# DAILY RAINFALL DATA VALIDATION: IMERG, CHIRPS, AND GAUGES.

*Isabela Taici Lopes Gonçalves Horta*[1], *Luana Albertani Pampuch Bortolozo*[2], *Renata Pacheco Quevedo*[3], *Frederico Fábio Mauad*[4], *Adriano Rogério Bruno Tech*[5]

[1]São Carlos School of Engineering, University of São Paulo, Brazil; [2]Environmental Engineering Department, São Paulo State University, Brazil; [3]Remote Sensing Post-Graduate Course, Earth Observation and Geoinformatics Division, National Institute for Space Research, Brazil; [4]São Carlos School of Engineering, University of São Paulo, Brazil; and [5]Department of Basic Sciences, University of Sao Paulo, Brazil.

## ABSTRACT

Daily rainfall data are essential for improving society's resilience. Rain Gauges are the conventional instrument to measure rainfall. However, it is uncommon to have long time series without gaps. This fact forces researchers to look for other data sources, such as satellite data. This study aims to validate daily rainfall estimated by CHIRPS and IMERG based on gauge measurements between 2000 and 2018. The chosen study area is the Paraiba Valley and North Coast (PVNC) region, located in Sao Paulo state, Brazil. Statistical measures such as Pearson Coefficient, Mean Absolute Error, Mean Squared Error and Index of Agreement demonstrate how well the estimations fit the real data. Results indicated a low but significant Pearson correlation; a low Mean Absolute Error; a high Mean Squared Error, and a regular Index of Agreement. Satellite data is more trustworthy in the plain and inland portion of PVNC.

***Keywords*** — rainfall, Satellite-based rainfall, goodness-of-fit, gauge, statistical metrics.

## 1. INTRODUCTION

Daily rainfall data are essential for improving society's resilience. Important services like the planning for electricity generation, water supply, or disaster prevention are reliant on daily rain information.

Gauges are the most conventional instrument for measuring rain and the only method that represents the real precipitated water [1]. Unfortunately, the rainfall station network in Brazil is geographically and temporally scarce.

As a solution to this problem, it is possible to obtain rainfall estimation on a daily scale through satellites. Products such as the Integrated Multi-satellitE Retrievals for GPM (IMERG) and the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) have provided scientists and companies with valuable information.

Nonetheless, satellite data also have limitations related to temporal and spatial resolutions and relief influence [2]. Therefore, it is necessary to validate satellite data before relying on research on these data. The validation tests present the accuracy of estimated data (satellite data) compared to the real data (rainfall gauge data) and how much the satellite data underestimates or overestimates the precipitation amount. Thus, it is possible to know whether or not satellite data can replace rain gauge measurements.

Studies have shown a better correlation between satellite data and gauges data in coarser temporal and spatial scales. For example, the use of satellite data for the entire South American continent in monthly studies is well accepted. Regarding satellite products comparison, CHIRPS commonly presents better results for daily data, while IMERG products commonly sub-estimate the rain. However, it is recommended to validate data before relying on estimates for refined scales [3, 4, 5].

The validation of continuous and countable variables, such as rainfall, can be done through statistical tests using goodness-of-fit measures [6]. Validation studies have used Mean Absolute Error, Normalized Mean Square Error, Pearson's Coefficient, and Willmott Index, among other statistics [7, 8, 9].

Considering that, this study aims to answer the question: how well do CHIRPS and IMERG daily estimates fit the real daily rainfall values for the Paraiba Valley and North Coast (PVNC) region?

## 2. MATERIAL AND METHODS

### 2.1. Study Area

The PVNC region of São Paulo state comprises 39 municipalities covering 16,175 km², comprising 2,264,594 inhabitants. This region registers rainfall-related disaster events, such as landslides and floods, due to the mountainous relief and floodplains occupied by cities, resulting in economic and human losses.

This study used data from 21 rain gauges between 2000 and 2018 distributed along the PVNC region. Satellite data refer to the same geographic coordinates as the gauges. The dataset comprises gauges, IMERG, and CHIRPS rainfall information and will be analyzed individually, considering each point on the map. Figure 1 highlights the study area, the points from which the collected data, and the respective slope.

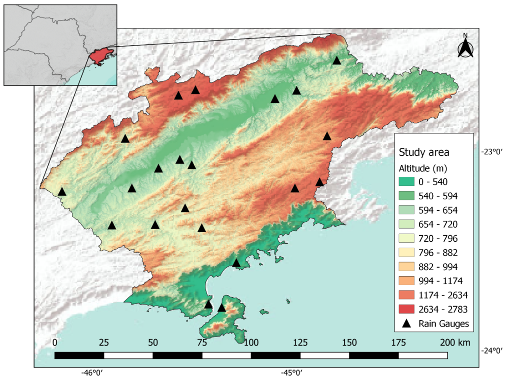

**Figure 1. PVNC altitude and gauges localization.**

Regarding the precipitation regime, the mean annual rainfall in the Paraiba's Valley portion is 1400 mm; 70% of this amount is concentrated in the spring-summer semester. The rainfall regime is driven mainly by extratropical atmospheric systems, frontal systems, and Mountain ranges. On the North Coast portion, the rainfall is higher, exceeding 2000 mm per year, due to the influences from the sea, the relief, and the humidity of the Atlantic Forest [10, 11].

### 2.2. Rain Gauge Stations

Automatic, telemetric, and conventional gauges managed by the Brazilian National Water Agency (ANA) provided the data for this study. After collecting the data, ANA performs a pre-analysis to ensure that the rain gauge data is consistent. This evaluation was carried out in three steps: seasonal filter, identification of graphical distortions, and comparison with data from their dataset. The qualified data was denominated as "consisted," and the non-qualified data as "non-consisted" [12].

For this study, we collected the "consisted" data using the HydroBr package [13] in python language.

### 2.3. IMERG

The IMERG data collection is produced by the National Aeronautics and Space Administration (NASA) through the Global Precipitation Measurement (GPM). Currently, there are seven versions of the algorithm, the last one being released in August 2022. The data have global coverage, between 60º S and 60º N, and are available in a 30 minutes time scale and 0,1º spatial scale [14].

For this study, we used daily data from version 6, Late Run, collected through the Google Earth Engine (GEE) platform, considering the parameters "scale = 11132 meters", which corresponds to the original scale, and "reducer = mean", which corresponds to the mean value between the pixels in the chosen geographical location.

### 2.4. CHIRPS

CHIRPS products are available for almost the entire globe (50°S to 50°N), with high resolution (0.05°), in daily, pentad, and monthly scales. The rain information is a composite of data collected by the Thermal Infrared band, the Tropical Rainfall Measuring Mission Multi-satellite Precipitation Analysis, version 7, the Cold Cloud Duration, and selected surface stations. More than 11,000 rain gauge stations in Brazil are used to reduce the CHIRPS estimates bias [15].

We collected the daily CHIRPS data using the GEE platform, considering the parameters "scale = 5566", which corresponds to the original scale, and "reducer = mean", which corresponds to the mean value between the pixels in the geographical location chosen.

### 2.5. Methodology Flowchart

The methodology is summarized in the flowchart below.

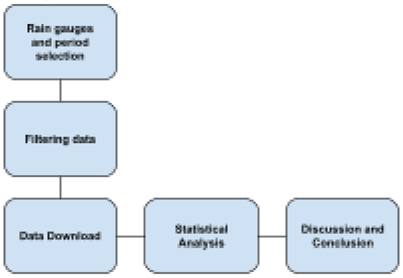

**Figure 2. Methodology steps.**

### 2.6. Statistical Analysis

Pearson's Coefficient (r), Mean squared error (MSE), Mean absolute error (MAE), and Index of agreement (d) indicate the goodness-of-fit between satellite data and gauges. Each formula is described below, where "y" is the rain gauge data, and "ŷ" is the satellite data.

#### *2.6.1. Pearson's coefficient*

Pearson's coefficient measures the linear association between the predicted and observed data. It ranges from -1 to +1, where negative results indicate an inverse relationship between the variables and positive results suggest a direct relationship. Values near 0 indicate a low relationship between the variables [16]. The formula is given by:

$$r = \frac{n\sum y\hat{y} - \sum y \sum \hat{y}}{\sqrt{n\sum y^2 - (\sum y)^2}\ \sqrt{n\sum \hat{y}^2 - (\sum \hat{y})^2}}$$

#### *2.6.2. Mean squared error*

The Mean squared error value ranges from zero to infinity and increases exponentially with an increase in error. The desired value is zero, as much closer to zero, the better the model's performance [16]. Its formula is given by:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

*2.6.3. Mean absolute error*

The Mean Absolute Error is the simplest loss measure. It is considered the arithmetic average of the absolute errors. Results point to just the extension of the errors but not the direction. The higher the value, the worse the accuracy of the model [16]. The formula is given by:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

*2.6.4. Index of Agreement (Willmott 1982)*

Developed by Willmott in 1982, the Index of Agreement (d) varies between 0 and 1, in which 0 represents no model agreement to the observed values, and 1 represents the best model agreement. An important characteristic is that d is sensitive to extreme values [17, 18, 19]. The formula is given by:

$$d = 1 - \frac{\sum_{i=1}^{n}(y - \hat{y})^2}{\sum_{i=1}^{n}(|\hat{y} - \bar{y}| + |y - \bar{y}|)^2}$$

## 3. RESULTS

Results are presented in this section. The maps on the left depict the statistics measures between IMERG and gauges, while the maps on the right depict the statistics measures between CHIRPS and gauges. The r values are presented first, then the MAE, posteriorly the MSE, and lastly, the d.

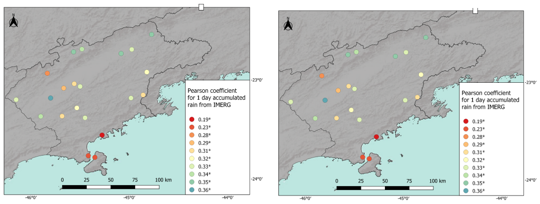

**Figure 3. Imerg and CHIRPS r.**

Both satellite data presented low but significant Pearson coefficients, which was expected because of the long time series. The values range between 0.18 and 0.37. In comparison to the IMERG, the CHIRPS presented higher Pearson's coefficients.

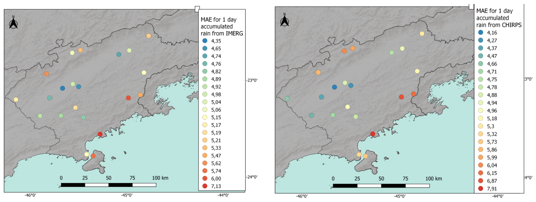

**Figure 4. Imerg and CHIRPS MAE.**

The values of Mean Abslolute Error ranged from 4.16 mm to 7,91 mm. In comparison to IMERG, CHIRPS presented more gauges with lower MAE values.

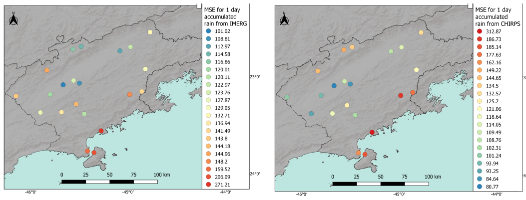

**Figure 5. Imerg and CHIRPS MSE.**

The Mean Squared Error values range from 80.77 to 312.87. In comparison to IMERG, CHIRPS presented a wider range.

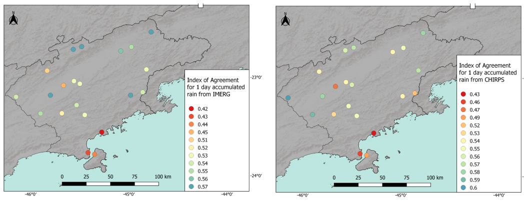

**Figure 5. Imerg and CHIRPS d.**

The Index of Agreement values ranges from 0.42 to 0.6. In comparison to IMERG, CHIRPS presented higher Index of Agreement values.

## 4. DISCUSSIONS

Daily rainfall is one of the hardest variables to be estimated. This paper presented a validation study using IMERG, CHIRPS, and rain gauge data. Regarding variation across the area, it is possible to notice better results in plain terrains inside the continent than in steep terrains on the coast.

Pearson Coefficient results demonstrated a low linear correlation but significant (p-value <= 0.05). Low Pearson's values were already expected since the time series is big. Similar results were found for the Sapucaí River watershed, north of the PVNC region [3]. MAE resulted in low values, which means there are no big differences between the predicted and the real value. MSE resulted in high values. Since some discrepancies between satellite and gauge data occur, this statistical measure increases exponentially. Most of the Index of Agreement values were higher than 0.5, which means a good agreement between the model and the real data.

## 5. CONCLUSIONS

Given the results across the study area, it is recommended to rely on some other source than daily satellite data for uses in the PVNC mountainous and littoral portion. However, these products can be used as complementary data. Moreover, it is more trustworthy in the plain and inland portion of PVNC.

It is shown that CHIRPS and IMERG predictions are similar, despite CHIRPS having more accurate daily rainfall values slightly. Better results should appear as long as satellite data is used on a coarser scale, such as the bibliography cites it.

It means it is important to install and maintain rain gauges for the superficial rain measures, despite their financial cost. Or increase the algorithm to correct the satellites' biases.

For future studies, it is suggested to delimitate groups of data according to climate seasons and calculate the goodness-of-fit measures to analyze the variation of the predictions in each season.

## 8. REFERENCES

[1] 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. Rev. Geophys. 56, 79–107. https://doi.org/10.1002/2017RG000574

[2] Sharifi, E., Brocca, L., 2022. Monitoring precipitation from space: progress, challenges, and opportunities, in: Precipitation Science. Elsevier, pp. 239–255. https://doi.org/10.1016/B978-0-12-822973-6.00021-4

[3] Silva, C.B., Silva, M.E.S., Ambrizzi, T., Tommaselli, J.T.G., Patucci, N.N., Mataveli, G.A.V., Lima, B.S., Correa, W.C., 2019. PRECIPITAÇÃO NA AMÉRICA DO SUL – DADOS OBTIDOS POR ESTAÇÕES METEOROLÓGICAS AUTOMÁTICAS E POR SISTEMAS ORBITAIS. ABClima 25. https://doi.org/10.5380/abclima.v25i0.58813

[4] Teodoro, T.A., Passos, R.B. dos, Silva, B.A., Silva, B.C. da, 2020. Análise das Estimativas da Precipitação Diária do Produto GPM-IMERG na Bacia Hidrográfica do Rio Sapucaí, Região Sudeste do Brasil. Anu. Inst. Geociênc. 43. https://doi.org/10.11137/2020_2_449_459

[5] Paredes-Trejo, F.J., Barbosa, H.A., Lakshmi Kumar, T.V., 2017. Validating CHIRPS-based satellite precipitation estimates in Northeast Brazil. Journal of Arid Environments 139, 26–40. https://doi.org/10.1016/j.jaridenv.2016.12.009

[6] Biecek, P., Burzykowski, T., 2021. Explanatory model analysis: explore, explain, and examine predictive models, 1st ed, Chapman & Hall/CRC Data Science Series. CRC Press, Boca Raton.

[7] Pirmoradian, R., Hashemi, H., Fayne, J., 2022. Performance evaluation of IMERG and TMPA daily precipitation products over CONUS (2000–2019). Atmospheric Research 279, 106389. https://doi.org/10.1016/j.atmosres.2022.106389

[8] López-Bermeo, C., Montoya, R.D., Caro-Lopera, F.J., Díaz-García, J.A., 2022. Validation of the accuracy of the CHIRPS precipitation dataset at representing climate variability in a tropical mountainous region of South America. Physics and Chemistry of the Earth, Parts A/B/C 127, 103184. https://doi.org/10.1016/j.pce.2022.103184

[9] Gentilucci, M., Bufalini, M., D’Aprile, F., Materazzi, M., Pambianchi, G., 2021. Comparison of Data from Rain Gauges and the IMERG Product to Analyse Precipitation in Mountain Areas of Central Italy. IJGI 10, 795. https://doi.org/10.3390/ijgi10120795.

[10] SILVA, Rodrigo César da; FISH, Gilberto. CENÁRIOS HIDROCLIMÁTICOS FUTUROS (2011-2040) PARA A REPRESA DE PARAIBUNA, SP, BRASIL:: subsídios para a transposição entre bacias hidrográficas. Revista Geociências, São Paulo, v. 38, n. 2, p. 587-597, 2019. Disponível em: https://www.periodicos.rc.biblioteca.unesp.br/index.php/geociencias/article/view/12944. Acesso em: 25 out. 2022.

[11] Milanesi, Marcos Alexandre Identificação de Unidades Climáticas na Ilha de São Sebastião (SP) / Marcos Alexandre Milanesi ; orientador Emerson Galvani. - São Paulo, 2016. 305 f.

[12] AGÊNCIA NACIONAL DE ÁGUAS. Manual de operação da sala de situação da ANA e para apoio aos estados. Setembro de 2013. Disponível em <https://progestao.ana.gov.br/progestao-1/certificacao/documentos-apoio-certificacao/modelo_manual-de-operacao-da-sala-de-situacao_ana.docx> Acesso em 4 de outubro de 2022.

[13] Carvalho, W.M.D., 2020. HydroBr: A Python package to work with Brazilian hydrometeorological time series. https://doi.org/10.5281/ZENODO.3931027

[14] Huffman, G.J., Bolvin, D.T., Braithwaite, D., Hsu, K.-L., Joyce, R.J., Kidd, C., Nelkin, E.J., Sorooshian, S., Stocker, E.F., Tan, J., Wolff, D.B., Xie, P., 2020. Integrated Multi-satellite Retrievals for the Global Precipitation Measurement (GPM) Mission (IMERG), in: Levizzani, V., Kidd, C., Kirschbaum, D.B., Kummerow, C.D., Nakamura, K., Turk, F.J. (Eds.), Satellite Precipitation Measurement, Advances in Global Change Research. Springer International Publishing, Cham, pp. 343–353. https://doi.org/10.1007/978-3-030-24568-9_19

[15] Funk, C., Peterson, P., Landsfeld, M. et al. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. Sci Data 2, 150066 (2015). https://doi.org/10.1038/sdata.2015.66

[16] Kirch, W. (Ed.), 2008. Pearson’s Correlation Coefficient, in: Encyclopedia of Public Health. Springer Netherlands, Dordrecht, pp. 1090–1091. https://doi.org/10.1007/978-1-4020-5614-7_2569

[17] Padhma Sahithya Muniraj, 2021. End-to-End Introduction to Evaluating Regression Models. URL https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/#h2_7 (accessed 10.05.22).

[18] López-Bermeo, C., Montoya, R.D., Caro-Lopera, F.J., Díaz-García, J.A., 2022. Validation of the accuracy of the CHIRPS precipitation dataset at representing climate variability in a tropical mountainous region of South America. Physics and Chemistry of the Earth, Parts A/B/C 127, 103184. https://doi.org/10.1016/j.pce.2022.103184

[19] Willmott, C.J., 1981. ON THE VALIDATION OF MODELS. Physical Geography 2, 184–194. https://doi.org/10.1080/02723646.1981.10642213