Universidade Federal da Paraíba

Centro de Tecnologia

PROGRAMA DE PÓS-GRADUAÇÃO EM

ENGENHARIA CIVIL E AMBIENTAL

– MESTRADO –

BRAZILIAN DAILY RAINFALL GRIDDED DATA FROM A QUALITY CONTROLLED DATASET

Por

José Lindemberg Vidal-Barbosa

Dissertação de Mestrado apresentada à Universidade Federal da Paraíba para obtenção do grau de Mestre

João Pessoa – Paraíba **Março de 2024**

Universidade Federal da Paraíba

Centro de Tecnologia

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA CIVIL E AMBIENTAL – MESTRADO –

BRAZILIAN DAILY RAINFALL GRIDDED DATA FROM A QUALITY CONTROLLED DATASET

Dissertação submetida ao Programa de Pós-Graduação em Engenharia Civil e Ambiental da Universidade Federal da Paraíba, como parte dos requisitos para a obtenção do título de Mestre.

José Lindemberg Vidal-Barbosa

Orientador: Cristiano das Neves Almeida
Coorientador: Guillaume Francis Bertrand

João Pessoa – Paraíba **Março de 2024**

Catalogação na publicação
Seção de Catalogação e Classificação

V648b Vidal-Barbosa, José Lindemberg.
Brazilian daily rainfall gridded data from a quality controlled dataset / José Lindemberg Vidal-Barbosa. - João Pessoa, 2024.
48 f. : il.

Orientação: Cristiano das Neves Almeida.
Coorientação: Guillaume Francis Bertrand.
Dissertação (Mestrado) - UFPB/CT.

1. Precipitação. 2. Controle de qualidade. 3. Dados diários em grade. 4. Interpolação. I. Almeida, Cristiano das Neves. II. Bertrand, Guillaume Francis. III. Título.

UFPB/BC CDU 551.577(043)

Elaborado por Larissa Silva Oliveira de Mesquita - CRB-15/746

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

BRAZILIAN DAILY RAINFALL GRIDDED DATA FROM A QUALITY CONTROLLED DATASET

JOSÉ LINDEMBERG VIDAL-BARBOSA
Dissertação aprovada em 27 de março de 2024.
Período Letivo: 2023.2

gov.br Documento assinado digitalmente
CRISTIANO DAS NEVES ALMEIDA
Data: 02/04/2024 07:44:58-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Cristiano das Neves Almeida – UFPB
Orientador

gov.br Documento assinado digitalmente
GUILLAUME FRANCIS BERTRAND
Data: 02/04/2024 08:17:36-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Guillaume Francis Bertrand – UFPB
Examinador Interno

gov.br Documento assinado digitalmente
GERALD NORBERT SOUZA DA SILVA
Data: 02/04/2024 09:09:37-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Gerald Norbert Souza da Silva – UFPB
Examinador Interno

gov.br Documento assinado digitalmente
SAULO AIRES DE SOUZA
Data: 17/04/2024 17:10:06-0300
Verifique em https://validar.iti.gov.br

Dr. Saulo Aires de Souza - ANA
Examinador Externo

João Pessoa/PB
2024

A minha mãe, Narciza Vidal, que sempre acreditou no meu potencial e fez tudo o que estava ao seu alcance para que os meus sonhos se realizassem, e a todos aqueles que dedicam grande parte de sua vida à ciência em meio a tantas adversidades.

AGRADECIMENTOS

A minha mãe, Maria Narciza de Lima Vidal, e a meu irmão, José Lucas Vidal Barbosa, deixo um agradecimento especial por me apoiarem incondicionalmente durante o percurso desta trajetória. Vocês são parte desta história!

A meus amigos, minhas amigas e familiares mais próximos, por compreenderem as minhas ausências ao longo desses últimos anos. Mesmo com a distância geográfica e, por vezes, emocional, muitos de vocês sempre estiveram muito presentes. Amo vocês!

Ao meu orientador, Professor Dr. Cristiano das Neves Almeida, que acreditou em mim e me deu a oportunidade de mostrar todo meu comprometimento com a educação, com a ciência e com este trabalho. Agradeço imensamente pela sua confiança e dedicação. Serei eternamente grato por todo conhecimento compartilhado e pelas suas inestimáveis contribuições a minha carreira acadêmica.

Ao meu coorientador, Professor Dr. Guillaume Francis Bertrand, e ao Professor Dr. Victor Hugo Rabelo Coelho pelos conhecimentos compartilhados, trabalhos desenvolvidos e pelas preciosas contribuições para o desenvolvimento deste trabalho.

Aos professores do PPGECAM, pela sua inabalável dedicação ao ensino em tempos tão desafiadores. Aos meus professores da educação básica, que despertaram em mim o amor pela ciência desde tenra idade.

Aos membros do grupo LARHENA, que têm construído o alicerce para que meu trabalho fosse possível ao longo de anos, especialmente aos meus amigos Filipe Lemos, Cinthia Abreu e Eduardo Patriota, por serem tão prestativos. Vocês nunca me deixaram desamparado, mesmo nas horas mais difíceis.

Aos meus líderes profissionais, especialmente a Katy Mendonça, Bruno de Marco e Rodrigo Toshimi, que sempre foram grandes apoiadores, especialmente pela compreensão e flexibilidade no cumprimento da minha jornada de trabalho neste período de aprofundamento acadêmico.

Ao Professor Dr. Gerald Norbert Souza da Silva e ao Dr. Saulo Aires de Souza, membros da Banca Examinadora, pela disponibilidade, apontamentos e compromisso, que guiaram a confecção final deste trabalho. Agradeço à Universidade Federal da Paraíba (UFPB). A todos que de alguma forma contribuíram para a realização deste trabalho, o meu profundo respeito e muito obrigado!

RESUMO

Dados meteorológicos precisos são cruciais para avaliar os impactos da variabilidade espaço-temporal das mudanças climáticas sobre hidrologia, agroecossistemas etc. Este trabalho aborda a importância de registros de precipitação de alta qualidade no cenário climático atual, enfatizando sua relevância não apenas no domínio científico e técnico, mas também para instituições públicas que gerenciam redes pluviométricas. O principal objetivo deste estudo é desenvolver grades de alta resolução (0,25° × 0,25°) de precipitação diária, utilizando dados de mais de 11.000 estações no período de 1961 a 2020. O conjunto de dados é proveniente da Rede Hidrometeorológica Nacional (RHN), submetido a um procedimento de controle de qualidade automático. O procedimento de controle de qualidade automático envolve duas etapas consecutivas: Controle de Qualidade Básico e Controle de Qualidade Absoluto. Avaliações mensais categorizam a qualidade das estações como Muito Baixa, Baixa, Aceitável, Boa ou Excelente, e posteriormente como Alta Qualidade (HQ – *high quality*) e Baixa Qualidade (LQ – *low quality*). A avaliação da metodologia foi conduzida utilizando um conjunto de dados inspecionado visualmente do CEMADEN (Centro Nacional de Monitoramento e Alertas de Desastres Naturais). Os resultados mostraram uma precisão de 98,4% na identificação correta de estações de alta qualidade. De um total de mais de 103 milhões de registros diários, aproximadamente 1,6% foram classificados com qualidade muito baixa ou baixa e, subsequentemente, descartados. O método de interpolação do Inverso da potência das distâncias (IDW – *Inverse Distance Weighting*) foi empregado nos demais 101 milhões de registros diários para produzir uma grade de dados de precipitação diária de alta resolução de 1961 a 2020. As estatísticas de validação cruzada dos dados interpolados mostraram um desempenho superior aos trabalhos anteriores sobre o mesmo conjunto de dados e os dados em grade representaram bem as duas normais climáticas de referência para o período.

PALAVRAS-CHAVE: controle de qualidade, precipitação, dados diários em grade, interpolação

ABSTRACT

Accurate meteorological data are crucial for assessing the impacts of spatiotemporal variability in climate change on hydrology, agroecosystems, etc. This work addresses the significance of high-quality precipitation records in the current climate scenario, emphasizing their importance not only in the scientific and technical domains, but also for public institutions managing pluviometric networks. The primary objective of this study was to develop high-resolution grids (0.25° × 0.25°) of daily precipitation, utilizing data from more than 11,000 stations spanning 1961 to 2020. The dataset was sourced from the Brazilian National Hydrometeorological Network (RHN) was subjected to an automatic quality control procedure. The automatic quality control procedure involves two consecutive steps: Basic Quality Control and Absolute Quality Control. Monthly quality assessments categorized station quality as Very Low, Low, Acceptable, Good, or Excellent, and later as High Quality (HQ) and Low Quality (LQ). The methodology was evaluated using a visually inspected dataset from the Brazilian National Center for Monitoring and Early Warnings of Natural Disasters (CEMADEN). The results showed an accuracy of 98.4% in correctly identifying high-quality stations. Out of over 103 million daily records, approximately 1.6% were flagged as very low or low quality and subsequently discarded. The inverse distance weighting (IDW) interpolation method was employed for the remaining 101 million daily records to produce high-resolution gridded data of daily precipitation from 1961 to 2020. The cross-validation statistics of the interpolated data performed better than those of previous studies on the same dataset, and the gridded data estimations represented both reference climate normals well.

KEYWORDS: quality control, precipitation, daily gridded data, interpolation

SUMMARY

1. INTRODUCTION

Understanding how rainfall behavior changes over time is indispensable for measuring its influence on other environmental parameters and processes in the water cycle, such as soil moisture, evapotranspiration, and groundwater recharge, and natural hazards, such as soil erosion, landslides, floods, and droughts. (Karoly *et al.,* 2003; Souza *et al.,* 2012; Lin *et al.,* 2020; Ghorbanian *et al.,* 2022). In addition to natural processes, rainfall also impacts human activities, such as public health management, agriculture and livestock, urban drainage, public infrastructure design, power generation, and water resource management. (Hacker *et al.,* 2020; Phosri, 2022; Romero *et al.,* 2020; Sokolovskaya *et al.,* 2023).

Despite its importance, the accurate measurement of rainfall remains a challenge. Among the three most common methods used to measure rainfall, rain gauges are more precise than weather radar or orbital remote sensing estimation products (Li *et al.,* 2017). However, the biggest disadvantage of rain gauges is that they can only provide spot measurements; thus, a dense network with well-distributed stations is required to cover large areas with adequate quality, increasing rainfall monitoring costs for equipment installation, maintenance, and operation (Villarini *et al.,* 2008; Hofstra *et al.,* 2009; Chen *et al.,* 2016; Zeng *et al.,* 2018; Merino *et al.,* 2021).

However, limited gauge coverage is commonly observed in many parts of the world, with low-density networks and poor quality data available, particularly in developing countries and sparsely populated areas such as the Brazilian countryside (Buarque *et al.,* 2011; Xu *et al.,* 2013; Murara *et al.,* 2019). An uneven distribution of gauges limits the data coverage over a region unless the data from station-sparse areas can be interpolated to produce a gridded precipitation dataset that can be used in hydrology, climate change, and meteorological studies (Golian *et al.,* 2019; Harris *et al.,* 2020; Bárdossy *et al.,* 2021; Han *et al.,* 2022).

Gauges are designed to operate continuously; however, occasional failures in data collection, processing, and transmission can result in intermittent or unavailable recording of information for extended periods (Sieck *et al.,* 2007). Gauge data can also contain anomalous, repetitive, missing values, and improper null values caused by a range of circumstances such as mechanical problems, electrical faults, power failures, power instabilities, data transmission interruptions, clogging, incorrect sets of time zones or reading of data time, equipment defects, and human errors (Robertson *et al.,* 2015; Ribeiro *et al.,* 2021). These errors may not be visually noticed in a large raw dataset and may cause significant negative interference in hydrological models calibrated using these rainfall datasets (Liu *et al.,* 2018).

Therefore, quality control procedures (QCP) are important for identifying a wider range of missing values and uncertainties to ensure data quality, consistency, and reliability (Hamada *et al.,* 2011; Qi *et al.,* 2016; Yatagai *et al.,* 2020; Jeong *et al.,* 2021; Lewis *et al.,* 2018; Lewis *et al.,* 2021). Several researchers have developed qualitative analysis routines for rainfall data to generate consistent datasets. In countries with greater data availability, there are several studies on QCPs, but most of them focus on temporal ranges of a few decades, apply the method at local and regional scales, or present a method that is difficult to replicate in a data scarcity scenario commonly found in developing countries (Yang *et al.,* 2006; Sciuto *et al.,* 2009; Delahaye *et al.,* 2015; Blenkinsop *et al.,* 2017; Liu *et al.,* 2018; Capozzi *et al.,* 2023).

It is necessary to simplify quality control procedures without losing much information. While developing countries still have deficient rainfall datasets that require high computational demands (Schmidt *et al.,* 2023), rainfall extremes in those regions still need to be addressed, as the efficiency of applications with precipitation data as input depends on its measurement consistency.

For instance, the study carried out by Blenkisop *et al*. (2017) in the United Kingdom (UK) used flags to identify dubious values (particularly extreme events) over hourly rainfall data. They analyzed data from approximately 1600 rain gauges from three different data sources. Although the study successfully built a quality-controlled dataset, two out of the three sources of information that they used were previously submitted to quality control procedures (QCP); thus, this step makes it unsuitable for countries with high-quality data scarcity.

Lewis *et al*. (2018) applied a similar methodology, supplementing the previous study by comparing the hourly data with those of the neighboring gauges, adding another four flags for this purpose. Overall, 3.4% of the hourly data were excluded after the application of this QC procedure; however, as in Blenkinsop's study, many variables used for the flagging were based on a large amount of previous data and studies in the area, including climate UK specificities.

In Catalonia (Northeastern Spain), Llabrés-Brustenga *et al*. (2019) developed a QCP for daily rainfall data applied to more than 1,700 stations on a yearly scale and obtained satisfactory results. This methodology can be divided into three steps: (1) a basic quality control for the detection and deletion of physically impossible values; (2) an absolute quality control where every single time series is tested individually for completeness, occurrence of gaps, distribution of gaps, outliers, etc.; and (3) a relative quality control that evaluates the quality of each daily rainfall value collected by a station based on its similarity to the values collected by neighboring stations. Estévez *et al*. (2022) evaluated the same methodology with fewer adaptations in the

semi-arid region of Andalusia (Southern Spain), a heterogeneous area of study with a predominant semiarid climate, but with other climate conditions as well, such as arid and dry subhumid.

In Brazil, the largest daily rainfall dataset is available on the HidroWeb Portal (https://snirh.gov.br/hidroweb/), a web tool that offers access to a database of hydrometeorological data from the Brazilian National Hydrometeorological Network (RHN). It currently hosts data from more than 12,000 rain gauges from 1855 to the present. With a comprehensive description of the dataset and the application of a QCP, it is possible to determine which stations have the target quality for the development of a gridded dataset of daily rainfall data on a national scale.

On a national scale, Meira *et al*. (2022) developed the first automatic QCP on a sub-hourly scale annually applied to the dataset of the Brazilian National Center for Monitoring and Alerting Natural Disasters (Centro Nacional de Monitoramento e Alertas de Desastres Naturais – CEMADEN) with promising results, correctly identifying 93.9% of pre-defined high quality (HQ) data series and 79.5% of pre-defined poor quality (PQ) data series from more than 3,000 gauges per year for a seven-year period (2014 – 2020).

With the development of new data analysis tools, QCPs have incorporated resources that ensure greater reliability and agility in data processing, resulting in refined data products that are suitable for producing high-quality and long-term gridded datasets (Caesar *et al.,* 2006; Bertoni and Tucci, 2007; Oliveira *et al.,* 2010).

Long-term and high-quality gridded datasets of observed rainfall data with good spatial and temporal scales are important for several types of hydrometeorology research, such as the validation of climate models and satellite-derived estimation models, detection of human influences on climate change, and evaluation of hydrological cycles (Chen *et al.,* 2017; Gallant *et al.,* 2018; Sun *et al.,* 2018; Beck *et al.,* 2019; Huang *et al.,* 2019; Lewis *et al.,* 2019; Pritchard *et al.,* 2023).

There are a variety of products available for precipitation, some of which are based solely on satellite data, others solely on ground-based data, and others combine satellite and ground-based data. Unlike most of the precipitation products, PRISM (Parameter-elevation Regressions on Independent Slopes Model) is an American model that uses ground-based data combined with *digital elevation model* and other *geographical datasets* to estimate monthly and event-based climatic parameters. It is assumed that elevation is the most important factor in the spatial distribution of climatic variables such as precipitation (Daly *et al.,* 2008).

Liebmann and Allured (2005) published the first gridded data of daily precipitation for South America from 1940 to 2003 on a 1.0° by 1.0° grid based on ground-based data from almost 8,000 stations. Most of them were missing some observations throughout the given recording period; all of them were heterogeneously spatialized over the continent, only located east of the Andes Mountains, with the Brazilian territory having the largest quantity and the highest station density on the continent. Basic quality control was addressed by excluding missing periods, incorrectly located stations, and suspected high values that might indicate accumulated rain before the missing days.

Jones *et al*. (2012) developed a daily precipitation grid dataset for the southernmost areas of South America from 1961 to 2000 on a 0.5° × 0.5° grid, focusing on quality control of the daily precipitation dataset, but only comprehending the catchment area of the La Plata Basin. More than half of the 8,000 potential station data points could not be used because they presented several issues, such as insufficient series, duplicated stations, or gross errors.

For the Brazilian territory, Silva *et al.* (2007) analyzed gauge-only precipitation techniques used by the Climate Prediction Center (CPC) of the National Oceanic and Atmospheric Administration (NOAA) of the federal government of the United States to produce historical gridded daily precipitation analyses for Brazil from 2000 (1948-2000) and 2005 datasets, where the daily gauge data passes through several types of quality control, and then a modified Cressman scheme is used as an interpolation method. Subsequently, station observations and gridded results from 12 selected stations were compared to verify accuracy, assess the quality control system, understand extreme precipitation events, and enhance data usability.

Rozante et al. (2010) developed MERGE, a gridded dataset of daily precipitation at 0.25° spatial resolution that consists of combining data from more than 4,00 rain gauges of monitoring networks operated by different Brazilian agencies with data from the GPM-IMERG-EARLY satellite, a substitute for TRMM-TMPA after its discontinuation, although it now presents a higher spatial resolution of 0.1° (Rozante *et al.,* 2020).

Xavier *et al*. (2016) published a meteorological gridded dataset, including precipitation (pr), on high-resolution grids (0.25° × 0.25°) of 3,625 rain gauges and 735 weather stations for the period 1980–2013. A simple quality verification was applied to precipitation to remove physically impossible values, such as values lower than 0 mm and higher than 450 mm (Liebmann and Allured, 2005). Xavier *et al*. (2017) updated the precipitation variable of their previous work using 9,259 rain gauges and extended the range by two more years from 1980 to

2015. Among the two interpolation methods assessed in this updated version, Angular Distance Weighting (ADW) performed better than Inverse Distance Weighting (IDW). In a more recent paper, Xavier *et al*. (2022) presented the latest version of the Brazilian gridded meteorological dataset using observed data from 11,473 rain gauges and 1,252 weather stations from 1961 to 2020. Among the interpolation methods evaluated in this version, Angular Distance Weighting (ADW) and Inverse Distance Weighting (IDW) performed better than the others. As in the first gridded meteorological data from 2016, they only removed physically impossible values from the precipitation in both updates.

In this framework, this master's thesis aims to develop a daily rainfall gridded dataset at the Brazilian territory scale based on quality-controlled data exclusively from the observed precipitation data available online (i.e., throughout HidroWeb). The application of a replicable QCP to the largest Brazilian daily rainfall dataset available for the construction of gridded daily data would improve the quality of hydrological information and its sub-products at a national level (Morbidelli *et al.*, 2020). To achieve this objective, this research was organized along three research axes: (i) collection of historical rainfall data and analysis of the spatiotemporal evolution of the gauge network, (ii) control of data quality throughout a quality index, and (iii) use of this quality-controlled ground-based dataset to develop a high-resolution precipitation grid over Brazil.

2. STUDY AREA

This work was carried out for the whole Brazilian territory, being approximately 8.5 million km^2 and spanning between latitudes 5°16′N – 33°45′S and longitudes 34°47′W – 73°59′W (IBGE, 2017). It is currently divided into twenty-six states and a Federal District. Brazil is the fifth-largest country by area in the world and the largest South American country, corresponding to approximately half of the continent's land.

These large dimensions contribute to great climatic variability, which involves variability in the precipitation patterns. According to the Köppen climate classification (Figure 1), three of the five main climate groups occur in Brazil: A (tropical), B (arid), and C (temperate), and consequently most of their subdivisions. The average annual precipitation has an irregular distribution throughout the territory, ranging from less than 350 mm in semi-arid areas to more than 3,000 mm in tropical areas, with strong seasonal variation throughout the year (Alvares *et al.*, 2013).

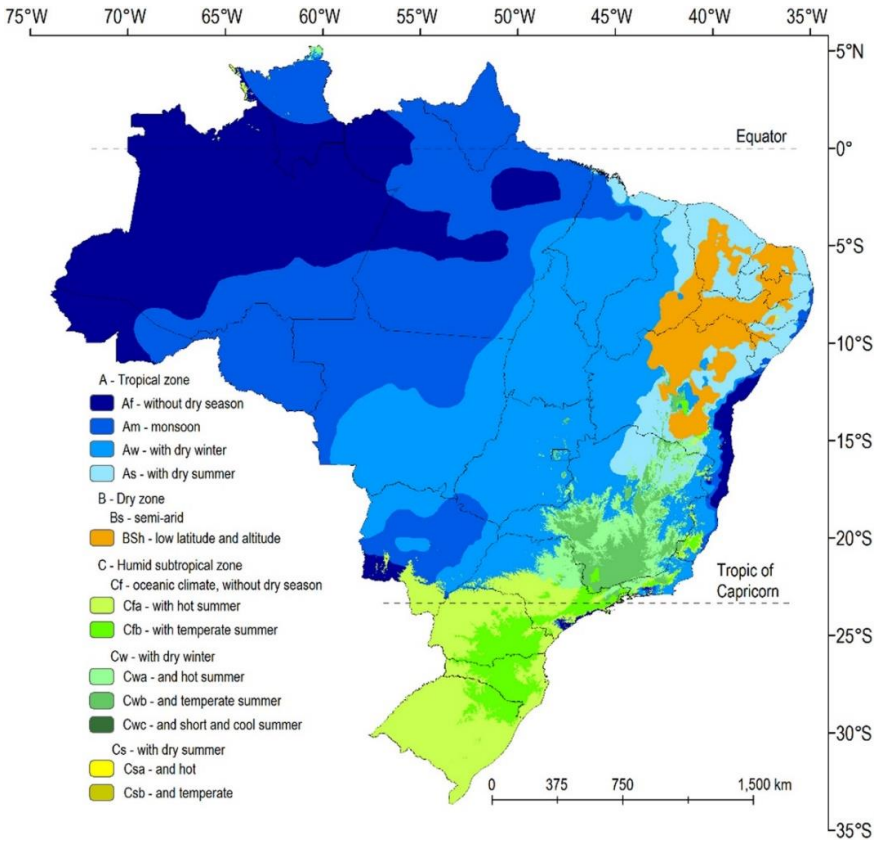


Figure 1 – Köppen climate classification for Brazil. Source: Alvares *et al*. (2013)

3. METHODOLOGY

3.1 Daily Rainfall Dataset

The daily rainfall database is available on the Hidroweb Portal, with individual data series from almost 12,000 stations. All the data were downloaded in January 2023 and updated throughout the year. HidroWeb Portal is an online tool of the Brazilian National Information System on Water Resources (Sistema Nacional de Informações sobre Recursos Hídricos, SNIRH) and provides access to a database containing all the information collected by the Brazilian National Hydrometeorological Network (Rede Hidrometeorológica Nacional, RHN), coordinated by the Brazilian National Water Agency (ANA). It is organized into river levels, flows, rainfall, climatology, water quality, and sediments (ANA, 2019; 2020).

Data were collected using HydroBR, an open-source package in the Python programming language developed by Carvalho (2020). This package provides a direct connection to the hydrometeorological time series stored in the HidroWeb Portal and allows data to be downloaded in the form of unique files for each station, pre-processing, and plotting

hydrometeorological data. Raw data were automatically stored in Hierarchical Data Format (HDF) files, specifically designed to organize large amounts of data and to make the file reading process much faster than traditional file formats (Koranne, 2011). These files were used to list the available stations, their geographical coordinates, periods of operability, percentage of missing values, and years of operation without gaps.

After collecting the individual files of precipitation data, the main data table was consolidated into a single two-dimensional database, totaling approximately 130 million daily records from 1855 to 2020. The structure of the raw data consists of three columns: (i) code, station code in text format; (ii) date, date of recorded data in date format; and (iii) value, daily precipitation value in millimeters (mm) in decimal format. Subsequently, to organize the data in a useful manner, a descriptive data table with additional information for each station, such as city, state, latitude, longitude, and responsible agency, was extracted from the database.

Owing to the extensive size and spatial reach of the dataset, its temporal scope had to be refined. The number of active stations reached only 3,000 measurement points in the 1960s. The inaugural station of Amapá State was established in 1962. Previous studies utilizing the HidroWeb dataset have included data from 1961 to the present day. In alignment with this precedent, the current research concentrates on the period from 1961 to 2020.

3.2 Automatic Quality Control Procedure

QCPs are fundamental to creating reliable and consistent datasets. The proposed method applied in this study is simplified from Llabrés-Brustenga *et al*. (2019) and Estévez *et al*. (2022), adapting the first two major steps:

a) Basic Quality Control: detection of physically impossible data, such as negative values, invalid records, and extreme daily rainfall events;
b) Absolute Quality Control: Perform a series of procedures based on single rain gauge tests to assign a quality index (Q) and a quality label to each station monthly, not yearly, to reduce data loss.

3.2.1 Basic Quality Control

The initial phase involves the identification and elimination of specific erroneous data points. This refers to instances such as the occurrence of physically implausible precipitation quantities that include negative values and values that exceed extreme events within the study region. The upper limit threshold of 450.0 mm.day-1 was adopted for the whole temporal

coverage of the chosen dataset (1961 to 2020), the same value chosen by previous works within the same study domain over Brazil (Liebmann & Allured, 2005; Xavier *et al.,* 2016; Xavier *et al.,* 2017; Xavier *et al.,* 2022).

3.2.2 Absolute Quality Control

During this phase, each series is subjected to the estimation of a quality index (Q) for each month according to Equation (1):

$$Q = \frac{P+Q_1+Q_2+Q_3}{4} \quad (1)$$

This index is derived considering various factors, each of which represents common problems found in daily precipitation data, including the annual percentage of data coverage, pattern of gaps within the dataset, variation in rainfall records on different days of the week, and presence of outliers (Llabrés-Brustenga *et al.,* 2019; Estévez *et al.,* 2022).

The parameter P (availability) represents the percentage of non-null values in an annual series, dividing the number of available daily data by the number of days in a year (365 days or 366 days in leap years).

Parameter Q_1 (gap) measures the distribution of gaps around the period, penalizing larger gaps over smaller ones (Equation 2):

$$Q_1 = 100 - 100\,\frac{\left(2n_{gap}+L_{gap}^{max}\right)}{n} \quad (2)$$

where n_{gap} is the number of null daily values, L_{gap}^{max} is the length of the largest range of null days and n is the number of active days per period.

The parameter Q_2 (weekday) evaluates whether the occurrence of precipitation days in each year is the same for each day of the week, as there is no indication that rainfall follows weekly patterns (Equation 3):

$$Q_2 = 100 - 100\,CV \quad (3)$$

where CV is the ratio of the standard deviation of the occurrence of rainy days for each label day of the week throughout the year (how many rainy Sundays, rainy Mondays, etc.) divided

by its mean value (rainy days in throughout a year divided by the seven days of a week). As the number of rainy days for each day approaches uniformity, the CV tends toward zero. Its fundamental premise rests on the assumption of independence between the frequency of precipitation events and the respective days of the week within a given period (Manola *et al.*, 2019).

Notably, a reduction in the value of this index is indicative of the detection of a particular day of the week experiencing notably fewer instances of rainfall than its counterparts throughout the observed timeframe.

Consequently, the interpretation of the coefficient of variation assumes significance, wherein a diminished value connotes a homogenous distribution of rainy days across the days of the week, thus implying a lack of systematic errors in data recording or collection practices. Conversely, an elevated coefficient of variation signals heightened variability in the distribution pattern, warranting further scrutiny to ascertain and rectify any underlying systemic discrepancies or biases in the dataset (Sanchez-Lorenzo *et al.*, 2012).

Parameter Q_3 (outlier) represents the proportion of days in which the recorded values did not exceed the prpre-definedhreshold for outliers relative to the total number of days. This assessment was conducted monthly, and the interquartile range (IQR) and quartiles were calculated considering the total series, not only for the given year.

In previous studies, the threshold chosen was extreme outliers equal to three times the IQR above the third quartile of the rainfall distribution (Llabrés-Brustenga *et al.*, 2019; Estévez *et al.*, 2022). However, to increase the sensitivity of this parameter, the standard threshold for outliers was maintained at one and a half times the IQR above the third quartile of the rainfall distribution.

The Quality Index (Q) of each station versus each month is the average value of the following components: P (availability), Q_1 (gap), Q_2 (weekday), and Q_3 (outlier). The final parameters Q (quality index) and P (availability) are used to assign qualitative labels to data points hierarchically, such as “Excellent Quality”, “Good Quality”, “Acceptable Quality”, “Low Quality” and “Very Low Quality”, following the rules of Table 1, where both conditions must be true to get the better Qualitative Label as possible.

The qualitative labels were grouped as High Quality (“Excellent Quality”, “Good Quality”, “Acceptable Quality”) and Low Quality (“Low Quality” and “Very Low Quality”) to establish the absolute quality of each station versus month data point.

Table 1 – Quality Label rules based on P (availability) and Q (quality index)

Qualitative Labels	P (availability)	Q (quality index)	Absolute Quality
Excellent Quality	≥ 99	≥ 90	High Quality
Good Quality	≥ 95	≥ 85	High Quality
Acceptable Quality	≥ 90	≥ 80	High Quality
Low Quality	-	≥ 50	Low Quality
Very Low Quality	-	-	Low Quality

3.2.3 Quality Control Procedure Evaluation

An independent dataset of daily precipitation from the Brazilian National Center for Monitoring and Early Warnings of Natural Disasters (CEMADEN) was subjected to Visual Inspection (REF-VI) as a reference dataset during the performance evaluation of the automatic QCP. The analyzed rain gauges per year (station-year series) were previously submitted for the validation of missing days, which removed 9,545 station-year series (41.9% of the station-year series) with more than 60 days of missing data from REF-VI because they were automatically classified as poor-quality gauges. The quality of the remaining data instances was identified in REF-VI by individually observing the station-year series.

The validation process used a confusion matrix that delineated four distinct outcomes: 1) true positives, denoting accurate predictions of the HQ gauges; 2) true negatives, denoting correct predictions of the LQ gauges. Additionally, errors were classified as: 3) False positives (type I error), when the method incorrectly misclassified a gauge as HQ; and 4) False negatives (type II error), representing instances when the method incorrectly misclassified a gauge as LQ.

The other three metrics related to the confusion matrix evaluated were: 1) precision, to verify false positives; 2) accuracy, to test gauge classification; and 3) recall, also known as sensitivity, to measure the strictness of the method by incorrectly flagging many gauges. By applying these metrics, it is possible to evaluate the results and efficiency of the method to flag as many LQ gauges as possible without incorrectly flagging HQ gauges.

3.3 Interpolation Method

Xavier *et al*. (2016) tested six interpolation methodologies for daily precipitation from the same data source: (1) average inside the area of the pixel (AVERAGE), (2) natural interpolation (NATURAL), (3) thin-plate spline (THINPLATE), (4) inverse distance weighting

(IDW), (5) angular distance weighting (ADW), and (6) ordinary point kriging (OPK). The best overall skill scores indicated that ADW and IDW were the best interpolation methods.

Xavier *et al*. (2017) used the IDW and ADW methods to develop new gridded daily precipitation data and found that the ADW interpolation method performed slightly better than the IDW interpolation method. Xavier *et al*. (2022), based on the same data and methodology as Xavier *et al*. (2016), also proposed to reassess both IDW and ADW methods for the interpolation of precipitation data. During cross-validation, it was shown that both methods had very similar statistics.

IDW method was chosen to create a gridded dataset of daily and monthly values that covers the expanse of Brazil, with a spatial resolution set at 0.25° by 0.25° (equivalent to about 27.75 km by 27.75 km on the Equator Line). The main advantages of IDW are its simpler equations and lower computational demands. The generated dataset was structured in the Network Common Data Form (NetCDF) format, incorporating grid coordinates, dates, and interpolated estimations of precipitation for each cell.

Within the context of the IDW method, the interpolated value at a specific location is determined by a weighting factor (see Equation (4), denoted by W_n) that exhibits an inverse relationship with the distance separating the point in question from the data originating from the nth neighboring station (Chen *et al.*, 2017).

$$W_n = \frac{1}{d_n^p} \tag{4}$$

The variable d represents the geodesic distance between the station n and a designated point, where p is equal to 2 (inverse squared distance weighting), which is the same value used in previous studies (Dirks *et al.*, 1998; Goovaert, 2000; Lloyd, 2005; Ahrens, 2006; Ly *et al.*, 2011; Xavier *et al.*, 2016; Xavier *et al.*, 2017; Xavier *et al.*, 2022). The determination of the appropriate stations for interpolation at a point involves consideration of the five nearest stations.

The IDW interpolation method is used to create a gridded dataset of daily and monthly values that cover the expanse of Brazil, with a spatial resolution set at 0.25° × 0.25° (equivalent to approximately 27.75 km × 27.75 km on the Equator Line). The generated dataset was structured in the Network Common Data Form (NetCDF) format, incorporating grid coordinates, dates, and interpolated estimations for precipitation for each cell.

3.4 Cross-validation

Cross-validation is a widely employed technique to assess the accuracy of interpolation methods. The comparison between the observed data (X) and our interpolated estimates (Y) was based on the utilization of the following statistical metrics: Pearson correlation coefficient (R), bias, root mean square error (RMSE), mean absolute error (MAE), compound relative error (CRE), critical success index (CSI), and percent correct (PC) (Hofstra *et al.*, 2008; Wilks, 2011; Xavier *et al.*, 2016).

$$R = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}\sqrt{(X_i - \overline{X})^2}\sqrt{(Y_i - \overline{Y})^2}} \quad (5)$$

$$\text{Bias} = \overline{Y} - \overline{X} \quad (6)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_i - Y_i)^2}{n}} \quad (7)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|X_i - Y_i| \quad (8)$$

$$CRE = \frac{\sum_{i=1}^{n}(X_i - Y_i)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \quad (9)$$

$$CSI = \frac{a}{a + b + c} \quad (10)$$

$$PC = \frac{a + d}{a + b + c + d} \quad (11)$$

In Equations 5 to 9, $\overline{X}$ and $\overline{Y}$ are the mean values of X and Y, respectively, and n is the number of observed daily values. For Equations 10 and 11, CSI and PC are verification measures for categorical forecast performance where a is the number of correct forecasts (hits), b is the number of forecasts not observed (false alarms), c is the number of events not forecasted but occurring (missed forecast), and d is the number of events that were not forecasted and did not occur (correct rejections) (Hofstra *et al.*, 2008). For the definition of "wet days" or "dry days", a "wet day" has a daily precipitation equal to or higher than 0.5 mm.

3.5 Climate Normal Validation

The Climatological Normals were derived by calculating the averages of meteorological variables. Climatologists regularly employ these normals to contextualize recent weather conditions within a historical framework. Climate Normals are commonly featured in local weather news segments for comparison with daily weather conditions.

In addition to weather and climate comparisons, normals find applications in numerous domains, such as regulation by energy companies, energy load forecasting, crop selection and planting times, construction planning, and architectural design.

In 1992, the Brazilian National Institute of Meteorology (INMET), known as the National Meteorology Department of the Ministry of Agriculture and Agrarian Reform, initiated the publication of Climatological Normals, with the first edition covering the period 1961-1990.

Climatological normals from two periods (1961–1990 and 1991–2020) were extracted from the INMET portal (https://portal.inmet.gov.br/normais) and statistically compared with estimated climate normals using Gridded Data.

4. RESULTS AND DISCUSSION

4.1 Observed Data

According to data collected from HidroWeb, the first data record of the pluviometric station registered in Brazil was from the “MINERAÇÃO MORRO VELHO" (station code 01943000), located in the city of Nova Lima (Minas Gerais), which began its precipitation series on January 31, 1855, and continued to record data until July 31, 2018.

Unlike Minas Gerais, as shown in Figure 2, the other Brazilian Federal Units (*Unidades Federativas* – UFs, also known as Brazilian states) took longer to start recording data from their monitoring networks.

The Southeast region maintained the lead with São Paulo and Paraná states in 1888 and 1889, respectively, followed by Rio de Janeiro state in 1900, and Paraíba state, the first UF in the Northeast region to start recording data, also in 1900. In 1907, Bahia became the sixth Brazilian state, and the second in the northeast to begin recording precipitation data.

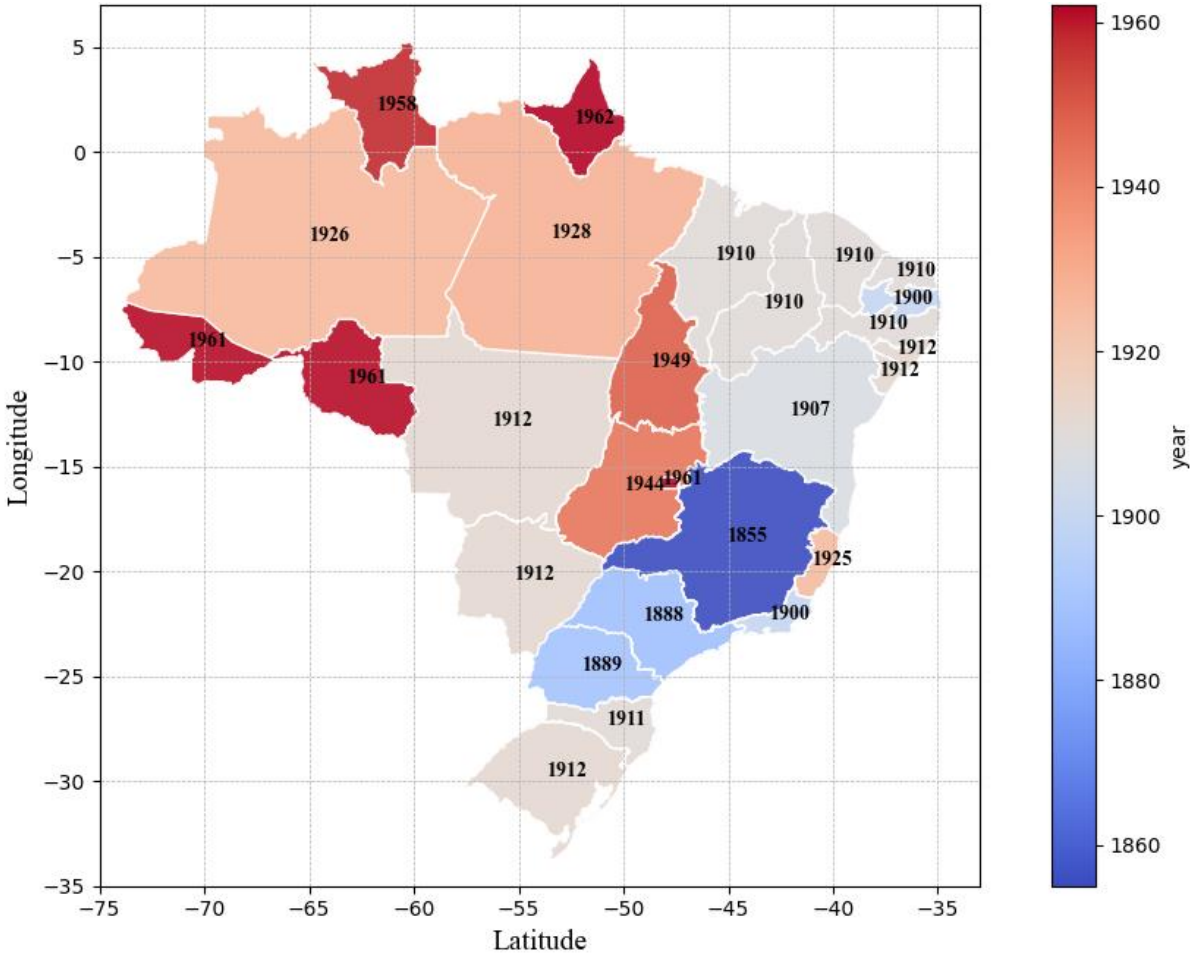


Figure 2 – Year of the first registered data record by Brazilian states in the HidroWeb dataset

In the 1910s and the 1920s, around a dozen states from different regions of the country started recording precipitation data, except for the Northern Region (Amapá state, 1962), Goiás state, and the Federal District, which until then did not exist as a federal unit. Excluding the Northern Region, the UFs that took the longest time to start recording data, as mentioned above, were Goiás state in 1944 and the Federal District in 1961, immediately after the foundation of the new Brazilian capital city, Brasília, in 1960.

According to the historical background presented thus far, temporal coverage focuses on the period from 1961 to 2020. This analysis incorporates data from the remaining 11,128 from the 11,726 original stations, resulting in a comprehensive dataset of 103,027,099 from the 128,943,487 original records of daily precipitation.

As shown in Figure 3, considering the temporal coverage of the study, only 46.4% (5,441) of the stations had more than 30 years of recorded data, which is the minimum period required to calculate the climatic normal for a given region (WMO, 2017).

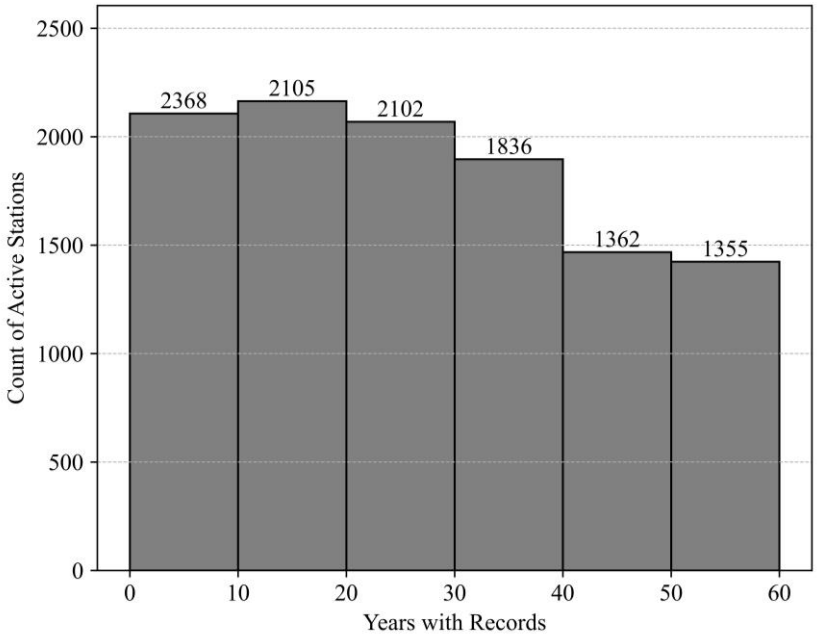


Figure 3 – Distribution of stations by the number of years with data records (1961 – 2020)

The spatial distribution and date series length of each gauge are displayed on the Brazilian state map in Figure 4. A higher density of longer precipitation series was found over the Ceará state, São Paulo state, Paraná state, and Federal District, all cited states were already known for having a good gauge density in general.

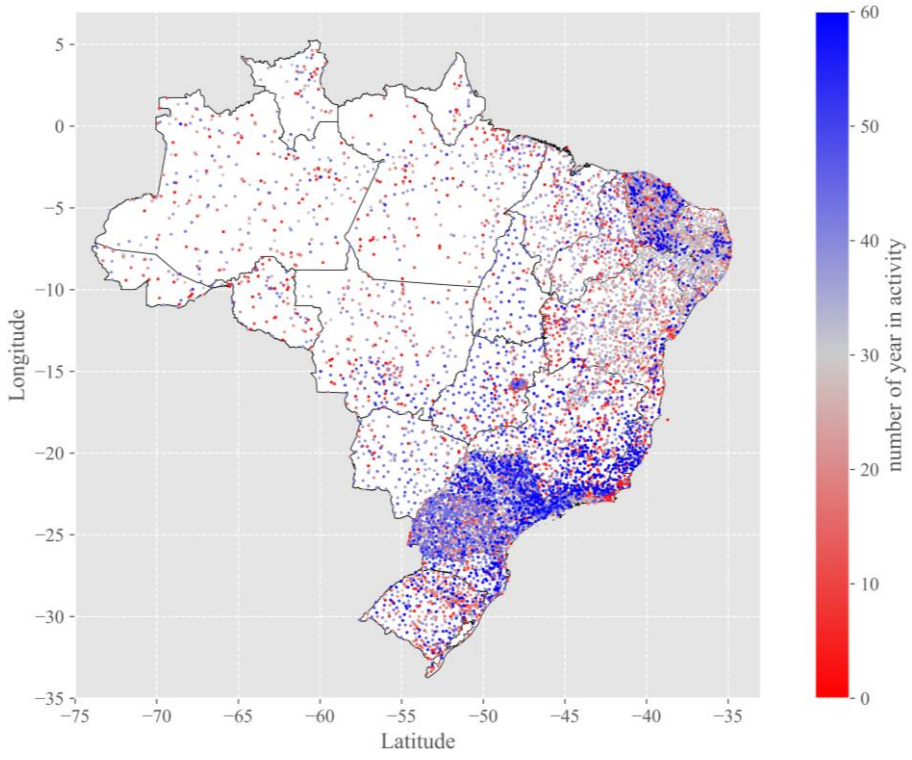


Figure 4 – Spatial Distribution of gauges by years of activity between 1961 to 2020

As shown in Figure 5, the distribution of stations shows significant temporal variability. In 1976, the number of active stations exceeded 6,000, peaking at 6,127 in 1985. From 1986 onward, the number of stations consistently declined, dropping to less than 5,000 stations at the end of the 1990s, less than 4,000 stations in 2016, and around 3,000 stations in 2020. As data from 2021 and 2022 have shown low availability, we opted for the removal of those couple years from the final analysis.

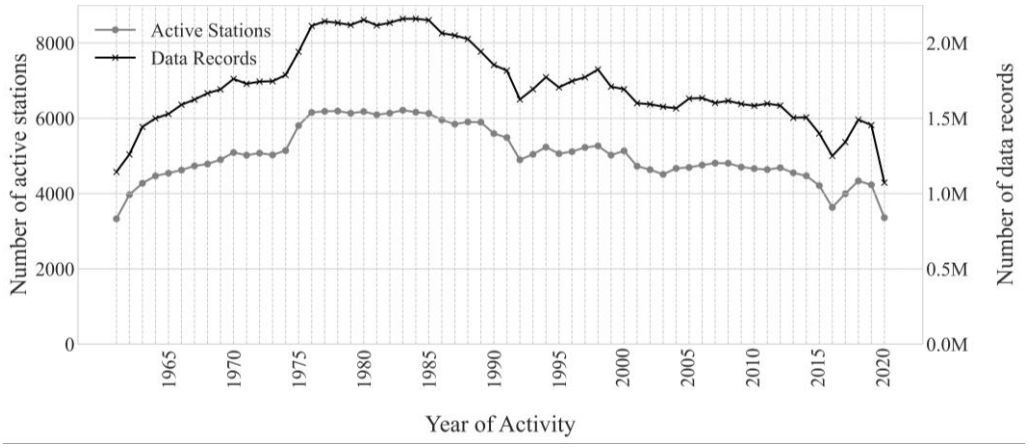


Figure 5 – Number of gauges with recorded data per year of operation from 1961 to 2020

Figure 5 also shows an increase in daily precipitation records from 1961 to 1975, followed by a stable period until 1985, a steady decrease until 1994, a relatively stable period until 2012 with significant fluctuation, and sharp decreases observed after 2014 and 2018. In the last few years of the time series (2014-2020), there was a notable decrease in the daily precipitation records, reaching levels comparable to those observed in the 1990s and the 2000s.

There is noticeable spatial variability in the distribution of stations across the national territory, which directly influences station density among Brazilian states. In absolute numbers, Brazilian states like São Paulo, Bahia, Paraná, Ceará and Minas Gerais are among the ones that have more than 1,000 stations with recorded data from 1961 to 2020 (Figure 6).

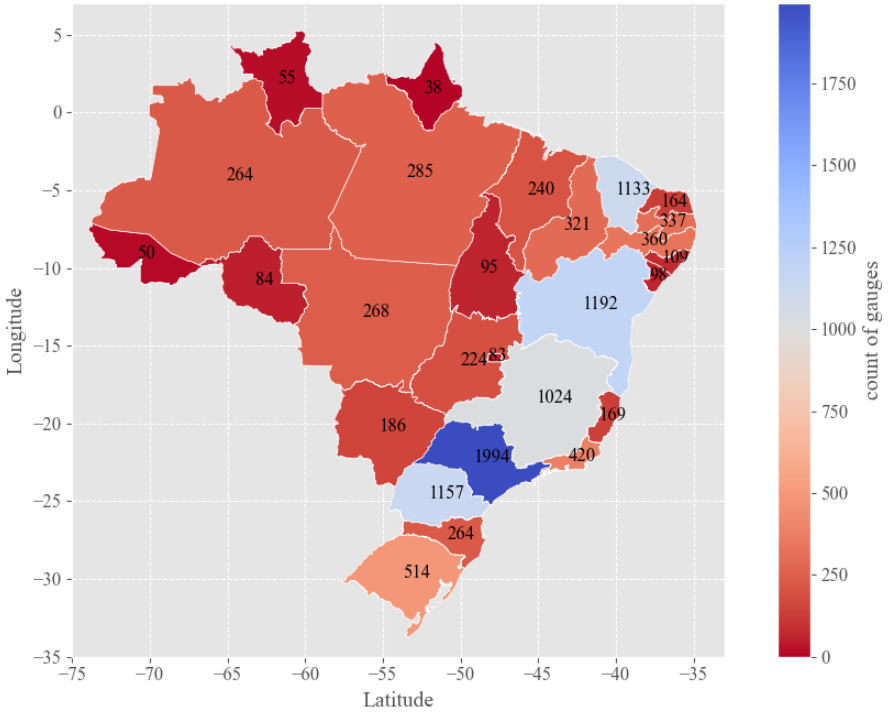


Figure 6 – Number of gauges with recorded data by Brazilian Federative Units

Regarding to the count of active gauges of each responsible agency, the Department of Water and Electric Power (Departamento de Águas e Energia Elétrica – DAEE, São Paulo state), the Water and Land Institute (Instituto Água e Terra – ÁGUASPARANÁ, Paraná state), the Ceará Meteorology and Water Resources Foundation (Fundação Cearense de Meteorologia e Recursos Hídricos – FUNCEME, Ceará state), and the Executive Agency for Water Management (Agência Executiva de Gestão das Águas – AESA, Paraíba State) are entities that operate at the state level, increasing the station density in their states (Figure 7).

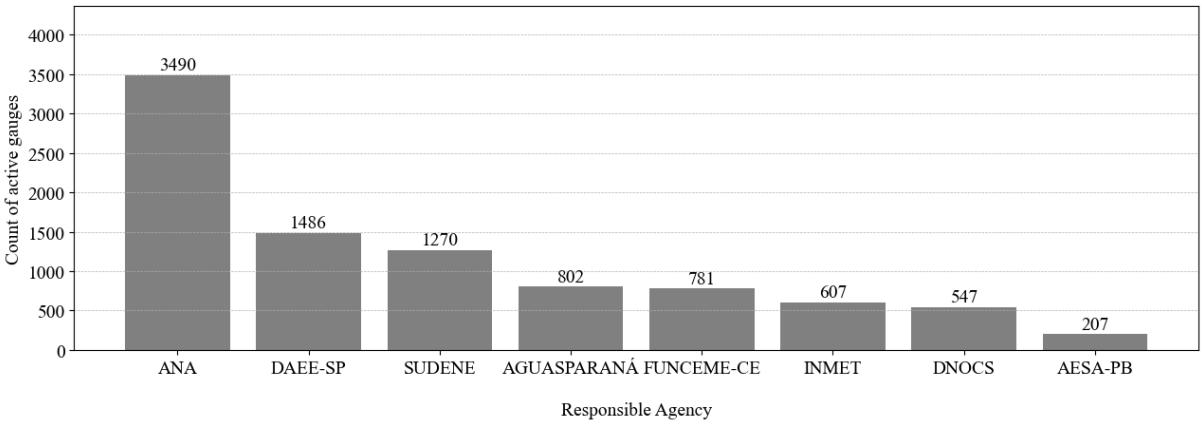


Figure 7 – Number of stations with data records by responsible agency (above 200 gauges)

The Federal District (1 station per 69 km²) and Rio de Janeiro State (1 station per 104 km²) ranked first and second in station density, respectively. São Paulo State (1 station per 124 km²), Ceará State (1 station per 131 km²), Paraíba State (1 station per 167 km²), and Paraná State (1 station per 172 km²) ranked third, fourth, fifth, and sixth in station density, respectively. The second group of states has monitoring networks of independent state agencies, which consequently increases the total number of gauges with available data (Figure 8).

The variation in the number of rain gauges and gauge density between the different Brazilian states can be attributed to a combination of geographical, climatic, and socioeconomic factors. Brazil is characterized by diverse ecosystems ranging from dense Amazon rainforests to arid regions in the northeast. States with more extensive and varied geographical features have a greater need for a higher density of rain gauges to capture the heterogeneity of precipitation patterns across different landscapes.

Additionally, states with more developed economies and urbanization have more resources to invest in meteorological infrastructure. The current network may also be influenced by historical weather patterns, with regions prone to extreme weather events or significant agricultural activities requiring more extensive gauge networks.

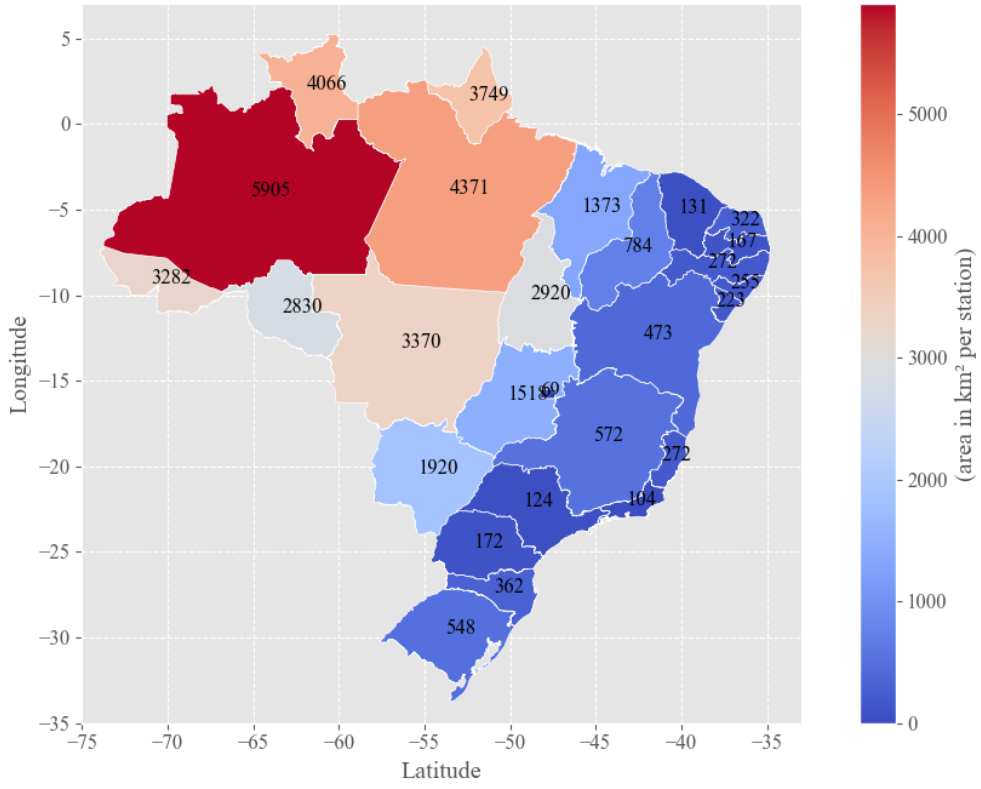


Figure 8 – Density of gauges by Brazilian state (area in km² per station)

4.2 Quality Control Evaluation

An independent dataset from CEMADEN, which included 6,130,973 daily records for a period of seven years from 2014 to 2020, was used to evaluate the proposed automatic QCP. The CEMADEN dataset underwent automatic QCP, where the data were divided into 22,773 station-year series to be analyzed year-by-year.

As shown in Figure 9, in the first row, the presented matrix indicates that from the 7,345 station-year series that belongs to the HQ class of Visual Inspection, approximately 98.4% (7,226 station-year series) were correctly classified by the QCP (True Positive), while only 1.6% (119 station-year series) were incorrectly classified as LQ and would have been removed from the final dataset (False Negative). In the second row, from the 15,428 station-year series that belong to the LQ class of Visual Inspection, approximately 68.3% (10,543 station-year series) were correctly classified as LQ (True Negative), and 31.7% (4,885 station-year series) were misclassified as HQ (False Positive). The results align with the main objective of the proposed QCP, which aims to minimize the exclusion of HQ gauge data from the final dataset.

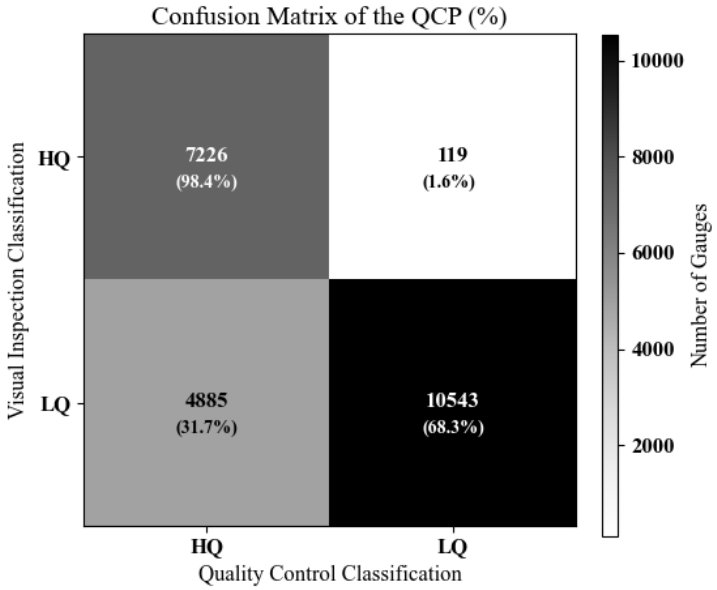


Figure 9 – Confusion Matrix for the Automatic QCP applied to the daily CEMADEN's data

The discernment between True Negatives and False Positives emerged as a noteworthy indicator, revealing that nearly 70% of the LQ station-year series were effectively identified and excluded through automated processes. Only a minimal fraction of the HQ station-year series (approximately 1.6%) was discarded from the remaining records.

Meira *et al*. (2022) obtained better results (78,5% vs. 68,3% from the QCP) when correctly classifying the LQ station-year series and slightly worse results (5.8%) when incorrectly classifying the HQ station-year series using the same dataset, but on a subhourly scale. The efficacy of the QCP is in selectively targeting and maintaining superior quality rain gauges while simultaneously discerning and mitigating the inferior quality counterparts.

A comparison between the scores from the confusion matrix and those from Meira *et al*. (2022) is presented in Table 2. The **accuracy** of 0.78 shows the overall correctness of the classifier, considering both true positive and true negative predictions in relation to the total instances. Although accuracy provides a general measure of performance, it may not be sufficient in cases of imbalanced datasets where one class dominates. In such instances, precision and recall are crucial for more nuanced evaluation. A **precision** of 0.60 indicates that when the model predicts a positive class, it is approximately 60% of the time. This metric is particularly useful where the cost of false positives is high, thereby emphasizing the importance of precise predictions. The **recall**, or true positive rate, is approximately 0.98, which means that the model successfully identifies approximately 98% of the station-year series belonging to HQ. Recall is particularly relevant when the cost of false negatives is high, as it gauges the model's ability to capture all instances of the positive class. In summary, accuracy suggests an overall effectiveness, whereas precision and recall offer a more nuanced understanding.

Table 2 – Comparison of classification scores with Meira *et al*. (2022)

Scores	**Proposed A-QCP (daily)**	**Meira *et al*. (2022) (subdaily)**
Accuracy	0.78	0.84
Precision	0.60	0.69
Recall	0.98	0.94

4.3 Quality Index (Q)

After excluding all data values lower than 0 mm and higher than 450 mm, approximately 0.0041% of the total dataset was discarded, which corresponds to 4,222 daily data records. The remaining dataset contained 103,022,877 daily records. In 2023, recent extreme events of daily rainfall in the littoral region of São Paulo surpassed 600 mm a day, a value two times higher than the previous record of daily precipitation for this region, which suggests that eventual updates need to reassess the value of the upper limit threshold of this step (Carmona *et al.*, 2023).

Absolute Quality Control step was applied on a monthly scale (approximately 3.4 million monthly series), with an average of 306.5 months verified for each of the 11,128 stations, equivalent to an average of 25.5 analyzed years.

As shown in Figure 10, the completeness index (P) and the gap index (Q_1) are highly correlated. The Pearson correlation coefficient between these two components was equal to 0.79, and the p-value was 0.00. In general, if the p-value is less than the chosen significance level (commonly 0.05), the null hypothesis is rejected, and it is concluded that there is a statistically significant correlation.

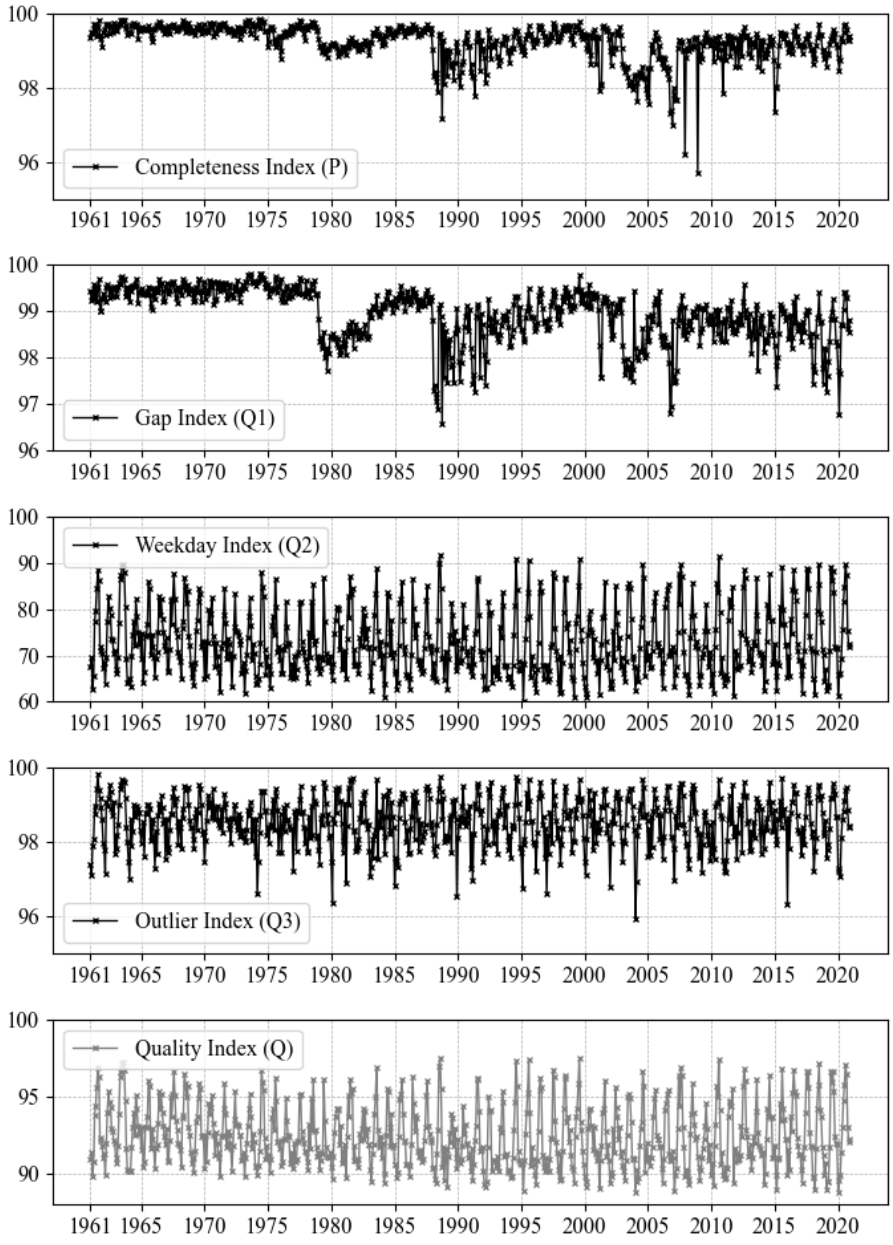


Figure 10 – Monthly average quality index (Q) and its four components.

The Weekday Index (Q_2) has a very similar fluctuation behavior to the Outlier Index (Q_3), although they span different minimum and maximum values. The Pearson correlation coefficient between these two components was equal to 0.76 and the p-value equal to. Daily skipping of measurements can result in unexpected outliers caused by accumulation. Table 3 shows the descriptive statistics of the quality index and its four components grouped by month.

Table 3 – Descriptive Statistics of Q grouped by month (720 months, 60 years)

Quality Index Component	Min	Max	Mean	Standard Deviation
Completeness Index (P)	95.70	99.84	99.23	0.49
Gap Index (Q_1)	96.59	99.82	98.93	0.59
Weekday Index (Q_2)	60.27	91.50	72.89	7.14
Outlier Index (Q_3)	95.94	99.81	98.55	0.65
Quality Index (Q)	0.98	0.94	92.40	2.00

4.4 Cross-validation

Although only the statistical results of Xavier *et al*. (2022) are in the time coverage used in the quality-controlled dataset, the cross-validation statistics of the other previous works are similar (Xavier *et al.*, 2016; Xavier *et al.*, 2017). We performed cross-validation statistics for each day of the period (1961 – 2020) to evaluate the temporal performance of precipitation estimation. Table 4 presents a comparison between the cross-validation statistics results of the quality-controlled dataset and those of previous studies.

Table 4 – Cross-validation statistics and comparison with previous works (IDW method)

IDW Statistics	Xavier *et al*. (2016) [34 years]	Xavier *et al*. (2017) [36 years]	Xavier *et al*. (2022) [60 years]	Quality-Controlled Dataset [60 years]
R	0.609	0.633	0.642	*0.860*
Bias	0.004	0.003	0.006	*0.000*
RMSE	9.141	8.822	8.470	*5.526*
CRE	0.666	0.632	0.621	*0.265*
MAE	3.709	3.366	3.192	*2.060*
PC	0.783	0.798	0.802	*0.824*
CSI	0.534	0.530	0.529	*0.585*

Figure 11a–f shows the cross-validation statistical results and precipitation for the first day of each decade as examples, presenting 2D histograms for the observed precipitation versus their corresponding precipitation estimations for all available rain gauges on each selected day, providing a visual representation of the data distribution and frequency in two dimensions.

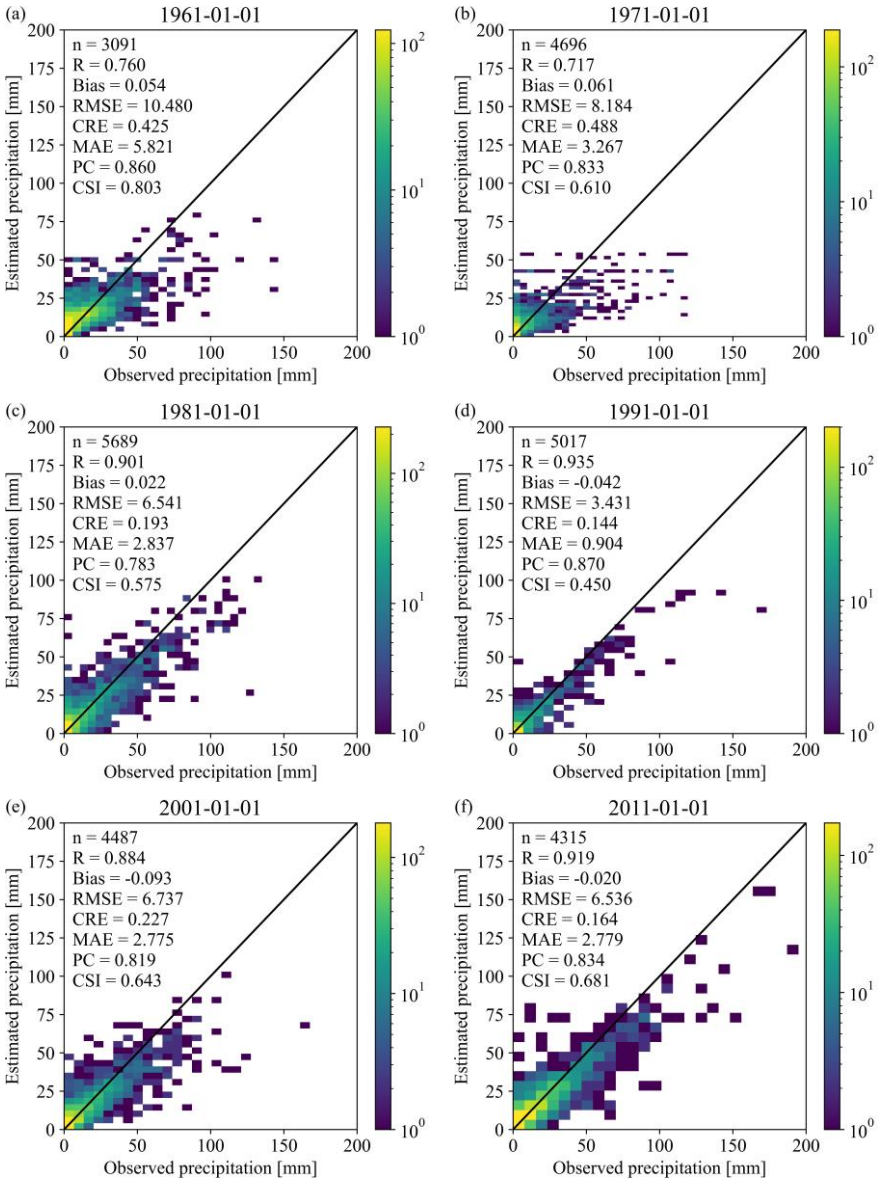


Figure 11 – 2D histogram plots of observed precipitation and their estimations, where the diagonal line represents the ideal correlation between observed precipitation and estimated precipitation, and the assigned color of each cell is based on the frequency of the precipitation value on the first day of each decade. (a) January 1, 1961; (b) January 1, 1971; (c) January 1, 1981; (d) January 1, 1991; (e) January 1, 2001; and (f) January 1, 2011.

Figure 12a–c presents the statistical results of R (correlation coefficient), Bias, and RMSE (root mean square error) for each day of the given period of 21,915 days from 1961 to 2020 for each year and by day of the year in chronological order (DOY). The daily average value of R is 0.846, ranging from 0.378 to 0.987. The bias ranged from -0.263 to 0.312, and its daily average was approximately 0.003. The average value of RMSE was 4.973 mm.day^{-1} and ranges from 0.473 to 19.230.

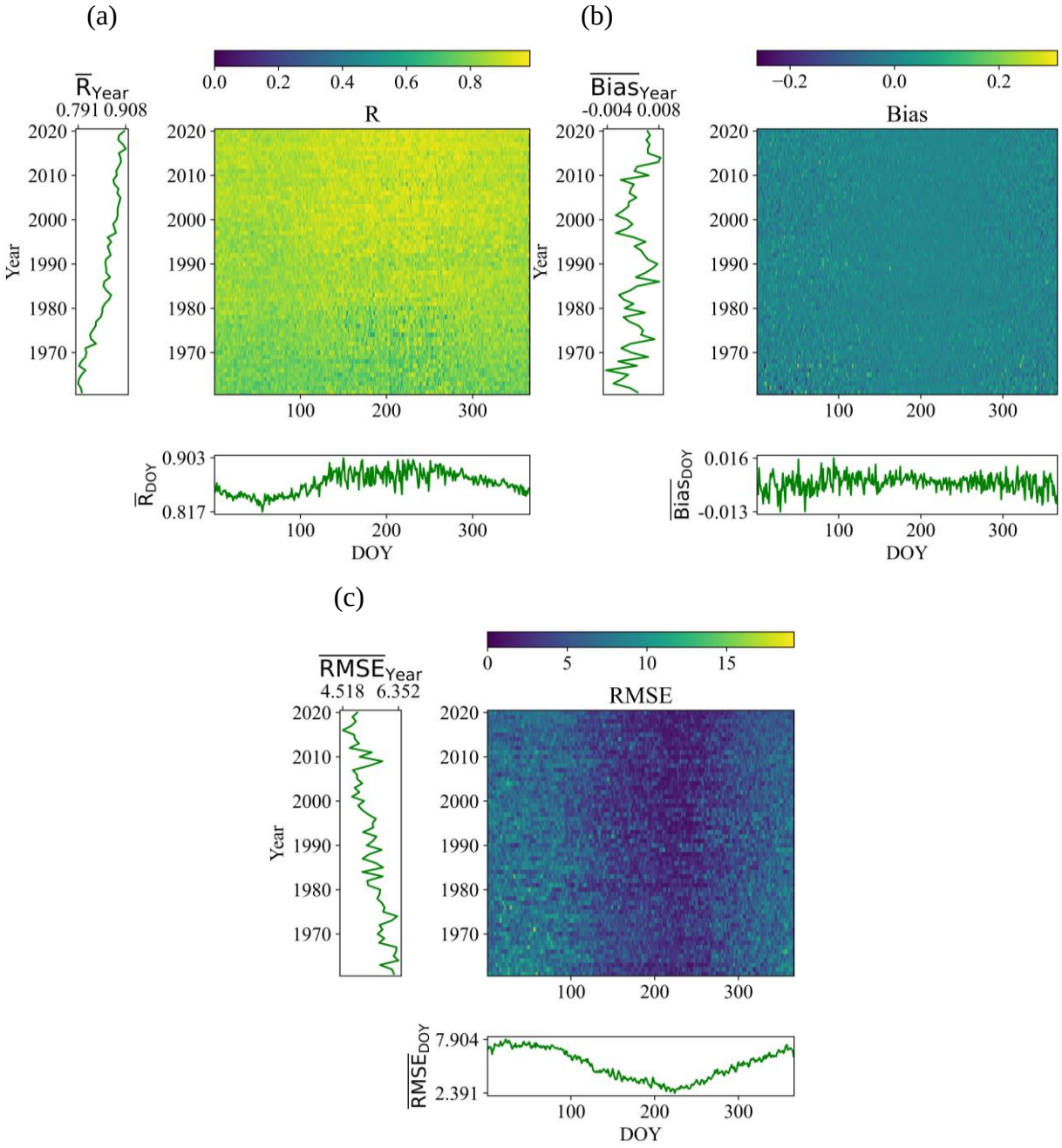


Figure 12 – Cross-validation statistics of daily averages for each day (main plot), each year (left plot), and by DOY (bottom plot) for (a) R, (b) bias, and (c) RMSE.

Both R and RMSE statistics have improved over the years, while Bias has a fluctuating trend, alternating between positive and negative tendencies over the years. All three statistical results exhibit sensitivity to a seasonal pattern in the same period of the general rainy season from late November to early May; RMSE values reach their highest points (peak), R values show their lowest points (valley), and Bias values tend to exhibit extremes.

4.5 Climate Normals vs. Gridded Data

A comprehensive analysis comparing 60 years (1961–2020) of daily rainfall gridded data with two climate normals provided by the Instituto Nacional de Meteorologia (INMET). The first climate normal spans the period from 1961 to 1990, whereas the second encompasses the years from 1991 to 2020. The analysis involved assessing the Pearson Coefficient Correlation (R), Bias, and Root Mean Square Error (RMSE) between the two climate normals and the gridded data for each period separately. Gridded data were aggregated per month and station to generate statistical results. As shown in Table 5, both periods show similar statistics.

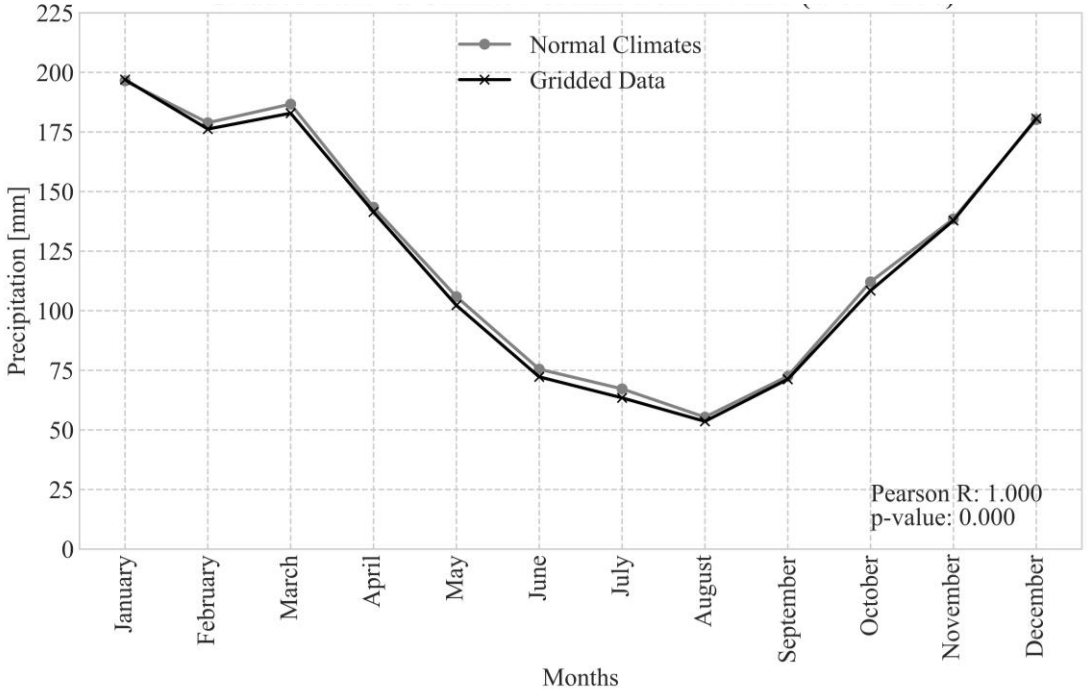
Table 5 – Comparison of Climate Normals and Gridded Data

Gridded Data Statistics	**Climate Normal (1961 – 1990)**	**Climate Normal (1991 – 2020)**
R	0.962	0.979
Bias	-2.138 mm	-4.980 mm
RMSE	24.722 mm	20.463 mm

The correlation coefficient (R) between the gridded data and the reference values (climate normal) for both periods was high, but there was a systematic underestimation (negative bias). A lower RMSE value suggests that the measured rainfall data are closer to the reference values from the climate normals for the later period, indicating higher accuracy or precision, and there is less error or uncertainty associated with the measured rainfall data in the later period (1991-2020) compared to the earlier period (1961-1990).

Monthly aggregated values from the gridded data presented a similar behavior to the climate normals over the studied period. As shown in Figure 13a–b, both monthly representations of the estimated data follow the same behavior as the reference data, with a correlation coefficient (R) virtually equal to 1. In other words, the two variables were linearly related, indicating a direct and strong relationship between them.

(a) 1961 – 1990



(b) 1991 – 2020

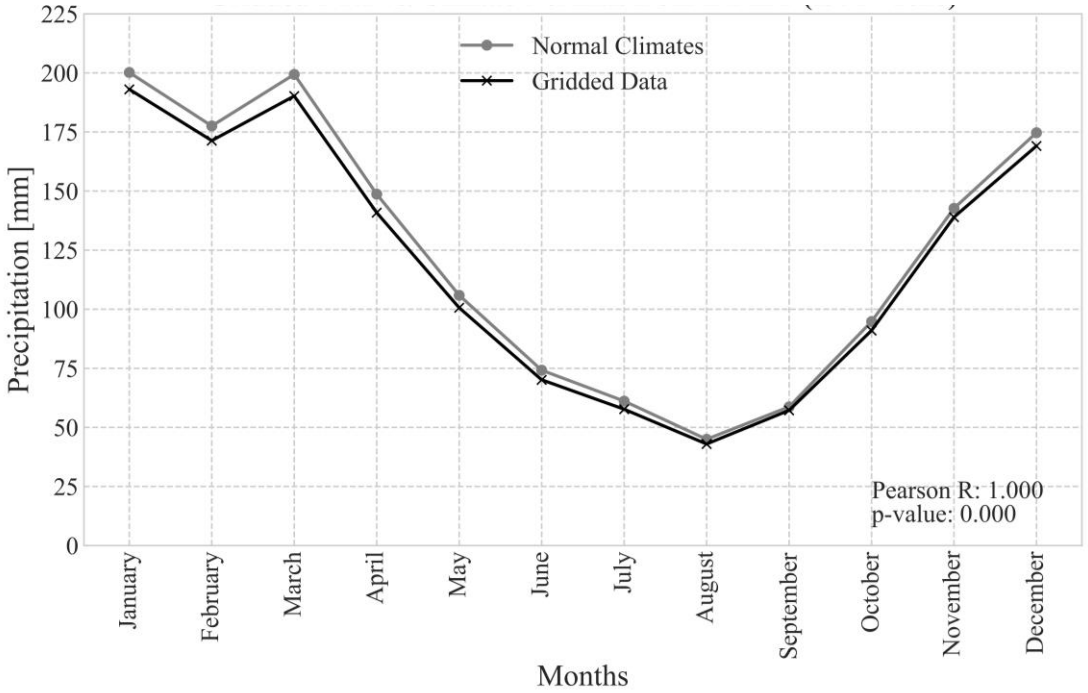


Figure 13 – Climate normals aggregated by month and gridded data estimations for the period of (a) 1961 to 1990 and (b) 1991 to 2020.

5. CONCLUSIONS AND RECOMMENDATIONS

This study developed a new gridded daily precipitation dataset for Brazil from a quality-controlled dataset based on a 60-year time series from more than 11,000 stations across the Brazilian land, and provided it to researchers from several scientific disciplines, such as hydrology, water resources, environmental science, agricultural science, disaster management, climate change research, urban planning, ecology and conservation, and renewable energy.

The procedures and thresholds used in the automatic QCP and QCP evaluations were adopted equally for the entire Brazilian territory. The automatic QCP originally sought to maintain as many HQ stations as in the final dataset. The results of the QCP evaluation were satisfactory because the confusion matrix showed that only 119 (1.6%) of the 7,345 ground-truth HQ gauges would be removed from the testing dataset; however, to achieve this result, the automatic QCP could not be overly stringent. As a result, ground-truth LQ gauges (31.7%) would have been mistakenly kept in the testing dataset after the QCP evaluation. After applying the automatic QCP over the original dataset (approximately 103 million data records), the analysis of the Quality Index (Q) and its components revealed insightful patterns. The Weekday Index (Q2) demonstrated the highest sensitivity, which is crucial for confirming the accurate registration of data records on specific days of the week, especially on weekends.

The IDW method was chosen to generate a new gridded precipitation dataset covering mainland Brazil with a daily temporal resolution and a 0.25° × 0.25° spatial resolution for the period of 1961 to 2020. The cross-validation method used 101,961,839 daily records from 21,915 days of precipitation data. The cross-validation statistics, when compared with those of previous studies, showed favorable results, with improvements in all statistical results. The statistical outcomes demonstrated improvements in R (0.860) and RMSE (5.526 mm) over the years, while Bias (0.000 mm) exhibited a fluctuating trend, alternating between positive and negative tendencies. Moreover, the statistical results exhibited sensitivity to seasonal patterns, aligning with the general rainy season from late November to early May.

This study not only contributes to a new quality-controlled precipitation dataset, but also provides a refined dataset with reliability and applicability of precipitation data for broader scientific research and applications. The gridded data also performed well, perfectly representing the behavior of the climate normals from INMET in both analyzed periods.

Subsequent refinements and recommendations are suggested for future investigations: (i) divide the spatial coverage of the study area into smaller regions and implement the procedure in homogenous regions characterized by similar climate and rainfall regimes, thereby

enhancing the adjustment of thresholds for each rainfall regime; (ii) combine satellite data and digital elevation models into the interpolation method to increase data density and interpolation quality; and (iii) reassess the value of the upper limit threshold according to recent extreme daily rainfall events.

6. REFERENCES

Agência Nacional de Águas e Saneamento Básico (ANA). (2019). *HidroWeb v3.2.7*. Brasil. https://www.snirh.gov.br/hidroweb/apresentacao

Agência Nacional de Águas e Saneamento Básico (ANA). (2020). *Centro de Memórias da ANA*. Brasil. https://memoria.ana.gov.br

Ahrens, B. (2006). Distance in spatial interpolation of daily rain gauge data. *Hydrology and Earth System Sciences*, 10(2), 197–208. https://doi.org/10.5194/hess-10-197-2006

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

Bárdossy, A., Modiri, E., Anwar, F., & Pegram, G. (2021). Gridded daily precipitation data for Iran: A comparison of different methods. *Journal of Hydrology: Regional Studies*, 38, 100958–100958. https://doi.org/10.1016/j.ejrh.2021.100958

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

Bertoni, J. C., & Tucci, C. E. M. (2007). Precipitação. In C. E. M. TUCCI (Ed.), *Hidrologia: ciência e aplicação* (pp. 177-241). UFRGS.

Blenkinsop, S., Lewis, E., Chan, S. C., & Fowler, H. J. (2017). Quality-control of hourly rainfall dataset and climatology of extremes for the UK. *International Journal of Climatology*, 37(2), 722–740. https://doi.org/10.1002/joc.4735

Buarque, D. C., de Paiva, R. C. D., Clarke, R. T., & Mendes, C. A. B. (2011). A comparison of Amazon rainfall characteristics derived from TRMM, CMORPH and the Brazilian national rain gauge network. *Journal of Geophysical Research*, *116*(D19). https://doi.org/10.1029/2011jd016060

Caesar, J., Alexander, L., & Vose, R. (2006). Large-scale changes in observed daily maximum and minimum temperatures: Creation and analysis of a new gridded data set. *Journal of Geophysical Research*, *111*(D5). https://doi.org/10.1029/2005jd006280

Capozzi, V., Annella, C., & Budillon, G. (2023). Classification of daily heavy precipitation patterns and associated synoptic types in the Campania Region (southern Italy).

Atmospheric Research, *289*, 106781–106781. https://doi.org/10.1016/j.atmosres.2023.106781

Carmona, M. I. A., Moroz, C. B., Ferrer, J. V., Oberhagemann, L., Mohor, G. S., Skålevåg, A., … Thieken, A. (2023). A Multi-Hazard Perspective on The São Sebastião-SP Event in February 2023: What Made It a Disaster? *Proceedings of the XXV Brazilian Water Resources Symposium*, XXV-SBRH0287. Brazilian Association for Water Resources (ABRHidro).

Carvalho, W. M. de. (n.d.). *HydroBr*: a Python package to work with Brazilian hydrometeorological time series. Zenodo. Retrieved October 26, 2023, from http://dx.doi.org/10.5281/ZENODO.3931027

Chen, F., & Li, X. (2016). Evaluation of IMERG and TRMM 3B43 Monthly Precipitation Products over Mainland China. *Remote Sensing*, *8*(6), 472. https://doi.org/10.3390/rs8060472

Chen, T., Ren, L., Yuan, F., Yang, X., Jiang, S., Tang, T., Liu, Y., Zhao, C., & Zhang, L. (2017). Comparison of Spatial Interpolation Schemes for Rainfall Data and Application in Hydrological Modeling. Water, *9*(5), 342. https://doi.org/10.3390/w9050342

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., Curtis, J., & Pasteris, P. P. (2008). Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology*, *28*(15), 2031–2064. https://doi.org/10.1002/joc.1688

Delahaye, F., Kirstetter, P.-E., Dubreuil, V., Machado, L. A. T., Vila, D. A., & Clark, R. (2015). A consistent gauge database for daily rainfall analysis over the Legal Brazilian Amazon. *Journal of Hydrology*, *527*, 292–304. https://doi.org/10.1016/j.jhydrol.2015.04.012

Dirks, K. N., Hay, J. E., Stow, C. D., & Harris, D. (1998). High-resolution studies of rainfall on Norfolk Island. *Journal of Hydrology*, 208(3-4), 187–193. https://doi.org/10.1016/s0022-1694(98)00155-3

Estévez, J., Llabrés-Brustenga, A., Casas-Castillo, M. C., García-Marín, A. P., Kirchner, R., & Rodríguez-Solà, R. (2022). A quality control procedure for long-term series of daily precipitation data in a semiarid environment. *Theoretical and Applied Climatology*, *149*(3-4), 1029–1041. https://doi.org/10.1007/s00704-022-04089-2

Gallant, A., Sadinski, W., Brown, J., Senay, G., & Roth, M. (2018). Challenges in Complementing Data from Ground-Based Sensors with Satellite-Derived Products to Measure Ecological Changes in Relation to Climate—Lessons from Temperate Wetland-Upland Landscapes. *Sensors*, 18(3), 880. https://doi.org/10.3390/s18030880

Ghorbanian, A., Mohammadzadeh, A., Jamali, S., & Duan, Z. (2022). Performance Evaluation of Six Gridded Precipitation Products throughout Iran Using Ground Observations over the Last Two Decades (2000–2020). *Remote Sensing*, 14(15), 3783. https://doi.org/10.3390/rs14153783

Golian, S., Javadian, M., & Behrangi, A. (2019). On the use of satellite, gauge, and reanalysis precipitation products for drought studies. *Environmental Research Letters*, 14(7), 075005. https://doi.org/10.1088/1748-9326/ab2203

Goovaerts, P. (2000). Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology*, 228(1-2), 113–129. https://doi.org/10.1016/s0022-1694(00)00144-x

Hacker, K. P., Sacramento, G. A., Cruz, J. S., de Oliveira, D., Nery, N., Lindow, J. C., Carvalho, M., Hagan, J., Diggle, P. J., Begon, M., Reis, M. G., Wunder, E. A., Ko, A. I., & Costa, F. (2020). Influence of Rainfall on Leptospira Infection and Disease in a Tropical Urban Setting, Brazil. *Emerging Infectious Diseases*, *26*(2), 311–314. https://doi.org/10.3201/eid2602.190102

Hamada, A., Arakawa, O., & Yatagai, A. (2011). An automated quality control method for daily rain-gauge data. *Global Environmental Research*, *15*(2), 183–192.

Han, J., Miao, C., Gou, J., Zheng, H., Zhang, Q., & Guo, X. (2022). A new daily gridded precipitation dataset based on gauge observations across mainland China. *Reviews of Geophysics*, 56(1), 79–107. https://doi.org/10.5194/essd-2022-373

Harris, I., Osborn, T. J., Jones, P., & Lister, D. (2020). Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Scientific Data*, 7(1). https://doi.org/10.1038/s41597-020-0453-3

Hofstra, N., Haylock, M., New, M., Jones, P., & Frei, C. (2008). Comparison of six methods for the interpolation of daily European climate data. *Journal of Geophysical Research*, *113*(D21). https://doi.org/10.1029/2008jd010100

Hofstra, N., New, M., & McSweeney, C. (2009). The influence of interpolation and station network density on the distributions and trends of climate variables in gridded daily data. *Climate Dynamics*, 35(5), 841–858. https://doi.org/10.1007/s00382-009-0698-1

Huang, Y., Bárdossy, A., & Zhang, K. (2019). Sensitivity of hydrological models to temporal and spatial resolutions of rainfall data. Hydrology and Earth System Sciences, 23(6), 2647–2663. https://doi.org/10.5194/hess-23-2647-2019

Instituto Brasileiro de Geografia e Estatística (IBGE). (2017). *Divisão regional do Brasil em regiões geográficas imediatas e regiões geográficas intermediárias: 2017* (p. 80).

Instituto Nacional de Meteorologia (INMET). (n.d.). Normais Climatológicas. Retrieved March 17, 2024, website: https://portal.inmet.gov.br/servicos/normais-climatol%C3%B3gicas

Jeong, G., Yoo, D.-G., Kim, T.-W., Lee, J.-Y., Noh, J.-W., & Kang, D. (2021). Integrated Quality Control Process for Hydrological Database: A Case Study of Daecheong Dam Basin in South Korea. *Water*, *13*(20), 2820–2820. https://doi.org/10.3390/w13202820

Jones, P. D., Lister, D. H., Harpham, C., Rusticucci, M., & Penalba, O. (2012). Construction of a daily precipitation grid for southeastern South America for the period 1961-2000.

International Journal of Climatology, *33*(11), 2508–2519. https://doi.org/10.1002/joc.3605

Karoly, D. J., Braganza, K., Stott, P. A., Arblaster, J. M., Meehl, G. A., Broccoli, A. J., & Dixon, K. W. (2003). Detection of a Human Influence on North American Climate. *Science*, *302*(5648), 1200–1203. https://doi.org/10.1126/science.1089159

Koranne, S. (2011). Hierarchical Data Format 5: HDF5. In: Handbook of Open Source Tools. Springer, Boston, MA. https://doi.org/10.1007/978-1-4419-7719-9_10

Lemos, F. C., Coelho, V. H. R., Freitas, E. da S., Tomasella, J., Bertrand, G. F., Meira, M. A., … Almeida, C. das N. (2023). Spatiotemporal distribution of precipitation and its characteristics under tropical atmospheric systems of Brazil: Insights from a large sub-hourly database. *Hydrological Processes*, 37(11). https://doi.org/10.1002/hyp.15017

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

Lewis, E., Pritchard, D., Villalobos-Herrera, R., Blenkinsop, S., McClean, F., Guerreiro, S., Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Rustemeier, E., & Fowler, H. J. (2021). Quality control of a global hourly rainfall dataset. *Environmental Modelling & Software*, 144, 105169. https://doi.org/10.1016/j.envsoft.2021.105169

Lewis, E., Quinn, N., Blenkinsop, S., Fowler, H. J., Freer, J., Tanguy, M., Hitt, O., Coxon, G., Bates, P., & Woods, R. (2018). A rule based quality control method for hourly rainfall data and a 1 km resolution gridded hourly rainfall dataset for Great Britain: CEH-GEAR1hr. *Journal of Hydrology*, 564, 930–943. https://doi.org/10.1016/j.jhydrol.2018.07.034

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

Liebmann, B., & Allured, D. (2005). Daily Precipitation Grids for South America. *Bulletin of the American Meteorological Society*, *86*(11), 1567–1570. https://doi.org/10.1175/BAMS-86-11-1567

Lin, Q., Wang, Y., Glade, T., Zhang, J., & Zhang, Y. (2020). Assessing the spatiotemporal impact of climate change on event rainfall characteristics that influence landfall occurrences based on multiple GCM projections in China. *Climatic Change*, *162*(2), 761–779. https://doi.org/10.1007/s10584-020-02750-1

Liu, S., Li, Y., Pauwels, N., & Walker, J. P. (2018). Impact of Rain Gauge Quality Control and Interpolation on Streamflow Simulation: An Application to the Warwick Catchment, Australia. *Frontiers in Earth Science*, *5*. https://doi.org/10.3389/feart.2017.00114

Llabrés-Brustenga, A., Rius, A., Rodríguez-Solà, R., Casas-Castillo, M. C., & Redaño, A. (2019). Quality control process of the daily rainfall series available in Catalonia from

1855 to the present. *Theoretical and Applied Climatology*, *137*(3-4), 2715–2729. https://doi.org/10.1007/s00704-019-02772-5

Lloyd, C. D. (2005). Assessing the effect of integrating elevation data into the estimation of monthly precipitation in Great Britain. *Journal of Hydrology*, 308(1-4), 128–150. https://doi.org/10.1016/j.jhydrol.2004.10.026

Ly, S., Charles, C., & Degré, A. (2011). Geostatistical interpolation of daily rainfall at catchment scale: the use of several variogram models in the Ourthe and Ambleve catchments, Belgium. *Hydrology and Earth System Sciences*, *15*(7), 2259–2274. https://doi.org/10.5194/hess-15-2259-2011

Manola, I., Steeneveld, G., Uijlenhoet, R., & Holtslag, A. A. M. (2019). Analysis of urban rainfall from hourly to seasonal scales using high-resolution radar observations in the Netherlands. *International Journal of Climatology*, 40(2), 822–840. https://doi.org/10.1002/joc.6241

Meira, M. A., Freitas, E. S., Coelho, V. H. R., Tomasella, J., Fowler, H. J., Ramos Filho, G. M., Silva, A. L., & Almeida, C. das N. (2022). Quality control procedures for subhourly rainfall data: An investigation in different spatio-temporal scales in Brazil. *Journal of Hydrology*, 613, 128358. https://doi.org/10.1016/j.jhydrol.2022.128358

Merino, A., García-Ortega, E., Navarro, A., Fernández-González, S., Tapiador, F. J., & Sánchez, J. L. (2021). Evaluation of gridded rain-gauge-based precipitation datasets: Impact of station density, spatial resolution, altitude gradient and climate. *International Journal of Climatology*, 41(5), 3027–3043. https://doi.org/10.1002/joc.7003

Morbidelli, R., García-Marín, A. P., Mamun, A. A., Atiqur, R. M., Ayuso-Muñoz, J. L., Taouti, M. B., Baranowski, P., Bellocchi, G., Sangüesa-Pool, C., Bennett, B., Oyunmunkh, B., Bonaccorso, B., Brocca, L., Caloiero, T., Caporali, E., Caracciolo, D., Casas-Castillo, M. C., G.Catalini, C., Chettih, M., & Kamal Chowdhury, A. F. M. (2020). The history of rainfall data time resolution in a wide variety of geographical areas. *Journal of Hydrology*, *590*, 125258. https://doi.org/10.1016/j.jhydrol.2020.125258

Murara, P. G., Acquaotta, F., Garzena, D., & Fratianni, S. (2019). Daily precipitation extremes and their variations in the Itajaí River Basin, Brazil. *Meteorology and Atmospheric Physics*, *131*(4), 1145–1156. https://doi.org/10.1007/s00703-018-0627-0

Oliveira, L. F. C. de, Fioreze, A. P., Medeiros, A. M. de M., & Silva, M. A. S. da. (2010). Comparison of metodologias of preenchimento de falhas de séries históricas de precipitação pluvial precipitation annual. *Revista Brasileira de Engenharia Agrícola e Ambiental*, *14*(11), 1186–1192. https://doi.org/10.1590/s1415-43662010001100008

Phosri, A. (2022). Effects of rainfall on human leptospirosis in Thailand: evidence of a multiprovince study using a distributed lag nonlinear model. *Stochastic Environmental Research and Risk Assessment*, *36*(12), 4119–4132. https://doi.org/10.1007/s00477-022-02250-x

Pritchard, D., Lewis, E., Blenkinsop, S., Velasquez, L. P., Whitford, A., & Fowler, H. J. (2023). An Observation-Based Dataset of Global Sub-Daily Precipitation Indices (GSDR-I). *Scientific Data*, 10(1). https://doi.org/10.1038/s41597-023-02238-4

Qi, Y., Martinaitis, S., Zhang, J., & Cocks, S. (2016). A Real-Time Automated Quality Control of Hourly Rain Gauge Data Based on Multiple Sensors in MRMS System. *Journal of Hydrometeorology*, *17*(6), 1675–1691. https://doi.org/10.1175/jhm-d-15-0188.1

Ribeiro, A. S., Almeida, M. C., Cox, M. G., Sousa, J. A., Martins, L., Loureiro, D., Brito, R., Silva, M., & Soares, A. C. (2021). Role of measurement uncertainty in the comparison of average areal rainfall methods. *Metrologia*, *58*(4), 044001. https://doi.org/10.1088/1681-7575/ac0d49

Robertson, D., Bennett, J. T., & Wang, Q. (2015). A strategy for quality controlling hourly rainfall observations and its impact on hourly streamflow simulations. In T. Weber, M. J. McPhee, & R. S. Anderssen (Eds.), *MODSIM2015, 21st International Congress on Modelling and Simulation* (pp. 490–496). Modelling and Simulation Society of Australia and New Zealand. https://doi.org/10.36334/modsim.2015.l4.robertson

Romero, P. E., González, M. H., Rolla, A. L., & Losano, F. (2020). Forecasting annual precipitation to improve the operation of dams in the Comahue region, Argentina. *Hydrological Sciences Journal*, 65(11), 1974–1983. https://doi.org/10.1080/02626667.2020.1786570

Rozante, J. R., Gutierrez, E. R., Fernandes, A. de A., & Vila, D. A. (2020). Performance of precipitation products obtained from combinations of satellite and surface observations. *International Journal of Remote Sensing*, 41(19), 7585–7604. https://doi.org/10.1080/01431161.2020.1763504

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

Sanchez-Lorenzo, A., Laux, P., Hendricks Franssen, H.-J. ., Calbó, J., Vogl, S., Georgoulias, A. K., & Quaas, J. (2012). Assessing large-scale weekly cycles in meteorological variables: a review. *Atmospheric Chemistry and Physics*, 12(13), 5755–5771. https://doi.org/10.5194/acp-12-5755-2012

Schmidt, L., Schäfer, D., Geller, J., Lünenschloss, P., Palm, B., Rinke, K., Rebmann, C., Rode, M., & Bumberger, J. (2023). System for automated Quality Control (SaQC) to enable traceable and reproducible data streams in environmental science. *Environmental Modelling & Software*, 169, 105809. https://doi.org/10.1016/j.envsoft.2023.105809

Sciuto, G., Bonaccorso, B., Cancelliere, A., & Rossi, G. (2009). Quality control of daily rainfall data with neural networks. *Journal of Hydrology*, *364*(1-2), 13–22. https://doi.org/10.1016/j.jhydrol.2008.10.008

Sieck, L. C., Burges, S. J., & Steiner, M. (2007). Challenges in obtaining reliable measurements of point rainfall. *Water Resources Research*, *43*(1). https://doi.org/10.1029/2005wr004519

Silva, V. B. S., Kousky, V. E., Shi, W., & Higgins, R. W. (2007). An Improved Gridded Historical Daily Precipitation Analysis for Brazil. *Journal of Hydrometeorology*, 8(4), 847–861. https://doi.org/10.1175/jhm598.1

Sokolovskaya, N., Vaughn, C., Jahangiri, H., Smith, V., Wadzuk, B., Ebrahimian, A., & Nyquist, J. (2023). Variability of urban drainage area delineation and runoff calculation with topographic resolution and rainfall volume. *Water Science & Technology*, *87*(6), 1349–1366. https://doi.org/10.2166/wst.2023.072

Souza, W. M., Azevedo, P. V. de, & Araújo, L. E. de. (2012). Classificação da Precipitação Diária e Impactos Decorrentes dos Desastres Associados às Chuvas na Cidade do Recife-PE. *Revista Brasileira de Geografia Física*, *5*(2), 250. https://doi.org/10.26848/rbgf.v5i2.232788

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., & Hsu, K. (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

Villarini, G., Mandapaka, P. V., Krajewski, W. F., & Moore, R. J. (2008). Rainfall and sampling uncertainties: A rain gauge perspective. *Journal of Geophysical Research*, *113*(D11). https://doi.org/10.1029/2007jd009214

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (3rd ed., Vol. 100). Academic Press.

World Meteorological Organization. (2017). *WMO Guidelines on the Calculation of Climate Normals (WMO-No. 1203)*. WMO.

Xavier, A. C., King, C. W., & Scanlon, B. R. (2016). Daily gridded meteorological variables in Brazil (1980-2013). *International Journal of Climatology*, *36*(6), 2644–2659. https://doi.org/10.1002/joc.4518

Xavier, A. C., King, C. W., & Scanlon, B. R. (2017). An update of Xavier, King and Scanlon (2016) daily precipitation gridded data set for the Brazil. *Simpósio Brasileiro de Sensoriamento Remoto,* 562–569.

Xavier, A. C., Scanlon, B. R., King, C. W., & Alves, A. T. (2022). New improved Brazilian daily weather gridded data (1961–2020). *International Journal of Climatology*, *42*(16), 8390–8404. https://doi.org/10.1002/joc.7731

Xu, H., Xu, C.-Y., Chen, H., Zhang, Z., & Li, L. (2013). Assessing the influence of rain gauge density and distribution on hydrological model performance in a humid region of China. *Journal of Hydrology*, *505*, 1–12. https://doi.org/10.1016/j.jhydrol.2013.09.004

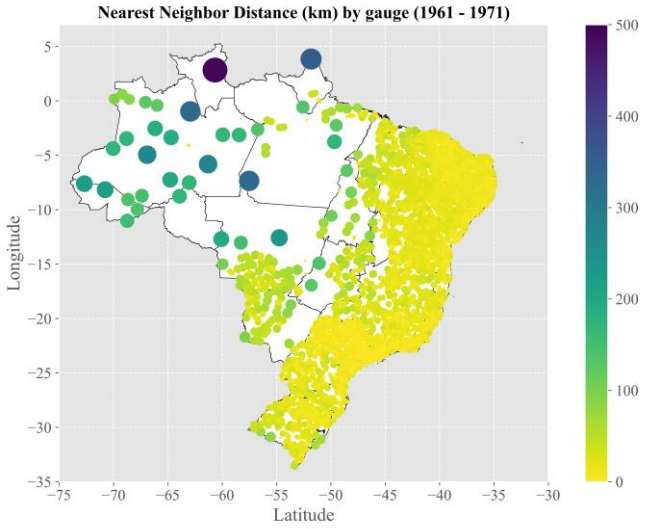
Yang, C.-C., Chandler, R. E., Isham, V., & Wheater, H. (2006). Quality control for daily observational rainfall series in the UK. *Water and Environment Journal*, *20*(3), 185–193. https://doi.org/10.1111/j.1747-6593.2006.00035.x

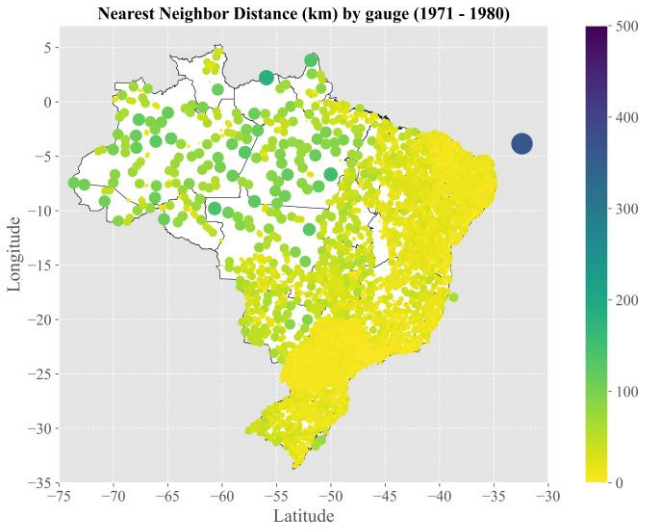
Yatagai, A., Maeda, M., Khadgarai, S., Masuda, M., & Xie, P. (2020). End of the Day (EOD) Judgment for Daily Rain-Gauge Data. *Atmosphere*, *11*(8), 772. https://doi.org/10.3390/atmos11080772

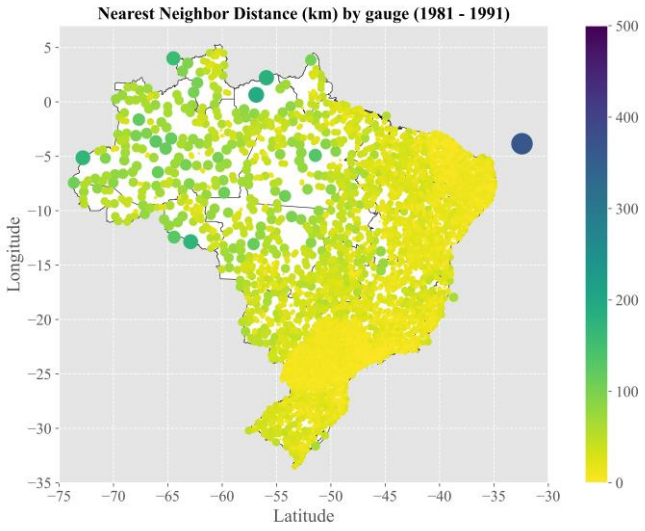
Zeng, Q., Chen, H., Xu, C.-Y., Jie, M.-X., Chen, J., & Liu, J. (2018). The effect of rain gauge density and distribution on runoff simulation using a lumped hydrological modelling approach. *Journal of Hydrology*, *563*, 106–122. https://doi.org/10.1016/j.jhydrol.2018.05.058

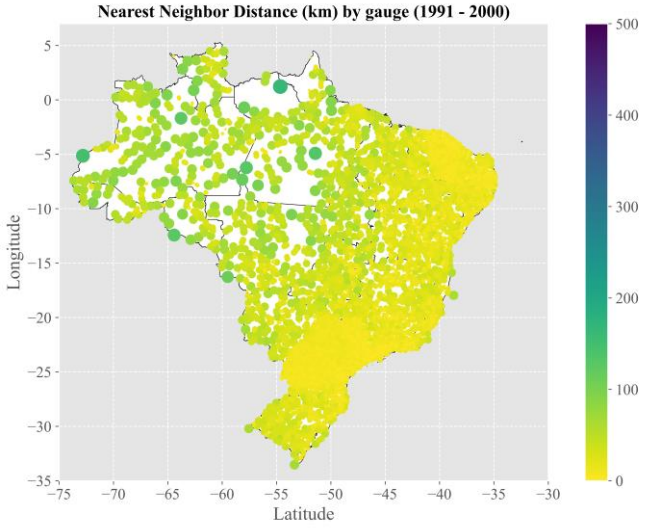
Appendix A

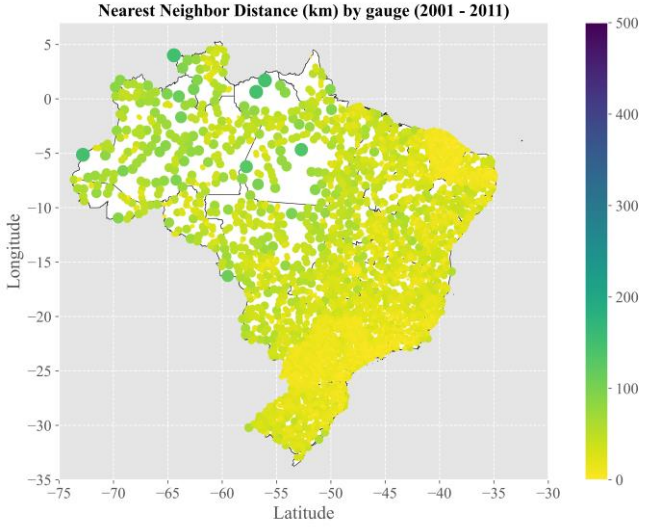
Nearest neighbor of each gauge by decade

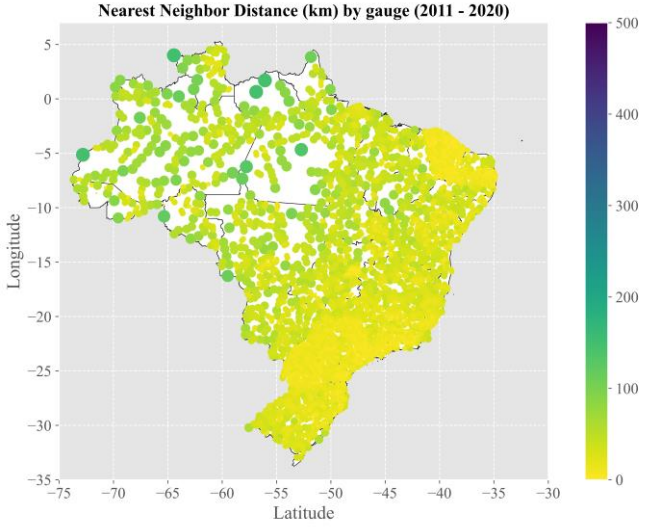






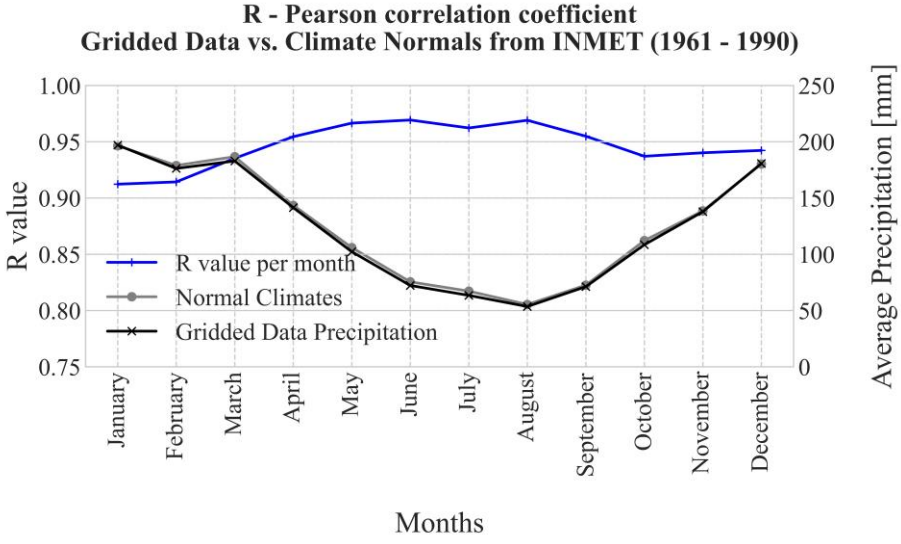

Nearest Neighbor Distance (km) by gauge (1991 - 2000)
Longitude
Latitude
5
0
−5
−10
−15
−20
−25
−30
−35
−75
−70
−65
−60
−55
−50
−45
−40
−35
−30
500
400
300
200
100
0

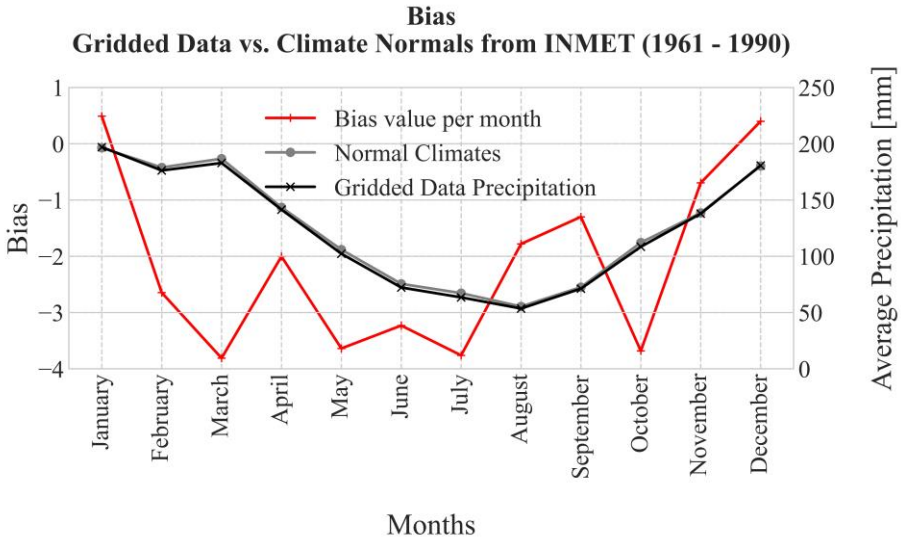

Nearest Neighbor Distance (km) by gauge (2001 - 2011)
Longitude
Latitude
5
0
−5
−10
−15
−20
−25
−30
−35
−75
−70
−65
−60
−55
−50
−45
−40
−35
−30
500
400
300
200
100
0

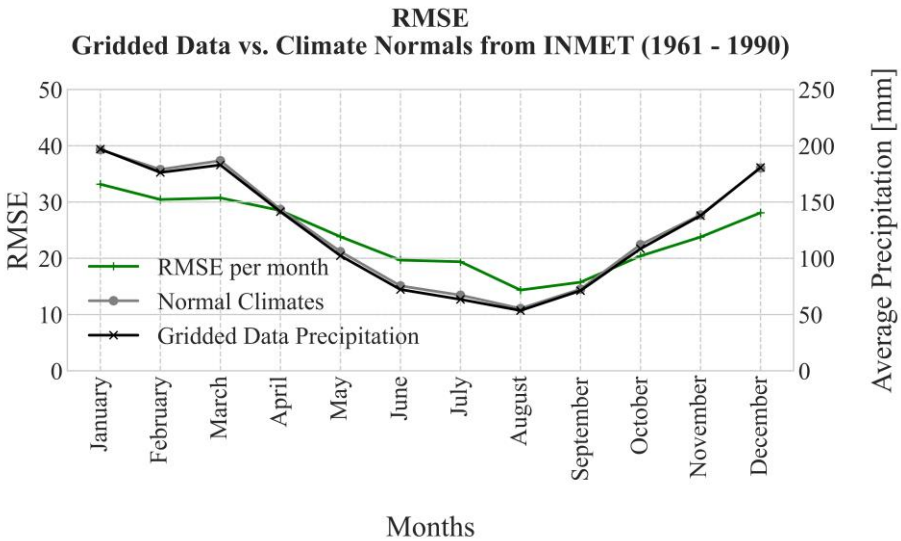

Nearest Neighbor Distance (km) by gauge (2011 - 2020)
Longitude
Latitude
5
0
−5
−10
−15
−20
−25
−30
−35
−75
−70
−65
−60
−55
−50
−45
−40
−35
−30
500
400
300
200
100
0

Appendix B

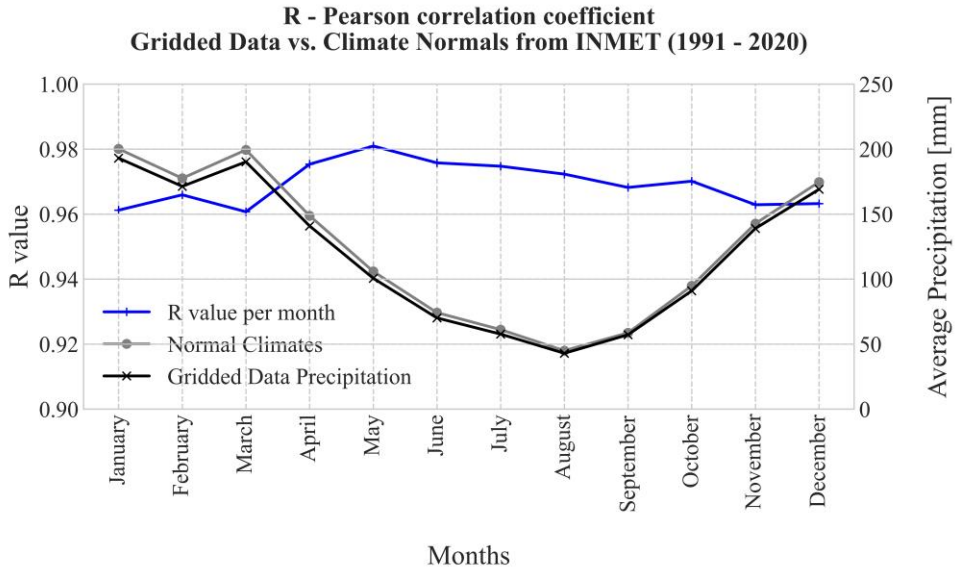
Comparison of climate norms and gridded data: Statistical results aggregated by month.
(1961 – 1990)

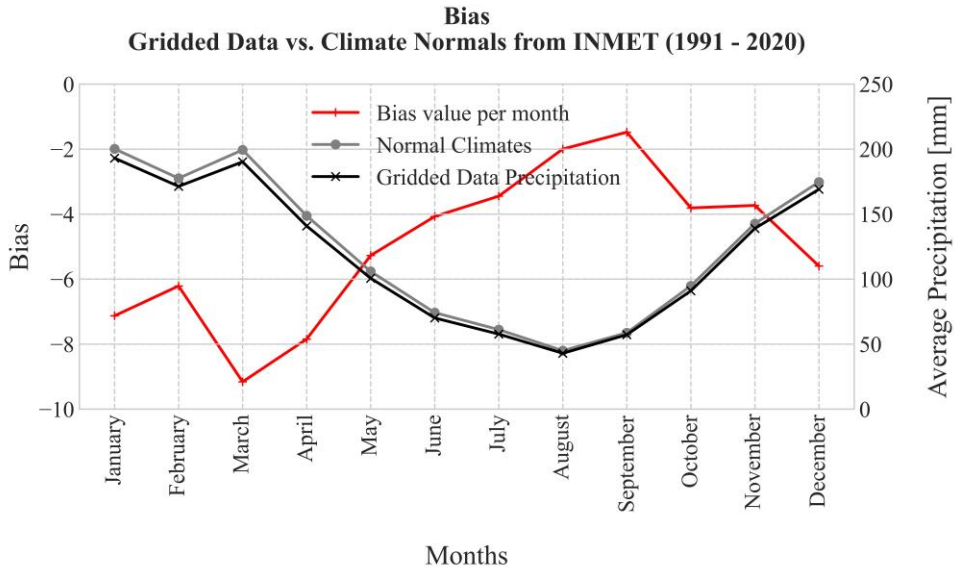


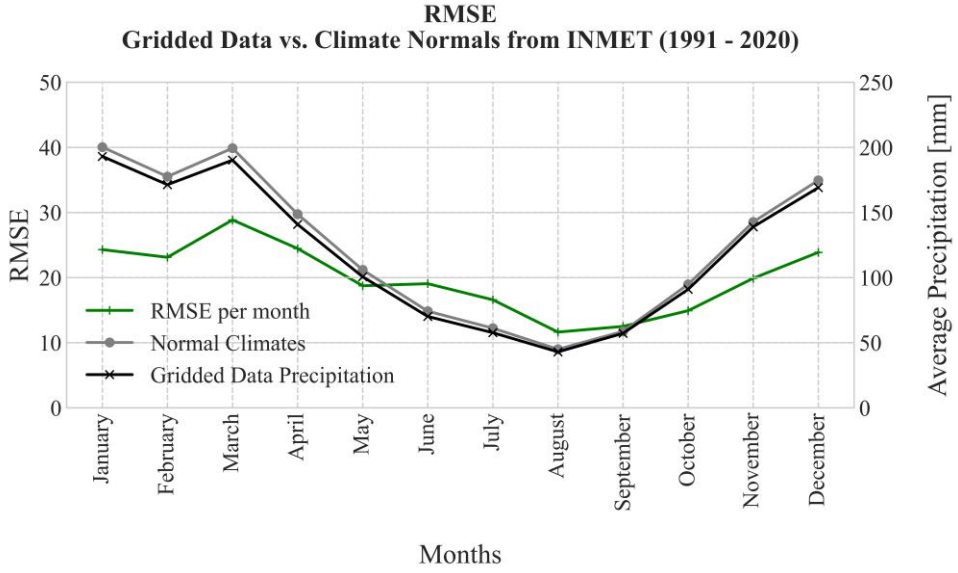




(1991 – 2020)

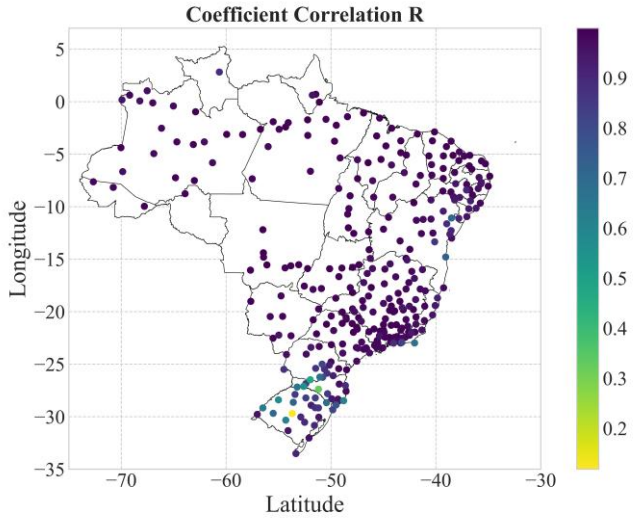

R - Pearson correlation coefficient
Gridded Data vs. Climate Normals from INMET (1991 - 2020)
R value
Average Precipitation [mm]
R value per month
Normal Climates
Gridded Data Precipitation
January
February
March
April
May
June
July
August
September
October
November
December
Months

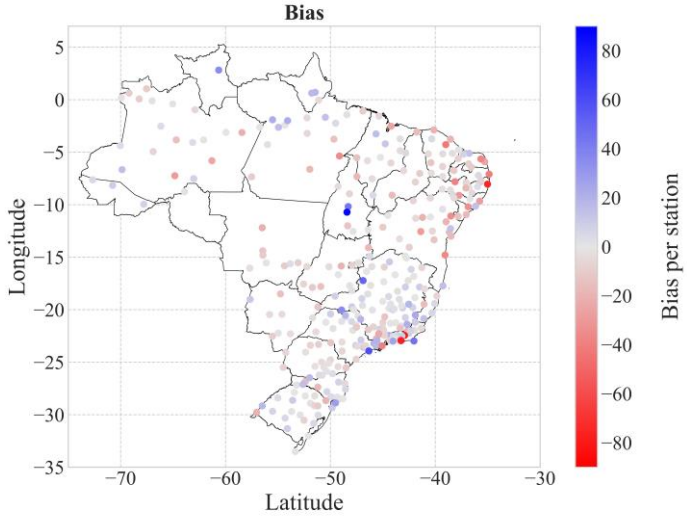

Bias
Gridded Data vs. Climate Normals from INMET (1991 - 2020)
Bias
Average Precipitation [mm]
Bias value per month
Normal Climates
Gridded Data Precipitation
January
February
March
April
May
June
July
August
September
October
November
December
Months

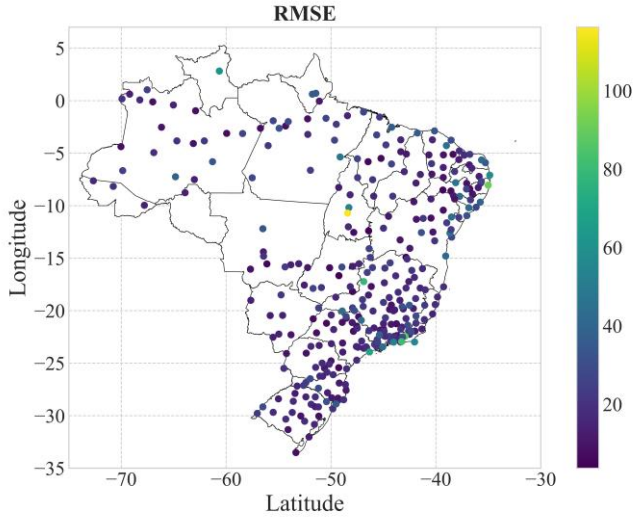

RMSE
Gridded Data vs. Climate Normals from INMET (1991 - 2020)
RMSE
Average Precipitation [mm]
RMSE per month
Normal Climates
Gridded Data Precipitation
January
February
March
April
May
June
July
August
September
October
November
December
Months

Appendix C

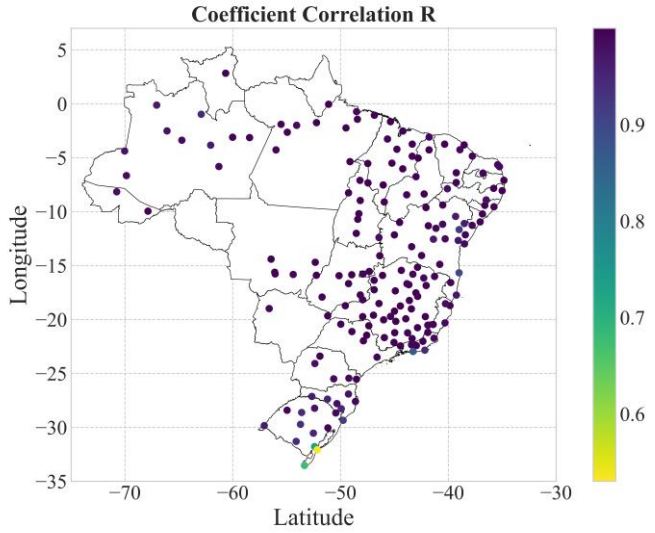
Comparison of Climate Normals and Gridded Data: geospatial results (1961 – 1990)

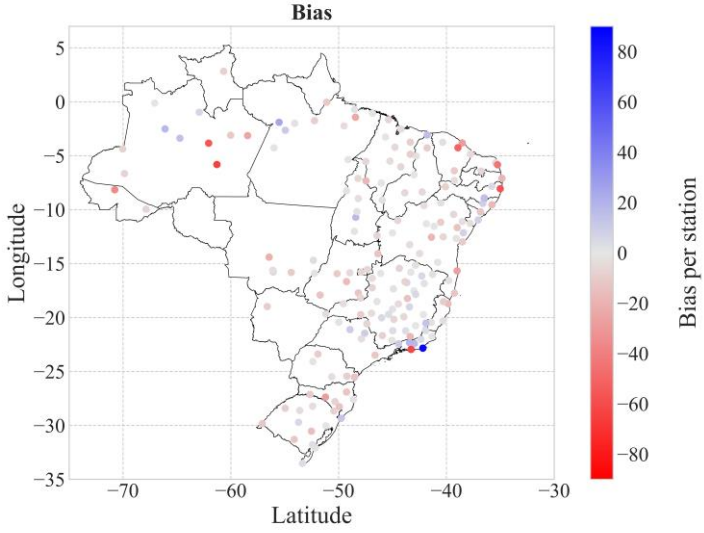


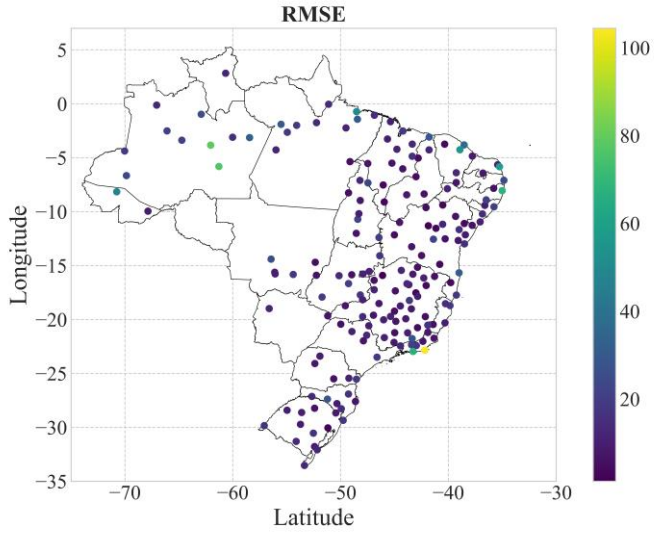




(1991 – 2020)


Coefficient Correlation R
Longitude
Latitude
5
0
−5
−10
−15
−20
−25
−30
−35
−70
−60
−50
−40
−30
0.9
0.8
0.7
0.6


Bias
Longitude
Latitude
Bias per station
80
60
40
20
0
−20
−40
−60
−80


RMSE
Longitude
Latitude
100
80
60
40
20