options for merging satellites and observed rainfall data in the Brahmaputra river basin (China Bhutan, Nepal, India, and Bangladesh) and over complex terrains (Peruvian and Colombian Andes in South America, and the Blue Nile in East Africa). At the same time, Papacharalampous et al. (2023a) pointed out that model Xtreme GB performed better in the United States of America. Therefore, our study tested the combination of machine learning models (e.g. SML, DML, and MML) to improve the daily precipitation estimates of the IMERG Early Run product without using observed data. From this combination, DML performed worst in generating the IMERG BraMaL product with daily precipitation estimates. This performance of a DML model diverges from the results found by Zhang et al. (2021), which pinpointed a better performance of a DML

model (RF-ANN an RF-RF) compared with an SML model (RF), mainly because of the greater capacity of the DML model in describing variations of precipitations, according to the authors. Our results point out that the SML and MML models exhibited improved performances, with MML being slightly greater as using a weighted average of several models, including KNN, the better SML used in the comparisons, corroborating the findings obtained by Freitas et al. (2024).

It is worth highlighting that, conversely to the previous studies (e.g. Bhuiyan et al., 2019, 2020b; Papacharalampous et al., 2023a, 2023b), our results were based on the test dataset, without directly relying on field data. Nonetheless, they significantly improved in reducing random and systematic errors for daily precipitation estimates. This suggests the model was successfully calibrated, allowing its application in other regions where gauge data are unavailable or present low quality (Ajiboye et al., 2015). Additionally, the model presented reliable results on the independent validation dataset since we used extreme precipitation events (lowest and highest) for the training dataset, avoiding data overfitting.

### 5.4.2. Potentialities of IMERG-BraMMaL as a new estimator of daily precipitation

The selection of satellite-based precipitation products for scientific research purposes and management practices requires the consideration of their performance in distinct spatial and temporal scales as well as the direct influence of several factors (e.g. rainfall regimes, climate, altitude) that can affect its agreement with the ground-based rainfall data (Bhuiyan et al., 2017, 2019, 2020a). For instance, the studies by Rozante et al. (2018), Gadelha et al (2019), Freitas et al. (2020), and Ramos Filho et al. (2022) identified considerable under and overestimations of satellite-based precipitation products in Brazil, which were more accentuated in the North, Central-West, and Northeast regions of the country. Other studies worldwide also observed general trends of under and overestimations of the satellite-based estimates, especially at short time resolutions (e.g. Beck et al., 2019; Peng et al., 2021; Wati et al., 2021; Xiang et al., 2021). A general trend of daily underestimation has also been found to occur for some global precipitation products in our analyses, especially for MSWEP and PERSIANN-CDR (see Fig. 25), with most values located below the 1:1 line in the scatterplots, which is consistent with other works (e.g. Du et al., 2023; Fallah et al., 2020); Guo et al., 2019).

Conversely to other global products, IMERG BraMMaL presents smaller errors and better agreement with the daily observed ground-based data (Fig. 25), representing advances for precipitation estimates from space. Likewise to the IMERG BraMaL product, such errors

were much-reduced, now daily, in regions in Brazil where the IMERG Final Run largely underestimates the rain gauge data, as in the Northeast coast, which is associated with the difficulties of the passive microwave sensors in detecting warm-rain systems over land (Freitas et al., 2020; Gadelha et al., 2019; Rozante et al., 2018). In the challenging context of precisely estimating precipitation regionally, IMERG BraMMaL can be an important tool for providing a more reliable regional daily series, expanding the potential for hydrological analysis and simulations, and allowing a more accurate water balance modelling, being of large interest for water resource, agricultural and risk management. Compared to other main global products, IMERG BraMMaL also has the potential to have a higher latency (3 weeks) when compared, for instance, with the IMERG Final Run (3.5 months; Huffman et al., 2019) and PERSIANN-CDR (3 months; Ashouri et al., 2015) products, which is still lower than CHIRPS (<5 days; Funk et al., 2015) and MSWEP (3 hours; Beck et al., 2017).

# 6 CONCLUSIONS AND RECOMMENDATIONS

This thesis proposed two precipitation products, based on machine learning techniques and meteorological re-analysis data (MERRA-2), to improve the satellite-based IMERG Early Run product. IMERG BraMaL was the first product, based on regression models to estimate monthly precipitation. Secondly, the IMERG BraMMaL product was proposed based on an evaluation of single and combined (double and multiple) machine learning models to improve the first product in terms of temporal resolution (i.e. from monthly to daily) and precipitation estimations (i.e. lower errors and better agreements compared with the observed data). Overall, both products presented better and more accurate daily and monthly precipitation estimates in Brazil compared to the IMERG Early and Final Run products in almost all analyses and evaluated metrics, especially IMERG BraMMaL. Superior performance statistics were also identified when the results of the proposed products were compared with other global precipitation products (CHIRPS, PERSIANN-CDR, and MSWEP). The seven main findings of this study are summarised as follows:

1. The two proposed products (IMERG BraMaL and IMERG BraMMaL) have a higher latency (i.e. 3 weeks) when compared to the IMERG Early Run product due to its dependency on the MERRA-2 product, but a much faster availability to end users when compared to the IMERG Final Run product (i.e. 3.5 months latency).
2. Once calibrated, the IMERG BraMaL and IMERG BraMMaL generation models do not depend on any field data, relying only on satellite-based precipitation (IMERG Early Run) and re-analysis (MERRA-2) data, allowing its application in areas where rain gauge data are unavailable or present low quality.
3. The proposed models presented errors unrelated to any local features (e.g. climate, terrain, or precipitation regimes) but equally distributed throughout Brazil. This characteristic of the model enables its application in other regions, especially under tropical and subtropical climates.
4. The IMERG BraM(M)aL products much improved the precipitation estimates in regions where, historically, the satellite-based products (e.g. TMPA and IMERG) largely underestimate the observed data (e.g. along the Northeast coast of Brazil due to the topographic forcing that favours warm-rain process systems which cannot be detected very well by passive microwave sensors over land).

5. The MML model performed better than the SML and DML models in most analyses, being chosen as the basis for the improved daily IMERG BraMMaL.
6. The GB model was the best for classifying rain events and the KNN for the regression of daily rain data to generate the IMERG BraMMaL.
7. Overall, the IMERG BraMMaL product presented more accurate daily and monthly precipitation estimates in Brazil compared with IMERG BraMaL.

Based on its spatiotemporal resolution and latency, we recommend the IMERG BraMMaL product as an effective data that can be used for multiple applications such as water management, watershed rainfall-runoff modelling, water budget accounting, and drought analysis and forecast. For further research, aiming to improve our understanding of the practical problems for enhancing the satellite-based precipitation products and the various algorithmic solutions to this problem, we recommend: i) an investigation of spatial and temporal patterns of precipitation regimes since the errors of the satellite-based products can follow them; ii) an evaluation of the predictive performance of the various models, combined with the incorporation of more input data that can influence the precipitation phenomenon; iii) the improvement of the product to estimate precipitation on a sub-daily basis; and iv) the recalibration of the model with update input dataset to create updated versions of the IMERG BraMMaL product.

## DATABASE

The IMERG BraM(M)aL products are available for download as NetCDF files at http://imergbramal.net, along with the text files for the Brazilian territory. The calibrated models are also available for download at the same website and can be used for other regions.

## REFERENCES


Ajiboye, A. R., Abdullah-Arshah, R., Qin, H., & Isah-Kebbe, H. (2015). EVALUATING THE EFFECT OF DATASET SIZE ON PREDICTIVE MODEL USING SUPERVISED LEARNING TECHNIQUE. *International Journal of Computer Systems & Software Engineering*, *1*(1), 75–84. https://doi.org/10.15282/ijsecs.1.2015.6.0006

Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*.

Alvares, C. A., Stape, J. L., Sentelhas, P. C., De Moraes Gonçalves, J. L., & Sparovek, G. (2013). Köppen's climate classification map for Brazil. *Meteorologische Zeitschrift*, *22*(6), 711–728. https://doi.org/10.1127/0941-2948/2013/0507

Amari, S.-I. (1993). Backpropagation and stochastic gradient descent method. In *Neurocomputing* (Vol. 5).

Ashouri, H., Hsu, K. L., Sorooshian, S., Braithwaite, D. K., Knapp, K. R., Cecil, L. D., Nelson, B. R., & Prat, O. P. (2015). PERSIANN-CDR: Daily precipitation climate data record from multisatellite observations for hydrological and climate studies. *Bulletin of the American Meteorological Society*, *96*(1), 69–83. https://doi.org/10.1175/BAMS-D-13-00068.1

Assiri, M. E., & Qureshi, S. (2022). A Multi-Source Data Fusion Method to Improve the Accuracy of Precipitation Products: A Machine Learning Algorithm. *Remote Sensing*, *14*(24). https://doi.org/10.3390/rs14246389

Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning. *Decision Analytics Journal*, *3*, 100071. https://doi.org/10.1016/j.dajour.2022.100071

Beck, H. E., Pan, M., Roy, T., Weedon, G. P., Pappenberger, F., Van Dijk, A. I. J. M., Huffman, G. J., Adler, R. F., & Wood, E. F. (2019). Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS. *Hydrology and Earth System Sciences*, *23*(1), 207–224. https://doi.org/10.5194/hess-23-207-2019

Beck, H. E., Van Dijk, A. I. J. M., Levizzani, V., Schellekens, J., Miralles, D. G., Martens, B., & De Roo, A. (2017). MSWEP: 3-hourly 0.25° global gridded precipitation (1979-2015) by merging gauge, satellite, and reanalysis data. *Hydrology and Earth System Sciences*, *21*(1), 589–615. https://doi.org/10.5194/hess-21-589-2017

Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., Van Dijk, A. I. J. M., Weedon, G. P., Brocca, L., Pappenberger, F., Huffman, G. J., & Wood, E. F. (2017). Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. *Hydrology and Earth System Sciences*, *21*(12), 6201–6217. https://doi.org/10.5194/hess-21-6201-2017

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., Van Dijk, A. I. J. M., McVicar, T. R., & Adler, R. F. (2019). MSWep v2 Global 3-hourly 0.1° precipitation: Methodology and quantitative assessment. *Bulletin of the American Meteorological Society*, *100*(3), 473–500. https://doi.org/10.1175/BAMS-D-17-0138.1

Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., Schamm, K., Schneider, U., & Ziese, M. (2013). A description of the global land-surface precipitation data products of the Global Precipitation Climatology Centre with sample applications including centennial (trend) analysis from 1901-present. *Earth System Science Data*, *5*(1), 71–99. https://doi.org/10.5194/essd-5-71-2013

Behrangi, A., & Wen, Y. (2017). On the spatial and temporal sampling errors of remotely sensed precipitation products. *Remote Sensing*, *9*(11). https://doi.org/10.3390/rs9111127

Belabid, N., Zhao, F., Brocca, L., Huang, Y., & Tan, Y. (2019). Near-real-time flood forecasting based on satellite precipitation products. *Remote Sensing*, *11*(3). https://doi.org/10.3390/rs11030252

Belyadi, H., & Haghighat, A. (2021). *Machine Learning Guide for Oil and Gas Using Python* (Vol. 1). Gulf Professional Publishing.

Bhuiyan, M. A. E., Anagnostou, E. N., & Kirstetter, P. E. (2017). A Nonparametric Statistical Technique for Modeling Overland TMI (2A12) Rainfall Retrieval Error. *IEEE Geoscience and*

*Remote Sensing Letters*, *14*(11), 1898–1902. https://doi.org/10.1109/LGRS.2017.2728658
Bhuiyan, M. A. E., Nikolopoulos, E. I., & Anagnostou, E. N. (2019). Machine learning–based blending of satellite and reanalysis precipitation datasets: A multiregional tropical complex terrain evaluation. *Journal of Hydrometeorology*, *20*(11), 2147–2161. https://doi.org/10.1175/JHM-D-19-0073.1
Bhuiyan, M. A. E., Nikolopoulos, E. I., Anagnostou, E. N., Quintana-Seguí, P., & Barella-Ortiz, A. (2018). A nonparametric statistical technique for combining global precipitation datasets: Development and hydrological evaluation over the Iberian Peninsula. *Hydrology and Earth System Sciences*, *22*(2), 1371–1389. https://doi.org/10.5194/hess-22-1371-2018
Bhuiyan, M. A. E., Yang, F., Biswas, N. K., Rahat, S. H., & Neelam, T. J. (2020a). Machine Learning-Based Error Modeling to Improve GPM IMERG Precipitation Product over the Brahmaputra River Basin. *Forecasting*, *2*(3), 248–266. https://doi.org/10.3390/forecast2030014
Bhuiyan, M. A. E., Yang, F., Biswas, N. K., Rahat, S. H., & Neelam, T. J. (2020b). Machine Learning-Based Error Modeling to Improve GPM IMERG Precipitation Product over the Brahmaputra River Basin. *Forecasting*, *2*(3), 248–266. https://doi.org/10.3390/forecast2030014
BISHOP, C. (2006). *Pattern recognition and machine learning* (Vol. 1). Springer google schola.
Bishop, C. M. (2006). *Pattern Recognition and Machine Learning* (1st ed.). Springer New York, NY.
Blenkinsop, S., Fowler, H. J., Barbero, R., Chan, S. C., Guerreiro, S. B., Kendon, E., Lenderink, G., Lewis, E., Li, X.-F., Westra, S., Alexander, L., Allan, R. P., Berg, P., Dunn, R. J. H., Ekström, M., Evans, J. P., Holland, G., Jones, R., Kjellström, E., … Tye, M. R. (2018). The INTENSE project: using observations and models to understand the past, present and future of sub-daily rainfall extremes. *Advances in Science and Research*, *15*, 117–126. https://doi.org/10.5194/asr-15-117-2018
Bonnema, M., Sikder, S., Miao, Y., ChenXiaodong, Hossain, Rahman, F., Mahbubur, I. A. P. S. M., & Lee, H. (2016). Water Resources Research. *Water Resources Research*, *10*(1002), 4095–4115. https://doi.org/10.1111/j.1752-1688.1969.tb04897.x
Breiman, L. (2001). *Random Forests* (Vol. 45).
Breugem, A. J., Wesseling, J. G., Oostindie, K., & Ritsema, C. J. (2020). Earth-Science Reviews Meteorological aspects of heavy precipitation in relation to floods – An overview. *Earth-Science Reviews*, *204*(January), 103171. https://doi.org/10.1016/j.earscirev.2020.103171
Brocca, L., Filippucci, P., Hahn, S., Ciabatta, L., Massari, C., Camici, S., Schüller, L., Bojkov, B., & Wagner, W. (2019). SM2RAIN-ASCAT (2007-2018): Global daily satellite rainfall data from ASCAT soil moisture observations. *Earth System Science Data*, *11*(4), 1583–1601. https://doi.org/10.5194/essd-11-1583-2019
Brocca, L., Moramarco, T., Melone, F., & Wagner, W. (2013). A new method for rainfall estimation through soil moisture observations. *Geophysical Research Letters*, *40*(5), 853–858. https://doi.org/10.1002/grl.50173
Brocca, L., Pellarin, T., Crow, W. T., Ciabatta, L., Massari, C., Ryu, D., Su, C.-H., Rüdiger, C., & Kerr6, Y. (2016). Rainfall estimation by inverting SMOS soil moisture estimates: A comparison of different methods over Australia. *Journal of Geophysical Research: Atmospheres RESEARCH*, *121*(5), 3010–3028. https://doi.org/10.1002/2016JD025382.Received
Brunetti, M. T., Melillo, M., Gariano, S. L., Ciabatta, L., Brocca, L., Amarnath, G., & Peruccacci, S. (2021). Satellite rainfall products outperform ground observations for landslide prediction in India. *Hydrology and Earth System Sciences*, *25*(6), 3267–3279. https://doi.org/10.5194/hess-25-3267-2021
Brunetti, M. T., Melillo, M., Peruccacci, S., Ciabatta, L., & Brocca, L. (2018). How far are we from the use of satellite rainfall products in landslide forecasting? *Remote Sensing of Environment*, *210*, 65–75. https://doi.org/10.1016/j.rse.2018.03.016
Chen, C., He, M., Chen, Q., Zhang, J., Li, Z., Wang, Z., & Duan, Z. (2022). Triple collocation-based error estimation and data fusion of global gridded precipitation products over the Yangtze River basin. *Journal of Hydrology*, *605*. https://doi.org/10.1016/j.jhydrol.2021.127307
Chen, H., Chen, A., Xu, L., Xie, H., Qiao, H., Lin, Q., & Cai, K. (2020). A deep learning CNN architecture applied in smart near-infrared analysis of water pollution for agricultural irrigation resources. *Agricultural Water Management*, *240*(April), 106303. https://doi.org/10.1016/j.agwat.2020.106303

Chen, H., Yong, B., Kirstetter, P. E., Wang, L., & Hong, Y. (2021). Global component analysis of errors in three satellite-only global precipitation estimates. *Hydrology and Earth System Sciences*, *25*(6), 3087–3104. https://doi.org/10.5194/hess-25-3087-2021

Cortes, C., Vapnik, V., & Saitta, L. (1995). Support-Vector Networks Editor. In *Machine Leaming* (Vol. 20). Kluwer Academic Publishers.

Das, S., Wang, Y., Gong, J., Ding, L., Munchak, S. J., Wang, C., Wu, D. L., Liao, L., Olson, W. S., & Barahona, D. O. (2022). A Comprehensive Machine Learning Study to Classify Precipitation Type over Land from Global Precipitation Measurement Microwave Imager (GPM-GMI) Measurements. *Remote Sensing*, *14*(15). https://doi.org/10.3390/rs14153631

Dee, D. P., Balmaseda, M., Balsamo, G., Engelen, R., Simmons, A. J., & Thépaut, J. N. (2014). Toward a consistent reanalysis of the climate system. *Bulletin of the American Meteorological Society*, *95*(8), 1235–1248. https://doi.org/10.1175/BAMS-D-13-00043.1

Derin, Y., & Yilmaz, K. K. (2014). Evaluation of multiple satellite-based precipitation products over complex topography. *Journal of Hydrometeorology*, *15*(4), 1498–1516. https://doi.org/10.1175/JHM-D-13-0191.1

Di Nunno, F., Granata, F., Pham, Q. B., & de Marinis, G. (2022). Precipitation Forecasting in Northern Bangladesh Using a Hybrid Machine Learning Model. *Sustainability (Switzerland)*, *14*(5). https://doi.org/10.3390/su14052663

Du, J., Yu, X., Zhou, L., Ren, Y., & Ao, T. (2023). Precipitation Characteristics across the Three River Headwaters Region of the Tibetan Plateau: A Comparison between Multiple Datasets. *Remote Sensing*, *15*(9). https://doi.org/10.3390/rs15092352

Du, Y., & Xie, Z. Q. (2020). *Spatial Scales of Heavy Meiyu Precipitation Events in Eastern China and Associated Atmospheric Processes Geophysical Research Letters*. 1–9. https://doi.org/10.1029/2020GL087086

Fallah, A., Rakhshandehroo, G. R., Berg, P., Sungmin, O., & Orth, R. (2020). Evaluation of precipitation datasets against local observations in southwestern Iran. *International Journal of Climatology*, *40*(9), 4102–4116. https://doi.org/10.1002/joc.6445

Freitas, E. da S., Coelho, V. H. R., Xuan, Y., Melo, D. de C. D., Gadelha, A. N., Santos, E. A., Galvão, C. de O., Ramos Filho, G. M., Barbosa, L. R., Huffman, G. J., Petersen, W. A., & Almeida, C. das N. (2020). The performance of the IMERG satellite-based product in identifying sub-daily rainfall events and their properties. *Journal of Hydrology*, *589*. https://doi.org/10.1016/j.jhydrol.2020.125128

Friedman, J. H. (2001). 999 REITZ LECTURE GREEDY FUNCTION APPROXIMATION: A GRADIENT BOOSTING MACHINE 1. In *The Annals of Statistics* (Vol. 29, Issue 5).

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., & Michaelsen, J. (2015). The climate hazards infrared precipitation with stations - A new environmental record for monitoring extremes. *Scientific Data*, *2*. https://doi.org/10.1038/sdata.2015.66

Gadelha, A. N., Coelho, V. H. R., Xavier, A. C., Barbosa, L. R., Melo, D. C. D., Xuan, Y., Huffman, G. J., Petersen, W. A., & Almeida, C. das N. (2019). Grid box-level evaluation of IMERG over Brazil at various space and time scales. *Atmospheric Research*, *218*(August 2018), 231–244. https://doi.org/10.1016/j.atmosres.2018.12.001

Garcez, L. N., & Alvarez, G. A. (1988). *Hidrologia* (2. ed. rev. e atual.).

Gardner, M. W., & Dorling, S. R. (1998). ARTIFICIAL NEURAL NETWORKS (THE MULTILAYER PERCEPTRON)-A REVIEW OF APPLICATIONS IN THE ATMOSPHERIC SCIENCES. In *Atmospheric Environment* (Vol. 32, Issue 14).

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A. M., Gu, W., … Zhao, B. (2017). The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *Journal of Climate*, *30*(14), 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1

Goodfellow, I., Yoshua Bengio, & Aaron Courville. (2016). *Deep learning*. MIT press.

Guo, D., Wang, H., Zhang, X., & Liu, G. (2019). Evaluation and analysis of grid precipitation fusion products in Jinsha river basin based on China meteorological assimilation datasets for the SWAT model. *Water (Switzerland)*, *11*(2). https://doi.org/10.3390/w11020253

Gupta, V., Jain, M. K., Singh, P. K., & Singh, V. (2020). An assessment of global satellite-based precipitation datasets in capturing precipitation extremes: A comparison with observed precipitation dataset in India. *International Journal of Climatology*, *40*(8), 3667–3688. https://doi.org/10.1002/joc.6419

Habib, E., Krajewski, W. F., & Kruger, A. (2001). Sampling Errors of Tipping-Bucket Rain Gauge Measurements. *Journal of Hydrologic Engineering*, *6*(2), 159–166. https://doi.org/10.1061/(ASCE)1084-0699(2001)6:2(159)

Harris, I., Jones, P. D., Osborn, T. J., & Lister, D. H. (2014). Updated high-resolution grids of monthly climatic observations - the CRU TS3.10 Dataset. *International Journal of Climatology*, *34*(3), 623–642. https://doi.org/10.1002/joc.3711

Hastie, T. , Tibshirani, R. , & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). New York: springer.

Hegerl, G. C., Black, E., Allan, R. P., Ingram, W. J., Polson, D., Trenberth, K. E., Chadwick, R. S., Arkin, Beena Balan Sarojini, P. A., Becker, A., Durack, P. J., Easterling, D., Fowler, H. J., Kendon, J., Huffman, G. J., Liu, C., Marsh, R., Osborn, T. J., Stott, P. A., Vidale, P.-L., … Zhang, X. (n.d.). *CHALLENGES IN QUANTIFYING CHANGES IN THE GLOBAL WATER CYCLE*. https://doi.org/10.1175/BAMS-D-13-00212.2

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, *18*(7), 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527

Houze, R. A. (2012). Orographic effects on precipitating clouds. *Reviews of Geophysics*, *50*(1), 1–47. https://doi.org/10.1029/2011RG000365

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., & Tan, J. (2019). *Integrated Multi-satellitE Retrievals for GPM (IMERG) Technical Documentation*.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Jiang, S., Wei, L., Ren, L., Xu, C. Y., Zhong, F., Wang, M., Zhang, L., Yuan, F., & Liu, Y. (2021). Utility of integrated IMERG precipitation and GLEAM potential evapotranspiration products for drought monitoring over mainland China. *Atmospheric Research*, *247*. https://doi.org/10.1016/j.atmosres.2020.105141

Jones, P. D., Briffa, K. R., Osborn, T. J., Moberg, A., & Bergström, H. (2002). Relationships between circulation strength and the variability of growing-season and cold-season climate in northern and central Europe. *Holocene*, *12*(6), 643–656. https://doi.org/10.1191/0959683602hl577rp

Kidd, C., Becker, A., Huffman, G. J., Muller, C. L., Joe, P., Skofronick-Jackson, G., & Kirschbaum, D. B. (2017). So, how much of the Earth's surface is covered by rain gauges? *Bulletin of the American Meteorological Society*, *98*(1), 69–78. https://doi.org/10.1175/BAMS-D-14-00283.1

Kidd, C., & Huffman, G. (2011). Global precipitation measurement. In *Meteorological Applications* (Vol. 18, Issue 3, pp. 334–353). John Wiley and Sons Ltd. https://doi.org/10.1002/met.284

Kidd, C., & Levizzani, V. (2011). Status of satellite precipitation retrievals. *Hydrology and Earth System Sciences*, *15*(4), 1109–1116. https://doi.org/10.5194/hess-15-1109-2011

Kidd, C., Levizzani, V., Turk, J., & Ferraro, R. (2009). Satellite precipitation measurements for water resource monitoring. *Journal of the American Water Resources Association*, *45*(3), 567–579. https://doi.org/10.1111/j.1752-1688.2009.00326.x

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Kumar, A., Ramsankaran, R., Brocca, L., & Munoz-Arriola, F. (2019). A Machine Learning Approach for Improving Near-Real-Time Satellite-Based Rainfall Estimates by Integrating Soil Moisture. *Remote Sensing*, *11*(19), 1–20. https://doi.org/10.3390/rs11192221

Lewis, E., Fowler, H., Alexander, L., Dunn, R., Mcclean, F., Barbero, R., Guerreiro, S., Li, X. F., & Blenkinsop, S. (2019). GSDR: A global sub-daily rainfall dataset. *Journal of Climate*, *32*(15), 4715–4729. https://doi.org/10.1175/JCLI-D-18-0143.1

Li, N., Tang, G., Zhao, P., Hong, Y., Gou, Y., & Yang, K. (2017). Statistical assessment and hydrological utility of the latest multi-satellite precipitation analysis IMERG in Ganjiang River basin. *Atmospheric Research*, *183*, 212–223. https://doi.org/10.1016/j.atmosres.2016.07.020

Liu, W., Wang, P., Meng, Y., Zhao, C., & Zhang, Z. (2020). Cloud spot instance price prediction using kNN regression. *Human-Centric Computing and Information Sciences*, *10*(1).

https://doi.org/10.1186/s13673-020-00239-5
Llauca, H., Lavado-casimiro, W., León, K., Jimenez, J., Traverso, K., & Rau, P. (2021). Assessing near real-time satellite precipitation products for flood simulations at sub-daily scales in a sparsely gauged watershed in Peruvian andes. *Remote Sensing*, *13*(4), 1–18. https://doi.org/10.3390/rs13040826
Ma, Z., Xu, J., Zhu, S., Yang, J., Tang, G., Yang, Y., Shi, Z., & Hong, Y. (2020). AIMERG: A new Asian precipitation dataset (0.1°/half-hourly, 2000-2015) by calibrating the GPM-era IMERG at a daily scale using APHRODITE. *Earth System Science Data*, *12*(3), 1525–1544. https://doi.org/10.5194/essd-12-1525-2020
Markonis, Y., Papalexiou, S. M., Martinkova, M., & Hanel, M. (2019). Assessment of Water Cycle Intensification Over Land using a Multisource Global Gridded Precipitation DataSet. *JGR Atmospheres*, *124*, 11175–11187. https://doi.org/10.1029/2019JD030855
Marra, F., & Morin, E. (2015). Use of radar QPE for the derivation of Intensity-Duration-Frequency curves in a range of climatic regimes. *Journal of Hydrology*, *531*, 427–440. https://doi.org/10.1016/j.jhydrol.2015.08.064
McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*.
Mei, Y., Anagnostou, E. N., Nikolopoulos, E. I., & Borga, M. (2014). Error analysis of satellite precipitation products in mountainous basins. *Journal of Hydrometeorology*, *15*(5), 1778–1793. https://doi.org/10.1175/JHM-D-13-0194.1
Meng, Q., Chen, W., Wang, Y., Ma, Z. M., & Liu, T. Y. (2019). Convergence analysis of distributed stochastic gradient descent with shuffling. *Neurocomputing*, *337*, 46–57. https://doi.org/10.1016/j.neucom.2019.01.037
Michaelides, S., Levizzani, V., Anagnostou, E., Bauer, P., Kasparis, T., & Lane, J. E. (2009). Precipitation: Measurement, remote sensing, climatology and modeling. *Atmospheric Research*, *94*(4), 512–533. https://doi.org/10.1016/j.atmosres.2009.08.017
Mitchell, T. M. (1977). *Machine Learning* (McGraw-Hill Science).
Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2020). FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, *36*(1), 86–92. https://doi.org/10.1016/j.ijforecast.2019.02.011
Mu, Y., Liu, W., Liu, X., & Fan, W. (2017). Stochastic Gradient Made Stable: A Manifold Propagation Approach for Large-Scale Optimization. *IEEE Transactions on Knowledge and Data Engineering*, *29*(2), 458–471. https://doi.org/10.1109/TKDE.2016.2604302
Munier, S., Aires, F., Schlaffer, S., Prigent, C., Papa, F., Maisongrande, P., & Pan, M. (2014). Combining data sets of satellite-retrieved products for basin-scale water balance study: 2. Evaluation on the Mississippi basin and closure correction model. *Journal of Geophysical Research*, *119*(21), 12,100-12,116. https://doi.org/10.1002/2014JD021953
New, M., Todd, M., Hulme, M., & Jones, P. (2001). Precipitation measurements and trends in the twentieth century. *International Journal of Climatology*, *21*(15), 1889–1922. https://doi.org/10.1002/joc.680
Ning, S., Song, F., Udmale, P., Jin, J., Thapa, B. R., & Ishidaira, H. (2017). Error Analysis and Evaluation of the Latest GSMap and IMERG Precipitation Products over Eastern China. *Advances in Meteorology*, *2017*(March 2014). https://doi.org/10.1155/2017/1803492
Ochoa-Rodriguez, S., Wang, L. P., Willems, P., & Onof, C. (2019). A Review of Radar-Rain Gauge Data Merging Methods and Their Potential for Urban Hydrological Applications. In *Water Resources Research* (Vol. 55, Issue 8, pp. 6356–6391). Blackwell Publishing Ltd. https://doi.org/10.1029/2018WR023332
Oliveira, R., Maggioni, V., Vila, D., & Morales, C. (2016). Characteristics and diurnal cycle of GPM rainfall estimates over the Central Amazon region. *Remote Sensing*, *8*(7). https://doi.org/10.3390/rs8070544
Papacharalampous, G., Tyralis, H., Doulamis, A., & Doulamis, N. (2023a). Comparison of Machine Learning Algorithms for Merging Gridded Satellite and Earth-Observed Precipitation Data. *Water (Switzerland)*, *15*(4). https://doi.org/10.3390/w15040634
Papacharalampous, G., Tyralis, H., Doulamis, A., & Doulamis, N. (2023b). Comparison of Tree-Based Ensemble Algorithms for Merging Satellite and Earth-Observed Precipitation Data at the Daily

Time Scale. *Hydrology*, *10*(2). https://doi.org/10.3390/hydrology10020050
Papacharalampous, G., Tyralis, H., Doulamis, N., & Doulamis, A. (2023c). Ensemble Learning for Blending Gridded Satellite and Gauge-Measured Precipitation Data. *Remote Sensing*, *15*(20). https://doi.org/10.3390/rs15204912
Pellarin, T., Louvet, S., Gruhier, C., Quantin, G., & Legout, C. (2013). A simple and effective method for correcting soil moisture and precipitation estimates using AMSR-E measurements. *Remote Sensing of Environment*, *136*(September), 28–36. https://doi.org/10.1016/j.rse.2013.04.011
Pellet, V., Aires, F., Munier, S., Fernández Prieto, D., Jordá, G., Arnoud Dorigo, W., Polcher, J., & Brocca, L. (2019). Integrating multiple satellite observations into a coherent dataset to monitor the full water cycle - Application to the Mediterranean region. *Hydrology and Earth System Sciences*, *23*(1), 465–491. https://doi.org/10.5194/hess-23-465-2019
Peng, J., Liu, T., Huang, Y., Ling, Y., Li, Z., Bao, A., Chen, X., Kurban, A., & De Maeyer, P. (2021). Satellite-based precipitation datasets evaluation using gauge observation and hydrological modeling in a typical arid land watershed of central asia. *Remote Sensing*, *13*(2), 1–26. https://doi.org/10.3390/rs13020221
Pinto, N. L. de S., Holtz, A. C. T., Martins, J. A., & Gomide, F. L. S. (1976). *Hidrologia básica* (Editora Blucher).
Prakash, S., Mitra, A. K., Pai, D. S., & AghaKouchak, A. (2016). From TRMM to GPM: How well can heavy rainfall be detected from space? *Advances in Water Resources*, *88*, 1–7. https://doi.org/10.1016/j.advwatres.2015.11.008
Quinlan, J. R. (1986). Induction of Decision Trees. In *Machine Learning* (Vol. 1).
Rafieeinasab, A., Norouzi, A., Seo, D. J., & Nelson, B. (2015). Improving high-resolution quantitative precipitation estimation via fusion of multiple radar-based precipitation products. *Journal of Hydrology*, *531*, 320–336. https://doi.org/10.1016/j.jhydrol.2015.04.066
Raghav. (2022). *Data Science with Raghav*. Https://Www.Datasciencewithraghav.Com/2022/10/28/How-to-Avoid-under-Fitting-and-over-Fitting-While-Training-a-Neural-Network/.
Raj, R., Saharia, M., Chakma, S., & Rafieinasab, A. (2022). Mapping rainfall erosivity over India using multiple precipitation datasets. *Catena*, *214*. https://doi.org/10.1016/j.catena.2022.106256
Rajagopalan, B., & Lall, U. (1999). A k-nearest-neighbor simulator for daily precipitation and other weather variables. *Water Resources Research*, *35*(10), 3089–3101. https://doi.org/10.1029/1999WR900028
Ramadhan, R., Yusnaini, H., Marzuki, M., Muharsyah, R., Suryanto, W., Sholihun, S., Vonnisa, M., Harmadi, H., Ningsih, A. P., Battaglia, A., Hashiguchi, H., & Tokay, A. (2022). Evaluation of GPM IMERG Performance Using Gauge Data over Indonesian Maritime Continent at Different Time Scales. *Remote Sensing*, *14*(5). https://doi.org/10.3390/rs14051172
Ramos Filho, G. M., Coelho, V. H. R., Freitas, E. da S., Xuan, Y., & Almeida, C. das N. (2021). An improved rainfall-threshold approach for robust prediction and warning of flood and flash flood hazards. *Natural Hazards*, *105*(3), 2409–2429. https://doi.org/10.1007/s11069-020-04405-x
Rozante, J. R., Moreira, D. S., de Goncalves, L. G. G., & Vila, D. A. (2010). Combining TRMM and surface observations of precipitation: Technique and validation over South America. *Weather and Forecasting*, *25*(3), 885–894. https://doi.org/10.1175/2010WAF2222325.1
Rozante, J. R., Vila, D. A., Chiquetto, J. B., Fernandes, A. de A., & Alvim, D. S. (2018). Evaluation of TRMM/GPM blended daily products over Brazil. *Remote Sensing*, *10*(6). https://doi.org/10.3390/rs10060882
Sadeghi, M., Nguyen, P., Hsu, K., & Sorooshian, S. (2020). Improving near real-time precipitation estimation using a U-Net convolutional neural network and geographical information. *Environmental Modelling and Software*, *134*. https://doi.org/10.1016/j.envsoft.2020.104856
Samuel, A. L. (1959). *Machine learning* (Vol. 1). The Technology Review.
Satgé, F., Ruelland, D., Bonnet, M. P., Molina, J., & Pillco, R. (2019). Consistency of satellite-based precipitation products in space and over time compared with gauge observations and snow-hydrological modelling in the Lake Titicaca region. *Hydrology and Earth System Sciences*, *23*(1), 595–619. https://doi.org/10.5194/hess-23-595-2019
Sattari, M. T., Avram, A., Apaydin, H., & Matei, O. (2023). Evaluation of Feature Selection Methods in Estimation of Precipitation Based on Deep Learning Artificial Neural Networks. *Water Resources*

*Management*, *37*(15), 5871–5891. https://doi.org/10.1007/s11269-023-03563-4
Schneider, U., Ziese, M., Meyer-Christoffer, A., Finger, P., Rustemeier, E., & Becker, A. (2016). The new portfolio of global precipitation data products of the Global Precipitation Climatology Centre suitable to assess and quantify the global water cycle and resources. *Proceedings of the International Association of Hydrological Sciences*, *374*, 29–34. https://doi.org/10.5194/piahs-374-29-2016
Sehad, M., Lazri, M., & Ameur, S. (2017). Novel SVM-based technique to improve rainfall estimation over the Mediterranean region (north of Algeria) using the multispectral MSG SEVIRI imagery. *Advances in Space Research*, *59*(5), 1381–1394. https://doi.org/10.1016/j.asr.2016.11.042
Sengoz, C., Ramanna, S., Kehler, S., Goomer, R., & Pries, P. (2023). Machine Learning Approaches to Improve North American Precipitation Forecasts. *IEEE Access*, *11*, 97664–97681. https://doi.org/10.1109/ACCESS.2023.3309054
Silva Lelis, L. C., Duarte Bosquilia, R. W., & Duarte, S. N. (2018). Assessment of precipitation data generated by GPM and TRMM satellites. *Revista Brasileira de Meteorologia*, *33*(1), 153–163. https://doi.org/10.1590/0102-7786331004
Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., & Hsu, K. L. (2018). A Review of Global Precipitation Data Sets: Data Sources, Estimation, and Intercomparisons. *Reviews of Geophysics*, *56*(1), 79–107. https://doi.org/10.1002/2017RG000574
Sungmin, O., Foelsche, U., Kirchengast, G., Fuchsberger, J., Tan, J., & Petersen, W. A. (2017). Evaluation of GPM IMERG Early, Late, and Final rainfall estimates using WegenerNet gauge data in southeastern Austria. *Hydrology and Earth System Sciences*, *21*(12), 6559–6572. https://doi.org/10.5194/hess-21-6559-2017
SUTTON, R. S., & BARTO, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.
Tan, J., Petersen, W. A., Kirchengast, G., Goodrich, D. C., & Wolff, D. B. (2018). Evaluation of global precipitation measurement rainfall estimates against three dense gauge networks. *Journal of Hydrometeorology*, *19*(3), 517–532. https://doi.org/10.1175/JHM-D-17-0174.1
Tang, G., Ma, Y., Long, D., Zhong, L., & Hong, Y. (2016). Evaluation of GPM Day-1 IMERG and TMPA Version-7 legacy products over Mainland China at multiple spatiotemporal scales. *Journal of Hydrology*, *533*, 152–167. https://doi.org/10.1016/j.jhydrol.2015.12.008
Thornes, J., Bloss, W., Bouzarovski, S., Cai, X., Chapman, L., Clark, J., Dessai, S., Du, S., van der Horst, D., Kendall, M., Kidd, C., & Randalls, S. (2010). Communicating the value of atmospheric services. *Meteorological Applications*, *17*(2), 243–250. https://doi.org/10.1002/met.200
Tyralis, H., Papacharalampous, G., Doulamis, N., & Doulamis, A. (2023). Merging Satellite and Gauge-Measured Precipitation Using LightGBM With an Emphasis on Extreme Quantiles. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *16*, 6969–6979. https://doi.org/10.1109/JSTARS.2023.3297013
Varma, A. K. (2018). Measurement of Precipitation from Satellite Radiometers (Visible, Infrared, and Microwave): Physical Basis, Methods, and Limitations. In *Remote Sensing of Aerosols, Clouds, and Precipitation* (pp. 223–248). Elsevier. https://doi.org/10.1016/B978-0-12-810437-8.00011-6
Wanders, N., Wada, Y., & Van Lanen, H. A. J. (2015). Global hydrological droughts in the 21st century under a changing hydrological regime. *Earth System Dynamics*, *6*(1), 1–15. https://doi.org/10.5194/esd-6-1-2015
Wang, C., Jia, Z., Yin, Z., Liu, F., Lu, G., & Zheng, J. (2021). Improving the Accuracy of Subseasonal Forecasting of China Precipitation With a Machine Learning Approach. *Frontiers in Earth Science*, *9*. https://doi.org/10.3389/feart.2021.659310
Wang, R., Chu, H., Liu, Q., Chen, B., Zhang, X., Fan, X., Wu, J., Xu, K., Jiang, F., & Chen, L. (2023). Application of Machine Learning Techniques to Improve Multi-Radar Mosaic Precipitation Estimates in Shanghai. *Atmosphere*, *14*(9). https://doi.org/10.3390/atmos14091364
Wang, Y., Kong, Y., Chen, H., & Zhao, L. (2020). Improving daily precipitation estimates for the Qinghai-Tibetan plateau based on environmental similarity. *International Journal of Climatology*, *40*(12), 5368–5388. https://doi.org/10.1002/joc.6523
Wang, Z., Zhong, R., Lai, C., & Chen, J. (2017). Evaluation of the GPM IMERG satellite-based precipitation products and the hydrological utility. *Atmospheric Research*, *196*, 151–163. https://doi.org/10.1016/j.atmosres.2017.06.020
Wati, T., Hadi, T. W., Sopaheluwakan, A., & Hutasoit, L. M. (2021). Evaluation gridded precipitation

datasets in Indonesia. *IOP Conference Series: Earth and Environmental Science*, *893*(1). https://doi.org/10.1088/1755-1315/893/1/012056

Wehbe, Y., Temimi, M., & Adler, R. F. (2020). Enhancing precipitation estimates through the fusion of weather radar, satellite retrievals, and surface parameters. *Remote Sensing*, *12*(8). https://doi.org/10.3390/RS12081342

Wolfensberger, D., Gabella, M., Boscacci, M., Germann, U., & Berne, A. (2021). RainForest: a random forest algorithm for quantitative precipitation estimation over Switzerland. *Atmospheric Measurement Techniques*, *14*(4), 3169–3193. https://doi.org/10.5194/amt-14-3169-2021

Xiang, Y., Chen, J., Li, L., Peng, T., & Yin, Z. (2021). Evaluation of eight global precipitation datasets in hydrological modeling. *Remote Sensing*, *13*(14). https://doi.org/10.3390/rs13142831

Xu, L., Chen, N., Moradkhani, H., Zhang, X., & Hu, C. (2020). Improving Global Monthly and Daily Precipitation Estimation by Fusing Gauge Observations, Remote Sensing, and Reanalysis Data Sets. *Water Resources Research*, *56*(3). https://doi.org/10.1029/2019WR026444

Xu, R., Tian, F., Yang, L., Hu, H., Lu, H., & Hou, A. (2017). Ground validation of GPM IMERG and trmm 3B42V7 rainfall products over Southern Tibetan plateau based on a high-density rain gauge network. *Journal of Geophysical Research*, *122*(2), 910–924. https://doi.org/10.1002/2016JD025418

Yang, X., & Chen, Z. (2023). Assessing the effects of time series on precipitation forecasting performance from complexity perspective. *Theoretical and Applied Climatology*, *154*(3–4), 973–986. https://doi.org/10.1007/s00704-023-04616-9

Yang, Y., & Luo, Y. (2014). Using the back propagation neural network approach to bias correct TMPA data in the arid region of northwest China. *Journal of Hydrometeorology*, *15*(1), 459–473. https://doi.org/10.1175/JHM-D-13-041.1

Yu, C., Shao, H., Hu, D., Liu, G., & Dai, X. (2023). Merging precipitation scheme design for improving the accuracy of regional precipitation products by machine learning and geographical deviation correction. *Journal of Hydrology*, *620*. https://doi.org/10.1016/j.jhydrol.2023.129560

Yuan, F., Zhang, L., Soe, K. M. W., Ren, L., Zhao, C., Zhu, Y., Jiang, S., & Liu, Y. (2019). Applications of TRMM- and GPM-era multiple- satellite precipitation products for flood simulations at sub-daily scales in a sparsely gauged watershed in Myanmar. *Remote Sensing*, *11*(2). https://doi.org/10.3390/rs11020140

Zhang, L., Li, X., Zheng, D., Zhang, K., Ma, Q., Zhao, Y., & Ge, Y. (2021). Merging multiple satellite-based precipitation products and gauge observations using a novel double machine learning approach. *Journal of Hydrology*, *594*. https://doi.org/10.1016/j.jhydrol.2021.125969

Zhou, C., Gao, W., Hu, J., Du, L., & Du, L. (2021). Capability of imerg v6 early, late, and final precipitation products for monitoring extreme precipitation events. *Remote Sensing*, *13*(4), 1–23. https://doi.org/10.3390/rs13040689

## APPENDICES

Figure S1 shows the overall mean values of MRAE and ME for the three models that presented the best performances (i.e. KNN, GB, and ANN), divided by seven precipitation intervals (i.e. 1-5, 5-10, 10-15, 15-25, 25-50, 50-75, 75-350 mm/day). GB, RNA, and ANN presented almost the same performance for precipitation intervals from 1-5 to 10-15 mm/day, with GB exhibiting slight superiority. Such intervals represent 71% of the precipitation events but correspond only to 30% of precipitation depth. For precipitation intervals from 15-50 to 50-350 mm/day, KNN exhibited better performance. Conversely, such intervals represent only 29% of the precipitation events but correspond to 70% of precipitation depth. Therefore, KNN was chosen as the best SML model.

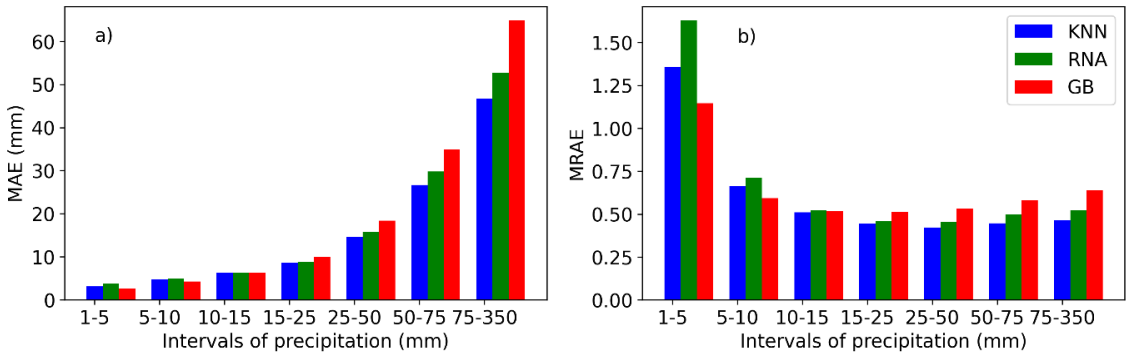


**Figure S1** - (a) MAE and (b) MRAE values of KNN, RNA, and GB single machine learning (SML) models by precipitation intervals of 1-5, 5-10, 10-15, 15-25, 25-50, 50-75, and 75-350 mm.

The overall mean values of the efficiency metrics for the three DML models that presented the best performances (i.e. GB, SVM, and SGCD), divided by six precipitation intervals (i.e. 5-10, 10-15, 15-25, 25-50, 50-75, 75-350 mm/day), are shown in Figure S2. Overall, the three models performed similarly for all metrics and precipitations intervals, with GB presenting a slight superiority of ~7.5%, on average, when compared to SVM for precipitation intervals from 1-5 to 5-15 mm/day (i.e. representing 70% of precipitation events). On the other hand, SGDC exhibited improved performances of ~5.2%, on average, when compared to SVM for precipitation intervals higher than 25-50 mm/day (i.e. 30% of precipitation events). Unlike the previous analysis performed for regression models, which considers the magnitude and errors of the rain, a precise number of classifications is more important for DML models. As shown in Figure S2, the classifications tend to improve as precipitation intervals increase, meaning correctly classifying rain events with lower volumes is a bigger challenge. Thus, models with low rates of false positives for smaller precipitation intervals and without a high number of false negatives for larger precipitation

intervals are shown to be more appropriate. Therefore, we chose GB as the best classification model to identify rain and non-rain events.

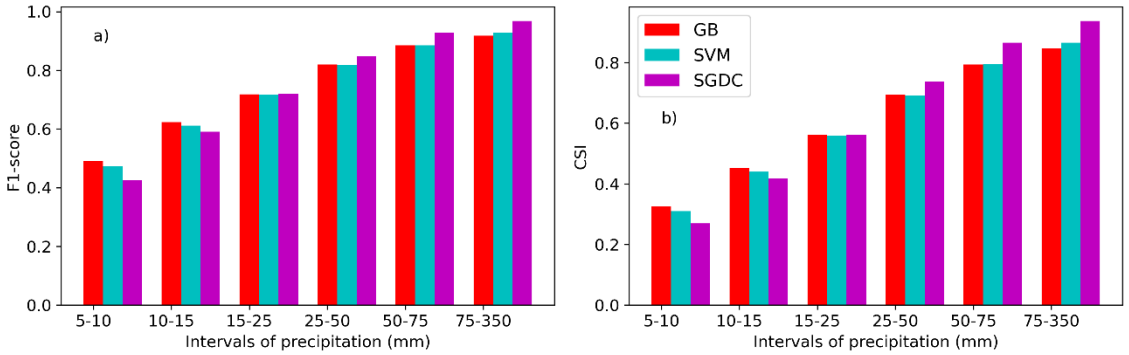


**Figure S2** - (a) F1-score and (b) CSI values of CB, SVM, and SGD double machine learning (DML) models classified by precipitation intervals of 5-10, 10-15, 15-25, 25-50, 50-75, and 75-350 mm.

The mean statistical error indices of the three DML models that performed better (i.e. KNN, SGDR, and GB), divided by seven precipitation intervals (i.e. 5-10, 10-15, 15-25, 25-50, 50-75, 75-350 mm/day), show that GB presented lower MAE (4.23 and 3.80 mm) and MRAE (1.73 and 0.52) for shorter intervals (i.e. 1-5 and 5-10 mm) when compared to KNN and SGDR (Figure S3). These two shorter intervals represent 48% of the rainfall events but only 17% of the depth. Overall similar performance of the three methods can be observed for the time intervals between 10-15 and 25-50 mm, with a slight superiority of KNN. Conversely, these intervals correspond to 44% of rainfall events and 54% of rainfall depth. For rainfall intervals higher than 25-50 mm, which represent 8% of events and 29% of depth, KNN exhibited lower errors (e.g. MAE = 59.24 mm and MRAE = 0.58 for 75-350 mm), especially compared to GB (e.g. MAE = 68.31 mm and MRAE = 0.66 for 75-350 mm). Following the same criteria for choosing the SML model, KNN was selected for the DML regression since presented better results for rainfall intervals representing 52% of the events and 83% of the depths.

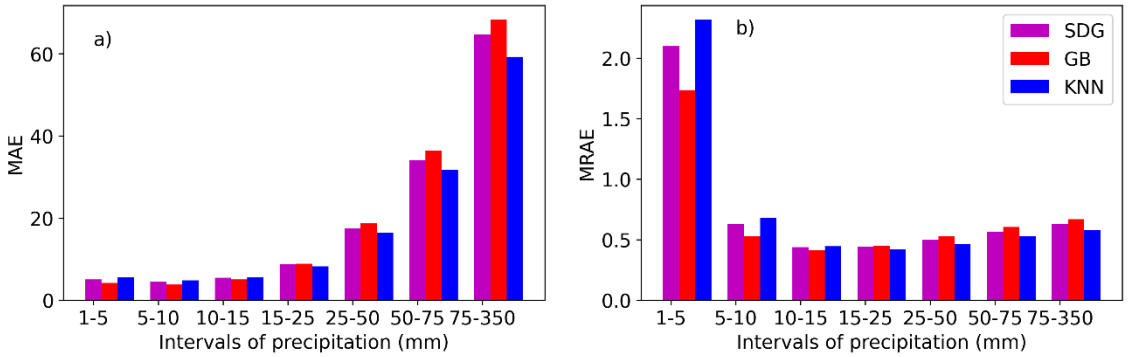


**Figure S3** - (a) MAE, (b) RMSE, (c) MRE, and (d) MRAE values of KNN, RNA, and GB double machine learning (DML) models classified by precipitation intervals of 1-5, 5-10, 10-15, 15-25, 25-50, 50-75, and 75-350 mm.

The error analysis by precipitation intervals shows that IMERG BraMaL considering SML and MML models presented lower absolute and relative than the IMERG Final Run product for almost all precipitation ranges, especially between 0 and 170 mm (Figure S4). However, the MRAE of IMERG BraMaL based on the SML model was slightly lower for the precipitation intervals between 10 and 50 mm. On the other hand, the IMERG BraMaL product based on the DML model exhibited the worst performance for almost all precipitation intervals, mainly higher than 50 mm, with MRAE and MAE values superior to those observed for the IMERG Final Run product.

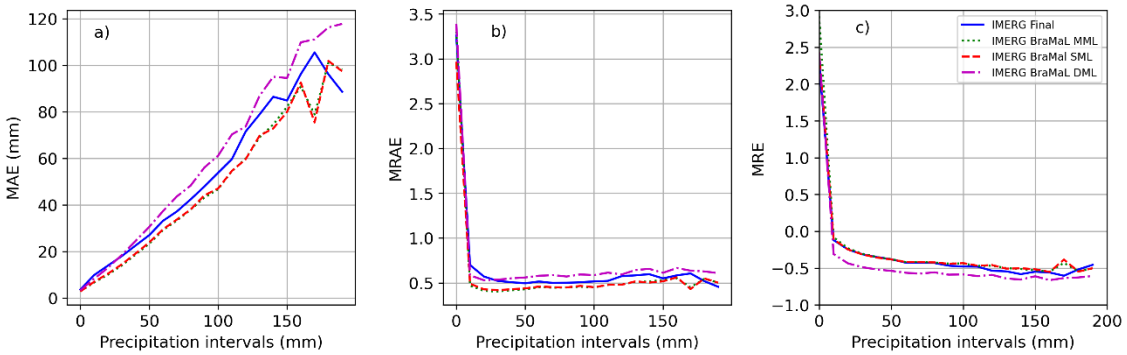


**Figure S4** - (a) MAE, (b) MRAE and (c) MRE values of the IMERG FInal, IMERG BraMaL SML, IMERG BraMaL DML and IMERG BraMaL MML products for different monthly precipitation intervals.

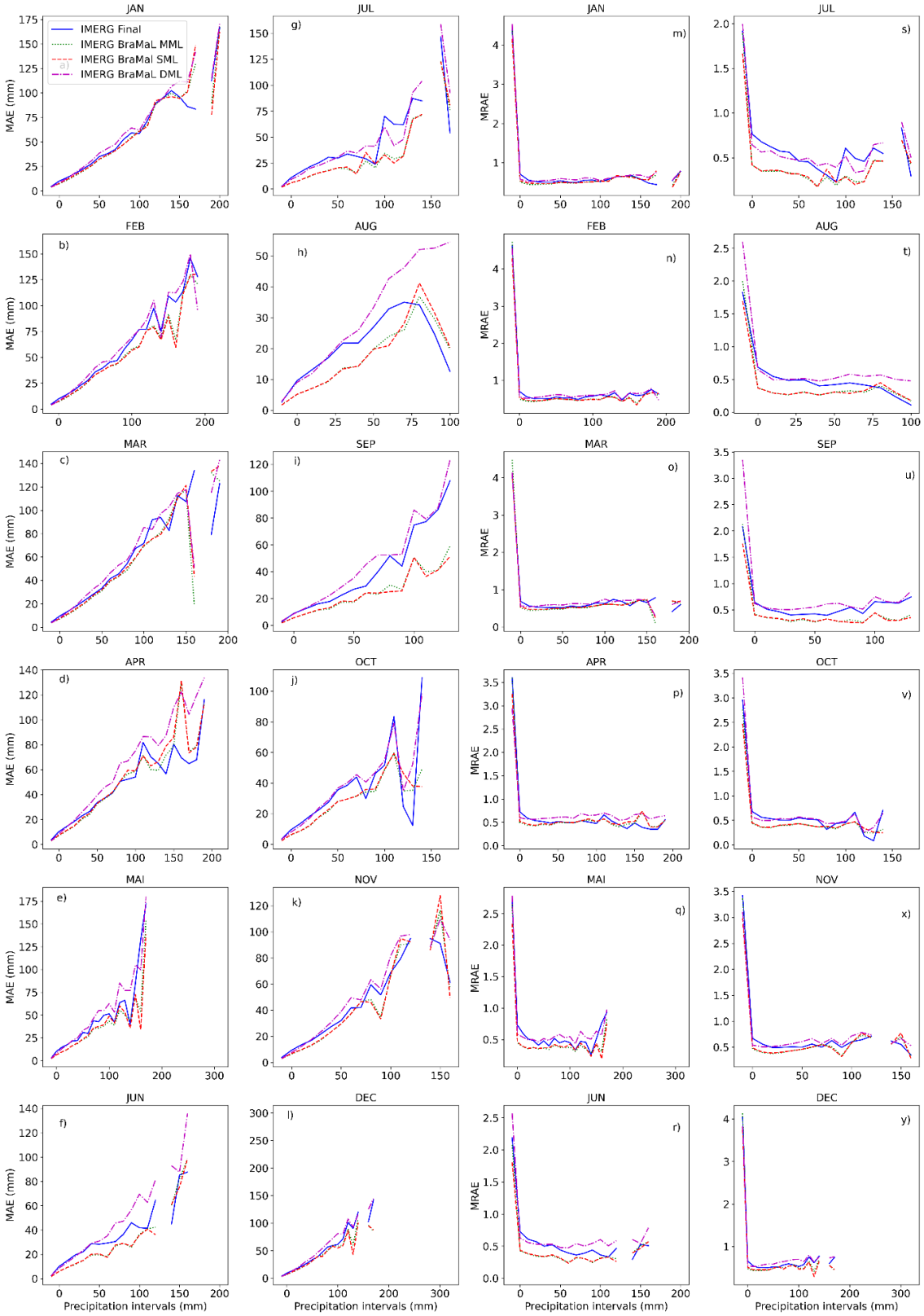


**Figure S5** – Monthly variations of the IMERG Final Run, IMERG BraMaL SML, DML and MML products for different intervals of precipitation: MAE from (a) January to December (l) and MRAE from (m) January to (y) December.

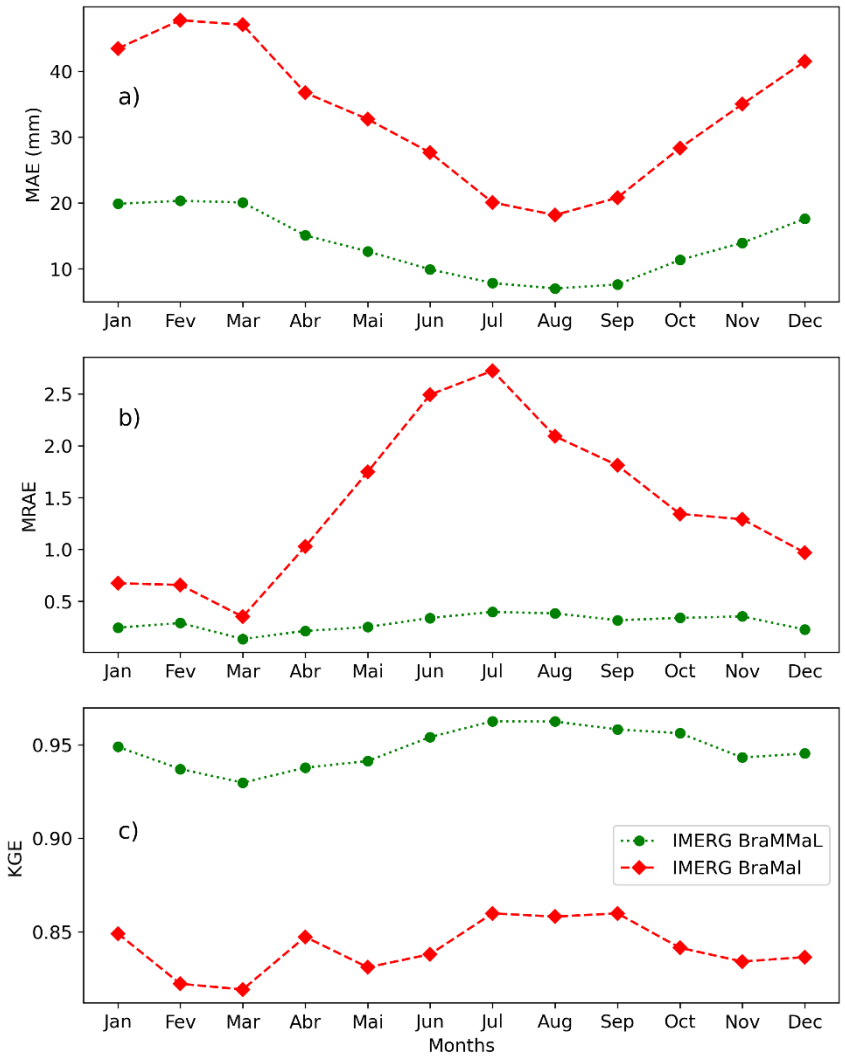


**Figure S6** – Monthly variations of (a) MAE, (b) MRAE, and (c) KGE for the IMERG BraMMaL and IMERG BraMaL.