University of São Paulo
"Luiz de Queiroz" College of Agriculture

Regionalization of hydrological variables for the Paraná state, Brazil

Jéssica Garcia Nascimento

Thesis presented to obtain the degree of Doctor in Science.
Area: Agricultural Systems Engineering

Piracicaba
2021

Jéssica Garcia Nascimento
Agricultural and Environmental Engineer

Regionalization of hydrological variables for the Paraná state, Brazil

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Advisor:
Prof. Dr. **SERGIO NASCIMENTO DUARTE**

Thesis presented to obtain the degree of Doctor in Science.
Area: Agricultural Systems Engineering

Piracicaba
2021

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Nascimento, Jessica Garcia

Regionalization of hydrological variables for the Paraná State, Brazil / Jéssica Garcia Nascimento. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2021.

94 p.

Tese (Doutorado) - - USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. Modelos hidrológicos 2. Regressões lineares múltiplas 3. Machine learning 4. Balanço hídrico I. Título

With love to my parents, Valéria and Alex,
my brothers Edson, Bruno, and Hugo,
and my Lalis.

ACKNOWLEDGMENT

God.

My family for always loving and supporting me unconditionally. Thanks especially to my parents, Valéria and Alex, my brothers Edson, Bruno, and Hugo, and the women of my life: Maria Eulália, Gilmara, Patrícia, e Leandra.

The "Luiz de Queiroz" College of Agriculture – University of São Paulo (ESALQ/USP) and the Department of Agricultural System Engineering for the opportunity and educational excellence.

The Coordination for the Improvement of Higher Educational Personnel (CAPES) for the scholarship to conduct this research in Brazil and the United States, through the Program for Institutional Internationalization PrInt program.

The University of Nebraska-Lincoln and the Daugherty Water for Food Global Institute, for the partnership and scholarship to conduct this research.

My advisor Dr. Sergio N. Duarte for encouraging me throughout graduate school, beginning in the Master's course. For all exchange of knowledge, patience, and friendship.

My supervisor Dr. Christopher M. U. Neale and his team, for supporting this research, having and supporting me as a visiting student, and for allowing me to be part of their projects at the Daugherty Water for Food Global Institute, University of Nebraska.

My friends Adriano Pacheco, Hugo Ricardo, Luísa Lelis, Thaís Charles, Pedro Sampaio (Rony), for all their support, friendship, and good moments.

My friends Nelida Quiñonez and Daniel Althoff, for the friendship, talks, and knowledge exchanges.

All my friends, especially Helizani Bazame, Camila Netto, Larissa Moura, Julia Fontes, Francyne Garcia, Cleverson Freitas, Abel Torres, Isabella Martins, Carla Chiles, Thuane Barbosa, Fernanda Granja, Maurício Leite, Joana Peloia, Rebeca Barros for their constant support.

My friends from Lincoln, Nebraska, especially Naisi Dave, Ivo Zution, Wellington Araújo, Camila Braga, Atanaele Bernardo, Wilman Iglesias, Gabirel Shimada and Aline Leite.

My friends from Biosystems Engineering Department (LEB), especially Fernanda Lamede, Débora Pantojo, Danielle Morais, Rubmara Oliveira, Tarcio Rocha, Luiz Sobenko, Rodolfo Armando, Diego Souza, Wagner Bombardelli, Nathalia Lopes, Alex Nunes, Ailson Maciel, Luciano Sobral, Elisabeth Carnevskis, Otávio Neto, Júnior Campos, Ana Sátiro, Ângelo Azevedo, Brenda Pinheiro, Juliana Martins, Carlos and Timóteo Barros.

Employees from LEB, especially Ângela Silva, Antônio Agostinho, Beatriz Novaes, Davilmar Colevatti, Gilmar Grigolon, and Paula Bonassa.

Everyone who supported me in any way throughout my Ph.D. period.

"Se as coisas são inatingíveis... ora!
Não é motivo para não querê-las...
Que tristes os caminhos, se não fora
A presença distante das estrelas!"
Mário Quintana

SUMMARY

RESUMO

Regionalização de variáveis hidrológicas para o Estado do Paraná, Brasil

O conhecimento das vazões e descargas dos rios é essencial para a gestão dos recursos hídricos, uma vez que estes parâmetros representam a disponibilidade de água nas bacias hidrográficas. Em geral, a vazão média de longo período (Qm), e as vazões excedidas ou igualadas em 90% e 95% do tempo (Q90 e Q95, respectivamente), as quais representam vazões mínimas, e a descarga anual (Q) são frequentemente usadas no gestão de recursos hídricos. As informações sobre esses parâmetros podem ser consideradas um desafio, principalmente em países em desenvolvimento, onde o monitoramento por estações fluviométricas é limitado em termos de densidade e frequência de observações. Assim, a predição de variáveis hidrológicas em bacias hidrográficas não monitoradas pode ser realizada por meio de modelos que permitem a relação das variáveis de interesse com variáveis descritoras, como por exemplo por Regressões Lineares Múltiplas (MLR), que consiste no método mais antigo e amplamente utilizado para problemas de regionalização, e novas técnicas como abordagens de aprendizado de máquina, por exemplo, o *Random Forest* (RF). Adicionalmente, o balanço hídrico pode ser usado como modelo para estimar a Q dos rios em bacias hidrográficas. Nesse contexto, os produtos de sensoriamento remoto contituem promissoras fontes de dados de precipitação (PPT) e evapotranspiração (ET), com alta resolução espacial e temporal, que podem ser usados em modelos hidrológicos, melhorando seu desempenho. Em um cenário geral, este estudo teve como objetivo analisar o desempenho de produtos integrados Multi-satellitE Retrievals for GPM (IMERG) para estimar a PPT sobre o Estado do Paraná, Brasil. Em seguida, os produtos mensais do IMERG foram utilizados juntamente com descritores morfológicos de bacias hidrográficas, para construir modelos hidrológicos para a predição de vazões (Q90, Q95 e Qm) em 81 bacias hidrográficas no Estado do Paraná, Brasil. Por fim, os produtos mensais IMERG e de ET do algorítimo ALEXI (Atmosphere-Land Exchange Inverse) foram utilizados para a predição de Q em 28 bacias hidrográficas do Estado do Paraná, por meio da equação do balanço hídrico. Os modelos apresentaram boa performance para a previsão de variáveis hidrológicas, o que comprovou a importância do sensoriamento remoto e dos modelos hidrológicos para a gestão dos recursos hídricos.

Palavras-chave: Modelos hidrológicos, Regressões lineares múltiplas, *Machine learning* Balanço hídrico

ABSTRACT

Regionalization of hydrological variables for the Paraná state, Brazil

The knowledge of the flows and discharge in the rivers is essential for water resources management since it represents the availability of water in the watersheds. Generally, the long-term average flows (Qm), flows exceeded or equaled in 90% and 95% of the time (Q90 and Q95, respectively), which represents minimum flows, and annual discharge (Q) are frequently used in the water resources management. The information about those parameters can be considered a challenge, especially in developing countries where monitoring by gauges is limited in terms of density and frequency of observations. Thus, hydrological models can be applied to predicted flows in unmonitored watersheds, as the Multiple Linear Regressions (MLR) method, the oldest and widely used method for regionalization problems, and new techniques as machine learning approaches, for example, the Random Forest (RF). Additionally, the water balance can be used to estimate the annual discharge of rivers in watersheds. In this context, remote sensing products offer precipitation (PPT) and evapotranspiration (ET) products with great spatial and temporal coverage, which can be used in hydrological models, improving its performance. In a general scenario, this study aimed to analyze the performance of Integrated Multi-satellitE Retrievals for GPM (IMERG) products to estimate the PPT over Paraná state, Brazil. Subsequently, the IMERG monthly products were used with watershed morphological descriptors to build hydrological models for predicting the flows (Q90, Q95, and Qm) in 81 watersheds in Paraná state, Brazil. Lastly, we predicted the Q using the IMERG monthly products and the ALEXI (Atmosphere-Land Exchange Inverse) ET over 28 watersheds in Paraná state, by water balance equation. The models performed very well to predict the hydrological variables, which demonstrated the importance of remote sensing and hydrological models in water resources management.

Keywords: Hydrological models, Multiple regressions, Machine Learning, Water balance

1. GENERAL INTRODUCTION

Over the past few decades, climate change and a growing population have concerned experts regarding water availability for its multiple uses. The water management must be efficient in minimizing possible conflicts, balancing the demands for varied human activities, and avoiding the rise of environmental degradation.

The availability of water in a watershed is given by the knowledge of the flow in the rivers. In general, reference flows are adopted in water resources management considering minimum flow aiming at low risk for the rivers (Harris et al., 2000). The minimum flows can be extracted from flow duration curves, which the frequency at which a specific flow value flow is equaled or exceeded, and is an important tool used to determine the reference values of the amount of water in a watershed. Two reference values frequently used are Q90 and Q95 ($m^3 s^{-1}$), which represent flows exceeded or equaled in 90% and 95% of the time, respectively, and represent minimum flows (Pereira et al., 2016).

The long-term average flow (Qm, $m^3 s^{-1}$) is another essential hydrological parameter that informs about the energy potential of the watershed and represents the highest flow that can be regularized. Qm is frequently used for calculating the regularization volume when projecting a reservoir and is obtained by averaging monthly values over a year, and then by averaging all the year's values.

Compilation of the information about the flow in watersheds can be a challenge in the water resources management in some countries, where the number of gauges, when compared to the areas to be monitored, is small (Swain and Patra, 2017). Especially, in developing countries in South America, monitoring by gauges is limited in terms of infrastructure, maintenance, density, and frequency of observations (Hobouchian et al. 2017, Salio et al. 2015).

However, hydrological models allow optimization of the information obtained in monitored sites for the prediction in unmonitored watersheds, particularly through obtaining flow estimates by transferring historical data between watersheds, using models, and regionalization techniques. Multiple linear regression (MLR) is one of the oldest and widely used methods for predicting hydrological variables, but new techniques like machine learning have shown interesting results in hydrological studies. An example is the Random Forest (RF) technique, which can work with big data and solve various hydrological problems. The annual river discharge (Q, mm $year^{-1}$) also is an important hydrological parameter used as a tool in water resources management. In addition to the Q90, Q95, and Qm, annual river discharge can be obtained alternatively by hydrological models, which uses descriptors to relate the volume

of water in the watersheds. A simple model consists of the water balance that computes the amount of water, based on the subtraction of the outflow by the evapotranspiration (ET), over watersheds from the inflow by the precipitation (PPT).

Independently of the approach to flow or discharge estimates, the most important parameter considered as a descriptor in models is the PPT because they derive ultimately from this event. (Poof et al., 1997). The remote sensing products from satellites represent a promising alternative to the conventional rain gauges due to its PPT data are free and available spatially and temporally continuous over large areas (Liu et al., 2015). The latest version of the Integrated Multi-satellitE Retrievals for GPM (IMERG) algorithm (Version 6), combines the reanalysis of precipitation estimated by satellites between 2000-2014 by TMPA and in the subsequent period by GPM, totaling 19 years of information so far. Its products have a spatial resolution of 0.1° and a 30-minutes temporal resolution (Huffman et al., 2019). In this way, hydrological models using its products can be carried out with greater accuracy.

Following PPT, the ET represents the second most important parameter in the water balance in watersheds and is an important input in hydrological models. The measurement of ET is, usually, not by directly methods, and several methods can be used to obtain this component. Regarding the approaches that allow mapping the ET spatially, the Two-Source Energy Balance (TSEB) allows the ET estimative using remote sensing products (Norman et al. 1995; Kustas and Norman, 2000; Li et al., 2005). The ALEXI (Atmosphere-Land Exchange Inverse) model is based on the TSEB (Anderson et al., 2007a; Anderson et al., 2007b) and was designed to minimize sensitivity to errors in inputs land-surface temperature and air temperature boundary conditions in ET estimation. Thus, the ALEXI model considers a spatially and physically realistic representation of land-atmosphere exchange over vegetation and cover conditions, using high temporal resolution products from geostationary satellites (Anderson et al., 2011).

Recently, the application of remote sensing in Hydrology has provided the development of models with good performance for monitoring and estimating variables of interest. Additionally, these tools allow the spatialization of important variables, such as PPT and ET, which are extremely relevant for the improvement of hydrological models for estimating essential information in the management of water resources.

In this context, this study aims to analyze the performance of hydrological models using remote sensing products for estimating the flows (Q90, Q95, and Qm) and Q over watersheds in Paraná state, Brazil. For that purpose, we first analyzed the performance of the daily and monthly IMERG products to estimate the PPT over the Paraná state. Afterwards, we

used its monthly products and watersheds descriptors to build MLR and machine learning (RF) models to predict the Q90, Q95, and Qm in 81 watersheds in Paraná. Lastly, we used the monthly IMERG PPT and ALEXI ET to estimate the annual discharge in watersheds in Paraná state using the water balance equation.

References

Anderson, M. C., Kustas, W. P., Norman, J. M., Hain, C. R., Mecikalski, J. R., Schultz, L., Gonzalez-Dugo, M. P., Cammalleri, C., d'Urso, G., Pimstein, A., and Gao, F. (2011). Mapping daily evapotranspiration at field to continental scales using geostationary and polar orbiting satellite imagery. 15, 223–239. https://doi.org/10.5194/hess-15-223-2011

Anderson, M.C., Norman, J. M., Mecikalski, J.R., Otkin, J.A., Kustas, W.P., 2007a. A climatological study of evapotranspiration and moisture stress across the continental United States based on thermal remote sensing: 1. Model formulation. Journal of Geophysical Research, 112, D10117. https://doi.org/10.1029/2006JD007506

Anderson, M.C., Norman, J.M., Mecikalski, J.R., Otkin, J.P., Kustas, W.P., 2007b. A climatological study of evapotranspiration and moisture stress across the continental U.S. based on thermal remote sensing: II. Surface moisture climatology. Journal of Geophysical Research, 112, D11112. https://doi.org/10.1029/2006JD007507

Harris, N. M., Gurnell, A. M., Hannah, D. M., & Petts, G. E. (2000). Classification of river regimes: a context for hydroecology. Hydrological Processes, 14 (16-17), 2831-2848. https://doi.org/10.1002/1099-1085

Hobouchian, M. P., Salio, P., Skabar, Y. G., Vila, D., Garreaud, R. (2017). Assessment of satellite precipitation estimates over the slopes of the subtropical Andes. Atmospheric Research, *190*, 43-54. https://doi.org/10.1016/j.atmosres.2017.02.006

Huffman, G.J., Bolvin, D.T., Nelkin, E.J., Tan J. 2019. Integrated Multi-satellitE Retrievals for GPM (IMERG) Technical Documentation, https://docserver.gesdisc.eosdis.nasa.gov/public/project/GPM/IMERG_doc.06.pdf.

Kustas, W. P.; Norman, J. M. (2000). A two-source energy balance approach using directional radiometric temperature observations for sparse canopy covered surfaces. Agronomy Journal, 92(5), 847-854. https://doi.org/10.2134/agronj2000.925847x

Li, F., Kustas, W., Prueger, J.H., Neale, C.M., Jackson, T.J., 2005. Utility of remote sensing–based two-source energy balance model under low-and high-vegetation cover conditions. Journal of Hydrometeorology, 6 (6), 878-891. https://doi.org/10.1175/JHM464.1

Liu, J., Duan, Z., Jiang, J., Zhu, A. 2015. Evaluation of three satellite precipitation products TRMM 3B42, CMORPH, and PERSIANN over a subtropical watershed in China. Advances in Meteorology, 2015. https://doi.org/10.1155/2015/151239

Norman, J.M.; Kustas, W.P.; Humes, K.S. 1995. Source approach for estimating soil and vegetation energy fluxes in observations of directional radiometric surface temperature. Agricultural and Forest Meteorology, 77 (3-4), 263-293. https://doi.org/10.1016/0168-1923(95)02265-Y

Pereira, D. D. R., Martinez, M. A., da Silva, D. D., Pruski, F. F. (2016). Hydrological simulation in a basin of typical tropical climate and soil using the SWAT Model Part II: Simulation of hydrological variables and soil use scenarios. Journal of Hydrology: Regional Studies, 5, 149-163. https://doi.org/10.1016/j.ejrh.2015.11.008

Poff, N. L., Allan, J. D., Bain, M. B., Karr, J. R., Prestegaard, K. L., Richter, B. D., Stromberg, S., Stromberg, J. C. (1997). The natural flow regime. BioScience, 47(11), 769-784. https://doi.org/10.2307/1313099

Salio, P., Hobouchian, M. P., Skabar, Y. G., Vila, D. (2015). Evaluation of high-resolution satellite precipitation estimates over southern South America using a dense gauge network. Atmospheric Research, 163, 146-161. https://doi.org/10.1016/j.atmosres.2014.11.017

Swain, J. B., Patra, K. C. (2017). Streamflow estimation in ungauged catchments using regionalization techniques. Journal of Hydrology, 554, 420-433. https://doi.org/10.1016/j.jhydrol.2017.08.054

2. EVALUATING THE LATEST IMERG PRODUCTS OVER THE PARANÁ STATE, BRAZIL

ABSTRACT

The lack of measurement of precipitation in large areas using fine-resolution data is a limitation in water management, particularly in developing countries. However, Version 6 of the Integrated Multi-satellitE Retrievals for GPM (IMERG) has provided a new source of precipitation information with high spatial and temporal resolution. In this study, the performance of the GPM products (Final run) in the state of Paraná, located in the southern region of Brazil, from June 2000 to December 2018 was evaluated. The daily and monthly products of IMERG were compared to the gauges data spatial distributed across the study area. Quantitative and qualitative metrics were used to analyze the performance of IMERG products to detect precipitation events and anomalies. In general, the products performed positively in the estimation of monthly rainfall events, both in volume and spatial distribution, and demonstrated limited performance for daily events and anomalies, mainly in mountainous regions (coast and southwest). This may be related to the orographic rainfall in these regions, associating the intensity of the rain, and the topography. IMERG products can be considered as a source of precipitation data, especially on a monthly scale. Product calibrations are suggested for use on a daily scale and for time-series analysis.

Keywords: Remote sensing; Satellite; GPM; Performance evaluation

2.1. Introduction

Precipitation plays a fundamental role in the hydrological cycle. It is considered the main water source input in the soil water balance and runoff and is used as an input in hydrological and climatological modeling. In the management of water resources, knowledge of the volume and intensity of precipitation is essential for the prediction of floods and droughts, the distribution of water for urban and industrial uses, and the planning of irrigation in agriculture and hydraulic infrastructure.

Precipitation can be measured by gauges, sensors onboard satellites, and radars (Shen and Xiong, 2016, Guo et al., 2016, Kucera et al., 2013). Precipitation gauges are fundamental instruments, and their observations are considered as a reference in many studies (Tapiador et al. 2012). However, to represent spatiotemporal variability of intensity and type of occurrence of precipitation, a dense measuring network is necessary with a long-period information, which unfortunately is not the reality in many regions of the world (Hou et al., 2014). In South

American countries, monitoring by gauges is limited in terms of infrastructure, maintenance, density, and frequency of observations (Hobouchian et al. 2017, Salio et al. 2015).

Regarding indirect methods, climate radars provide precipitation estimates with high spatial and temporal resolution but have limited accuracy in mountainous regions and cold climates (Falck et al. 2018, Hou et al. 2014). On the other hand, satellite estimates of precipitation provide vast spatial and temporal coverage and are freely available. Over the last two decades, several satellite precipitation products have been developed, such as: Tropical Rainfall Measuring Mission (TRMM; Kummerow et al., 1998); Remotely Sensed Precipitation Estimation from Information using Artificial Neural Networks (PERSIANN; Sorooshian et al., 2000); Climate Prediction Center Morphing Method (CMORPH; Joyce et al., 2004); Global Satellite Mapping of Precipitation (GSMaP; Mega et al., 2014); Climate Hazards Group Infrared Precipitation with Stations (CHIRPS; Funk et al., 2015), and Multi-Source Weighted-Ensemble Precipitation (MSWEP; Beck et al., 2017).

The TRMM Multi-satellite Precipitation Analysis (TMPA) algorithm combines precipitation estimates from satellite systems with data measured on the Earth's surface to provide a calibrated final product and with the "best" satellite estimate (Huffman et al., 2007). Successor to TRMM, the Global Precipitation Measurement (GPM) was launched in 2014, on a joint mission between NASA (National Aeronautics and Space Administration) and JAXA (Japan Aerospace Exploration Agency) and offers products to this day. The GPM constellation consist of the first Dual-frequency phased array Precipitation Radar, and a GPM Microwave Imager, which represent the most advanced versions compared to the Precipitation Radar (PR) and the TRMM Microwave Imager (TMI), on board the TRMM satellite (Wang et a., 2018). Relevant improvements in the GPM products include an increase in latitudinal coverage (global coverage of 60° N/S) and the detection of heavy rain, light rain, and snow (Hou et al., 2014, Hobouchian et al., 2017). In the era of GPM, the Integrated Multi-satellite Retrievals for GPM (IMERG) algorithm operates with the objective of calibrating, uniting, and interpolating satellite precipitation estimates with data from gauges (Huffman et al., 2015).

The latest version of the IMERG algorithm (Version 6), made available to the public in October 2019, combines the reanalysis of precipitation estimated by satellites between 2000-2014 by TMPA and in the subsequent period by GPM, totaling 19 years of information so far. Its products have a spatial resolution of 0.1° and a 30-minutes temporal resolution (Huffman et al., 2019). In this way, trend analysis and analysis of extreme events can be carried out with greater accuracy. In addition, the performance of climatological and hydrological models can be improved with greater detail of recent precipitation information.

Currently, the performance of previous versions of IMERG in estimating precipitation has been analyzed in comparison to TRMM data and those obtained by gauges in a large number of studies (for example, El Kenawy et al., 2015; Melo et al., 2015; Hobouchian et al., 2017; Fang et al., 2019 and Gadelha et al., 2019) and allowed advances in the application of remote sensing in determining the volume and behavior of precipitation in several countries. However, when comparing the performance of IMERG (Version 5) concerning TMPA (Version 7), Liu (2016) found better performance of IMERG in estimating precipitation on a global scale. The better performance of IMERG's products also was observed by Rozante et al. (2018) concerning the TMPA in estimating precipitation in all regions of Brazil. Recently, Chen et al (2020) observed better performance of IMERG-Late Version 6 products compared to IMERG-Early, GSMaP-NRT, GSMaP-MVK, TMPA-RT, and PERSIANN-CCS products, on a global scale. Thus, studies evaluating the performance of IMERG Version 6 products are important and promising at regional scales.

Brazil is a country of continental dimensions, with different micro-climates and rainfall patterns throughout its territory. Among its states, Paraná is in the south of the country, in the Paraná River Basin region, the most important socio-economically hydrologic region in Brazil (Zandonadi et al., 2016). This basin has the largest hydroelectric infrastructure in the country, which is responsible for approximately 44.6% of the electric energy production and transmission system in Brazil (ANA, 2017). Thus, the determination of precipitation in this region is essential for forecasting with hydrological and climatological monitoring models. Also, it is important for this region to detect anomalies related to excess or deficit of precipitation, which can compromise the supply of energy and hydraulic structures.

Thus, the objectives of the study are: (1) to evaluate the performance of IMERG's (Version 6) daily and monthly products and (2) to evaluate IMERG's (Version 6) performance in detecting monthly anomalies in the Paraná state using observations from a dense network of precipitation gauges. This study is expected to provide a reference for the use of IMERG (Version 6) products in monthly and daily temporal resolutions and further contribute to improvements of the satellite precipitation algorithm.

2.2. Material and methods

2.2.1. Study region

The study area is the state of Paraná in the south region of Brazil (Figure 1). Paraná occupies an area of 199,315 km^2 and covers 399 municipalities (IBGE, 2011). According to the Köppen classification, carried out by Alvares et al. (2013), the state is located in the transition from tropical and subtropical climates where the humid subtropical Cfa (hot summer) and Cfb (warm summer) predominate in 61.7% and 37.0%, respectively across the state. Because of its extensive area, there is a great diversity in terms of climate, soil types, vegetation, and agricultural use. The main biomes that constitute the state are the Atlantic Forest and the Cerrado. The predominant agricultural crops are maize, soy, and sugar cane. Most of Paraná relief is found at altitudes above 600 m (Figure 1), subdivided into four Morpho-sculptural Units (Santos et al. 2006).

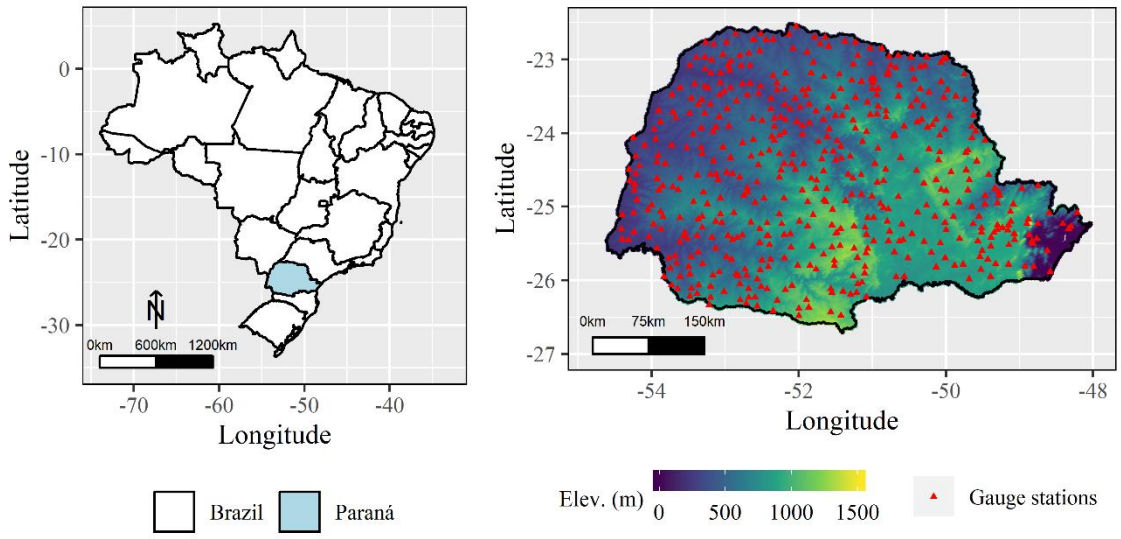

Figure 1. Study area and gauges used to validate the IMERG estimates of precipitation at daily and monthly time-steps.

Precipitation in Paraná varies spatially, with an annual average between 1300 - 2200 mm. Summer is the season with the highest rainfall in South America, including the subtropical region (Grimm et al. 2007; Grimm 2011).

2.2.2. Data

2.2.2.1. Observed data: Ground gauge

The precipitation data used were acquired from 511 gauges (Figure 1) of the National Water Agency (ANA), through the Hidroweb Portal (http://www.snirh.gov.br/hidroweb/serieshistoricas). All available daily data were used for analysis on a daily scale, while, for analysis on a monthly scale, monthly totals were eliminated when there were more than 5% of daily failures in the corresponding month. The daily data was accumulated to produce the monthly information. The analyzed time series was from June 1, 2000 to December 31, 2018.

2.2.2.2. Estimated data: IMERG

Precipitation data from remote sensing were acquired as daily and monthly values with spatial temporal resolution of 0.1°, from the satellite constellation of the Global Precipitation Measurement mission (GPM), IMERG Version 6 product, distributed by Goddard Earth Sciences Data and Information Services Center Distributed Active Archive Center (GES DISC DAAC), available online on: http://mirador.gsfc.nasa.gov/com. The daily and monthly products "Final Run" were used, and the time series analyzed was coincident with that of the gauges (June 1, 2000 to December 31, 2018).

The IMERG algorithm operates to intercalibrate, merge, and interpolate all satellite microwave precipitation estimates, microwave-calibrated infrared estimates, gauge observations and other data from potential sensors from the TRMM and GPM eras (Huffman et al., 2019). The "Final Run" product includes microwave-infrared estimates without gauge adjustment and the calibrated product based on the Global Precipitation Climatology Centre monthly gauge analysis (Tang et al., 2020). In general, the "Final Run" products present bias correction and more accurate results than the other products supplied almost in real time (Early and Late Run) (Su et al., 2019).

2.2.3. Performance analyses

The data quantitative assessment was performed using the correlation coefficient (CC), the determination coefficient (R^2), the mean error (MBE), the mean absolute error (MAE) and

the root of the mean square error (RMSE). The data qualitative assessment was performed using the categorical skills metrics: probability of detection (POD), false alarm ratio (FAR), and critical success index (CSI). POD indicates the fraction of rain events detected correctly with the total number of events detected by satellite; FAR measures the fraction of occurrences of unreal events among the total number of events detected by satellite; CSI denotes the proportion of rain events correctly detected by satellite to the total number of observed events. Such metrics are used in several studies to assess the performance of satellite products (Tang et al., 2016; Chen et al., 2018; Fang et al. 2019; Su et al., 2019). The equations for the metrics used are shown in Table 1. The rainfall threshold was considered as amounts higher than 1 mm day^{-1}.

Table 1. Summary of statistical indices used to evaluate the satellite precipitation products.

Index	Unit	Equation *	Best value
Correlation coefficient (CC)	-	$CC = \frac{\sum_{i=1}^{n}(P_i - \bar{P})(O_i - \bar{O})}{\sqrt{\sum_{i=1}^{n}(P_i - \bar{P})^2 \sqrt{\sum_{i=1}^{n}(O_i - \bar{O})^2}}}$	1
Determination coefficient (R^2)	-	$R^2 = \frac{\{\sum_{i=1}^{n}(P_i - \bar{P})(O_i - \bar{O})\}^2}{\sum_{i=1}^{n}(P_i - \bar{P})^2 \sum_{i=1}^{n}(O_i - \bar{O})^2}$	1
Mean error (MBE)	mm	$MBE = \frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)$	0
Mean absolute error (MAE)	mm	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert P_i - O_i \rvert$	0
Root of the mean square error (RMSE)	mm	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2}$	0
Probability of detection (POD)	-	$POD = \frac{Hits}{Hits + Misses}$	1
Critical success index (CSI)	-	$CSI = \frac{Hits}{Hits + FalseAlarm + Misses}$	1
False alarm ratio (FAR)	-	$FAR = \frac{FalseAlarm}{Hits + FalseAlarm}$	0

* where, Oi is the observed data of gauges of order i; Pi is the estimated order data (IMERG) of order i; Hits are the days when IMERG and the station recorded rain; False Alarm are the days when IMERG recorded rain, but the gauges did not; Misses are the days when IMERG did not register rain, but the gauges did.

2.2.4. Analysis of anomalies

For the analysis of anomalies, the monthly values of the gauges and the monthly products of IMERG were used. The investigation of the volume of precipitation for each month

concerning its average (2000-2018), was carried out by calculating the normalized anomalies of precipitation with standard deviation (Eq.1) (Aragão et al., 2007):

$$X_{Anomaly} = \frac{(X_i - \overline{X}_{2000\text{-}2018})}{\sigma_{2000\text{-}2018}} \quad (1)$$

where, Xi is the month of the year analyzed, X is the monthly average of the 2000 - 2018 series, and σ is the monthly standard deviation of the time series. The mean ($\overline{X}$) and standard deviation (σ) were calculated using Eq. (2) and (3), respectively:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X - \overline{X})^2}{n}} \quad (2)$$

$$\overline{X} = \frac{\sum_{i=1}^{n}(X_i)}{n} \quad (3)$$

Significant anomalies (95% confidence interval) were $X_{Anomaly}$ values greater than or equal to 1.96 and less than or equal to -1.96 as adopted by Silva Junior et al. (2018).

2.3. Results and discussion

2.3.1. Temporal and spatial distribution of precipitation

Figure 2 shows the average monthly precipitation (mm month^{-1}) observed by gauges and estimated by IMERG, between June 2000 and December 2018. The volume and spatial distribution of observed and estimated precipitation were consistent over all months of the year. The rainfall distribution density curve, which relates the precipitated volume to the observation frequency, had a similar distribution, being very similar between the observed data from the gauges and those estimated by IMERG.

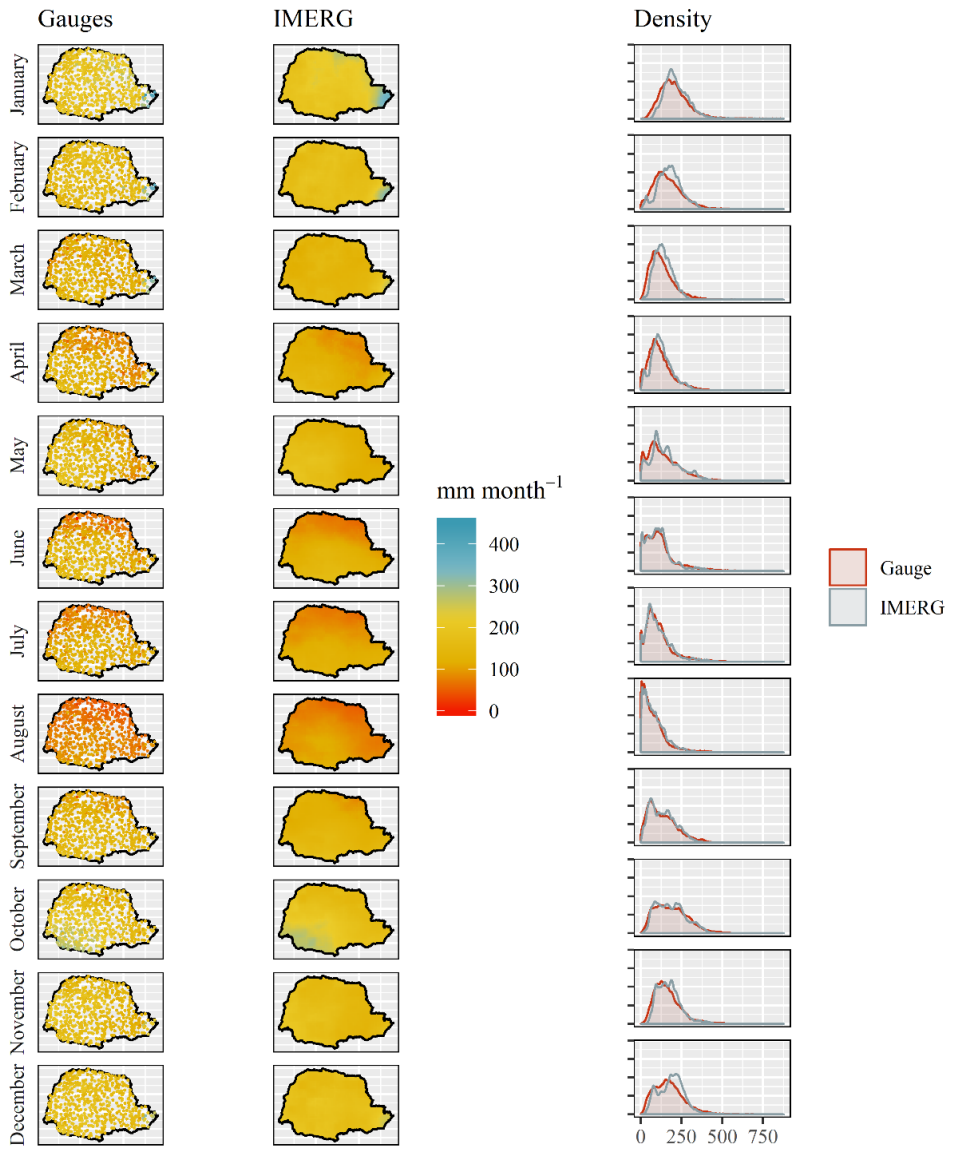

Figure 2. Average monthly rainfall observed during the study period for gauges and by IMERG, and density curve of monthly observations.

The highest frequency of precipitation observations occurs between 125 - 200 mm month^{-1} from October to March which is the wet season and is well spatially distributed in Paraná (Figure 2). In January and February, precipitation above 300 mm month^{-1} occurs in the coastal area. In October, the same behavior is observed in the southwest of the state. The IMERG data overestimated the high values of monthly precipitation recorded by the gauges, presenting a higher frequency of the monthly precipitation peaks in the wet season (Figure 2).

The dry season occurs between April and September, during which the precipitation decreases from the south to the north across the state, towards the Tropic of Capricorn, with a greater frequency observed between 0 - 125 mm month^{-1}. In this period, the months of July and

August stand out as the driest months of the year, reaching frequency of observation close to 0 mm month^{-1}.

In the summer months, the Paraná State region typically receives a humidity air mass that moves from the Amazon to the southwestern Atlantic region, defined as the South Atlantic Convergence Zone (SACZ, Hirata and Grimm, 2016), which is directly connected to the South American monsoon system along a northwest-southeast axis. The precipitation observed in the southwest region in October, on the other hand, is related to convective and frontal complexes (Boulanger et al., 2005; Zandonadi, et al., 2016). The summer wet season, and the spatial and temporal distribution in the study area presented herein is corroborated by previous studies (Grimm et al., 1998, Boulanger et al., 2005, Terassi and Galvani, 2017; Zandonadi, et al., 2016) confirming the precipitation estimates through IMERG through remote sensing approaches.

2.3.2. Daily and monthly spatial evaluation of y IMERG products

The dispersion the total daily and monthly precipitation of the IMERG products versus the observed data are presented in Figure 3. For the daily values, it is observed that the IMERG overestimates the values, with a low coefficient of determination (R^2 = 0.19). For monthly values, precipitation values are close to observed, with R^2 of 0.73.

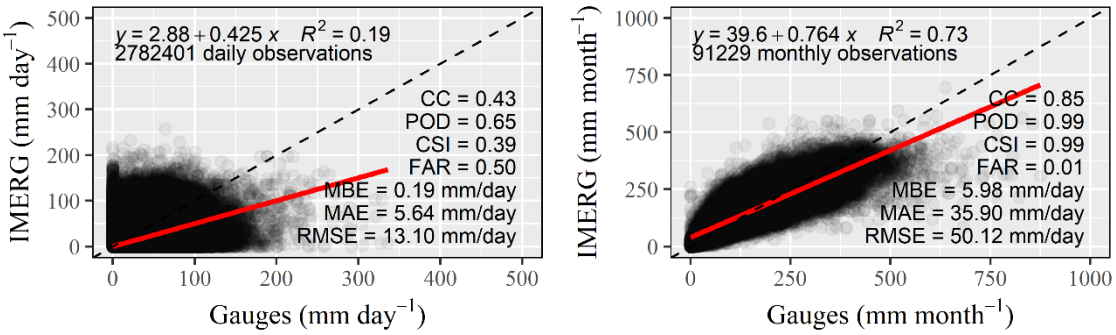

Figure 3. Regressions between daily and monthly data observed by gauges and IMERG.

Based on the general summaries of the metrics used in this study, presented in Figure 3, IMERG shows better performance for estimating monthly precipitation. The high CC value (0.85) indicates a strong correlation between the monthly products of IMERG and the precipitation gauge data, which shows its ability to quantify monthly precipitation in the humid subtropical region. The lower accuracy is observed for daily products, (CC=0.43) indicating

lower correlation between satellite precipitation data and pluviometers. This is mainly due to the high variability of precipitation over small areas on a daily scale.

Our findings indicate that both IMERG products (daily and monthly) overestimate the observed precipitation, with a bias (MBE) of 0.19 and 5.98 mm, in the daily and monthly products, respectively. Regarding the errors, IMERG presented MAE of 5.64 and 35.90 mm and accuracy (RMSE) of 13.10 and 50.12 mm for daily and monthly products, respectively. The better performance of the monthly precipitation product throughout Brazil, was also confirmed by Gadelha et al. (2019). The authors also observed an average of CC of 0.93 and RMSE of 23.20 mm when comparing IMERG version 5 in the state of Paraná. The better performance observed by Gadelha et al. (2019), is possibly due to the methods used for interpolation of observed capture data, when assimilating them to the IMERG forecast fields. Typically, interpolation methods do not capture a large spatial variability of rain and the estimation is complex due to the spatial discontinuity (Hewitson and Crane, 2005). The interpolation leads to the smoothing of high and low peaks of precipitation in a region, improving the relationship between observed and modeled values.

As for the ability to detect rain at monthly resolution, IMERG has an almost perfect performance, with POD and CSI of 0.99 (very close to ideal, 1) and FAR of 0.01 (very close to ideal, 0). In daily resolution, IMERG demonstrated limited capacity to detect rain events, with CSI of 0.39, detection probability of 65% and the risk of false alarm of 50%.

Intensity and volume on a daily scale provide important information in hydrological applications, such as frequency analysis, daily precipitation event detection, and irrigation planning. In this way, the inferior metrics of the daily products requires attention by the user and previous calibrations of the products at this temporal scale. Melo et al. (2015) also observed a better performance of the monthly estimates of the precipitation when analyzing the TRMM products in Brazil. According to the authors, the monthly estimates are less affected by systematic errors than daily estimates. The IMERG products are calibrated using monthly data of in situ gauging stations of the Global Precipitation Climatology Centre (GPCC) network (Anjum et al., 2018), which also can explain the better performance of products on a monthly compared to the daily scale.

The geographic mesoregions of Paraná, separated by the Brazilian Institute of Geography and Statistics (IBGE) and shown in Figure 4, were considered for the regional analysis of the performance of the IMERG products over the study area. The spatial distribution of the metrics for the Paraná regions at daily and monthly scale are summarized in Table 2.

Less accuracy was observed in the southwest, west, central-west and metropolitan regions in daily and monthly products.

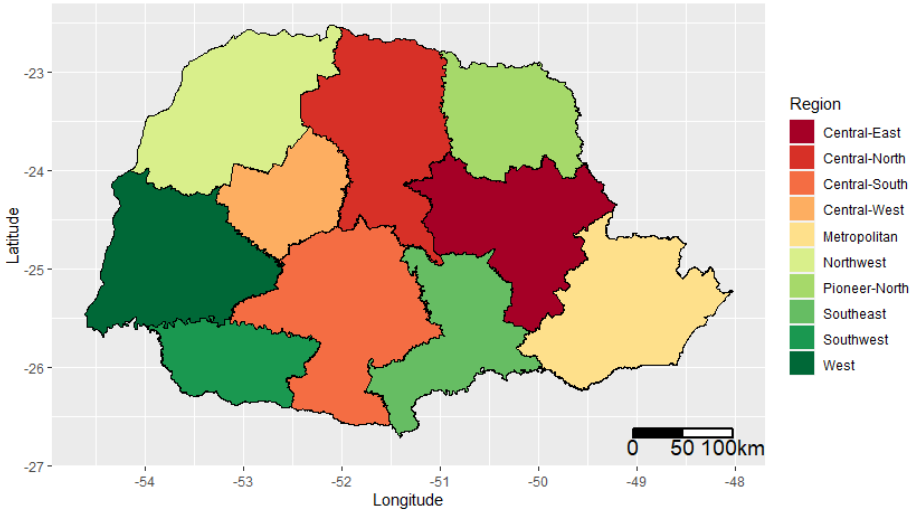

Figure 4. Mesoregions of Paraná state separated by the Brazilian Institute of Geography and Statistics (IBGE).

Table 2. Summary of error metrics on a daily and monthly scale of the satellite precipitation products in Paraná regions. The metrics were calculated based on mean area precipitation.

	RMSE	MAE	MBE	RMSE	MAE	MBE
Region		**(mm day^{-1})**			**(mm month^{-1})**	
Central-South	13.70	6.19	0.02	48.50	36.30	0.83
Central-West	13.40	5.79	0.14	51.20	37.80	4.44
Central-East	11.90	5.24	0.18	43.20	31.90	5.52
Metropolitan	13.10	5.99	0.44	56.90	41.80	13.20
Northwest	12.40	5.14	0.31	50.50	36.90	9.37
Central-North	12.20	5.17	0.14	44.40	32.10	4.25
Pioneer-North	11.30	4.66	0.26	46.90	33.40	7.75
West	14.30	6.12	0.08	55.40	39.50	2.38
Southeast	12.60	5.58	0.31	41.60	31.10	9.41
Southwest	14.80	6.50	0.17	48.20	35.80	5.54

A spatial distribution of the error metrics for estimating daily and monthly rainfall by IMERG, for each gauge in the state of Paraná, are shown in Figure 5. Corroborating the results presented in Table 2, the MBE and MAE values are spatially well distributed in the study area with close values to the averages shown in Figure 3, for daily and monthly data. However, less

accurate metrics were observed in the coastal areas (eastern part of the state), where IMERG presents greater disagreements in some stations. Since the IMERG pixel covers an extensive area, large variability in precipitation are masked in areas where orographic effects are prevalent. This is evident near the coastal region where there are abrupt changes in elevation (Figure 1).

The change in a short distance between the ocean and coastal mountains can induce failure of the satellite sensor to discriminate the adjacent pixels of land and water, generating signal contamination in coastal areas and resulting in poor performance in estimating precipitation (El Kenawy et al., 2015). The limitation of the precipitation estimates by satellites (Global Precipitation Climatology Project - GPCP) in orographic regions also was observed in the Andes by Schumacher et al (2020). According to the authors, the spatial resolution (2.5°) and dependence on passive microwaves and infrared precipitation recoveries used by satellites may imply the worst performance of representing precipitation in locations with orographic type precipitation. Despite presenting higher spatial resolution (0.1°), IMERG products also showed less accuracy for estimating rainfall in orographic regions.

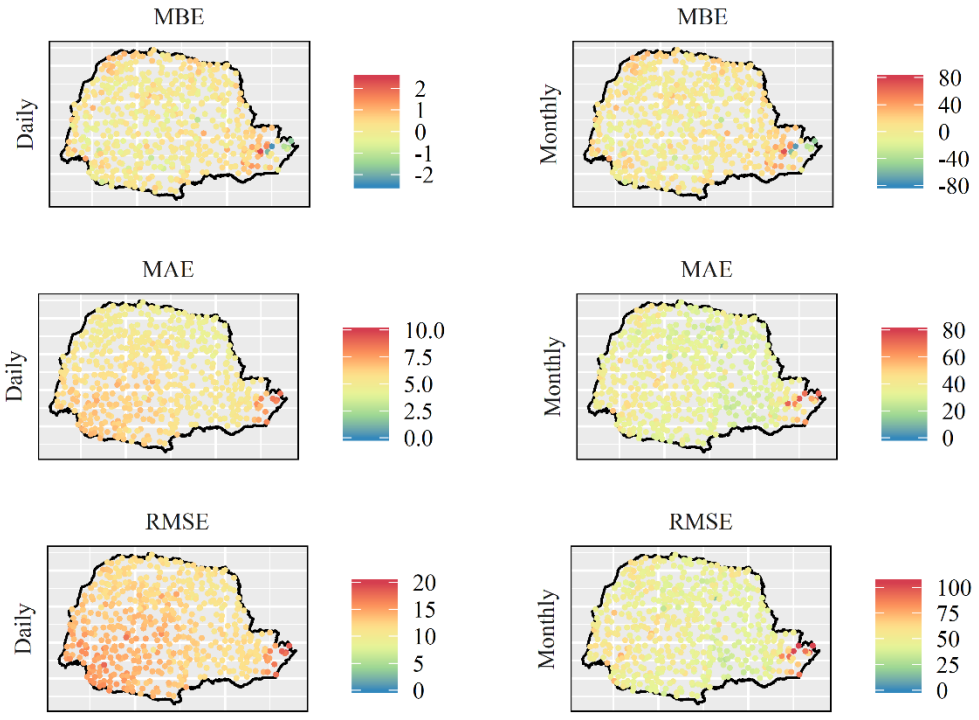

Figure 5. Spatial distribution of error metrics on a daily and monthly scale.

The distribution of POD, CSI, and FAR (Figure 6) showed good performance of IMERG's products in detecting monthly rain events throughout Paraná, with values very close

to ideals in the entire area. The performance of IMERG's products in estimating daily rainfall also was homogeneous across the state area, with the worst performance on the coast, in agreement with the statistical metrics.

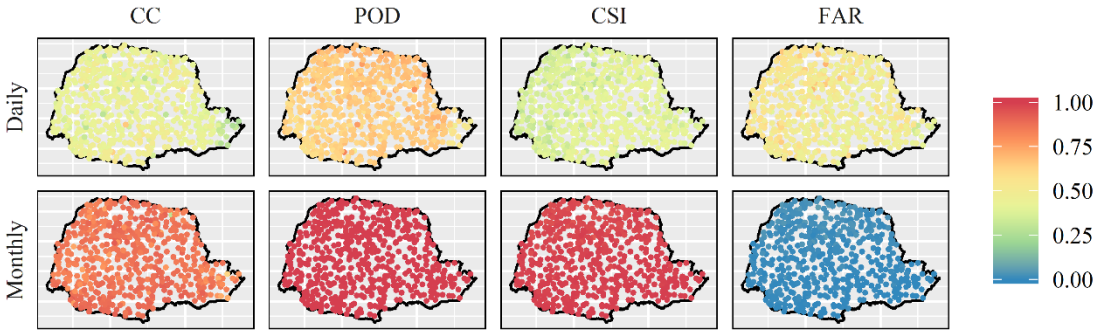

Figure 6. Spatial distribution of the qualitative metrics of IMERG performance on a daily and monthly scale.

In the mountainous region, rainfall tends to be underestimated towards the ocean (east aera), where orographic rains occur. On the opposite side, after the sudden change in elevation, it tends to overestimate the precipitation. Corroborating these results, Duan et al. (2015) also observed trends of underestimation of precipitation by the TRMM (2A25 version 7) product in regions of orographic rainfall and overestimation in regions of valleys or flat areas in southeastern Appalachians. According to the authors, this behavior occurs due to the spatial resolution and the correction of soil characteristics made by the satellite. Another possibility for underestimating precipitation in the coastal region is that precipitation occurs while the top of the cloud is still relatively warm. Satellites are unable to fully identify rain, as heat exceeds infrared thresholds and the lower amount of ice in the air makes detection by passive microwave sensors difficult, and thus satellite products detect only part of the precipitation. (Dinku et al., 2008; Karaseva et al., 2012; Guo et al., 2016).

In addition, in mountainous regions, precipitation is extremely variable and there are changes in rainfall distribution over short distances (Navarro et al., 2019), which can result in a representation of precipitation with less accuracy in these areas, since satellite products have the limitation of estimating precipitation considering the pixel size when compared to gauges that measures the precipitation in situ.

Also, in the coastal area, the highest peak of precipitation occurs in the summer and is related to the predominant role of the Atlantic Tropical Mass (Vanhoni and Mendonça, 2008), which finds the mountain as a physical barrier and culminates in orographic rains, with great

volume over a short duration interval. Such an event may not be fully captured by satellites, as observed by Tian et al., (2009) on the west coast of the United States, in the analysis of the precipitation of six satellite products (AFWA, TMPA -3B42, TMPA - 3B42RT, CMORPH, PERSIANN, and NRL) and Su et al., 2019 in China, with IMERG Version 5 products. Despite notable improvements in the Version 6 of IMERG's algorithm over previous versions (detailed by Tang et al., 2020), Navarro et al. 2019 also observed less accuracy of IMERG Version 6 in estimating rainfall on the coastline of the Adriatic Sea, in Europe. Navarro et al. (2020) also related the limitation of estimating precipitation in the Ebro River basin, in Spain, in an area where weather was dominated by the advection of wet maritime air masses. Thus, the measurement of the precipitation over coastal locations still poses a challenge and deserves further research.

Concerning the daily and monthly RMSE, higher values are found in the southwest and coastal areas of Parana, which correspond to areas with high volumes of rain in the autumn and summer, respectively, and the highest volumes of annual rainfall in Paraná (Figure 3). Thus, the precipitation estimates by IMERG performed better in the drier areas of the state.

2.3.3. Rainfall anomalies between 2000 and 2018

The spatial distribution of the performance of IMERG's monthly resolution products in detecting anomalies observed by the gauges is shown in Figure 7. In general, IMERG showed limited capacity for detecting anomalies across the state, considering +/-1.96 monthly standard deviation. The best performance was observed in the south-central region of Paraná, with POD above 0.75, CSI above 0.50 and FAR below 0.50. The worst performance occurred in the northeast region, the region with the lowest annual rainfall, and in the coastal and southwest regions, which corresponds to the regions with the highest annual rainfall in the state (Figure 1), agreeing with the worst performance of the daily and monthly product metrics (Figure 5).

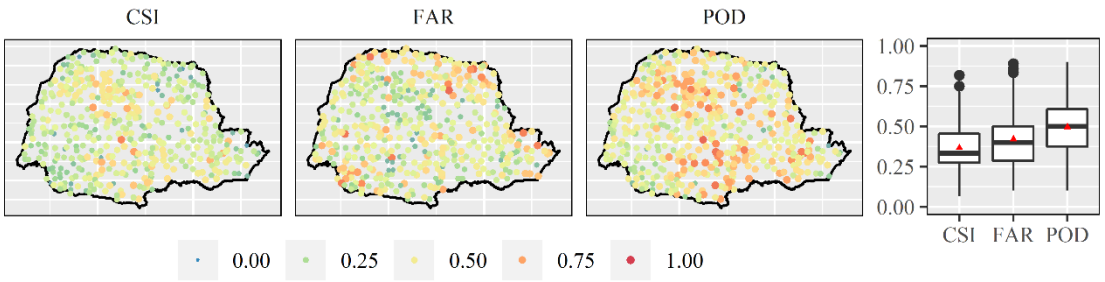

Figure 7. Spatial distribution of performance metrics for anomaly detection.

The northeastern area of the state is concentrated in the central part of the Paraná River basin, close to the climate transition line (subtropical and tropical) that separates several active climate systems in an area of greater atmospheric instability (Zandonadi et al 2016). Thus, the precipitation estimate by the satellite may present a greater limitation (overestimation or underestimation) in this area due to atmospheric conditions.

As previously mentioned in the discussion of the performance of daily and monthly products IMERG, high intensity orographic rainfall occurs along the coast, which is estimated with less accuracy by satellite-based algorithms, resulting in poor performance in detecting anomalies. The southwestern region of Paraná has favorable conditions for the formation of severe storms and hail, which occur very quickly (Brooks et al., 2003; Martins et al., 2017; Beal et al., 2020). In September and October, the highest amounts of hail formation are observed in days per month (Beal et al., 2020), which exactly coincides with the period of greatest rainfall in the region (Figure 3). Thus, as on the coast of Paraná, the sensitivity of IMERG in detecting anomalies in this region may have a lower performance compared to other regions.

In the analysis of the boxplot of the metrics (POD, CSI, and FAR) used to assess the performance of IMERG in detecting anomalies, a relative deviation from the mean and extreme values of CSI and FAR were observed (Figure 6). Therefore, eight stations were selected randomly, one in each region of the state, to detect anomalies by the gauges and by IMERG, between the years 2000 and 2018 (Figure 8).

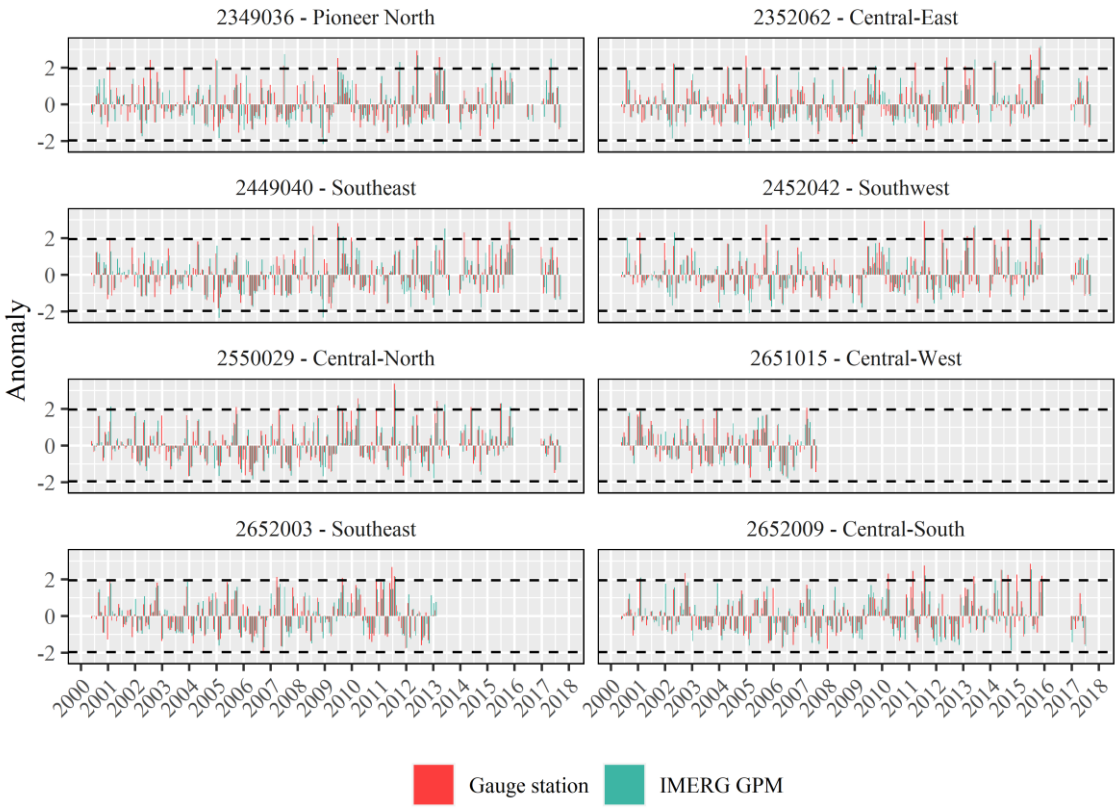

Figure 8. Anomalies for nine stations selected at random. The dashed lines represent the values of -1.96 and 1.96 σ (monthly standard deviation).

According to Grimm (2011), positive anomalies in rainfall can occur during the summer under conditions of El Niño Southern Oscillation (ENSO) in the south of Brazil. During the study period, two El Niño events classified as ”moderate,” between 2002-2003 and 2009-2010, and one classified as ”strong,” between 2015-2016 (Golden Gate Weather Services, 2020) occurred. During these periods, positive anomalies were detected by IMERG and gauges in all stations.

In the other years, IMERG and gauges detected some anomalies in all nine stations analyzed. However, only IMERG detected negative anomalies. Su et al. (2019) reported that IMERG version 5 products tend to underestimate precipitation amounts for rainfall rates 40 – 75 mm day^{-1}, but overestimate precipitation amounts for high rainfall rates (> 80 mm day^{-1}). The anomalies detected in this study by the satellite may have occurred because under or overestimation rain events. Thus, the use of IMERG products in anomaly studies must consider their variable performance for this purpose, requiring calibrations and prior data assessments.

2.4. Conclusions and final considerations

In this study, the performance of IMERG Version 6 products in estimating daily and monthly precipitation were evaluated in comparison with the data observed by 511 gauges distributed in the state of Paraná in Brazil. The results showed better metrics for monthly precipitation. In summary, the main findings of this study were:

i. The volume and spatial distribution of observed and estimated rainfall are consistent across all months of the year in the monthly products of IMERG Version 6, with similar rainfall distribution density curves.

ii. IMERG Version 6 has a good relationship between precipitation estimates and those observed by gauges on the monthly time scale, with high correlation and accuracy, and low errors in statistical metrics. However, a lower performance was observed in estimating rainfall in regions with abrupt changes in topography along the coast, related to the less accuracy to estimate orographic affected rainfall.

iii. The monthly products of IMERG Version 6 performed very close to perfect considering qualitative assessments for the detection of precipitation events in this time scale throughout the study area.

iv. The daily estimates of IMERG Version 6 were limited in representing the rainfall observed by the gauges, with little correlation between the data and low values of rain event detection rates. Although the gauges are direct observations and considered references, it is known there is a great spatial variability in daily data, which is the probable cause of the low performance.

v. The detection of anomalies by the monthly products of IMERG Version 6 showed limited performance over the years analyzed and the study area, probably due to the topography and rainfall regime in the northeast, coast, and southeast.

Based on the results presented here, IMERG Version 6 can be used as a source of monthly precipitation data over the territory of Paraná. However, on a daily scale, prior calibration of the product is recommended to ensure the good performance of the estimate on this time scale, especially for mountainous areas. Future improvements to IMERG Version 6 products may increase its accuracy and favor its application for the detection of rain in coastal areas and anomalies. Also, studies that consider seasonal analyses and other time scales (hourly and half-hourly), areas with complex topographies, and the other products of IMERG Version 6 (Early and Late Run) are strongly recommended.

References

Alvares, C. A., Stape, J. L., Sentelhas, P. C., de Moraes, G., Leonardo, J., Sparovek, G., 2013. Köppen's climate classification map for Brazil. Meteorologische Zeitschrift, 22, 711-728. https://doi.org/10.1127/0941-2948/2013/0507

ANA. Conjuntura dos recursos hídricos no Brasil 2017; relatório pleno / Agência Nacional de Águas. Brasília. 2017. http//www.ana.gov.br (assesed 18 March 2020).

Anderson, L.O.; Malhi, Y.; Aragão, L.E.O.C.; Ladle, R.; Arai, E.; Barbier, N.; Phillips, O. Remote sensing detection of droughts in Amazonian forest canopies. New Phytol. 2010, 187, 733–750.

Anjum M. N., Ding Y. J., Shangguan D. H., Ahmad I., Ijaz M. W., Farid H. U., Yagoub Y. E., Zaman M., AdnanM. 2018. Performance evaluation of latest Integrated Multi-satellitE Retrievals for GPM (IMERG) (IMERG) over the northern highlands of Pakistan. Atmos. Res., 205,134-146. https://doi.org/10.1016/j.atmosres.2018.02.010

Aragão, L.E.O.C., Malhi, Y., Roman-Cuesta, R.M., Saatchi, S., Anderson, L.O., Shimabukuro, Y.E. 2007. Spatial patterns and fire response of recent Amazonian droughts. Geophysical Research Letters, 34, 7. https://doi.org/10.1029/2006GL028946

Beal, A., Hallak, R., Martins, L. D., Martins, J. A., Biz, G., Rudke, A. P., Tarley, C. R. 2020. Climatology of hail in the triple border Paraná, Santa Catarina (Brazil) and Argentina. Atmos. Res. 234, 104747. https://doi.org/10.1016/j.atmosres.2019.104747

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

Boulanger, J. P., Leloup, J., Penalba, O., Rusticucci, M., Lafon, F., Vargas, W. 2005. Observed precipitation in the Paraná-Plata hydrological basin: long-term trends, extreme conditions, and ENSO teleconnections. Clim. Dyn., 24, 393-413. https://doi.org/10.1007/s00382-004-0514-x

Brooks, H. E., Lee, J. W., Craven, J. P. 2003. The spatial distribution of severe thunderstorm and tornado environments from global reanalysis data. Atmos. Res., 67, 73-94. https://doi.org/10.1016/S0169-8095(03)00045-0.

Chen, C., Chen, Q., Duan, Z., Zhang, J., Mo, K., Li, Z., Tang, G. 2018. Multiscale comparative evaluation of the GPM IMERG v5 and TRMM 3B42 v7 precipitation products from 2015 to 2017 over a climate transition area of China. Remote Sens., 10, 944. https://doi.org/10.3390/rs10060944

Chen, H., Yong, B., Shen, Y., Liu, J., Hong, Y., Zhang, J. 2020. Comparison analysis of six purely satellite-derived global precipitation estimates. J. Hydrol., 581, 124376. https://doi.org/10.1016/j.jhydrol.2019.124376

Dinku, T., Chidzambwa, S., Ceccato, P., Connor, S. J., Ropelewski, C. F. 2008. Validation of high-resolution satellite rainfall products over complex terrain. Int. J. Remote. Sens., 29, 4049-4110. https://doi.org/10.1080/01431160701772526

Duan, Y., Wilson, A. M., Barros, A. P. 2015. Scoping a field experiment: error diagnostics of TRMM precipitation radar estimates in complex terrain as a basis for IPHEx2014. Hydrol. Earth Syst. Sci., 19. https://doi.org/10.5194/hess-19-1501-2015

El Kenawy, A. M., Lopez-Moreno, J. I., McCabe, M. F., Vicente-Serrano, S. M. 2015. Evaluation of the TMPA-3B42 precipitation product using a high-density gauge network over complex terrain in northeastern Iberia. Glob. Planet Change, 133, 188-200. https://doi.org/10.1016/j.gloplacha.2015.08.013

Falck, A. S., Maggioni, V., Tomasella, J., Diniz, F. L. R., Mei, Y., Beneti, C. A., Herdies D. L., Neundorf R., Caram R. O., Rodriguez, D. A. 2018. Improving the use of ground-based radar rainfall data for monitoring and predicting floods in the Iguaçu river basin. J. Hydrol., 567, 626-636. https://doi.org/10.1016/j.jhydrol.2018.10.046

Fang, J., Yang, W., Luan, Y., Du, J., Lin, A., Zhao, L. 2019. Evaluation of the TRMM 3B42 and GPM IMERG products for extreme precipitation analysis over China. Atmos. Res. 223, 24-38. https://doi.org/10.1016/j.atmosres.2019.03.001

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., Michaelsen J. 2015. The climate hazards infrared precipitation with stations —a new environmental record for monitoring extremes. Sci. Data, 2, 1-21. https://doi.org/10.1038/sdata.2015.66

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

Grimm, A. M. 2011. Interannual climate variability in South America: impacts on seasonal precipitation, extreme events, and possible effects of climate change. Stoch. Env. Res. Risk Assess., 25, 537-554, https://doi.org/10.1007/s00477-010-0420-1

Grimm, A. M., Ferraz, S. E., Gomes, J. 1998. Precipitation anomalies in southern Brazil associated with El Niño and La Niña events. J. Clim., 11, 2863-2880. https://doi.org/10.1175/1520-0442(1998)011<2863:PAISBA>2.0.CO;2

Grimm, A.M., Pal, J., Giorgi F. 2007. Connection between spring conditions and peak summer monsoon rainfall in South America: role of soil moisture, surface temperature, and topography in eastern Brazil. J. Clim., 20, 5929-5945. https://doi.org/10.1175/2007JCLI1684.1

Guo, H., Chen, S., Bao, A., Behrangi, A., Hong, Y., Ndayisaba, F., Hu, J., Stepanian, P. M. 2016. Early assessment of integrated multi-satellite retrievals for global precipitation measurement over China. Atmos. Res., 176, 121-133. https://doi.org/10.1016/j.atmosres.2016.02.020

Hewitson, B. C., Crane, R. G. 2005. Gridded area-averaged daily precipitation via conditional interpolation. J. of Clim., 18, 41-57. https://doi.org/10.1175/JCLI3246.1

Hirata, F. E., Grimm, A. M. 2016. The role of synoptic and intraseasonal anomalies in the life cycle of summer rainfall extremes over South America. Clim. Dyn., 46, 3041-3055, 2016.

Hobouchian, M. P., Salio, P., Skabar, Y. G., Vila, D., Garreaud, R. 2017. Assessment of satellite precipitation estimates over the slopes of the subtropical Andes. Atmos. Res., 190, 43-54. https://doi.org/10.1016/j.atmosres.2017.02.006

Hou, A. Y., Kakar, R. K., Neeck, S., Azarbarzin, A. A., Kummerow, C. D., Kojima, M., Oki R., Nakamura K., Iguchi, T. 2014. The global precipitation measurement mission. Bull. Am. Meteorol. Soc., 95, 701-722. https://doi.org/10.1175/BAMS-D-13-00164.1

Huffman, G. J., Adler, R., Bolvin, D. T., Guojun G., Nelkin, E. J., Bowmans K. P., Hong Y., Stocker, E. F., Wolff, D. B. 2007. The TRMM multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. J. Hydrometeorol., 8, 38-55. https://doi.org/10.1175/JHM560.1

Huffman, G. J., Bolvin, D.T., Nelkin E.J. 2015. Integrated Multi-satellitE Retrievals for GPM (IMERG) Technical Documentation. NASA/GSFC Code 612, 47. http://pmm.nasa.gov/sites/default/files/document_files/IMERG_doc.pdf (Accessed 02 February 2020).

Huffman, G.J., Bolvin, D.T., Nelkin, E.J., Tan J. 2019. Integrated Multi-satellite Retrievals for GPM (IMERG) Technical Documentation, https://docserver.gesdisc.eosdis.nasa.gov/public/project/GPM/IMERG_doc.06.pdf.

Joyce, R. J., Janowiak, J. E., Arkin, P. A., Xie, P. 2004. CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. J. Hydrometeorol, 5, 487-503. https://doi.org/10.1175/1525-7541(2004)005<0487:CAMTPG>2.0.CO;2

Karaseva, M. O., Prakash, S., Gairola, R. M. 2012. Validation of high-resolution TRMM-3B43 precipitation product using gauge measurements over Kyrgyzstan. Theor. Appl. Climatol., 108, 147-157. https://doi.org/10.1007/s00704-011-0509-6

Kucera, P. A., Ebert, E. E., Turk, F. J., Levizzani, V., Kirschbaum, D., Tapiador, F. J., Loew A., Borsche, M. 2013. Precipitation from space: Advancing Earth system science. Bull. Am. Meteorol. Soc., 94, 365-375. https://doi.org/10.1175/BAMS-D-11-00171.1

Kummerow, C., Barnes, W., Kozu, T., Shiue, J., Simpson, J. 1998. The tropical rainfall measuring mission (TRMM) sensor package. J. Atmos. Ocean. Technol., 15, 809-817. https://doi.org/10.1175/1520-0426(1998)015<0809:TTRMMT>2.0.CO;2

Martins, J. A., Brand, V. S., Capucim, M. N., Felix, R. R., Martins, L. D., Freitas, E. D., Gonçalves F. L. T., Hallak R., Dias, M. A. F. S., Cecil, D. J. 2017. Climatology of destructive hailstorms in Brazil. Atmos. Res., 184, 126-138. https://doi.org/10.1016/j.atmosres.2016.10.012

Mega, T., Ushio, T., Takahiro, M., Kubota, T., Kachi, M., Oki, R. 2018. Gauge-adjusted global satellite mapping of precipitation. IEEE Transactions on Geoscience and Remote Sensing, 57, 1928-1935. https://doi.org/10.1109/TGRS.2018.2870199

Melo, D. D. C., Xavier, A. C., Bianchi, T., Oliveira, P. T., Scanlon, B. R., Lucas, M. C., Wendland, E. 2015. Performance evaluation of rainfall estimates by TRMM Multi-satellite Precipitation Analysis 3B42V6 and V7 over Brazil J. Geophys. Res. Atmos., 120, 9426-9436. https://doi.org/10.1002/2015JD023797

Navarro A., García-Ortega E., Merino A., Sánchez L. J., Kummerow C., Tapiador J. F. 2019. Assessment of IMERG precipitation estimates over Europe. Remote Sens., 11, 2470. https://doi.org/10.3390/rs11212470

Navarro, A., García-Ortega, E., Merino, A., Sánchez, J. L., Tapiador, F. J. 2020. Orographic biases in IMERG precipitation estimates in the Ebro River basin (Spain): The effects of rain gauge density and altitude. Atmos. Res., 244, 105068. https://doi.org/10.1016/j.atmosres.2020.105068

Pedron, I. T., Silva Dias, M. A., de Paula Dias, S., Carvalho, L. M., Freitas, E. D. 2017. Trends and variability in extremes of precipitation in Curitiba–Southern Brazil. Int. J. of Clim., 37, 1250-1264. https://doi.org/10.1002/joc.4773

Rozante, J. R., Vila, D. A., Barboza Chiquetto, J., Fernandes, A. D. A., Souza Alvim, D. 2018. Evaluation of TRMM/GPM blended daily products over Brazil. Remote Sens., 10. https://doi.org/10.1016/j.atmosres.2018.12.011

Salio, P., Hobouchian, M. P., Skabar, Y. G., Vila, D. 2015. Evaluation of high-resolution satellite precipitation estimates over southern South America using a dense gauge network. Atmos. Res., 163, 146-161. https://doi.org/10.1016/j.atmosres.2014.11.017

Santos, L. J. C., Oka-Fiori, C., Canali, N. E., Fiori, A. P., da Silveira, C. T., da Silva, J. M. F., Ross, J. L. S. 2006. Mapeamento geomorfológico do Estado do Paraná. Rev. Bras. Geomorfologia, 7, 3-12. http://dx.doi.org/10.20502/rbg.v7i2.74

3. REGIONALIZING MINIMUM AND LONG-TERM FLOWS BY RANDOM FOREST AND MULTIPLE LINEAR REGRESSIONS: A CASE STUDY IN THE PARANÁ STATE, BRAZIL

ABSTRACT

The knowledge of the flows in the rivers is essential for water resources management since it represents the availability of water in the watersheds. In general, the long-term average flows (Qm) and flows exceeded or equaled in 90% and 95% of the time (Q90 and Q95, respectively), which represents minimum flows, frequently adopted as reference in the water resources management. The information about the flow in watersheds can be considered a challenge, especially in developing countries where monitoring by gauges is limited in terms of density and frequency of observations. Regionalization models can be applied to predicted flows in unmonitored watersheds. The Multiple Linear Regressions (MLR) method is the oldest and widely used method for this purpose. The Random Forest (RF) is a machine learning approach recently applied in hydrological studies for flow prediction. This study aimed to analyze the performance of MLR and random forest RF, to predict the reference flows in 81 watersheds in Paraná state, in Brazil. The MLR method was applied in five homogeneous regions in Paraná. For RF method the performance of subset watershed descriptors was tested and showed better performance compared when all descriptors were used. The MLR method outperformed the RF, but their performances were similar in terms of error metrics. Thus, RF can be proposed as a new method for reference flow predictions in subtropical regions.

Keywords: Machine learning, Multiple regressions, Hydrological models, Regression tree, Water management.

3.1. Introduction

Water resources management has, as main objectives, the guarantee of life and human development based on the availability of water for multiple uses. The water management must be efficient in minimizing possible conflicts, balancing the demands for different human activities, and avoiding the progress of environmental degradation.

The knowledge of the flow in the rivers is essential for water resources management, as it informs about the water availability in a watershed. In general, reference flows are adopted in water resources management considering minimum flow aiming a low risk for the rivers (Harris et al., 2000).

A flow duration curve is an important tool to determine reference values of water availability in a watershed. It provides the frequency at which a specific flow value is equaled or exceeded. Two reference values frequently extracted from this curve in the water resources management are Q90 and Q95, which are flows exceeded or equaled in 90% and 95% of the time, respectively, and represent minimum flows (Pereira et al., 2016).

The long-term average flow (Qm) is another essential hydrological parameter which informs about the energy potential of the watershed and represents the highest flow that can be regulated. Among its most varied applications, Qm is the base for calculating the regularization volume when designing a reservoir and is obtained by averaging monthly values over a year and then averaging all the year's values.

The information about the flow in watersheds is a challenge in the water resources management over the countries, where the number of gauges is very small comparing to the areas to be monitored (Swain and Patra, 2017). This is true, especially in developing countries in South America, where monitoring by gauges is limited in terms of infrastructure, maintenance, density, and frequency of observations (Hobouchian et al. 2017, Salio et al. 2015).

However, the optimization of information obtained in monitored sites can be used for prediction in unmonitored watersheds, through statistical approaches that allow hydrological parameters transfer (Sivapalan et al., 2003; Swain; Patra, 2017; Requena et al., 2018). In general, flow predictions consist of transferring historical data between watersheds using models and regionalization techniques.

Multiple linear regression (MLR) is one of the oldest and most widely used methods for hydrological variable predictions in unmonitored sites, using environmental descriptors associated with the watersheds (Swain; Patra, 2017). In this method, dependent variables are calibrated to relate climatic and landscape attributes, to build empirical relationships that can be used to predict variables in unmonitored watersheds (Zhang et al., 2015).

The flow prediction in the MLR presents, in general, good performance, mainly when applied in hydrologically homogeneous regions where climatological and topographic characteristics are similar, and the hydrological responses tend also to be similar, even if these regions are not located side by side geographically (Smakhtin, 2001). In hydrologically homogeneous regions, linear functions provide a good approximation to regional models, but it is not the same reality in a broader area (Li et al., 2010), and alternatives methods could predict better the flows in heterogeneous regions.

Recently, machine learning techniques interested the hydrologists, due to their ability to work with big data and solve the most diverse problems. Random forest (RF) is a machine

learning technique developed by Breiman (2001) that can be used for prediction and classification purposes. The RF regression can work with nonlinear relationships between variables, combining many regression trees by drawing several bootstrap samples from the original training data and analyzing the decision trees (Breiman, 2001; Xu et al. 2019; Booker and Woods, 2014).

A decision tree can be defined as a hierarchical analysis diagram, in which each internal node represents an independent variable, the branch represents the test outcome, and each terminal (leaf) node represents a decision (Xu et al., 2019). The decision rules for node splits are tuned aiming to optimize the homogeneity of the dependent variable. More details can be found in Loh (2011) and Tyralis et al. (2019).

The use of RF in water resources is recent (Tyralis et al., 2019). The RF has been using on water price prediction (Xu et al., 2019), regionalization of hourly hydrological model parameters (Saadi et al., 2019), large-scale flood discharge simulation (Schoppa et al., 2020), and several hydrological parameters and signatures (Booker and Woods, 2014; Addor et al., 2018; Booker and Snelder, 2012). However, few studies focused on applying RF to predict specific quantiles along the flow duration curve (for example, Schnier and Cai, 2014).

Due to the importance of specifics exceedance frequency in the permanence curves (Q90 and Q95) and the long-term average (Qm) in the water resources management, we analyzed the performance of the MLR and RF models methods in predicting these flows. For that purpose, we used a large-scale sample with 81 watersheds in Paraná state, in Brazil.

The flow regionalization is essential for the water management and some studies provided models to predict the flow over São Paulo (Liazi et al., 1988; Wolff et al., 2014), Santa Catarina (Wolff, 2017), and Minas Gerais (IGAM, 2012) states, in Brazil. However, studies aiming the flow prediction across other states are needed.

We focused on the reference flows because they are used as tools in the area this study was applied to. Thereby, we could effectively contribute to improving the water resources management, providing models, and identifying the relevant descriptors for the flow prediction in the region. Additionally, we aimed to contribute our knowledge and capabilities in Hydrology by answering the following questions:

i. Does the RF overperform the MLR in watershed large-scale, in the references flows prediction?

ii. Does the low-cost of RF, in terms of time of execution and free software implementation, support its application for predicting reference flows?

iii. Could a subset of landscape and climatic descriptors be used for the long-term flow and minimum flows prediction by RF models?

3.2. Material and methods

3.2.1. Study region

The study area consists of the Paraná state, in the Southern region of Brazil (Figure 9). Paraná is the fifth most populous state in Brazil, with about 11.4 milion inhabitants, in an area of 199,315 km² and 399 municipalities (IBGE, 2019). The study area is represented by two main Brazilian biomes: Atlantic Forest and Cerrado. Soils are highly weathered, the Ferralsols (30.8%) is the predominant soil type, followed by Entisols (22.2%), Acrisols (15.5%)), Nitisols (15.2%), and Cambisols (10.6%) (Bhering et al., 2008). Land use is mainly composed of natural forest (28%), planted forest (6%), forest species cultivated for commercial purposes, grassland (14%), and short, medium, and perennial agricultural (35%) (MapBiomas, 2019). In the agricultural scenario, Paraná is the second largest Brazilian region that produces and sells grains, mainly soybeans (Carmello and Sant'anna Neto, 2016). Other crops also are produced in the state, such as coffee, corn, beans, cotton, wheat, cassava, sugarcane, and various fruits (Aparecido et al., 2016).

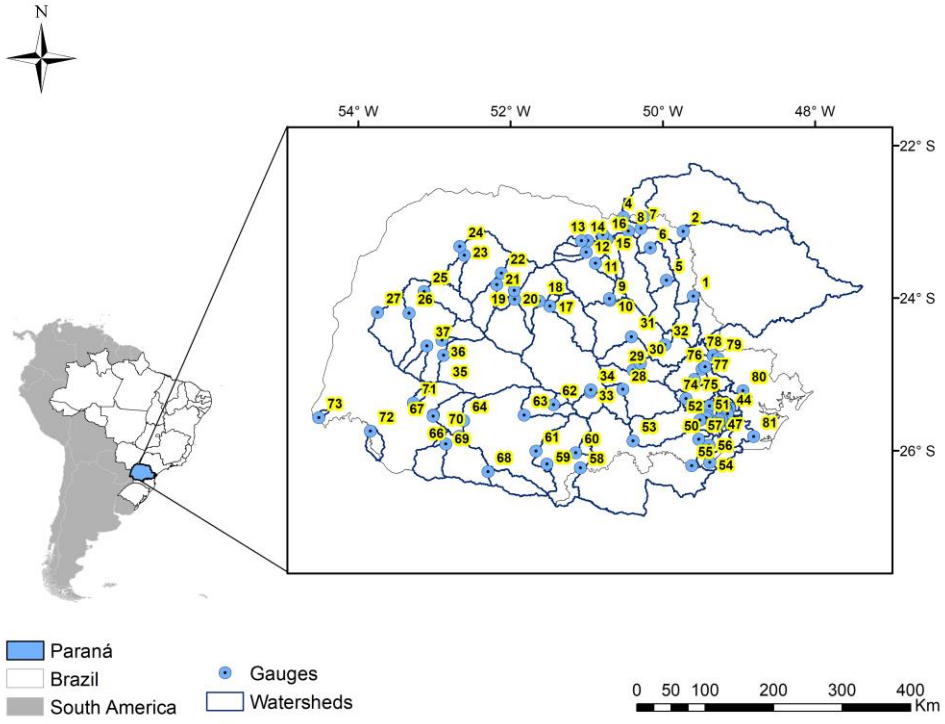

Figure 9. Study area, hydrometric gauges position, and watersheds in the Paraná state. The numbers of each watershed are located at its fluviometric gauge.

The north of Paraná is located in the Tropic of Capricorn, so the state comprises both tropical and subtropical climates. According to Köppen classification, the humid subtropical Cfa (hot summer) and Cfb (warm summer) are the predominant climates (Alvares et al., 2013). The wet season occurs from October to March while the drought period is between April and September.

Regarding Paraná hydrography, it is composed of three Brazilian hydrographic regions: Southeast Atlantic and South Atlantic, both located on the coast of the State, and Paraná River Basin in the entire continental region of the State, occupying the largest area. The Paraná River Basin represents the most important Brazilian hydrological region in the socio-economic aspect, with the largest hydroelectric park, responding to approximately 44.6% of the electric energy production and transmission system in the country (Zandonadi et al., 2016; ANA, 2017).

3.2.2. Data

3.2.2.1. Hydrometric Gauges

The daily flow data were acquired from the National Water Agency (ANA), through the Hidroweb platform (http://www.snirh.gov.br/hidroweb/publico/medicoes_historicas_abas.jsf using) using 81 hydrometric gauges as shown in Figure 9. We selected watersheds that presented time series with at least 15 years of data and less than 10% of gaps, between the years 1920 and 2015. The date period of the gauges data was not necessarily common between all the watersheds, as suggested by Hosking and Wallis (1997). The watersheds (Figure 9) were delimitated in ArcGIS (ESRI, 2014) using the *Acr Hidro Tools* over the Digital Elevation Model (DEM) in Paraná (Figure 10b). The watersheds areas varied between 24.38 and 45700.02 km^2 and were concentrated in the east and central regions of the state.

The gap-filling in the daily flow series was performed using the R environment (R CORE TEAM, 2019), through the *mtsdi* package, which is based on the Maximized Hope algorithm for the imputation of missing values in normal, multivariate time series, using spatial and temporal correlation structures (Junger and Leon, 2018). In this process, the spatial proximity of the watersheds was considered for the input of the missing data.

Daily flow data were added to result in monthly data for calculating the flows analyzed in this study. The average monthly flow in a watershed resulted in the long-term average flow (Qm). We obtained the minimum flows from permanence curves, organizing the monthly data in descending order based on their magnitude. Then, the data were adjusted to probabilistic distributions, for example, Weibull, Pearson 3, Log-Pearson 3, and Log-normal 3, in the software EasyFit 5.5 (MathWave, 2010).

When using a probabilistic distribution as a model to describe water flow rates, it is necessary to apply an adherence test to verify the adequacy of the procedure. Thus, we used the Kolmogorov-Smirnov adherence test. The adjustment of the probability distribution is presented in the Appendix. From the permanence curves, values corresponding to flows with 90% and 95% (Q90 and Q95, respectively) of permanence were obtained. In order to disregard the effect of the size of the watersheds on flows, the Qm, Q90, and Q95 were divided by the area of the respective watershed, resulting in specific flows.

3.2.2.2. Precipitation data and watershed descriptors

Climatic variables and landscape characteristics of the watersheds represent the main important descriptors of the watershed's response (Saadi et al., 2019). As climate descriptor was considered the mean annual precipitation (PPT, mm) of the watersheds, which were obtained from the monthly satellite constellation of the Global Precipitation Measurement mission (GPM), IMERG Version 6 "Final Run" (http://mirador.gsfc.nasa.gov/com). The monthly products, with a spatial resolution of 0.1° (10 km), were added to result in annual products from January 1, 2001 to December 31, 2018. The annual products were averaged resulting in the mean annual precipitation and from it was extracted the mean value for each watershed. The rainfall in Paraná varies spatially, with an annual average between 1291 and 2237 mm (Figure 10a).

As a landscape descriptor, morphological watershed characteristics were considered such as the length of the drainage area (A, km2), watershed mean altitude (ALT, m), main thalweg length (TL, km), watershed mean slope (WS, %), and the main thalweg slope (TS, %). They were obtained from the Digital Elevation Model (DEM) in a Geographic Information System (GIS). Due to the long period of the data analyzed (>15 years) and the not mandatory coincidence between the period of watersheds data, using the land use as descriptors would be very difficult to include the land use as a descriptor and it was not considered in this study. A description of the 81 watershed characteristics is presented in the Appendix.

The DEM was generated using the products from the earth terrain mapping mission by the SRTM (Shuttle Radar Topography Mission). The DEM spatial resolution is 90 meters (https://www.cnpm.embrapa.br/projetos/relevobr/download/index.htm). The altitude in Paraná varies from 1 and 1830 m, where the north, west, and coastal represent the regions with low altitudes (Figure 10b).

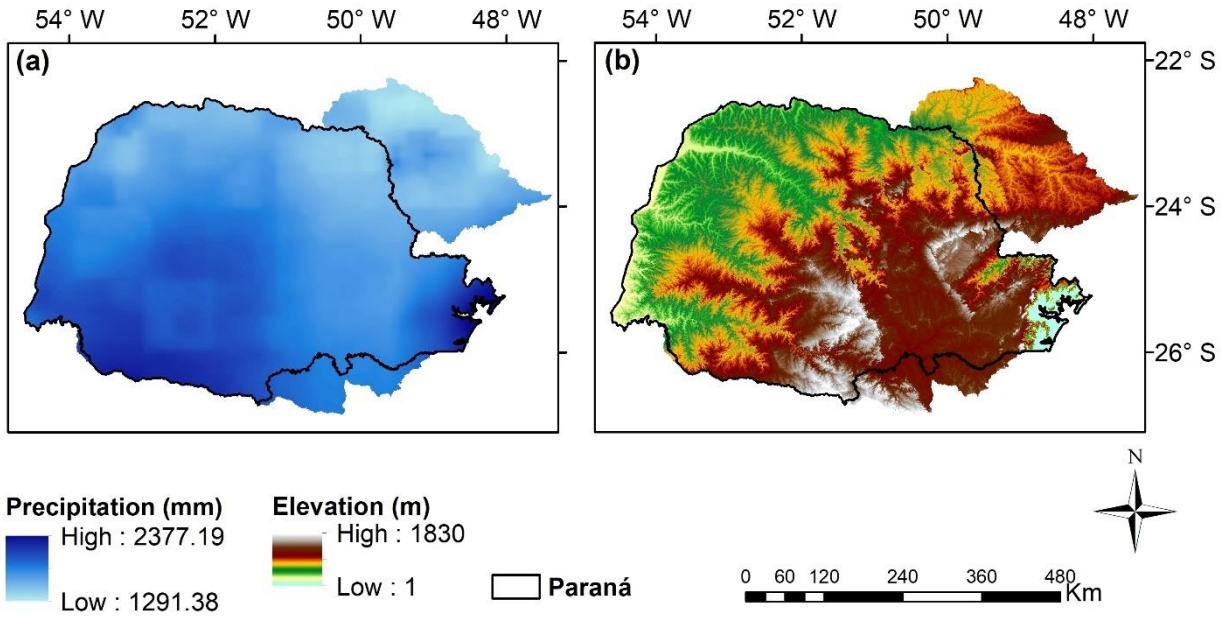

Figure 10. Mean annual precipitation (mm) and digital elevation model (m) in Paraná state, Brazil.

3.2.3. Prediction of hydrological variables in unmonitored watersheds

3.2.3.1. Multiple Regression (MLR) in homogeneous hydrological regions

The classical approaches of regionalization are calibrating the hydrological model with several watersheds and use linear regression to relate each model parameter to their attributes (Wagener and Wheater, 2006). In the regionalization of hydrological variables by multiple linear regressions (MLR), the Qm, Q90, and Q95 specifics were considered as dependent variables, and the climatic (PPT) and landscape descriptors (WS, TS, TL) as independent variables. The drainage area was not considered as a descriptor because we are considering the specific flows (flows/area). A general MLR model to estimate a hydrological parameter is represented by the Eq. 4:

$$Y = a_0 x_1^{a1} + x_2^{a2} + x_k^{ak} + \varepsilon \quad (4)$$

where Y is the dependent variable, x_i with i = 1, 2, …, k are the descriptors (independent variables), a_i i = 0, 1, 2, …, k are the regional coefficients, and ε is the model residuals.

Before applying the MLR method, the Paraná state was separated into hydrologically homogeneous regions or clusters. The clusters are regions with similar hydrological

characteristics and similar watershed response, which imply in the better performance of regionalization models. Several methodologies are found in the literature for the delimitation of homogeneous regions, such as the separation of regions considering geographical, political, or administrative limits, which despite being easily obtained, do not guarantee hydrological homogeneity (Rao and Srinivas, 2006). According to Smakhtin (2001), the flow responses tend to be similar in homogeneous regions in terms of climate, geology, topography, vegetation, and soil. Thereby, the models to transfer information between similar hydrological regions (regionalization) tend to present better performance.

The hydrological regions were obtained by cluster analysis in R software (R Core Team, 2019). We used the PPT, A, ALT, TL, WS, and TS as descriptors. The latitude (X) and longitude (Y) of the watershed centroids also were included to obtain clusters spatially continuous (Rao and Srinivas, 2006). Other studies included similar descriptors in the identification of homogeneous regions (Farsadnia et al., 2014; Elesbon et al., 2015; Silva, 2018).

The geographical distance between the watersheds was calculated between their centroids, to represent the runoff as a response of the total drainage area (Chebbi et al., 2017). We used the k-means algorithm and the Euclidean distance between the mentioned variables to obtain the clusters. The process consists first in the standardization of variables, using the Eq 5:

$$z = \frac{x - \mu}{\sigma} \tag{5}$$

where z is the standardized value, x is the value of the observed variable, μ is the mean, and σ is the standard deviation of the values in the data set.

The Euclidean distance in a hyperspace of the watersheds descriptors was calculated through the Eq. 6:

$$di,j = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{in} - x_{jn})^2} \tag{6}$$

where d is the distance between the watershed descriptors i and j in the Euclidean space and n is the number of independent variables.

The clustering was performed using the k-means algorithm, which divides the characteristic vectors into 'C' groups, minimizing an objective function through iteration to optimize the process, using the Eq. 7 (Hartigan and Wong, 1979, Beskow et al., 2016):

$$F = \sum_{i=1}^{n} \sum_{k=1}^{C} d(y_i; v_k) \quad (7)$$

where $d(y_i; v_k)$ is the Euclidean distance from the characteristic vector y_i to the centroid of cluster v_k, n is the number of characteristic vectors.

The k-means methodology requires the initial number of clusters definition. As this value is not previously known, the Elbow algorithm was used to estimate the optimal number of groups. This method consists of a visual graph analysis where the percentage of variance is explained as a function of the number of clusters. The graph shows the error decreasing as the number of cluster increase until it reaches a plateau when the increasing of clusters does not contribute with the error decreasing anymore. At this moment is possible to observe the optimal number of groups that are distant from each other, represented by an “elbow” of the curve (Bholowalia and Kumar, 2014).

The K-Nearest Neighbors (KNN) algorithm was used for spatializing the clusters in the Paraná, which predict a response variable Yt for each target geographical unit t (i.e. a pixel or a group of pixels) using the values Yi of the same variable measured in the field in locations corresponding to the k nearest neighbor units, though the Eq. 8 (Chirici et al., 2012):

$$Y_t = \frac{\sum_{i=1}^{k} W_{t,i} Y_i}{\sum_{i=1}^{k} W_{t,i}} \quad (8)$$

where the weight W is inversely proportional to the distance between the units t and i, calculated in the multidimensional feature space created on the feature space variables (Paraná).

The multidimensional distance was calculated by the Euclidean distance (Eq. 6), considering the feature space variability in the geographic units (pixels or groups of pixels). The package *kknn* (Schliep et al., 2016) was used for the implementation of the algorithm in the R software (R Core Team, 2019).

3.2.3.2. Random Forest

Random forest (RF) is a machine learning algorithm that predicts or classifies a dependent variable (or response variable) using a set of independent variables (or descriptors), based on a large number of classification and decision trees (Saadi et al., 2019). The RF method

is executed in four steps, according to Xu et al. (2019). First, the bootstrap method produces *ntree* (number of trees to grow) subset samples from the original training dataset. The second is the regression tree growth in each bootstrap sample, where randomly select samples with *mtry* independent variables at each splitting node and determine the optimal split based on this subset, and this process ends when the stopping criterion (nodesize) is reached. Third, the predictions over *ntree* decision trees are obtained through the mean of the dependent variable values at the leaf nodes for the individual tree. Lastly, the average of the prediction by *ntree* gives the final prediction.

In general, machine learning techniques are applied in hydrological big data studies, for example for daily and monthly scale predictions (Saadi et al., 2019, Rezaie-Balf et al., 2019). However, due to the importance of the reference flows, obtained from permanence curves built using a long period of data for water management, in this study we were interested if the RF method can be used to predict hydrological variables. The three main reasons why the RF was selected for predictions of the flows in Paraná are: the algorithm can handle highly correlated predictor variables, it can effectively handle small sample sizes, and it can capture nonlinear relationships between attributes and hydrological variables (Tyralis et al., 2019; Addor et al., 2018).

The original dataset containing 81 watersheds descriptors and dependent variables was divided into two groups: training (60 samples) and validation (21 samples). The training group was used to construct the RF model, while the validation set was used to test the performance of the RF model to predict the dependent variables at unseen locations. Usually, the selection of the training and validation groups is random. However, stratified sampling was used to ensure that in both datasets, samples from all regions of Paraná (north, south, east, and west) were considered in the construction and validation of the model.

The RF algorithm was implemented in R software using the *caret* package (Wing et al., 2019). In this study, the package default parameters were used, *ntree* = 500, with 25 bootstrap interactions. The *mtry* = 2 was selected aiming the optimal model using the smallest value of RMSE.

Besides the parameters, the variable importance (*varImp*) was used to describe the descriptor's importance in predicting the dependent variables. In the *caret* package, the variable importance for regression is based on the MSE for each tree, permuting the variable. Then, the differences are averaged and normalized by the standard error to give the importance value (Kuhn, 2012).

In the RF method, the PPT, WS, TS, TL, and ALT were used as descriptors. Also, we considered the mean of clay and sand content (%) to analyze the performance model improvement when using a greater number of descriptors. The soil descriptors were obtained from the Soil Geographic Databases compendium (https://soilgrids.org/), at a spatial resolution of 250 meters. For that purpose, we considered the clay and sand content average in the 0-100 cm depth. First analyzed was the RF performance using all the descriptors. Then, the best descriptor group was selected through cross-validation, removing each descriptor at a time.

3.2.3.3. Model Performance Analyses

For MLR classification regarding the performance of the model in each cluster, we used the Nash-Sutcliffe Efficiency (NSE), percent bias (pbias), and the index of agreement (d), as shown in Table 3. The equations, ranges, optimal values, and as well as advantages and disadvantages of the statistical measures used in this hydrological study, are presented by Moriasi et al. (2015).

Aiming to evaluate a general performance of the MRL model performance predicting the reference flows of the whole Paraná area, we used the coefficient of determination (R^2), Lin's concordance correlation coefficient (CCC), the mean absolute error (MAE), the root of the mean square error (RMSE), and statistical bias (bias).

The RF training performance was evaluated using the R^2, RMSE, and MAE, while RF validation performance was evaluated using the R^2, CCC, MAE, RMSE, and bias. The metrics were calculated using the *hydroGOF* (Zambrano-Bigiarini, 2020), *tdr* (Lamigueiro, 2018) and *ithir* (Malone, 2019) packages in R software. The equations for the metrics used are shown in Table 4.

Table 3. Criteria for assessing the statistical performance recommended for hydrological watershed models.

Parameters	**Performance**			
	Very good	**Good**	**Satisfactory**	**Not Satisfactory**
NSE	$> 0{,}80$	$0{,}70 < \text{NSE} \leq 0{,}80$	$0{,}50 < \text{NSE} \leq 0{,}70$	$\leq 0{,}50$
pbias (%)	$< \pm 5$	$\pm 5 \leq \text{pbias} \leq \pm 10$	$\pm 10 \leq \text{pbias} \leq \pm 15$	$\text{pbias} \geq \pm 15$
d	$> 0{,}90$	$0{,}85 < \text{d} \leq 0{,}90$	$0{,}75 < \text{d} < 0{,}85$	$\text{d} \leq 0{,}75$

Table 4. Summary of statistical indices used to evaluate the satellite precipitation products.

Index	Unit	Equation *	Best value
Correlation coefficient (CCC)	-	$CCC = \frac{2\,\rho\,\sigma_O\,\sigma_S}{\sigma_O^2 + \sigma_S^2 + (\bar{O} - \bar{S})^2}$	1
Determination coefficient (R^2)	-	$R^2 = \frac{\{\sum_{i=1}^{n}(S_i - \bar{S})(O_i - \bar{O})\}^2}{\sum_{i=1}^{n}(S_i - \bar{S})^2 \sum_{i=1}^{n}(O_i - \bar{O})^2}$	1
Mean absolute error (MAE)	mm	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert S_i - O_i\rvert$	0
Root of the mean square error (RMSE)	mm	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)^2}$	0
Nash-Sutcliffe Efficiency (NSE)	-	$NSE = 1 - \frac{\sum(O_i - S_i)^2}{\sum(O_i - \bar{O})^2}$	1
percent bias (pbias)	-	$pbias = \left[\frac{\sum_{i=1}^{n}(O_i - S_i)}{\sum_{i=1}^{n} O_i} * 100\right]$	0
index of agreement (d)	-	$d = 1 - \frac{\sum(S_i - O_i)^2}{\sum(\lvert S_i - \bar{O}\rvert + \lvert O_i - \bar{O}\rvert)^2}$	1

* where, Oi is the observed flow of hydrometric gauges of order i; Si is the estimated order flow (MRL and RF) of order i, $\bar{O}$ and $\bar{S}$ are the average of observed and estimated flow by hydrometric gauges and models (MRL and RF), respectively; ρ Pearson's correlation coefficient; σ_O and σ_S are the standard deviations of observed and estimated flow.

3.2.4. Results and Discussion

3.2.4.1. Multiple Linear Regression (MLR) models in hydrological regions

The Elbow algorithm indicated five clusters in Paraná, considering the climate and landscape descriptors variables. Thus, using the k-means methodology, the Paraná was grouped in five hydrologically homogeneous, as showed in Figure 11.

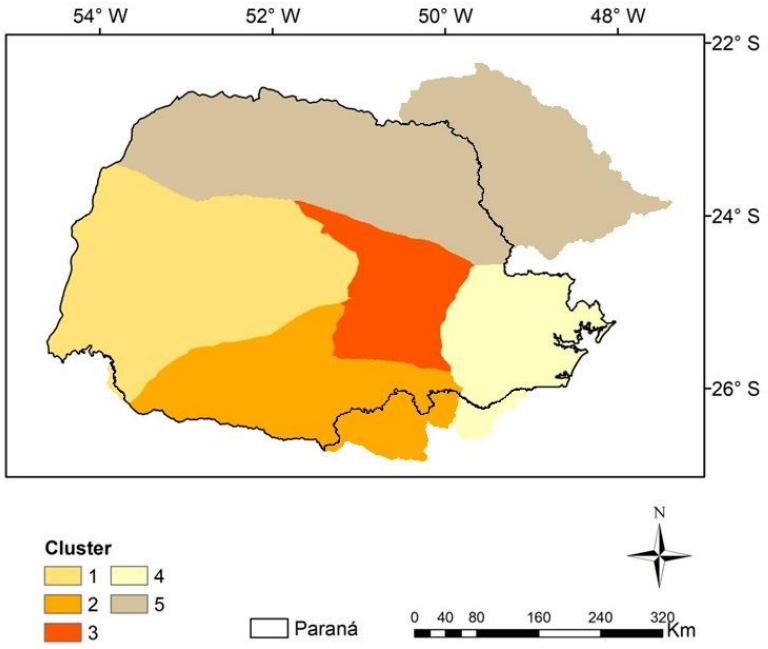

Figure 11. Hydrological homogeneous regions (clusters) in Paraná state, Brazil.

In hydrological regionalization of many watersheds distributing in a large area, identifying homogeneous regions is usually the most difficult part of the analysis and it requires a subjective judgment (Farsadnia et al., 2014). The spatial delimitation was coincidently also the spatial distribution of the Paraná geographic regions (north, south, central, east, and west) and displayed coherence with the mean annual precipitation and digital elevation model maps (Figure 10).

In the north and west of Paraná, lower mean annual precipitation, flattened landscapes with lower altitudes, and physiographic characteristics of the watersheds, were grouped in clusters 1 and 5. Cluster 2, in the south of the state, presented high altitudes and high mean annual precipitation. In cluster 3, the altitude and precipitation were high and homogeneous along the area. The cluster 4 considered the similar characteristics of the climate and landscape descriptor of the watersheds near to the coastal region but located in the higher altitude areas. In this cluster, the mean annual precipitation is high close to the ocean and decreases towards the continent.

A similar method was adopted by Silva (2018) in clustering hydrologically homogenous regions in Paraná. The authors considered as dependent variables Q7,10, Qm, Q90, and Q95, and the watersheds morphometric characteristics as descriptors. They identified

four clusters in the state, using the Fuzzy C-Means, but the spatial continuity was not prioritized as in this study.

Table 5 shows the MLR models and their performances for Q95, Q90, and Qm predictions. In all clusters, the MLR models showed good performance. The NSE is considered as a robust method and it can incorporate measurement uncertainty (Moriasi et al., 2015), which can explain the lower performance classified by this method. The index d is a standardized measure of the degree of model prediction error, and it showed for all clusters the predicted flows were close to the observed value in the gauges, as the perfect agreement is 1.

Besides the information about the performance, the pbias shows the percentage of the model over or underestimated, when it presents positive and negative values, respectively. Only in cluster 5 all the models underestimate the flows, in the other clusters the models showed heterogeneity in terms of pbias. It is important to highlight that a good performance in each cluster does not imply that the MLR models will perform similarly outside of the region in with they were fitted, which requires attention when using the equations presented in this study.

Table 5. Summary of the performance metrics of MLR models to estimate the Q95, Q90, and Qm in the clusters.

Cluster	Variable	NSE	pbias	d	MLR models
1	Q95	0.744	-0.104	0.921	$Q95_1 = 3.94x10^{-2} - 1.68\ x10^{-5}\ PPT - 9.20\ x10^{-4}\ WS + 3.58\ x10^{-3}\ TS^{*}$
1	Q90	0.736	-0.110	0.918	$Q90_1 = 4.95x10^{-2} - 2.13\ x10^{-5}\ PPT - 1.07\ x10^{-3}\ WS + 4.23\ x10^{-3}\ TS^{*}$
1	Qm	0.681	0.139	0.896	$Qm_1 = -0.28 + 1.06\ x10^{-4}\ PPT^{*}$
2	Q95	0.792	-0.016	0.939	$Q95_2 = -5.76\ x10^{-2} + 2.26\ x10^{-5}\ PPT + 2.09\ x10^{-3}\ WS + 1.59\ x10^{-8}\ TS^{*}$
2	Q90	0.740	0.049	0.921	$Q90_2 = -7.18\ x10^{-2} + 3.01\ x10^{-5}\ PPT + 2.31\ x10^{-3}\ WS + 1.67\ x10^{-8}\ TS^{*}$
2	Qm	0.647	-0.091	0.887	$Qm_2 = -0.199 + 9.68\ x10^{-5}\ PPT + 4.69\ x10^{-3}\ WS^{*}$
3	Q95	0.885	0.122	0.969	$Q95_3 = 6.44\ x10^{-2} - 3.20\ x10^{-5}\ PPT - 1.02\ x10^{-3}\ WS + 3.68\ x10^{-3}\ TS - 9.23\ x10^{-9}\ TL^{*}$
3	Q90	0.909	0.050	0.976	$Q90_3 = 5.91\ x10^{-2} - 2.74\ x10^{-5}\ PPT - 1.25\ x10^{-3}\ WS + 4.34\ x10^{-3}\ TS - 1.01\ x10^{-8}\ TL^{*}$
3	Qm	0.780	-0.152	0.934	$Qm_3 = -0.139 + 1.04\ x10^{-4}\ PPT - 2.35\ x10^{-3}\ WS^{*}$
4	Q95	0.584	0.083	0.853	$Q95_4 = -3.55\ x10^{-2} + 2.10\ x10^{-5}\ PPT + 1.91\ x10^{-3}\ TS^{*}$
4	Q90	0.625	0.015	0.873	$Q90_4 = -5.79\ x10^{-2} + 3.34\ x10^{-5}\ PPT + 2.66\ x10^{-3}\ TS^{*}$
4	Qm	0.724	-0.083	0.917	$Qm_4 = -0.388 + 2.06\ x10^{-4}\ PPT + 4.16\ x10^{-3}\ WS^{*}$
5	Q95	0.821	-0.540	0.948	$Q95_5 = -3.42\ x10^{-2} + 1.99\ x10^{-5}\ PPT - 1.05\ x10^{-4}\ WS + 3.31\ x10^{-3}\ TS + 1.42\ x10^{-8}\ TL^{*}$
5	Q90	0.759	-0.790	0.927	$Q90_5 = -3.59\ x10^{-2} + 2.07\ x10^{-5}\ PPT + 1.27\ x10^{-5}\ WS + 3.53\ x10^{-3}\ TS + 1.39\ x10^{-8}\ TL^{*}$
5	Qm	0.775	-0.252	0.932	$Qm_5 = -2.91\ x10^{-2} + 1.87\ x10^{-5}\ PPT + 1.51\ x10^{-3}\ WS^{*}$

* each descriptor variable in the equation was statistically significant at 90% confidence interval, in the p-value test. NSE = Nash Sutcliffe Efficiency, d = index of agreement, pbias = percent bias, MLR models = Multiple Linear Regressions equations for estimate the dependent variables, PPT = mean annual precipitation (mm), WS = watershed mean slope (%), TS = main thalweg slope (%), and TL = length of the main thalweg (km).

As mentioned before, the descriptors variables used were selected aiming the best performance of the models in each cluster and considering a statistical significance of 90% confidence interval, in *p-value* test. Thus, in each MLR model, a group of descriptors was selected to estimate the dependent variable. In the global scenario, the mean annual precipitation was considered as a common explicative variable, present in all MLR models. In fact, the precipitation is a descriptor directly related with the Qm, Q90, and Q95, since all flows in watersheds derive ultimately from precipitation (Poof et al., 1997).

The flow in watersheds also represents a response of the physical factors including climate, geology, topography, land use, all of which can impact the magnitude and timing of flow during and after precipitation events (Poof et al., 1997). Therefore, the landscape descriptors were important to explain the Q95, Q90, and Qm, where the WS slope was the most common in the MLR models, followed by the TS and TL.

Table 6 shows the concordance and error indexes of predicted flows when considering the predicted flows by the MLR models in all clusters ($Q95_{all}$, $Q90_{all,}$ and Qm_{all}) as a unique dataset. to describe the applicability and performance of the MLR in the study area. In this analysis, the minimum ($Q95_{all}$ and $Q90_{all}$) and the long-term average (Qm_{all}) predicted flows that presented good collinearity with the values observed in the gauges, with R^2 and CCC higher than 0.74 and 0.83, respectively.

Table 6. Summary of error and statistics metrics of the MLR models to estimate Q95, Q90, and Qm in all clusters.

Cluster	Flow	R^2	CCC	MAE	RMSE	bias
All	$Q95_{all}$	0.741	0.840	$1.17E^{-3}$	$1.64E^{-3}$	$2.52E^{-6}$
All	$Q90_{all}$	0.740	0.835	$1.50E^{-3}$	$2.17E^{-3}$	$8.89E^{-8}$
All	Qm_{all}	0.749	0.845	$7.51E^{-3}$	$1.27E^{-2}$	$1.40E^{-5}$

R^2= coefficient of determination, CCC= concordance correlation coefficient, MAE= Mean Absolute Error ($m^3\ s^{-1}\ km^{-2}$), RMSE= Root Mean Square Error ($m^3\ s^{-1}\ km^{-2}$), bias= statistical bias ($m^3\ s^{-1}\ km^{-2}$).

The dispersion of observed and predicted values is presented in Figure 12. The positive bias suggest that the models slightly overestimate the flows when the flows prediction in each cluster in a unique data analysis was considered (Table 5). In a global scenario, the MLR models tend to underestimate the flows higher than 0.020 $m^3\ s^{-1}\ km^{-2}$ for the Qm_{all}, $Q90_{all}$ and $Q95_{all}$, and overestimate values lower than this rate (Figure 12).

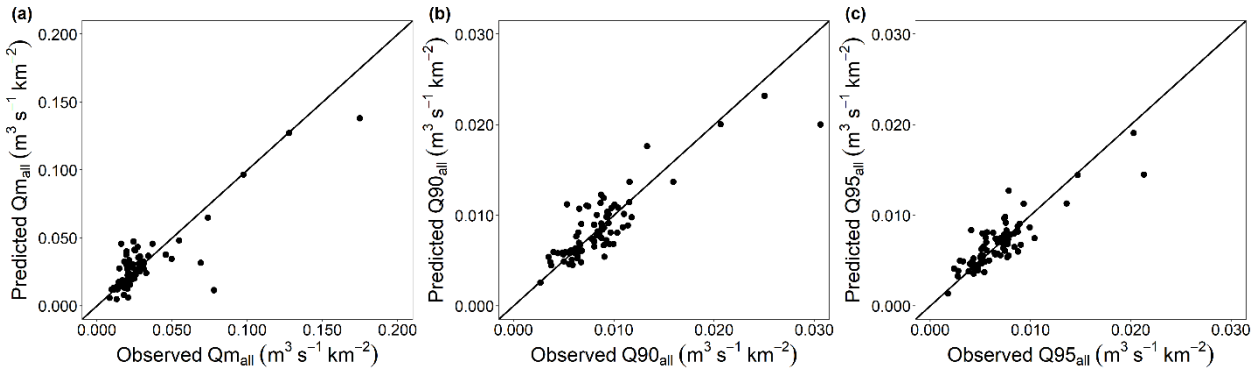

Figure 12. Regressions between predicted flows by MLR models and data observed by gauges, considering the prediction in all clusters as a unique dataset. (a) Qm is the long-term average flow, and (b) Q90 and (c) Q95 are the flows exceeded or equaled in 90% and 95% of the time.

The small value of MAE and RMSE (< 7.51E^{-3} and 2.17E^{-3}, respectively) in the global scenario, reflects the good performance of the MLR models in estimating the flow in each cluster, as mentioned before. The slightly higher errors of the Qm_{all} compared to the $Q90_{all}$ and $Q95_{all}$ can be explained by the difference in the magnitude of the long-term average flow when compared to the minimum flows.

The good performance of MLR models to predict hydrological variables also was verified in several hydrological studies (for example, Lelis et al., 2020, Silva et al., 2019, Lopes et al., 2017, Swain and Patra, 2017, and Boscarello et al., 2016). The main explanation for MLR models' success is the application of the method within homogeneous regions, where most watersheds characteristics are similar, and the fact that linear function generally provides a good approximation to regional models (Li et al., 2010).

3.2.5. Random forest (RF)

The RF was applied to predict flows from a long-term dataset average and permanence curve, based on watershed descriptors. First, we considered all (A) descriptors to analyze the RF training and validation performance. Then, we selected the best (B) group of descriptors to predict the Qm, Q90, and Q95 in the RF, removing one descriptor at a time in cross-validation. The errors and performance metrics of the RF prediction results in training (calibration) and validation, using all (A) and best (B) descriptors are shown in Table 7.

Table 7. Summary of error and performance metrics of the RF training and validation to predict the Q95, Q90, and Qm in Paraná.

Train						
	R^2		MAE		RMSE	
Flow	**A**	**B**	**A**	**B**	**A**	**B**
$Q95_{train}$	0.875	0.887	$1.01x10^{-3}$	$1.05x10^{-3}$	$8.06x10^{-3}$	$1.67x10^{-3}$
$Q90_{train}$	0.890	0.895	$1.25x10^{-3}$	$1.26x10^{-3}$	$2.18x10^{-3}$	$2.14x10^{-3}$
Qm_{train}	0.905	0.903	$4.42x10^{-3}$	$4.36x10^{-3}$	$1.64x10^{-3}$	$7.55x10^{-3}$

Validation										
	R^2		CCC		MAE		RMSE		bias	
Flow	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**
$Q95_{val}$	0.483	0.454	0.603	0.464	$1.81x10^{-3}$	$1.43x10^{-3}$	$6.70x10^{-3}$	$1.77x10^{-3}$	$3.42x10^{-3}$	$1.49x10^{-4}$
$Q90_{val}$	0.115	0.463	0.349	0.460	$2.02x10^{-3}$	$1.49x10^{-3}$	$2.46x10^{-3}$	$1.97x10^{-3}$	$4.32x10^{-4}$	$2.19x10^{-4}$
Qm_{val}	0.056	0.521	0.267	0.590	$4.43x10^{-3}$	$4.61x10^{-3}$	$2.26x10^{-3}$	$7.13x10^{-3}$	$2.93x10^{-4}$	$4.25x10^{-3}$

A is the result considering group with all descriptors in the RF; B is the result considering the group of best descriptors in the RF. $Q95_{train}$, $Q90_{train}$, and Qm_{train} are the predicted flow in the training ($m^3\ s^{-1}\ km^{-2}$), and $Q95_{test}$, $Q90_{test}$, and Qm_{test} are the predicted flow in the validation ($m^3\ s^{-1}\ km^{-2}$). R^2= coefficient of determination, MAE= Mean Absolute Error ($m^3\ s^{-1}\ km^{-2}$), CCC= concordance correlation coefficient, RMSE= Root Mean Square Error ($m^3\ s^{-1}\ km^{-2}$), bias= statistical bias ($m^3\ s^{-1}\ km^{-2}$).

When all (A) descriptors were considered in flow predictions, we noticed a good performance of the train over the 60 watersheds, with $R^2 > 0.87$, and lower error metrics (MAE < 4.42E-3 and RMSE < $8.06E^{-3}$ $m^3\ s^{-1}\ km^{-2}$). However, regarding the validation performance, the group A of descriptors exhibited a very low ability to predict the flows outside the training domain (over the 21 watershed samples), especially for the $Q90_{val}$, and $Q95_{val}$ ($R^2 \leq 0.115$ and CCC ≤ 0.349).

The group of the best (B) predictors comprised the descriptors: WS, TS, and ALT as landscape, PPT as climatic, and X and Y as watersheds spatial centroids coordinates. In this case, the RF training resulted in a very good prediction of the average and minimum flows in the samples. The descriptors could explain more than 88% ($R^2 > 0.88$) of the Qm_{train}, $Q90_{train}$, and $Q95_{train}$ in Paraná, with MAE and RMSE lower than $4.61E^{-3}$ and $7.55E^{-3}$ $m^3\ s^{-1}\ km^{-2}$, respectively.

Additionally, the RF built using the group B of descriptors, could explain more than 45% ($R^2 > 0.45$) of the Qm_{val}, $Q90_{val}$, and $Q95_{val}$ in the validation, with a CCC ≥ 0.460. Similar performance was observed in others hydrological studies using RF, for example, Saadi et al. (2019) regionalizing hourly hydrological parameters, Addor et al. (2018) predicting watersheds

hydrological signatures, Carlisle et al. (2010) predicting hydrological parameters in regional and national scales, and Schnier and Cai (2014) predicting complete flow duration curves.

The R^2 and CCC metrics evidencing a superior predictive power of the RF when using the group B of descriptors, and the regionalization capability of the RF models to predict the reference flows in unmonitored watershed. Thus, we considered the RF built with group B of descriptors for the regionalization of the Qm_{val}, $Q90_{val,}$ and $Q95_{val}$ in Paraná, and only the results obtained from these models are discussed hereafter.

Corroborating the better performance of RF models when filtering the descriptors, Xu et al. (2019) found a similar or better RF predictive performance using a subset of descriptors, in the prediction of the water rights price, in the United States. According to the authors, this result proves the ability of the ensemble bagging algorithm to obtain better predictive performance, and the advantage of filtering out the descriptors, due that the algorithm is robust to work with a low number of variables.

Regarding a practical application of the RF in the water resources management, the algorithm's ability to perform similarly or better using a less number of descriptors variable could be advantageous, making the prediction of hydrological variables easier and faster, especially in sites where limited information is available.

In general, a slightly better performance was observed in the prediction of the Qm in terms of the R^2 compared to Q90 and Q95. A hypothesis for the higher performance is that Qm was obtained by averaging annual values of long-term data, while Q90 and Q95 were obtained from permanence curves, adjusted for each watershed. In this way, the average flows prediction could perform better compared to the minimum flows since the average amortizes extreme values.

The relative importance of individual variables, the dispersion between estimated and observed data, as well as the Qm, Q90, and Q95 density curves in the RF (using group B descriptors) are presented in Figure 13. The positive bias values (Table 7) for the Qm_{val}, $Q90_{val,}$ and $Q95_{val}$ ($4.25E^{-3}$, $2.19E^{-4}$, and $1.49E^{-4}$ $m^3\,s^{-1}\,km^{-2}$, respectively) demonstrated a general overestimation of RF models to predict the flows analyzed. In fact, this behavior can be observed in Figure 13, where RF overestimated almost all values of Qm. For Q90 and Q95 a tendency to overestimate the values lower than 0.075 $m^3\,s^{-1}\,km^{-2}$ and to underestimate higher values was noticed.

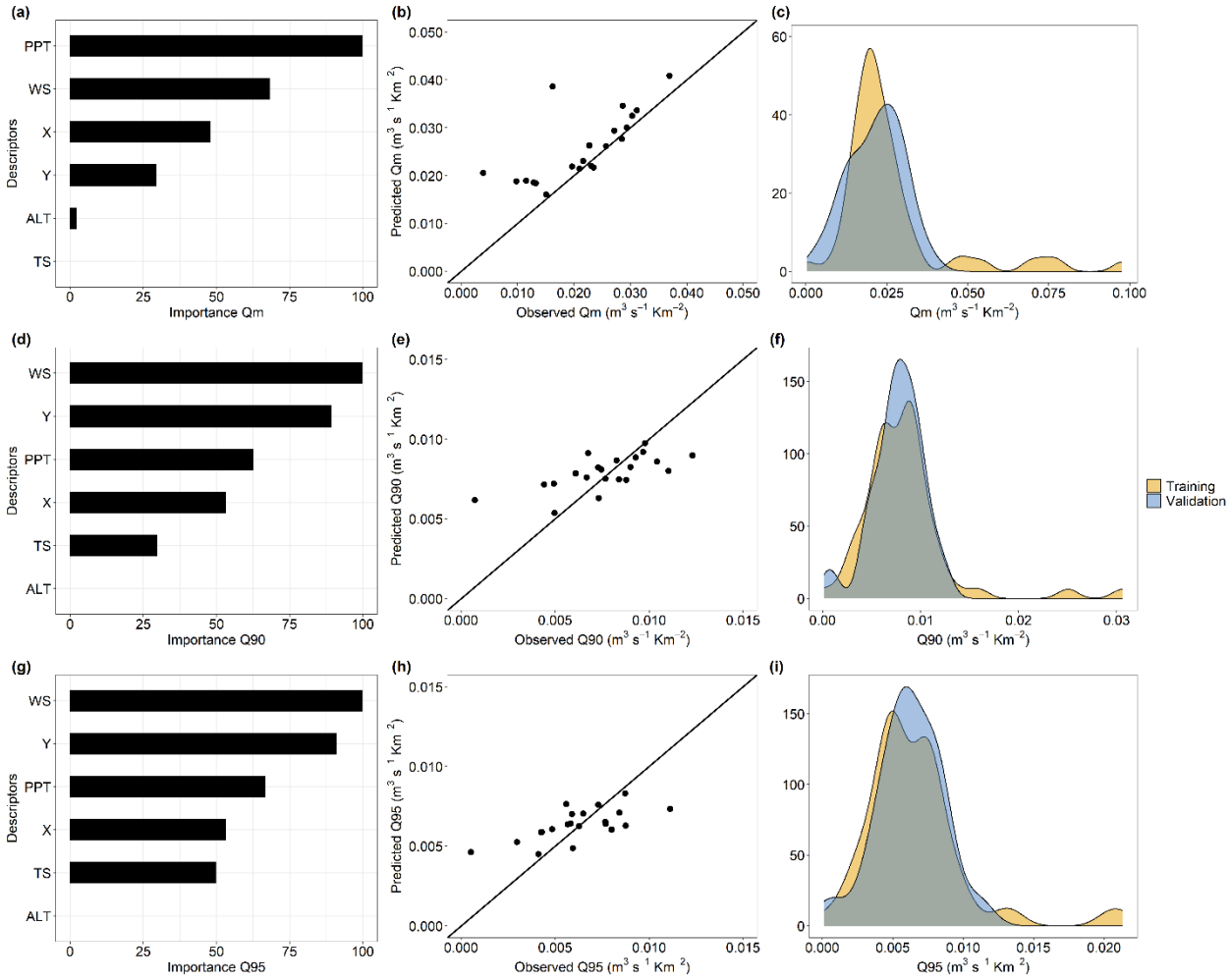

Figure 13. Relative importance of descriptors (a, d, and g), dispersion between the estimated and observed data (b, c, and h), and density curves in the RF training and validation (c, f, and i), in the Qm, Q90, and Q95 prediction, respectively. PPT = mean annual precipitation (mm), WS = watershed mean slope (%), TS = main thalweg slope (%), and ALT = watershed mean altitude (m), X = watershed centroid's latitude, and Y = watershed centroid's longitude.

In a general analysis, the climatic (PPT), landscape (WS), and the centroid spatial coordinates (X and Y) were the descriptors more representative in the flow prediction. For Qm, the PPT was the most important descriptor (100%), followed by WS (68.33%), X (48.06%), Y (29.42%), and ALT (2.25%). The TL (0%) was not representative to describe the Qm in RF.

The ranking of the important variables was the same, in sequence from the most important to less important for the Q90 and Q95: WS (100% for both), Y (89.42 and 91.09%, respectively), PPT (62.59 and 66.67%, respectively), X (53.28% for both), and TS (29.90 and 49.87%, respectively). For the minimum flows, the ATL was not a significant descriptor in RF. The greater importance of precipitation compared to landscape descriptors for the average compared to minimum flows in the RF model was also observed by Carlisle et al. (2010).

Besides RF define the variable importance, interpreting the relative importance of each descriptor is not an easy task due to the nature of both dependent variables and descriptors and their inter-correlations (Saadi et al., 2019).

The relevant importance of PPT and WS descriptors in the Qm, Q90, and Q95 prediction is coherent with the results in the MLR model prediction. As aforementioned, the importance of the PPT to the flows is obvious due to the flows is a result of the rainfall. However, the PPT influence the average and minimum flows differently and participated with importance of individual variables ranking prediction.

Among several flow-rainfall interactions, the process of obtaining the flows seems reasonable to assist in understanding why PPT is the first in the Qm ranking of important descriptors and third in the Q90 and Q95 ranking. The Qm was obtained by averaging the annual values over the data period. Thus, it includes the flow measured during the wet season, when PPT effectively affects the flow (Belhassan, 2011). In contrast, Q90 and Q95 are minimum flows and usually occur during the dry season, when the groundwater is, in general, the primary source of water in the watersheds (Belhassan, 2011, Gilfedder, et al., 2012). In this scenario, the WS was the most important descriptor, and its greater relevance when compared to the other physical factors, could be because it represents an average of the whole drainage area.

The density curves, which relate the flows to the observation frequency (Figure 13), had a similar distribution in the training and validation. However, the frequency peak in the training was concentrated close to 0.020 $m^3 s^{-1} km^{-2}$, while in validation was close to 0.025 $m^3 s^{-1} km^{-2}$ in the Qm prediction by the RF model. Thus, the training in samples with a greater peak in lower flow rates could explain the overestimation in the validation prediction, as the predictive power of RF completely depends on the training experience (Schoppa et al., 2020).

In contrast, the peaks in the validation were between two peaks observed in the training in the prediction of the Q90 and Q95, what could explain the under and overestimation of Q90 and Q95 predicted flows by the RF.

3.2.5.1. General analysis of MLR and RF in flow prediction

In a global analysis, the MLR outperformed RF in predicting the Qm, Q90, and Q95, with $R^2 > 0.74$ and CCC > 0.83 compared to $R^2 > 0.45$ and CCC > 0.46, respectively. The main hypothesis for the MLR outperformance is that the model was applied in hydrological homogeneous regions, where the flows tend to present similar responses and can be explained by linear relationships with the descriptors.

Regarding the error metrics, RF presented very similar values of MAE and RMSE in the Q95 prediction, and lower values of the statistics errors in the Qm and Q90 prediction compared to the MLR. Thus, the error metrics could support using the RF model as a tool to predict hydrological variables in sites with a density and frequency limited of information.

The MLR and RF resulted in positive bias, which reflects a slightly overestimation by the models. When a model that underestimates the flow is used for water management, it contributes to the conservation of the water resources (Lelis et al., 2020), because it shows a lower volume of disposable water to be used. Thus, when applying both models to the water resources management, the hydrologist should consider this in the hydrological analysis in Paraná.

As main advantages cited by Tyralis et al. (2019) about using RF in hydrology, four of them were very relevant in our study: (1) RF does not require the previous hydrological homogeneous clustering, which requires efforts of hydrologists, besides it consists in a subjective process; (2) RF is a robust method and there are many packages to use in free software; (3) RF can work with a small number of descriptors and samples and; (4) RF can handle highly correlated predictor variables. Thus, besides MLR presented better performance, the error metrics, and the advantages mentioned support RF for predicting reference flows and indicate this method for hydrological studies.

3.3. Conclusions and final considerations

The prediction of hydrological variables is essential for water resources management, especially in developing countries were the gauges data is limited by the density and frequency. We analyzed the MLR and RF methods to predict the long-term average (Qm) and minimum flows from permanence curves (Q90 and Q95) in the Paraná state, Brazil. For that purpose, the climatic and landscape descriptors were used.

MLR is the most traditional method to predict the flow in unmonitored sites. In general, it performs better when applied in clusters, where the hydrological variables respond similarly. We found five clusters in Paraná and the MLR models performed very well predicting the reference flows.

RF is a machine learning algorithm that is being increasingly applied for different purposes in Hydrology. As advantages, the method does not require the delimitation of homogeneous regions, and the RF was applied considering the whole Paraná's area. In this method, WS, TS, and ALT as landscape, PPT as climatic, and X and Y as watersheds spatial

centroids coordinates represented the group of descriptors that resulted in RF best prediction considering the dataset and study area. RF performance could be improved by increasing the number of watersheds samples and/or data sample (e.g., decreasing the time scale of flow analysis for monthly or daily). Additionally, the RF could show better performance if package parameters were modified. Thus, further studies could focus on that purpose.

Besides MLR overperformed RF in this study, the error metrics (MAE and RMSE) were very similar in both methods for the Q95 prediction, and for the Qm and Q90 prediction, RF resulted in lower values compared to MLR. Thus, RF in the references flows prediction is promising. Due to its low-cost, RF, in terms of time of execution and free software implementation, could be applied as an encouraging tool for the reference flows prediction.

References

Addor, N., Nearing, G., Prieto, C., Newman, A. J., Le Vine, N., Clark, M. P. (2018). A ranking of hydrological signatures based on their predictability in space. Water Resources Research, 54(11), 8792-8812. https://doi.org/10.1029/2018WR022606

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

ANA. Conjuntura dos recursos hídricos no Brasil 2017; relatório pleno/ Agência Nacional de Águas. Brasília. 2017. Available online on http//www.ana.gov.br Assessed in 03/08/2020.

Aparecido, L.E.D.O., Rolim, G.D.S., Richetti, J., Souza, P. S. D., Johann, J. A. (2016). Köppen, Thornthwaite and Camargo climate classifications for climatic zoning in the State of Paraná, Brazil. Ciência e Agrotecnologia, 40(4), 405-417. https://doi.org/10.1590/1413-70542016404003916.

Belhassan, K. (2011). Relationship between river flow, rainfall and groundwater pumpage in Mikkes Basin (Morocco). Iranian Journal of Earth Sciences, 3(2), 98-107.

Beskow, S., de Mello, C. R., Vargas, M. M., Corrêa, L. D. L., Caldeira, T. L., Durães, M. F., & de Aguiar, M. S. (2016). Artificial intelligence techniques coupled with seasonality measures for hydrological regionalization of Q90 under Brazilian conditions. Journal of Hydrology, 541, 1406-1419. https://doi.org/10.1016/j.jhydrol.2016.08.046

Bholowalia, P., Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. International Journal of Computer Applications, *105*(9). https://doi.org/10.5120/18405-9674

Booker, D. J., Snelder, T. H. (2012). Comparing methods for estimating flow duration curves at ungauged sites. Journal of Hydrology, *434*, 78-94. https://doi.org/10.1016/j.jhydrol.2012.02.031

Booker, D. J., Woods, R. A. (2014). Comparing and combining physically based and empirically-based approaches for estimating the hydrology of ungauged catchments. Journal of Hydrology, 508, 227-239. https://doi.org/10.1016/j.jhydrol.2013.11.007

Boscarello, L., Ravazzani, G., Cislaghi, A., Mancini, M. (2016). Regionalization of flow-duration curves through catchment classification with streamflow signatures and physiographic–climate indices. Journal of Hydrologic Engineering, 21(3), 05015027. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001307

Carlisle, D. M., Falcone, J., Wolock, D. M., Meador, M. R., Norris, R. H. (2010). Predicting the natural flow regime: models for assessing hydrological alteration in streams. River Research and Applications, 26(2), 118-136. https://doi.org/10.1002/rra.1247

Carmello, V., Sant'anna Neto, J.L., (2016). Rainfall variability and soybean yield in Paraná State, Southern Brazil. International Journal of Environmental & Agriculture Research 2 (1), 86-97.

Chang, W., Chen, X. (2018). Monthly rainfall-runoff modeling at watershed scale: a comparative study of data-driven and theory-driven approaches. Water, 10(9), 1116. https://doi.org/10.3390/w10091116

Chebbi, A., Bargaoui, Z. K., Abid, N., & da Conceição Cunha, M. (2017). Optimization of a hydrometric network extension using specific flow, kriging, and simulated annealing. Journal of Hydrology, 555, 971-982. https://doi.org/10.1016/j.jhydrol.2017.10.076

Chirici, G., Corona, P., Marchetti, M., Mastronardi, A., Maselli, F., Bottai, L., Travaglini, D. (2012). K-NN FOREST: a software for the non-parametric prediction and mapping of environmental variables by the k-Nearest Neighbors algorithm. European Journal of Remote Sensing, 45(1), 433-442. https://doi.org/10.5721/EuJRS20124536

E.S.R.I. ArcGIS 10.2. 2 for Desktop. Environmental Systems Research Institute, Redlands, CA, USA (2014).

Elesbon, A.A., Silva, D.D.D., Sediyama, G.C., Guedes, H. A., Ribeiro, C.A., Ribeiro, C.B.D.M. Multivariate statistical analysis to support the minimum streamflow regionalization. Engenharia Agrícola, 35(5), 838-851. https://doi.org/10.1590/1809-4430

Farsadnia, F., KaMLRood, M. R., Nia, A. M., Modarres, R., Bray, M. T., Han, D., Sadatinejad, J. (2014). Identification of homogeneous regions for regionalization of watersheds by two-level self-organizing feature maps. Journal of Hydrology, *509*, 387-397. https://doi.org/10.1016/j.jhydrol.2013.11.050

Gilfedder, M., Rassam, D. W., Stenson, M. P., Jolly, I. D., Walker, G. R., Littleboy, M. (2012). Incorporating land use changes and surface–groundwater interactions in a simple catchment water yield model. Environmental Modelling & Software, 38, 62-73. DOI https://doi.org/10.1016/j.envsoft.2012.05.005

Harris, N. M., Gurnell, A. M., Hannah, D. M., & Petts, G. E. (2000). Classification of river regimes: a context for hydroecology. Hydrological Processes, 14(16-17), 2831-2848. https://doi.org/10.1002/1099-1085

Hartigan, J. A., Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100-108. DOI https://doi.org/10.2307/2346830

Hobouchian, M. P., Salio, P., Skabar, Y. G., Vila, D., Garreaud, R. (2017). Assessment of satellite precipitation estimates over the slopes of the subtropical Andes. Atmospheric Research, *190*, 43-54. https://doi.org/10.1016/j.atmosres.2017.02.006

Hosking, J. R. M.; Wallis, J. R. Regional Frequency Analysis: An Approach Based on L-moments. 1997. Cambridge University Press, UK, 1997.

IBGE- Instituto Brasileiro de Geografia e Estatística. Sinopse do Censo Demográfico 2010. Rio de Janeiro, 2011. Uploded in 2019. Available online on: https://www.in.gov.br/en/web/dou/-/resolucao-n-3-de-26-de-agosto-de-2019-212912380. Assessed in 08/05/2020.

IGAM. Minas Gerais Water Management Institute. Estudo de regionalização de vazão para o aprimoramento do processo de outorga no Estado de Minas Gerais. Grupo de Pesquisas em Recursos Hídricos da UFV. Belo Horizonte, 2012.

Junger, W.; Leon, P. Multivariate time series data imputation. R package version 0.3.5. 2018. <https://CRAN.R-project.org/package=mtsdi>. Assessed in 10/01/2019.

Knn S. K., Hechenbichler K., Lizee A. 2016. Knn: Weighted k-Nearest Neighbors. Available online on: https://github.com/KlausVigo/kknn. Assessed in 08/05/2020.

Kuhn, M. (2012). Variable importance using the caret package. Journal of Statistical Software. Available online on http://btr0x2.rz.unibayreuth.de/math/statlib/R/CRAN/doc/vignettes/caret/caretVarImp.pdf. Assessed in 09/27/2020.

Lamigueiro O. P. tdr: Target Diagram, 2018. Available online on: http://github.com/oscarperpinan/tdr. Assessed in 08/15/2020.

Lelis, L. C. S., Nascimento, J. G., Duarte, S. N., Pacheco, A. B., Bosquilia, R. W. D., Wolff, W. (2020). Assessment of hydrological regionalization methodologies for the upper Jaguari River basin. Journal of South American Earth Sciences, *97*, 102402. https://doi.org/10.1016/j.jsames.2019.102402

Li, M., Shao, Q., Zhang, L., Chiew, F. H. (2010). A new regionalization approach and its application to predict flow duration curve in ungauged basins. Journal of Hydrology, 389(1-2), 137-145. https://doi.org/10.1016/j.jhydrol.2010.05.039

Liazi, A.; Conejo, J. L.; Palos, J. C. F.; Cintra, P. S. (1988). Regionalização Hidrológica no Estado de São Paulo. São Paulo: Revista Águas e Energia Elétrica – DAEE, 5(14), 4-10.

Loh, W. Y. (2011). Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(1), 14-23.14–23. https://doi.org/10.1002/widm.8

Lopes, T. R., Zolin, C. A., Prado, G. D., Paulino, J., Almeida, F. T. D. (2017). Regionalization of maximum and minimum flow in the Teles Pires basin, Brazil. Engenharia Agrícola, 37(1), 54-63. http://dx.doi.org/10.1590/1809-4430-eng.agric.v37n1p54-63/2017

Malone, B. ithir: Functions and Algorithms Specific to Pedometrics, 2019. Available online on: http://R-Forge.R-project.org. Assessed in 09/01/2020.

MathWave Technologies. EasyFit 3.6, 2017.

Moriasi, D. N.; Gitau, M. W.; Pai, N.; Daggupati, P. (2015) Hydrologic and water quality models: Performance measures and evaluation criteria. Transactions of the ASABE, 58 (6), 1763-1785. https://doi.org/10.13031/trans.58.10715

Pereira, D. D. R., Martinez, M. A., da Silva, D. D., Pruski, F. F. (2016). Hydrological simulation in a basin of typical tropical climate and soil using the SWAT Model Part II: Simulation of hydrological variables and soil use scenarios. Journal of Hydrology: Regional Studies, 5, 149-163. https://doi.org/10.1016/j.ejrh.2015.11.008

Poff, N. L., Allan, J. D., Bain, M. B., Karr, J. R., Prestegaard, K. L., Richter, B. D., Stromberg, S., Stromberg, J. C. (1997). The natural flow regime. BioScience, 47(11), 769-784. https://doi.org/10.2307/1313099

MapBiomas – Coleção [Version 5] da Annual Series of Coverage and Land Use Maps of Brazil. Available online on: https://mapbiomas.org/. Assessed in 10/01/2020

Rao, A.R.; Srinivas, V.V. Regionalization of watersheds by fuzzy cluster analysis. Journal of Hydrology, 318, 57-79, 2006. https://doi.org/10.1016/j.jhydrol.2005.06.004

Requena, A. I., Chebana, F., Ouarda, T. B. (2018). A functional framework for flow-duration-curve and daily streamflow estimation at ungauged sites. Advances in Water Resources, 113, 328-340 https://doi.org/10.1016/j.advwatres.2018.01.019

Rezaie-Balf, M., Fani Nowbandegani, S., Samadi, S. Z., Fallah, H., Alaghmand, S. (2019). An ensemble decomposition-based artificial intelligence approach for daily streamflow prediction. Water, 11(4), 709. https://doi.org/10.3390/w11040709

Saadi, M., Oudin, L., Ribstein, P. (2019). Random Forest Ability in Regionalizing Hourly Hydrological Model Parameters. Water, 11(8), 1540. https://doi.org/10.3390/w11081540

Salio, P., Hobouchian, M. P., Skabar, Y. G., Vila, D. (2015). Evaluation of high-resolution satellite precipitation estimates over southern South America using a dense gauge network. Atmospheric Research, 163, 146-161. https://doi.org/10.1016/j.atmosres.2014.11.017

Schnier, S., Cai, X. (2014). Prediction of regional streamflow frequency using model tree ensembles. Journal of Hydrology, *517*, 298-309. https://doi.org/10.1016/j.jhydrol.2014.05.029

Schoppa, L., Disse, M., Bachmair, S. (2020). Evaluating the Performance of Random Forest for Large-Scale Flood Discharge Simulation. Journal of Hydrology, 125531. https://doi.org/10.1016/j.jhydrol.2020.125531

Silva, A. C. G. (2018). Identification of hydrologically homogeneous regions by fuzzy c-means group in the state of Paraná. 2018. 80p. MSc thesis. State University of the West of Paraná Cascavel, Brazil

Silva, R. D. S. E., Blanco, C. J. C., Pessoa, F. C. L. (2019). Alternative for the regionalization of flow duration curves. Journal of Applied Water Engineering and Research, 7(3), 198-206. https://doi.org/10.1080/23249676.2019.1611493

Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J.J., Mendiondo E.M., O'Connell, P.E., Oki, T., Pomeroy, J.W., Schertzer, D., S., Uhlenbrook, S., Zehe E. (2003). IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. Hydrological sciences journal, 48(6), 857-880. DOI https://doi.org/10.1623/hysj.48.6.857.51421

Smakhtin, V. U. Low flow Hydrology: a review. Journal of hydrology, 240 (3), 147-186, 2001. https://doi.org/10.1016/S0022-1694(00)00340-1

Swain, J. B., Patra, K. C. (2017). Streamflow estimation in ungauged catchments using regionalization techniques. Journal of Hydrology, 554, 420-433. https://doi.org/10.1016/j.jhydrol.2017.08.054

Tyralis, H., Papacharalampous, G., Langousis, A. (2019). A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water, 11(5), 910. https://doi.org/10.3390/w11050910

Wagener, T., Wheater, H. S. (2006). Parameter estimation and regionalization for continuous rainfall-runoff models including uncertainty. Journal of Hydrology, 320(1-2), 132-154. https://doi.org/10.1016/j.jhydrol.2005.07.015

Wing, J. M. K.C., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., et al. caret: Classification and Regression Training, 2020. Available online on: https://github.com/topepo/caret/. Assessed in 09/01/2020.

Wolff, W. (2017). Hydrologic Regionalization of Sant Catarina State: a seasonal and geostatistical approach based in models. 2017. 79 p. Ph.D. thesis. Luiz de Queiroz College of Agriculture/USP, Piracicaba, Brazil.

Wolff, W., Duarte, S. N., Mingoti, R. (2014). Nova metodologia de regionalização de vazões, estudo de caso para o Estado de São Paulo. Brazilian Journal of Water Resources – RBRH, 19(4), 21-33.

Xu, Z., Lian, J., Bin, L., Hua, K., Xu, K., Chan, H. Y. (2019). Water Price Prediction for Increasing Market Efficiency Using Random Forest Regression: A Case Study in the Western United States. Water, 11(2), 228. https://doi.org/10.3390/w11020228

Zambrano-Bigiarini, M. Package 'hydroGOF'. Goodness-of-fit Functions for Comparison of Simulated and Observed, 2020. URL: https://github.com/hzambran/hydroGOF

Zandonadi, L., Acquaotta, F., Fratianni, S., Zavattini, J. A. (2016). Changes in precipitation extremes in Brazil (Paraná River basin). Theor. Appl. Climatol., 123, 741-756. https://doi.org/10.1007/s00704-015-1391-4

Zhang, Y.; Vaze, J.; Chiew, F. H.; Li, M. Comparing flow duration curve and rainfall–runoff modelling for predicting daily runoff in ungauged catchments. Journal of Hydrology, 525, 72-86, 2015. https://doi.org/10.1016/j.jhydrol.2015.03.043

4. DISCHARGE PREDICTION BY USING REMOTE SENSING DATA IN PARANÁ STATE, BRAZIL

ABSTRACT

This study aimed to estimate the annual discharge (Q) based on annual precipitation (PPT) and evapotranspiration (ET), using the simple water balance equation across 28 watersheds in Paraná state, Brazil. Integrated Multi-satellitE Retrievals for GPM (IMERG) PPT and Atmosphere-Land Exchange Inverse (ALEXI) ET data were used as remote sensing products in the water balance components, while Q were estimated as the residual. The estimated Q values were compared to the observed data from fluviometric gauges in each watershed. We analyzed the effect of the area, slope, and vegetation coverage in Q estimation performance, though the performance and error indices: coefficient of determination (R^2), Pearson correlation coefficient (r), error root of the mean square error (RMSE), mean absolute error (MAE), and percent bias (pbias). A similar performance was observed over different areas and percent of slope. The vegetation coverage, however, affected the model performance. The water balance model using remote sensing products performed better in watersheds with more percentage of forest and pasture (>25 and 15%, respectively), and less of soybean (≤15%). In a general scenario, the model overestimated the Q in the watersheds, which is reasonable since we had not considered changes in water storage in the soil. The main advantages of this methodology are the simplicity and good performance of the model, as well as the free PPT and ET data, especially in regions where fluviometric gauges data are absent. The estimative of the Q, using a simple water budget using remote sensing product, provides an important hydrological tool for water resources management and possibility of use the same approach in other watersheds, with different climatologic and topography characteristics.

Keywords: ALEXI, IMERG, Hydrological models, Water balance

4.1. Introduction

The water resources management has as a main objective to ensure life and human development, as well as guarantee the availability of water for its multiple uses. Efficient governmental actions seek to minimize possible conflicts over this natural resource, avoid the advance of environmental degradation, and balance the demands of water for varied human activities and the maintenance of ecological uses.

The knowledge of the river's discharge allows the study of water availability over time. However, this task can be considered challenging in the water resources management,

especially in developing countries, due to the small number of gauges, when compared to the large number and the area of the watersheds (Swain and Patra, 2017), and also due to the absence of records of historical series with a data long period.

Conventional methods for predicting river discharge require a great amount of hydrological and meteorological data, the measurement of these data is expensive and time-consuming, and consists of a difficult process (Singh et al., 2018). Thus, the hydrological community has been applying several methodologies to estimate the river's discharge, using the available data of fluviometric gauges to build hydrological models, based on precipitation and evapotranspiration over the watersheds. A simple model to estimate the discharge is the water balance over the watershed, which is the budget of water computed by the precipitation as the inflow of water, minus the evapotranspiration, discharge, and positive soil water storage, representing the water outflow into the watersheds. Thus, the discharge can be obtained when the other components of the water balance are known.

Precipitation data (PPT), as well as discharge, can be obtained by rain gauges, which also demands high investment for its installation, and due to the territorial size of countries such as Brazil, monitoring by the punctual method over the whole territory is extremely difficult. Also, many of the historical series of equipment already installed have a few years of data and failures in certain periods, which require the estimation of the data in periods with gaps. For the PPT estimative, the hydrologists must consider its spatial heterogeneity variability over time, which can make conventional measurements in rain gauges spatial limited in terms of application in hydrological models (Duan and Bastiaanssen, 2013).

In this context, studies using new technologies and products are necessary to improve the performance of hydrological models. Remote sensing products from satellites represent a promising alternative to the conventional rain gauges due to PPT data available spatially and temporally continuous over large areas (Liu et al., 2015).

In the last few years, satellite PPT products have been used in many hydrological studies and have shown good performance in different approaches (Stisen and Sandholt, 2010; Moreno et al., 2012; Liu et al., 2015; Xu et al., 2015; Wu et al., 2018). For example, Li et al. (2012) compared the PPT obtained by rain gauges and the TRMM satellite (Tropical Rainfall Measuring), in the simulation of hydrological processes and the water balance of a basin located in China, and verified a good performance of the simulation model of monthly flow using the precipitation coming from the satellite.

The latest version of the IMERG algorithm (Version 6), made available to the public in October 2019, combines the reanalysis of PPT estimated by satellite TMPA and in the GPM.

Its products have a spatial resolution of 0.1° and a 30-minutes temporal resolution (Huffman et al., 2019). In this way, the IMERG products could give great accuracy to hydrological models.

Evapotranspiration (ET) is the combination of evaporation from soil and open water and plant transpiration (Ukkola and Prentice, 2013), and represents an important component of water balance and watersheds management. The direct ET measurement is not usual, and several methods can be used to obtain this component, e.g., equations, surface energy, and water balance methods, which vary in terms of complexity and data requirements, as described by Zhang et al. (2016). Regarding the approaches that allow mapping the ET spatially, the Two-Source Energy Balance (TSEB) allows the ET estimative using remote sensing products (Norman et al. 1995; Kustas and Norman, 2000; Li et al., 2005). The ALEXI (Atmosphere-Land Exchange Inverse) model is based on the TSEB (Anderson et al., 2007a; Anderson et al., 2007b) and was designed to minimize sensitivity to errors in inputs land-surface temperature and air temperature boundary conditions in ET estimation. Thus, the ALEXI model considers a spatially and physically realistic representation of land-atmosphere exchange over vegetation and cover conditions, using high temporal resolution products from geostationary satellites (Anderson et al., 2011).

The TSEB has been used in several studies to estimate turbulent fluxes over sparse and heterogeneous vegetation (Hssaine et al., 2018), in drought and ET approaches (Anderson and Kustas, 2008), monitoring of ET over irrigated and rainfall crops, as wheat, corn, and soybean (Choi et al., 2009, Boulet, et al., 2015; Diarra et al., 2017; Bigeard et al., 2019), ET estimating over a complex mountain region (Elfarkh, et al. 2020) and tree-grass ecosystem in semiarid conditions (Burchard-Levine et al., 2020). However, few studies used TSEB in the ALEXI model applied in Hydrology, focused on the discharge estimation by water balance in watersheds.

The application of remote sensing in Hydrology has provided the development of models with good performance for monitoring and estimating variables of interest. Additionally, these tools allow the spatialization of important variables, such as PPT, which were previously collected punctually. Thus, studies using this technology are extremely relevant for the improvement of hydrological models and techniques for estimating essential information in the management of water resources. In this context, this study aims to analyze the performance of estimating the annual discharge over watersheds using PPT and ET, obtained by remote sensing, by the water balance equation.

4.2. Material and methods

4.2.1. Study region

The study area consisted of 28 watersheds in Paraná state, in the Southern region of Brazil (Figure 14). It is the fifth most populous state in Brazil, with approximately 11.4 million citizens, in an area of 199,315 km² and 399 municipalities (IBGE, 2019). According to the Köppen classification, the state is located in the transition from tropical to subtropical climates, where the humid subtropical Cfa (hot summer) and Cfb (warm summer) predominate in 61.7% and 37.0%, respectively across the state area (Alvares et al., 2013a). The mean values of maximum and mnimum annual temperatures are 22.3 and 11.9 °C, respectively, while the mean annual air temperature is 16.9 °C (Alvares et al., 2013b). The temperature varies with the longitude, with higher values near the Tropic of Capricorn, in the north of the state (Santos et al., 2019).

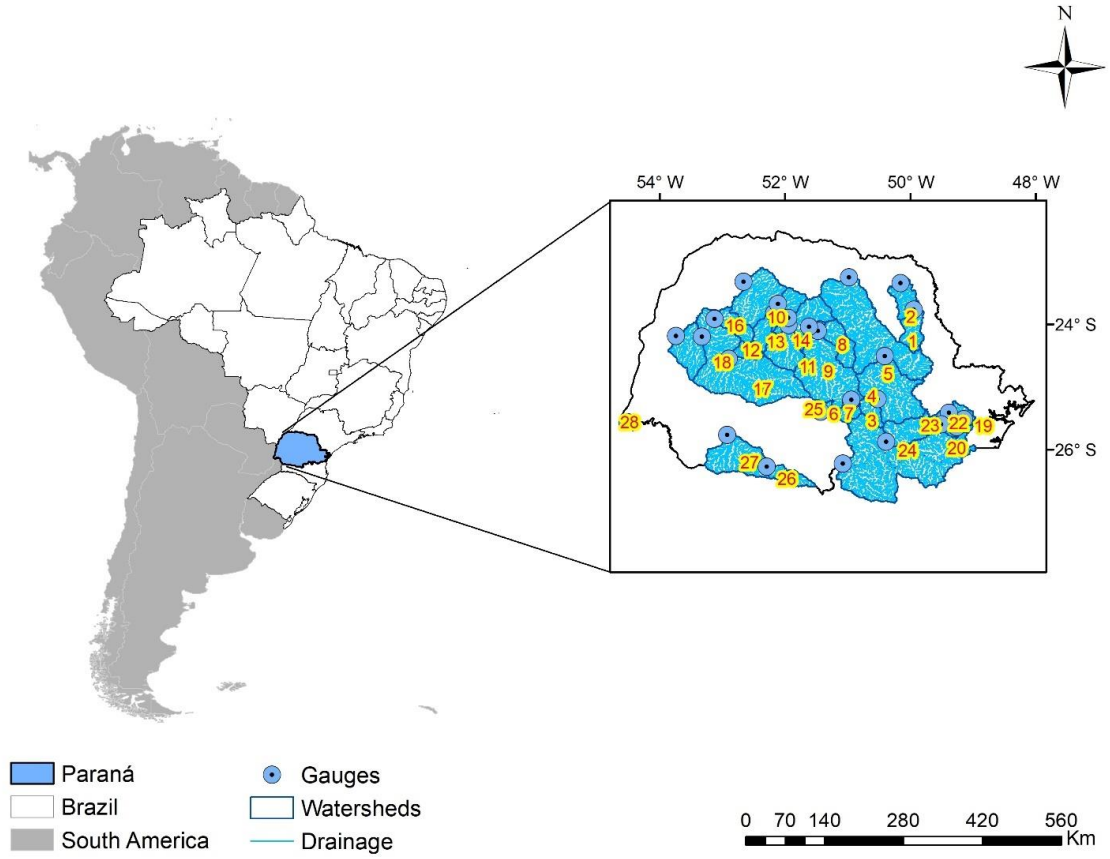

Figure 14. Study area, fluviometric gauges, and watersheds studied over the Paraná state, Brazil. The number of each watershed is located at its fluviometric gauge.

The main Brazilian biomes present in the state are the Atlantic Forest and Cerrado. The soils are highly weathered, with Ferralsols (30.8%) being the predominant soil type, followed by Entisols (22.2%), Acrisols (15.5%)), Nitisols (15.2%), and Cambisols (10.6%) (Bhering et al., 2008).

The land use in Paraná is mainly composed of natural forest (28%), planted forest (6%), pasture (14%), and short, medium, and perennial agricultural (35%) (Projeto MapBiomas, 2019). The state is the second-largest Brazilian region that produces and sells grains, mainly soybeans (Carmello and Sant'anna Neto, 2016).

The lower altitudes are in the coastal region and at the western area of the Paraná (Figure 15). The higher altitudes are concentrated in the central, southern, and eastern regions of the state. The rainfall in Paraná varies spatially, with an annual average between 1450 – 2352 mm (Figure 15). Summer is the season with the highest rainfall in South America, including the subtropical region (Grimm et al. 2007; Grimm 2011).

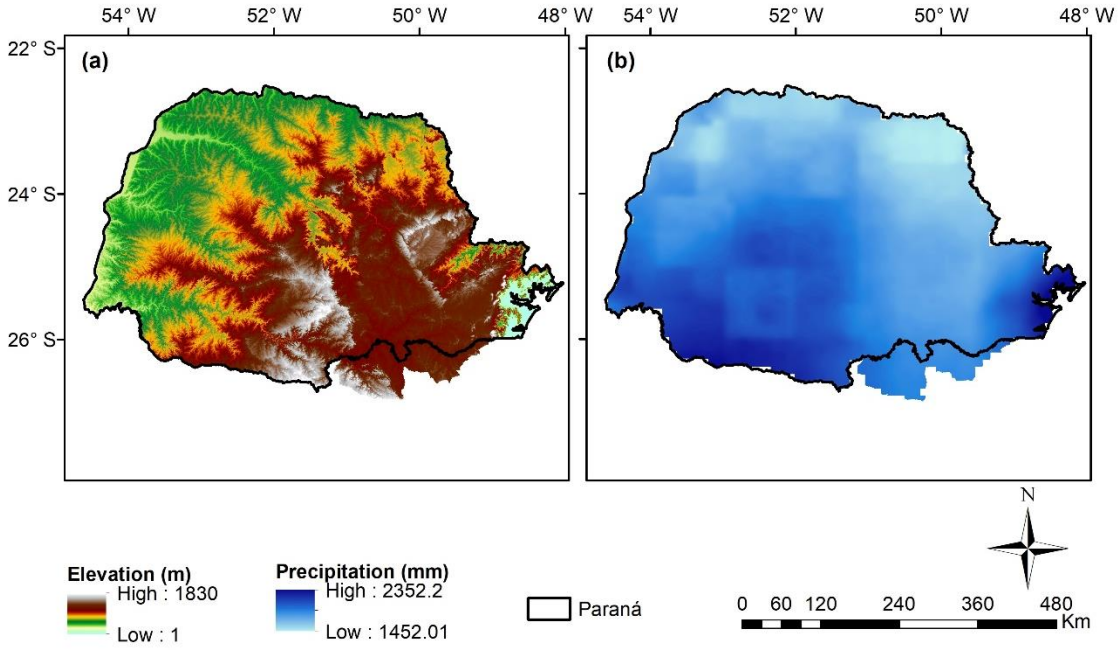

Figure 15. Mean annual precipitation (mm) and elevation (m) in Paraná state, Brazil.

4.2.2. Data

4.2.2.1. Observed data: Ground gauge

The discharge data used were acquired from 28 fluviometric gauges (Figure 14) of the National Water Agency (ANA), through the Hidroweb platform

(http://www.snirh.gov.br/hidroweb/serieshistoricas). The daily data in each gauge were added to result in monthly information. The analyzed time series was from January 1, 2013, to December 31, 2014, and fluviometric gauges data did not present gaps in this period.

The watersheds delimitation and characteristics, as area and slope, for example, were obtained using the Digital Elevation Model (DEM) in ArcGIS software. The watersheds description is presented in Table 8.

Table 8. The average of annual PPT (mm), and physical characteristics as altitude (m), area (km²) and slope (%), and percentage of land use of forest, soybean, and pasture, over the 28 watersheds area in Paraná state, Brazil. The code is the number of the fluviometric gauge in the Hidroweb platform and the ID is the number used to spacial identification in Figure 14.

Code	ID	PPT (mm)	Altitude (m)	Area (km²)	Slope (%)	Forest (%)	Soybean (%)	Pasture (%)
64360000	1	1592.80	565.99	2016.69	11.50	49.10	0.73	29.12
64362000	2	1556.18	901.42	4101.26	9.93	18.77	12.43	27.31
64442800	3	1742.36	623.31	1361.38	7.65	43.79	29.52	13.74
64465000	4	1717.50	442.53	8924.66	8.14	36.15	33.20	15.59
64507000	5	1693.30	596.08	22032.96	9.79	17.62	13.87	9.03
64619950	6	1810.39	862.21	1051.30	10.40	32.98	12.07	11.05
64620000	7	1810.39	500.63	1087.72	10.28	53.37	16.79	17.92
64652000	8	1796.40	708.56	2610.87	14.91	18.40	3.14	13.13
64655000	9	1851.62	677.06	12701.24	13.06	20.69	9.45	22.52
64659000	10	1924.35	691.75	3284.93	14.37	21.50	27.11	42.77
64660500	11	1852.36	933.62	19433.30	12.93	17.29	10.35	20.96
64673000	12	1869.64	930.74	1547.08	7.27	10.24	33.37	6.93
64675002	13	1844.65	935.68	23102.31	12.13	31.05	25.68	19.59
64685000	14	1821.47	926.20	28404.66	11.00	12.59	12.35	16.20
64785000	15	1937.90	922.79	1137.31	7.19	14.26	73.22	7.55
64810000	16	1777.33	931.73	2039.88	6.20	22.51	15.08	16.18
64820000	17	1893.72	931.49	17430.99	10.74	25.96	17.39	18.66
64830000	18	1875.04	895.16	20962.35	9.92	11.54	17.70	8.37
65010000	19	1949.19	675.47	105.28	7.32	19.89	32.60	27.66
65017035	20	1873.83	913.24	63.29	5.76	24.52	3.84	3.84
65021770	21	1748.82	534.69	24.38	8.26	31.99	11.62	2.36
65025000	22	1861.22	732.31	2403.73	5.90	19.12	8.78	0.00
65060000	23	1781.37	832.87	6016.70	7.12	20.88	7.42	3.23
65310000	24	1805.56	845.75	24046.51	8.37	32.13	8.50	3.05
65809000	25	1899.28	700.36	312.28	1.80	41.51	1.52	4.05
65925000	26	2090.70	1046.71	1657.01	1.76	47.36	15.15	6.21
65960000	27	2111.35	1028.24	6693.92	2.20	31.54	12.95	15.89
65996000	28	1919.35	797.17	134.75	4.21	15.54	32.41	11.87
Minimum	-	1556.18	442.53	24.38	1.76	10.24	0.73	0.00
Maximum	-	2111.35	1046.71	28404.66	14.91	53.37	73.22	42.77
Average	-	1836.00	788.71	7667.46	8.58	26.51	17.79	14.10

4.2.2.2. Estimated precipitation data: IMERG

Precipitation data by remote sensing was in monthly and 0.1° spatial-temporal resolution, respectively, from the satellite constellation of the Global Precipitation Measurement mission (GPM), IMERG Version 6 product, distributed by Goddard Earth Sciences Data and Information Services Center Distributed Active Archive Center (GES DISC DAAC), available online on http://mirador.gsfc.nasa.gov/com. The monthly products "Final Run" were applied, and the time series analyzed was coincident with that of the fluviometric gauges (January 1, 2013, to December 31, 2014).

The IMERG algorithm operates to intercalibrate, merge, and interpolate all satellite microwave precipitation estimates, microwave-calibrated infrared estimates, gauge observations, and other data from potential sensors from the TRMM and GPM eras (Huffman et al., 2019). The "Final Run" product includes microwave-infrared estimates without gauge adjustment and the calibrated product based on the Global Precipitation Climatology Centre monthly gauge analysis (Tang et al., 2020). In general, the "Final Run" products present bias correction and more accurate results than the other products supplied almost in real-time (Early and Late Run) (Su et al., 2019).

4.2.2.3. Estimated evapotranspiration data: ALEXI

The ALEXI (Atmosphere-Land Exchange Inverse) model is based on the Two-Source Energy Balance (TSEB, Norman, et al., 1995; Kustas and Norman, 2000). The ALEXI was developed by USDA-ARS (U.S. Department of Agriculture - Agricultural Research Service) and is applied in several programs and studies coordinated by Daugherty Water for Food Global Institute (DWFI), at the University of Nebraska-Lincoln, in the United States.

The TSEB obtains the ET using a land-surface representation to partition surface fluxes between the canopy and the soil (Cawse-Nicholson and Anderson, 2018), which treats the land surface as a mosaic of soil and vegetation elements with different temperatures, fluxes, and atmospheric coupling.

According to Elfarkh et al. (2020), two input variables derived from remote sensing instruments are key to the TSEB model: the first is the surface temperature, which is used in sensible heat flux estimative, and the second is the vegetated cover fractional, which controls the partitioning of energy between surface vegetation and the underlying soil. The soil and

canopy energy budgets are computed separately in the TSEB, using the Eq. 9, Eq.10, and are added in the Eq. 11 (Anderson et al., 2005):

$$RNs = Hs + LEs + G \quad (9)$$

$$RNc = Hc + LEc \quad (10)$$

$$RN = RNs + RNc \quad (11)$$

where RN, RNc, and RNs are the total, canopy, and soil net radiation, respectively, Hc and Hs, and LEc and LEs are the canopy and soil sensible and latent heat, respectively, and G represents the ground heat conduction. The canopy is assumed to transpire at his potential rate using the Priestley-Taylor (1972) equation (Eq. 12)

$$LEc = \alpha_{PT} f_G \frac{\Delta}{\Delta\gamma + \Delta} RNc \quad (12)$$

where α_{PT} is the Priestley-Taylor (Priestley and Taylor, 1972) coefficient, f_G is the fraction of green vegetation, γ is the thermodynamic psychrometric constant (approximately 67 Pa K^{-1}), and Δ is the slope of the temperature saturation vapor pressure curve.

The TSEB model uses observation of radiometric surface temperature (Trad) to calculate the energy balance at the time of the satellite overpass (Diarra et al., 2017), then the model breaks down total LE into estimates of soil evaporation (LESs) and canopy transpiration (LEc). The TSEB partitions the composite surface radiometric temperature (Trad), obtained from thermal measurements into characteristic soil and canopy temperatures (Ts and TC), considering the vegetation cover fraction apparent at the sensor view angle, f(θ), using the Eq. 13 and Eq. 14, respectively:

$$Trad = [f_\theta T_c^{\,4} + (1 - f_\theta) T_s^{\,4}]^{1/4} \quad (13)$$

where T_c and T_s are the canopy and the soil temperatures, respectively; f_θ is the fractional vegetation cover. The soil conduction heat is calculated by Eq. 6:

$$f_\theta = 1 - \exp\left(\frac{-0.5\,\Omega(\theta)\,LAI}{\cos\theta}\right) \quad (14)$$

where $\Omega(\theta)$ is a view angle dependent clumping factor, assigned by vegetation class, and LAI is the leaf area index (Anderson et al., 2005). The soil conduction heat is calculated by Eq. 15 (Li et al., 2005):

$$G = 0.35\ RNs \tag{15}$$

where RNs are the soil net radiation. The soil and canopy sensible heat fluxes (Hs and Hc, respectively) are obtained using Eq. 16 and Eq. 17:

$$Hs = \rho_{Cp} \frac{T_s - T_a}{r_{ah} + r_{sh}} \tag{16}$$

$$Hc = \rho_{Cp} \frac{T_c - T_a}{r_{ah}} \tag{17}$$

where ρ_{Cp} is volumetric heat capacity (J m^{-3} K^{-1}) of air, T_a is the air temperature (K) measured at 2 m, r_{ah} is the aerodynamic resistance (s m^{-1}), and r_{sh} is the resistance (s m^{-1}) to heat flow above the soil surface.

ALEXI applies the TSEB two times a day, during the morning, after one hour of sunrise, and one hour before local noon (Cawse-Nicholson et al., 2020), measuring the land surface temperature through images collected by a geostationary satellite (Castelli et al., 2018). A TSEB and ALEXI mathematical formulation as well as a detailed methodology are present by Anderson et al. (2007a), Anderson et al. (2007b), and Cawse-Nicholson and Anderson (2018).

The ALEXI model can be applied to a wide range of canopy and moisture stress conditions, including partially vegetated surfaces, at high-temporal resolution (hourly) information provided by geostationary satellites (Anderson et al., 2007a; Anderson et al., 2007b).

The ALEXI products are available on-line on the “Global EvapoTranspiration” (GloDET - https://glodet.nebraska.edu/index.html#/) after validated over the globe in a 15º x 15º grid, using eddy covariance (EC) measurements, which obtains fluxes in direct measurements, averaging the product of fluctuating vertical velocity and the transported quantity, e.g., humidity (Wang and Dickinson, 2012). We used the ALEXI daily product for the Paraná state, in a 375 meter of spatial resolution. The daily data were added to the results

in monthly information. The time series analyzed were coincident with that of the fluviometric gauges and IMERG data, from January 1, 2013, to December 31, 2014.

4.2.2.4. Water balance and discharge prediction and performance analyses

The water balance in a watershed is calculated considering a system where PPT is the main water inflow, and Q, ET, and the positive water storage in the soil (Δs) are the outflow of water in the watershed (Eq. 18). PPT is the higher and independent term in the water balance, followed by the ET, closely linked with vegetation characteristics (Zhang et al., 2001). Over the annual-scale, the Δs change in water storage can be negligible in watersheds (Zhang et al., 2001; Liu et al., 2016), especially in tropical regions where there is no effect of snowmelt, and during a period when the soil use and the water multiple uses are very similar over the year. In this context, Q in the watersheds analyzed in this study were obtained using Eq. 19:

$$PPT = ET + Q + \Delta s \quad (18)$$

$$Q = PPT - ET \quad (19)$$

where Q is the discharge (mm) predicted using the remote sensing products, PPT is the annual precipitation (mm) from IMERG, ET is the evapotranspiration (mm) ALEXI products, and Δs is the water storage in the soil (mm), not considered in this study.

As the source of PPT and ET were remote sensing products, we applied Eq. 19 in a map algebra in SIG software. First, we downscaled the IMERG to the same spatial resolution of the ALEXI products, 375 m. Then we subtracted the IMERG (PPT) and ALEXI (ET) pixels. The estimated Q (mm) was obtained averaging values for each watershed, using the zonal tool.

We evaluated whether the watershed characteristics could affect the estimation performance. Thus, we separated the watersheds into two groups, considering the watersheds area: small (area ≤ 5000 km^2), and big (area >5000 km^2), and in two groups considering watershed's slope: flat-soft wavy (≤8%) and wavy-strong wavy (>8%). The slope classification criteria were based on EMBRAPA criteria (1979).

Additionally, we evaluated the vegetation effect over the estimation performance, and for this purpose we grouped the watersheds in six groups, considering the percentage of forest (≤25% and >25%), soybean (≤15% and >15%), and pasture (≤15% and >15%). The percentage

of land use type was selected considering its average over the 28 watersheds, as shown in Table 8.

To evaluate the performance and accuracy of estimating the Q using remote sensing, we compared the watershed estimated data with the fluviometric gauges (ANA), using statistic metrics. We used the coefficient of determination (R^2), Pearson correlation coefficient (r), error root of the mean square error (RMSE), mean absolute error (MAE), and percent bias (pbias), using the packages *hydroGOF* and Metrics in Rstudio software.

4.3. Results and Discussion

4.3.1. Water balance for the watersheds

The average PPT of the 28 watersheds in 2013 and 2014 are presented in Figure 16. We evaluated the average precipitation in all watersheds considering the mean annual PPT of 18 years, aiming to analyze if the 2013 and 2014 consisted of dry years for those watersheds. With exception of the watersheds 2 (northeast) in 2014, and 19, 20, 21, and 22 (near the coastal region) in 2013 and 2014, the PPT annual average was superior to the average of 18 years, which indicates, in the general scenario, no deficit in pluviometry for those watersheds.

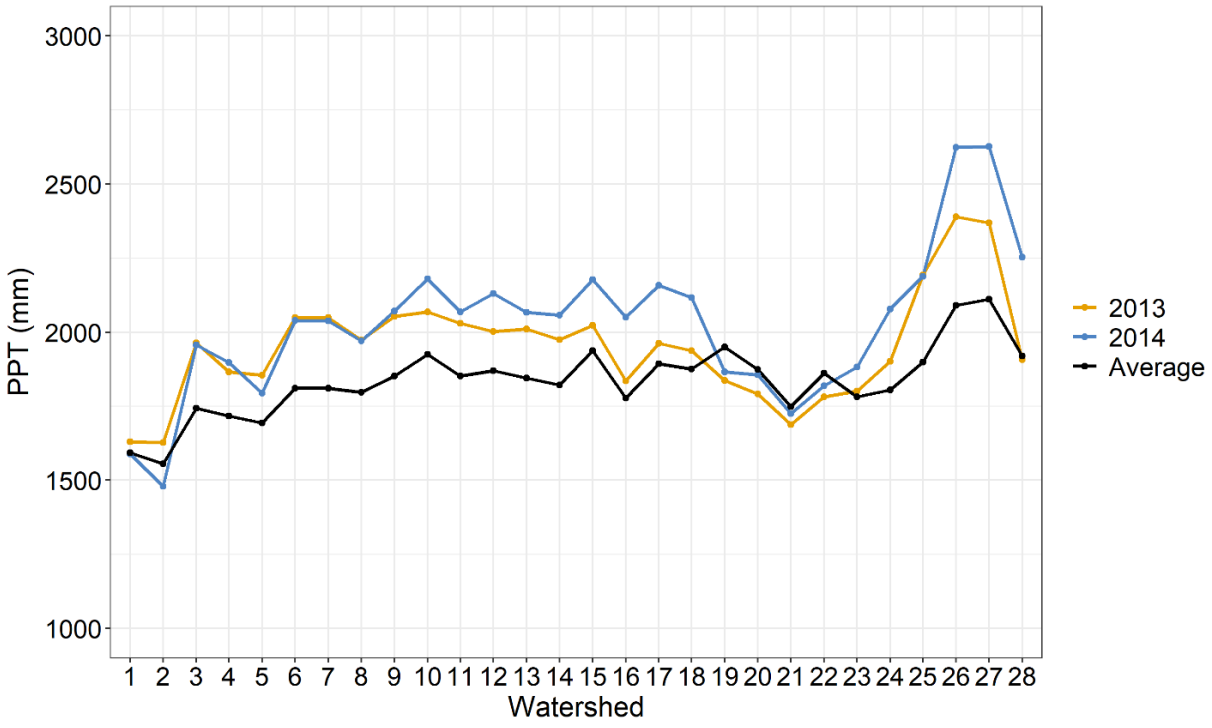

Figure 16. Average precipitation over the watersheds studied, considering an 18-year data series and the years 2013 and 2014.

Figure 17 shows the mean values of ET and PPT for all the 28 watersheds, and the estimated Q obtained from remote sensing products, compared to the observed Q in the fluviometric gauges. The wet period occurred from September to March, with the higher values of PPT. From April to August the drought period was observed , except for June, when the amount of PPT for the watersheds was greater than the values observed in the other months in 2013 and 2014.

The higher values of ET were noticed from the spring and summer (September to March), coinciding with the highest PPT value, when the solar radiation flux is about 40-50% larger than in fall and winter (Martins et al., 2008). The ET variability is strongly controlled by the water and energy balance components such as precipitation, solar radiation, air temperature, relative humidity, wind speed, and vegetation cover (Granata, 2019). In fact, the energy required for the ET process is provided by solar radiation and air temperature (Granata, 2019), which explains the higher ET over the summer and spring.

In the summer of the Southern Hemisphere tropics and subtropics, where Paraná is inserted, PPT is more related to humidity from air mass that moves from the Amazon to the southwestern Atlantic, defined as the South Atlantic Convergence Zone (SACZ, Hirata and Grimm, 2016). During the other seasons, the PPT is related to convective and frontal complexes, when the incursion of frontal systems increases the nebulosity and results in the reduction of solar radiation (Martins et al., 2008). Consequently, the ET decreases in this period, due to the low solar radiation energy (Figure 17).

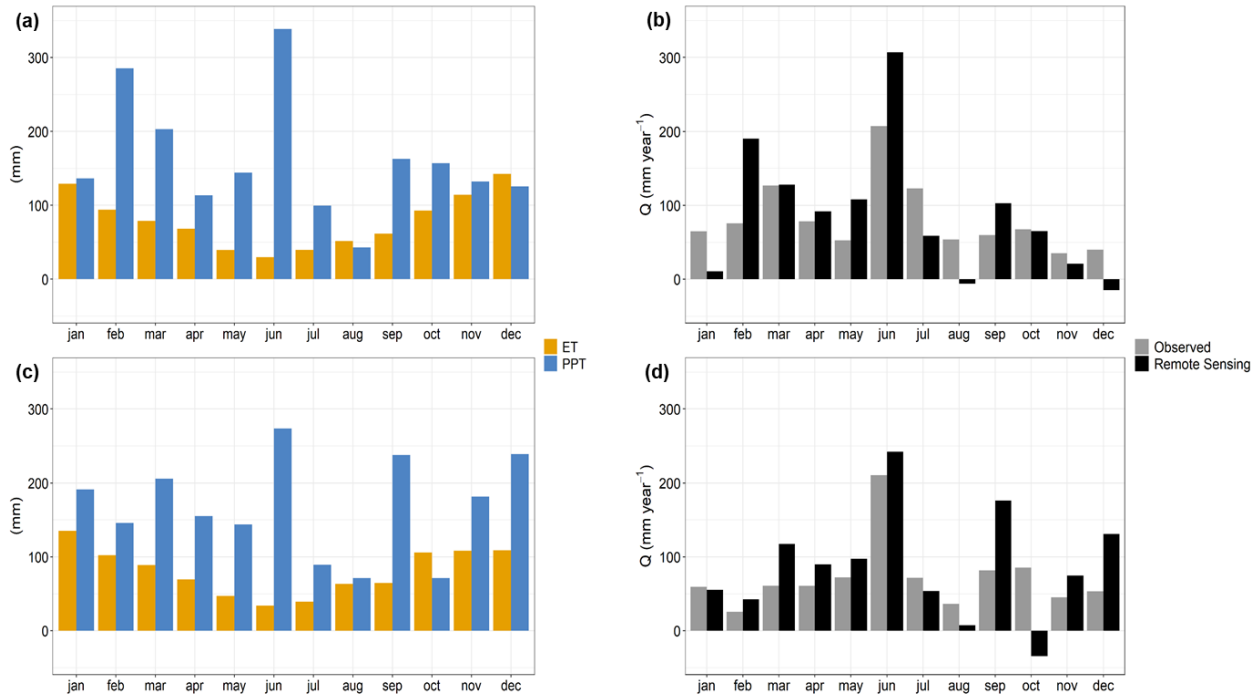

Figure 17. Average PPT and ET in 2013 (a) and 2014 (c) of the 28 hydrographic basins and the distribution of Q estimated by remote sensing and the observed fluviometric gauges in the years 2013 (b) and 2014 (d).

In the wet season, the PPT value, as well as ET, was higher due to the greater solar radiation energy explained above. The Q estimated was obtained subtracting the PPT from the ET products. During the dry period, the groundwater is, in general, the primary source of water in the watersheds (Belhassan, 2011, Gilfedder, et al., 2012), and the Q observed in the watershed was lower than in the wet seasons, but always with a positive value. Thus, the Q estimated presented positive values during the months when the amount of PPT was greater than the ET, basically all the months of the years, and negative values when the ET rate was greater than the PPT, as in August and December of 2013, and October of 2014 (Figure 17).

4.3.2. Annual discharge (Q) estimation using remote sensing products

Figure 18 shows the annual Q estimated from water balance using the PPT and ET products from remote sensing compared to the observed values by fluviometric gauges in watersheds, for the years 2013 and 2014. The water balance equation performed well, explaining more than 62% ($R^2 > 0.62$) of the Q in the watersheds, with a correlation (r) higher than 0.79 and accuracy (RMSE) of 200.360 mm $year^{-1}$, and a tendency to overestimate the Q in 12.9% of the annual values, considering the analyses of 2013 and 2014 as a unique dataset.

Regarding the individual analysis in 2013 and 2014, in both years the water balance equation explained more than 63% ($R^2 > 0.63$) of the Q in the watersheds, with a correlation (r) higher than 0.79. In 2014 the higher value of RMSE (234.94 against 158.40 mm $year^{-1}$) and pbias (12.9 against 4.8 %) could has as a hypothesis the temporal distribution of the PPT over the year, compared to 2013. Besides the average of PPT and ET in both years was very similar (Table 8), in 2013 February, March and June presented greater depth when compared to the other months in this year (Figure 17), even in the wet season.

On the other hand, in 2014 greater volumes in June and September was observed, however, they were closer to those observed during the wet season, compared to the peaks observed in 2013 and the PPT values during the wet season. The infiltration of the water for groundwater recharge is more accentuated after PPT events, and especially in wet season (Yenehun et al., 2020). Thus, the similar distribution of the annual PPT over the months could propitiate less variability of the amount of water storage of the water in the soil and the groundwater recharge through the gradual infiltration process. Thereby, since we negligence the storage term in the water balance, it is comprehensible (and expected) an overestimation of the Q over the watersheds.

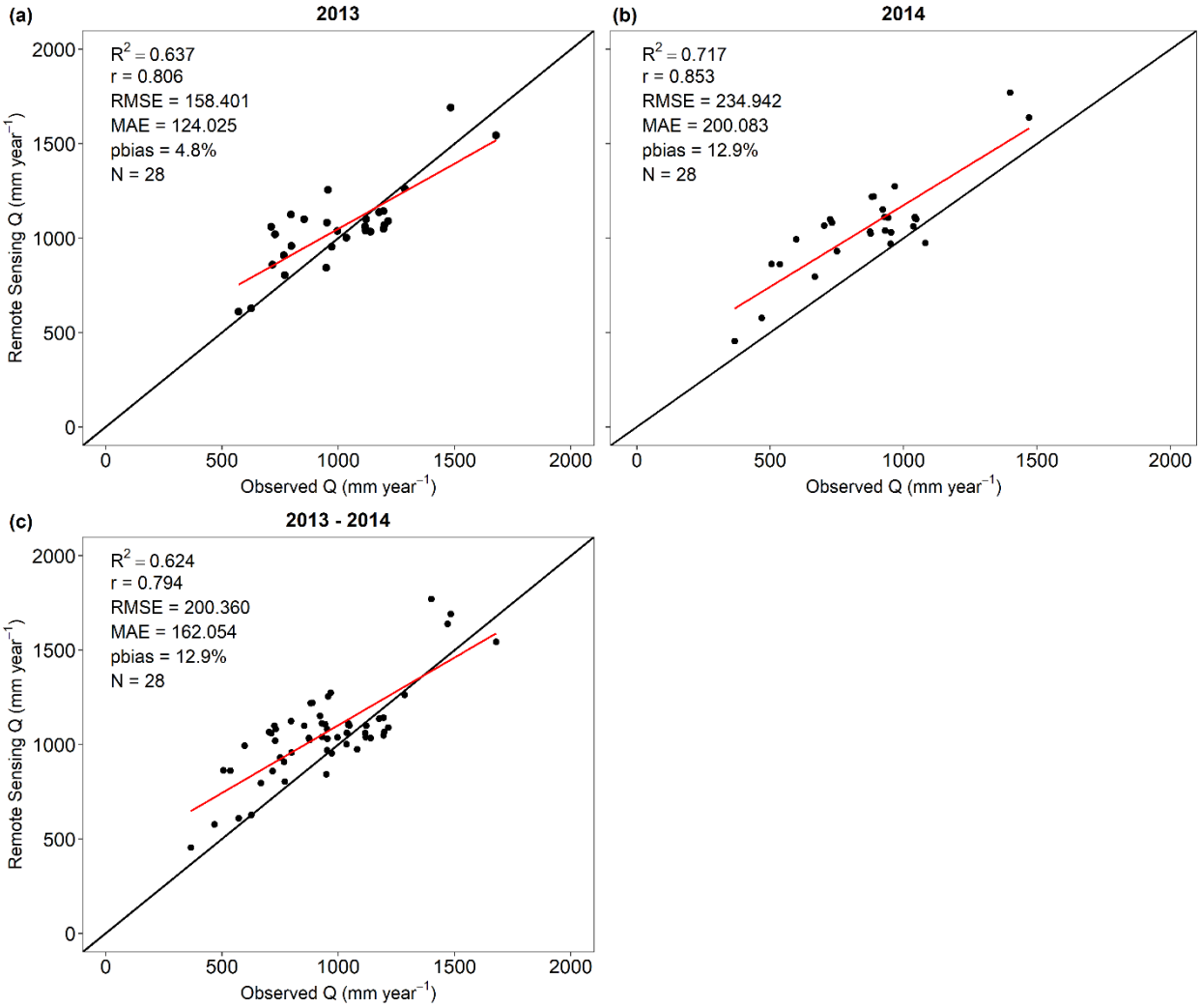

Figure 18. Comparison of the Q (mm year^{-1}) from water balance using remote sensing against observed data from fluviometric gauges in watersheds in Paraná state, Brazil, for 2013 (a), 2014 (b), and 2013-2014 (c). R² is the coefficient of determination, r is the Pearson correlation coefficient, RMSE is the root of the mean square error, MAE is the mean absolute error, pbias is the percent bias (pbias), and N is the number of watersheds.

Additionally, we did not evaluate in this study the uncertainty of the remote sensing products over the watersheds area, and this could be a second hypothesis for the overestimation of the Q. We admitted the good performance of the IMERG to estimate the PPT in the Paraná region observed by Gadelha et al. (2019). Thus, the Q overestimate also could be a result of the overestimate by the PPT products.

In fact, Behrangi et al. (2011) found that satellites products (TMPA-RT, TMPA-V6, CMORPH, PERSIANN, and PERSIANN-adj) significantly overestimate over warm months, during the spring and summer, and underestimate in cold season, resulting in over-and under-estimation of discharge forecast, respectively, in the United States. Wang et al. (2018), also observed overestimations of discharges prediction because of the overestimate of IMERG PPT

products in China. In the context of this study, Gardelha et al. (2019) observed overestimation by the IMERG algorithm compared to rain fluviometric gauges in the Paraná region, and it could be a reasonable explanation of the Q overestimate over the watersheds.

Regarding the ET, Cawse-Nicholson et al. (2020) analyzed the uncertain of the disALEXI (a downscale of ALEXI) over areas with agriculture and non-agriculture land-use type in the United States. According to the authors, based on expert knowledge of the algorithm's developer, it tends to overestimate the ET, and land the surface temperature, albedo, and leaf area index (LAI), represent uncertainty in disALEXI ET. Thus, ALEXI ET could result in an uncertain Q estimative, but not in the overestimation of this hydrological parameter, because of the water balance Q = PPT – ET and higher ET implies lower values of Q.

Aiming to analyze if the area and mean slope of the watersheds could affect the Q estimative, we grouped the watersheds considering areas smaller and bigger than 5000 km², and mean slope greater an lower than 8%, as shown in Figure 19.

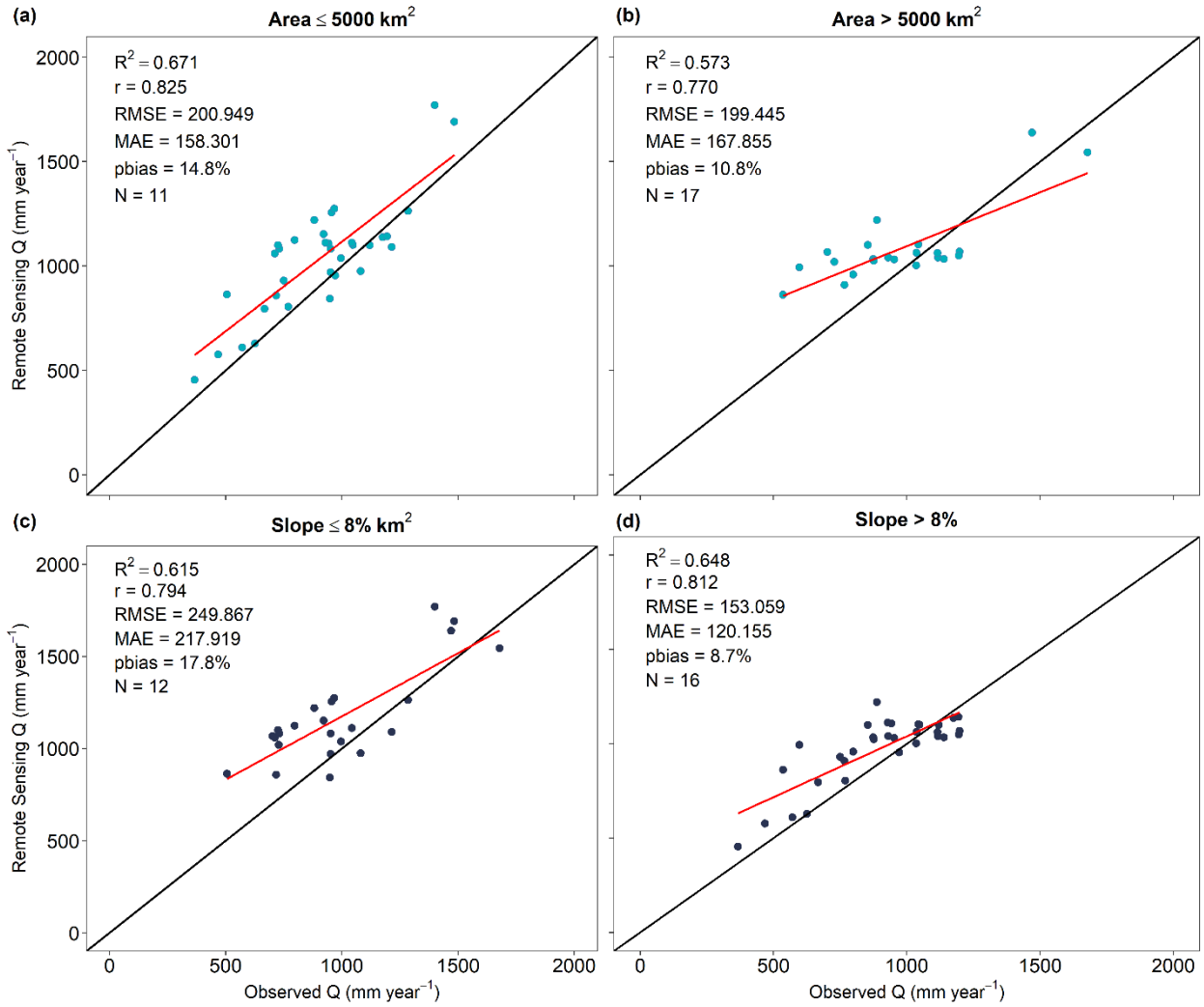

Figure 19. Comparison of the Q (mm year^{-1}) from water balance using remote sensing against observed data by fluviometric gauges in watersheds in Paraná state, Brazil. R^2 is the coefficient of determination, r is the Pearson correlation coefficient, RMSE is the root of the mean square error, MAE is the mean absolute error, pbias is the percent bias (pbias), and N is the number of watersheds present in the groups with (a) area ≤ 5000 km^2, (b) area > 5000 km^2, (c) slope ≤ 8%, (d) slope > 8%.

The estimation of the Q by the water balance equation performed very similarly in all the area and slope groups (Figure 19). The statistic metrics of Q estimated and observed in fluviometric gauges were very close, with R^2 of 0.671 and 0.573 and r 0.828 and 0.770, in the watersheds smaller and larger than 5000 km², respectively. In both areas group, the RMSE was approximate 200 mm year^{-1}, with MAE ranging from 158.30 to 167.86 mm year^{-1} and with the same tendency of overestimating (pbias = 10.8%). According to Zhou et al. (2015), large watersheds tend to present lower hydrological responses mainly due to their more complex landforms, greater buffering capacities, and longer residence times. In this study, we analyzed a long period, and this effect could be not as important for annual responses. Another concern was having good estimative in the small watershed due to the pixel size, but the high spatial

resolution of the ET and PPT products (365 m and 10 km, respectively), were shown to be sufficient for the Q estimation.

In the slope groups, the Q estimative performed even more similarly with R^2 of 0.615 and 0.648 and r 0.794 and 0.812, in the watersheds with average slope greater and lower 8%, respectively. The water balance in watersheds with flat-soft wavy (≤8%) presented a tendency of overestimation in almost 10 times larger than wavy-strong wavy slope (>8%), with pbias of 17.8 and 8.7%, respectively. The same tendency was observed for the RMSE and MAE, of 249.87 and 217.92 mm $year^{-1}$ in watersheds group with an average slope lower than 8%, and 153.06 and 120.16 mm $year^{-1}$ for slope greater than 8%, respectively. The watersheds with higher values of slopes in general present lower water retention ability due to its faster water movement (Zhou et al., 2015), especially in afforested areas (Šatalová and Kenderessy, 2017). On the other hand, the flat terrain over the watershed area contributes to the longer residence time of the water and its infiltration through the soil, which increases the storage term value, not considered in this study. Thus, the overestimation tends to be greater in these groups of watersheds.

Water infiltration through the soil also is related to its land use. Figure 20 shows the Q estimation in the watersheds grouped considering the percentage of forest (≤25% and >25%), soybean (≤15% and >15%), and pasture (≤15% and >15%). The adjusted models into the watershed's groups with a percentage of land use equal or higher than 25% (Figure 20b) of forests and 15% of pasture (Figure 20f) presented a higher coefficient of determination (0.754 and 0.802, respectively) compared to the models with watersheds with areas below these limits (Figures 7a and 7e).

The Q estimation over the watersheds with the land use of soybean in an area of less than 15% (Figure 20c) presented a higher coefficient of determination (0.697) than those with an area above this limit (Figure 20d). The accuracy and errors, in terms of RMSE and MAE, were very similar between the forests and soybean watersheds groups (approximately 200 and 160 mm $year^{-1}$, respectively), as well as the tendency of overestimating (pbias value). However, in the pasture groups the Q estimative was more accurate, with minor errors, and less overestimated in the group of pasture >15%, with RMSE approximately 117 mm $year^{-1}$, MAE 102.9 and pbias 17.5% smaller than the group of pasture ≤15% (Figure 20e and 7e).

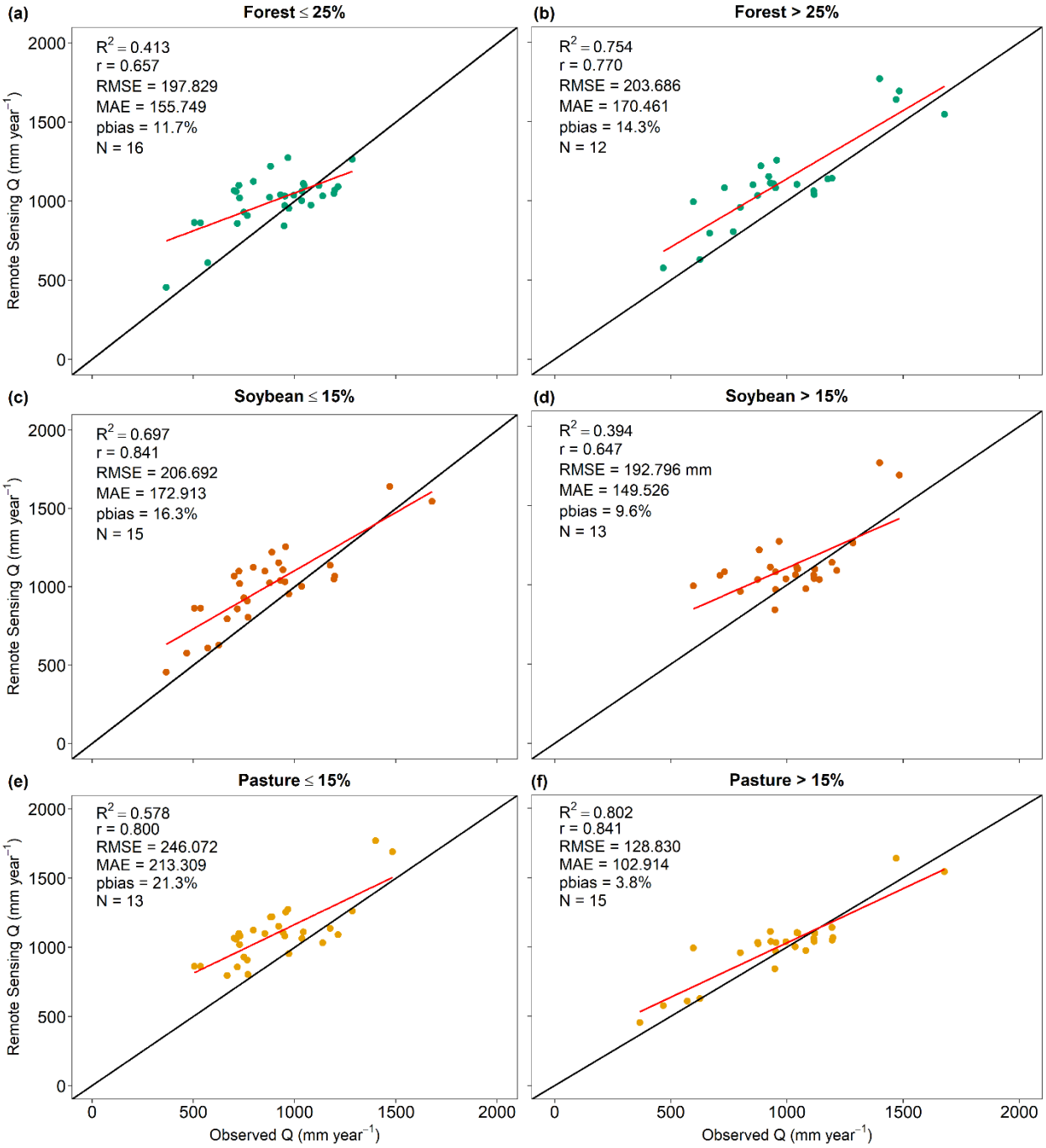

Figure 20. Comparison of the Q (mm year^{-1}) from water balance using remote sensing against observed data by fluviometric gauges in watersheds in Paraná state, Brazil. R^2 is the coefficient of determination, r is the Pearson correlation coefficient, RMSE is the root of the mean square error, MAE is the mean absolute error, pbias is the percent bias (pbias), and N is the number of watersheds present in the groups with (a) forest ≤ 25%, (b) forest > 25%, (c) soybean ≤ 15%, (d) soybean > 15%, (e) pasture ≤ 15%, and (e) pasture > 15%.

The importance of the forest to the discharge maintenance and regularization in watersheds is known. Cecílio et al. (2019), working with modeling the influence of forest cover on discharges in the Soil & Water Assessment Tool (SWAT) software, observed that

afforestation over the upper regions of the watershed can increase minimum discharge. In a general scenario, forest soils present higher total water storage capacity and more large pores, which allow the water infiltration and percolation, resulting in less runoff (Cheng and Lu, 2002). Additionally, the forest increases the infiltration and retention of the water by reducing the water runoff as a result of physic obstacles, longer discharge paths, and pores in surface soil (Zhou et al., 2015).

The effect filter of forest and pasture areas is reported in the literature (Menezes et al., 2016). The vegetation filter is related to the greater interception of rain and the reduction of the speed of surface runoff, due to the greater soil coverage that is provided by forests and pastures. Both are very important, in different magnitudes, for the hydrologic cycle components, like an infiltration, ET, soil moisture, annual water yield, peak, and low discharge in the watersheds (Cheng and Lu, 2002).

Regarding the filter effect, the forest and pasture provide greater coverage stability throughout the year than in areas with soybean crops, due to its permanence in the field. In soybean cultivation areas, because it is an annual crop, there is a greater dynamic regarding soil cover. Thus, this greater stability of ground cover reflects in the models with greater performance. Grouping the watersheds by the percentage of land use (Figure 20), is a way to understand and improve the discharge estimates through the water balance with the use of remote sensing products.

This study did not explore the performance of the ALEXI over each land use, because it was not our main objective. We intended to present a brief panorama of the performance of Q estimative using only remote sensing products and investigate the model performance over watershed with different land uses. However, other studies reported the better performance of the algorithm and overestimation tendency due to its components and different architecture and spacing of vegetations in the field (Anderson et al., 2005, Kustas and Anderson, 2009, Semmens et al., 2016), with could be a key to future hydrological studies.

4.4. Conclusion and final considerations

In developing countries, fluviometric gauges data are limited in density and frequency, and the use of free data, in hydrological models, as the remote sensing products, represents a tool for water resources management. We analyzed the Q annual estimation by water balance using the IMERG PPT and ALEXI ET products for 28 watersheds in Paraná state, in Brazil.

The water balance for the Q estimation for the years 2013 and 2014, using only remote sensing products, performed very well and presented a good correlation with observed data in fluviometric gauges in the same period. We investigated the effect of the area and slope over the model performance, grouping the watersheds with similar characteristics, and the results showed similar performance over those groups.

Additionally, we analyzed the effect of the land uses, grouping the watersheds in percentages of forest, soybean, and pasture. We found a better performance of the model in watersheds that present a higher percentage of forest and/or pasture, and less percentage of soybean, which could be related to the filter effect of forest and pasture and to the lower time of permanence of the soybean in the field.

The limitations of this study were the availability of fluviometric gauge data, because only a few watersheds presented data available in 2013 and 2014 (and none after this period), coincident to the period of the ALEXI consisted data, from 2013 to now. The uncertainty of the IMERG and ALEXI products could result in an uncertainty of the water balance estimation. However, the positive performance of the model in estimating the Q using only remote sensing products supports the recommendation of this method for watershed hydrological predictions.

References

Alvares, C.A., Stape, J. L., Sentelhas, P.C., de Moraes, G., Leonardo, J., Sparovek, G., 2013a. Köppen's climate classification map for Brazil. Meteorologische Zeitschrift, 22, 711-728. https://doi.org/10.1127/0941-2948/2013/0507

Alvares, C. A., Stape, J. L., Sentelhas, P. C., Moraes Gonçalves, J. L., 2013b. Modeling monthly mean air temperature for Brazil. Theoretical and Applied Climatology, 113 (3-4), 407-427. https://doi.org/10.1007/s00704-012-0796-6

Anderson, M., Kustas, W., 2008. Thermal remote sensing of drought and evapotranspiration. Eos, Transactions American Geophysical Union, 89 (26), 233-234. https://doi.org/10.1029/2008EO260001

Anderson, M.C., Norman, J. M., Mecikalski, J.R., Otkin, J.A., Kustas, W.P., 2007a. A climatological study of evapotranspiration and moisture stress across the continental United States based on thermal remote sensing: 1. Model formulation. Journal of Geophysical Research, 112, D10117. https://doi.org/10.1029/2006JD007506

Anderson, M.C., Norman, J.M., Mecikalski, J.R., Otkin, J.P., Kustas, W.P., 2007b. A climatological study of evapotranspiration and moisture stress across the continental U.S. based on thermal remote sensing: II. Surface moisture climatology. Journal of Geophysical Research, 112, D11112. https://doi.org/10.1029/2006JD007507

Anderson, M.C., Norman, J.M., Kustas, W.P., Li, F., Prueger, J.H., Mecikalski, J.R., 2005. Effects of vegetation clumping on two–source model estimates of surface energy fluxes from an agricultural landscape during SMACEX. Journal of Hydrometeorology, 6 (6), 892-909. https://doi.org/10.1175/JHM465.1

Anderson, M. C., Kustas, W. P., Norman, J. M., Hain, C. R., Mecikalski, J. R., Schultz, L., Gonzalez-Dugo, M. P., Cammalleri, C., d'Urso, G., Pimstein, A., and Gao, F. (2011). Mapping daily evapotranspiration at field to continental scales using geostationary and polar orbiting satellite imagery. 15, 223–239. https://doi.org/10.5194/hess-15-223-2011

Behrangi, A., Khakbaz, B., Jaw, T.C., AghaKouchak, A., Hsu, K., & Sorooshian, S., 2011. Hydrologic evaluation of satellite precipitation products over a mid-size basin. Journal of Hydrology, 397 (3-4), 225-237. https://doi.org/10.1016/j.enpol.2008.02.014

Belhassan, K. (2011). Relationship between river flow, rainfall, and groundwater pumpage in Mikkes Basin (Morocco). Iranian Journal of Earth Sciences, 3(2), 98-107.

Bigeard, G., Coudert, B., Chirouze, J., Er-Raki, S., Boulet, G., Ceschia, E., Jarlan, L., 2019. Ability of a soil-vegetation-atmosphere transfer model and a two-source energy balance model to predict evapotranspiration for several crops and climate conditions. Hydrology and Earth System Sciences, 23 (12), 5033-5058. https://doi.org/10.5194/hess-23-5033-2019

Bosquilia, R.W., Neale, C.M., Duarte, S.N., Longhi, S.J., Ferraz, S.F.D.B., Muller-Karger, F.E., Mccarthy, M.J. 2018a. Evaluation of evapotranspiration variations as a function of relief and terrain exposure through multivariate statistical analysis. Ecohydrology & Hydrobiology, 19 (2), 307-315. https://doi.org/10.1016/j.ecohyd.2018.11.001

Bosquilia, R.W., Neale, C.M., Duarte, S.N., Longhi, S.J., Ferraz, S.F.D.B., Muller-Karger, F.E., Mccarthy, M.J., 2018b.Temporal evaluation of evapotranspiration for sugar cane, planted forest and native forest using Landsat 8 images and a two-source energy balance. Computers and Electronics in Agriculture, 151, 70-76. https://doi.org/10.1016/j.compag.2018.06.003

Boulet, G., Mougenot, B., Lhomme, J.P., Fanise, P., Lili-Chabaane, Z., Olioso, A., Bahir, M., Rivalland, V., Jarlan, L., Merlin, O., Coudert, B., Er-Raki, S., Lagouarde, J.-P., 2015. The SPARSE model for the prediction of water stress and evapotranspiration components from thermal infra-red data and its evaluation over irrigated and rainfed wheat. Hydrology and Earth System Sciences Discussions, 19, 4653-4672. https://doi.org/10.5194/hess-19-4653-2015

Burchard-Levine, V., Nieto, H., Riaño, D., Migliavacca, M., El-Madany, T. S., Perez-Priego, O., Carrara, A., Martín, M. P., 2020. Seasonal adaptation of the thermal-based two-source energy balance model for estimating evapotranspiration in a semiarid tree-grass ecosystem. Remote Sensing, 12 (6), 904. https://doi.org/10.3390/rs12060904

Castelli, M., Anderson, M.C., Yang, Y., Wohlfahrt, G., Bertoldi, G., Niedrist, G., Notarnicola, C., 2018. Two-source energy balance modeling of evapotranspiration in Alpine grasslands. Remote Sensing of Environment, 209, 327-342. https://doi.org/10.1016/j.rse.2018.02.062

Cawse-Nicholson, K., Anderson, M. C., 2018. ECOsystem Spaceborne Thermal Radiometer Experiment on Space Station (ECOSTRESS) Mission. California Institute of Technology.

Cawse-Nicholson, K., Braverman, A., Kang, E.L., Li, M., Johnson, M., Halverson, G., Anderson M., Hain, C., Gunson, M., Hook, S., 2020. Sensitivity and uncertainty quantification for the ECOSTRESS evapotranspiration algorithm–DisALEXI. International Journal of Applied Earth Observation and Geoinformation, 89, 102088. https://doi.org/10.1016/j.jag.2020.102088

Cecílio, R.A., Pimentel, S.M., Zanetti, S.S., 2019. Modeling the influence of forest cover on streamflows by different approaches. Catena, 178, 49-58. https://doi.org/10.1016/j.catena.2019.03.006

Cheng, J.D., Lin, L.L., Lu, H.S., 2002. Influences of forests on water flows from headwater watersheds in Taiwan. Forest Ecology and Management, 165 (1-3), 11-28. https://doi.org/10.1016/S0378-1127(01)00626-0

Choi, M., Kustas, W.P., Anderson, M.C., Allen, R.G., Li, F., Kjaersgaard, J.H., 2009. An intercomparison of three remote sensing-based surface energy balance algorithms over a corn and soybean production region (Iowa, US) during SMACEX. Agricultural and Forest Meteorology, 149(12), 2082-2097. https://doi.org/10.1016/j.agrformet.2009.07.002

Duan, Z, Bastiaanssen, W. G. M., 2013. First results from Version 7 TRMM 3B43 precipitation product in combination with a new downscaling–calibration procedure. Remote Sensing of Environment, 131, 1-13. https://doi.org/10.1016/j.rse.2012.12.002

Elfarkh, J., Ezzahar, J., Er-Raki, S., Simonneaux, V., Ait Hssaine, B., Rachidi, S., Brut, A., Rivalland, V., Khabba, S., Chehbouni, A., Jarlan, L., 2020. Multi-Scale Evaluation of the TSEB Model over a Complex Agricultural Landscape in Morocco. Remote Sensing, 12 (7), 1181. https://doi.org/10.3390/rs12071181

EMBRAPA - Empresa Brasileira de Pesquisa Agropecuária, 1979. In: Reunião de Levantamento de Solos. Rio de Janeiro, EMBRAPA. Available on http://www.bdpa.cnptia.embrapa.br/

Euclydes, H. P., Ferreira, P. A., Rubert, O. A. V., Santos, R. D., 2001. Regionalização hidrológica na bacia do alto São Francisco a montante da barragem de Três Marias, Minas Gerais. Revista Brasileira de Recursos Hídricos, 6 (2), 81-105. Available on https://www.abrhidro.org.br/SGCv3/publicacao.php?PUB=1&ID=41&SUMARIO=621

Geli, H.M., Neale, C.M., 2012. Spatial evapotranspiration modelling interface (SETMI). Remote Sensing and Hydrology, p. 171-174.

Gilfedder, M., Rassam, D. W., Stenson, M. P., Jolly, I. D., Walker, G. R., Littleboy, M. (2012). Incorporating land use changes and surface–groundwater interactions in a simple catchment water yield model. Environmental Modelling & Software, 38, 62-73. DOI https://doi.org/10.1016/j.envsoft.2012.05.005

Gharbia, S.S., Smullen, T., Gill, L., Johnston, Pilla, F., 2018. Spatially distributed potential evapotranspiration modeling and climate projections. Science of the Total Environment, 633, 571-592. https://doi.org/10.1016/j.scitotenv.2018.03.208

Granata, F. 2019. Evapotranspiration evaluation models based on machine learning algorithms -A comparative study. Agricultural Water Management, 217, 303-315. https://doi.org/10.1016/j.agwat.2019.03.015

Heuvelmans, G., Muys, B., Feyen, J., 2006. Regionalisation of the parameters of a hydrological model: Comparison of linear regression models with artificial neural nets. Journal of Hydrology, 319, (1-4), 245-265. https://doi.org/10.1016/j.jhydrol.2005.07.030

Hssaine, B.A., Ezzahar, J., Jarlan, L., Merlin, O., Khabba, S., Brut, A., Er-Raki, S., Elfarkh, J., Cappelaere, B., Chehbouni, G., 2018. Combining a two-source energy balance model driven by MODIS and MSG-SEVIRI products with an aggregation approach to estimate turbulent fluxes over sparse and heterogeneous vegetation in Sahel region (Niger). Remote Sensing, 10 (6), 974. https://doi.org/10.3390/rs10060974

Huffman, G.J., Bolvin, D.T., Nelkin, E.J., Tan J. 2019. Integrated Multi-satellite Retrievals for GPM (IMERG) Technical Documentation. IMERG Tech Document. Available on https://docserver.gesdisc.eosdis.nasa.gov/public/project/GPM/IMERG_doc.06.pdf

Kustas, W. P.; Norman, J. M. (2000). A two source energy balance approach using directional radiometric temperature observations for sparse canopy covered surfaces. Agronomy Journal, 92(5), 847-854. https://doi.org/10.2134/agronj2000.925847x

Li, F., Kustas, W., Prueger, J.H., Neale, C.M., Jackson, T.J., 2005. Utility of remote sensing–based two-source energy balance model under low-and high-vegetation cover conditions. Journal of Hydrometeorology, 6 (6), 878-891. https://doi.org/10.1175/JHM464.1

Li, X.H., Zhang, Q, Xu., C.Y., 2012. Suitability of the TRMM satellite rainfalls in driving a distributed hydrological model for water balance computations in Xinjiang catchment, Poyang lake basin. Journal of Hydrology, 426, 28-38. https://doi.org/10.1016/j.jhydrol.2012.01.013

Liu, J., Duan, Z., Jiang, J., Zhu, A. 2015. Evaluation of three satellite precipitation products TRMM 3B42, CMORPH, and PERSIANN over a subtropical watershed in China. Advances in Meteorology, 2015. https://doi.org/10.1155/2015/151239

Liu, W., Wang, L., Zhou, J., Li, Y., Sun, F., Fu, G., Li, X., Sang, Y. F., 2016. A worldwide evaluation of basin-scale evapotranspiration estimates against the water balance method. Journal of Hydrology, 538, 82-95. http://dx.doi.org/10.1016/j.jhydrol.2016.04.006

Martins, F.R., Pereira, E.B., Silva, S.A. B., Abreu, S.L. Colle, S., 2008. Solar energy scenarios in Brazil, Part one: Resource assessment. Energy Policy, 36 (8), 2853-2864. https://doi.org/10.1016/j.enpol.2008.02.014

Menezes, J.P.C., Bittencourt, R.P., Farias, M.D.S., Bello, I.P., Fia, R., Oliveira, L.F.C.D., 2016. Relação entre padrões de uso e ocupação do solo e qualidade da água em uma bacia hidrográfica urbana. Engenharia Sanitária e Ambiental, 21 (3), 519-534. http://dx.doi.org/10.1590/S1413-41522016145405

Moreno, H.A.; Vivoni, E.R.; Gochis, D.J., 2012.Utility of quantitative precipitation estimates for high resolution hydrologic forecasts in mountain watersheds of the Colorado Front Range. Journal of Hydrology, v. 438, p. 66-83. https://doi.org/10.1016/j.jhydrol.2012.03.019

Norman, J.M.; Kustas, W.P.; Humes, K.S. 1995. Source approach for estimating soil and vegetation energy fluxes in observations of directional radiometric surface temperature. Agricultural and Forest Meteorology, 77 (3-4), 263-293. https://doi.org/10.1016/0168-1923(95)02265-Y

Pagliero, L., Bouraoui, F., Diels, J., Willems, P., Mcintyre, N., 2019. Investigating regionalization techniques for large-scale hydrological modelling. Journal of Hydrology, 570, 220-235. https://doi.org/10.1016/j.jhydrol.2018.12.071

Priestley, C.H.B., Taylor, R.J., 1972. On the assessment of surface heat flux and evaporation using large-scale parameters. Monthly Weather Review, 100 (2), 81-92. https://doi.org/10.1175/1520-0493(1972)100<0081:OTAOSH>2.3.CO;2

Razavi, T., Coulibaly, P., 2012. Streamflow prediction in ungauged basins: review of regionalization methods. Journal of Hydrologic Engineering, 18 (8), 958-975. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000690

Robert B. Daugherty Water for Food Global Institute. (Year). GloDET: Global daily evapo-transpiration. University of Nebraska. Lincoln, NE, USA. <glodet.nebraska.edu>.

Santos, L.C., José, J.V., Bender, F.D., Alves, D.S., Nitsche, P.R., Reis, E.F., Coelho, RD., 2019. Climate change in the Paraná state, Brazil: responses to increasing atmospheric CO_2 in reference evapotranspiration. Theoretical and Applied Climatology, 1-14. https://doi.org/10.1007/s00704-019-03057-7

Šatalová, B., Kenderessy, P. 2017. Assessment of water retention function as tool to improve integrated watershed management (case study of Poprad river basin, Slovakia). Science of the Total Environment, 599, 1082-1089. https://doi.org/10.1016/j.scitotenv.2017.04.227

Semmens, K.A., Anderson, M.C., Kustas, W.P., Gao, F., Alfieri, J.G., McKee, L., Prueger, J.H., Hain, C.R., Cammalleri, C., Yang, Y., Xia, T., Sanchez L., Alsina, M.M., Vélezg, M., 2016. Monitoring daily evapotranspiration over two California vineyards using Landsat 8 in a multi-sensor data fusion approach. Remote Sensing of Environment, 185, 155-170. https://doi.org/10.1016/j.rse.2015.10.025

Singh, V.P., Yadav, S., Yadava, R.N., 2018. Hydrologic Modeling: Select Proceedings of ICWEES-2016. Vol. 81. Springer. Available on https://www.springer.com/gp/book/9789811058004

Sivapalan, M., Takeuchi, K., Franks, S.W., Gupta, V.K., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J.J., Mendiondo, E.M., O'Connell, P.E., Oki, T., Pomeroy, J.W., Schertzer, D., Uhlenbrook, S., Zehe E., 2003. IAHS Decade on predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. Hydrological Sciences Journal, 48 (6), 857-880. https://doi.org/10.1623/hysj.48.6.857.51421

Stisen, S., Sandholt, I., 2010. Evaluation of remote-sensing-based rainfall products through predictive capability in hydrological runoff modelling. Hydrological Processes: An International Journal, 24 (7), 879-891. https://doi.org/10.1002/hyp.7529

Swain, J. B., Patra, K. C., 2017. Streamflow estimation in ungauged catchments using regionalization techniques. Journal of Hydrology, 554, 420-433. https://doi.org/10.1016/j.jhydrol.2017.08.054

Wang, K., Dickinson, R. E., 2012. A review of global terrestrial evapotranspiration: Observation, modeling, climatology, and climatic variability. Reviews of Geophysics, 50 (2), 1-54. https://doi.org/10.1029/2011RG000373

Wang, C., Tang, G., Han, Z., Guo, X., Hong, Y. 2018. Global intercomparison and regional evaluation of GPM IMERG Version-03, Version-04 and its latest Version-05 precipitation products: Similarity, difference, and improvements. J. Hydrol., 564, 342-356. https://doi.org/10.1016/j.jhydrol.2018.06.064

Wolff, W., Duarte, S. N., Mingoti, R., 2014. Nova metodologia de regionalização de vazões, estudo de caso para o Estado de São Paulo. Revista Brasileira de Recursos Hídricos, 19 (4), 21-33. Available on https://www.abrhidro.org.br/SGCv3/publicacao.php?PUB=1&ID=173&SUMARIO=4887

Wu, Z., Zhang, Y., Sun, Z., Lin, Q., He, H., 2018. Improvement of a combination of TMPA (or IMERG) and ground-based precipitation and application to a typical region of the East China Plain. Science of the Total Environment, 640, 1165-1175. https://doi.org/10.1016/j.scitotenv.2018.05.272

Xu, S., Wu, C., Wang, L., Gonsamo, A., Shen, Y., Niu, Z. 2015. A new satellite-based monthly precipitation downscaling algorithm with non-stationary relationship between precipitation and land surface characteristics. Remote Sensing of Environment, 162, 119-140. https://doi.org/10.1016/j.rse.2015.02.024

Yang, X., Magnusson, J., Rizzi, J., Xu, C.Y., 2018. Runoff prediction in ungauged catchments in Norway: comparison of regionalization approaches. Hydrology Research, 49 (2), 487-505. https://doi.org/10.2166/nh.2017.071

Yenehun, A., Nigate, F., Belay, A. S., Desta, M. T., Van Camp, M., and Walraevens, K. (2020). Groundwater recharge and water table response to changing conditions for aquifers at different physiography: The case of a semi-humid river catchment, northwestern highlands of Ethiopia. Science of The Total Environment, 748, 142243. https://doi.org/10.1016/j.scitotenv.2020.142243

Zhang, L., Dawes, W.R., Walker, G.R., 2001. Response of mean annual evapotranspiration to vegetation changes at catchment scale. Water Resources Research, 37 (3), 701-708. https://doi.org/10.1029/2000WR900325

Zhang, K., Kimball, J. S., & Running, S. W. (2016). A review of remote sensing based actual evapotranspiration estimation. Wiley Interdisciplinary Reviews: Water, 3(6), 834-853. https://doi.org/10.1002/wat2.1168

Zhou, G., Wei, X., Chen, X., Zhou, P., Liu, X., Xiao, Y., Sun, G., Scott, D.F., Zhou, S., Han, L., Su, Y., 2015. Global pattern for the effect of climate and land cover on water yield. Nature communications, 6 (1), 1-9. https://doi.org/10.1038/ncomms6918

APPENDIX

Table 9. Landscape and climatic characteristics, and distribution adjusted of permanence curves for the 81 watersheds located in Paraná state, Brazil.

Continue…

Lon	Lat	Code	ID	Distribution	Area (km²)	ALT (m)	TS (%)	WS (%)	TL (km)	PPT (mm)	Clay (%)	Sand (%)	Cluster
-49.67	-24.3	64242000	1	Pearson 5 (3P)	1688.59	901.4	2.65	10.95	115658.02	1613.16	0.37	0.39	5
-48.91	-23.77	64270080	2	Burr	27760.94	808.83	1.49	9.22	394285	1584.23	0.36	0.41	5
-49.07	-23.48	64345075	3	Burr	39583.19	729.91	1.5	8.28	481945.73	1544.3	0.36	0.42	5
-49.12	-23.44	64345080	4	Burr	41330.69	642.2	1.46	8.13	515555.51	1545.05	0.36	0.42	5
-49.96	-24.1	64360000	5	Dagum	2016.69	565.99	2.27	11.5	120293.29	1592.8	0.38	0.37	5
-50.03	-23.84	64362000	6	Lognormal (3P)	4101.26	901.42	1.8	9.93	203688.57	1556.18	0.38	0.34	5
-50.04	-23.7	64370000	7	Log-Pearson 3	5612.88	652.93	1.76	9.89	262606.16	1536.66	0.39	0.34	5
-50.38	-23.69	64390000	8	Pearson 5	3443.69	929.59	2.11	10.43	238075.18	1542.62	0.39	0.3	5
-50.37	-24.81	64491000	9	Burr	16118.71	979.11	1.87	8.95	338881.55	1704.2	0.42	0.27	3
-50.6	-23.92	64491260	10	Fatigue Life	356.47	243.94	2.75	13.04	48110.58	1621.86	0.39	0.28	5
-50.75	-23.6	64501950	11	Lognormal (3P)	562.74	623.86	2.28	9.9	65300.37	1586.06	0.42	0.25	5
-51.23	-23.46	64504550	12	Burr (4P)	327.91	461.08	1.76	9.94	65543.16	1676.59	0.5	0.21	5
-50.53	-24.51	64507000	13	Gen. Gamma (4P)	22032.96	596.08	2.15	9.79	487155.46	1693.3	0.43	0.26	3
-51.16	-23.26	64507100	14	Dagum (4P)	181.51	857.89	1.96	6.75	28712.19	1658.32	0.35	0.16	5
-50.65	-23.45	64508020	15	Fatigue Life (3P)	941.36	668.5	2.23	9.59	97371.78	1527.68	0.41	0.29	5
-50.67	-23.41	64508500	16	Pearson 5 (3P)	1122.59	896.27	2	9.31	115722.12	1530.42	0.42	0.29	5
-51.14	-24.29	64652000	17	Gen. Gamma (4P)	2610.87	708.56	2.98	14.91	159765.66	1796.4	0.42	0.25	3
-51.27	-24.72	64655000	18	Fatigue Life (3P)	12701.24	677.06	2.05	13.06	344675.14	1851.62	0.48	0.21	1
-52.01	-24.41	64659000	19	Burr (4P)	3284.93	691.75	2.88	14.37	189138.04	1924.35	0.47	0.23	1
-51.46	-24.52	64660500	20	Fatigue Life	19433.3	933.62	1.94	12.93	414621.24	1852.36	0.48	0.21	1
-52.35	-24.12	64673000	21	Burr	1547.08	930.74	1.94	7.27	127534.36	1869.64	0.53	0.19	1
-51.57	-24.42	64675002	22	Fatigue Life (3P)	23102.31	935.68	1.88	12.13	463950.78	1844.65	0.48	0.21	1
-52.72	-23.67	64682000	23	Gen. Gamma (4P)	813.72	945.84	1.27	5.9	66844.19	1702.9	0.28	0.57	5
-51.73	-24.26	64685000	24	Fatigue Life (3P)	28404.66	926.2	1.77	11	553699.55	1821.47	0.45	0.26	1

Continue…

Lon	Lat	Code	ID	Distribution	Area (km²)	ALT (m)	TS (%)	WS (%)	TL (km)	PPT (mm)	Clay (%)	Sand (%)	Cluster
-52.8	-24.01	64810000	25	Pearson 6 (4P)	2039.88	931.73	1.26	6.2	106520.01	1777.33	0.39	0.38	3
-52.71	-24.6	64820000	26	Fatigue Life (3P)	17430.99	931.49	2.86	10.74	513961.19	1893.72	0.51	0.21	1
-52.85	-24.57	64830000	27	Log-Pearson 3	20962.35	895.16	2.69	9.92	571243.44	1875.04	0.51	0.21	1
-50.62	-25.36	64442800	28	Gen. Gamma (4P)	1361.38	623.31	1.15	7.65	73861.88	1742.36	0.49	0.21	3
-50.37	-25.27	64447000	29	Fatigue Life	5725.55	926.61	1.1	7.63	136813.95	1720.77	0.44	0.27	3
-50.08	-24.95	64453000	30	Dagum	1056.77	711.44	1.51	7.64	88361.86	1707.61	0.45	0.23	3
-50.36	-25.11	64465000	31	Gen. Gamma (4P)	8924.66	442.53	1.34	8.14	220179.13	1717.5	0.43	0.27	3
-49.93	-24.52	64477020	32	Dagum	210.09	1096.67	1.35	9.95	29401.17	1673.02	0.39	0.31	3
-50.98	-25.4	64619950	33	Lognormal (3P)	1051.3	862.21	2.49	10.4	54507.62	1810.39	0.5	0.19	3
-50.98	-25.39	64620000	34	Fatigue Life (3P)	1087.72	500.63	2.39	10.28	59555.73	1810.39	0.5	0.19	3
-52.36	-24.92	64776100	35	Fatigue Life	7656.53	952.6	3.32	14.69	369043.04	1974.21	0.55	0.15	1
-52.61	-24.44	64785000	36	Log-Pearson 3	1137.31	922.79	1.77	7.19	113964.56	1937.9	0.57	0.17	1
-53.22	-24.88	64790000	37	Dagum (4P)	697.21	957.52	2.35	9.42	97496.09	1914.25	0.57	0.15	1
-49.07	-25.41	65003950	38	Dagum (4P)	164.02	931.48	1.2	6.74	21339.21	1971.98	0.41	0.19	4
-49.07	-25.41	65004995	39	Pearson 5	165.04	1151.02	1.17	6.73	22675.8	1971.98	0.41	0.19	4
-49.17	-25.35	65006055	40	Burr (4P)	86.8	1055.62	1.14	6.03	23933.78	1881.69	0.35	0.16	4
-49.09	-25.42	65006075	41	Dagum (4P)	378.05	948.53	1.6	6.88	31994.7	1929.61	0.41	0.19	4
-49.23	-25.37	65007045	42	Log-Pearson 3	110.14	860.59	1.5	6.28	26233.34	1820.51	0.2	0.09	4
-49.12	-25.42	65009000	43	Dagum (4P)	559.76	761.64	1.47	6.28	35781.58	1918.81	0.35	0.16	4
-49.06	-25.57	65010000	44	Weibull (3P)	105.28	675.47	1.39	7.32	29126.69	1949.19	0.48	0.22	4
-49.11	-25.65	65015400	45	Fatigue Life (3P)	258.31	624.88	1.08	6.96	32101.52	1946.3	0.47	0.24	4
-49.14	-25.5	65017006	46	Lognormal (3P)	1154.32	598.17	1.12	5.96	56456.02	1919.5	0.35	0.17	4
-49.25	-25.72	65017035	47	Pearson 5 (3P)	63.29	913.24	1.08	5.76	20256.16	1873.83	0.43	0.33	4
-49.32	-25.42	65019700	48	Burr	260.36	901.55	1.52	6.52	54427.37	1810.77	0.26	0.13	4
-49.4	-25.37	65021770	49	Burr (4P)	24.38	534.69	1.87	8.26	13062.44	1748.82	0.5	0.24	4
-49.38	-25.43	65023000	50	Burr (4P)	166.97	788.94	2.04	7.81	34885.82	1756.94	0.4	0.24	4
-49.38	-25.44	65024000	51	Pearson 6	179.74	737.76	1.92	7.59	38516.07	1756.94	0.39	0.24	4
-52.8	-24.01	64810000	25	Pearson 6 (4P)	2039.88	931.73	1.26	6.2	106520.01	1777.33	0.39	0.38	3

Continue…

Lon	Lat	Code	ID	Distribution	Area (km²)	ALT (m)	TS (%)	WS (%)	TL (km)	PPT (mm)	Clay (%)	Sand (%)	Cluster
-49.68	-25.62	65060000	53	Fatigue Life	6016.7	832.87	1.12	7.12	218559.94	1781.37	0.41	0.26	4
-49.2	-26.07	65090000	54	Fatigue Life (3P)	772.22	589.07	1.94	10.43	62414.62	1873.25	0.47	0.25	4
-49.28	-26.1	65094500	55	Pearson 5 (3P)	1165.95	797.72	1.91	9.69	80172.84	1853.62	0.46	0.26	4
-49.24	-25.88	65135000	56	Fatigue Life (3P)	601.18	808.78	1.05	7.84	55200.95	1872.88	0.46	0.29	4
-49.31	-25.89	65136550	57	Lognormal (3P)	951.55	888.85	1.19	7.77	75004.7	1841.4	0.45	0.31	4
-50.09	-26.04	65310000	58	Fatigue Life (3P)	24046.51	845.75	1.22	8.37	352305.5	1805.56	0.48	0.22	2
-51.52	-26.23	65365000	59	Lognormal (3P)	71.29	616.81	2.54	2.54	14.93	2001.38	0.61	0.11	2
-51.02	-25.98	65415000	60	Inv. Gaussian	326.18	671	2.32	2.32	12.05	1838.03	0.55	0.15	2
-50.32	-26.05	65774400	61	Burr (4P)	29969.06	668.72	2.51	9.54	172819.72	1829.1	0.49	0.21	2
-51.34	-25.32	65809000	62	Inv. Gaussian (3P)	312.28	700.36	1.8	1.8	9.43	1899.28	0.63	0.09	2
-51.52	-25.38	65815050	63	Burr (4P)	2222.6	664.07	1.46	1.46	7.97	1950.19	0.63	0.09	2
-50.83	-25.95	65883051	64	Log-Logistic (3P)	43711.66	629.6	2.52	10.16	45803.98	1878.16	0.52	0.18	2
-52.64	-25.84	65883070	65	Burr (4P)	41.71	800.65	1.72	1.72	9.03	1976.25	0.6	0.13	2
-50.91	-25.93	65894991	66	Burr (4P)	45697.57	641.55	2.16	10.21	319355.16	1881.9	0.52	0.18	2
-50.91	-25.93	65894992	67	Burr	45700.02	888.62	2.16	10.21	320668.33	1881.9	0.52	0.18	2
-51.95	-26.43	65925000	68	Log-Pearson 3	1657.01	1046.71	1.76	1.76	7.12	2090.7	0.6	0.12	2
-53.05	-26.15	65955000	69	Fatigue Life (3P)	1725.46	937.91	2.36	2.36	10.92	2138.69	0.58	0.13	2
-52.57	-26.21	65960000	70	Inv. Gaussian (3P)	6693.92	1028.24	2.2	2.2	9.85	2111.35	0.59	0.13	2
-53.15	-25.24	65971010	71	Pearson 5 (3P)	385.07	910.87	2.69	12.96	65174.94	1982.79	0.57	0.14	1
-53.79	-25.9	65990550	72	Gamma (3P)	821.16	937.56	2.03	9.06	85121.83	2048.35	0.55	0.13	1
-54.47	-25.53	65996000	73	Log-Logistic (3P)	134.75	797.17	0.93	4.21	24630.32	1919.35	0.61	0.11	1
-49.65	-25.4	81019300	74	Dagum	211.96	918.67	3.93	16.61	28232.04	1689.58	0.44	0.27	4
-49.57	-25.27	81080000	75	Burr (4P)	1362.58	978.18	5.22	17.5	78880.42	1687.09	0.46	0.25	4
-49.55	-25.23	81102000	76	Pearson 5 (3P)	1688.96	866.18	4.92	17.91	107067.51	1682.58	0.46	0.25	4
-49.63	-25.11	81107000	77	Pearson 5 (3P)	3268.76	1076	4.87	16.56	115761.45	1681.96	0.46	0.25	4
-49.51	-24.68	81125000	78	Burr (4P)	418.57	882.57	3.47	14.23	52572.48	1663.14	0.42	0.34	4

Lon	Lat	Code	ID	Distribution	Area (km²)	ALT (m)	TS (%)	WS (%)	TL (km)	PPT (mm)	Clay (%)	Sand (%)	Cluster
-49.56	-25.05	81135000	79	Log-Pearson 3	4594.16	853.26	4.67	17.17	144779.08	1687.08	0.45	0.27	4
-49.05	-25.28	81299000	80	Dagum	562.93	892.48	2.58	12.11	53521.46	1871.87	0.5	0.22	4
-48.97	-25.78	82234000	81	Johnson SB	778.71	915.57	4.39	18.37	65931.26	1984.75	0.48	0.23	4

where Area= watershed area (km²), Alt= watershed mean altitude (m), PPT = mean annual precipitation (mm), WS = watershed mean slope (%), TS = main thalweg slope (%), and TL = length of the main thalweg (km).