Federal Public Service
Ministry of Education
Federal University of Mato Grosso do Sul
Postgraduate Program in Environmental Technologies

André Almagro

TOWARDS A BETTER UNDERSTANDING OF CATCHMENT HYDROLOGY IN BRAZIL

Campo Grande, MS
September, 2021

Federal University of Mato Grosso do Sul
Faculty of Engineering, Architecture and Urbanism, and Geography
Postgraduate Program in Environmental Technologies

André Almagro

TOWARDS A BETTER UNDERSTANDING OF CATCHMENT HYDROLOGY IN BRAZIL

Thesis submitted to the Faculty of Engineering, Architecture and Urbanism, and Geography, Federal University of Mato Grosso do Sul, in partial fulfillment of the requirements for the Degree of Doctor in Science: Water Resources and Environmental Sanitation.

Advisor: Prof. Dr. Paulo Tarso Sanches de Oliveira
Co-Advisor: Prof. Dr. Antônio Alves Meira Neto

Approved in: 03/09/2021

Examination board

Dr. Paulo Tarso Sanches de Oliveira
President

Prof. Dr. Edson Cezar Wendland
EESC/USP

Prof. Dr. Marcos Heil Costa
UFV

Dr. Noemi Vergopolan
Princeton University/NOAA

Prof. Dr. Walter Collischonn
IPH/UFRGS

Campo Grande, MS
September, 2021

DEDICATION

To my parents, my brother, and my wife, who always believed, encouraged, and supported me.

ACKNOWLEDGMENTS

First, I would like to thank my dad (João), my mom (Maria) and brother (Felipe) for the support and patience in all my life. Especially, I would like to thank my wife Lígia, who always believed, supported, helped, and took care of me. It was a long road until we got here, and you were by my side at all times, even when it would mean being thousands of kilometers far from you for 6 months.

To my advisor, Prof. Dr. Paulo Tarso, who is my greatest inspiration and supporter. Thanks for the friendship, orientation, support, availability, patience. You believed and made me believe in my potential. I will always be grateful to work with someone who always wants to share knowledge and push you, without vanity and selfishness. Know that you changed the lives of many people with your example of life. To my co-advisor, Dr. Antônio Alves Meira Neto, who is one of the smartest and coolest guys that I've ever met. Thanks so much for our always productive conversations and the opportunity you gave me to work with the best in the world in the hydrology field.

To Dr. Peter A. Troch (The University of Arizona) for receiving me in Tucson, AZ, during my sandwich program, and for being such an example of seriousness, competence, and scientist. I would like to extend my gratitude to my colleagues (especially Dr. Hannes Bauser and Dr. Minseok Kim) from the Hydrology and Atmospheric Sciences Department and Biosphere 2, from the University of Arizona, for the opportunity to work in one of the best hydrology departments in the world and also for the opportunity to get to know in detail the largest hydrology laboratory in the world.

I am also thankful for some of the scientists around the world that, directly or indirectly, contributed to the development of this work: Dr. Rafael Rosolem, Dr. Carlos Nobre, Dr. Stefan Hagemann, Dr. Luca Brocca, Dr. Thirtankar Roy, Dr. Murugesu Sivapalan, and Dr. Ross Woods. I am also deeply grateful to Prof. Dr. Teodorico Alves for the partnership, unconditional support, shared experiences, for being a friend since the undergraduate and with whom I learned so much about life, people, and science, and to Dr. Jamil Anache for being a friend and inspiration all through his academic journey. I am also grateful to Dr. José Marcato Junior for the support given through my doctorate and for helping in my sandwich program process.

To my partners in Hydrology, Erosion and Sediment Laboratory (HErosS) and Hydrology and Water Security research group: Dr. Dulce Rodrigues, Camila Couto, Carina Colman, Karina Pinheiro, Raquel Godoy. I am also thankful to my dear friend Ronaldo Ferreira, from NHL Stenden, who inspires me, has always encouraged me, for the partnership of years, in undergraduate, master's, and doctorate. Special Thanks to my friend from EESC/USP, Jullian Sone and Gabriela Gesualdo, for the conversations and for helping to process the CMIP6 projections, and to Pedro Zamboni for helping me with the machine learning coding and setup.

Finally, I would like to thank the professors, colleagues, technical staff, and secretaries from the FAENG and the Federal University of Mato Grosso do Sul, for the support during the doctoral course and to Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for the financial support.

“We cannot solve our problems with the same thinking we used when we created them”

(Albert Einstein)

RESUMO

Almagro, A. (2021). **Em direção a um melhor entendimento da hidrologia de bacia no Brasil**. Tese de Doutorado, Faculdade de Engenharias, Arquitetura e Urbanismo, e Geografia, Universidade Federal de Mato Grosso do Sul, Campo Grande, MS. Brasil.

A hidrologia de bacias no Brasil ainda é pouco explorada. O comportamento hidrológico, similaridades e funcionamento das bacias hidrográficas brasileiras são desconhecidos e devem ser investigados. É importante compreender os processos hidrológicos e as respostas às mudanças no clima para aumentar a capacidade de adaptação aos extremos hidrológicos. O principal objetivo desta tese de doutorado é aprimorar a compreensão da hidrologia de bacias no Brasil. A avaliação de GCMs/RCMs mostrou uma boa concordância (viés até 10%) das simulações anuais dos produtos downscaled sobre a Amazônia e Cerrado e grandes vieses (até 40%) no Pampa. Verificou-se que o modelo HadGEM2-ES é capaz de reproduzir médias de longo prazo para grandes áreas e que o Eta RCM melhorou as simulações do modelo MIROC5. Construiu-se o CABra, que é um conjunto de dados de longo prazo para 735 bacias hidrográficas brasileiras. Os resultados da avaliação dos produtos de sensoriamento remoto mostraram que eles são melhores que o ERA5 na estimativa da chuva. A vazão e assinaturas hidrológicas são mais bem estimadas com o SM2RAIN-ASCAT e GPM+SM2RAIN. A classificação das bacias mostrou a existência de seis grupos de bacias: "não-sazonais", "secas", "floresta tropical", "savana", "extremamente seca" e "extremamente úmida". A vazão nos grupos é controlada principalmente pelo índice de aridez. Os resultados desta tese de doutorado fornecem material de referência para impulsionar o estudo de hidrologia de bacias, exploração de novas hipóteses e, assim, avançar a compreensão do comportamento hidrológico das bacias.

Palavras-chave: bacia hidrográfica, classificação, agrupamento, satélite, atributos, mudanças climáticas, cenários, precipitação, vazão.

ABSTRACT

Almagro, A. (2021). **Towards a better understanding of the Catchment hydrology in Brazil.** Doctoral Thesis, Faculty of Engineering, Architecture and Urbanism, and Geography, Federal University of Mato Grosso do Sul, Campo Grande, MS. Brazil.

The catchment hydrology in Brazil is still poorly explored. The hydrological behavior, similarities, and functioning of Brazilian catchments are unknown and must be investigated. It is important to better understand the hydrological processes and responses to changes in climate to increase the ability to deal with hydrological extremes. The main objective of this doctoral thesis is to improve the understanding of catchment hydrology in Brazil. The GCMs/RCMs evaluation showed a good agreement (bias up to 10%) of downscaled annual simulations over the Amazon and Cerrado and large biases (reaching 40%) in the Pampa. I showed that HadGEM2-ES can represent long-term means for large areas and Eta RCM improves MIROC5 simulations. I also presented the CABra, which is a multi-source large-sample dataset including long-term data for 735 Brazilian catchments in eight attribute classes. The results of the satellite rainfall products evaluation showed that they performed better than ERA5 in estimating precipitation against ground observations. Streamflow and hydrologic signatures are better modeled with SM2RAIN-ASCAT and GPM+SM2RAIN. The catchment classification showed the existence of six groups of similar catchments: “non-seasonal”, “dry”, “rainforest”, “savannah”, “extremely-dry”, and “extremely-wet”. The streamflow into groups is mainly driven by the aridity index. The results found in this doctoral thesis provide benchmark material to benefit catchment hydrology investigations, exploration of new hypotheses and thereby advance our understanding of catchments’ behavior.

Keywords: catchment, large-sample, classification, clustering, satellite, catchment attributes, climate change, scenarios, precipitation, streamflow.

TABLE OF CONTENTS

ACKNOWLEDGMENTS .... **iv**

RESUMO .... **vii**

ABSTRACT .... **viii**

LIST OF FIGURES .... **xiii**

LIST OF TABLES .... **xix**

GENERAL INTRODUCTION **20**

1. Background and problem statement 20
2. Objectives 23

2.1. General objective 23

2.2. Specific objectives 23

3. Organization of thesis 24
4. References 25

CHAPTER 1 .... **27**

PERFORMANCE EVALUATION OF ETA/HADGEM2-ES AND ETA/MIROC5 PRECIPITATION SIMULATIONS OVER BRAZIL **27**

Abstract 27

1. Introduction 28
2. Material and methods 31

2.1. Brazil: Study area and biomes 31

2.2. Data acquisition and processing 33

2.3. Metrics to evaluate the simulated precipitation 34

2.4. Regional and spatial analysis 35

3. Results and discussion 36

3.1. Spatial patterns on annual and seasonal precipitation 36

3.2. Long-term means and annual variability at the biomes 41

3.3. Long-term mean seasonal precipitation 44

3.4. Investigating the origin of the biases 48

4. Conclusions 52
5. Acknowledgments 53
6. References 53

CHAPTER 2..**61**

CABRA: A NOVEL LARGE-SAMPLE DATASET FOR BRAZILIAN CATCHMENTS **61**

Abstract..61

1. Introduction..62

2. The CABra dataset..63

2.1. Overview..63

2.2. Catchment delineation and topography..65

2.2.1. Uncertainty and limitations..70

2.3. Climate..70

2.3.1. Methodology..70

2.3.2. Results and discussion..75

2.3.3 Uncertainty and limitations..78

2.4. Streamflow and hydrologic signatures..78

2.4.1. Methodology..78

2.4.2. Results and discussion..81

2.4.3. Uncertainty and limitations..84

2.5. Groundwater..84

2.5.1. Methodology..84

2.5.2. Results and discussion..85

2.5.3. Uncertainty and limitations..87

2.6. Soil..88

2.6.1. Methodology..88

2.6.2. Results and discussion..89

2.6.3. Uncertainty and limitations..91

2.7. Geology..91

2.7.1. Methodology..91

2.7.2. Results and discussion..92

2.7.3. Uncertainty and limitations..94

2.8. Land-cover..94

2.8.1. Methodology..94

2.8.2. Results and discussion..96

2.8.3. Uncertainty and limitations 98
2.9. Hydrologic disturbance 99
2.9.1. Methodology..99
2.9.2. Results and discussion 101
2.9.3. Uncertainty and limitations 104
3. Comparison with CAMELS-BR and broader implications for hydrological studies 105
4. Conclusions 109
5. Acknowledgments..110
6. References 110
CHAPTER 3...123
ASSESSMENT OF BOTTOM-UP SATELLITE RAINFALL PRODUCTS IN ESTIMATING RIVER DISCHARGE AND HYDROLOGIC SIGNATURES IN BRAZILIAN CATCHMENTS 123
Abstract...123
1. Introduction 124
2. Study area..126
3. Datasets 127
3.1. Ground observations 127
3.2. Rainfall products..128
4. Methodology 129
4.1. Hydrologic modeling 129
4.2. Hydrologic signatures 131
4.3. Performance evaluation on precipitation estimates and hydrologic modeling 132
5. Results and discussion...133
5.1. Precipitation estimates over Brazil...133
5.2. Streamflow estimates on Brazilian catchments...136
5.3. Reliability of remote sensing products on hydrologic signatures 145
6. Conclusions 148
7. Acknowledgments..149
8. References 149
CHAPTER 4...154
THE DOMINANT ATTRIBUTES OF STREAMFLOW VARIABILITY ON

BRAZILIAN CATCHMENTS **154**
Abstract 154
1. Introduction 155
2. Material and methods 157
2.1. Study area and dataset 157
2.2. Hydrologic signatures 158
2.3. Cluster analysis to group by streamflow variability similarities 159
2.4. Identifying dominant attributes to streamflow variability by random forest analysis ... 160
3. Results and discussion 163
3.1. Groups with similar hydrologic behavior 163
3.2. Dominant attributes to streamflow variability in Brazilian catchments 171
3.2.1. Group 1 – Non-seasonal catchments 172
3.2.2. Group 2 – Dry catchments 174
3.2.3. Group 3 – Rainforest catchments 175
3.2.4. Group 4 – Savannah catchments 176
3.2.5. Group 5 – Extremely-dry catchments 177
3.2.6. Group 6 – Extremely-wet catchments 179
4. Conclusions 180
5. Acknowledgments 182
6. References 182
GENERAL CONCLUSIONS **188**
APPENDIX **191**

LIST OF FIGURES

CHAPTER 1...**27**

Figure 1: Brazilian biomes map and their respective mean monthly precipitation shown as bar plots. Monthly means were calculated for the 1980-2013 period from daily dataset originally developed by Xavier et al. (2016). The red lines represent a threshold used as a criterion to identify the dry months (< 100 mm)...32

Figure 2: Spatial distribution of the annual precipitation (P) over the Brazilian biomes for a) Observed dataset; b) HadGEM2-ES dataset; c) Eta/HadGEM2-ES dataset; d) MIROC5 dataset; and e) Eta/MIROC5 dataset...37

Figure 3: Spatial distribution of the PBIAS on annual precipitation (P) and on the annual coefficient of variation (CV) over the Brazilian biomes. a) and e) represents the PBIAS and CV on HadGEM2-ES simulations; b) and f) represents the PBIAS and CV on Eta/HadGEM2-ES simulations; c) and g) represents the PBIAS and CV on MIROC5 simulations; and d) and h) represents the PBIAS and CV on Eta/MIROC5 simulations. 39

Figure 4: Absolute biases (BIAS) in each season (DJF, MAM, JJA and SON) simulated precipitation in Brazilian biomes. a) to d) represent the BIAS for the HadGEM2-ES for all seasons; e) to h) represent the BIAS for the Eta/HadGEM2-ES for all seasons; i) to l) represent the BIAS for the MIROC5 for all seasons; and m) to p) represent the BIAS for the Eta/MIROC5 for all seasons. Shades of blue indicate a positive BIAS while shades of red indicate a negative BIAS. 41

Figure 5: Long-term mean monthly rainfall for the 1980-2005 period for observations (black dashed line), HadGEM2-ES (dark blue line), Eta/HadGEM2-ES (light blue line), MIROC5 (dark red line), and Eta/MIROC5 (light red line) simulations in the a) Amazon, b) Atlantic Forest, c) Cerrado, d) Caatinga, e) Pampa, and f) Pantanal. The red reference line represents a threshold used as a criterion to identify the dry season (< 100 mm).......................................43

Figure 6: Annual characteristics of precipitation over the Brazilian biomes for the 1980-2005 period. The mean annual precipitation (a) is presented in absolute values to differ the magnitudes between the biomes. The annual variability (b) is calculated by the division of annual standard deviation by the annual mean precipitation. 44

Figure 7: PBIAS on the long-term seasonal precipitation for the 1980-2005 period simulated by models against observations in the a) Amazon, b) Atlantic Forest, c) Cerrado, d) Caatinga, e) Pampa, and f) Pantanal. 45

Figure 8: Taylor diagrams of seasonal mean precipitation over a) Amazon, b) Atlantic Forest, c) Cerrado, d) Caatinga, e) Pampa, and f) Pantanal for simulations and observations. Means of the observed seasonal precipitation are marked as a black star. The azimuth and the radial distance from the origin of the plot represents the correlation coefficient and the standard deviation (mm) of simulated data in relation to the observed value, respectively. 47

Figure 9: Mean bias error for precipitation simulations of a) HadGEM2-ES, b) Eta/HadGEM2-ES, c) MIROC5, and d) Eta/MIROC5 for the 1980-2005 period...48

Figure 10: Heatmap of the relative error ($\in j$) of precipitation properties for Brazilian biomes

(Amazon – AMZ, Atlantic Forest – MAT, Cerrado – CER, Caatinga – CAA, Pampa – PAM, and Pantanal – PAN). The lower the error value, the better the model represents the dry season, seasonal and annual precipitation. The “Overall” refers to an integrated evaluation of models to simulate all the previous properties. 50

CHAPTER 2........ 61

Figure 1: Study delineation for the CABra dataset organization. From ANA’s database, 735 were selected to integrate our dataset due to its high consistency and long time series of streamflow. 64

Figure 2: Location map of the streamflow gauges and CABra catchments. a. Streamflow gauges coordinates of CABra catchments; b. The 735 CABra catchments boundaries; c. The 12 hydrographic regions of Brazil; d. The six main biomes of Brazil; e. Level of consistency of the streamflow gauges records for each biome. 66

Figure 3: Spatial distribution of the topography attributes of the CABra catchments. a. Stream order of Brazilian rivers; b. Area of the catchments, in km²; c. Mean elevation of the catchments, in m; d. Mean slope of the catchments, in percentage; e. Elevation of the streamflow gauge, in m. 69

Figure 4: Spatial distribution of climate indices of the CABra catchments. a. Mean daily precipitation, in mm day^{-1}; b. Precipitation seasonality, dimensionless; c. Aridity index, dimensionless; d. Mean daily solar radiation, in MJ m^{2} day^{-1}; e. Mean daily minimum temperature, in ºC; f. Mean daily maximum temperature, in ºC. 76

Figure 5: Spatial distribution of the PET calculated from three different methods of the CABra catchments. a. Penman-Monteith method; b. Priestley and Taylor method; c. Hargreaves method. 77

Figure 6: Distribution of the CABra catchments in the Budyko framework from the three different climate dataset of CABra: REF, ERA5 and ENS. Values of E were estimated from the relation P = E + Q, considering long-term means. 77

Figure 7: Spatial distribution of the hydrological signatures of the CABra catchments. a. Mean daily streamflow, in mm day^{-1}; b. Runoff ratio, dimensionless; c. Baseflow index, dimensionless; d. half-flow day, in day of the year; e. The slope of the flow duration curve, dimensionless; f. Elasticity of daily streamflow, dimensionless; g. Low streamflow, in mm day^{-1}; h. Frequency of low streamflow events, in days year^{-1}; i. Duration of low streamflow events, in days; j. High streamflow, in mm day^{-1}; k. Frequency of high streamflow events, in days year^{-1}; l. Duration of high streamflow events, in days. 83

Figure 8: Spatial distribution of the groundwater attributes of the CABra catchments. a. Water table depth, in m; b. Height Above Nearest Drainage, in m; c. Type of aquifer bedrock. d. Number of experimental wells; e. Static level, in m; f. Dynamic level, in m. 87

Figure 9: Spatial distribution of the soil attributes of the CABra catchments. a. The most common type of soil in the catchment; b. The class of texture based on USDA classification; c. The clay fraction of the soil, in percentage; d. The sand fraction of the soil, in percentage; e. The silt fraction of the soil, in percentage; f. The organic carbon content of the soil, in permille; g. The bulk density of the soil, in g cm^{-3}; h. The depth to soil bedrock, in m. 90

Figure 10: Spatial distribution of geology attributes of the CABra catchments. a. Most common lithology class in the catchment; b. Porosity, dimensionless; c. Saturated permeability, in m^2; d. Saturated hydraulic conductivity, in $m\ s^{-1}$. 93

Figure 11: Spatial distribution of the land-cover and land-use attributes of the CABra catchments. a. The most common land-cover type in the catchment; b. Forest fraction of land-cover, in percentage; c. Grass fraction of land-cover, in percentage; d. Shrub fraction of land-cover, in percentage; e. Crops fraction of land-cover, in percentage; f. Urban fraction of land-cover, in percentage. 97

Figure 12: Spatial distribution of the seasonal NDVI of the CABra catchments. a. NDVI in summer season (DJF); b. NDVI in autumn season (MAM); c. NDVI in the winter season (JJA); d. NDVI in the spring season (SON). 98

Figure 13: Spatial distribution of the hydrologic disturbance attributes of CABra catchments. a. Distance from urban cover to the streamflow gauge, in km; b. Urban fraction of land-cover, in percentage; c. Crops fraction of land-cover, in percentage; d. The number of reservoirs in the catchment; e. Reservoir fraction of land-cover, in percentage; f. The total volume of the reservoirs in the catchment, in km^3; g. The capacity of the reservoirs in the catchment to regulate the mean annual streamflow, dimensionless; h. Multi-purpose water demand in the catchment, in mm $year^{-1}$; i. Hydrologic disturbance index (HDI) of the catchment, dimensionless. The HDI is a weighted relationship between all the anthropogenic factors of the catchments. 103

Figure 14: Hydrological signatures as predictors of the Hydrological Disturbance Index. The random forest regressor algorithm assess how much each signature increase the error of a HDI prediction when randomly sorted. The higher the deviaton caused by a predictor, the higher is the influence of the hydrological signature on the HDI. 104

Figure 15: Scatter plots and correlation coefficients between hydrological signatures of CABra and CAMELS-BR catchments. There were 607 catchments and 13 hydrological signatures overlapped in both datasets..107

CHAPTER 3...123

Figure 1: Location map containing the 520 catchments in which the remote sensing products were evaluated. For each biome, the long-term precipitation and streamflow are represented in the subplots. The shaded area represents the range of daily streamflow for a given biome. ..127

Figure 2: Spatial distribution of performance scores for the three different rainfall products (SM2RASC, GPMSM2R, and ERA5) in estimating daily precipitation over Brazil: a), b) and c) bias; d), e) and f) root mean square error (RMSE); and g), h) and i) correlation coefficient with a 0.05 significance level. 134

Figure 3: Radar chart for area-averaged performance scores of rainfall products on Brazilian biomes. Best values of the area-averaged bias (aBIAS), area-averaged root mean squared error (aRMSE), and area-averaged correlation coefficient (aCC) are converging to the center of the triangle, making the lower the area, the better the performance. 135

Figure 4: Performance scores for the three different rainfall products in simulating daily river discharge in Brazilian catchments in calibration phase. a), b) and c) are the mean bias (BIAS-Q) for SM2RASC, GPMSM2R, and ERA5, respectively. d), e) and f) are the root mean squared error (RMSE-Q) for SM2RASC, GPMSM2R, and ERA5, respectively. g), h) and i) are the

Nash-Sutcliffe Efficiency (NSE-Q) for SM2RASC, GPMSM2R, and ERA5, respectively. j), k) and l) are the Kling-Gupta Efficiency (KGE-Q) for SM2RASC, GPMSM2R, and ERA5, respectively. 137

Figure 5: Performance scores for the three different rainfall products in simulating daily river discharge in Brazilian catchments in validation phase. a), b) and c) are the mean bias (BIAS-Q) for SM2RASC, GPMSM2R, and ERA5, respectively. d), e) and f) are the root mean squared error (RMSE-Q) for SM2RASC, GPMSM2R, and ERA5, respectively. g), h) and i) are the Nash-Sutcliffe Efficiency (NSE-Q) for SM2RASC, GPMSM2R, and ERA5, respectively. j), k) and l) are the Kling-Gupta Efficiency (KGE-Q) for SM2RASC, GPMSM2R, and ERA5, respectively. 139

Figure 6: Box charts for comparison between calibration and validation phases of hydrological modeling using the MISDc model and the three different rainfall products. a) BIAS-Q results for hydrological modeling; b) RMSE-Q results for hydrological modeling; c) NSE-Q results for hydrological modeling; d) KGE-Q results for hydrological modeling....................................140

Figure 7: Performance scores, in terms of KGE, during the calibration period for the SM2RAIN-ASCAT, GPM+SM2RAIN, and ERA5, with varying areas of the catchment. The histograms are representing the number of catchments with <1,000 km², 1,000 – 10,000 km², 10,000 – 100,000 km², and >100,000 km²...141

Figure 8: Daily river discharge simulated during calibration and validation phases by MISDc with SM2RASC (red line), GPMSM2R (blue line), and ERA5 (green line) rainfall products as inputs. The Black dashed line is the observed daily river discharge. a) and b) Daily river discharge for "Óbidos" streamflow gauge. c) and d) Daily river discharge for "Araguantins" streamflow gauge. e) and f) Daily river discharge for the "Vila Urucuia" streamflow gauge. g) and h) Daily river discharge for "Campos (Ponte Municipal)" streamflow gauge..................143

Figure 9: Performance of the satellite and reanalysis rainfall products against ground-based rainfall products during the calibration period. a) to c) are the results of the Skill Score (SS), dimensionless, of the SM2RAIN-ASCAT, GPM+SM2RAIN, and ERA5 rainfall products, respectively, against Xavier et al. (2016) dataset. d) shows the best performance between SM2RAIN-ASCAT, GPM+SM2RAIN, and ERA5 in estimating daily river discharge in Brazilian catchments...145

Figure 10: Hydrologic signatures from daily river discharge simulated by MISDc with ground-based, SM2RASC, GPMSM2R, and ERA5 rainfall products as inputs. The panel on the left side represents the ability of the rainfall product to represent each hydrologic signature, in terms of coefficient of determination, for a given biome. The panel on the right side represents a rainfall product ranking of ability to represent the hydrologic signature in each biome. 147

CHAPTER 4...154

Figure 1: Random forest basic principles and functioning (Wang et al., 2015)......................162

Figure 2: Groups with similar hydrological behavior, based on a catchment classification of their hydrological signatures. A) is the spatial distribution of the six groups over Brazilian biomes: Amazon, Atlantic Forest, Cerrado, Caatinga, Pampa, and Pantanal; b) is the biplot of the principal component analysis of the hydrological signatures from the six hydrological groups; c) is the distribution of the hydrological groups in the space of Budyko...............................163

Figure 3: Spatial distribution (upper panels) and daily regime curves (lower panels) from water balance components (precipitation, actual evapotranspiration, streamflow, and changes in water storage) for each of the hydrological groups. A) is the Non-seasonal group; b) is the Dry group; c) is the Rainforest group; d) is the Savannah group; e) is the Extremely-dry group; and f) is the Extremely-wet group all the units are in mm. 165

Figure 4: Boxplots of the hydrological signatures characterizing the catchments within each of the six hydrological groups. A) is the mean streamflow, in mm.day-1; b) is the runoff coefficient, dimensionless; c) is the coefficient of variation of the streamflow, dimensionless; d) is the baseflow index, dimensionless; e) to h) are the percentiles of daily streamflow, in mm.day-1; i) and j) are the frequency and duration of low-flows, respectively; k) and l) are the frequency and duration of the high-flows, respectively; m) is the frequency of zero-flow; n) is the half-flow day; and o) is the streamflow elasticity to the precipitation, dimensionless. 169

Figure 5: Boxplots of catchment attributes signatures characterizing the catchments within each of the six hydrological groups. A) to c) are the attributes related to the topography: the area, in km², the mean elevation, in m, and the slope, in %; d) to f) are the attributes related to the climate: mean precipitation, in mm.day-1, the precipitation seasonality, dimensionless, and the aridity index, dimensionless; g) to i) are the attributes related to the soil: fraction of sand, in %, the bulk density, in %, and the soil depth to the bedrock, in m; j) to l) are the attributes related to the geology: height above nearest drainage, in m, subsurface porosity, dimensionless, and the subsurface permeabilty, in m²; m) to p) are the attributes related to the land-cover: the fractions of forest cover, crops cover, and grass cover, respectively, in %, and the NDVI during the summer season, dimensionless; and q) to r) are the attributes related to the anthropogenic modification: the hydrological disturbance index, dimensionless, and the number of reservoirs in the catchment...170

Figure 6: Diagram of the classes of attributes controlling the streamflow variability through the hydrological groups in Brazilian catchments. 172

Figure 7: Diagram of the main controls of hydrological behavior of Group 1 – Non-seasonal catchments. The size of each circle indicates the importance of a given catchment attribute to a given hydrological signature (the higher the circle, the higher the importance). The color of each circle indicates the correlation between a given catchment attribute to a given hydrological signature...173

Figure 8: Diagram of the main controls of hydrological behavior of Group 2 – Dry catchments. The size of each circle indicates the importance of a given catchment attribute to a given hydrological signature (the higher the circle, the higher the importance). The color of each circle indicates the correlation between a given catchment attribute to a given hydrological signature. 175

Figure 9: Diagram of the main controls of hydrological behavior of Group 3 – Rainforest catchments. The size of each circle indicates the importance of a given catchment attribute to a given hydrological signature (the higher the circle, the higher the importance). The color of each circle indicates the correlation between a given catchment attribute to a given hydrological signature...176

Figure 10: Diagram of the main controls of hydrological behavior of Group 4 – Savannah catchments. The size of each circle indicates the importance of a given catchment attribute to a

given hydrological signature (the higher the circle, the higher the importance). The color of each circle indicates the correlation between a given catchment attribute to a given hydrological signature....177

Figure 11: Diagram of the main controls of hydrological behavior of Group 5 – Extremely-dry catchments. The size of each circle indicates the importance of a given catchment attribute to a given hydrological signature (the higher the circle, the higher the importance). The color of each circle indicates the correlation between a given catchment attribute to a given hydrological signature....178

Figure 12: Diagram of the main controls of hydrological behavior of Group 6 – Extremely-wet catchments. The size of each circle indicates the importance of a given catchment attribute to a given hydrological signature (the higher the circle, the higher the importance). The color of each circle indicates the correlation between a given catchment attribute to a given hydrological signature....180

LIST OF TABLES

CHAPTER 1 27

Table 1: Properties and the variables (k) considered to calculate the errors (Ek, j) for each property..35

CHAPTER 2 61

Table 1: General attributes of the CABra catchments. 65

Table 2: Topography attributes of the CABra catchments. 67

Table 3: Daily series of meteorological variables and climate indices for the CABra catchments. 74

Table 4: Hydrological signatures of the CABra dataset. 80

Table 5: Groundwater attributes of the CABra catchments. 85

Table 6: Soil attributes of the CABra catchments. 89

Table 7: Geology attributes of CABra catchments. 92

Table 8: Land-cover attributes of CABra catchments. 95

Table 9: Hydrologic disturbance attributes of CABra catchments. 101

CHAPTER 3 123

Table 1: Description of the MISDc parameters and initial conditions range for calibration. .130

Table 2: Hydrological signatures chosen for this study. 131

Table 3: Area-averaged performance scores of rainfall products on Brazilian biomes on simulating precipitation. Best scores for each biome are highlighted in bold. Values of Observed rainfall are in mm.day-1, aBIAS are in mm.day-1 (and in %), aRMSE are in mm.day-1 (and in %), and aCC are dimensionless. 135

CHAPTER 4 154

Table 1: Hydrological signatures chosen for this study. 158

Table 2: Catchment attributes chosen for dominant attributes analysis. 161

GENERAL INTRODUCTION

1. Background and problem statement

The catchment is the main unit of interest in the hydrology field. They are complex poorly defined systems that present significant variability in space, time, and processes (McDonnell and Woods, 2004), mostly due to the uniqueness of each place (Beven, 2000). Even so, a catchment presents at least a level of self-organization, in where its geomorphology, soils, and vegetation are adaptive to (and a result of) the landscape co-evolution (Dooge, 1986; Blöschl and Sivapalan, 1995; Sivapalan, 2005; Troch *et al.*, 2013). The identification and classification of the catchment's hydrological behavior is the first step towards a better understanding and comprehension of many complex levels involved in catchment hydrological processes (McDonnell and Woods, 2004), providing insights into hydrological behavior and functioning of the catchments (Wagener *et al.*, 2007). Decoding patterns in observations inevitably relies on a catchment classification capable of predicting the dominant controls on the water fluxes (Sivapalan, 2005), providing structure to hydrology science (Wagener *et al.*, 2008). One of the primary steps to investigate the hydrological behavior is data collection. As much as we can collect, organize and process hydrological data, further we reach in hydrological science (Beven, 2000).

When the large-sample approach is employed, it's possible to identify the dominant hydrological functions e verify the similarity between catchments, grouping and categorizing them (Lyon e Troch, 2010; Wagener et al., 2007). The degree of similarity (or dissimilarity) between catchments is a key factor to understand why and how certain patterns or hydrological behavior occurs (Gottschalk, 1985). The understanding of hydrological functions and their causes is essential to classify, transfer information to ungauged catchments, develop theories and explain a phenomenon, and assess environmental changes impacts, such as climate and land-use and land-cover changes (Sawicz *et al.*, 2011).

Regardless of being one of the most important countries to the global water fluxes, Brazil has a scarce allocation of funding for hydro-meteorological monitoring, which creates great challenges for proper knowledge and monitoring of its water resources, including precipitation. There is a lack of gauge-stations for precipitation monitoring, as shown in Xavier et al. (2016), with stations density lower than recommended by the World Meteorological Organization (WMO), making the satellite rainfall products an important tool for water resources monitoring in Brazil. However, few studies (Paredes-Trejo *et al.*, 2018, 2019) investigated the suitability of these products in simulating the rainfall in Brazil, and none addressed broader hydrological applications, such as river discharge and hydrologic signatures estimations across the country. This kind of continuous monitoring would enable us to construct a continuous and high-quality large-sample dataset for hydrological studies.

In a further stage of hydrological behavior assessment, we should insert the climate change projections to investigate the expected trajectories of the streamflow for a given scenario. The hydroclimatic variables widely used to determine the hydrological behavior can be projected for future periods by complex Earth system models under specific scenarios, the Global Climate Models (GCMs). The GCMs simulate, on a large scale, the atmospheric dynamics with acceptable accuracy and have been used as the primary tool of the scientific community to support climate change studies. However, to provide large-scale global datasets, the GCMs have a coarse spatial resolution (~100-250km), making unsuitable their applicability to local-scale studies. To address this problem, these datasets need to be downscaled by the employment of Regional Climate Models (RCM). Thus, simulations and projections are capable of providing detailed information, according to local or regional forcings (Giorgi, 1990). Therefore, the ability of GCM/RCM in simulating historical spatial distribution and patterns of climatologies tells us a lot about their projections for the future, making it indispensable to investigate the possible biases inherent to these products.

The GCMs provide hydrometeorological data projecting the possible climate trajectories through the 21st century, following emission scenarios proposed for the Intergovernmental Panel on Climate Change (IPCC) Assessment Reports (AR). By the employment of the projected climate change hydrometeorological data into hydrological models, we can estimate the expected trajectory of streamflow (and its components) for each emission scenario. This is especially important and valuable when a large-sample dataset is available. We can use this large-scale dataset to project changes in the hydrological behavior of catchments (Hoomehr *et al.*, 2016). The knowledge of climate change impacts becomes more important when there is a contrary trend between water availability and water consumption in strategic catchments, as shown in Rodrigues et al. (2015). Few studies investigated the performance of GCMs/RCMs in simulating the rainfall patterns (Chou *et al.*, 2014a, 2014b; Avila-Diaz *et al.*, 2020) and none of them employed the most recent climate change products available for CMIP6 to investigate the impacts of a changing climate over the hydrological behavior of Brazilian catchments.

In Brazil, there is a gap in large-sample studies looking for the hydrological behavior, similarity degree, and categorization of catchments, especially because it took a long time to build large-sample databases of catchment attributes and hydrometeorological data. Additionally, there is a lack in search for new products and technologies to continuously monitoring the water resources and to investigate the projected climate change impacts over the water resources. Due to the importance of water resources to water-food-energy security in Brazil, a deeper investigation to understand the hydrological behavior of the Brazilian catchments and the projected impacts is needed. In this context, in this doctoral thesis, I seek to provide an overview and a better understanding of the catchment hydrology in Brazil, by answer the following questions: how suitable are the GCMs and RCMs in simulating the rainfall over Brazil? How do satellite rainfall products perform in simulating the rainfall in Brazil? How accurate is to estimate the streamflow from hydrological modeling using satellite rainfall products? How do the catchments in Brazil behave? Is there a similarity between catchment behavior that allows to classify and group the catchments? What are the main drivers of streamflow variability in the catchment groups? To address these questions, I worked with multi-source (observed, remote sensing, reanalysis, climate projections), multi-scale (local, regional, continental), and diverse available technology (GIS, machine learning, programming) datasets.

2. Objectives

2.1. General objective

The main objective of this study is to investigate the hydrological behavior of Brazilian catchments, their similarities, and their responses to climate change.

2.2. Specific objectives

i. To evaluate the performance of climate change models (GCM and RCM) in representing hydroclimatic variables by their simulations and correct the systematic errors (bias).
ii. To collect, process, synthesize and create a large-sample dataset of hydrometeorological and geophysical data from Brazilian catchments.
iii. To calibrate a conventional hydrological model and test satellite rainfall products' ability to simulate the daily river discharge in Brazilian catchments.
iv. To determine the actual hydrological behavior of the Brazilian catchments by a catchment classification using hydrological signatures grouping them by hydrological similarity.
v. To determine the main drivers and controls of the streamflow variability in each catchment group.

3. Organization of thesis

The thesis is organized into four chapters. **Chapter 1** provides an overview of the hydrometeorological data of the two widely used RCMs (Eta/HadGEM2-ES and Eta/MIROC5) and their driver GCMs (HadGEM2-ES and MIROC5) in the Brazilian territory. We evaluated the quality of simulations during the historical period (1980-2005), comparing with observed data and assessing their ability to capture the main features of the rainfall in seasonal and annual scales, which is of main importance for further climate change investigations. Aside from identifying their systematic errors, we investigated the possible causes and indicated the best products for each Brazilian biome.

To create a novel large-sample dataset for Brazilian catchments, in **Chapter 2** we collected, processed, and synthesized more than 100 attributes for 735 catchments in Brazil. Along with the attributes, we provided daily measurements of hydrological and meteorological data for 30 years of record (1980-2010), catchment boundaries, area, and drainage data. The collection and processing methods are discussed along with the limitations for each of our multiple data sources.

In order to keep continuous monitoring of hydrometeorological data in the Brazilian catchments, in **Chapter 3**, we assessed the ability of two satellite rainfall products (SM2RAIN-ASCAT and GPM+SM2RAIN) in simulating the precipitation over Brazil. Moreover, we calibrated and validated a conventional hydrological model – Modello Idrologico SemiDistribuito in continuo (MISDc) – to evaluate the reliability of this approach to estimate the daily river discharge and hydrological signatures in Brazilian catchments.

Finally, in **Chapter 4**, we performed a catchment classification over Brazilian catchments based on their hydrological behavior similarity. To do so, we used a clustering method that considered 15 hydrological signatures of the catchments to perform the grouping. Additionally, using a set of 18 catchment attributes and a random forest algorithm, we investigated the main drivers of the hydrological behavior through the catchment groups.

4. References

Avila-Diaz A, Benezoli V, Justino F, Torres R, Wilson A. 2020. Assessing current and future trends of climate extremes across Brazil based on reanalyses and earth system model projections. *Climate Dynamics* **55** (5–6): 1403–1426 DOI: 10.1007/s00382-020-05333-z

Beven KJ. 2000. Uniqueness of place and process representations in hydrological modelling. *Hydrology and Earth System Sciences* **4** (2): 203–213 DOI: 10.5194/hess-4-203-2000

Blöschl G, Sivapalan M. 1995. Scale issues in hydrological modelling: A review. *Hydrological Processes* **9** (3–4): 251–290 DOI: 10.1002/hyp.3360090305

Chou SC, Lyra A, Mourão C, Dereczynski C, Pilotto I, Gomes J, Bustamante J, Tavares P, Silva A, Rodrigues D, et al. 2014a. Evaluation of the Eta Simulations Nested in Three Global Climate Models. *American Journal of Climate Change* **03** (05): 438–454 DOI: 10.4236/ajcc.2014.35039

Chou SC, Lyra A, Mourão C, Dereczynski C, Pilotto I, Gomes J, Bustamante J, Tavares P, Silva A, Rodrigues D, et al. 2014b. Assessment of Climate Change over South America under RCP 4.5 and 8.5 Downscaling Scenarios. *American Journal of Climate Change* **03** (05): 512–527 DOI: 10.4236/ajcc.2014.35043

Dooge JCI. 1986. Looking for hydrologic laws. *Water Resources Research* **22** (9S): 46S-58S DOI: 10.1029/WR022i09Sp0046S

Giorgi F. 1990. Simulation of Regional Climate Using a Limited Area Model Nested in a General Circulation Model. *Journal of Climate* **3** (9): 941–963 DOI: 10.1175/1520-0442(1990)003<0941:SORCUA>2.0.CO;2

Gottschalk L. 1985. Hydrological regionalization of Sweden. *Hydrological Sciences Journal* **30** (1): 65–83 DOI: 10.1080/02626668509490972

Hoomehr S, Schwartz JS, Yoder DC. 2016. Potential changes in rainfall erosivity under GCM climate change scenarios for the southern Appalachian region, USA. *Catena* **136**: 141–151 DOI: 10.1016/j.catena.2015.01.012

Lyon SW, Troch PA. 2010. Development and application of a catchment similarity index for subsurface flow. *Water Resources Research* **46** (3): 1–13 DOI: 10.1029/2009WR008500

McDonnell JJ, Woods R. 2004. On the need for catchment classification. *Journal of*

Hydrology **299** (1–2): 2–3 DOI: 10.1016/j.jhydrol.2004.09.003

Paredes-Trejo F, Barbosa H, dos Santos CAC. 2019. Evaluation of the performance of SM2RAIN-derived rainfall products over Brazil. *Remote Sensing* **11** (9): 1–28 DOI: 10.3390/rs11091113

Paredes-Trejo F, Barbosa HA, Spatafora LR. 2018. Assessment of SM2RAIN-derived and state-of-the-art satellite rainfall products over Northeastern Brazil. *Remote Sensing* **10** (7) DOI: 10.3390/rs10071093

Rodrigues DBB, Gupta H V, Mediondo EM, Oliveira PTS. 2015. Assessing uncertainties in surface water security: An empirical multimodel approach. *Water Resources Research* **51** (11): 9013–9028 DOI: 10.1002/2014WR016691.Received

Sawicz K, Wagener T, Sivapalan M, Troch PA, Carrillo G. 2011. Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA. *Hydrology and Earth System Sciences* **15**: 2895–2911 DOI: 10.5194/hess-15-2895-2011

Sivapalan M. 2005. Pattern, Process and Function: Elements of a Unified Theory of Hydrology at the Catchment Scale. In *Encyclopedia of Hydrological Sciences*, Anderson M (ed.).John Wiley: London; 193–219. DOI: 10.1002/0470848944.hsa012

Troch PA, Carrillo G, Sivapalan M, Wagener T, Sawicz K. 2013. Climate-vegetation-soil interactions and long-term hydrologic partitioning: signatures of catchment co-evolution. *Hydrology and Earth System Sciences* **17** (6): 2209–2217 DOI: 10.5194/hess-17-2209-2013

Wagener T, Sivapalan M, McGlynn B. 2008. Catchment Classification and Services-Toward a New Paradigm for Catchment Hydrology Driven by Societal Needs. In *Encyclopedia of Hydrological Sciences*1–12. DOI: 10.1002/0470848944.hsa320

Wagener T, Sivapalan M, Troch P, Woods R. 2007. Catchment Classification and Hydrologic Similarity. *Geography Compass* **1** (4): 901–931

Xavier AC, King CW, Scanlon BR. 2016. Daily gridded meteorological variables in Brazil (1980-2013). *International Journal of Climatology* **2659** (October 2015): 2644–2659 DOI: 10.1002/joc.4518

CHAPTER 1

PERFORMANCE EVALUATION OF ETA/HADGEM2-ES AND ETA/MIROC5 PRECIPITATION SIMULATIONS OVER BRAZIL

Almagro, A., Oliveira, P.T.S., Rosolem, R., Hagemann, S., Nobre, C.A. Performance evaluation of Eta/HadGEM2-ES and Eta/MIROC5 precipitation simulations over Brazil, Atmospheric Research, 244, 105053, https://doi.org/10.1016/j.atmosres.2020.105053. (Impact factor 2021: 5.369)

Abstract

Climate change effects can have significant impacts worldwide. Extreme events can modify water availability and agricultural production, making climate change planning an essential task. The National Institute for Space Research (INPE in Portuguese) in Brazil has made a large dataset of regional climate model outputs (simulations and projections) available, which opens up many possibilities of carrying out high-resolution climate change studies. However, there is still no performance evaluation of the model-derived rainfall output against high-resolution ground-based observation data considering the Brazilian biomes. This paper attempts to fill this gap and evaluates the simulated precipitation throughout Brazil. We used gridded observed precipitation data and historical climate simulations from the Model for Interdisciplinary Research on Climate, version 5 (MIROC5) and from the Hadley Center Global Environment Model, version 2 (HadGEM2-ES), which were downscaled by the Eta RCM (Regional Climate Model). For the overlapping period (1980-2005), there is a good agreement (PBIAS up to 10%) of downscaled annual simulations for the Amazon and Cerrado biomes and large biases (reaching 40%) in the Pampa biome, compared to the observations. Our results showed that HadGEM2-ES is capable of representing long-term mean monthly precipitation for large areas well, such as the Amazon and Cerrado. Furthermore, the Eta RCM has considerably improved the driving GCM MIROC5 simulations. In conclusion, we recommend using the HadGEM2-ES simulations for the Amazon, Eta/HadGEM2-ES for the Atlantic Forest, Cerrado, and Pampa, and Eta/MIROC5 for the Caatinga and Pantanal. Our study provides an overview of two downscaled simulation datasets in Brazil that may help verify the models' suitability for further climate change assessments.

Keywords: Climate change, general circulation model, rainfall, regional climate model.

1. Introduction

Effects of climate change (e.g., warming of the atmosphere, extreme weather events contributing to a lack or excess of water) have significant socio-economic and environmental impacts worldwide (Hansen and Cramer, 2015). In Brazil, more severe and frequent hydrometeorological events (e.g., droughts, floods, landslides) are expected. Such extreme events can significantly alter water availability and agricultural production (PBMC, 2013). To ensure appropriate climate change planning and mitigation policies, accurate projections of climate changes are an essential but yet challenging task for the scientific research community due to the large number of climate-sensitivity factors that need to be considered (McNutt, 2013).

Future climate projections are usually produced by Global Climate Models (GCMs) or Earth System Models (ESM), which resolve the physics and dynamics of the Earth System as a whole (IPCC, 2013). GCMs/ESMs are the most advanced scientific tools for simulating responses of global climate regarding to variations of greenhouse gas concentration and to support climate change studies (Mello et al., 2015). The community-wide use of GCMs/ESMs is widely recognized in the fifth phase of the Coupled Model Intercomparison Project (CMIP5). The CMIP5 initiative comprises new sets of climate model experiments, which are coordinated by the World Climate Research Programme and include more than 50 complex GCMs/ESMs (Knutti and Sedláček, 2012; Taylor et al., 2012). Climate data are produced under different Representative Concentration Pathway scenarios (RCP 8.5, 6, 4.5, 2.6 Wm^{-2}) of the 5th Assessment Report (AR5) from the IPCC (IPCC, 2014). These pre-determined climatic conditions have been widely used to understand the climate and project climate changes on Earth (Zhao et al., 2013).

To provide a large volume of climatic data at global coverage, GCM historical simulations and future projections are usually produced at relatively coarser resolution (grid sizes in the order of 100 km-200 km) to assess climate change impacts. To evaluate the potential impacts of climate change on regional scales at finer spatial resolution (grid sizes in the order of 20 km), simulations and projections from GCMs are usually downscaled by employing Regional Climate Models (RCMs) (Giorgi, 1990; Maraun et al., 2017). The National Institute for Space Research (INPE) developed four sets of downscaled products based on the Eta RCM for Brazil, parts of South America and adjacent oceans, forced with both RCP 8.5 and RCP 4.5

scenarios obtained from the AR5 taken from global simulations and projections from two GCMs/ESMs, namely HadGEM2-ES and MIROC5 GCMs, respectively (Chou et al., 2014a). Choosing GCMs/ESMs was based on satisfactory performance in resolving precipitation and atmospheric circulation over South America, and also importantly due to ease accessibility of the data on public domain (Brazil, 2016; Flato et al., 2013). These downscaling simulations were performed in support of strategic climate change studies and the Brazilian Third National Communication to the United Nations Framework Convention on Climate Change (Brasil, 2016). Before the CMIP6 new datasets become fully available, those data based on AR5 scenarios are the latest and most advanced products in terms of spatial resolution available for climate change studies in South America. As a result, these regional-scale scenarios have been adopted for the Brazilian National Adaptation Plan for Climate Change.

For over 30 years, the RCMs have satisfied the need for high-spatial resolution climatologies for climate change impacts assessments, overcoming the inability of the GCMs to deal with it. During this time, we had phases of development, maturing, and exploration of contradictions and limitations (Tapiador et al., 2019). Due to the large public availability, the comparison between simulations of the RCMs against their driving-GCM turned into a common and necessary procedure to assess and evaluate the added values of the dynamical downscaling employed. For South America, some of the state-of-the-art of added value from RCMs simulations were performed using CORDEX experiments from the CMIP5 datasets. Llopart et al. (2019) assessed the added value of a pair of multi-model ensembles of RCMs and GCMs. In terms of precipitation, they found an added value of the RCMs in the coastal portion of the South Atlantic Convergence Zone (SACZ) and the Intertropical Convergence Zone (ITCZ), showing the clear improvement of the finer resolution in resolving convective schemes, such as the convergence zones. In their study, Falco et al. (2018) found that all the multi-model ensembles analyzed could reproduce the main features of seasonal climatology, while the individual analyses presented more biases, showing the need for individual assessment of the models. Moreover, they concluded that added value of RCMs simulations over historical periods were not consistent, being found only in certain combinations of model-region. The most noticeable degrading of simulations was found on winter climatology. Solman & Blázquez (2019) evaluated the ability of RCMs and their corresponding driving-GCM in reproducing the precipitation spatial distribution and behavior in a multi-temporal scale over South America.

According to their results, spatial patterns and seasonal means are close related with large-scale atmospheric circulation, such as SACZ and ITCZ, and for some regions large biases (underestimating or overestimating) in rainfall amount were found for a group of RCMs. On common point was found for all above mentioned studies: this kind of analysis is the basis for further work using climatic projections. The ability of GCM/RCM in simulate historical spatial distribution and patterns of climatologies tell us a lot about their projections for the future.

As mentioned above, to increase the degree of confidence in these model projections, their simulations need to be evaluated compared to historical observations. However, a proper investigation of Eta/HadGEM-ES and MIROC5 has not yet been fully addressed. A study conducted by Chou et al. (2014b) is the only one that has previously evaluated these RCM products (Eta/HadGEM2-ES and Eta/MIROC5). Their analyses use long-term monthly and seasonal mean fields of temperature and precipitation from 1961 to 1990 against a relatively coarse resolution CRU TS 3.1 global gridded dataset (Mitchell and Jones, 2005), focusing mainly on summer and winter seasons. However, the results were obtained by averaging the simulations over oversized areas with heterogeneous hydroclimatic characteristics and multiple precipitation regimes, limiting the representation of specific local climatic characteristics. When averaging precipitation over oversized areas or regions, we can mask some regional characteristics and local features, and the use of large areas to analyze precipitation seems to be viable only when using spatial distribution. As we can see in supplementary Figure S1, the averaged observed precipitation (1980-2005) over North and Central-West of Brazil (administrative regions) does not reproduce the precipitation regime of any biome within this area (Amazon, Cerrado, and Pantanal). The use of the smaller and more coherent areas for averaging precipitation was done before. For the European territory, Christensen & Christensen (2007) and Dosio & Paruolo (2011) used eight sub-areas taking into account topography and climate features. Gregory et al. (1991) adopted nine spatially coherent precipitation regions for analyzing area-averaged statistics of precipitation in Great Britain. On a global scale, some of the world biomes were adopted by Huxman et al. (2004) for an average rain-use efficiency in aboveground net primary production using global precipitation data. Thus, we claim that hydrometeorological divisions are the most appropriate, such as biomes or hydrographic regions. Therefore, to improve the RCM evaluation across Brazil, a more suitable division should consider areas with similar ecoclimatic dynamics and characteristics. In this context, the

Brazilian biomes appear as a viable alternative. These biomes were defined by the Ministry of Environment in Brazil and are large ecosystems with relatively similar and uniform climate, vegetation and biodiversity, relative to the overall extent of the country (Brown and Maurer, 1989; Coutinho, 2016).

The objective of this study is to evaluate the Eta/HadGEM2-ES and Eta/MIROC5 performance to represent long-term monthly and seasonal mean precipitation over key Brazilian biomes. We compared the simulated precipitation of the driving GCMs (HadGEM2-ES and MIROC5) and their respective downscaled datasets (Eta/HadGEM2-ES and Eta/MIROC5) against a high-resolution gridded dataset derived from observational networks in Brazil. In addition, we investigated the possible origin of biases in the simulated precipitation.

2. Material and methods

2.1. Brazil: Study area and biomes

Given its large spatial extent (8,511,000 km^2) including a large range of elevation (sea level to 2900 m altitude) and heterogeneous patterns of precipitation seasonality, Brazil's diverse vegetation types are classified into six main biomes (see Figure 1). We followed previous definitions (Li et al., 2006; Murray-Tortarolo et al., 2017; Myneni et al., 2007) that consider 100 mm per month as a threshold for defining a month within the dry season to verify if the simulations are capable of estimating the onset, duration, and termination of both rainy and dry seasons. This value is a global precipitation threshold that considers the amount to begin runoff (Zhang et al., 2004), to maintain the vegetation growth (Li et al., 2006; Murray-Tortarolo et al., 2017), and sensitivity analysis (Murray-Tortarolo et al., 2017).

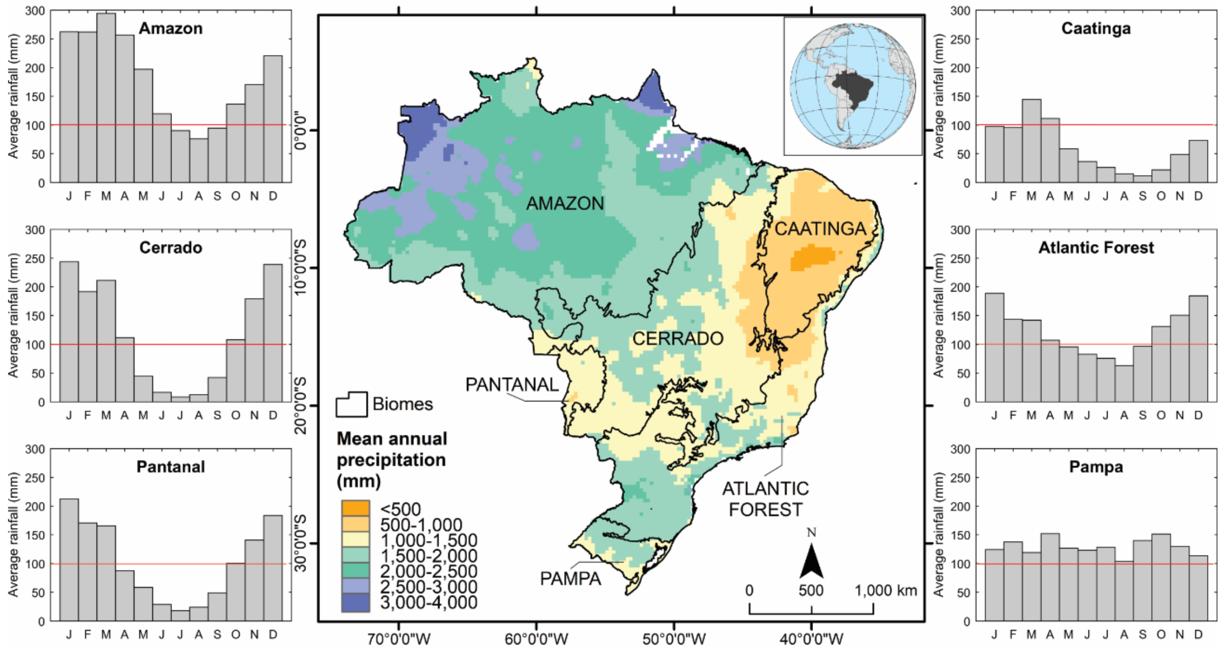

Figure 1: Brazilian biomes map and their respective mean monthly precipitation shown as bar plots. Monthly means were calculated for the 1980-2013 period from daily dataset originally developed by Xavier et al. (2016). The red lines represent a threshold used as a criterion to identify the dry months (< 100 mm).

The Amazon is the largest tropical biome in the world, consisting of a densely vegetated rainforest with the highest annual mean precipitation (average annual precipitation of 2.3 m and greater than 4 m/year in some portions of western Amazon) and a short dry season (Sombroek, 2001). The Cerrado biome mainly consists of woodlands and savanna and is crucial for water and food supplies, for maintaining ecological services and for economic activities in Brazil. Moreover, it is considered one of the most important biomes in Brazil related to food-energy-water security (Oliveira et al., 2014). The Pantanal is one of the largest flooded areas in the world and it is the most intact biome in Brazil (Junk et al., 2006). It has very defined dry and rainy seasons, and the flood cycles – caused not by an excess of precipitation but due to drainage deficiency – are the most important ecological phenomenon (Ribas and Schoereder, 2007). The Caatinga biome is characterized by a semi-arid region in the Northeast of Brazil. It comprises mostly secondary vegetation (herbaceous and arboreous) and presents the lowest values of annual precipitation with a severe dry season – about 70% of the annual precipitation occurs in February-April period (Menezes et al., 2012; Pinheiro et al., 2013). The Atlantic Forest is characterized by rainforest cover in the coastal area and the semi-deciduous forest in the continental area with very defined wet and dry seasons (Morellato and Haddad, 2000). Finally, the Pampa biome is located in the South of Brazil and has no defined dry season. Natural grasslands are predominant, with tree formations and sparse shrub, and it is referred to as "Campos" (Lupatini et al., 2013; Roesch et al., 2009).

2.2. Data acquisition and processing

We used rainfall data from the Eta Regional Climate Model, available from INPE (http://projeta.cptec.inpe.br/). These products were generated using climate forcing data derived from the r1i1p1 ensemble member of two GCMs/ESMs: the British HadGEM2-ES and the Japanese MIROC5 (both original GCM/ESM products available at https://esgf-data.dkrz.de/search/esgf-dkrz/). The MIROC5 is an atmosphere-ocean general circulation model (AOGCM), with a 1.4ºx1.4º spatial resolution for the atmospheric parcel, that brought several improvements on Intertropical Convergence Zone (ITCZ) and El Niño Southern Oscillation (ENSO) simulations (Watanabe et al., 2010). The HadGEM2-ES is also a coupled AOGCM, with an atmospheric resolution of 1.875ºx1.25º, that brings some components generated interactively by the model, instead of being assigned as boundary conditions (Jones et al., 2011).

The baseline period is defined from 1961 through 2005. The INPE dataset was produced at approximately 20 km spatial grid resolution for South America and with two different Representative Concentration Pathways (RCP4.5 and RCP8.5), respectively. The Eta RCM downscaling procedure is described in more detail in Chou et al. (2014a, 2014b).

To evaluate the simulated monthly precipitation from both datasets, we compared model-derived precipitation against a gridded-interpolated product derived from observations (0.25º by 0.25º spatial resolution), developed by Xavier et al. (2016), and available at http://careyking.com/data-downloads/). This product used observed precipitation data derived from approximately 4,000 rain gauges from the Brazilian Water Agency (ANA), the National Institute of Meteorology (INMET), and the Water and Electric Energy Department of São Paulo state (DAEE/SP) from 1980-2013. Furthermore, this reference dataset has been extensively applied in many fields of study, such as evaluation of remote sensing products (Melo et al., 2015; Paredes-trejo et al., 2018, 2017; Paredes-trejo and Barbosa, 2017), vegetation response to rainfall variability (Souza et al., 2016), impacts of climatic extremes (Melo et al., 2016), and climate change assessments (Almagro et al., 2017).

We applied an interpolation method on the global climate model outputs and on the reference gridded observations to a common spatial resolution of the Eta RCM spatial resolution, which is the focus of this study. We applied a first-order conservative remapping method (Jones, 1999) using the Climate Data Operators (CDO) to preserve the main characteristics of each dataset and to ensure that any area average – area of a pixel or an area of a biome – would be similar to the original dataset , allowing more detailed comparisons when considering the different resolutions and introducing the less possible kind of error associated to the remapping. As we show in the supplementary material (Figure S2), the highest mean error added to the dataset due the first-order conservative remap is about 0.01 mm, a value lower than the uncertainty associated to the most of common ground rain gauges observations and to our reference dataset (Villarini et al., 2008; Xavier et al., 2016). The same procedure was widely done in previous studies (Diaconescu et al., 2015; Nadeem et al., 2019; Wu et al., 2020).

2.3. Metrics to evaluate the simulated precipitation

To evaluate the performance and quality of simulated precipitation against the observed data product, we used the following statistical metrics: Percentage Bias (PBIAS), Root Mean Squared Error (RMSE), Correlation Coefficient (CC), and Coefficient of Variation (CV) (Equations 1, 2, 3 and 4).

$$BIAS = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{P_{sim\,i} - P_{obs\,i}}{P_{obs\,i}}\right) \qquad 1$$

$$CC = \frac{n(\sum_{i=1}^{n} P_{sim\,i}\,P_{obs\,i}) - (\sum_{i=1}^{n} P_{sim\,i})(\sum_{i=1}^{n} P_{obs\,i})}{\sqrt{\left[n\sum_{i=1}^{n} {P_{sim\,i}}^{2} - (\sum_{i=1}^{n} P_{sim\,i})^{2}\right]\left[n\sum_{i=1}^{n} {P_{obs\,i}}^{2} - (\sum_{i=1}^{n} P_{obs\,i})^{2}\right]}} \qquad 2$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_{obs\,i} - P_{sim\,i})^{2}}{n}} \qquad 3$$

$$CV = \frac{\sigma}{P} \qquad 4$$

where *P* is the long-term mean precipitation from observations "*obs*" and simulations "*sim*"; σ is the standard deviation of the annual precipitation; and *n* is the number of points in each biome.

We also ranked the performance of GCMs and RCMs to simulate some proprieties of precipitation, such as the rainy and dry periods, seasonal and annual precipitation for each biome. To do so, we followed the methodology proposed by Deidda et al. (2013), which is commonly applied on GCM-RCM comparison (see Mascaro et al., 2018). This methodology calculates a single dimensionless error metric, the $\in_j$, for each precipitation property combining multiple variables that characterizes that property (Table 1). Then, models are ranked by the value of $\in_j$ (lower rank means better performance), calculated by Equation 5.

Table 1: Properties and the variables (k) considered to calculate the errors ($\boldsymbol{E_{k,j}}$) for each property

Property	Variable, *k*	Error, $\boldsymbol{E_{k,j}}$
Dry-season	Number of dry months, DM	$\lvert DM_{obs} - DM_j \rvert$
Seasonal cycle	Seasonal root mean square error, RMSE	$\lvert RMSE_{obs} - RMSE_j \rvert$
	Seasonal correlation coefficient, CC	$1 - CC_j$
Annual cycle	Mean annual P, P	$\lvert \bar{P}_{obs} - \bar{P}_j \rvert$
	Coefficient of variation of annual P, CV	$\lvert CV_{obs} - CV_j \rvert$

$$\in_j = \sqrt{\sum_{k=1}^{S} \left(\frac{E_{k,j}}{\sum_{i=1}^{N} E_{k,j}} \right)^2} \qquad 5$$

where $\in_j$ is the dimensionless error for the *j* (*j*=1,…,*N*) simulation; $E_{k,j}$ is the error between observed and simulated values of the variable *k* (*k*=1,…,*S*), and dividing it by the sum of the errors of all models, we obtain a dimensionless contribution for the error of variable *k*. Then, summing and taking the square root of the error parcel for all variables *S*, we reach the $\in_j$ for the rank.

2.4. Regional and spatial analysis

Using the baseline period (1980-2005) from simulations and observations, we calculated the long-term precipitation averages at monthly, seasonal (December, January and February – DJF; March, April and May – MAM; June, July and August – JJA; and September, October and

November – SON), and annual scales for each grid point for all datasets across Brazilian biomes (Figure 1). Then, we performed two distinct analyses with regards to regional averages and spatial distributions and patterns using the metrics described in Subsection 2.3.

We computed the PBIAS, CC and CV separately for each biome, considering grid points within the biome to represent biome-average quantity (regional analysis) and the PBIAS and CV on a grid point scale for the whole of Brazil (spatial analysis). The first analysis identifies a general behavior for each biome while the second one enables an investigation of the biases at the same time, as well as spatial patterns and their possible causes. To assess the reliability of the models to represent the wet and dry months, we computed the biome-specific long-term mean for each month of the year.

3. Results and discussion

3.1. Spatial patterns on annual and seasonal precipitation

In this section, we present the results of the spatial distribution of the biases among observations (Figure 2a) and models' simulations/projections (Figure 2b to 2e) and observations for the entire Brazilian territory. The analysis identifies the spatial patterns of precipitation, and consequently the biases (Figure 3) in relation to the observed annual means.

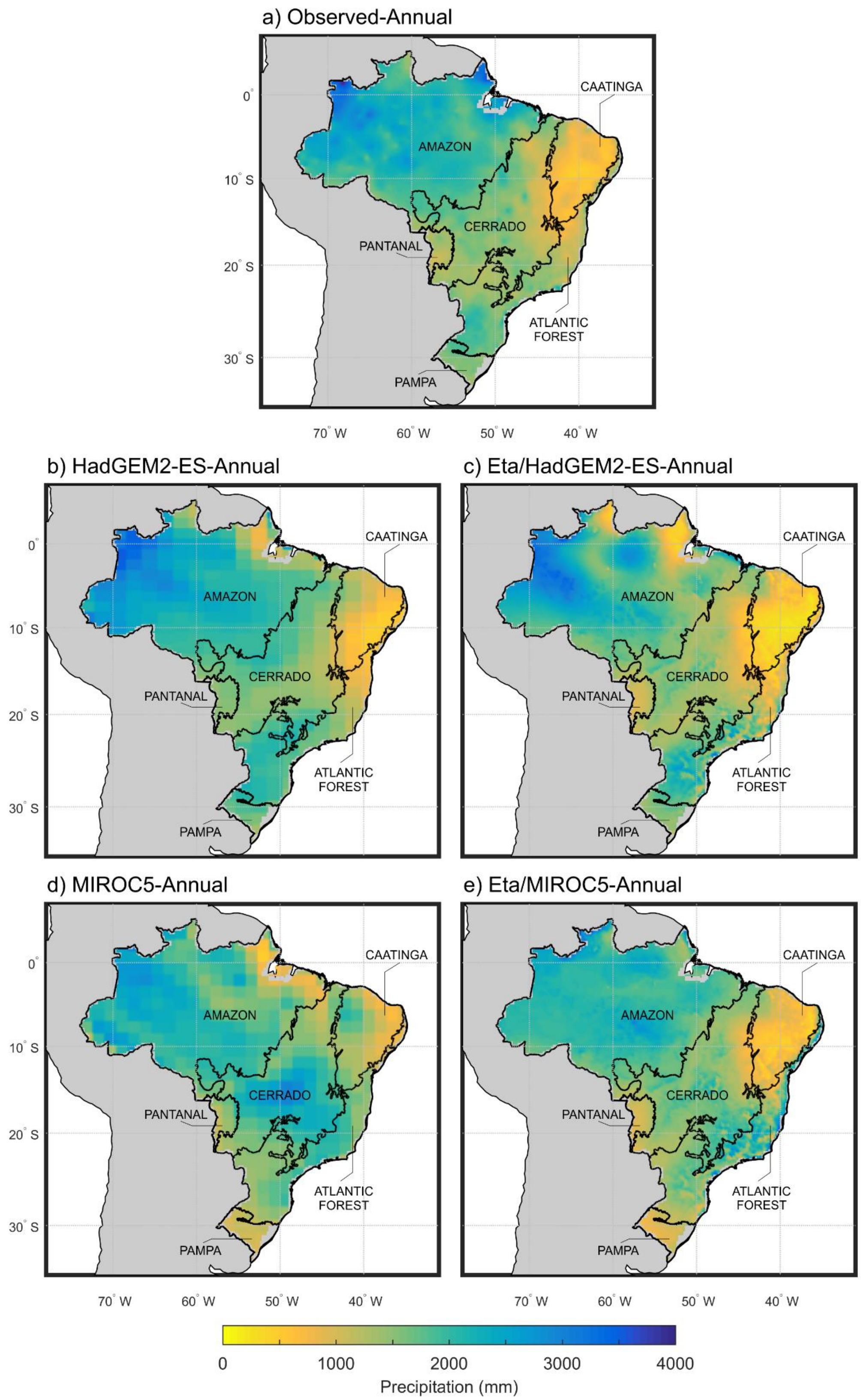

Figure 2: Spatial distribution of the annual precipitation (P) over the Brazilian biomes for a) Observed dataset; b) HadGEM2-ES dataset; c) Eta/HadGEM2-ES dataset; d) MIROC5 dataset; and e) Eta/MIROC5 dataset.

The HadGEM2-ES presented an overall positive bias throughout Brazil with negative biases observed in some areas in the northern region. The downscaling process barely corrected the biases on annual precipitation simulations, but a negative bias was generated in the Caatinga biome portion, where the lowest values of annual precipitation occur and a small change in rainfall corresponds to a relatively large percentual change. At the same time, in the coastal area of the Northeast region, in the Atlantic Forest biome, a negative bias was spread to the Eta simulations. Even with the 20 km resolution, the Eta RCM is not capable of capturing the rainfall system neither improve the representation of the phenomenon. This is also true with respect to sea-breeze induced rainfall along the Amazonian coastal zone. Due to the coarse resolution, we cannot expect that the GCMs/ESMs capture sea-breeze induced rainfall. However, we expect that Eta RCM would at least improve the simulations, but it did not.

For the MIROC5 simulations, extremely high values of PBIAS (which reached up to 200%) were found in the midwestern and northeastern areas of Brazil. However, the Eta/MIROC5 downscaled simulations reduced the biases while remaining just a few grid points with positive values in the coast of the southeastern region. These results show the inability of MIROC5 to simulate mean annual precipitation over a large area of Brazil. At the same time, the results demonstrate the great improvement of Eta/MIROC5 in relation to its original coarse-scale GCM product.

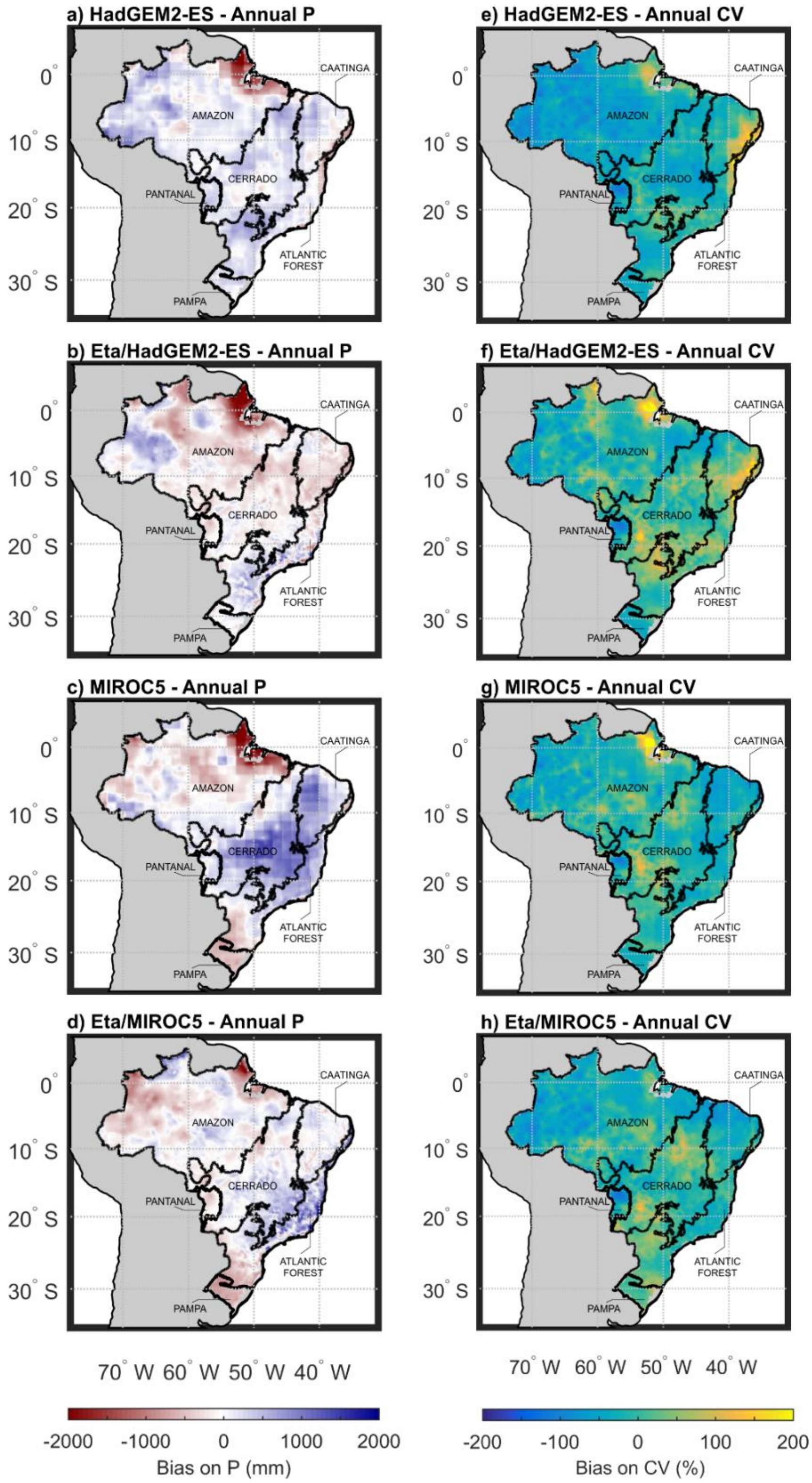

Figure 3: Spatial distribution of the PBIAS on annual precipitation (P) and on the annual coefficient of variation (CV) over the Brazilian biomes. a) and e) represents the PBIAS and CV on HadGEM2-ES simulations; b) and f) represents the PBIAS and CV on Eta/HadGEM2-ES simulations; c) and g) represents the PBIAS and CV on MIROC5 simulations; and d) and h) represents the PBIAS and CV on Eta/MIROC5 simulations.

Figure 4 shows the bias on the long-term mean precipitation for each season (DJF, MAM, JJA, and SON) calculated for all evaluated model products against observations (see also Supplementary Figure S2). As we can note in the DJF and MAM seasons, there is an overestimation of all simulations for the extreme north of the Amazon, which can be related to the inability of the models to capture the sea-breeze influence on the rainfall systems in this part of Brazil. The Amazonian coastal regions of Amapá and Pará states are strongly influenced by sea-breeze and are affected by it up to 300 days a year. The negative bias observed in this region could be related to the wrong simulation of this phenomenon, which can be expected by a GCM (due to its coarse resolution). However, it was expected that this local feature would be captured and reproduced by the Eta model. For other locations and biomes, the simulations significantly improved the use of Eta RCM, especially for the MIROC5, showing the good suitability of Eta RCM downscaling process on these seasons over Brazil. During the JJA season, high-pressure systems (< 1,013 hPa), or anticyclonic, dominate the subtropical region (see Figure S5), making the formation of clouds more difficult and blocking the occurrence of rainfall. The amount of precipitation in much of Brazil is very low, except for the northern part of the Amazon and southern part of the Pampa. In general, both GCMs represented these features of the dry season well and we noted more improvement of the Eta RCM in the MIROC5 data. Despite this, due to the very low rainfall amounts observed in the JJA months, any minimal over/underestimation generates an expressive relative bias, as we can see in Figure S6. For the SON season, the simulations presented low biases in absolute terms and higher biases in relative terms, in the same way as JJA. The downscaling process did not improve the simulations of HadGEM2-ES, maintaining the spatial behavior of the biases. For the MIROC5 simulations, there was a great improvement in the Amazon biome and a change in the signal of the biases in the Caatinga biome. Once again, this is due to low precipitation totals in this biome.

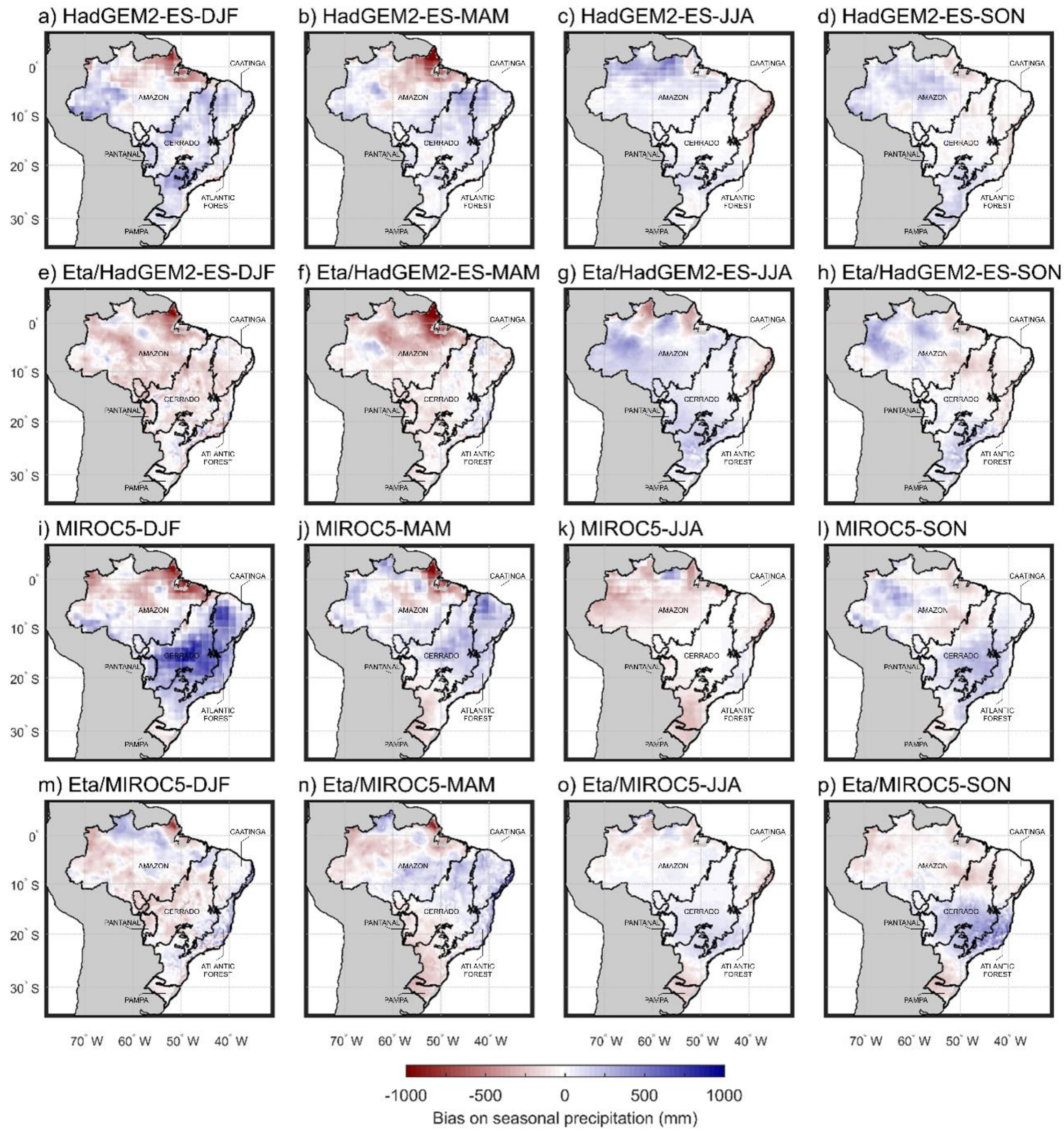

Figure 4: Absolute biases (BIAS) in each season (DJF, MAM, JJA and SON) simulated precipitation in Brazilian biomes. a) to d) represent the BIAS for the HadGEM2-ES for all seasons; e) to h) represent the BIAS for the Eta/HadGEM2-ES for all seasons; i) to l) represent the BIAS for the MIROC5 for all seasons; and m) to p) represent the BIAS for the Eta/MIROC5 for all seasons. Shades of blue indicate a positive BIAS while shades of red indicate a negative BIAS.

3.2. Long-term means and annual variability at the biomes

In this section, we present and discuss the results of mean monthly and annual precipitation, and the annual variability during the 1980-2005 period for each Brazilian biome. The long-term mean monthly precipitation analysis identifies how well the models can simulate precipitation patterns and the dry/rainy seasons. Figure 5 shows the comparison of the downscaled and driver-family models' simulations against the observations. We used an error bar (standard deviation) of observations to create a range of acceptable values for the simulations. Eta/HadGEM2-ES generally underestimates the mean monthly rainfall in the rainy season and represents it well in the dry season. On the other hand, HadGEM2-ES overestimates the means for rainy and underestimates means for dry seasons, but most of these under/overestimations are inside the range of acceptable values. For larger areas such as the Atlantic Forest, the Amazon and Cerrado, the GCM means are closer to observations than RCM simulations in the rainy season (DJF to MAM). In the Amazon, Cerrado and Pantanal biomes, Eta/HadGEM2-ES has more negative precipitation biases during the rainy season, which reach almost -50%. Considering the error bar of observations, the Eta RCM simulated the rainy and dry seasons over the biomes well, except for the Pantanal (one month longer) and Pampa (no defined dry season). Throughout the Amazon, we verified very large error bars due its large area and spatial variability of rainfall regimes. Taking this into account, downscaled models were capable of simulating the short dry season (three months), showing close values to the observations and a significant drop from the rainy season rain. In the Pampa biome, where there is no defined dry season, the HadGEM2-ES model was capable of capturing this characteristic but, at the same time, produced considerable errors in the DJF and SON months. These large errors found in the Pampa are related to the coarse resolution of the GCMs, which makes Pampa's area relatively small for 100-200 km simulations. In the Caatinga, most of the GCM and RCM simulations were considered acceptable, except for HadGEM2-ES in February and Eta/HadGEM2-ES in January. Moreover, the dry season and distribution of rainfall over the year were well simulated by all models.

The mean monthly rainfall simulated by Eta/MIROC5 is overestimated in the rainy season and underestimated in the dry season in the Atlantic Forest and Caatinga biomes, while in the Cerrado and Pantanal, the opposite pattern is observed. In relation to the original MIROC5 product, the downscaled Eta/MIROC5 version mostly improved the values, thereby showing a clear added value to the downscaling process. The Amazon means are underestimated in all months of the year, and there is a noticeable improvement of the values simulated by Eta/HadGEM2-ES in the DJF and JJA months. The poorest simulation was performed for the Pampa biome, where the model cannot represent the absence of a dry season. Moreover, simulations for this biome have the largest difference to the observations with just two months of simulations inside the error bar of the observations. In general, Eta/MIROC5 can capture the rainy and dry season except for the Pampa biome. Finally, the Eta/MIROC5 is generally drier than the driving GCM in the wet season and wetter in the dry season.

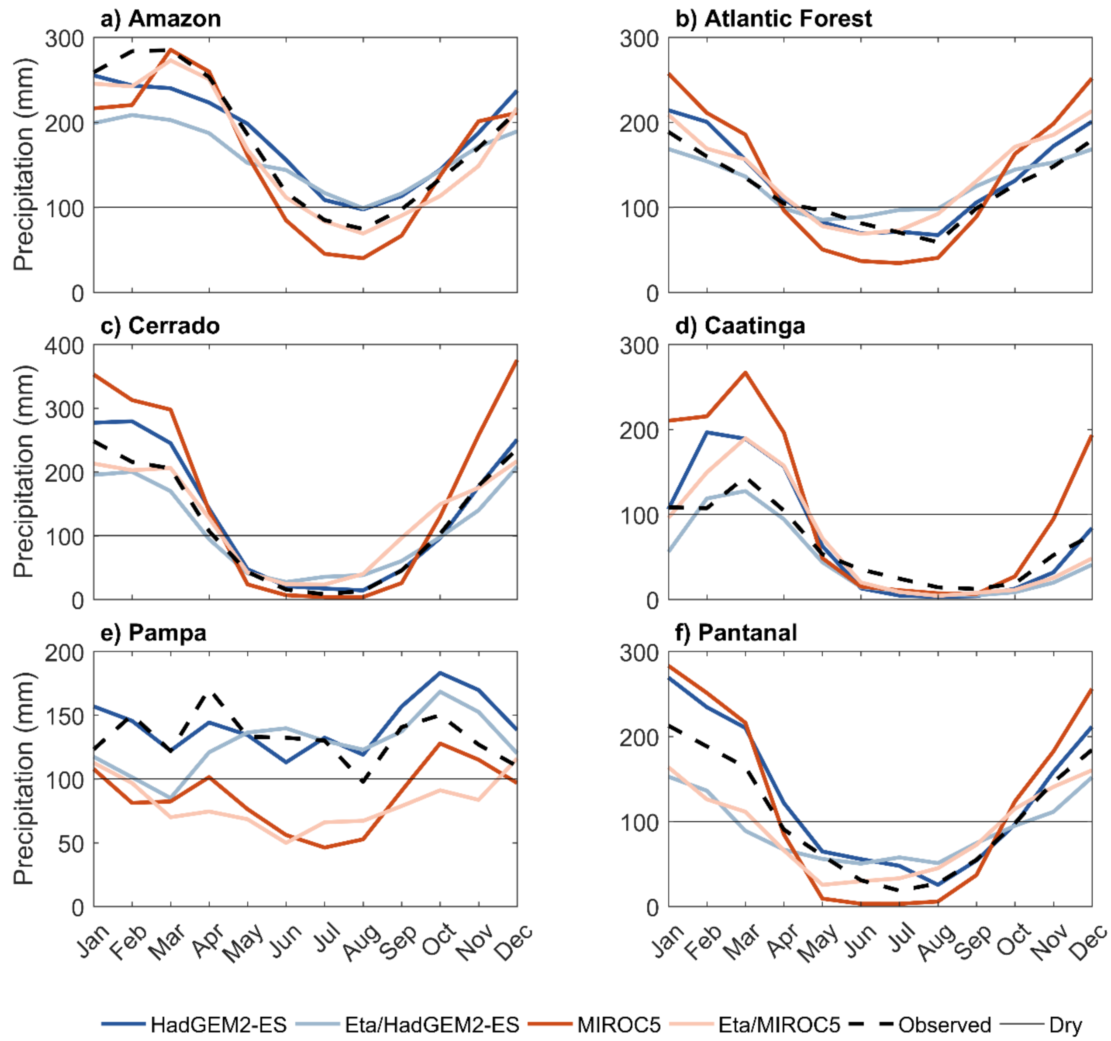

Figure 5: Long-term mean monthly rainfall for the 1980-2005 period for observations (black dashed line), HadGEM2-ES (dark blue line), Eta/HadGEM2-ES (light blue line), MIROC5 (dark red line), and Eta/MIROC5 (light red line) simulations in the a) Amazon, b) Atlantic Forest, c) Cerrado, d) Caatinga, e) Pampa, and f) Pantanal. The red reference line represents a threshold used as a criterion to identify the dry season (< 100 mm).

In terms of seasonality, for most biomes, the Eta/HadGEM2-ES is drier than HadGEM2-ES in the wet season, but the wetter behavior in the dry season is less obvious than the Eta/MIROC5 one. In general, the downscaling process applied by the Eta RCM improved the long-term mean monthly values, but those on the MIROC5 were more notable than for

HadGEM2-ES. The simulations of MIROC5 were originally not as good as the one of the simulations of HadGEM2-ES and this led to a more notable improvement of the downscaling process for the first model. At the same time, our analysis of the annual cycle clearly showed that the downscaled simulations are more suitable for the biomes where there are large amounts and well-defined rainy and dry seasons.

For the mean annual rainfall during the period 1980-2005, Eta/HadGEM2-ES performed well for the Atlantic Forest (+2%), Caatinga (-6%), and Pampa (-3%), while Eta/MIROC5 shows lower biases over the Amazon (-7%), Cerrado (+6%) and Pantanal (-14%) (Figure 6a). We found a poor performance of Eta/HadGEM2-ES for the mean annual rainfall in the Cerrado with negative biases up to -26%, and Eta/MIROC5 in the Pampa biome with underestimates up to -39%. No model was capable of capturing the annual variability (Figure 6b) for the Pantanal and just the HadGEM2-ES captured it for the Caatinga biome. These two biomes presented the highest values of observed CV once they presented the lowest observed amounts of annual precipitation.

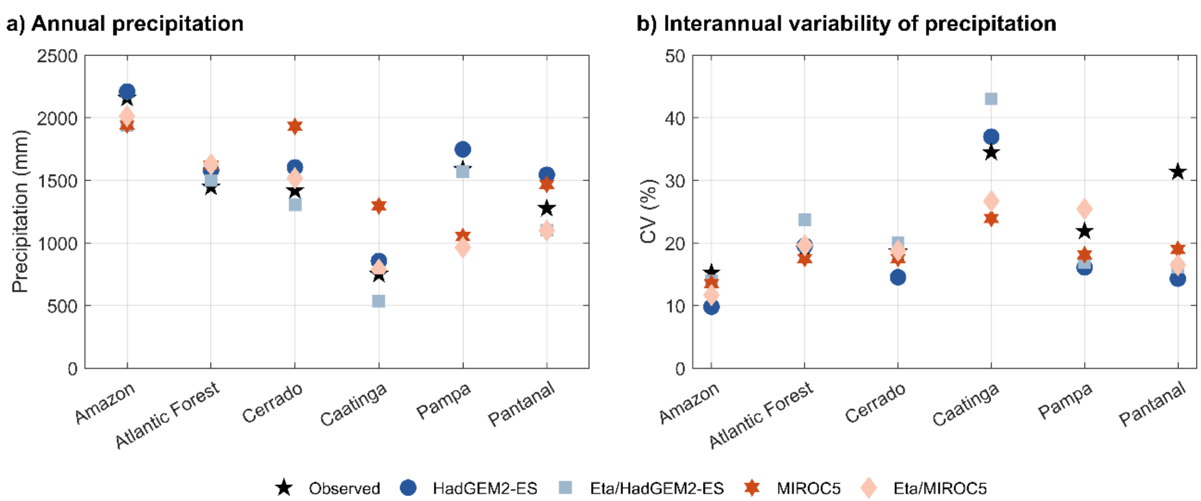

Figure 6: Annual characteristics of precipitation over the Brazilian biomes for the 1980-2005 period. The mean annual precipitation (a) is presented in absolute values to differ the magnitudes between the biomes. The annual variability (b) is calculated by the division of annual standard deviation by the annual mean precipitation.

3.3. Long-term mean seasonal precipitation

Figure 7 shows the percent bias (PBIAS) of HadGEM2-ES, Eta/HadGEM2-ES, MIROC5 and Eta/MIROC5 in terms of amount in seasonal precipitation simulations. In general, Eta/HadGEM2-ES underestimates the rainfall in DJF and MAM in the Brazilian biomes, while the JJA and SON rainfalls are overestimated in the Amazon, Atlantic Forest and Pampa. In the Cerrado and Pantanal, the largest biases occurred in the JJA season (dry period), which were increased by the downscaling process (Eta RCM) (Figure 7c and 7f). All seasons in the Caatinga were underestimated by the HadGEM2-ES, in terms of the amount in the season. On the other hand, Eta/MIROC5 underestimates rainfall throughout all the seasons in the Amazon and Pampa. Overestimates were simulated for MAM, JJA and SON in the Cerrado, DJF and MAM in the Caatinga, JJA and SON in the Pantanal and for all seasons in the Atlantic Forest, with lower overestimates in the dry season. The Eta/MIROC5 shows some improvements of the rainfall biases compared with its driving GCM, especially in the rainy seasons (DJF and MAM) for all biomes. These improvements were more distinct than for Eta/HadGEM2-ES, where improvements were restricted to some season/biome combinations.

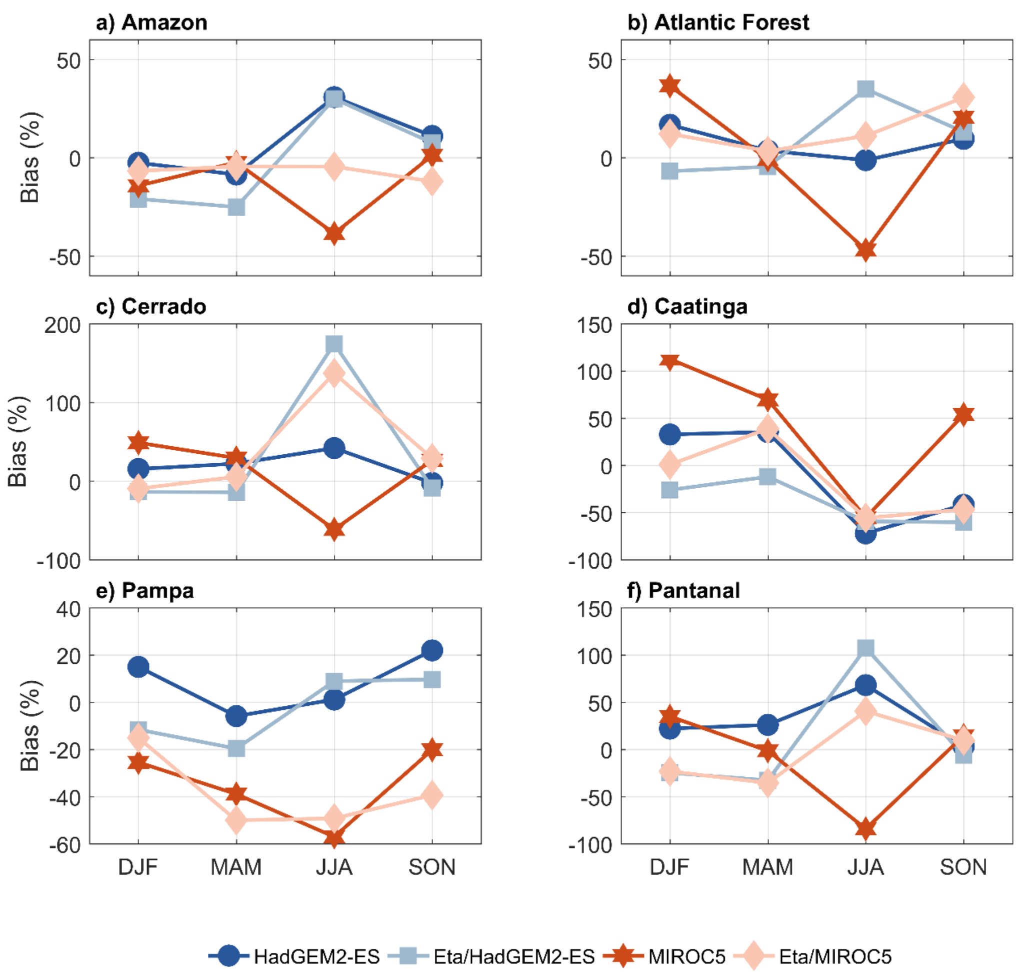

Figure 7: PBIAS on the long-term seasonal precipitation for the 1980-2005 period simulated by models against observations in the a) Amazon, b) Atlantic Forest, c) Cerrado, d) Caatinga, e) Pampa, and f) Pantanal.

Correlation coefficients (Equation 3) are presented in Figure 8 for the GCM and RCM simulations. The results for the Eta/GCM corroborate those presented by Chou et al. (2014b), who found spatial correlations above 0.50 for all simulations in their regional analysis for the whole of Brazil. These better values of seasonal means compared to mean monthly values are expected, once that spatial errors are being reduced when averaging (Pierce et al., 2009). In general, Eta/HadGEM2-ES simulates the mean seasonal precipitation better than the Eta/MIROC5 in the Atlantic Forest, Caatinga, Cerrado and Pampa. We highlight the smaller correlation of Eta/HadGEM2-ES simulations in the Pantanal during the JJA (dry season) and minimal correlation (up to 0.07) found between Eta/MIROC5 and observations in Pampa during the MAM and SON and Pantanal's JJA. On the other hand, both models show good results (up to 0.95) in simulating the seasonal cycle of precipitation for the Amazon, Caatinga, and Cerrado biomes. Once again, we noted better improvements of Eta RCM for MIROC5 than HadGEM2-ES and the best performance of GCM simulations for large areas such as the Amazon and Cerrado.

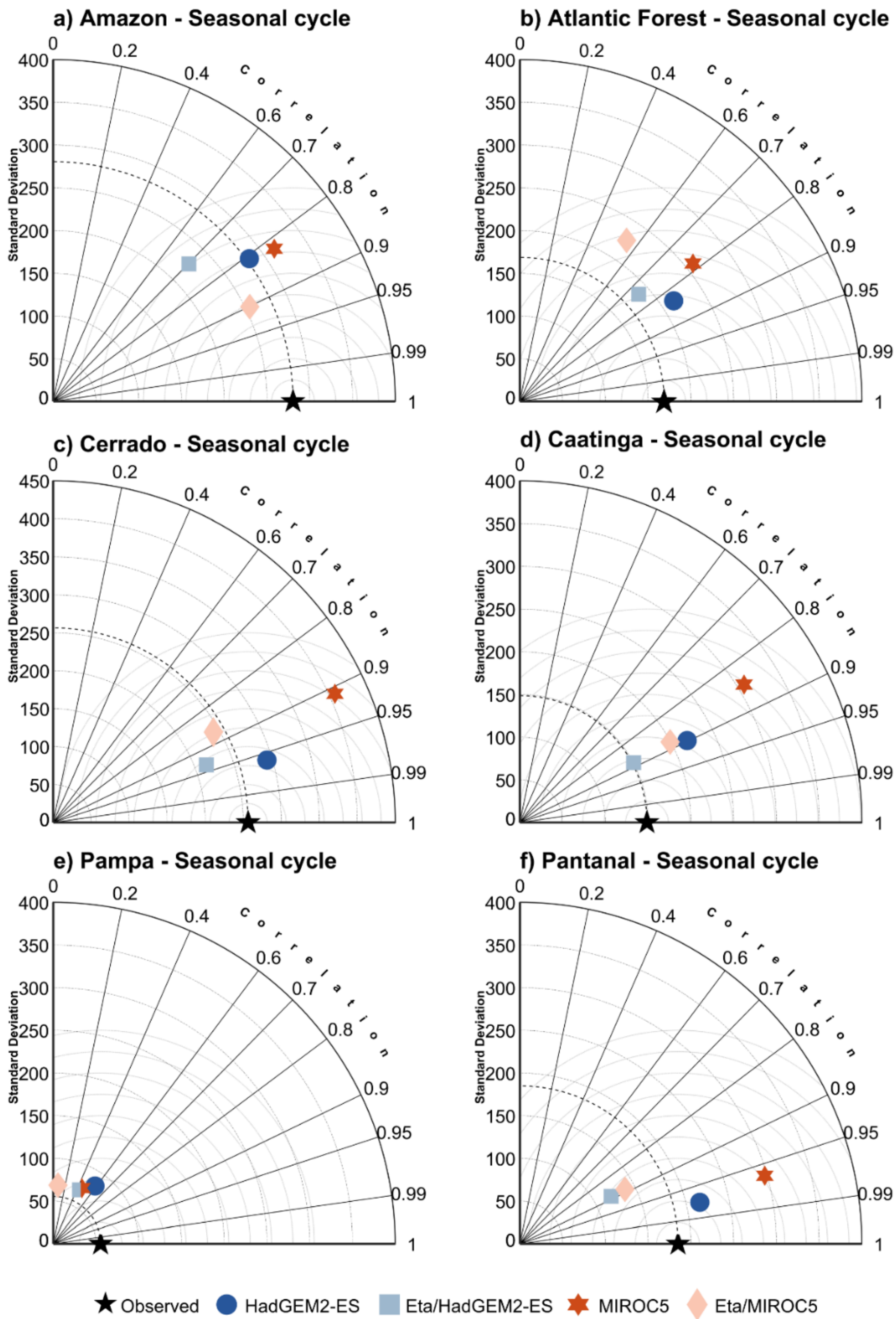

Figure 8: Taylor diagrams of seasonal mean precipitation over a) Amazon, b) Atlantic Forest, c) Cerrado, d) Caatinga, e) Pampa, and f) Pantanal for simulations and observations. Means of the observed seasonal precipitation are marked as a black star. The azimuth and the radial distance from the origin of the plot represents the correlation coefficient and the standard deviation (mm) of simulated data in relation to the observed value, respectively.

3.4. Investigating the origin of the biases

We consider that precipitation biases in the Eta/HadGEM2-ES and Eta/MIROC5 simulations for Brazilian biomes may have three possible reasons: a) they are produced by the GCM and not corrected by Eta RCM; b) they are produced by the Eta RCM and are absent in the CGM; or c) they are related to uncertainty in the observations. Figure 9 provides some insight concerning these reasons.

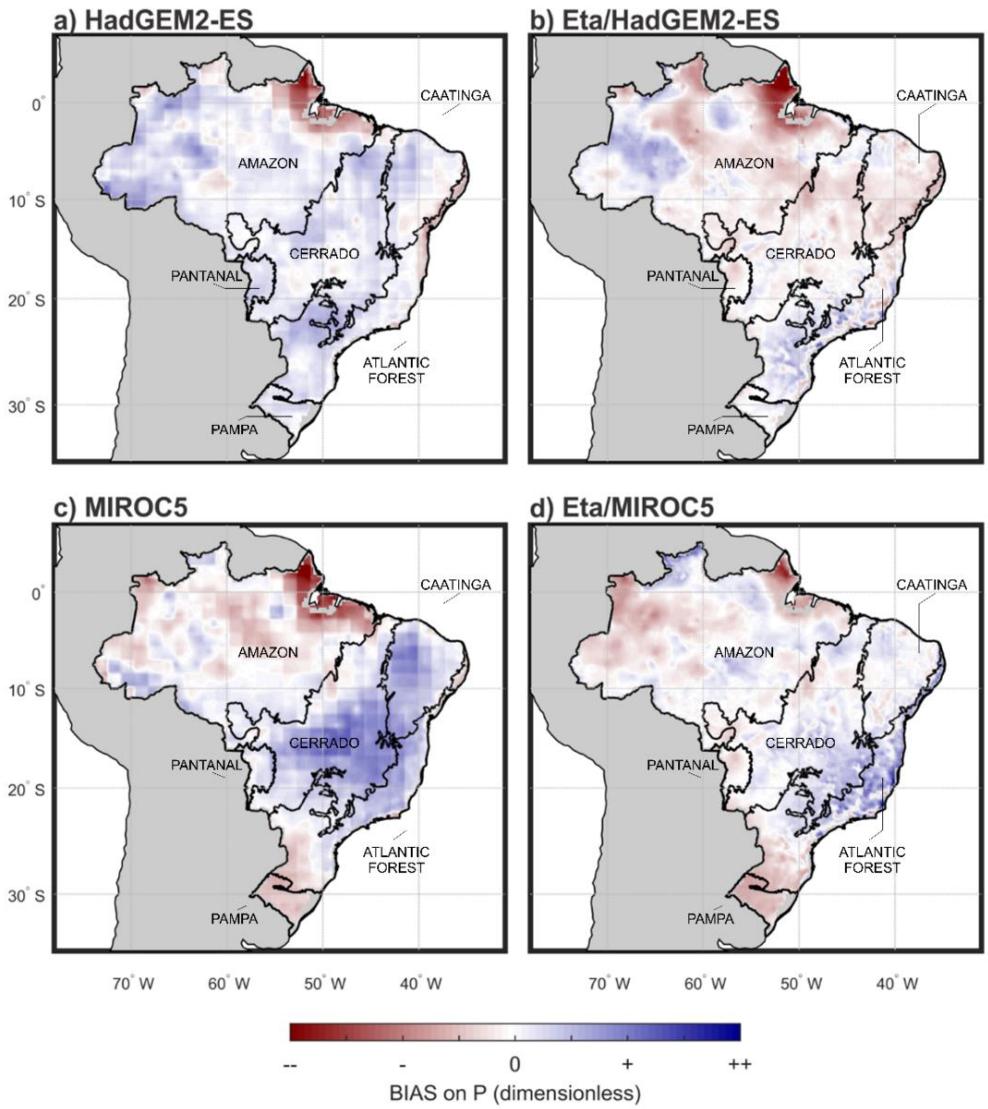

Figure 9: Mean bias error for precipitation simulations of a) HadGEM2-ES, b) Eta/HadGEM2-ES, c) MIROC5, and d) Eta/MIROC5 for the 1980-2005 period.

Figures 9a and 9b show that simulations of HadGEM2-ES and Eta/HadGEM2-ES have a positive bias in the western part of the Amazon and the southern part of the Atlantic Forest and negative biases in the northeastern area of the Atlantic Forest and Caatinga biomes. In turn, Fig. 9c and 9d indicate negative biases in the northern part of the Amazon, the southern region of the Atlantic Forest and all throughout the Pampa biome, and strong positive biases in the Cerrado and the central part of the Atlantic Forest. These biases are likely related to inherent driving GCM biases that can be carried out from GCM to RCM via the lateral boundary conditions (Ehret et al., 2012; Xu and Yang, 2015). The mean bias error estimated for the Pantanal is positive for the GCM simulations and negative for the RCM simulations. This kind of error, and the lower biases seen in Figure 3 of Section 3.1 for HadGEM2-ES instead of Eta/HadGEM2-ES, could be explained by the downscaling process, but more in-depth analysis in the data generated in the downscaling process and also in the physical processes involved leading to deforestation need to be made. According to Chou et al. (2014b), the Eta RCM is especially suited for regions with steep topography (particularly because of the Eta vertical coordinate). Some of the surface physical processes of the Pantanal biome may not be simulated accurately by the Eta RCM, leading to errors in the precipitation outputs. We noted that all simulated data show a high negative mean bias error in the extreme north of Brazil. A logical reason is the failure of the GCMs to capture local features such as the sea breeze-induced rainfall. The sea breeze circulation typifies a mesoscale atmospheric system from coastal areas. It is a specific local wind system (from sea to land) due to thermal differences between land and sea surfaces, which leads to low-level pressure anomalies. In the tropics, the mesoscale diurnal processes, such as the sea breeze, are particularly important and may occur in 3 out of every 4 days (Ahrens, 2010; National Research Council, 1992). As shown in Kousky (1980), the Amazonian coastal area is highly influenced by the sea breeze, with the formation and propagation of the line of convective activity inland. Thereof, we can relate the negative bias found in this region to the inability of the GCMs to capture this mesoscale system, resulting in lower amounts simulated than the observed ones. Moreover, using Eta RCM did not resolve this local feature. A logical reason is the uncertainty of the observations. The rain gauge stations in Brazil are not equally distributed over the biomes and the northern part of Brazil has the lowest density of stations (Xavier et al., 2016). Moreover, interpolation methods for generating the observational grid have uncertainties that can impact our results and must be considered. The inability of the GCMs to capture the local sea breeze influenced rainfall along with the observational gap in the northern part of the Amazon, resulting in a strong negative bias.

Based on the biases of the models to simulate temporal and spatial precipitation patterns (Equation 5), we calculated their suitability for the Brazilian biomes, which is graphically represented in a heatmap (Figure 10). The heatmap divides the errors into classes and each class has a color, ranging from blue (best) to yellow (worst). In general, models were capable of simulating the phase and the amplitude of the rainy and dry seasons for the large and well-defined season biomes. We highlight the excellent performance of models to simulate the dry season in the Amazon, Atlantic Forest, and Cerrado biomes. The Caatinga and Pantanal presented significant improvements on this property for the Eta/MIROC5. For the seasonal cycle, there was an overall good performance of models. MIROC5 presented the highest errors (worst performance) for Cerrado and Caatinga biomes, but the downscaling process improved their ability to simulate seasonal precipitation. Related to the annual precipitation (represented by the mean annual precipitation and annual variability), as well as for seasonal precipitation, the GCM MIROC5 was not capable of capturing the main characteristics of the Cerrado and Caatinga biomes. For all properties, simulations were improved by using Eta RCM in HadGEM2-ES data for the Pampa biome. For the same biome, MIROC5's simulations did not improve by using Eta RCM. In a general view of simulated precipitation over Brazil, the Eta RCM improved the results of HadGEM2-ES for many biomes, except for the Amazon and Caatinga, where the original GCM is more suitable than the downscaled data. Related to the MIROC5 family, the GCM simulations were improved in the Amazon, Cerrado, Caatinga and Pantanal by the Eta RCM. For the Atlantic Forest and Pampa, the Eta RCM worsened the simulations.

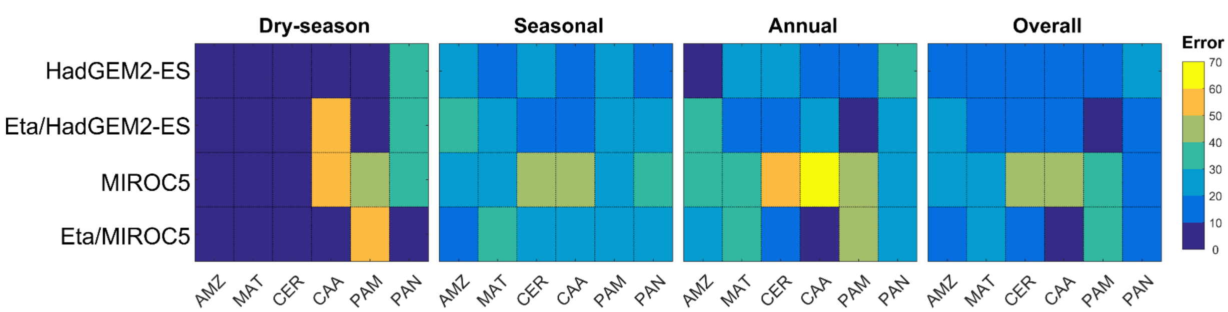

Figure 10: Heatmap of the relative error ($\in_j$) of precipitation properties for Brazilian biomes (Amazon – AMZ, Atlantic Forest – MAT, Cerrado – CER, Caatinga – CAA, Pampa – PAM, and Pantanal – PAN). The lower the error value, the better the model represents the dry season, seasonal and annual precipitation. The "Overall" refers to an integrated evaluation of models to simulate all the previous properties.

As we showed above, it is not a rule that the downscaling procedure will provide more suitable values of precipitation. The driving GCM HadGEM2-ES proved to be more suitable than Eta/HadGEM2-ES for large biomes such as the Amazon and Caatinga. At the same time, Eta/MIROC5 significantly improved the monthly means for almost all biomes and, consequently, the annual totals. For the Pampa biome, only the HadGEM2-ES family was capable of simulating the precipitation on acceptable levels. These results must be considered when the projections of these models are used. Moreover, our results support previous studies that aimed at the same RCM-GCM evaluation. Liang et al. (2008) observed very high spatial correlation between RCM minus GCM differences in precipitation and temperature between present and future climates, indicating that a major portion of the biases found on simulations (either for RCM and GCM) are systematically propagated into their future projections. In addition, they concluded that, even the uncertainty of future climate projections is sensitive to present climate simulation biases, there is no linear relationship between simulation and projections biases, depending on regions and models. We can infer that a model that better reproduces the present climate leads to more confidence in the physical and dynamical processes considered and represented by this model under boundary conditions applied, such as the historical GHG concentration. Considering that only the boundary conditions and scenarios are changed for project future climate, we can also expect a good representation of the climate for a given scenario by the model. And even with the advances in model developments and computational power, biases are still occurring (and sometimes increasing), and the identification of their causes is an actual need for assessing future impacts (Addor and Seibert, 2014). This highlights the importance of a more accurate assessment of the origin and incidence of models' biases, using adequate regions for the evaluation. As shown in Teutschbein & Seibert (2012), there is always a best bias correction method for a group of regions – achieving the best mean statistical results – but it is not always the best for all regions.

As is well known, GCMs and RCMs suffer from substantial biases, especially regarding precipitation (Flato et al., 2013; Kotlarski et al., 2014), and climate model precipitation usually needs to be bias corrected before these data are used for impact assessments. The most accurate choice of regions of assessment implies in a more accurate choice of bias correction method to be applied in future projections, enhancing climate change impact studies.

4. Conclusions

We evaluated the performance of the downscaled precipitation data from higher-resolution RCM simulations driven by two coarser-resolution GCM products (Eta/HadGEM2-ES and Eta/MIROC5). Both products have been used in the Brazilian Third National Communication to the UN Framework on Climate Change. We statistically analyzed the long-term means of the simulated precipitation compared to high-resolution observation-based gridded products to better understand the reliability of these simulations. Our analysis was conducted for the six main Brazilian biomes in order to consider areas with rather homogeneous (eco)climatological patterns and we evaluated the precipitation simulations in terms of monthly and seasonal means, thereby considering the separation into rainy and dry seasons. To the best of our knowledge, it is the most appropriate evaluation of high-resolution climate change datasets of precipitation for large areas in Brazil.

For the long-term mean monthly analysis, HadGEM2-ES and Eta/HadGEM2-ES simulated rainy and dry seasons very well in the Amazon, Atlantic Forest and Cerrado biomes. This result expresses the potential reliability of the GCM to simulate mean fields of precipitation in large areas. The GCMs require lower time and computational effort to process long-term data for large areas than RCMs and in this case, HadGEM2-ES presents itself as a viable alternative for larger Brazilian biomes. In turn, Eta/MIROC5 showed great improvements when compared to its driving-GCM MIROC5. In most cases, for all biomes, the downscaling brought the simulated means close to the observational means. In the Pampa biome, no model was able to represent the mean monthly precipitation well. However, in some cases, the biases embedded in the model simulations interfered in the identification and duration of rainy/dry seasons. The long-term mean seasonal analysis showed that the Eta RCM modifies the range of precipitation, with less reliability of models to simulate means in the dry season (JJA and SON). According to our heatmap, we recommend the following model for each biome: HadGEM2-ES for the Amazon, Eta/HadGEM2-ES for the Atlantic Forest, the Cerrado, and the Pampa, and Eta/MIROC5 for the Caatinga and the Pantanal.

The development of regional climate models for Brazil increases the country's ability to better understand the impacts of climate change. However, these data must be used with caution, as RCM simulations have systematic errors. Our results show that Eta/HadGEM2-ES and

Eta/MIROC5 data for Brazil have various biases, which can be originated from the driving GCMs, introduced by the downscaling RCM, and related to uncertainties in the observational data. As these models project rainfall data for the future as well, it is expected that these biases are also present in these projections and if these data are not corrected, any hydrological application will be compromised. When corrected, the climate change simulations and projections become a valuable tool for increasing resilience and decreasing environmental, social, and economic vulnerability.

5. Acknowledgments

This study was supported by grants from the Ministry of Science, Technology, Innovation and Communication – MCTIC and National Council for Scientific and Technological Development – CNPq [grants numbers 441289/2017-7 and 306830/2017-5]. This study was also financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and CAPES Print. Rafael Rosolem kindly acknowledges support from the Natural Environment Research Council (NERC) under the project "Brazilian Experimental datasets for MUlti-Scale interactions in the critical zone under Extreme Drought" (BEMUSED) [grant number NE/R004897/1]. Moreover, the authors are grateful to Dr. Chou Sin Chan of the National Institute of Space Research (INPE) for providing Eta HadGEM2-ES and MIROC5 data.

6. References

Addor, N., Seibert, J., 2014. Bias correction for hydrological impact studies – beyond the daily perspective. Hydrol. Process. 28, 4823–4828. https://doi.org/10.1002/hyp.10238

Ahrens, C.D., 2010. Essentials of meteorology: an invitation to the atmosphere, 6th ed.

Almagro, A., Oliveira, P.T.S., Nearing, M.A., Hagemann, S., 2017. Projected climate change impacts in rainfall erosivity over Brazil. Sci. Rep. 7, 1–12. https://doi.org/10.1038/s41598-017-08298-y

Brasil, 2016. Terceira Comunicação Nacional do Brasil à Convenção-Quadro das Nações Unidas sobre a Mudança do Clima. Ministério da Ciência, Tecnologia e Inovação.

Brazil, 2016. Modelagem Climática e Vulnerabilidades Setoriais à Mudança do Clima no Brasil. Ministério da Ciência, Tecnologia e Inovação, Brasília.

Brown, J.H., Maurer, B.A., 1989. Macroecology: The Division of Food and Space Among Species on Continents. Science (80-.). 243.

Chou, S.C., Lyra, A., Mourão, C., Dereczynski, C., Pilotto, I., Gomes, J., Bustamante, J., Tavares, P., Silva, A., Rodrigues, D., Campos, D., Chagas, D., Sueiro, G., Siqueira, G., Marengo, J., 2014a. Assessment of Climate Change over South America under RCP 4.5 and 8.5 Downscaling Scenarios. Am. J. Clim. Chang. 03, 512–527. https://doi.org/10.4236/ajcc.2014.35043

Chou, S.C., Lyra, A., Mourão, C., Dereczynski, C., Pilotto, I., Gomes, J., Bustamante, J., Tavares, P., Silva, A., Rodrigues, D., Campos, D., Chagas, D., Sueiro, G., Siqueira, G., Nobre, P., Marengo, J., 2014b. Evaluation of the Eta Simulations Nested in Three Global Climate Models. Am. J. Clim. Chang. 03, 438–454. https://doi.org/10.4236/ajcc.2014.35039

Christensen, J.H., Christensen, O.B., 2007. A summary of the PRUDENCE model projections of changes in European climate by the end of this century. Clim. Change 81, 7–30. https://doi.org/10.1007/s10584-006-9210-7

Coutinho, L.M., 2016. Biomas brasileiros, 1st ed. Oficina de textos, São Paulo.

Deidda, R., Marrocu, M., Caroletti, G., Pusceddu, G., Langousis, A., Lucarini, V., Puliga, M., Speranza, A., 2013. Regional climate models' performance in representing precipitation and temperature over selected Mediterranean areas. Hydrol. Earth Syst. Sci. 17, 5041–5059. https://doi.org/10.5194/hess-17-5041-2013

Diaconescu, E.P., Gachon, P., Laprise, E., 2015. On the Remapping Procedure of Daily Precipitation Statistics and Indices Used in Regional Climate Model Evaluation. J. Hydrometeorol. 16, 2301–2310. https://doi.org/10.1175/JHM-D-15-0025.1

Dosio, A., Paruolo, P., 2011. Bias correction of the ENSEMBLES high-resolution climate change projections for use by impact models: Evaluation on the present climate. J. Geophys. Res. 116, 1–22. https://doi.org/10.1029/2011JD015934

Ehret, U., Zehe, E., Wulfmeyer, V., Liebert, J., 2012. Should we apply bias correction to global and regional climate model data? HESS 16, 3391–3404. https://doi.org/10.5194/hess-16-3391-2012

Falco, M., Carril, A.F., Menéndez, C.G., Zaninelli, P.G., Li, L.Z.X., 2018. Assessment of CORDEX simulations over South America: added value on seasonal climatology and resolution considerations. Clim. Dyn. 52, 4771–4786. https://doi.org/10.1007/s00382-018-4412-z

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S.C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C., Rummukainen, M., 2013. Evaluation of Climate Models, in: Stocker, T.F., Qin, D., Plattner, G.K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P.. (Eds.), Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK and New York, USA, pp. 741–866. https://doi.org/10.1017/CBO9781107415324

Gibbs, H.K., Rausch, L., Munger, J., Schelly, I., Morton, D.C., Noojipady, P., Barreto, P., Micol, L., Walker, N.F., Gibbs, B.H.K., Rausch, L., Munger, J., Schelly, I., Morton, D.C., Noojipady, P., Barreto, P., Micol, L., Walker, N.F., Amazon, B., Cerrado, E., 2014. Brazil's Soy Moratorium. Sci. - Policy Forum Environ. Dev. 347, 377–378. https://doi.org/10.1126/science.aaa0181

Giorgi, F., 1990. Simulation of Regional Climate Using a Limited Area Model Nested in a General Circulation Model. J. Clim. https://doi.org/10.1175/1520-0442(1990)003<0941:SORCUA>2.0.CO;2

Gregory, J.M., Jones, P.D., Wigley, T.M.L., 1991. Precipitation in britain: an analysis of area-average data updated to 1989 1. Int. J. Climatol. 11, 331–345.

Hansen, G., Cramer, W., 2015. Global distribution of observed climate change impacts. Nat. Clim. Chang. 5, 182–185. https://doi.org/10.1038/nclimate2529

Huxman, T.E., Smith, M.D., Fay, P.A., Knapp, A.K., Shaw, M.R., Loik, M.E., Smith, S.D., Tissue, D.T., Zak, J.C., Weltzin, J.F., Pockman, W.T., Sala, O.E., Haddad, B.M., Harte, J., Koch, G.W., Schwinning, S., Small, E.E., Williams, D.G., 2004. Convergence across biomes to a common rain-use efficiency. Nature 1–4. https://doi.org/10.1038/nature02597.1.

IPCC, 2014. Climate Change 2013 – The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781107415324

IPCC, 2013. Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. https://doi.org/10.1017/CBO9781107415324

Jones, C.D., Hughes, J.K., Bellouin, N., Hardiman, S.C., Jones, G.S., Knight, J., Liddicoat, S., Connor, F.M.O., Andres, R.J., Bell, C., Boo, K.-O., Bozzo, A., Butchart, N., Cadule, P., Corbin, K.D., Doutriax-boucher, M., Friedlingstein, P., Gornall, J., Gray, L., Halloran, P.R., Hurtt, G.C., Ingram, W.J., Lamarque, J.F., Law, R.M., Meinshausen, M., Osprey, S., Palin, E.J., Parsons Chini, L., Raddatz, T., Sanderson, M.G., Sellar, A.A., Schurer, A., Valdes, P., Wood, N., Woodward, S., Yoshioka, M., Zerroukat, M., 2011. The HadGEM2-ES implementation of CMIP5 centennial simulations. Geosci. Model Dev. 4, 543–570. https://doi.org/10.5194/gmd-4-543-2011

Jones, P.D., 1999. First- and Second-Order Conservative Remapping Schemes for Grids in Spherical Coordinates. Mon. Weather Rev. 127, 2204–2210.

Junk, W.J., Cunha, C.N., Wantzen, K.M., Petermann, P., Strüssmann, C., Marques, M.I., Adis, J., 2006. Biodiversity and its conservation in the Pantanal of Mato Grosso, Brazil. Aquat. Sci. 68, 278–309. https://doi.org/10.1007/s00027-006-0851-4

Knutti, R., Sedláček, J., 2012. Robustness and uncertainties in the new CMIP5 climate model projections. Nat. Clim. Chang. 3, 1–5. https://doi.org/10.1038/nclimate1716

Kotlarski, S., Keuler, K., Christensen, O.B., Colette, A., Déqué, M., Gobiet, A., Goergen, K., Jacob, D., Lüthi, D., van Meijgaard, E., Nikulin, G., Schär, C., Teichmann, C., Vautard, R., Warrach-Sagi, K., Wulfmeyer, V., 2014. Regional climate modeling on European scales: a joint standard evaluation of the EURO-CORDEX RCM ensemble. Geosci. Model Dev. 7, 1297–1333. https://doi.org/10.5194/gmd-7-1297-2014

Kousky, V.E., 1980. Diurnal rainfall variation in Northeast Brazil. Mon. Weather Rev. 108.

Lapola, D.M., Martinelli, L.A., Peres, C.A., Ometto, J.P.H.B., Ferreira, M.E., Nobre, C.A., Aguiar, A.P.D., Bustamante, M.M.C., Cardoso, M.F., Costa, M.H., Joly, C.A., Leite, C.C., Moutinho, P., Sampaio, G., Strassburg, B.B.N., Vieira, I.C.G., 2014. Pervasive transition of the Brazilian land-use system. Nat. Clim. Chang. 4, 27–35. https://doi.org/10.1038/nclimate2056

Li, W., Fu, R., Dickinson, R.E., 2006. Rainfall and its seasonality over the Amazon in the 21st century as assessed by the coupled models for the IPCC AR4. J. Geophys. Res. Atmos. 111, 1–14. https://doi.org/10.1029/2005JD006355

Liang, X., Kunkel, K.E., Meehl, G.A., Jones, R.G., Wang, J.X.L., 2008. Regional climate models downscaling analysis of general circulation models present climate biases propagation into future change projections. Geophys. Res. Lett. 35, 1–5. https://doi.org/10.1029/2007GL032849

Llopart, M., Reboita, M.S., da Rocha, R.P., 2019. Assessment of multi-model climate projections of water resources over South America CORDEX domain. Clim. Dyn. 54, 99–116. https://doi.org/10.1007/s00382-019-04990-z

Lupatini, M., Suleiman, A.K.A., Jacques, R.J.S., Antoniolli, Z.I., Kuramae, E.E., de Oliveira Camargo, F.A., Roesch, L.F.W., 2013. Soil-Borne Bacterial Structure and Diversity Does Not Reflect Community Activity in Pampa Biome. PLoS One 8. https://doi.org/10.1371/journal.pone.0076465

Maraun, D., Shepherd, T.G., Widmann, M., Zappa, G., Walton, D., Gutiérrez, J.M., Hagemann, S., Richter, I., Soares, P.M.M., Hall, A., Mearns, L.O., 2017. Towards process-informed bias correction of climate change simulations. Nat. Clim. Chang. 7, 764–773. https://doi.org/10.1038/nclimate3418

Mascaro, G., Viola, F., Deidda, R., 2018. Evaluation of Precipitation From EURO-CORDEX Regional Climate Simulations in a Small-Scale Mediterranean Site. J. Geophys. Res. Atmos. 123, 1604–1625. https://doi.org/10.1002/2017JD027463

McNutt, M., 2013. Climate Change Impacts. Science (80-.). 341, 435–435. https://doi.org/10.1126/science.1243256

Mello, C.R., Ávila, L.F., Viola, M.R., Curi, N., Norton, L.D., 2015. Assessing the climate change impacts on the rainfall erosivity throughout the twenty-first century in the Grande River Basin (GRB) headwaters, Southeastern Brazil. Environ. Earth Sci. 73, 8683–8698. https://doi.org/10.1007/s12665-015-4033-3

Melo, D. de C.D., Scanlon, B.R., Zhang, Z., Wendland, E., Yin, L., 2016. Reservoir storage and hydrologic responses to droughts in the Paraná River basin, south-eastern Brazil. Hydrol. Earth Syst. Sci. 20, 4673–4688. https://doi.org/10.5194/hess-20-4673-2016

Melo, D.D.C.D., Xavier, A.C., Bianchi, T., Oliveira, P.T.S., 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

Menezes, R., Sampaio, E., Giongo, V., Pérez-Marin, A., 2012. Biogeochemical cycling in terrestrial ecosystems of the Caatinga Biome. Brazilian J. Biol. 72, 643–653. https://doi.org/10.1590/S1519-69842012000400004

Mitchell, T.D., Jones, P.D., 2005. An improved method of constructing a database of monthly climate observations and associated high-resolution grids. Int. J. Climatol. 25, 693–712. https://doi.org/10.1002/joc.1181

Morellato, L.P.C., Haddad, C.F.B., 2000. Introduction: The Brazilian Atlantic Forest. Biotropica 32, 786–792. https://doi.org/10.1111/j.1744-7429.2000.tb00618.x

Murray-Tortarolo, G., Jaramillo, V.J., Maass, M., Friedlingstein, P., Sitch, S., 2017. The decreasing range between dry- and wet-season precipitation over land and its effect on vegetation primary productivity. PLoS One 12, 1–11.

Myneni, R.B., Yang, W., Nemani, R.R., Huete, A.R., Dickinson, R.E., Knyazikhin, Y., Didan, K., Fu, R., Negron Juarez, R.I., Saatchi, S.S., Hashimoto, H., Ichii, K., Shabanov, N. V., Tan, B., Ratana, P., Privette, J.L., Morisette, J.T., Vermote, E.F., Roy, D.P., Wolfe, R.E., Friedl, M.A., Running, S.W., Votava, P., El-Saleous, N., Devadiga, S., Su, Y., Salomonson, V. V., 2007. Large seasonal swings in leaf area of Amazon rainforests. Proc. Natl. Acad. Sci. 104, 4820–4823. https://doi.org/10.1073/pnas.0611338104

Nadeem, I., Formayer, H., Yaqub, A., 2019. Effect of 1-km Subgrid Land-Surface Heterogeneity on the Multi-year Simulation of RCM-Modelled Surface Climate Over the Region of Complex Topography. Earth Syst. Environ. 3, 367–379. https://doi.org/10.1007/s41748-019-00116-x

National Research Council, 1992. Coastal meteorology: A review of the state of the science. Washington, DC. https://doi.org/10.17226/1991

Oliveira, P.T.S., Nearing, M.A., Moran, M.S., Goodrich, D.C., Wendland, E., Gupta, H. V., 2014. Trends in water balance components across the Brazilian Cerrado. Water Resour. Res. 50, 7100–7114. https://doi.org/10.1002/2013WR015202

Paredes-trejo, F., Barbosa, H., 2017. Evaluation of the SMOS-Derived Soil Water Deficit Index as Agricultural Drought Index in Northeast of Brazil. Water 9, 377.

https://doi.org/10.3390/w9060377

Paredes-trejo, F., Barbosa, H.A., Spatafora, L.R., 2018. Assessment of SM2RAIN-Derived and State-of-the-Art Satellite Rainfall Products over Northeastern Brazil. Remote Sens. 10, 1093. https://doi.org/10.3390/rs10071093

Paredes-trejo, F.J., Barbosa, H.A., Kumar, T.V.L., 2017. Validating CHIRPS-based satellite precipitation estimates in Northeast Brazil. J. Arid Environ. 139, 26–40. https://doi.org/10.1016/j.jaridenv.2016.12.009

PBMC, 2013. Executive summary: Impacts, vulnerabilities and adaptation to climate change. Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil.

Pierce, D.W., Barnett, T.P., Santer, B.D., Gleckler, P.J., 2009. Selecting global climate models for regional climate change studies. Proc. Natl. Acad. Sci. 106, 8441–8446. https://doi.org/10.1073/pnas.0900094106

Pinheiro, E.A.R., Costa, C.A.G., De Araújo, J.C., 2013. Effective root depth of the Caatinga biome. J. Arid Environ. 89, 1–4. https://doi.org/10.1016/j.jaridenv.2012.10.003

Ribas, C.R., Schoereder, J.H., 2007. Ant communities, environmental characteristics and their implications for conservation in the Brazilian Pantanal. Biodivers. Conserv. 16, 1511–1520. https://doi.org/10.1007/s10531-006-9041-x

Roesch, L.F.W., Vieira, F.C.B., Pereira, V.A., Schünemann, A.L., Teixeira, I.F., Senna, A.J.T., Stefenon, V.M., 2009. The Brazilian Pampa: A fragile biome. Diversity 1, 182–198. https://doi.org/10.3390/d1020182

Solman, S.A., Blázquez, J., 2019. Multiscale precipitation variability over South America: Analysis of the added value of CORDEX RCM simulations. Clim. Dyn. 53, 1547–1565. https://doi.org/10.1007/s00382-019-04689-1

Sombroek, W., 2001. Spatial and Temporal Patterns of Amazon Rainfall. Ambio 30, 388–396.

Souza, R., Feng, X., Antonino, A., Montenegro, S., Souza, E., Porporato, A., 2016. Vegetation response to rainfall seasonality and interannual variability in tropical dry forests. Hydrol. Process. 30, 3583–3595. https://doi.org/10.1002/hyp.10953

Tapiador, F.J., Navarro, A., Moreno, R., Sánchez, J.L., García-Ortega, E., 2019. Regional climate models: 30 years of dynamical downscaling. Atmos. Res. 235, 104785. https://doi.org/10.1016/j.atmosres.2019.104785

Taylor, K.E., Stouffer, R.J., Meehl, G. a., 2012. An overview of CMIP5 and the experiment design. Bull. Am. Meteorol. Soc. 93, 485–498. https://doi.org/10.1175/BAMS-D-11-00094.1

Teutschbein, C., Seibert, J., 2012. Bias correction of regional climate model simulations for hydrological climate-change impact studies : Review and evaluation of different methods. J. Hydrol. 456–457, 12–29. https://doi.org/10.1016/j.jhydrol.2012.05.052

Villarini, G., Mandapaka, P. V, Krajewski, W.F., Moore, R.J., 2008. Rainfall and sampling uncertainties : A rain gauge perspective. J. Geophys. Res. Atmos. 113, 1–12. https://doi.org/10.1029/2007JD009214

Watanabe, M., Suzuki, T., O'ishi, R., Komuro, Y., Watanabe, S., Emori, S., Takemura, T., Chikira, M., Ogura, T., Sekiguchi, M., Takata, K., Yamazaki, D., Yokohata, T., Nozawa, T., Hasumi, H., Tatebe, H., Komioto, M., 2010. Improved Climate Simulation by MIROC5: Mean States, Variability, and Climate Sensitivity. J. Clim. 23, 6312–6335. https://doi.org/10.1175/2010JCLI3679.1

Wu, M., Nikulin, G., Kjellström, E., Belušić, D., Lindstedt, D., 2020. The impact of regional climate model formulation and resolution on simulated precipitation in Africa. Earth 11, 377–394.

Xavier, A.C., King, C.W., Scanlon, B.R., 2016. Daily gridded meteorological variables in Brazil (1980-2013). Int. J. Climatol. 36, 2644–2659. https://doi.org/10.1002/joc.4518

Xu, Z., Yang, Z., 2015. A new dynamical downscaling approach with GCM bias corrections and spectral nudging. J. Geophys. Res. Atmos. 3063–3084. https://doi.org/10.1002/2014JD022958.Received

Zhang, P., Anderson, B., Barlow, M., Tan, B., Myneni, R.B., 2004. Climate-related vegetation characteristics derived from Moderate Resolution Imaging Spectroradiometer (MODIS) leaf area index and normalized difference vegetation index. J. Geophys. Res. Atmos. 109. https://doi.org/10.1029/2004JD004720

Zhao, Z.-C., Luo, Y., Huang, J.-B., 2013. A Review on Evaluation Methods of Climate Modeling. Adv. Clim. Chang. Res. 4, 137–144. https://doi.org/10.3724/SP.J.1248.2013.137

CHAPTER 2

CABRA: A NOVEL LARGE-SAMPLE DATASET FOR BRAZILIAN CATCHMENTS

Almagro, A., Oliveira, P.T.S., Meira Neto, A.A., Roy, T., Troch, P., 2021. CABra: a novel large-sample dataset for Brazilian catchments. Hydrology and Earth System Sciences, 25, 3105–3135. https://doi.org/10.5194/hess-25-3105-2021. (Impact Factor 2021: 5.748)

Abstract

In this paper, we present the Catchments Attributes for Brazil (CABra), which is a large-sample dataset for Brazilian catchments that includes long-term data (30 years) for 735 catchments in eight main catchment attribute classes (climate, streamflow, groundwater, geology, soil, topography, land-cover, and hydrologic disturbance). We have collected and synthesized data from multiple sources (ground stations, remote sensing, and gridded datasets). To prepare the dataset, we delineated all the catchments using the Multi-Error-Removed Improved-Terrain Digital Elevation Model and the coordinates of the streamflow stations provided by the Brazilian Water Agency, where only the stations with 30 years (1980-2010) of data and less than 10% of missing records were included. Catchment areas range from 9 to 4,800,000 km² and the mean daily streamflow varies from 0.02 to 9 mm day^{-1}. Several signatures and indices were calculated based on the climate and streamflow data. Additionally, our dataset includes boundary shapefiles, geographic coordinates, and drainage area for each catchment, aside from more than 100 attributes within the attribute classes. The collection and processing methods are discussed along with the limitations for each of our multiple data sources. The CABra intends to improve the hydrology-related data collection in Brazil and pave the way for a better understanding of different hydrologic drivers related to climate, landscape, and hydrology, which is particularly important in Brazil, having continental-scale river basins and widely heterogeneous landscape characteristics. In addition to benefitting catchment hydrology investigations, CABra will expand the exploration of novel hydrologic hypotheses and thereby advance our understanding of Brazilian catchments' behavior. The dataset is freely available at https://doi.org/10.5281/zenodo.4070146 and https://thecabradataset.shinyapps.io/CABra/.

Keywords: hydrology, climate, large-sample, database, big data.

1. Introduction

The integrated assessment of large-sample catchment attributes is fundamental for the description and classification of landscape properties, leading to an improved understanding of similarities (or dissimilarities) between catchments. Large-sample catchment hydrology is essential in terms of hydrological processes understanding (Addor et al., 2020; Beven et al., 2020). It provides an attractive venue for general inferences that would otherwise be impossible to study based on individual or small groups of catchments, aside from allowing the testing of new and existing hypotheses in hydrologic sciences (Addor et al., 2017; Gupta et al., 2014; Lyon and Troch, 2010; Wagener et al., 2007).

A classic example of a large catchment-scale dataset is the Model Parameter Estimation Experiment (MOPEX) (Duan et al., 2006; Schaake et al., 2006), with hydrologic time series from 438 catchments located within the continental US (CONUS). The MOPEX dataset has been used in several studies supporting theoretic and modeling advances in hydrologic sciences (Ao et al., 2006; Ren et al., 2016; Sawicz et al., 2011). A more recent example is the Catchment Attributes and MEteorological for Large-sample Studies (CAMELS, Addor et al. (2017)) consisting of a set of daily hydrometeorological time series data for 671 small- to medium-sized catchments for the CONUS, aside from several landscape and climate related attributes. The CAMELS initiative has been widely used and other large-sample datasets have been recently developed following the CAMELS format, such as CAMELS-GB for Great Britain, covering 671 catchments, CAMELS-CL for Chile, covering 516 catchments, and CAMELS-BR for Brazil, covering 897 catchments. A list of available large-sample datasets can be found in Addor et al. (2020).

Brazil is a country with continental dimensions, hosting a wide range of climates, soils, geology, and land-cover types. Despite covering almost 50% of South America and hosting between 12% and 18% of the world's renewable freshwater (Rodrigues et al., 2015; UNEP and ANA, 2007), Brazil suffers from scarce allocation of funds for hydrological monitoring services, which creates great challenges for the proper monitoring of the quality and quantity of its water resources. While the density of streamflow gauges falls below the standards recommended by the World Meteorological Organization (WMO) of 1 station for each 1,000 km², hydrologic observations are often discontinued and lack proper length (ANA, 2019a; WMO, 2010). An integrated dataset containing multiple levels of environmental information can be of extreme importance to leverage investigations in hydrology and related disciplines within the Brazilian territory.

Recently, two large-sample datasets for catchment attributes were developed for Brazil: the Catchment Attributes for Brazil (CABra) (first introduced in Oliveira et al., 2020) and the Catchment Attributes and MEteorology for Large-sample Studies (CAMELS-BR) (Chagas et al., 2020). Even though both datasets aim to fill the lack of hydrological data access in Brazil, the data sources, quality control, number, and types of attributes differ significantly. To address the similarities and differences between both datasets, an extensive discussion comparing CAMELS-BR and CABra is also presented in our study.

In this paper, we present the CABra dataset, which is a comprehensive, large-sample dataset for catchment attributes in Brazil. We have synthesized several multi-source data from eight main attribute classes (topography, climate, streamflow, groundwater, soil, geology, land-use and land-cover, and hydrologic disturbance) for 735 catchments in Brazil. Our dataset covers all Brazilian administrative and hydrographic regions as well as its biomes. We have delimited all the catchments using an error-corrected digital elevation model employing automatic drainage area delineation methods. For the area-averaged attributes, we have used national datasets from the Brazilian Water Agency (ANA), Brazilian Agricultural Research Corporation (EMBRAPA), and Xavier et al. (2016), and widely used global datasets, such as ERA5, SoilGrids250, Global Land Evaporation Amsterdam Model (GLEAM), Global Lithologic Map (GLiM), and GLobal HYdrogeology MaPS (GLHYMPS). Additionally, a hydrologic disturbance index was created to indicate the most human-impacted catchments. Finally, we discuss the spatial variabilities of the attributes and their limitations of application.

2. The CABra dataset

2.1. Overview

The CABra dataset is a multi-source, multi-temporal, and multi-spatial resolution large-sample dataset for catchment attributes for Brazilian catchments. Using an extensive local/global high-quality data collection, we developed CABra considering eight main classes of attributes: topography, climate, streamflow, groundwater, soil, geology, land-cover, and hydrological disturbance. Gridded datasets of various kinds were averaged onto the selected catchments located over Brazil and neighboring countries, in the case of transboundary catchments. Moreover, we provide daily time series from climate and streamflow variables for a 30-year period, covering the hydrological years from 1980 to 2010, as described in Fig. 1.

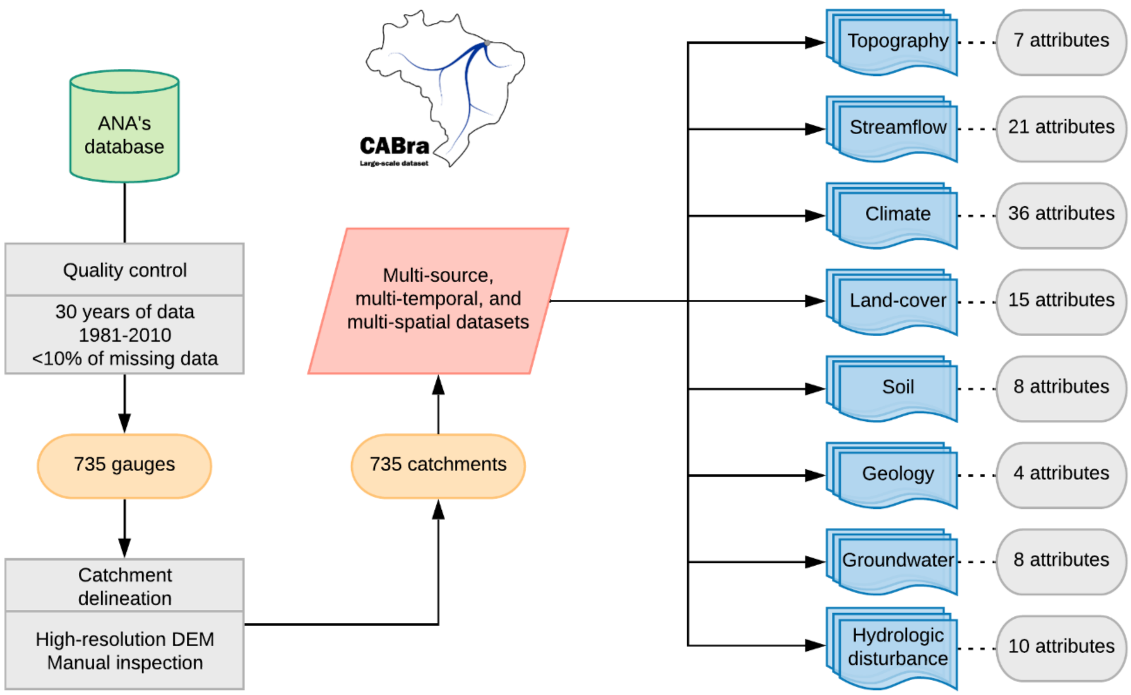

Figure 1: Study delineation for the CABra dataset organization. From ANA's database, 735 were selected to integrate our dataset due to its high consistency and long time series of streamflow.

The CABra dataset is recommended for a wide range of users for decision-making at multiple scales – local, national, or regional – covering all Brazilian biomes (Amazon, Cerrado, Atlantic Forest, Pantanal, Caatinga, and Pampa). CABra was created to ensure easy access to its information and provide high-quality data, with attributes useful for a variety of hydrometeorological modeling and assessments. Each catchment presents several attributes, ranging from the file information described in Table 1 to the attributes described throughout this article. Moreover, we made available all the geospatial data (shapefile of the boundaries) for the users.

Table 1: General attributes of the CABra catchments.

Type	**Attribute**	**Long name**	**Unit**
Identification	cabra_id	CABra's identification code of the streamflow gauge	-
	ana_id	ANA's identification code of the streamflow gauge	-
Location	longitude	Longitude coordinate of the streamflow gauge	dd
	latitude	Latitude coordinate of the streamflow gauge	dd
	gauge_hreg	Brazilian hydrographic region of the streamflow gauge	-
	gauge_biome	Brazilian biome of the streamflow gauge location	-
	gauge_state	Brazilian state of the streamflow gauge location	-
Quality	missing_data	Percentage of missing data	%
	series_length	Timeseries length of the streamflow gauge	years
	quality_index	Quality index of the CABra catchment records	-

- Means dimensionless

2.2. Catchment delineation and topography

Brazil does not have an official database for the national catchments boundaries, and the Brazilian Water Agency (ANA) does not make available its geospatial database. Because of this and to avoid uncertainties in the existing datasets for South America, we freshly generated all the CABra catchments boundaries used in this study. Digital Elevation Model (DEM) quality and resolution are crucial at this stage since all the post-analysis with the multi-source information utilized in the CABra dataset are area-averaged. For example, is well-known that errors in topographic indices, e.g., slope and catchment area and boundary, are dependent on and highly sensitive to DEM resolution and accuracy, and it is suggested that, if available, a high-resolution DEM should be used instead of a low-resolution DEM due the negative effects of terrain generalization caused by them (Mukherjee et al., 2012; Vaze et al., 2010; Wechsler, 2007; Zhou and Liu, 2004). We delineated the CABra catchments following the procedure described in Maidment (2002), using streamflow gauges location information from the ANA’s database and a high-resolution elevation product, i.e., the Multi-Error-Removed Improved-Terrain Digital Elevation Model with a 90-m spatial resolution at Equator (Yamazaki et al., 2017) (Fig. 2).

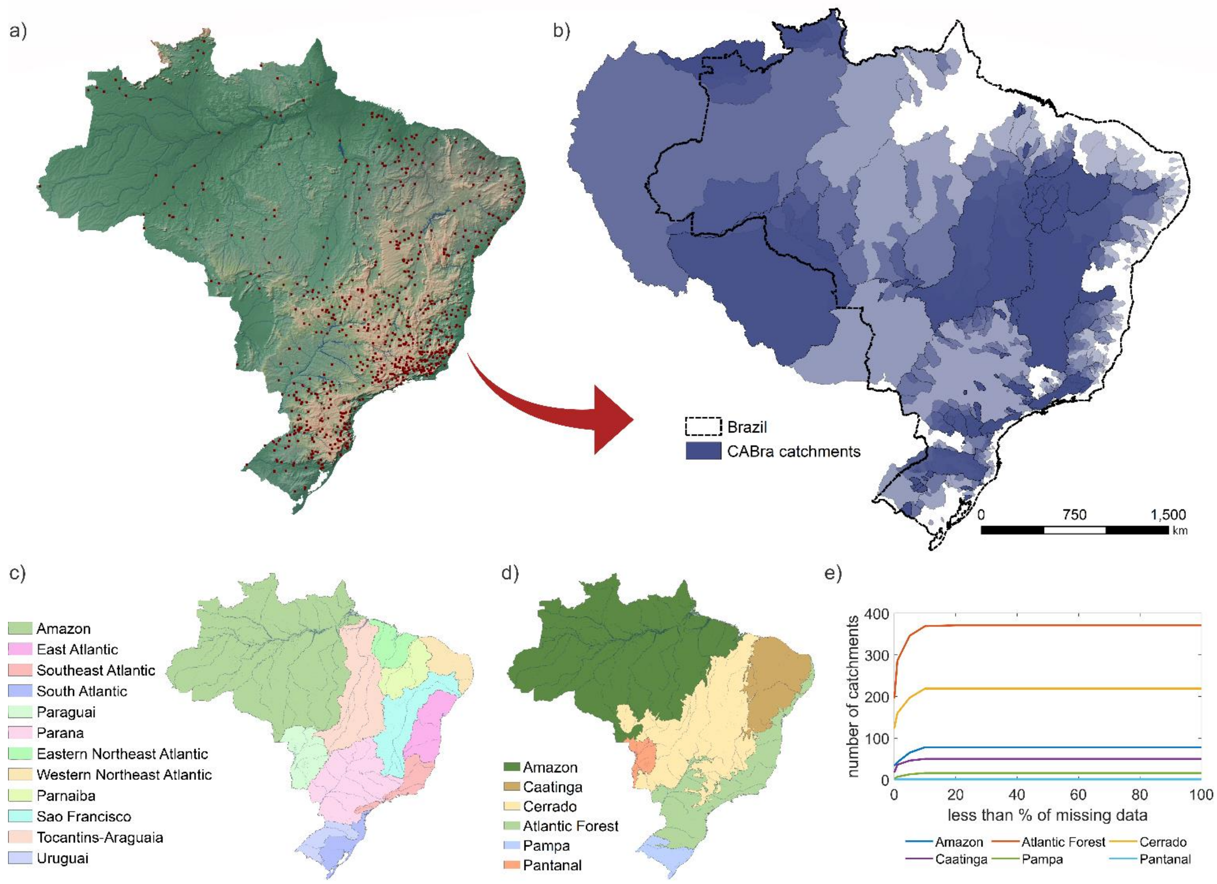

Figure 2: Location map of the streamflow gauges and CABra catchments. a. Streamflow gauges coordinates of CABra catchments; b. The 735 CABra catchments boundaries; c. The 12 hydrographic regions of Brazil; d. The six main biomes of Brazil; e. Level of consistency of the streamflow gauges records for each biome.

In the first stage, which we call "terrain processing", the DEM was sink-filled to avoid possible errors due to peaks or depressions. Then, the flow direction and flow accumulation were calculated, which indicates the direction and accumulation of flow, respectively, in each grid cell within the catchment. The next step was to define the stream network in the catchment. For the definition of a river stream, we considered a threshold of 100 cells accumulating water, and this value was chosen considering the DEM spatial resolution and the range of the size of the catchments. All the previous steps were run for the South America extension. Even though all outlets are located in the Brazilian territory, some of the drainage areas embrace larger areas outside of it. The second step was catchment delineation, where the products generated in the previous step and the coordinates of the streamflow gauges were used. Each streamflow gauge coordinate was first plotted as a point and the position of it to the stream network was checked and corrected, if necessary. The correction procedure was performed for 132 out of CABra catchments. Then, each corrected point was used as an outlet of the catchment and the delineation of the drainage area was performed using the ArcHydro tool. Aside from the catchments limits, perimeters, and areas, we also extracted the stream information, such as the stream network and hierarchy (Strahler, 1952, 1957). It is important to highlight that we manually inspected each catchment outlet and area to overcome the limitation of unchecked boundaries of another existing catchment datasets, such as Do et al. (2018), which is based on a DEM with a spatial resolution of 500-m. Moreover, this presented itself as a crucial procedure for an accurate delineation since several outlets' positions needed to be corrected to represent the real expected catchment boundary. Once the catchment boundaries were delimited, we calculated seven attributes related to the topography of each catchment: area, slope, maximum, minimum, and mean elevation, streamflow gauge elevation, and catchment order. The catchment boundaries and drainage network are also provided in CABra dataset.

Table 2: Topography attributes of the CABra catchments.

Type	**Attribute**	**Long name**	**Unit**
Elevation	elev_mean	Mean elevation of the catchment	m
	elev_max	Maximum elevation of the catchment	m
	elev_min	Minimum elevation of the catchment	m
	elev_gauge	Elevation of the streamflow gauge	m
Area	catch_area	Area of the catchment	km²
Slope	catch_slope	Mean slope of the catchment	%
Drainage	catch_order	Strahler order of the catchment	-

Figure 3 summarizes the topographic attributes for the CABra catchments. Catchment areas ranged from 9 to 4.8×10^{6} km² (Fig. 3a). This large range of areas shows how Brazilian hydrology can be, at the same time, local and continental, necessitating a better understanding of hydrologic processes on different scales. Many of the largest catchments are in the mainstream of one of the 12 hydrologic regions of Brazil, especially in the Amazon, Tocantins/Araguaia, São Francisco, Paraguay, and Paraná. The mean elevation of CABra catchments ranges from close to zero to up to 2000 m, with the highest values found in the southern and south-eastern portions. In turn, steepen areas can be found in the coastal and mountainous areas of the southeast and south (Fig. 3b and Fig. 3c). Most of the Brazilian catchments have a flat topography though, with a mean slope up to 10%. Figure 3d shows the gauge elevation. Note the difference between the gauge elevation and the mean catchment elevation in Fig. 3b. The gauge elevation considers only the elevation at the gauge position in the landscape, thereby proving only the local information, while the mean catchment elevation considers the average elevation for the entire catchment. An example of this difference is the largest CABra catchment, i.e., the Amazon. The mean elevation in the Amazon basin would be low, however, the western part of the basin has some of the highest peaks of the Andes, where the gauge elevation would be much higher.

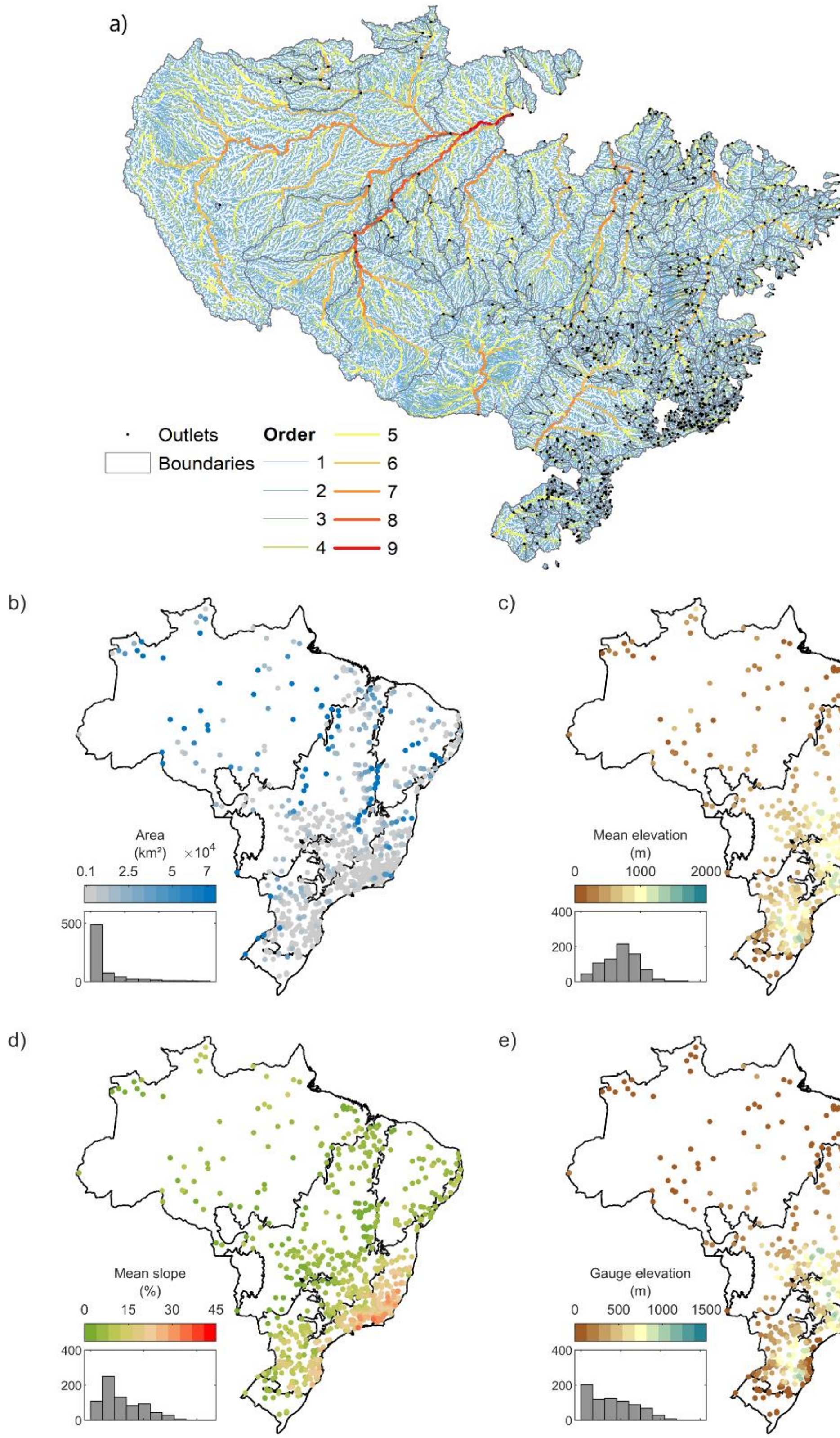

Figure 3: Spatial distribution of the topography attributes of the CABra catchments. a. Stream order of Brazilian rivers; b. Area of the catchments, in km²; c. Mean elevation of the catchments, in m; d. Mean slope of the catchments, in percentage; e. Elevation of the streamflow gauge, in m.

2.2.1. Uncertainty and limitations

The uncertainties related to the topography attributes are mainly related to the model terrain and streamflow gauges coordinates. The digital elevation model adopted for CABra catchments, developed by Yamazaki et al. (2017) is an improved product based on the composition of another baseline terrain products, such as the SRTM3 DEM, AW3D-30m DEM, and Viewfinder Panoramas DEM. Moreover, there are gaps in high-relief mountains and water bodies that were filled manually for the final MERIT-DEM product, leading to 72% of mapped area with height accuracy better than 2 m when slope < 10%. Regarding to streamflow gauges coordinates, there were inconsistencies between the location provided by ANA and the stream network generated using the MERIT-DEM. We corrected the pair of coordinates, by matching the point to the nearest stream network, in a way that the area error against ANA's area was minized. Regarding to the catchment delineation, the uncertainty related to the automatic procedure conducted at the SIG environment is mainly dependent on the accucarcy, but some authors found that channels heads (1st order catchments) are the most subjected to greatest uncertainties (Zandbergen, 2011).

2.3. Climate

2.3.1. Methodology

We present daily time series of area-averaged precipitation, minimum, maximum, and mean temperatures, solar radiation, relative humidity, wind speed, evapotranspiration, and potential evapotranspiration (calculated by Penman-Monteith, Priestley-Taylor, and Hargreaves methods). Moreover, we calculated several core climate indices, defined by the Climate and Ocean: Variability, Predictability, and Change project from the World Climate Research Programme (WCRP). Two main climate datasets were used in CABra. The first one, a high-resolution meteorological gridded dataset (0.25ºx0.25º), developed by Xavier et al. (2016) (here referred to as "REF") is based on the spatial interpolation of meteorological data from ~4,000 rain gauges and wheatear stations in Brazil, from the ANA, Brazilian Institute for Meteorology (INMET, in Portuguese), and Water and Power Department of São Paulo (DAEE/SP, in

Portuguese), covering the period from 1980 to 2015. From these sets of meteorological gauges, 2890 are limited to precipitation data. This dataset is available at http://careyking.com/data-downloads/. This product has a much finer spatial resolution and is based on a higher number of rain gauge stations than other widely used products (~4,000 stations for Brazil, in comparison to ~600 stations for South America in CRU TS3.1 product). However, the REF dataset covers only the Brazilian territory, while the CABra dataset has 20 catchments with upstream areas outside Brazil. To overcome this, we incorporated the ERA5 (Hersbach et al., 2020) climate data into the CABra dataset (here referred to as "ERA5").

ERA5 is the most recent version of climate reanalysis from the European Centre for Medium-Range Weather Forecasts (ECMWF) and provides hourly, daily, and monthly data on several atmospheric, sea, and land variables in a 0.25ºx0.25º spatial resolution grid, from 1950 to the present. As a reanalysis dataset, the ERA5 uses past observations and models to generate accurate and consistent time series of climate variables and parameters, being one of the widely used datasets in geosciences (Hersbach et al., 2020). To incorporate and produce a more reliable product for all the CABra catchments, we have generated an ensemble mean product (here referred to as "ENS") using both datasets beforementioned, i.e., REF and ERA5 climate products. The procedure was conducted in the Climate Data Operators (CDO, Schulzweida, 2019) and aimed to a better characterization and representation of the climate based on the two independent estimations, which generally imply in a more robust reproducibility of the phenomenon than in a single-member analysis (Abramowitz et al., 2018). Newman et al. (2015b) also found that ensemble product of precipitation and temperature still capture the main features of the variables and, moreover, improves the identification of extreme event frequency, and it is know that an ensemble usually outperforms individual forecasts (Bellucci et al., 2015; Solman et al., 2013; Tebaldi et al., 2005), being capable to detect internal variability and seasonal patterns. The ENS dataset generated here can be useful for climate-related analysis through the Brazilian territory, since it merges two high-resolution and high-quality products.

The precipitation seasonality (Woods, 2009), which indicates the timing of the precipitation seasonal cycle and the temperature seasonal cycle – values close to +1 indicates summer precipitation and values close to -1 indicates winter precipitation – was calculated for the ensemble product.

The actual evapotranspiration adopted in CABra is derived from the Global Land

Evaporation Amsterdam Model version 3 (GLEAM v3, Martens et al., 2017), which is a set of algorithms that estimates the many components of land evaporation based on satellite observations of climatic and environmental variables. The calculations of the actual evapotranspiration by GLEAM v3 take into account a potential evapotranspiration module (by Priestley and Taylor method), an interception loss module (by a Gash analytical model), and a stress module (by a semi-empirical relationship to root-zone moisture and vegetation optical depth). The GLEAM dataset is one of the most commonly used datasets on evapotranspiration applications (Forzieri et al., 2018; Schumacher et al., 2019; Zhang et al., 2016).

Even though the REF dataset presents a reference evapotranspiration product (calculated by Penman-Monteith method following the FAO-56 guidelines), it embraces only the Brazilian territory and did not comprise all the areas of the catchments included in the CABra dataset. To overcome this limitation, we calculated the daily potential evapotranspiration (PET) by three different widely used methods based on energy balance and transfer mass, radiation, and temperature, using meteorological variables from the ERA5 and the ensemble products as inputs. These three newly products are, to our knowledge, the most extensive datasets of potential evapotranspiration for Brazil, covering a larger period than the existent products, such as the one introduced in Althoff et al. (2020) and Xavier et al. (2016).

The first method was the FAO-56 Penman-Monteith equation (Allen et al., 1998), which is the standard for reference evapotranspiration, and assumes a hypothetical crop similar to a surface of small grass of uniform grass, actively growing and sufficiently watered. The FAO Penman-Monteith (PM) equation considers the energy budget and the aerodynamic and surface resistances of the crop and uses as inputs the solar radiation, air temperature, humidity, and 2m wind speed data (Equation 1).

$$PET_{PM} = \frac{0.408\Delta(R_n - G) + \gamma \frac{900}{T + 273} u_2 (e_s - e_a)}{\Delta + \gamma(1 + 0.34u_2)} \tag{1}$$

where PET_{PM} is the reference evapotranspiration, in mm day^{-1}, R_n is the net radiation, in MJ m^{-2} day^{-1}, G is the soil heat flux, in MJ m^{-2} day^{-1}, T is the mean daily temperature at 2m height, in ºC, u_2 is the wind speed at 2m height, in m s^{-1}, e_s is saturation vapor pressure, in kPa, e_a is the actual vapor pressure, in kPa, Δ is the slope vapor pressure curve, in kPa ºC^{-1}, and γ is the psychrometric constant, in kPa ºC^{-1}.

The radiation-based method chosen for the CABra dataset is the Priestley-Taylor equation (PT) (Priestley and Taylor, 1972). The PT considers that when large areas, such as catchments, are saturated, the main force that governs the evaporation is the net radiation, and under certain conditions, the knowledge of net radiation and the ground dryness is enough to determine the vapor and sensible heat fluxes at the surface. Moreover, it is one of the most commonly used models to estimate evapotranspiration due to its low number of inputs requirement (Maes et al., 2018; McMahon et al., 2013; Shuttleworth, 1996). The PT equation takes the following form:

$$PET_{PT} = \alpha \frac{\Delta}{\Delta + \gamma}(R_n - G) \quad 2$$

where PET_{PT} is the potential evapotranspiration, in mm day^{-1}, α is the Priestley-Taylor constant, dimensionless, R_n is the net radiation, in MJ m^{-2} day^{-1}, G is the soil heat flux, in MJ m^{-2} day^{-1}, Δ is the slope vapor pressure curve, in kPa ºC^{-1}, and γ is the psychrometric constant, in kPa ºC^{-1}. Considering that PT only considers daytime evapotranspiration and G is negligible during the daytime, we used $G = 0$ in our calculations.

Priestley & Taylor (1972) empirically determined α for many locations and conditions in the world, ranging between 1.08 and 1.34. The authors concluded the best estimation for α should be an overall mean of 1.26. However, it is known that the α value is scenario-dependent and its variability is not taken into account when using the mean value proposed in its development (Guo et al., 2007).

The third method adopted here is the Hargreaves equation. The method was developed by Hargreaves (1975) for irrigation planning and design and it is a temperature-based equation widely used to calculate the potential evapotranspiration due to its easy application and low inputs requirement (Equation 3).

$$PET_{HG} = 0.0135\ R_s(T_a + 17.8) \quad 3$$

where PET_{HG} is the potential evapotranspiration, in mm day^{-1}, R_s is the solar radiation, in MJ m^{-2} day^{-1}, and T_a is the daily mean temperature, in ºC.

From the climatic variables and attributes, we carried out an analysis of the annual water balance in the Budyko space, an empirical approach applied to the study of the hydrological behavior of catchments. The Budyko hypothesis (Budyko, 1948, 1974) considers that the ratio between the long-term annual actual evapotranspiration (ET) and precipitation (P) is a function of the ratio between the long-term potential evapotranspiration (PET) and precipitation (P). The Budyko framework has been used to assess global impacts of climate change on water resources (Berghuijs et al., 2017; Roderick et al., 2014), and to gain further insight on water balance controls at mean annual timescales (Donohue et al., 2007; Berghuijs et al., 2017; Meira Neto et al., 2020).

Table 3: Daily series of meteorological variables and climate indices for the CABra catchments.

Type	Attribute	Long name	Unit
Precipitation	p_ref	Daily precipitation from the REF dataset	mm day^{-1}
	p_era5	Daily precipitation from the ERA5 dataset	mm day^{-1}
	p_ens	Daily precipitation from the ENS dataset	mm day^{-1}
Temperature	tmax_ref	Daily maximum temperature from the REF dataset	ºC
	tmin_ref	Daily minimum temperature from the REF dataset	ºC
	tmax_era5	Daily maximum temperature from ERA5 dataset	ºC
	tmin_era5	Daily minimum temperature from ERA5 dataset	ºC
	tmax_ens	Daily maximum temperature from the ENS dataset	ºC
	tmin_ens	Daily minimum temperature from the ENS dataset	ºC
Solar radiation	srad_ref	Daily mean solar radiation from the REF dataset	MJ m² day^{-1}
	srad_era	Daily mean solar radiation from the ERA5 dataset	MJ m² day^{-1}
	srad_ens	Daily mean solar radiation from the ENS dataset	MJ m² day^{-1}
Wind	wnd_ref	Daily mean 2m wind speed from the REF dataset	m s^{-1}
	wnd_ era5	Daily mean 2m wind speed from the ERA5 dataset	m s^{-1}
	wnd_ ens	Daily mean 2m wind speed from the ENS dataset	m s^{-1}
Evaporation	et_act	Daily actual evapotranspiration from the GLEAM	mm day^{-1}
	pet_pm	Daily potential evapotranspiration (Penman-Monteith method)	mm day^{-1}
	pet_pt	Daily potential evapotranspiration (Priestley and Taylor method)	mm day^{-1}
	pet_hg	Daily potential evapotranspiration (Hargreaves method)	mm day^{-1}
Climate Indices	clim_p	Long-term mean daily precipitation	mm day^{-1}
	p_seasonality	Seasonality and timing of precipitation	-
	clim_rh	Long-term mean daily relative humidity	%

clim_tmin	Long-term mean daily minimum temperature	ºC
clim_tmax	Long-term mean daily maximum temperature	ºC
clim_et	Long-term mean daily actual evapotranspiration	mm day^{-1}
clim_pet	Long-term mean daily potential evapotranspiration	mm day^{-1}
aridity_index	Aridity index (clim_p/clim_pet) of the catchment	-
clim_srad	Long-term mean daily solar radiation	MJ m^2 day^{-1}
clim_quality	Quality index of climate indices (indicates the source meteorological daily series used for long-term mean calculation)	-

- Means dimensionless

2.3.2. Results and discussion

Figure 4 shows some of the climate attributes for the CABra dataset. Regarding the precipitation derived from our ensemble of Xavier et al. (2016) and ERA5 (Fig. 4a), we found the highest values, reaching up to 10 mm day^{-1}, in the northern portion, and the lowest values, below 1 mm day^{-1}, in the north-eastern portion. Despite the wide range in the daily precipitation, most of catchments (~80%) presented area-averaged precipitation between 3 and 6 mm day^{-1}.

Figure 4d shows the area-averaged solar radiation reaching the surface, ranging from 10 to 20 MJ m^2 day^{-1}, with most of the catchments with daily values higher than 15 MJ m^2 day^{-1}. The spatial distribution of solar radiation is reflected in the temperature values in CABra catchments (Fig. 4e and Fig. 4f). The southern and south-eastern portions present the lowest values of both the maximum and minimum temperatures. This is due to the lower values of solar radiation and high altitudes found in these regions of Brazil. Other areas of Brazil are located in higher latitudes and are subject to higher solar radiation, and due to its flat relief, the temperatures are higher than in the south. Figure 4b indicates that, in most of CABra catchments (~85%), the precipitation seasonal cycle is in timing with the temperature seasonal dynamics, which means that most of the precipitation occurs in the summer (seas > 0). There are only a few catchments in the northern portion of Brazil that have precipitation in the winter (seas < 0), and this can be explained by the high influence of sea breeze on convective precipitation in this region. According to Ahrens (2010) and Kousky et al. (1984), the Amazonian coastal area is highly influenced by the sea breeze, which can occur in 3 out of every 4 days, with the formation of convective activity inland.

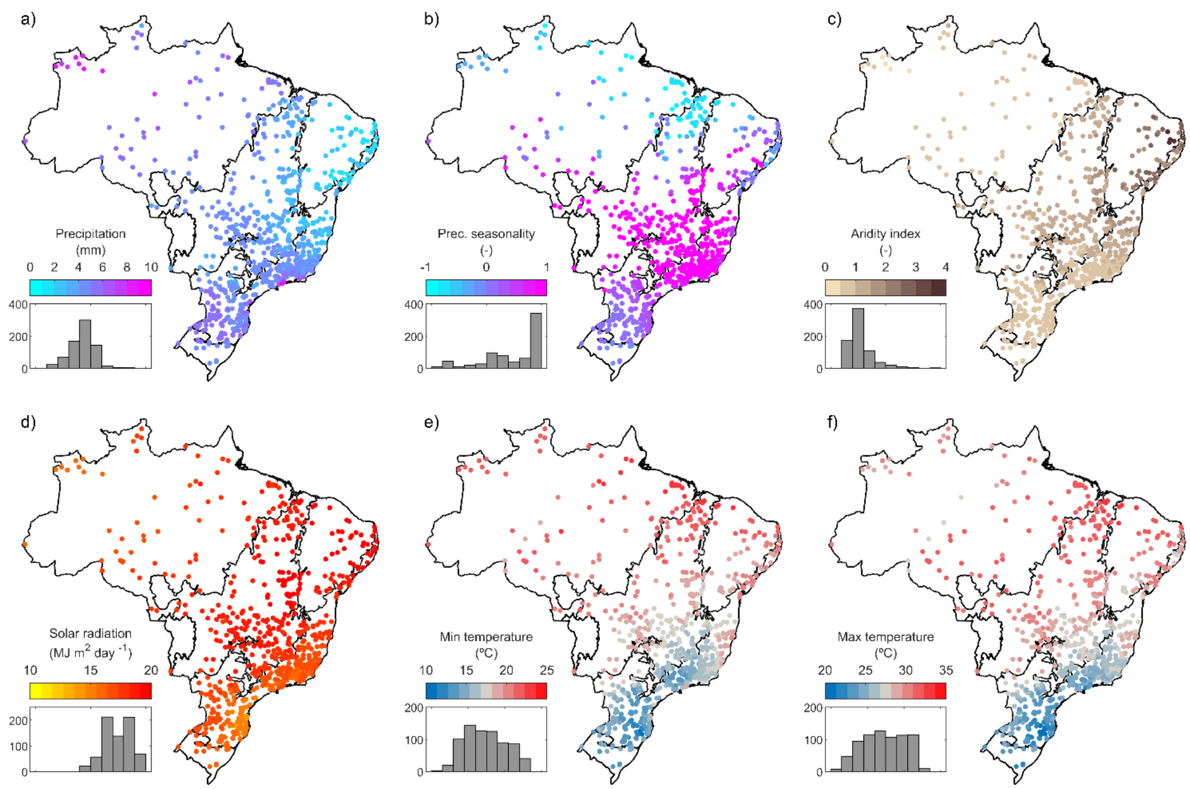

Figure 4: Spatial distribution of climate indices of the CABra catchments. a. Mean daily precipitation, in mm day^{-1}; b. Precipitation seasonality, dimensionless; c. Aridity index, dimensionless; d. Mean daily solar radiation, in MJ m^2 day^{-1}; e. Mean daily minimum temperature, in ºC; f. Mean daily maximum temperature, in ºC.

Our results of the computed potential evapotranspiration are presented in Fig. 5a, Fig. 5b, and Fig. 5c. They are related to three different methods for PET calculation, being: potential evapotranspiration for a reference crop using the Penman-Monteith equation; potential evapotranspiration by the Priestley-Taylor equation; and potential evapotranspiration by the Hargreaves equation. All the equations generated similar results of PET ranging from 3 to 6 mm day^{-1}, with similar spatial variability. The highest values were found for the north-eastern portion of Brazil, with the Penman-Monteith results being slightly higher than other equations. This could be related to the wind component in the method, which is not taken into account in the Priestley-Taylor and Hargreaves methods.

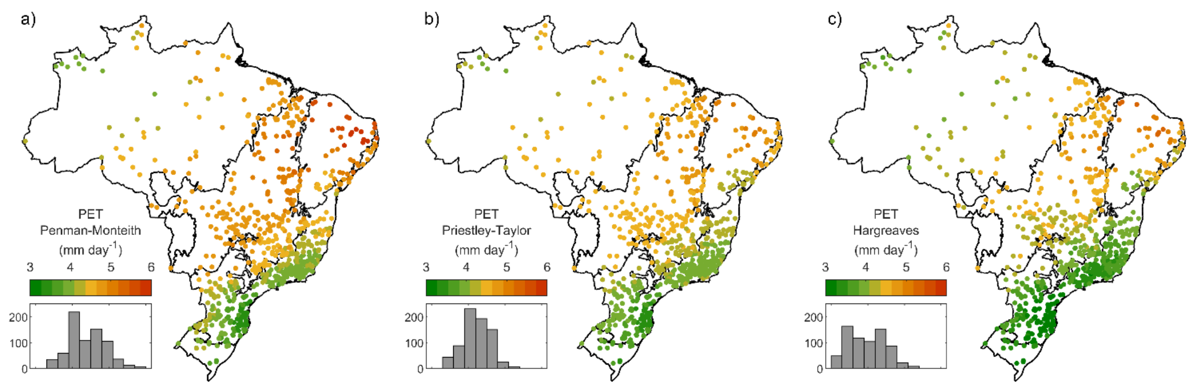

Figure 5: Spatial distribution of the PET calculated from three different methods of the CABra catchments. a. Penman-Monteith method; b. Priestley and Taylor method; c. Hargreaves method.

The Budyko framework (Budyko, 1948, 1974) shows that half of CABra catchments are water-limited and the other half are energy limited (Fig. 6). The lowest aridity index values are found in the Amazon and the Atlantic Forest, while the warmer and drier climate can be found in the Cerrado and Caatinga biomes. This may be correlated with the physiognomies of vegetation found in these biomes: tropical forests for the first group and grass and shrub for the second one, and especially, to the water availability and radiation incidence on these abovementioned biomes. Although we have found some outliers which are not explained by the Budyko hypothesis, most of the CABra catchments follow the expected behavior to the long-term mean water balance proposed by Budyko (1948, 1974). Moreover, we can note that the main climate features are captured by all the datasets, with catchments in Caatinga being more arid, followed by the Cerrado. The Atlantic Forest is in the same location at the Budyko space, while some catchments in Amazon only appears on ERA5 and ENS dataset, due to its extension outside REF. This shows the consistency between all datasets adopted in CABra.

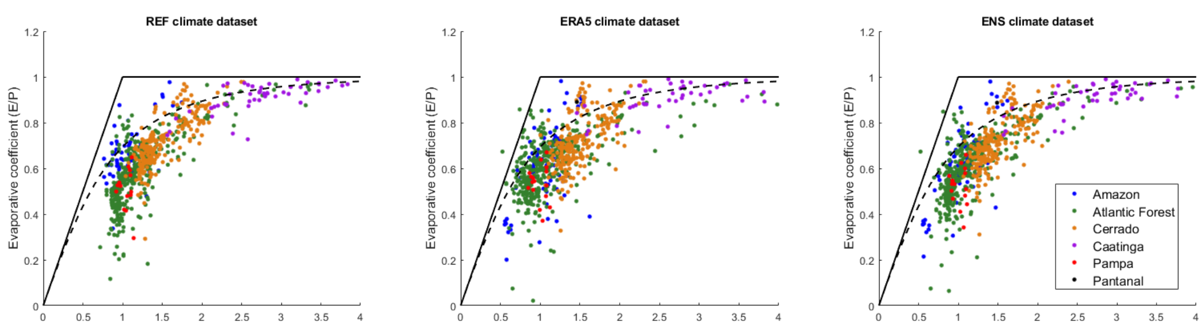

Figure 6: Distribution of the CABra catchments in the Budyko framework from the three different climate dataset of CABra: REF, ERA5 and ENS. Values of E were estimated from the relation P = E + Q, considering long-term means.

2.3.3 Uncertainty and limitations

The climate data provided by CABra dataset has limitations related to the number and spatial distribution of rainfall gauges in Brazilian territory that must be pointed. Since REF and ERA5 datasets are, respectively, ground-based and reanalysis gridded data, they are subject to uncertainties on the density of rainfall gauges network and in its post-processing procedures, which includes geospatial interpolation and data modelling and assimilation. In addition, REF dataset is not present in all of the 735 catchments due to its spatial extent, covering only the Brazilian territory. The quality of the data is presented for the users with a flag in the data though.

The potential evapotranspiration calculated for the CABra catchments are also subjected to uncertainties related to the equations chosen for the study and propagation of errors of input variables from climatic data. The golden standard for reference potential evapotranspiration is the Penman-Monteith method, and the main limitations are related to the other two methods: on the application of the Pristley & Taylor method, the requirement of the Priestley-Taylor constant α, which is related to the ratio between the actual evapotranspiration and the equilibrium evaporation rate (Eichinger et al., 1996), is one of the greatest sources of uncertainty because it is scenario-dependent and its variability is not considered by using the mean value ($\alpha = 1.26$) proposed in its development (Guo et al., 2007). On the other hand, the main limitation of Hargreaves equation for potential evapotranspiration is that the estimations are subject to error due to a large range of temperatures caused by weather fronts on a daily scale. On the other hand, it is a less biased model, when compared to other methods, when applied to small and not well-watered catchments (Hargreaves and Allen, 2003).

2.4. Streamflow and hydrologic signatures

2.4.1. Methodology

The CABra dataset provides daily streamflow records for 735 catchments in Brazil. We used data from streamflow gauges of ANA, where each gauge is related to one of the abovementioned catchments. This dataset is available in the HIDROWEB database (see http://www.snirh.gov.br/hidroweb/). ANA's database contains raw time series of dozens of thousands of gauges of streamflow, precipitation, water quality, and sediment discharge, with a consistency level for each observation. Due to the inconsistencies and missing records in the streamflow data provided by ANA, we implemented filters to take into account only the reliable data for the CABra dataset.

During our analysis, we found four main issues with ANA's database collected from HIDROWEB: (a) missing streamflow values for a period of the time series; (b) duplicate streamflow values with different consistency levels; (c) duplicate values with the same consistency level, and (d) duplicate dates with different values and consistent levels. In the first filter step, we overcame the last three issues by picking up only one of the duplicated values/dates based on the best level of consistency. The first issue is more complex and difficult to overcome as in some cases the missing data reaches almost 100% for some gauges. Since long time series of streamflow is needed for reliable hydrologic investigations, we defined a threshold for the selection of the streamflow gauges considered in the CABra dataset based on the following conditions: at least 30 years of data, comprising the hydrologic years from 1980 to 2010, with up to 10% of missing data. The application of these filters led to 735 streamflow gauges, and consequently, 735 catchments. During the analysis, we also noted inconsistences on streamflow gauges data, such as extremely high values (up to 1,000 mm day^{-1}) and unexpected changes on daily streamflow values. Such inconsistences can lead to an under/overestimation of signatures based on mean values (e.g., mean daily flow, aridity index, runoff ratio) and, when repeated for a long time, it can modify signatures based on the frequency and dynamics of streamflow (e.g., flow duration curve, high and low flows frequency and duration). To avoid carrying these issues to the signatures' calculation, we checked for outliers on the streamflow data by comparing each value to its neighbours. Elements with a value larger than five times the median of a sliding ten-elements window (centred in 'x') were considered as an invalid value (NaN).

After the employment of the filters, we calculated for the 735 selected catchments, a variety of hydrological signatures, which can provide a better understanding of the patterns of functionality and behavior of the catchments. From the quantification of hydrological characteristics, it is possible to explain the variability in responses to climate forcings. We selected hydrological signatures obtained from widely available hydrological series (see Table

4), as well as Sawicz et al. (2011) e Westerberg e McMillan (2015). A list with more hydrological signatures can be found in Yadav et al. (2007). All the hydrological signatures were calculated considering the hydrological years (October 1st – September 30th) from 1980 to 2010, as adopted by the Brazilian Water Agency in their annual reports (ANA, 2020a).

Table 4: Hydrological signatures of the CABra dataset.

Type	Attribute	Long name	Unit
Distribution	q_mean	Mean daily streamflow	mm day^{-1}
	q_1	Very low streamflow (1st quantile)	mm day^{-1}
	q_5	Low streamflow (5th quantile)	mm day^{-1}
	q_95	High streamflow (95th quantile)	mm day^{-1}
	q_99	Very high streamflow (99th quantile)	mm day^{-1}
Frequency and duration	q_hf	Frequency of high streamflow events	days y^{-1}
	q_hd	Duration of high streamflow events	days
	q_lf	Frequency of low streamflow events	days y^{-1}
	q_ld	Duration of low streamflow events	days
	q_hfd	Half-flow date	day of the year
	q_zero	Frequency of zero-flow events	days y^{-1}
Dynamics	baseflow_index	Baseflow index	-
	q_cv	Coefficient of variation of daily streamflow	-
	q_lv	Coefficient of variation of low-flows	-
	q_hv	Coefficient of variation of high-flows	-
	q_elasticity	Elasticity of daily streamflow	-
	fdc_slope	Slope of flow duration curve (between 33th and 66th percentiles)	-
Runoff	runoff_coef	Runoff ratio	-

- Means dimensionless

The hydrological signatures are based on the distribution of the streamflow, we have used the daily streamflow and its quantiles to define the mean daily streamflow, very low-, low-, high-, and very high-flows. For the calculation of frequency and duration of the streamflow, besides the number of days with no flow, the number of days was identified with 0.2 and 9 times the mean daily streamflow (low-flows and high-flows) and its number of days in sequence. The half-flow date corresponds to the day of the year in which the cumulated annual streamflow reaches half of the annual totals. The baseflow index was calculated using a recursive digital filter proposed by Lyne and Hollick (1979), presented in Ladson et al. (2013). Additionally, regarding to the dynamics of streamflow, we calculated the coefficients of variation of the streamflow (mean, low, and high), the streamflow elasticity proposed by (Sankarasubramanian et al., 2001), which indicates the impact of changes in precipitation on the streamflow, and the slope of flow duration curve between 33th and 66th quantiles, which is a good indicator of the perennial/non-perennial condition of the catchment. We also calculated the runoff coefficient for each catchment, which indicates how much of the precipitated water becomes streamflow by the simple ratio between mean daily streamflow and mean daily precipitation.

2.4.2. Results and discussion

Figure 7 shows the hydrologic signatures calculated for the CABra catchments for the period between the hydrologic years 1980 and 2010. The mean daily flow for the Brazilian catchments ranges from less than 1 mm day^{-1} to up to 9 mm day^{-1}, with an overall mean of 2 mm day^{-1}. The highest values were found in the extreme north of Amazon, where the daily flows reached 8 mm day^{-1} due to high amounts of precipitation through the year, and in the Atlantic Forest, in the southeast, where we also have steepness relief with higher values of the slope, providing the runoff instead of infiltration process. This can be seen in Fig. 7b, related to the runoff coefficient, where we noted the high values in the southern and north-western portions of Brazil. Most of the CABra catchments presented a runoff coefficient up to 0.5, though.

Our results also revealed that the Brazilian catchments are mainly dependent on the baseflow since all of them presented a baseflow index greater than 70%. The lowest values were found in the Caatinga biome, where we also found the lowest mean daily flows. The half-flow date (considering October 1st as the beginning of the hydrologic year) indicates that ~80% of Brazilian catchments reach half of the total accumulated annual flow in less than 200 days (Fig. 7d), showing the high correlation with the seasonal cycle of precipitation. The catchments with later dates of the half-flow day can be found in the Pampa biome, where there is no well-defined rainy/dry season, and in the Amazon, where the amounts of accumulated annual streamflow are too high and the peak of precipitation is near the end of the hydrologic year (Almagro et al., 2020). The analysis of the slope of the flow duration curve, in Fig. 7e, shows the lowest values in a great portion of Brazil, ranging from the Cerrado to the Atlantic Forest and Pampa biomes.

In our analyses, we also found zero values between the 33rd and 66th percentiles of the slope of flow duration curve in the north-eastern portion of Brazil, in the Caatinga biome, which indicates the existence of catchments with non-perennial rivers in that region, which are mainly dependent on direct runoff of rainfall. This can also be seen when analyzing Fig. 7f, related to the streamflow elasticity. The highest values, up to 4, are located in catchments within the same abovementioned region, indicating the strong dependence of those catchments on precipitation events to generate its streamflow. Moreover, we can note that most Brazilian catchments are inelastic to changes in precipitation. This fact can be explained by the high values of the baseflow index, which maintains the streamflow through the year. Fig. 7g, Fig. 7h, and Fig. 7i show the results related to the low flows of CABra catchments.

In general, Brazilian catchments present a low flow (5th quantile) lower than 1 mm day^{-1}, up to 50 days through the year, with a mean duration of up to 25 following days. Despite the mean values, we can note high values (up to 3 mm day^{-1}) in the Amazon. Additionally, higher values of frequency and duration of low flows can be found in the north-eastern portion of Brazil, with mean frequency reaching 150 days and mean duration reaching 100 days for some catchments. In turn, Fig. 7j, Fig. 7k, and Fig. 7l show the information about high flows in CABra catchments. Most CABra catchments present high flows up to 10 mm day^{-1}, but in some catchments, this value can reach 30 mm day^{-1}. As seen in the low flow analyses, the mean frequency of high flow does not exceed 50 days per year for most of the catchments. The frequency, instead, lasts for lower time, up to 10 days.

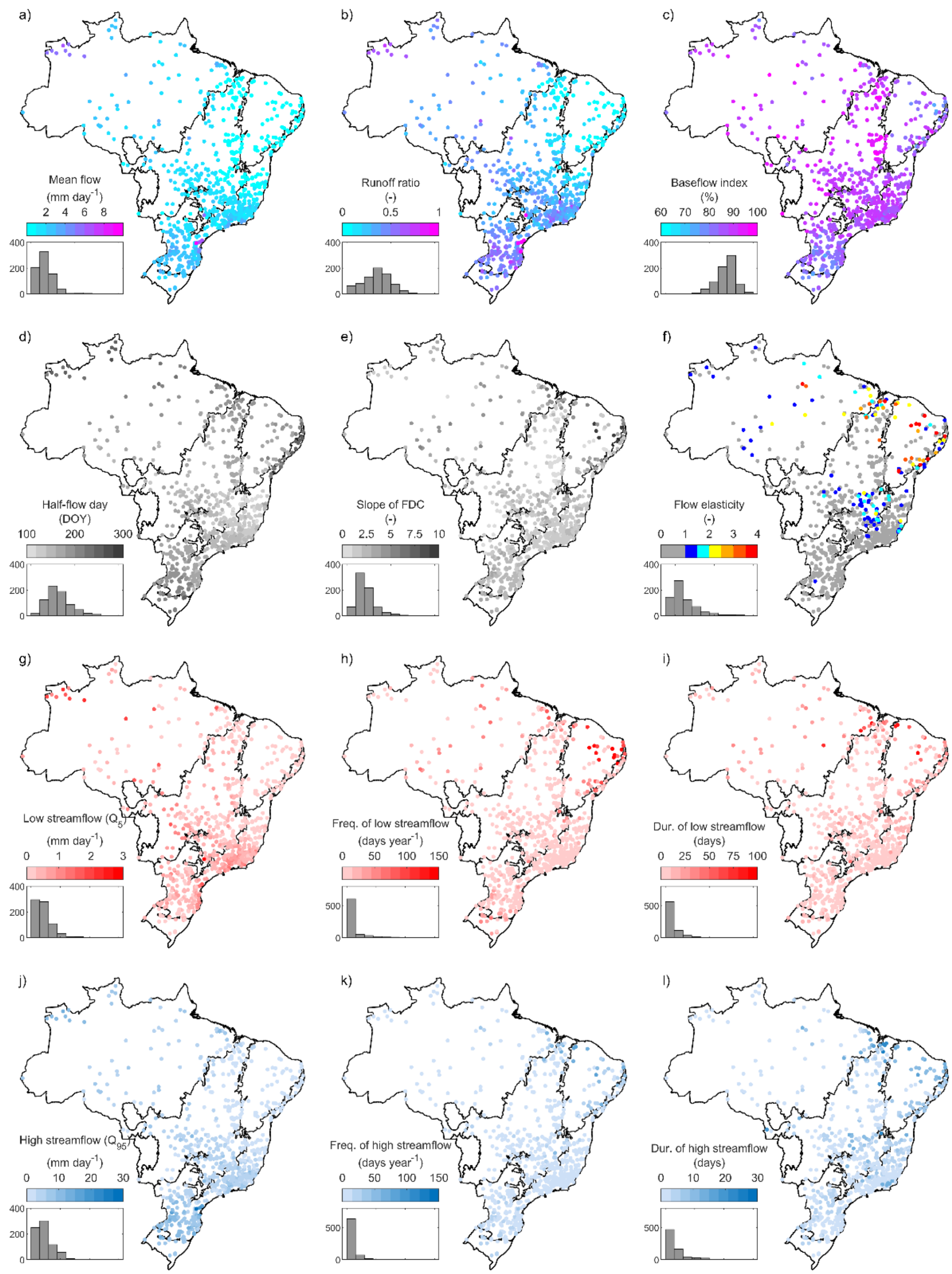

Figure 7: Spatial distribution of the hydrological signatures of the CABra catchments. a. Mean daily streamflow, in mm day^{-1}; b. Runoff ratio, dimensionless; c. Baseflow index, dimensionless; d. half-flow day, in day of the year; e. The slope of the flow duration curve, dimensionless; f. Elasticity of daily streamflow, dimensionless; g. Low streamflow, in mm day^{-1}; h. Frequency of low streamflow events, in days year^{-1}; i. Duration of low streamflow events, in days; j. High streamflow, in mm day^{-1}; k. Frequency of high streamflow events, in days year^{-1}; l. Duration of high streamflow events, in days.

2.4.3. Uncertainty and limitations

Uncertainties in the hydrologic signatures are mainly related to the daily streamflow data, which is, in turn, mainly related to the river discharge measurements and database maintenance by the ANA. Data collection and streamflow measurements are not the same in all catchments, varying from current meter to most advanced acoustic doppler profilers. The daily discharge of sections with well-established beds and a long enough series of measurements are estimated by rating curves, which are more susceptible to errors than direct measurements (Tomkins, 2014). Despite this, daily streamflow records are provided with a consistence level, which can be "raw", meaning that data was not quality checked, or "consistent", meaning that data was quality checked. The consistence level is provided along with each daily record in CABra dataset, allowing the user to identify the best and worst periods of streamflow measurements in each catchment. Although it is impossible to accurately measure the uncertainties (as much as eliminating them) in a large-sample dataset such as CABra dataset, it is important to indicate the possible sources, since they are widespread in any hydrological modeling. This way we can indicate the best periods for calibration/validation, increasing the reliability of the dataset and its application.

2.5. Groundwater

2.5.1. Methodology

The CABra dataset presents eight attributes regarding the groundwater at the catchments (Table 5). They are related to the water table (water table depth and height above the nearest drainage) and to the aquifer where the catchment is within (aquifer name and rock type). The first attribute is the area-averaged water table depth. This information was extracted from Fan et al. (2013), which is a global water table depth map generated using a climate-sea-terrain coupled model. The results were validated against observations and show the global patterns of shallow groundwater, making possible the understanding of how groundwater affects terrestrial ecosystems, such as the soil moisture and land hydrology, in a deficiency of rain (Fan et al., 2013; Lo et al., 2010).

The second attribute is the Height Above Nearest the Drainage (HAND), also related to the water table but is an indirect way to infer the water table depth. The HAND is a normalized drainage version of a digital elevation model, where the height is defined as the vertical distance from a hillslope (at the surface cell) to a respective "outlet-to-the-drainage" cell, as defined by Nobre et al. (2011). Considering the local gravitational potential, the HAND model shows robust correlations between soil water conditions and its values. Additionally, the authors created three classes to easily infer about the water table depth (if at the surface, shallow or deep) only using a digital elevation model, which is commonly a piece of difficult and scarce information on a large scale. We also present the aquifer in which the catchment is within (most of the area) and the most common type of rock of the aquifer. This information was provided by the ANA database, and it is important to the knowledge of the aquifer geology and its implication to groundwater storage and recharge. We also have included data from experimental wells on the CABra catchments, when available. The data was provided by the Integrated Groundwater Monitoring Network (RIMAS) from the Geological Survey of Brazil (CPRM) and includes the location of each well and its static and dynamic levels.

Table 5: Groundwater attributes of the CABra catchments.

Type	Attribute	Long name	Unit
Water table	catch_wtd	Water table depth	m
Height above nearest drainage	catch_hand	Height above the nearest drainage	m
	hand_class	Class of the height above the nearest drainage	-
Aquifers	aquif_name	Aquifer name	-
	aquif_type	Aquifer rock type	-
Wells	well_number	Number of experimental wells	-
	well_static	Static level of water table depth	m
	well_dynamic	Dynamic level of water table depth	m

- Means dimensionless

2.5.2. Results and discussion

Our analyses showed a close relationship between the water table depth from Fan et al. (2013) and the HAND. In the northern portion of Brazil, especially in the Amazon, we can find shallow water table depths, while in the south-eastern, especially in the Atlantic Forest, we noted the deepest values for the water table depths (see Fig. 8a and Fig. 8b). This could be related to the altitudes of each catchment since the HAND is a product derived from a digital elevation model. As a catchment lies at a high elevation, the water table depth is deeper than the other catchments in low elevations. This is particularly noted in the coastal area of the Atlantic Forest, which presents high altitudes and at the same time, is close to the sea level. Values of water table depth and HAND are also in accordance to the experimental wells for catchments where this analysis were possible to carry. Despite this, the low density of experimental wells shows the lack of field data abour groundwater in Brazil.

Figure 8c shows that most of the CABra catchments are dominated by fractured and porous rocks. The fractured rocks store the water in fractures, creating large pockets of water. The porous rocks store water in the soil pores (especially in sandy soils originated by sedimentary rocks), and it is common to find large amounts of water in them. The two of the world's largest aquifers are in Brazil and are porous, the Guarani Aquifer in the Cerrado biome, and the Alter do Chão Aquifer in the Amazon biome. The third aquifer type found in CABra catchments is the karstic one. This can be found in the São Francisco River Basin.

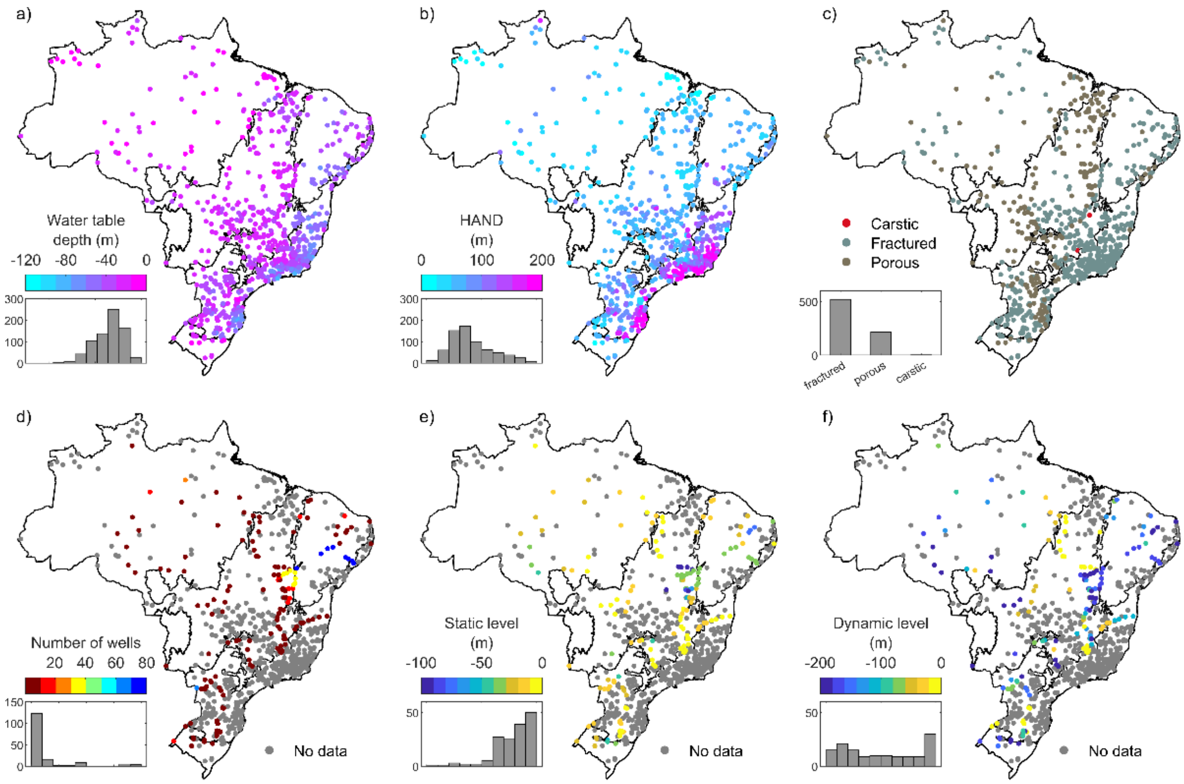

Figure 8: Spatial distribution of the groundwater attributes of the CABra catchments. a. Water table depth, in m; b. Height Above Nearest Drainage, in m; c. Type of aquifer bedrock. d. Number of experimental wells; e. Static level, in m; f. Dynamic level, in m.

2.5.3. Uncertainty and limitations

Due to the lack of a robust monitoring network for groundwater resources in Brazil, most of the data Fan et al. (2013) for covering the Brazilian territory is based on in situ observations of water table depth and groundwater model forced by climate, terrain and sea levels, only up the 2013 year. For South America, there were 34,508 observation sites, most of them in Brazil, but they are concentrated in the Atlantic coastal area, with few observations in most of the Brazilian area. Moreover, the global dataset provided by Fan et al. (2013) neglects local perched aquifers, groundwater pumping, irrigation, drainage, and any other complexity of human interaction. The HAND product, in turn, is not based on observations, but it is a simplified way to correlate the water table depth with terrain elevation, and it is mainly subject to errors in the digital elevation model used as input, especially in flat areas, where there are uncertainties during the flow direction determination (Nobre et al., 2011). The information of aquifers presented in the CABra dataset, provided by the Brazilian Water Agency, was developed with a previous and rigorous consistency analysis of geological and hydrogeological studies in Brazil, followed by the classification in three main classes, as fractured, carstic or porous. The mapping of aquifers systems was based on the analysis of consistency, adequacy, and reclassification of existing geological and hydrogeological information. The reclassification of polygons from geological units and their groupings, according to their hydrogeological characteristics. Data sources with different scales, which might an uncertainty source for the aquifers data. The sources and spatial map of the aquifers is not available through CABra dataset, where we only present the most common aquifer in each catchment.

2.6. Soil

2.6.1. Methodology

The CABra dataset has eight attributes related to the soil type, properties, and texture (Table 6). The soil type of the catchment presented here is the most common type for each catchment (bigger percentage of the different types) derived from the Brazilian soil map developed by the Brazilian Agricultural Research Corporation (EMBRAPA, in Portuguese) (Santos et al., 2011). To meet with the international standards for soil classification, we converted the classes to the widely used World Reference Base (WRB) (FAO, 2014). Due to the high importance of the knowledge of the soil depth, density, texture, and organic matter to the understanding of soil-water dynamics and root grow (Dexter, 2004; Saxton et al., 1986; Saxton and Rawls, 2006; Shirazi and Boersma, 1984), we also present the mean areal attributes for them. These fields were taken from the SoilGrids250m, a global high-resolution gridded soil information based on field measurements, data assimilation, and machine learning. This is the most detailed and accurate global soil product and is crucial for the development of large-scale studies in many fields (ecology, climate, hydrology). However, despite all the improvements brought by SoilGrids250m, the data still have limitations, and one of the biggest is the high uncertainty levels for some of its products, such as the depth to bedrock and coarse fragments. Besides, we also employed the United States Department of Agriculture (USDA) soil texture classification, which is a widely used method for soil definition based on the mechanical limits of soil particles. Moreover, previous studies showed that the USDA soil texture classification can potentially reflect other soil parameters and characteristics (Groenendyk et al., 2015; Twarakavi et al., 2010), making it a powerful tool with a low input requirement.

Table 6: Soil attributes of the CABra catchments.

Type	Attribute	Long name	Unit
Soil type	soil_type	Most common soil type	-
Soil depth	soil_depth	Soil depth to bedrock	m
Soil density	soil_bulkdensity	Soil bulk density	g cm^{-3}
Soil texture	soil_sand	Sand portion on soil first layer	%
	soil_silt	Silt portion on soil first layer	%
	soil_clay	Clay portion on soil first layer	%
	soil_textclass	Soil texture classification (USDA)	-
Organic content	soil_carbon	Organic carbon content on soil first layer	‰

- Means dimensionless

2.6.2. Results and discussion

The catchments presented 12 main soil classes, with the Ferrasols, Acrisols, and Nitisols being the most common soil types in more than 90% of the CABra catchments (Fig. 9a). The Ferrasols were the dominant soil type in approximately 75% of the catchments, typical of equatorial and tropical regions, which have an advanced stage of weathering of their constitutive material, being normally deep (>1m), well-drained, and acidic soils (high pH levels can occur in areas with a strong dry season, such as observed in the Caatinga biome). Acrisols are formed mainly by minerals, with an evident increase in the clay content from the surface to horizon B, with variable depth and drainage, but always with high acidity. The third most common soil type is the Nitisols, which have a clay texture, with a well-developed B horizon structure, and are usually deep and well-drained with moderate acidity (EMBRAPA, 2018).

We noted that most of the catchments present soil texture dominated by sand and clay (Fig. 9c, Fig. 9d, and Fig. 9e). South-eastern, northern, and central regions of Brazil are dominated by sandy clay loam soils, while the southern portion is dominated by clay, which can reach up to 80%, making this region one of the most productive in terms of agriculture in Brazil. By the employment of the USDA texture triangle, we found 6 classes: clay, clay loam, loam, sandy clay, sandy clay loam, and sandy loam (see Fig. 9b). The soils presenting a clay and clay loam texture are in the southern portion, especially where the Nitisols occur, which is also the region with a significant portion of the Brazilian agricultural production.

Most of the catchments present a mix of texture, the sandy clay loam, which covers from the south through the central to the northern regions of Brazil. There is a spatial correlation between the soil organic carbon, bulk density, and the distance to the bedrock, as we can see in Fig. 9f, Fig. 9g, and Fig. 9h. In the southern and south-eastern portions, especially in the Atlantic Forest biome, there is a combination of high soil organic carbon, low bulk density, and low distance to the bedrock. These characteristics, allied to the favorable climate, turned this region attractive to agriculture. On the other hand, other Brazilian regions present the opposite.

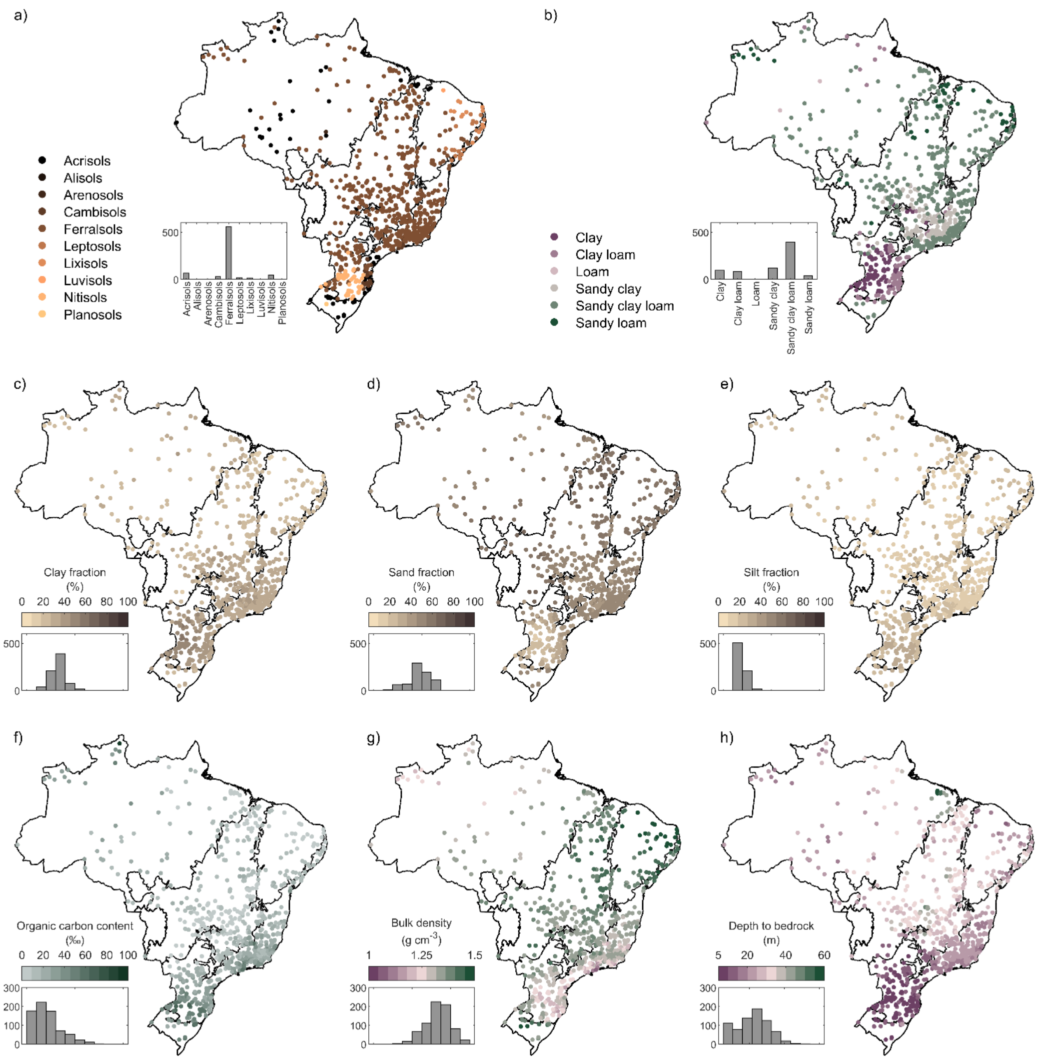

Figure 9: Spatial distribution of the soil attributes of the CABra catchments. a. The most common type of soil in the catchment; b. The class of texture based on USDA classification; c. The clay fraction of the soil, in percentage; d. The sand fraction of the soil, in percentage; e. The silt fraction of the soil, in percentage; f. The organic carbon content of the soil, in permille; g. The bulk density of the soil, in g cm^{-3}; h. The depth to soil bedrock, in m.

2.6.3. Uncertainty and limitations

The main limitation of the database used in CABra dataset as the source for soil attributes, the SoilGrids250 (Hengl et al., 2017), is related to the interpolation of a predicted data (through machine learning algorithms), which is based on soil profiles observed data. In this aspect, Brazil has a good starting point, with a dense and uniform distribution of in situ samples. However, authors state that, although most of the properties are unbiased, coarse fragments and depth to bedrocks present relatively high uncertainties, as well overestimations in low values of organic carbon content. Uncertainties are also related to the need for translation from the Brazilian classification system to the World Reference Base and USDA classification systems, where some information could be missed or misunderstood.

2.7. Geology

2.7.1. Methodology

The CABra dataset presents four attributes related to the geology of the catchments (Table 7), being the predominant lithology class, the porosity, the saturated permeability, and the saturated hydraulic conductivity. The lithology class is derived from the Global Lithologic Map (GLiM) (Hartmann and Moosdorf, 2012). The GLiM is a high-resolution global dataset that describes the geochemical, mineralogical, and physical properties of the rocks in 16 main lithological classes. Moreover, GLiM allows us to better understand the geology of smaller areas, such as our CABra catchments. Also, we are using a GLiM-derivate product of porosity and permeability named GLobal HYdrogeology MaPS (GLHYMPS), developed by Gleeson et al. (2014). The GLHYMPS is the first large-scale high-resolution mapping of porosity and permeability and fills a lack of robust and spatially distributed subsurface geology map.

The porosity is the void spaces in a material (soil in our case) controls how much fluid (water) can be stored in this material, or in the soil subsurface. The movement of the stored water in the soil is controlled by the permeability, which is the capacity of a porous material (again, soil) to transmit fluids. Both parameters are fundamental to the knowledge of fluid rate and its impacts on Earth’s subsurface. When using this kind of high-resolution data for large-scale studies, we can improve our understanding of the dynamics between groundwater and land surface. Considering the saturated hydraulic conductivity as one of the most important physical properties on the quantitative and qualitative assessment of the water movement in the soil, we presented its values in the CABra dataset. Following the assumption that the hydraulic conductivity is separable into the contributions of the porous matrix of the soil, and the density and viscosity of the fluid, we also estimated the saturated hydraulic conductivity of the CABra catchments using its relation to the permeability (Equation 4), as described in Grant (2005).

$$K = \frac{k\rho g}{\mu} \qquad (4)$$

where *K* is the saturated hydraulic conductivity, *k* is the saturated permeability, ρ is the density of the fluid, *g* is the gravitational constant (9.8 m s^{-2}), and μ is the viscosity of the fluid. In our study, we have considered the water as the fluid, so we have used ρ = 999.97 kg m^{-3}, and μ = 0.001 kg m^{-1} s^{-1}.

Table 7: Geology attributes of CABra catchments.

Type	**Attribute**	**Long name**	**Unit**
Lithology	catch_lith	Most common lithology class	-
Subsurface geology	sub_porosity	Porosity	-
	sat_permeability	Saturated permeability	m^2
	sat_hconduc	Saturated hydraulic conductivity	m s^{-1}

- Means dimensionless

2.7.2. Results and discussion

Related to the lithology class, the catchments present 10 different classes according to the GLiM dataset: siliciclastic sedimentary rocks, acid volcanic rocks, unconsolidated sediments, acid plutonic rocks, metamorphic rock, mixed sedimentary rocks, basic volcanic rocks, carbonate sedimentary rocks, intermediate volcanic rocks, and pyroclastic rocks (Fig. 10). We found that 35% of the catchments have the metamorphic rocks as the most common lithologic class, a result of continuous weathering on the original rock. These catchments are located especially in the southern portion of Brazil, in mountainous areas. Approximately 39% of CABra catchments are formed by sedimentary rocks, considering its subdivision in siliciclastic, unconsolidated, and mixed resulted from sediment deposition. They are mostly located in flat areas, such as in the Paraná River Basin and São Francisco River Basin, in the central and north-eastern portion of Brazil. 25% of catchments presents igneous rocks (plutonic and volcanic) as the most common lithology class, resulted from volcanic eruptions. These catchments are located mainly in the Atlantic Forest biome, although we can find some catchments in the Amazon.

In respect to the porosity, most CABra catchments presented values lower than 20%, with a mean value of 10%. Catchments in the Atlantic Forest presented the lowest values of the catchments set. Results regarding the saturated permeability and hydraulic conductivity reinforce the heterogeneity and random occurrence of these soil properties. As we can see in Fig. 10c and Fig. 10d, there is no well-defined spatial behavior for them. Saturated permeability ranges from -14 to -12 m² in log scale, with a mean of -13.4 m², while the saturated hydraulic conductivity presented a mean value of -6.4 m s^{-1} in log scale, vary between -10 to -4 m s^{-1} in log scale.

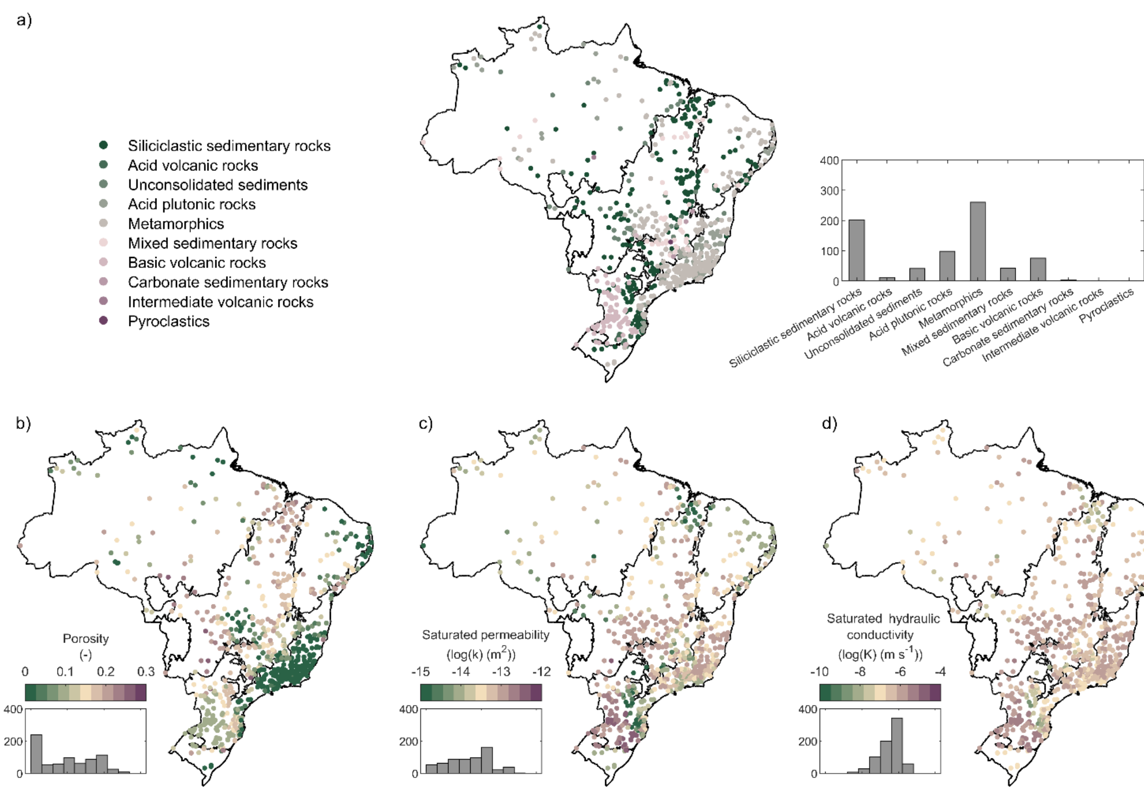

Figure 10: Spatial distribution of geology attributes of the CABra catchments. a. Most common lithology class in the catchment; b. Porosity, dimensionless; c. Saturated permeability, in m^2; d. Saturated hydraulic conductivity, in m s^{-1}.

2.7.3. Uncertainty and limitations

The geological map of the CABra dataset is derived from the GLiM dataset (Hartmann and Moosdorf, 2012), which is, in turn, the main source for the development of the hydrologeological map used in CABra dataset, the GLHYMPS (Gleeson et al., 2014). Authors state that the global lithological map is still subject to significant uncertainty in rock properties in some of its lithological classes, mainly because of the scale of the maps. About 14,6% of the map's area is covered by mixed sediments, explaining the large area subject to undistinguishable properties. In addition, the quality of literature used to identify lithology in rare locations may have introduced some uncertainty level on GLiM. As mentioned before, the GLiM map was employed as a basemap for GLHYMPS permeability product, implying that all uncertainty associated with GLiM might be propagated to it. Moreover, Gleeson et al. (2014) presents a uncertainty map of permeability, showing high standard deviation values for central portions of Brazil, especially in Tocantins-Araguaia catchments. Finally, authors also recommend a careful use of the dataset where unsaturated zone processes are dominant, since GLHYMPS only takes in account saturated permeability.

2.8. Land-cover

2.8.1. Methodology

The CABra dataset presents 15 attributes regarding the land-cover and land-use of the Brazilian catchments (Table 8). They are related to the area-averaged land-cover and land-use itself (dominant cover type, and the cover fractions of 9 main classes of use: bare soil, forest, grass, shrub, moss, crops, urban, snow, and water) and to the area-averaged intra-annual variability of the vegetation biomass, here represented by the Normalized Difference Vegetation Index. The land-cover and land-use map used in the CABra dataset is the Copernicus Global Land Cover, which has 100-m spatial resolution, is a result of a classification of the PROBA-V satellite observations of the year 2015 and follows the UN FAO Land Cover Classification System (Buchhorn et al., 2019) available at https://land.copernicus.eu/global/lcviewer.

As an indicator for the vegetation biomass of the land cover through the year, we are using the seasonal NDVI for each CABra catchment. The NDVI is widely used, easily accessible, and with high-temporal availability, which can be useful for many purposes on hydrology, from an annual precipitation cycle indicator to an input for soil erosion assessments. We adopted a product derived from the Long Term Statistics (LTS) based on the Normalized Difference Vegetation Index (NDVI) from the Copernicus Global Land services. This dataset is an NDVI mean for each month of the year during the 1999-2017 period, obtained from the SPOT-VGT and PROBA-V sensors in a 1-km spatial resolution, available at https://land.copernicus.eu/global/products/ndvi. The NDVI is obtained by calculating the spectral reflectance difference between red and near-infrared bands of the satellite image (Tucker, 1979) (Equation 5) and ranges from -1 to +1, with the highest values attributed to areas with greater vegetation cover.

$$\mathrm{NDVI} = \left(\frac{\mathrm{NIR} - \mathrm{RED}}{\mathrm{NIR} + \mathrm{RED}}\right) \quad (5)$$

where NIR is the surface spectral reflectance in the near-infrared band and RED is the surface spectral reflectance in the red band.

Table 8: Land-cover attributes of CABra catchments.

Type	**Attribute**	**Long name**	**Unit**
Land-cover and land-use	cover_main	Dominant cover type	-
	cover_bare	Bare soil fraction of cover	%
	cover_forest	Forest fraction of cover	%
	cover_grass	Grass fraction of cover	%
	cover_shrub	Shrub fraction of cover	%
	cover_moss	Moss fraction of cover	%
	cover_crops	Crops fraction of cover	%
	cover_urban	Urban fraction of cover	%
	cover_snow	Snow fraction of cover	%
	cover_waterp	Water fraction of cover (permanent)	%
	cover_waters	Water fraction of cover (seasonal)	%
Vegetation	ndvi_djf	DJF normalized difference vegetation index	-
	ndvi_mam	MAM normalized difference vegetation index	-
	ndvi_jja	JJA normalized difference vegetation index	-
	ndvi_son	SON normalized difference vegetation index	-

- Means dimensionless

2.8.2. Results and discussion

We observed that most of the Brazilian catchments are covered by forest and grassland (Fig. 11). The shrub is the dominant cover for most of Caatinga catchments, while the grass is the dominant one in the Cerrado (tropical savannah). The forest cover is dominant especially in the Amazon and Atlantic Forest, as these two biomes are known by tropical forest occurrence, but even though the forest cover is not the most common for all the CABra catchments, ~85% of them present at least 20% of it (Fig. 11b). The grass cover fraction presented values up to 40% of the area for most of the catchments but reached 60% in some cases (Fig. 11c). The highest values were found in the Cerrado and Atlantic Forest biomes, in central and south-eastern portions of Brazil.

Large areas of natural cover were converted to agricultural lands (including crops and pasture) in past years (Gibbs et al., 2010, 2014), and satellite sensors and classifiers algorithms cannot separate natural grassland and pasture/managed grasslands, as described in the PROBA-V documentation. Figure 11d gives us a better idea of this. Probably the fraction of the shrub cover of the Cerrado is the natural cover remaining for this biome since this is the expected type of vegetation. As seen in Fig. 11e, a few numbers of catchments present the crops as the dominant cover type, mostly in the central and southern region, but we can also see the great fraction of crop cover in the MATOPIBA region, one of the largest agriculture frontiers in Brazil (Gibbs et al., 2014; Pires et al., 2016; Spera et al., 2016). Figure 11f shows that there are only a few cases of urban catchments, within or close to major Brazilian cities that present this type of cover, showing that the CABra dataset is mainly composed of either natural or minimally (hydrologically) modified catchments.

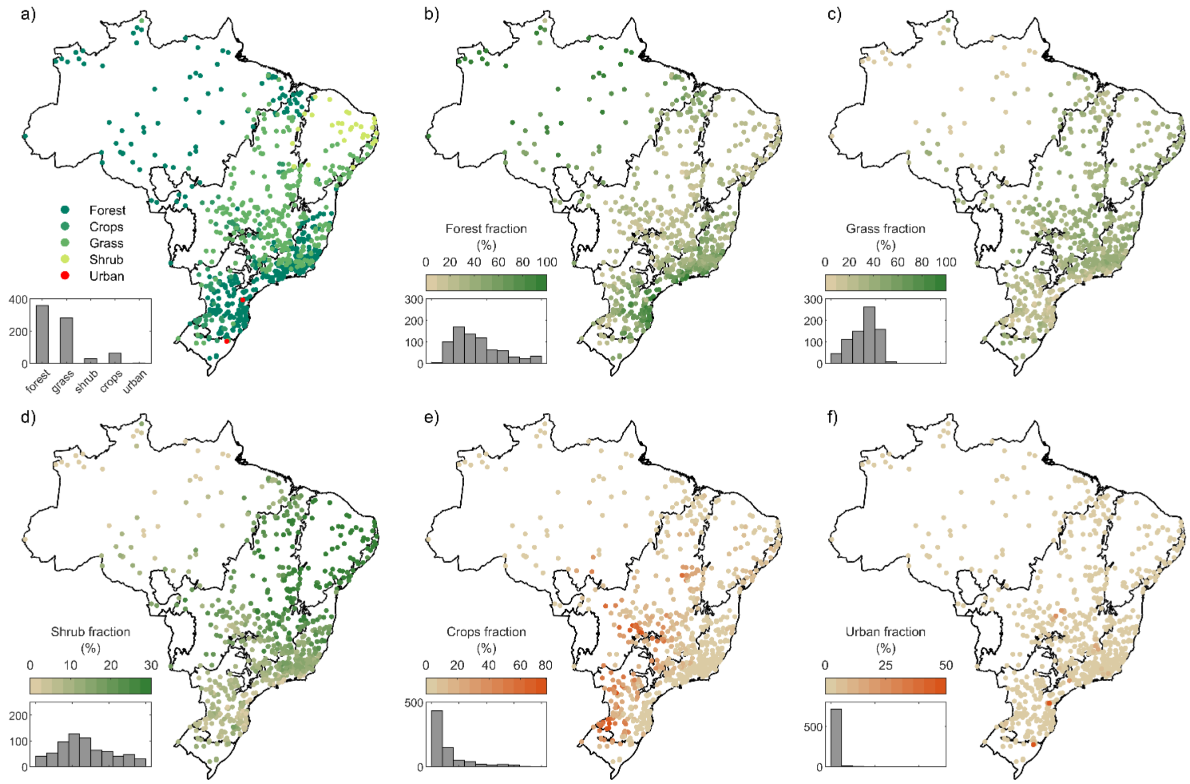

Figure 11: Spatial distribution of the land-cover and land-use attributes of the CABra catchments. a. The most common land-cover type in the catchment; b. Forest fraction of land-cover, in percentage; c. Grass fraction of land-cover, in percentage; d. Shrub fraction of land-cover, in percentage; e. Crops fraction of land-cover, in percentage; f. Urban fraction of land-cover, in percentage.

The seasonal variability of the NDVI can be seen in Fig. 12. Although the mean seasonal values for the entire country are similar (0.65 for DJF, 0.69 for MAM, 0.64 for JJA, and 0.56 for SON), the spatial variability of the NDVI values are noticeable. There is a clear relationship with the annual cycle of precipitation, and that is why it is so important to consider the seasons to analyze the NDVI. Higher values of NDVI occur in accordance with the seasonal cycle of precipitation in all the biomes, especially in DJF and MAM months. Even in the Amazon, we can see a considerable decrease in the NDVI values for the catchments in the dry seasons (JJA and SON) as well as the other biomes and regions of Brazil. NDVI reaches the lowest values at the end of the hydrological year and then starts to increase the values only at the beginning of the rainy season, i.e., DJF season. Intermediate values in the central portion of Brazil are much likely to be linked to agricultural production, leading the values to be lower than the natural cover.

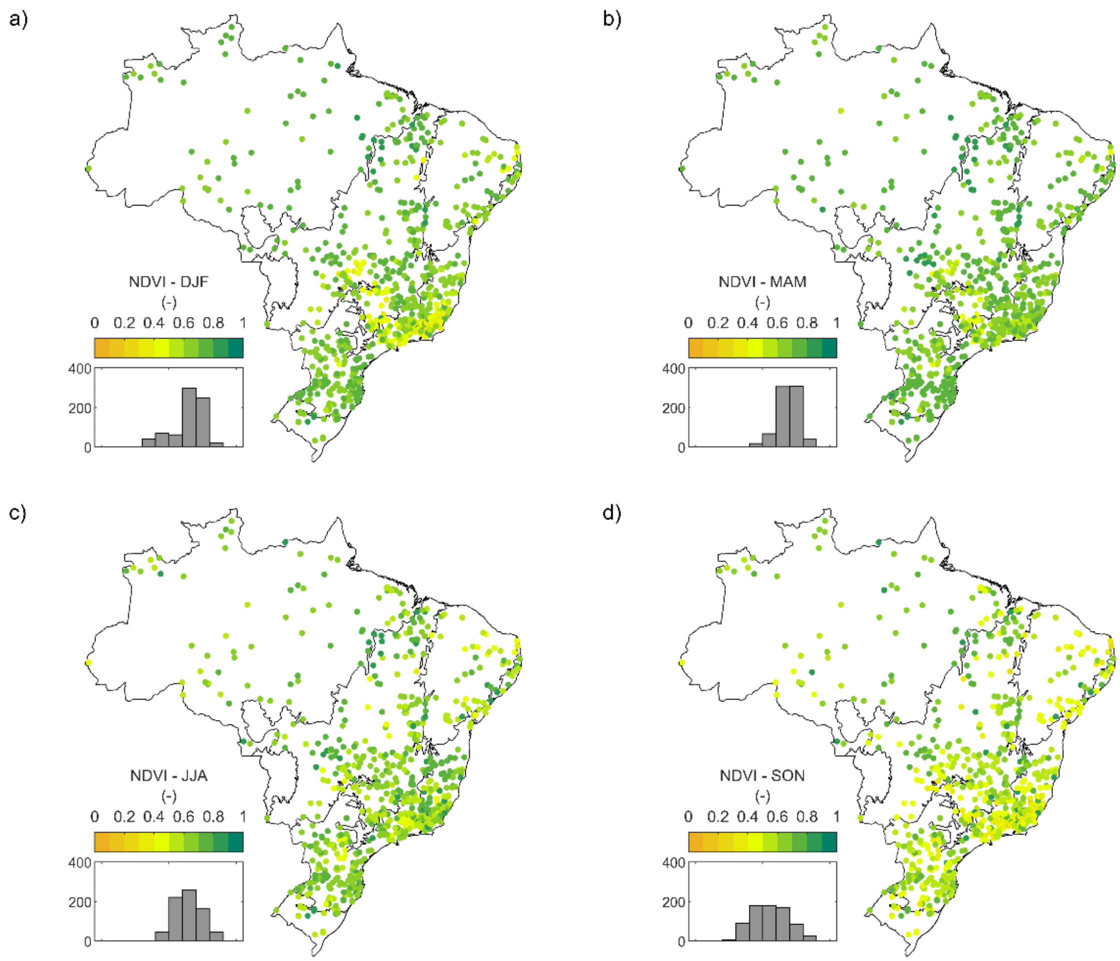

Figure 12: Spatial distribution of the seasonal NDVI of the CABra catchments. a. NDVI in summer season (DJF); b. NDVI in autumn season (MAM); c. NDVI in the winter season (JJA); d. NDVI in the spring season (SON).

2.8.3. Uncertainty and limitations

Although the CABra dataset presents one of the most high-accuracy spatial resolutions on a global scale, the data is related to the 2015 year, which is not within the 1980-2010 period adopted in the hydrological analyses. As authors from the Copernicus Global Land Cover (Buchhorn et al., 2019) state, the global land-cover data should be used with confidence but with careful and critical analysis by the users, due to the land changes commissions and omissions. Uncertainty analyses conducted in three aggregated classes (forest, crops, and natural vegetation) showed high accuracy in all regions of the world when compared with more than 200,000 samples points. Even though, there is some level of overestimation in the forest class, leading to a careful assessment of land-cover in Amazon and Atlantic Forest catchments. At the same time, due to the 100 m spatial resolution, small villages and highly fragmented landscapes might be indistinguishable and/or mixed with different classes.

NDVI dataset, also provided by Copernicus Global Land Cover, should be used as a qualitative indication of the biomass in the catchment, due to it relatively low spatial resolution (300 m). There are also uncertainties related to the radiometric calibration of the images, anisotropic surfaces, aside from the fact that the products did not consider adjacency effects and slope correction.

2.9. Hydrologic disturbance

2.9.1. Methodology

The CABra dataset presents 10 attributes related to the hydrologic disturbances on catchments water fluxes (Table 9). Anthropic changes in water flux patterns, which happens outside the range of natural flow and climate extremes, can directly impact the water availability and quality, stream channel geometry and sedimentation, and the equilibrium of ecosystems (Boulton et al., 1992; Coleman et al., 2011; Whited et al., 2007). Natural conditions of catchments are constantly modified by human interactions such as land-cover and land-use changes, flow regulation, water abstractions, soil impermeabilization, and many others, which can drastically alter the way hydrologic fluxes in the catchments respond. Then, our goal was to create a simple index, with easily accessible inputs, that is capable to measure how much disturbed a catchment is in relation to its hydrology. Since the beginning of CABra development, it was known that most of the catchments were minimally urbanized, but with some of them with changes in the original land-cover (conversion of natural vegetation to cropland/pasture). Some studies conducted in Brazil found that, besides the fact of the interference by the conversion of natural vegetation to pasture, this led to minimal changes in the surface hydrology of the catchment, being more relevant to groundwater recharge and soil chemistry (Bacellar, 2005; Lanza, 2015; Nepstad et al., 1994; Salemi et al., 2012). Additionally, it has been seen that the human-induced impact of the reservoirs can be more relevant than the natural ones, and can significantly alter natural hydrological processes (Zhao et al., 2016), leading to an increase/decrease of streamflow and hydrological droughts characteristics

(Wanders and Wada, 2015; Ye et al., 2003; Zhang et al., 2015). Moreover, Zhang et al. (2015) found that hydrologic vulnerability is also directly related to human water abstractions, but this can be compensated by streamflow regulation of the reservoirs. This led us to an integrated analysis of the reservoir regulation and human water abstract to reach the optimal balance on our index.

Based on the abovementioned, we have decided to use weighted information about the land cover, reservoirs, and water demand of each catchment. We considered the reservoir-based information with more impact: regulation capacity with 40%, number of reservoirs and its percentage of catchment area with 5% each. The second most impacting factor of the index is the non-natural land-cover in the catchment, which can lead to modify hydrological surface and subsurface processes, with 40% of the weights. Finally, the water abstraction of the catchment was pondered with 10%.

In the development of this index, we have considered the fraction of urban cover in each catchment, the distance to the nearest urban area of each catchment (considering any pixel of urban area), the number of reservoirs in each catchment (ANA, 2020b), the total volume of reservoirs in each catchment (ANA, 2020b), and its flow regulation capacity, the fraction of reservoir area of each catchment area (ANA, 2020b), and the annual water demand (ANA, 2019b). The equation related to the hydrologic disturbance index can be found in the following Equation 6:

$$HD_{index} = 0.4([U_C . U_D] + CR_C) + 0.05R_N + 0.05R_{\%A} + 0.4R_R + 0.1W_D \qquad 6$$

where HD_{index} is the hydrologic disturbance index, dimensionless; U_C is the normalized fraction of urban cover; U_D is the normalized distance to the nearest urban area; CR_C is the normalized fraction of crops cover; R_N is the normalized number of reservoirs; $R_{\%A}$ is the normalized percentage of catchment's area covered by reservoirs; R_R is the normalized reservoirs' regulation capacity of catchment's mean annual flow; and W_D is the normalized catchment's annual water demand.

Table 9: Hydrologic disturbance attributes of CABra catchments.

Type	Attribute	Long name	Unit
Reservoirs	res_number	Number of catchment's reservoirs	-
	res_area	Total area of catchment's reservoirs	km²
	res_area_%	Catchment's area percentage covered by reservoirs	%
	res_volume	Total volume of catchment's reservoirs	hm³
	res_regulation	Reservoir's regulation capacity of the mean annual flow	-
Water demand	water_demand	Water demand in the catchment	mm $year^{-1}$
Land-cover	cover_urban	Urban fraction of cover	%
	cover_crops	Crops fraction of cover	%
	dist_urban	Distance from gauge to nearest urban cover	km
Hydrologic disturbance	hdisturb_index	Index of hydrologic disturbance in the catchment	-

- Means dimensionless

The result is the hydrologic disturbance index (HDI), which will easily provide for CABra users the degree of human interactions that can modify water fluxes in each catchment. Additionally, we also applied a random forest algorithm for a regression analysis to show if and how the hydrological signatures are captured by the HDI.

2.9.2. Results and discussion

The results of the spatial distribution of the hydrological disturbance index and its components are shown in Fig. 13. Most CABra catchments are close to an urban cover (it can be a large city or a small village), with up to 10 km. However, we also could find catchments with up to 100 km of distance to the urban cover. As seen in Fig. 13b and Fig. 13c, most CABra catchments present a fraction of urban cover up to 10%, with high values close to large cities, and a fraction of crops cover up to 40%, with the highest values in central and southern portions. As these factors present a high weight on the hydrological disturbance index, they are a good clue of the most disturbed catchments.

Results from the reservoirs in CABra catchments are shown in Fig. 13d, Fig. 13e, Fig.

13f, and Fig. 13g. The number of reservoirs in the catchment ranges from zero to 48,404. Even though we found the largest number of reservoirs in a large catchment, this relationship is not linear. There are some catchments, especially in the São Francisco River Basin, which presents an extremely high number of reservoirs due to the low amounts of annual precipitation and intensive drought in the region. Moreover, catchments in the São Francisco River Basin presents the highest values of the total volume of reservoirs. These reservoirs are used for many anthropogenic purposes, such as hydroelectric power plants, irrigation, drinking water supply, fish-farming, and recreation. These high values of the total volume of reservoirs, especially in the drier regions, could lead to a strong streamflow regulation, as seen in Fig. 13g. In most of the CABra catchments, reservoirs can regulate up to 25% of the annual flow, but there are some cases in the Caatinga biome where the regulation capacity reaches up to ten times the annual flow, making these catchments susceptible to non-natural events.

The water demand on CABra catchments ranges from zero (in Amazon) to 171 mm $year^{-1}$ (in Caatinga) and it is related to drinking water supply and irrigation of agricultural areas (Fig. 13h). The integrated analysis of the above-mentioned attributes is shown in Fig. 13i, as the new hydrological disturbance index. Most of the CABra catchments present an index value of up to 0.2, indicating a low anthropic interference on water fluxes. Higher values, above 0.4, indicate catchments with some significant interference on water fluxes, which may be related to one or more terms of the equation. High values of the hydrological disturbance index in the central and southern portion of Brazil may be related to agriculture development, while in the south-eastern part, they may be related to urbanization, and in the north-eastern part, they may be related to the presence of numerous voluminous reservoirs. As expected, in the Amazon and mountainous areas of Atlantic Forest, low values were found. The creation of the hydrological disturbance index can be especially useful for the users of the CABra dataset, allowing them to quickly view the general state of the anthropogenic interferences on water fluxes, which is an important consideration in a wide range of studies.

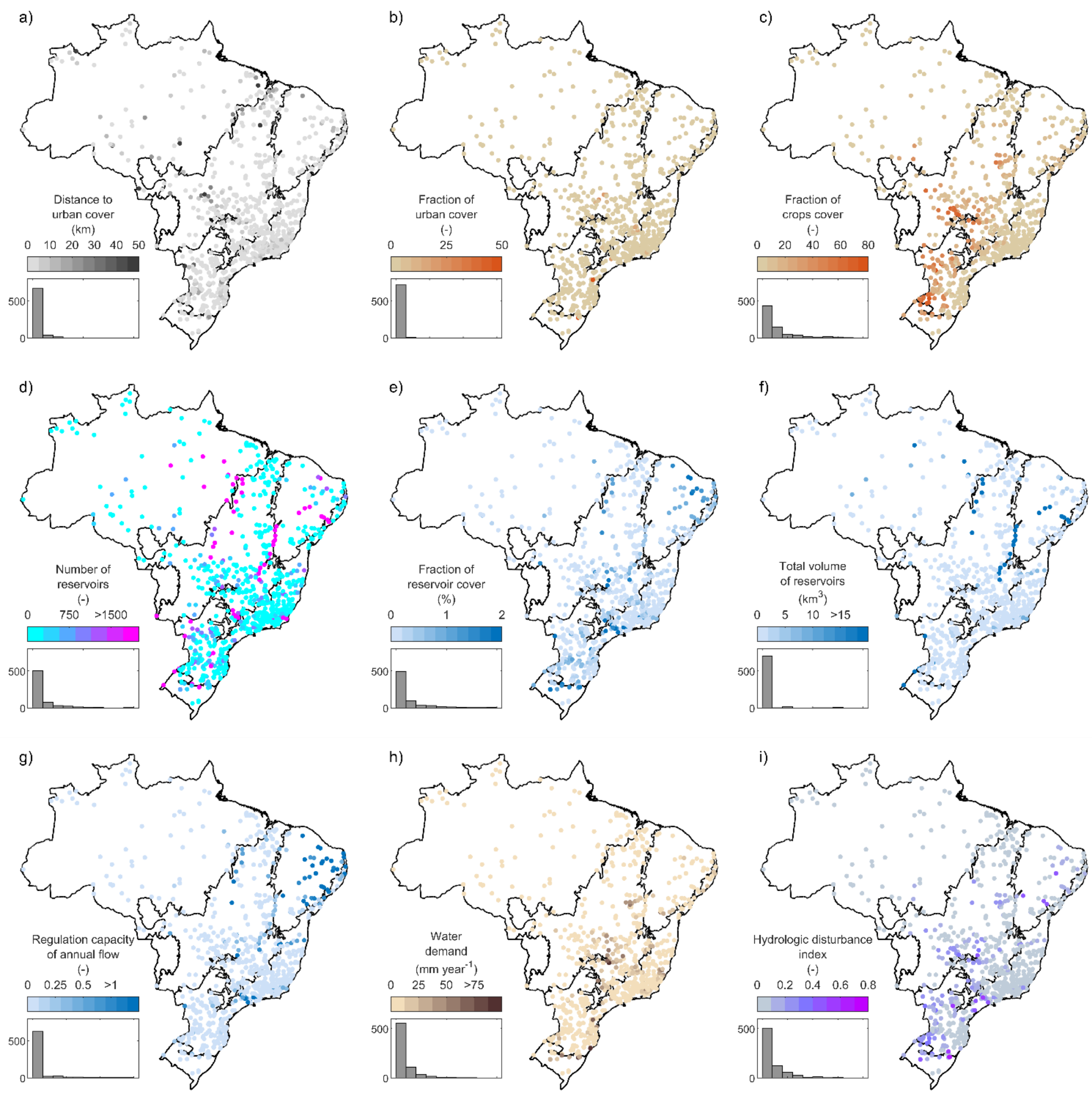

Figure 13: Spatial distribution of the hydrologic disturbance attributes of CABra catchments. a. Distance from urban cover to the streamflow gauge, in km; b. Urban fraction of land-cover, in percentage; c. Crops fraction of land-cover, in percentage; d. The number of reservoirs in the catchment; e. Reservoir fraction of land-cover, in percentage; f. The total volume of the reservoirs in the catchment, in km³; g. The capacity of the reservoirs in the catchment to regulate the mean annual streamflow, dimensionless; h. Multi-purpose water demand in the catchment, in mm year^{-1}; i. Hydrologic disturbance index (HDI) of the catchment, dimensionless. The HDI is a weighted relationship between all the anthropogenic factors of the catchments.

The random forest regressor algorithm (Figure 14) showed us the most relevant hydrological signatures captured by the Hydrologic Disturbance Index. About 25% of the variance of the HDI is explained by the Half-flow day and the Streamflow Elasticity, which are two signatures extremely sensitive to streamflow regulation and to the generation of runoff in the catchment. Our results show us that the index is capable to capture what it was intended to: catchments with higher values presents a large number or high regulation capacity of reservoirs, or a great percentage of non-natural areas. Medium values present some level of non-natural areas (pasture or crops), but there is not a high hydrological disturbance. Finally, lower values of HDI indicate minimally human-impacted catchments.

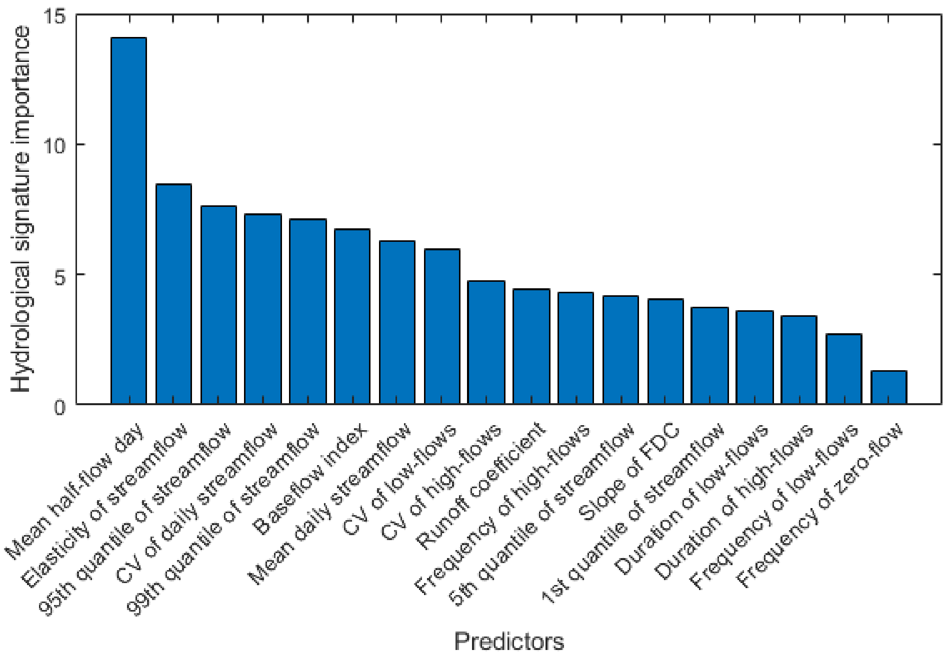

Figure 14: Hydrological signatures as predictors of the Hydrological Disturbance Index. The random forest regressor algorithm assess how much each signature increase the error of a HDI prediction when randomly sorted. The higher the deviation caused by a predictor, the higher is the influence of the hydrological signature on the HDI.

2.9.3. Uncertainty and limitations

Uncertainties in hydrological disturbance are mainly related to the components of the index. As mentioned before, there is a limitation of use in the land-cover maps for small villages, urban areas, fragmented areas, and transitional areas of croplands, due to the spatial resolution of the land-cover maps. Because of this, small areas of urban fraction (U_C), and consequently the distance to the urban area (U_D), and crops area (CR_C), might be undetected and this fraction of the index – representing 40% – not considered or underestimated. Another 50% of the HDI is derived from reservoir data, from the ANA database. Although the reservoirs' data have been extensively improved through the years, there are still uncertainties related to their many sources. Different sources do not use the same satellite products or methodology to identify and catalog the reservoirs. Additionally, latest inclusions of reservoirs were automatically made and there was not a quality check of these data. Due to the crucial importance of reservoirs to the HDI, unrealistic numbers, areas and volumes of reservoirs can lead to unrealistic values of the index. The last component considered here is the water demand (W_D), is a area-averaged estimation, which accounts to both consumptive and non-consumptive water abstractions, possible leading to higher values than real abstractions. Even representing, 10% of HDI composition, it should be taken into account in post-processing.

3. Comparison with CAMELS-BR and broader implications for hydrological studies

The CABra and the CAMELS-BR (Chagas et al., 2020) contain both large samples of hydroclimatic, landscape, and other attributes for Brazilian catchments. Their striking similarities in concept and goals highlight nothing but the urgent need for the creation of such a database for Brazilian catchments. However, it is important to notice that multiple differences between both datasets exist, as we will discuss below.

The first main difference between CABra and CAMELS-BR is related to the catchment delineation procedures adopted. CAMELS-BR uses the basin masks from the GSIM (Do et al., 2018) product, where a 500-m digital elevation model was used for the delineation of catchment boundaries and extraction of topographic indices. GSIM has a quality filter allowing for up to 50% of error in the catchment area when compared with ANA’s value, as described in Do et al. (2018). As previously explained, the CABra catchment boundaries (delineated using streamflow gauge location from ANA), use a high-definition (90-m) elevation product. We have manually inspected each of the 735 catchments to minimize further errors, correcting the geographic position of the outlet to coincide with the stream network, achieving a mean error of 2% against ANA’s areas. It is important to highlight that a suitable watershed delineation is of paramount importance for catchment hydrology studies because errors in these processes are further propagated for all computed attributes dependents on area and location. In addition, we provide the drainage network or CABra catchments.

Related to the daily streamflow data, in the CABra dataset we have retained catchments with less than 10% missing streamflow records over 30 hydrologic years (1980-2010) which resulted in the final selection of 735 catchments. On the other hand, CAMELS-BR contains 897 catchments with less than 5% missing data, while considering 20 hydrologic years (1990-2009). Additionally, CAMELS-BR also provides longer timeseries when available for the gauge. Our choice for a longer time series was predicated on the commonly adopted rationale which assumes 30 years as the basis for establishing long-term climatology as well as hydrologic indices (Huntingford et al., 2014; Tetzlaff et al., 2017), which we in turn believe will lead to better characterization of hydrological and climatological processes taking place. A correlation test between hydrological signatures of 607 overlapping catchments in CABra and CAMELS-BR datasets is shown in Figure 15. The signatures based only on daily streamflow values, such as daily mean streamflow (q_mean), 5^{th} and 95^{th} quantiles of daily streamflow (q_5 and q_95), are quite similar between CABra and CAMELS-BR, showing that both periods of analysis were capable to capture the streamflow patterns of the catchments. When comparing signatures related to frequency and duration of low and high streamflow events, we can note little variation but still good agreement between datasets. In this case, the distinct period for hydrological signatures calculation (1980-2010 in CABra, and 1990-2009 in CAMELS-BR) might be the cause of deviations. The slope of flow duration curve and runoff coefficient are in a very good agreement ($r^2 > 0.95$), demonstrating that both datasets are using precipitation products with good reliability. The streamflow elasticity and baseflow index have presented notable differences between CABra and CAMELS-BR. This might be due to the different components adopted in the equations of Woods (2009) and Ladson et al. (2013), which were implemented for elasticity and baseflow index calculations.

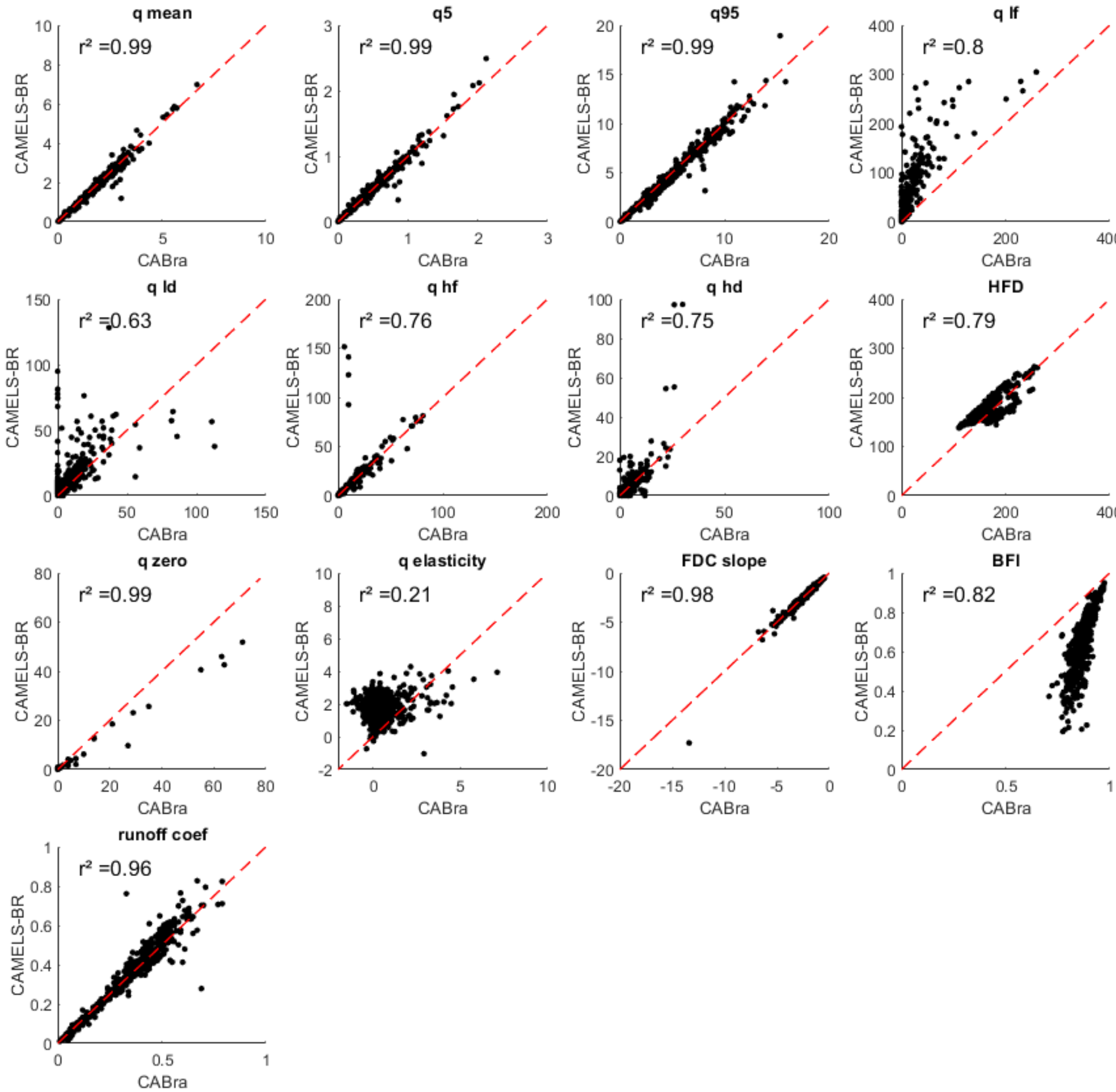

Figure 15: Scatter plots and correlation coefficients between hydrological signatures of CABra and CAMELS-BR catchments. There were 607 catchments and 13 hydrological signatures overlapped in both datasets.

Another important difference between both datasets is related to the choice of databases used for providing the daily meteorological time series and estimated the related indices. While CAMELS-BR uses three widely used gridded datasets (based on remote sensing/reanalysis/gauge blends of rainfall), i.e., the CHIRPS v2.0, CPC, and MSWEP v2.2, being the first one chosen for the climatic indices (because of its spatial resolution of 0.05ºx0.05º), the CABra uses the Xavier et al. (2016) dataset and the ERA5 reanalysis. The Xavier et al. (2016) dataset was produced based on observations from 3,625 rain gauges and

735 wheatear stations in the Brazilian territory and is extensively used as the ground-truth reference for the validation of precipitation products, including the CHIRPS, MSWEP, and the soil moisture satellite-corrected estimates (SM2RAIN, Brocca et al. (2014)) (Paredes-Trejo et al., 2018), the Global Precipitation Measurement (GPM, Hou et al. (2014)) (Gadelha et al., 2019), the Tropical Rainfall Measuring Mission (TRMM, Huffman et al. (2007)) (Melo et al., 2015). Other uses of this dataset include the evaluation of precipitation from downscaled-global circulation models (Almagro et al., 2020), as well as other meteorological variables used in regional studies (Battisti et al., 2019; Bender and Sentelhas, 2018; Monteiro et al., 2018), aside from being widely used for hydrological studies (Almagro et al., 2017; Avila-Diaz et al., 2020; Lima and AghaKouchak, 2017; Souza et al., 2016). The main limitation of Xavier's dataset is that it covers only Brazil.

Additional differences belonging to the meteorological time-series section are also worth noting. CAMELS-BR provides the model-based PET estimates extracted from the GLEAM product (Martens et al., 2017), while daily temperatures (maximum, minimum, and average) are the only PET-related variable provided in a daily time series format. The CABra dataset provides the computed PET following 3 widely used methods, along with all necessary variables for its computation, such as solar radiation, wind speed, temperature, and relative humidity. Our choice for the computation of PET instead of using model-based estimates should allow for more transparency and reproducibility of results obtained using our dataset. Also, the choice of providing a wider range of meteorological variables allows the user to estimate PET based on different methods while enhancing the reach of our dataset for studies that might benefit from additional meteorological variables.

While the soil and geology attributes of both CABra and CAMELS-BR are derived from the same data sources, (i.e., the SoilGrids250, the GLiM, and the GLHYMPS v2.0), CABra provides the following additional variables not available in CAMELS-BR: saturated permeability (saturated hydraulic conductivity for geology attribute), soil type, textural class, and soil bulk density – which can be used to estimate soil porosity. Regarding groundwater attributes, CABra contains rock type and name of the aquifer and water table depths from Fan et al., (2013) and the HAND estimates, while CAMELS-BR contains only the water table depth estimates from Fan et al., (2013).

In terms of land-cover attributes, CABra and CAMELS-BR present similar attributes, but the data source is different. CABra adopted a product with a higher spatial resolution (100-m against 300-m) and more recent observation (2015 against 2009) than in CAMELS-BR. Due to this better spatial resolution. we chose to use a most recent land cover, even it being outside of the timespan of the hydrologic time series. CABra also brings information about the seasonal vegetation biomass of the catchment, in terms of NDVI, which is not present in CAMELS-BR.

Finally, both datasets take into account the human influence within each catchment, which is essential to a holistic understanding of the catchment behavior due to anthropogenic interactions and a lack of most of the large-sample datasets (Addor et al., 2020). CAMELS-BR presents data about water use, the volume of reservoirs, and the degree of regulation of the reservoirs. However, there is no combination or integration of these attributes in a specific index or approach. On the other hand, CABra presents eight attributes, i.e., distance to urbanization, the fraction of non-natural land-cover (crops and urban areas), water demand, reservoirs' count, area, volume, and streamflow regulation capacity (the last two are also found in CAMELS-BR), which can affect the hydrologic behavior of the catchment in terms of water quantity, quality, and regulation. Additionally, we developed a new hydrologic disturbance index (HDI), which considers all the eight attributes above-mentioned. The HDI is a quantitative index of the level of anthropization, being reproducible and practical to identify a more or less human-impacted catchment.

4. Conclusions

In this study, we have collected, synthesized, organized, and made available more than 100 topography, climate, streamflow, groundwater, soil, geology, land-use, and land cover, and hydrologic disturbance attributes for 735 catchments in Brazil. To do so, we have used several sources, such as observed time series, observed and modeled gridded data, remote sensing data, and reanalysis data. Moreover, we have calculated some attributes for providing more accurate data than those available in the literature, including potential evapotranspiration, and providing inexistent data, such as the hydrological disturbance index. As this dataset deals with catchment-scale averaged attributes, we have paid particular attention to DEM resolution, catchment delineation, while also manually inspecting each of the CABra catchments.

The development of the CABra dataset opens up several opportunities to test and develop a hypothesis in a unique environment like Brazil, with its vast and rich diversity in hydrology and landscapes. Finding relationships between the catchments' attributes will enable hydrologists to identify the drivers of the water fluxes in the catchment. We hope our dataset will aid catchment classification efforts that will ultimately unravel the underlying dominant controls of Brazilian regional hydrology across space and time. At the same time, the CABra dataset covers fundamentally different hydroclimatologic and ecologic regions than those covered by other similar large-sample datasets (United States, Great Britain, Chile, etc.), being a complement for global assessments and expanding the possibility of the use of our dataset for multiple scientific areas, such as geology, agronomy, ecohydrology.

We intend to expand the CABra dataset in the future. Information and attributes related to relevant fields of work, such as soil erosion, ecology, biology, and chemistry, as well as climate change projections, will be added to the CABra dataset in future updates releases. Thus, CABra represents a robust multi-source data collection effort for Brazil and is intended to play a key role in advancing the scientific understanding of climate-landscape-hydrology interactions. As such, we hope it will guide large-sample hydrology investigations and pave the way for testing novel hypotheses by both the Brazilian and the international scientific community.

5. Acknowledgments

This study was supported by grants from the Ministry of Science, Technology, and Innovation – MCTI and National Council for Scientific and Technological Development – CNPq [grants numbers 441289/2017-7, 306830/2017-5, and 309752/2020-5]. This study was also financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and CAPES Print.

6. References

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R. and Schmidt, G. A.: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing, Earth Syst. Dyn. Discuss., 6, 1–

20, doi:10.5194/esd-2018-51, 2018.

Addor, N., Newman, A. J., Mizukami, N. and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrol. Earth Syst. Sci., 21, 5293–5313, doi:10.5194/hess-21-5293-2017, 2017.

Addor, N., Do, H. X., Alvarez-Garreton, C., Coxon, G., Fowler, K. and Mendoza, P. A.: Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges, Hydrol. Sci. J., 65(5), 712–725, doi:10.1080/02626667.2019.1683182, 2020.

Ahrens, C. D.: Essentials of Meteorology: an invitation to the atmosphere., 2010.

Allen, R. G., Pereira, L. S., Raes, D. and Smith, M.: FAO Irrigation and Drainage Paper No. 56 - Crop Evapotranspiration., 1998.

Almagro, A., Oliveira, P. T. S., Nearing, M. A. and Hagemann, S.: Projected climate change impacts in rainfall erosivity over Brazil, Sci. Rep., 7(8130), 1–12, doi:10.1038/s41598-017-08298-y, 2017.

Almagro, A., Oliveira, P. T. S., Rosolem, R. and Hagemann, S.: Performance evaluation of Eta/HadGEM2-ES and Eta/MIROC5 precipitation simulations over Brazil, Atmos. Res., 244(1 November 2020), 105053, 2020.

Althoff, D., Dias, S. H. B., Filgueiras, R. and Rodrigues, L. N.: ETo-Brazil: A Daily Gridded Reference Evapotranspiration Data Set for Brazil (2000–2018), Water Resour. Res., 56(7), 0–2, doi:10.1029/2020WR027562, 2020.

ANA: Conjuntura dos recursos hídricos no Brasil 2019: informe anual / Agência Nacional de Águas., 2019a.

ANA: Manual dos Usos Consuntivos de Água do Brasil., 2019b.

ANA: Conjuntura dos recursos hídricos no Brasil 2020: informe anual, Brasília. [online] Available from: http://conjuntura.ana.gov.br/static/media/conjuntura-completo.23309814.pdf, 2020a.

ANA: Technical Note N. 52/2020/SPR, Brasília., 2020b.

Ao, T., Ishidaira, H., Takeuchi, K., Kiem, A. S., Yoshitari, J., Fukami, K. and Magome, J.: Relating BTOPMC model parameters to physical features of MOPEX basins, J. Hydrol., 320(1–2), 84–102, doi:10.1016/j.jhydrol.2005.07.006, 2006.

Avila-Diaz, A., Benezoli, V., Justino, F., Torres, R. and Wilson, A.: Assessing current and future trends of climate extremes across Brazil based on reanalyses and earth system model projections, Clim. Dyn., 55(5–6), 1403–1426, doi:10.1007/s00382-020-05333-z, 2020.

Bacellar, L. de A. P.: O papel das florestas no regime hidrológico de bacias hidrográficas, Geo.br, 1, 1–39, 2005.

Battisti, R., Bender, F. D. and Sentelhas, P. C.: Assessment of different gridded weather data for soybean yield simulations in Brazil, Theor. Appl. Climatol., 135(1–2), 237–247, doi:10.1007/s00704-018-2383-y, 2019.

Bellucci, A., Haarsma, R., Gualdi, S., Athanasiadis, P. J., Caian, M., Cassou, C., Fernandez, E., Germe, A., Jungclaus, J., Kröger, J., Matei, D., Müller, W., Pohlmann, H., Salas y Melia, D., Sanchez, E., Smith, D., Terray, L., Wyser, K. and Yang, S.: An assessment of a multi-model ensemble of decadal climate predictions, Clim. Dyn., 44(9–10), 2787–2806, doi:10.1007/s00382-014-2164-y, 2015.

Bender, F. D. and Sentelhas, P. C.: Solar radiation models and gridded databases to fill gaps in weather series and to project climate change in Brazil, Adv. Meteorol., 2018, doi:10.1155/2018/6204382, 2018.

Berghuijs, W. R., Larsen, J. R., van Emmerik, T. H. M. and Woods, R. A.: A Global Assessment of Runoff Sensitivity to Changes in Precipitation, Potential Evaporation, and Other Factors, Water Resour. Res., 53(10), 8475–8486, doi:10.1002/2017WR021593, 2017.

Beven, K., Asadullah, A., Bates, P., Blyth, E., Chappell, N., Child, S., Cloke, H., Dadson, S., Everard, N., Fowler, H. J., Freer, J., Hannah, D. M., Heppell, K., Holden, J., Lamb, R., Lewis, H., Morgan, G., Parry, L. and Wagener, T.: Developing observational methods to drive future hydrological science: Can we make a start as a community?, Hydrol. Process., 34(3), 868–873, doi:10.1002/hyp.13622, 2020.

Boulton, A. J., Peterson, C. G., Grimm, N. B. and Fisher, S. G.: Stability of an aquatic macroinvertebrate community in a multiyear hydrologic disturbance regime, Ecology, 73(6), 2192–2207, doi:10.2307/1941467, 1992.

Brocca, L., Ciabatta, L., Massari, C., Moramarco, T., Hahn, S., Hasenauer, S., Kidd, R., Dorigo, W., Wagner, W. and Levizzani, V.: Soil as a natural rain gauge: Estimating global rainfall from satellite soil moisture data, J. Geophys. Res. Atmos., 119(9), 5128–5141, doi:10.1002/2014JD021489, 2014.

Buchhorn, M., Smets, B., Bertels, L., Lesiv, M., Tsendbazar, N.-E., Herold, M. and Fritz, S.: Copernicus Global Land Service: Land Cover 100m: epoch 2015: Globe, , doi:10.5281/ZENODO.3243509, 2019.

Budyko, M. I.: Evaporation under natural conditions, Israel Program for Scientific Translations, Jerusalem., 1948.

Budyko, M. I.: Climate and Life, Elsevier, New York., 1974.

Chagas, V. B. P., Chaffe, P. L. B., Addor, N., Fan, F. M., Fleischmann, A. S., Paiva, R. C. D. and Siqueira, V. A.: CAMELS-BR: hydrometeorological time series and landscape attributes for 897 catchments in Brazil, Earth Syst. Sci. Data, 12(3), 2075–2096, doi:10.5194/essd-12-2075-2020, 2020.

Coleman, J. C., Miller, M. C. and Mink, F. L.: Hydrologic disturbance reduces biological integrity in urban streams, Environ. Monit. Assess., 172(1–4), 663–687, doi:10.1007/s10661-010-1363-1, 2011.

Dexter, A. R.: Soil physical quality Part I. Theory, effects of soil texture, density, and organic matter, and effects on root growth, Geoderma, 120(3–4), 201–2014, doi:10.1016/j.geodermaa.2003.09.005, 2004.

Do, H. X., Gudmundsson, L., Leonard, M. and Westra, S.: The Global Streamflow Indices and Metadata Archive (GSIM)-Part 1: The production of a daily streamflow archive and metadata, Earth Syst. Sci. Data, 10(2), 765–785, doi:10.5194/essd-10-765-2018, 2018.

Donohue, R. J., Roderick, M. L. and McVicar, T. R.: On the importance of including vegetation dynamics in Budyko's hydrological model, Hydrol. Earth Syst. Sci., 11(2), 983–995, doi:10.5194/hess-11-983-2007, 2007.

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H. V., Gusev, Y. M., Habets, F., Hall, a., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O. N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T. and Wood, E. F.: Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, J. Hydrol., 320(1–2), 3–17, doi:10.1016/j.jhydrol.2005.07.031, 2006.

Eichinger, W. E., Parlange, M. B. and Stricker, H.: On the concept of equilibrium evaporation and the value of the Priestley-Taylor coefficient, Water Resour. Res., 32(1), 161–164, doi:10.1029/95WR02920, 1996.

EMBRAPA: Sistema brasileiro de classificação de solos., 2018.

Fan, Y., Li, H. and Miguez-Macho, G.: Global patterns of groundwater table depth, Science (80-.)., 339(6122), 940–943, doi:10.1126/science.1229881, 2013.

FAO: World reference base for soil resources 2014. International soil classification system for naming soils and creating legends for soil maps., 2014.

Forzieri, G., Alkama, R., Miralles, D. G. and Cescatti, A.: Response to Comment on "Satellites reveal contrasting responses of regional climate to the widespread greening of Earth," Science (80-.)., 360(6394), 1180–1184, doi:10.1126/science.aap9664, 2018.

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

Gibbs, H. K., Ruesch, A. S., Achard, F., Clayton, M. K., Holmgren, P., Ramankutty, N. and Foley, J. A.: Tropical forests were the primary sources of new agricultural land in the 1980s and 1990s, Proc. Natl. Acad. Sci., 107(38), 16732–16737, doi:10.1073/PNAS.0910275107, 2010.

Gibbs, H. K., Rausch, L., Munger, J., Schelly, I., Morton, D. C., Noojipady, P., Barreto, P., Micol, L., Walker, N. F., Gibbs, B. H. K., Rausch, L., Munger, J., Schelly, I., Morton, D. C., Noojipady, P., Barreto, P., Micol, L., Walker, N. F., Amazon, B. and Cerrado, E.: Brazil's Soy Moratorium, Sci. - Policy Forum Environ. Dev., 347(6220), 377–378, doi:10.1126/science.aaa0181, 2014.

Gleeson, T., Moosdorf, N., Hartmann, J. and van Beek, L. P. H.: A glimpse beneath earth's surface: GLobal HYdrogeology MaPS (GLHYMPS) of permeability and porosity, Geophys. Res. Lett., 41(11), 3891–3898, doi:10.1002/2014GL061184.Received, 2014.

Grant, S. A.: Hydraulic Properties, Temperature Effects, Encycl. Soils Environ., 4, 207–211, doi:10.1016/B0-12-348530-4/00379-9, 2005.

Groenendyk, D. G., Ferré, T. P. A., Thorp, K. R. and Rice, A. K.: Hydrologic-process-based soil texture classifications for improved visualization of landscape function, PLoS One, 10(6), 1–17, doi:10.1371/journal.pone.0131299, 2015.

Guo, X., Zhang, H., Kang, L., Du, J., Li, W. and Zhu, Y.: Quality control and flux gap filling strategy for Bowen ratio method: Revisiting the Priestley-Taylor evaporation model, Environ. Fluid Mech., 7(5), 421–437, doi:10.1007/s10652-007-9033-8, 2007.

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, a., Kumar, R., Clark, M. and

Andréassian, V.: Large-sample hydrology: A need to balance depth with breadth, Hydrol. Earth Syst. Sci., 18(2), 463–477, doi:10.5194/hess-18-463-2014, 2014.

Hargreaves, G. H.: Moisture Availability and Crop Production, Trans. ASAE, 18(5), 0980–0984, doi:10.13031/2013.36722, 1975.

Hargreaves, G. H. and Allen, R. G.: History and evaluation of Hargreaves evapotranspiration equation, J. Irrig. Drain. Eng., 129(1), 53–63, doi:10.1061/(ASCE)0733-9437(2004)130:5(447.2), 2003.

Hartmann, J. and Moosdorf, N.: The new global lithological map database GLiM: A representation of rock properties at the Earth surface, Geochemistry, Geophys. Geosystems, 13(12), 1–37, doi:10.1029/2012GC004370, 2012.

Hengl, T., De Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S. and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning., 2017.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S. and Thépaut, J. N.: The ERA5 global reanalysis, Q. J. R. Meteorol. Soc., 146(730), 1999–2049, doi:10.1002/qj.3803, 2020.

Hou, A. Y., Kakar, R. K., Neeck, S., Azarbarzin, A. A., Kummerow, C. D., Kojima, M., Oki, R., Nakamura, K. and Iguchi, T.: The global precipitation measurement mission, Bull. Am. Meteorol. Soc., 95(5), 701–722, doi:10.1175/BAMS-D-13-00164.1, 2014.

Huffman, G. J., Adler, R. F., Bolvin, D. T., Gu, G., Nelkin, E. J., Bowman, K. P., Hong, Y., Stocker, E. F. and Wolff, D. B.: The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales, J. Hydrometeorol., 8(1), 38–55, doi:10.1175/JHM560.1, 2007.

Huntingford, C., Marsh, T., Scaife, A. A., Kendon, E. J., Hannaford, J., Kay, A. L., Lockwood, M., Prudhomme, C., Reynard, N. S., Parry, S., Lowe, J. A., Screen, J. A., Ward, H. C., Roberts, M., Stott, P. A., Bell, V. A., Bailey, M., Jenkins, A., Legg, T., Otto, F. E. L., Massey, N., Schaller, N., Slingo, J. and Allen, M. R.: Potential influences on the United Kingdom's floods of winter 2013/14, Nat. Clim. Chang., 4(9), 769–777, doi:10.1038/nclimate2314, 2014.

Kousky, V. E., Kagano, M. T. and Cavalcanti, I. F. a: A review of the Southern Oscillation: oceanic-atmospheric circulation changes and related rainfall anomalies, Tellus A, 36 A(5), 490–504, doi:10.1111/j.1600-0870.1984.tb00264.x, 1984.

Ladson, A. R., Brown, R., Neal, B. and Nathan, R.: A standard approach to baseflow separation using the Lyne and Hollick filter, Aust. J. Water Resour., 17(1), 25–34, doi:10.7158/W12-028.2013.17.1, 2013.

Lanza, R.: Hidrologia comparativa e perda de solo e água em bacias hidrográficas cultivadas com eucalipto e campo nativo com pastagem manejada, Master Thesis, 150, 2015.

Lima, C. H. R. and AghaKouchak, A.: Droughts in Amazonia: Spatiotemporal Variability, Teleconnections, and Seasonal Predictions, Water Resour. Res., 53(12), 10824–10840, doi:10.1002/2016WR020086, 2017.

Lo, M. H., Famiglietti, J. S., Yeh, P. J. F. and Syed, T. H.: Improving parameter estimation and water table depth simulation in a land surface model using GRACE water storage and estimated base flow data, Water Resour. Res., 46(5), 1–15, doi:10.1029/2009WR007855, 2010.

Lyne, V. and Hollick, M.: Stochastic Time-Variable Rainfall-Runoff Modeling, in Hydrology and Water Resources Symposium, pp. 89–92, Institution of Engineers National Conference Publication, Perth., 1979.

Lyon, S. W. and Troch, P. A.: Development and application of a catchment similarity index for subsurface flow, Water Resour. Res., 46(3), 1–13, doi:10.1029/2009WR008500, 2010.

Maes, W. H., Gentine, P., Verhoest, N. E. C. and Miralles, D. G.: Potential evaporation at eddy-covariance sites across the globe, Hydrol. Earth Syst. Sci. Discuss., (i), 1–38, doi:10.5194/hess-2018-470, 2018.

Maidment, D. R.: Arc Hydro: GIS for Water Resources., 2002.

Martens, B., Miralles, D. G., Lievens, H., Van Der Schalie, R., De Jeu, R. A. M., Fernández-Prieto, D., Beck, H. E., Dorigo, W. A. and Verhoest, N. E. C.: GLEAM v3: Satellite-based land evaporation and root-zone soil moisture, Geosci. Model Dev., 10(5), 1903–1925, doi:10.5194/gmd-10-1903-2017, 2017.

McMahon, T. A., Peel, M. C., Lowe, L., Srikanthan, R. and McVicar, T. R.: Estimating actual, potential, reference crop and pan evaporation using standard meteorological data: A pragmatic synthesis, Hydrol. Earth Syst. Sci., 17(4), 1331–1363, doi:10.5194/hess-17-1331-2013, 2013.

Melo, D. D. C. D., Xavier, A. C., Bianchi, T., Oliveira, P. T. S., Scanlon, B. R., Lucas, M. C. and Wendland, E.: Performance evaluation of rainfall estimates by TRMM Multi-satellite Precipitation Analysis 3B42V6 and V7 over Brazil, J. Geophys. Res. Atmos., 120(18), 9426–9436, doi:10.1002/2015JD023797, 2015.

Monteiro, L. A., Sentelhas, P. C. and Pedra, G. U.: Assessment of NASA/POWER satellite-based weather system for Brazilian conditions and its impact on sugarcane yield simulation, Int. J. Climatol., 38(3), 1571–1581, doi:10.1002/joc.5282, 2018.

Mukherjee, S., Joshi, P. K., Mukherjee, S., Ghosh, A., Garg, R. D. and Mukhopadhyay, A.: Evaluation of vertical accuracy of open source Digital Elevation Model (DEM), Int. J. Appl. Earth Obs. Geoinf., 21(1), 205–217, doi:10.1016/j.jag.2012.09.004, 2012.

Nepstad, D. C., Carvalho, C. R. De, Davidson, E. A., Jipp, P. H., Lefebvre, P. A., Negrelros, G. H., Sllva, E. D., Stone, T. A., Trumbore, S. E. and Vieira, S.: The role of deep roots in the hydrological and carbon cycles of Amazonian forests and pastures, Nature, 372(December), 666–669, 1994.

Newman, A. J., Clark, M. P., Craig, J., Nijssen, B., Wood, A., Gutmann, E., Mizukami, N., Brekke, L. and Arnold, J. R.: Gridded ensemble precipitation and temperature estimates for the contiguous United States, J. Hydrometeorol., 16(6), 2481–2500, doi:10.1175/JHM-D-15-0026.1, 2015.

Nobre, A. D., Cuartas, L. A., Hodnett, M., Rennó, C. D., Rodrigues, G., Silveira, A., Waterloo, M. and Saleska, S.: Height Above the Nearest Drainage - a hydrologically relevant new terrain model, J. Hydrol., 404(1–2), 13–29, doi:10.1016/j.jhydrol.2011.03.051, 2011.

Oliveira, P. T. S., Almagro, A., Pitaluga, F., Meira Neto, A. A., Durcik, M. and Troch, P. A.: CABra: a novel large-scale dataset for Brazilian catchments, in AGU Fall Meeting, p. 12138., 2020.

Pires, G. F., Abrahão, G. M., Brumatti, L. M., Oliveira, L. J. C., Costa, M. H., Liddicoat, S., Kato, E. and Ladle, R. J.: Increased climate risk in Brazilian double cropping agriculture systems: Implications for land use in Northern Brazil, Agric. For. Meteorol., 228–229, 286–298, doi:10.1016/j.agrformet.2016.07.005, 2016.

Priestley, C. H. B. and Taylor, R. J.: On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters, Mon. Weather Rev., 100(2), 81–92, doi:10.1175/1520-0493(1972)100<0081:otaosh>2.3.co;2, 1972.

Ren, H., Hou, Z., Huang, M., Bao, J., Sun, Y., Tesfa, T. and Ruby Leung, L.: Classification of hydrological parameter sensitivity and evaluation of parameter transferability across 431 US MOPEX basins, J. Hydrol., 536, 92–108, doi:10.1016/j.jhydrol.2016.02.042, 2016.

Roderick, M. L., Sun, F., Lim, W. H. and Farquhar, G. D.: A general framework for understanding the response of the water cycle to global warming over land and ocean, Hydrol. Earth Syst. Sci., 18(5), 1575–1589, doi:10.5194/hess-18-1575-2014, 2014.

Rodrigues, D. B. B., Gupta, H. V., Serrat-Capdevila, A., Oliveira, P. T. S., Mario Mendiondo, E., Maddock, T. and Mahmoud, M.: Contrasting American and Brazilian systems for water allocation and transfers, J. Water Resour. Plan. Manag., 141(7), 1–11, doi:10.1061/(ASCE)WR.1943-5452.0000483, 2015.

Salemi, L. F., Groppo, J. D., Trevisan, R., Seghesi, G. B., Moraes, J. M., Ferraz, S. F. B. and Martinelli, L. A.: Consequências hidrológicas da mudança de uso da terra de floresta para pastagem na região da floresta tropical pluvial Atlântica, Ambient. e Agua - An Interdiscip. J. Appl. Sci., 7(3), 127–140, doi:10.4136/ambi-agua.927, 2012.

Sankarasubramanian, A., Vogel, R. M. and Limbrunner, J. F.: Climate elasticity of streamflow in the United States, Water Resour. Res., 37(6), 1771–1781, doi:10.1029/2000WR900330, 2001.

Santos, H. G., Carvalho Júnior, W., Dart, R. O., Áglio, M. L. D., Sousa, J. S., Pares, J. G., Fontana, A., Martins, A. L. S. and Oliveira, A. P. O.: O novo mapa de solos do Brasil: legenda atualizada, Embrapa Solos, 67 [online] Available from: https://www.embrapa.br/busca-de-publicacoes/-/publicacao/920267/o-novo-mapa-de-solos-do-brasil-legenda-atualizada, 2011.

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A. and Carrillo, G.: Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA, Hydrol. Earth Syst. Sci., 15, 2895–2911, doi:10.5194/hess-15-2895-2011, 2011.

Saxton, K. E. and Rawls, W. J.: Soil Water Characteristic Estimates by Texture and Organic Matter for Hydrologic Solutions, Soil Sci. Soc. Am. J., 70(5), 1569–1578, doi:10.2136/sssaj2005.0117, 2006.

Saxton, K. E., Rawls, W. J., Romberger, J. S. and Papendick, R. I.: Estimating Generalized Soil-water Characteristics from Texture, Soil Sci. Soc. Am. J., 50(4), 1031–1036, doi:10.2136/sssaj1986.03615995005000040039x, 1986.

Schaake, J., Cong, S. and Duan, Q.: The US mopex data set, IAHS-AISH Publ., (307), 9–28, 2006.

Schulzweida, U.: CDO User guide (1.9.6), , 2015, doi:10.5281/zenodo.2558193, 2019.

Schumacher, D. L., Keune, J., van Heerwaarden, C. C., Vilà-Guerau de Arellano, J., Teuling, A. J. and Miralles, D. G.: Amplification of mega-heatwaves through heat torrents fuelled by upwind drought, Nat. Geosci., 12(9), 712–717, doi:10.1038/s41561-019-0431-6, 2019.

Shirazi, M. A. and Boersma, L.: A Unifying Quantitative Analysis of Soil Texture, Soil Sci. Soc. Am. J., 48(1), 142–147, doi:10.2136/sssaj1984.03615995004800010026x, 1984.

Shuttleworth, W. J.: Evaporation, in Handbook of Hydrology, edited by D. R. Maidment, p. 824, McGraw-Hill Education., 1996.

Solman, S. A., Sanchez, E., Samuelsson, P., da Rocha, R. P., Li, L., Marengo, J., Pessacg, N. L., Remedio, A. R. C., Chou, S. C., Berbery, H., Le Treut, H., de Castro, M. and Jacob, D.: Evaluation of an ensemble of regional climate model simulations over South America driven by the ERA-Interim reanalysis: Model performance and uncertainties, Clim. Dyn., 41(5–6), 1139–1157, doi:10.1007/s00382-013-1667-2, 2013.

Souza, R., Feng, X., Antonino, A., Montenegro, S., Souza, E. and Porporato, A.: Vegetation response to rainfall seasonality and interannual variability in tropical dry forests, Hydrol. Process., 30(20), 3583–3595, doi:10.1002/hyp.10953, 2016.

Spera, S. A., Galford, G. L., Coe, M. T., Macedo, M. N. and Mustard, J. F.: Land-use change affects water recycling in Brazil’s last agricultural frontier, Glob. Chang. Biol., 22(10), 3405–3413, doi:10.1111/gcb.13298, 2016.

Strahler, A. N.: Hypsometric Area-Altitude Analysis of Erosional Topography, Bull.

Geol. Soc. Am., 63(11), 1117–1142, doi:10.1130/0016-7606(1952)63, 1952.

Strahler, A. N.: Quantitative Analysis of Watershed Geomorphology, Trans. ASAE, 38(6), 913–920, 1957.

Tebaldi, C., Smith, R. L., Nychka, D. and Mearns, L. O.: Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles, J. Clim., 18(10), 1524–1540, doi:10.1175/JCLI3363.1, 2005.

Tetzlaff, D., Carey, S. K., McNamara, J. P., Laudon, H. and Soulsby, C.: The essential value of long-term experimental data for hydrology and water management, Water Resour. Res., 53(4), 2598–2604, doi:10.1002/2017WR020838, 2017.

Tomkins, K. M.: Uncertainty in streamflow rating curves: Methods, controls and consequences, Hydrol. Process., 28(3), 464–481, doi:10.1002/hyp.9567, 2014.

Tucker, C. J.: Red and Photographic Infrared l , lnear Combinations for Monitoring Vegetation, Remote Sens. Environ., 8, 127–150, 1979.

Twarakavi, N. K. C., Šimůnek, J. and Schaap, M. G.: Can texture-based classification optimally classify soils with respect to soil hydraulics?, Water Resour. Res., 46(1), doi:10.1029/2009WR007939, 2010.

UNEP and ANA: GEO Brazil Water Resources., 2007.

Vaze, J., Teng, J. and Spencer, G.: Impact of DEM accuracy and resolution on topographic indices, Environ. Model. Softw., 25(10), 1086–1098, doi:10.1016/j.envsoft.2010.03.014, 2010.

Wagener, T., Sivapalan, M., Troch, P. and Woods, R.: Catchment Classification and Hydrologic Similarity, Geogr. Compass, 1(4), 901–931, 2007.

Wanders, N. and Wada, Y.: Human and climate impacts on the 21st century hydrological drought, J. Hydrol., 526, 208–220, doi:10.1016/j.jhydrol.2014.10.047, 2015.

Wechsler, S. P.: Uncertainties associated with digital elevation models for hydrologic applications : a review, Hydrol. Earth Syst. Sci., 11(4), 1481–1500, 2007.

Westerberg, I. K. and McMillan, H. K.: Uncertainty in hydrological signatures, Hydrol. Earth Syst. Sci., 19, 3951–3968, doi:10.5194/hess-19-3951-2015, 2015.

Whited, D. C., Lorang, M. S., Harner, M. J., Hauer, F. R., Kimball, J. S. and Stanford, J. A.: Climate, hydrologic disturbance, and succession: Drivers of floodplain pattern, Ecology, 88(4), 940–953, doi:10.1890/05-1149, 2007.

WMO: Guide to the Global Observing System., 2010.

Woods, R. A.: Analytical model of seasonal climate impacts on snow hydrology: Continuous snowpacks, Adv. Water Resour., 32(10), 1465–1481, doi:10.1016/j.advwatres.2009.06.011, 2009.

Xavier, A. C., King, C. W. and Scanlon, B. R.: Daily gridded meteorological variables in Brazil (1980-2013), Int. J. Climatol., 2659(October 2015), 2644–2659, doi:10.1002/joc.4518, 2015.

Xavier, A. C., King, C. W. and Scanlon, B. R.: Daily gridded meteorological variables in Brazil (1980-2013), Int. J. Climatol., 2659(October 2015), 2644–2659, doi:10.1002/joc.4518, 2016.

Yadav, M., Wagener, T. and Gupta, H. V.: Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, Adv. Water Resour., 30, 1756–1774, doi:10.1016/j.advwatres.2007.01.005, 2007.

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., Sampson, C. C., Kanae, S. and Bates, P. D.: A high-accuracy map of global terrain elevations, Geophys. Res. Lett., 44(11), 5844–5853, doi:10.1002/2017GL072874, 2017.

Ye, B., Yang, D. and Kane, D. L.: Changes in Lena River streamflow hydrology: Human impacts versus natural variations, Water Resour. Res., 39(7), 1–14, doi:10.1029/2003WR001991, 2003.

Zandbergen, P. a: Error Propagation Modeling for Terrain Analysis using Dynamic Simulation Tools in ArcGIS Modelbuilder, Geomorphometry, 57–60, 2011.

Zhang, R., Chen, X., Zhang, Z. and Shi, P.: Evolution of hydrological drought under the regulation of two reservoirs in the headwater basin of the Huaihe River, China, Stoch. Environ. Res. Risk Assess., 29(2), 487–499, doi:10.1007/s00477-014-0987-z, 2015.

Zhang, Y., Peña-Arancibia, J. L., McVicar, T. R., Chiew, F. H. S., Vaze, J., Liu, C., Lu, X., Zheng, H., Wang, Y., Liu, Y. Y., Miralles, D. G. and Pan, M.: Multi-decadal trends in global terrestrial evapotranspiration and its components, Sci. Rep., 6(December 2015), 1–12, doi:10.1038/srep19124, 2016.

Zhao, G., Gao, H., Naz, B. S., Kao, S. C. and Voisin, N.: Integrating a reservoir regulation scheme into a spatially distributed hydrological model, Adv. Water Resour., 98, 16–31, doi:10.1016/j.advwatres.2016.10.014, 2016.

Zhou, Q. and Liu, X.: Analysis of errors of derived slope and aspect related to DEM data properties, Comput. Geosci., 30(4), 369–378, doi:10.1016/j.cageo.2003.07.005, 2004.

CHAPTER 3

ASSESSMENT OF BOTTOM-UP SATELLITE RAINFALL PRODUCTS IN ESTIMATING RIVER DISCHARGE AND HYDROLOGIC SIGNATURES IN BRAZILIAN CATCHMENTS

Almagro, A., Oliveira, P. T. S., Brocca, L. 2021. Assessment of bottom-up satellite rainfall products on estimating river discharge and hydrologic signatures in Brazilian catchments, Journal of Hydrology, 603, 126897. https://doi.org/10.1016/j.jhydrol.2021.126897. (Impact Factor 2021: 5.722)

Abstract

Satellite rainfall products are one of the most valuable tools for water resources monitoring in data-scarce regions, due to their low latency and quasi-global range. However, there are still uncertainties associated with rainfall products performance used to estimate hydrologic signatures in several regions, such as Brazil. Here, we investigate the performance of three rainfall products in estimating daily precipitation, daily river discharge, and hydrologic signatures over Brazil: the SM2RAIN-ASCAT and the GPM+SM2RAIN satellite products, and the ERA5 reanalysis product. We used a subset of 520 catchments from the Catchments Attributes for Brazil (CABra) dataset and the hydrologic modeling was carried out using the MISDc hydrologic model. Satellite-based products performed better than ERA5 for most Brazilian biomes in estimating daily precipitation when compared with ground observations used as reference. Daily river discharge was also better modeled with SM2RAIN-ASCAT and GPM+SM2RAIN. Hydrologic modeling presented low values of bias and more than 80% of catchments with KGE > 0.5 in calibration. Lastly, hydrologic signatures were well estimated by SM2RAIN-ASCAT and GPM+SM2RAIN, and for some biomes (Atlantic Forest, Cerrado, and Caatinga) they are better predictors than ground-based observations. We showed that there is a significant added value when using SM2RAIN-ASCAT and GPM+SM2RAIN products in tropical catchments, allowing a high-quality continuous water resources monitoring even in data-scarce regions. Besides, our findings pave the way for a better understanding of hydrologic extremes (drought and floods) using these satellite rainfall products on multiple spatial and temporal scales.

Keywords: satellite, climate, streamflow, catchments, hydrologic signatures.

1. Introduction

Precipitation is a prerequisite for human life and development, and its monitoring is crucial for several applications in many fields, such as hydrology, climatology, geology, and agriculture. It is of great interest for understanding Earth's system as a whole (Barrett and Beaumont, 1994), as it is considered the most important variable in geosciences (Maggioni and Massari, 2018). Due to its high spatio-temporal variability, it is a hard variable to measure with a considerable uncertainty associated, and the ground-based rainfall monitoring generally suffers from a lack of station's density, especially in developing countries (Barrett and Beaumont, 1994; Brocca et al., 2020; Xavier et al., 2016).

Among the alternative options to ground-based stations currently available for precipitation monitoring (weather prediction models, weather radar, and satellite-based observations), the satellite-based products are the most comprehensive and viable (and sometimes the unique) source of information, with low cost and easy access for users in data-scarce regions (Brocca et al., 2019; Camici et al., 2020; Massari et al., 2020; Wagner et al., 2012). Products like the Tropical Rainfall Measuring Mission (TRMM) have been extensively used worldwide and it is one of the most known satellite rainfall products (Huffman et al., 2007).

The "bottom-up" approach for precipitation monitoring arose in recent years and uses the soil moisture of soil to estimate and improve the precipitation over land, contrary to the conventional and most commonly used approach "top-down" (Brocca et al., 2014). A set of papers using the Kalman Filter (KF) and satellite soil moisture have been published (Crow et al., 2009; Crow and Ryu, 2009; Crow and Zhan, 2007), showing improvements in rainfall and streamflow estimates. Pellarin et al. (2009, 2013) adopted an Antecedent Precipitation Index (API) constrained with the Advanced Microwave Scanning Radiometer for EOS (AMSR-E) to improve TRMM satellite-based rainfall product. Recently, Brocca et al. (2019) presented a new satellite-based rainfall product based on soil moisture, the SM2RAIN-ASCAT. This product is based on the implementation of the SM2RAIN algorithm (Brocca et al., 2014) to soil moisture observations from the Advanced SCATterometer (ASCAT). Unlike the previous ones, the SM2RAIN-ASCAT is a rainfall product estimated with soil moisture, and not corrected by soil moisture. The SM2RAIN-based rainfall products were also combined with the Integrated Multi-Satellite Retrievals for Global Precipitation Measurement (GPM-IMERG), resulting in an enhanced product named GPM+SM2RAIN (Massari et al., 2020).

The SM2RAIN-ASCAT and GPM+SM2RAIN are the most advanced satellite-based rainfall products that use the bottom-up approach so far. Both are low-latency (<1 day) with a global range that may be useful for continuous monitoring of water resources. Despite the facility and advantages of using the satellite-based rainfall products, we must keep in mind that still have uncertainties associated (Brocca et al., 2014), and these uncertainties are transferred to the applications, such as hydrological modeling. Thus, it is important to assess not only the reliability of estimating precipitation, but also the suitability for river discharge simulation (or any other application), as done in Brocca et al. (2020), Brunetti et al. (2021), and Camici et al. (2020). For instance, Brunetti et al. (2021) employed the SM2RAIN-ASCAT and the GPM products (and their integration) to perform landslide forecasting in India, establishing that the satellite rainfall products generally outperform ground-based observations in this goal.

Hydrological signatures are important indicators to provide an understanding of the hydrological functionality and behavior of a catchment (Sawicz et al., 2011). Because of this, the hydrologic signatures can also be used as indicators of hydrologic modeling ability in simulate catchment's responses variability (Euser et al., 2013). Therefore, if a combination of rainfall product and hydrologic model can provide accurate hydrologic signatures, it is a good indicator of product/model capability to reproduce the hydrological behavior of a catchment.

Brazil is a country covering almost 50% of South America's area, presenting heterogeneous hydrological, climatic, and geophysical attributes. It should be underlined that between 12% and 18% of the world's renewable freshwater flows in Brazilian rivers (Rodrigues et al., 2015; UNEP and ANA, 2007). Despite being one of the most important countries to the global water fluxes, Brazil has a scarce allocation of funding for hydro-meteorological monitoring, which creates great challenges for proper knowledge of its water resources, including precipitation. There is a lack of gauge-stations for precipitation monitoring, as shown in Xavier et al. (2016), with stations density lower than recommended by the World Meteorological Organization (WMO), making the satellite rainfall products an important tool for water resources monitoring in Brazil.

Some studies have been carried out to evaluate the performance of SM2RAIN-based rainfall products in Brazil. Paredes-Trejo et al. (2018) evaluated the performance of SM2RAIN-CCI (Ciabatta et al., 2018) product to simulate rainfall characteristics and temporal distribution over Northeastern Brazil, showing quite good performance on daily estimations. Later, Paredes-Trejo et al. (2019) expanded the previous analysis for entire Brazil and found that SM2RAIN-based products (-CCI and -ASCAT) can be effectively used for water resources and agricultural purposes on a daily scale. However, none of the abovementioned studies address broader applications of the satellite rainfall products, such as river discharge and hydrologic signatures estimations across the country. Thus, the performance of hydrological applications of satellite-based rainfall products that use a bottom-up approach is still unknown for Brazilian catchments.

To our knowledge, this is the first study evaluating and validating the performance of the bottom-up satellite rainfall products in simulating daily river discharge and hydrologic signatures for Brazilian catchments. Thus, the objective of this study is to assess the performance of the SM2RAIN-ASCAT and GPM+SM2RAIN rainfall products in simulating daily precipitation in Brazil. For that, at the catchment scale, the SM2RAIN-based products are compared with ground-based observations used as a reference, and the performance is evaluated against a widely-used reanalysis rainfall product, ERA5 (Hersbach et al., 2020). Moreover, the hydrological application of these data in simulating daily river discharge on 520 Brazilian catchments is investigated. We have used a large-sample catchment attributes dataset, the Catchment Attributes for Brazil (CABra), to calibrate and validate the simulations of the MISDc hydrological model (Brocca et al., 2011).

2. Study area

Our study adopts 520 catchments in Brazil as case studies. They are located in six different biomes throughout Brazil: Amazon, Atlantic Forest, Cerrado, Caatinga, Pampa, and Pantanal (see Figure 1). Brazil has continental dimensions, with ~8,500,000 km² ranging from longitudes 34W to 73W and latitudes 6N to 34S. Catchment areas range from 22 to 4,800,000 km², covering a wide range of climate patterns, land use, geology, and soil types. According to Almagro et al. (2020b), precipitation in Brazil ranges from 400 in the Caatinga, which is a semi-arid region comprised mostly by secondary vegetation (herbaceous and arboreous) with a severe dry season, to 4,000 mm year^{-1} in the Amazon, the largest tropical biome in the world, consisting of a densely vegetated rainforest. High values of precipitation are also observed in the Atlantic Forest, characterized by rainforest cover in the coastal area and the semi-deciduous forest in the continental. Intermediate amounts of precipitation (1,500 to 2,500 mm year^{-1}) are observed in the Cerrado (woodlands and savanna), the Pantanal (one of the largest flooded areas in the world), and in the Pampa biome (predominance of natural pastures, with tree formations and sparse shrub).

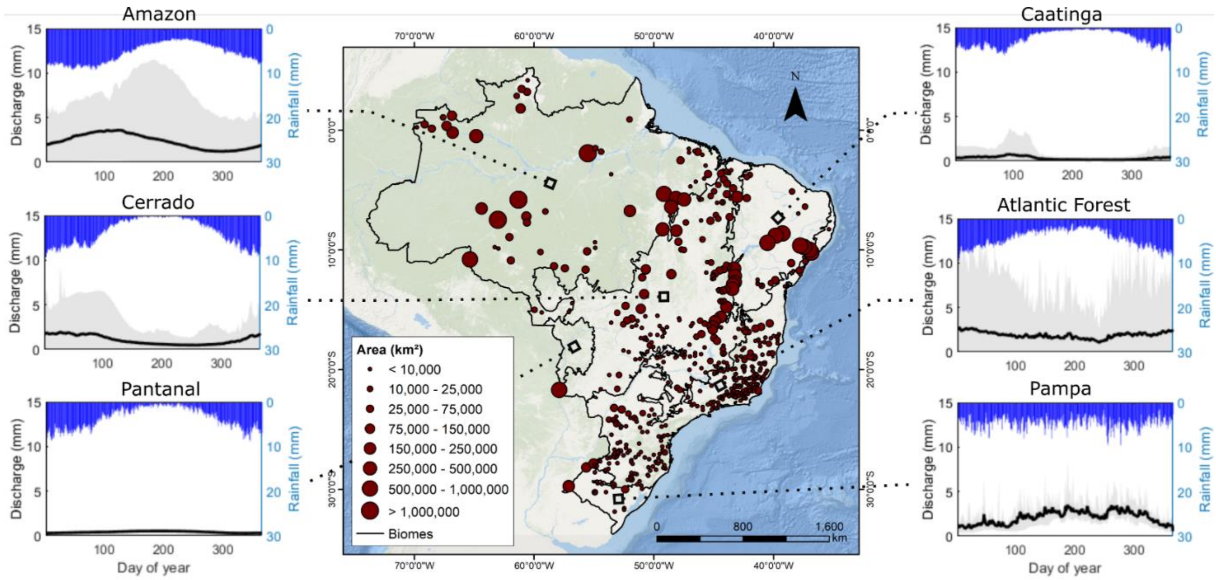

Figure 1: Location map containing the 520 catchments in which the remote sensing products were evaluated. For each biome, the long-term precipitation and streamflow are represented in the subplots. The shaded area represents the range of daily streamflow for a given biome.

3. Datasets

3.1. Ground observations

To evaluate the bottom-up remote sensing precipitation products, we used a gridded-interpolated product derived from observations developed by Xavier et al. (2016) for entire Brazil as a ground-truth reference. This product used observed precipitation data (and many other climatic variables) derived from approximately 4,000 rain gauges from the Brazilian Water Agency (ANA), the National Institute of Meteorology (Inmet), and the Water and Electric Energy Department of São Paulo state (DAEE/SP), presenting a 0.25ºx0.25º spatial resolution and covering the 1980-2015 period. Moreover, this reference dataset has been extensively applied in previous studies, such as evaluation of remote sensing products (Melo et al., 2015; Paredes-Trejo et al., 2018; Paredes-Trejo and Barbosa, 2017; Paredes-trejo et al., 2017), vegetation response to rainfall variability (Souza et al., 2016), impacts of climatic extremes

(Melo et al., 2016), and climate change historical simulations and projections (Almagro et al., 2017, 2020; Avila-Diaz et al., 2020). The Xavier dataset can be accessed through http://careyking.com/data-downloads/.

The observed river discharge for the Brazilian catchments is derived from the Catchment Attributes for Brazil (CABra) dataset, developed by Almagro et al. (2021). The CABra dataset is a large-sample dataset of catchment attributes for 735 catchments in Brazil. Authors collected, synthesized, organized, and made available a set of more than 100 catchment attributes for eight main classes: topography, climate, streamflow, groundwater, soil, geology, land-cover, and hydrological disturbance. Due to the continental scale of Brazil, there is high heterogeneity in the hydrological behavior of the catchments. CABra provides daily river discharge for each of its catchments, which was obtained from the ANA, covering the hydrological period from 1980 to 2010, with high-quality records (less than 10% of missing data). Moreover, CABra river discharge data were quality checked for inconsistencies and outliers on its data. Data from the CABra dataset is available at: https://doi.org/10.5281/zenodo.4070146. To increase the performance evaluation reliability, we have extended the time-series of river discharge from the CABra dataset to overlap the whole period of satellite monitoring, so the CABra dataset used in this study is an updated version of the published one, covering the 2007-2018 period.

3.2. Rainfall products

In this study, we adopted three different precipitation products to conduct the performance on precipitation estimates and hydrological modeling. Two of them are satellite-derived: SM2RAIN-ASCAT and the GPM+SM2RAIN. The last one is derived from reanalysis data, the ERA5.

The first remote sensing product used in this study is SM2RAIN-ASCAT. SM2RAIN-ASCAT is a gridded rainfall dataset based on soil moisture, which is commonly called as “bottom-up” method. The SM2RAIN algorithm is based on the inversion of the soil water balance equation to estimate the amount of water entering the soil, by using soil moisture as input. SM2RAIN-ASCAT rainfall product uses the satellite observations of soil moisture data from the Advanced SCATterometer (ASCAT), carried by three satellites from the MetOp family. The SM2RAIN-ASCAT precipitation product was developed by Brocca et al. (2019), presents a spatial resolution of 12.5 km and its latest release covers the 2007-2020 period. SM2RAIN-ASCAT has been extensively tested and applied throughout the world since its release and appears as a valuable product for short-latency rainfall monitoring. SM2RAIN-ASCAT is referred to here as "SM2RASC", and version 1.2 used in the study can be download at: https://zenodo.org/record/3635932.

The second satellite rainfall product used here is GPM+SM2RAIN. This product is a short-latency, daily 25km satellite-based rainfall product generated by integrating the IMERG-ER and SM2RAIN products. It was developed by Massari et al. (2020) by using the Optimal Linear Combination (OLC) approach (Bishop and Abramowitz, 2013) which optimally merges multiple estimates from the same climatological variable (rainfall in this case) by minimizing the error with a calibration dataset, the ERA5. GPM+SM2RAIN covers the 2007-2018 period, is referred to here as "GPMSM2R", and is available at: https://doi.org/10.5281/zenodo.3345322.

The third precipitation product used in our study is the ERA5 reanalysis (Hersbach et al., 2020). ERA5 is the latest version of climate reanalysis data from the European Centre for Medium-Range Weather Forecasts (ECMWF) and provides hourly, daily, and monthly data on several atmospheric, sea, and land variables. As a reanalysis dataset, ERA5 employs past observations and models to generate accurate and consistent time series of climate variables and parameters, being one of the most widely used datasets in geosciences. ERA5 precipitation is available at 36 km spatial resolution, from 1950 to the present. Here we also make use of the ERA5's temperature product for the hydrological modeling input. ERA5 products can be explored and downloaded through the Copernicus Climate Data Store (https://cds.climate.copernicus.eu/).

4. Methodology

4.1. Hydrologic modeling

To perform the hydrologic modeling, we have used the Modello Idrologico Semi-Distribuito in continuo – MISDc (Brocca et al., 2011), which is a semi-distributed continuous rainfall-runoff model. The MISDc is a parsimonious and reliable rainfall-runoff model to estimate the total streamflow considering the wetness conditions of the soil before the rainfall event. To calibrate its nine parameters (Table 1), the MISDc solves the objective function to maximize the Kling-Gupta efficiency (KGE) (Gupta et al., 2009) between the simulated and observed river discharge. The calibration period was taken in the first 7 years of data (Jan-2007 to Mar-2014), with a warm-up period of one year, while the validation period englobes the apr-2014 to dec-2018 period.

Table 1: Description of the MISDc parameters and initial conditions range for calibration.

Parameter	Description	Unit	Range
Wp	Maximum water capacity of the soil layer	mm	0.1-0.9
$Wmax_2$	Total water capacity of 2nd layer	mm	300-4000
m_1	The exponent of drainage for 1st layer	-	2-10
m_2	The exponent of drainage for 2nd layer	-	5-20
Ks_1	Hydraulic conductivity for 1st layer	mm h^{-1}	0.1-20
Ks_2	Hydraulic conductivity for 2nd layer	mm h^{-1}	0.01-65
γ	Coefficient of lag-time relationship	-	0.5-3.5
Kc	Parameter for potential evapotranspiration	-	0.4-2.0
α	Exponent of runoff	-	1-15
Cm	Snow module parameter	ºC day^{-1}	-*

* Removed from analysis due to non-existence of snow in Brazil.

MISDc model has been widely used in recent years. Brocca et al. (2020) conducted a study in 10 catchments in Europe, West Africa, and South Africa and found that the remote sensing products outperform the gauge-based rainfall products in simulating daily river discharge for western Africa catchments. In Europe, Camici et al. (2020) tested the accuracy of different satellite rainfall products against observations for 1318 catchments covering 23 countries using MISDc. The authors found that SM2RAIN-ASCAT is the most reliable product for river discharge simulation across Europe.

To minimize the uncertainties of the hydrological modeling, we filtered the total number of catchments, selecting only those with long and consistent series. From the extended version of the CABra dataset (Almagro et al., 2021), we have applied the following criteria: only perennial catchments with data covering all the observation period of the remote sensing products (12 years from Jan-2007 to Dec-2018), and with less than 5% of missing data. This procedure reduced the number of catchments from 735 to 520, but still covers all Brazilian biomes and regions.

4.2. Hydrologic signatures

To assess the long-term reliability of the satellite rainfall products, we calculated a series of hydrological signatures that quantifies the hydrological characteristics, providing a better understanding of hydrological behavior. These signatures can be used to explain the catchment's variability in responses to climate forcings. We calculated, for the 520 catchments, a subset of 11 hydrological signatures for the SM2RASC, GPMSM2R, ERA5, and the ground-based reference product, related to the magnitude, frequency, duration, timing, and rate of change of river discharge (Sawicz et al. (2011) and Westerberg and McMillan (2015)), listed in Table 2. The calculation of the hydrological series was carried out using the Toolbox for Streamflow Signatures in Hydrology (TOSSH), developed by Gnann et al. (2021).

Table 2: Hydrological signatures chosen for this study.

Attribute	Long name	Unit
Mean Q	Mean daily streamflow	mm.day^{-1}
Q5	Streamflow 5th quantile	mm.day^{-1}
Q95	Streamflow 95th quantile	mm.day^{-1}
Q 7-day min	7-day minimum streamflow	mm.day^{-1}
High Q frequency	Max streamflow frequency	y^{-1}
High Q duration	Max streamflow duration	days
Low Q frequency	Min streamflow frequency	y^{-1}
Low Q duration	Min streamflow duration	days
HFD	Mean half-flow date	day of the year
BFI	Baseflow index	-
FDC slope	The slope of the flow duration curve	-

4.3. Performance evaluation on precipitation estimates and hydrologic modeling

To evaluate the performance and quality of simulated precipitation against the observed data product during an overlap period (2007-2015) between them, we used the following statistical metrics: bias (BIAS), root mean squared error (RMSE), and correlation coefficient (CC) (Equations 1, 2, and 3).

$$BIAS = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{P_{sim\,i} - P_{obs\,i}}{P_{obs\,i}}\right) \quad 1$$

$$CC = \frac{n(\sum_{i=1}^{n} P_{sim\,i}\,P_{obs\,i}) - (\sum_{i=1}^{n} P_{sim\,i})(\sum_{i=1}^{n} P_{obs\,i})}{\sqrt{\left[n\sum_{i=1}^{n} {P_{sim\,i}}^2 - (\sum_{i=1}^{n} P_{sim\,i})^2\right]\left[n\sum_{i=1}^{n} {P_{obs\,i}}^2 - (\sum_{i=1}^{n} P_{obs\,i})^2\right]}} \quad 2$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_{obs\,i} - P_{sim\,i})^2}{n}} \quad 3$$

where P is the long-term mean precipitation from observations "*obs*" and simulations "*sim*"; σ is the standard deviation of the annual precipitation; n is the number of points for each biome.

To the river discharge simulations, we have included two more widely-used performance metrics to calibrate and evaluate hydrologic models with observed data: the Nash-Sutcliffe efficiency (Nash and Sutcliffe, 1970) and the Kling-Gupta efficiency (Gupta et al., 2009), described in Equations 4 and 5. Both metrics might be interpreted in the same way, but with different approaches. The closer to 1 is the NSE (and KGE), the better is the agreement between simulations and observations. When NSE (and KGE) < 0 (-0.41), it indicates that it is better to use the mean as a predictor.

$$NSE = 1 - \frac{\sum_{i=1}^{n}(Q_{sim\,i} - Q_{obs\,i})^2}{\sum_{i=1}^{n}(Q_{obs\,i} - \mu_{obs})^2} \quad 4$$

$$KGE = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{sim}}{\sigma_{obs}} - 1\right)^2 + \left(\frac{\mu_{sim}}{\mu_{obs}} - 1\right)^2} \quad 5$$

where Q is the daily river discharge from observations "obs" and simulations "sim"; σ is the standard deviation observations "obs" and simulations "sim"; and μ is the mean daily river discharge from observations "obs" and simulations "sim"; and 'r' is the linear correlation between observed and simulated daily river discharge, dimensionless.

Finally, we applied a Skill Score (SS) to evaluate the goodness-of-fit from the satellite and reanalysis rainfall products against the ground-based rainfall product in simulating the daily river discharge. The SS is presented in Jiang and Bauer-Gottwein (2019), and raises from the need to avoid the catchment variability due to different conditions of flow generation, and is represented by Equation 6:

$$SS = 1 - \frac{\sum(Q_{simS} - \mu_{obs})^2}{\sum(Q_{simG} - \mu_{obs})^2} \quad 6$$

where Q is the daily river discharge from simulations based on satellite and reanalysis rainfall products "simS"; and simulations based on ground-observed rainfall product "simG"; and μ is the mean daily river discharge from observations "obs".

5. Results and discussion

5.1. Precipitation estimates over Brazil

The results of the spatial variability of performance scores of remote sensing products against ground-based rainfall are presented in Figure 2. The SM2RASC bias of daily precipitation is close to zero in almost all Brazilian biomes, with exception of the Amazon where it shows overestimation in central Amazon and underestimation in western and eastern portions. The bias presented by the merged product GPMSM2R is slightly greater than SM2RASC, but still with most of the areas with values close to zero. On the other hand, the ERA5 rainfall product follows the same pattern of SM2RASC, underestimating/overestimating the precipitation in all Amazon, and close-to-zero bias on most of Brazil. In RMSE spatial analysis, SM2RASC and GPMSM2R perform similarly and the ERA5 product presents the highest values, up to 10 mm day^{-1}. The correlation coefficient presents better performance in satellite rainfall products than in ERA5 products, with the highest values in Amazon, Pampa, and Caatinga, showing its high reliability on estimating the precipitation in different regions of Brazil, with different climates and seasonal cycles.

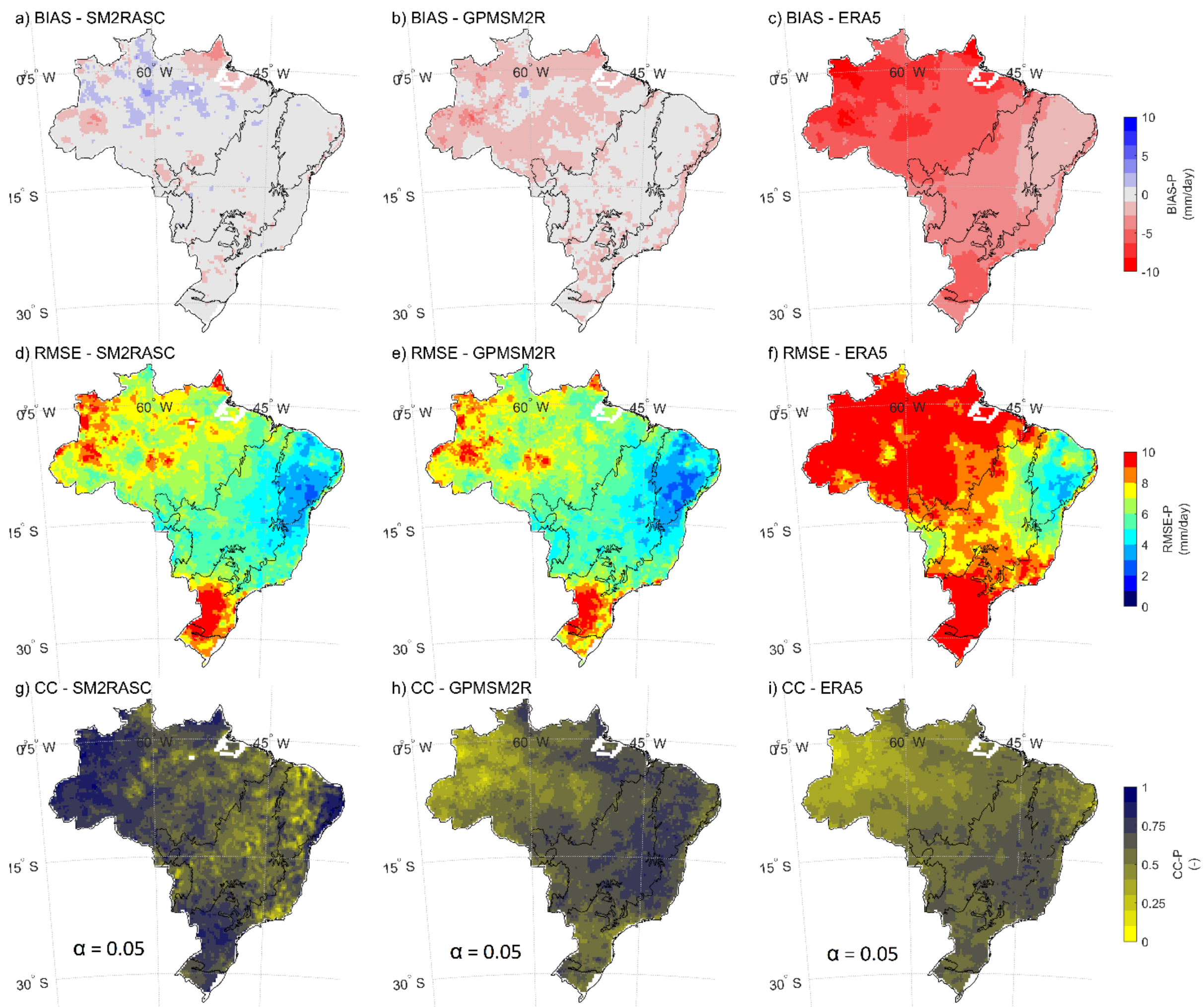

Figure 2: Spatial distribution of performance scores for the three different rainfall products (SM2RASC, GPMSM2R, and ERA5) in estimating daily precipitation over Brazil: a), b) and c) bias; d), e) and f) root mean square error (RMSE); and g), h) and i) correlation coefficient with a 0.05 significance level.

The added value of satellite rainfall products in relation to ERA5 can be seen in the radar chart in all regions and biomes of Brazil (Figure 3) and Table 3. The radar chart allows us to infer what is the most suitable product by the area of the polygon. Since better values of each performance skill are converging to the center of the chart, the lower the polygon's area, the better is the product represented by this polygon. For almost all biomes, on average, the satellite rainfall products perform better than the ERA5 estimations. The exception is the Caatinga, where ERA5 presents the best performances in the area-averaged values of BIAS and RMSE (referred to as aBIAS and aRMSE).

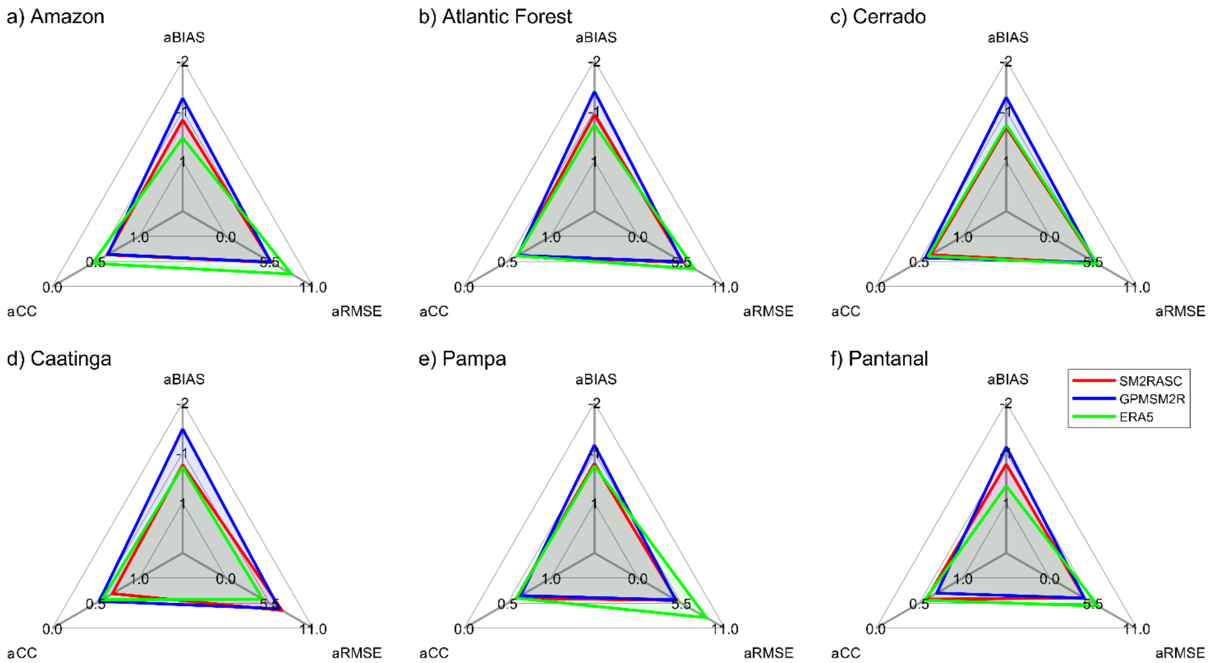

Figure 3: Radar chart for area-averaged performance scores of rainfall products on Brazilian biomes. Best values of the area-averaged bias (aBIAS), area-averaged root mean squared error (aRMSE), and area-averaged correlation coefficient (aCC) are converging to the center of the triangle, making the lower the area, the better the performance.

Table 3: Area-averaged performance scores of rainfall products on Brazilian biomes on simulating precipitation. Best scores for each biome are highlighted in bold. Values of Observed rainfall are in mm.day-1, aBIAS are in mm.day-1 (and in %), aRMSE are in mm.day-1 (and in %), and aCC are dimensionless.

Biome	Observed	Model	aBIAS	aRMSE	aCC
Amazon	6.08	SM2RAIN-ASCAT	**-0.23 (-3.8%)**	5.75 (94%)	0.63
		GPM+SM2RAIN	-0.88 (-14.4%)	**5.65 (93%)**	**0.64**
		ERA5	0.31 (5.1%)	8.29 (136%)	0.46
Atlantic Forest	4.07	SM2RAIN-ASCAT	-0.38 (-9.3%)	5.77 (142%)	0.61
		GPM+SM2RAIN	-1.08 (-26.5%)	**5.59 (137%)**	**0.62**
		ERA5	**-0.08 (-2.0%)**	7.17 (176%)	0.61
Cerrado	3.73	SM2RAIN-ASCAT	**-0.01 (-0.3%)**	5.93 (159%)	**0.63**
		GPM+SM2RAIN	-0.91 (-24.4%)	**5.77 (155%)**	0.57
		ERA5	-0.08 (-2.1%)	6.07 (163%)	0.60
Caatinga	1.81	SM2RAIN-ASCAT	-0.14 (-7.8%)	7.10 (394%)	**0.69**
		GPM+SM2RAIN	-1.21 (-67,2%)	6.73 (374%)	0.54
		ERA5	**-0.08 (-4,4%)**	**4.67 (259%)**	0.57
Pampa	4.31	SM2RAIN-ASCAT	-0.18 (-4.2%)	4.91 (114%)	0.60
		GPM+SM2RAIN	-0.73 (-16.9%)	**4.78 (111%)**	**0.65**

		ERA5	**-0.11 (-2.6%)**	8.68 (201%)	0.60
		SM2RAIN-ASCAT	**-0.16 (-5.1%)**	4.39 (140%)	0.58
Pantanal	3.13	GPM+SM2RAIN	-0.68 (-21.7%)	**4.38 (140%)**	**0.70**
		ERA5	0.49 (15.7%)	6.13 (196%)	0.56

5.2. Streamflow estimates on Brazilian catchments

The performance of the SM2RASC, GPMSM2R, and ERA5 rainfall products in simulating the river discharge of Brazilian catchments during the calibration period (carried from Jan-2007 to Mar-2014) from the MISDc hydrological model is illustrated in Figure 4. All products present a close-to-zero mean daily river discharge bias (BIAS-Q) in most of the catchments, with some overestimated values in the southern portion and underestimated values in the northeastern part, especially in the São Francisco river basin. This is probably due to the large reservoir's regulation capacity and due to the groundwater exploration in this region – as shown in Almagro et al. (2021) and Lucas et al. (2021), which is not taken into account in the hydrological modeling. On median terms (mBIAS-Q), the GPMSM2R and ERA5 present best performances with absolute mBIAS-Q = 0.001 mm day^{-1}, followed by the SM2RASC (mBIAS-Q = 0.006 mm day^{-1}). Spatial variability of mean daily river discharge root mean squared error (RMSE-Q) shows a similar pattern and median values ranging from 0.780 to 0.923 mm day^{-1}, with RMSE-Q ≤ 2 mm day^{-1} for more than 80% of the catchments. The highest values of RMSE-Q are found for the southern region, in the Pampa biome, reaching up to 5 mm day^{-1}. The mean Nash-Sutcliffe Efficiency for daily river discharge (NSE-Q) demonstrates a better performance of GPMSM2R against other products analyzed here, with median value of 0.614 and 68% of the catchments with NSE-Q > 0.5. Considering the SM2RASC and ERA5 products, we found median values of NSE-Q (mNSE-Q) of 0.501 and 0.453, and 51% and 42% of the catchments with NSE-Q > 0.5. In terms of the Kling-Gupta Efficiency on daily river discharge (KGE-Q), all products perform well, with median values (mKGE-Q) higher than 0.7. SM2RASC presented KGE-Q > 0.5 in 82% of the catchments (mKGE-Q = 0.748), while in GPMSM2R and ERA5 there were 92% (mKGE-Q = 0.807) and 85% (mKGE-Q = 0.727), respectively. Values of KGE-Q ≈ 0.5 in the southern portion of Brazil should be attributed to the absence of a well-defined seasonality of precipitation in this region, which is important to establish a temporal correlation between observations and simulations. Another point to note is the poor performance in terms of NSE-Q and KGE-Q for some catchments in the Caatinga biome, in northeastern Brazil. We can explain the poor performance by the large number of reservoirs in catchments located in this region, which again is not taken in account in the hydrological modeling.

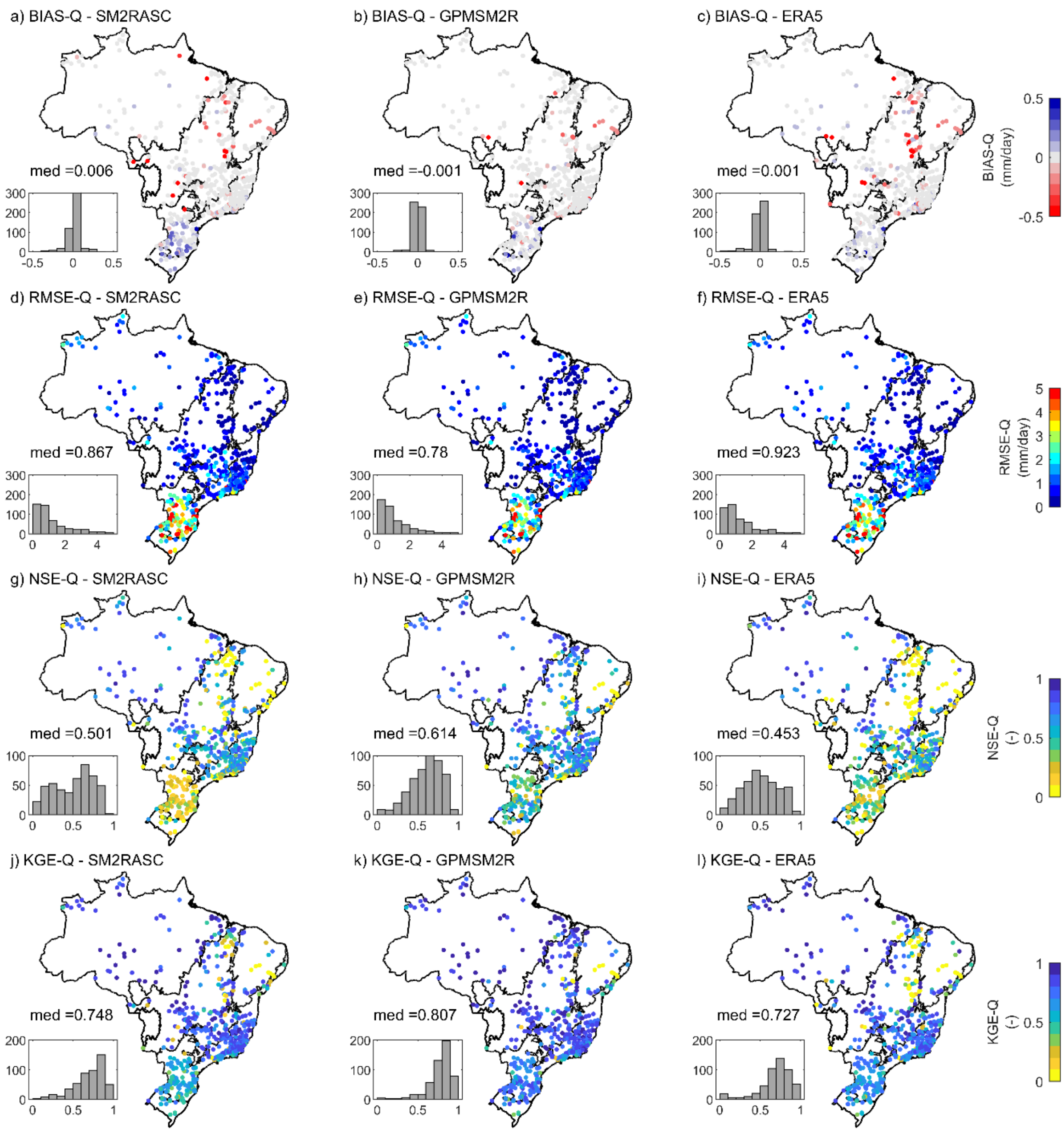

Figure 4: Performance scores for the three different rainfall products in simulating daily river discharge in Brazilian catchments in the calibration phase. a), b) and c) are the mean bias (BIAS-Q) for SM2RASC, GPMSM2R, and ERA5, respectively. d), e) and f) are the root mean squared error (RMSE-Q) for SM2RASC, GPMSM2R, and ERA5, respectively. g), h) and i) are the Nash-Sutcliffe Efficiency (NSE-Q) for SM2RASC, GPMSM2R, and ERA5, respectively. j), k) and l) are the Kling-Gupta Efficiency (KGE-Q) for SM2RASC, GPMSM2R, and ERA5, respectively.

In Figure 5, the results of performance scores on the validation period (carried from apr-2014 to dec-2018) from the hydrological modeling are shown. The same color scheme from the calibration period was kept for a direct comparison between modeling phases. The BIAS-Q from all rainfall products presents higher values than the calibration period, keeping the characteristic of underestimation in the southern portion (in the Pampa biome). SM2RASC presents mBIAS-Q value of 0.178 mm day^{-1}, while GPMSM2R and ERA5 present values of 0.156 mm day^{-1}, and 0.152 mm day^{-1}, respectively. Although median values indicate an overall overestimation for satellite rainfall products, there is high spatial variability across the whole country. Figures 5d, 5e, and 5f show that the validation period presented similar spatial patterns of the calibration period for Brazilian catchments, with the highest values in the southern portion. The values of mRMSE-Q are close to those obtained in the calibration period, being 0.888 mm day^{-1}, 0.708 mm day^{-1}, and 1.026 mm day^{-1}, for SM2RAIN, GPMSM2R, and ERA5, respectively. NSE-Q from the ERA5 product (mNSE-Q = 0.161) is considerably lower than those obtained for the daily river discharge estimated with satellite rainfall as inputs. GMPSM2R shows the best performance, with 40% of the catchments with NSE > 0.5 and mNSE-Q = 0.418. SM2RASC and ERA5 present similar results, with mNSE-Q = 0.204 and mNSE-Q = 0.161, respectively, and 21% of the catchments with NSE-Q > 0.5. In terms of KGE, satellite rainfall products also perform better than the ERA5 rainfall product. Employing the SM2RASC as input for modeling daily river discharge, we obtain 52% of the catchments with KGE > 0.5 in the validation period, with mKGE-Q = 0.510. GPMSM2R is slightly better, with mKGE-Q = 0.574 and 61% of the catchments with KGE-Q > 0.5. The worst performance, overall, is achieved by the ERA5 product (mKGE-Q = 0.487), even though 48% of the catchments with KGE-Q > 0.5.

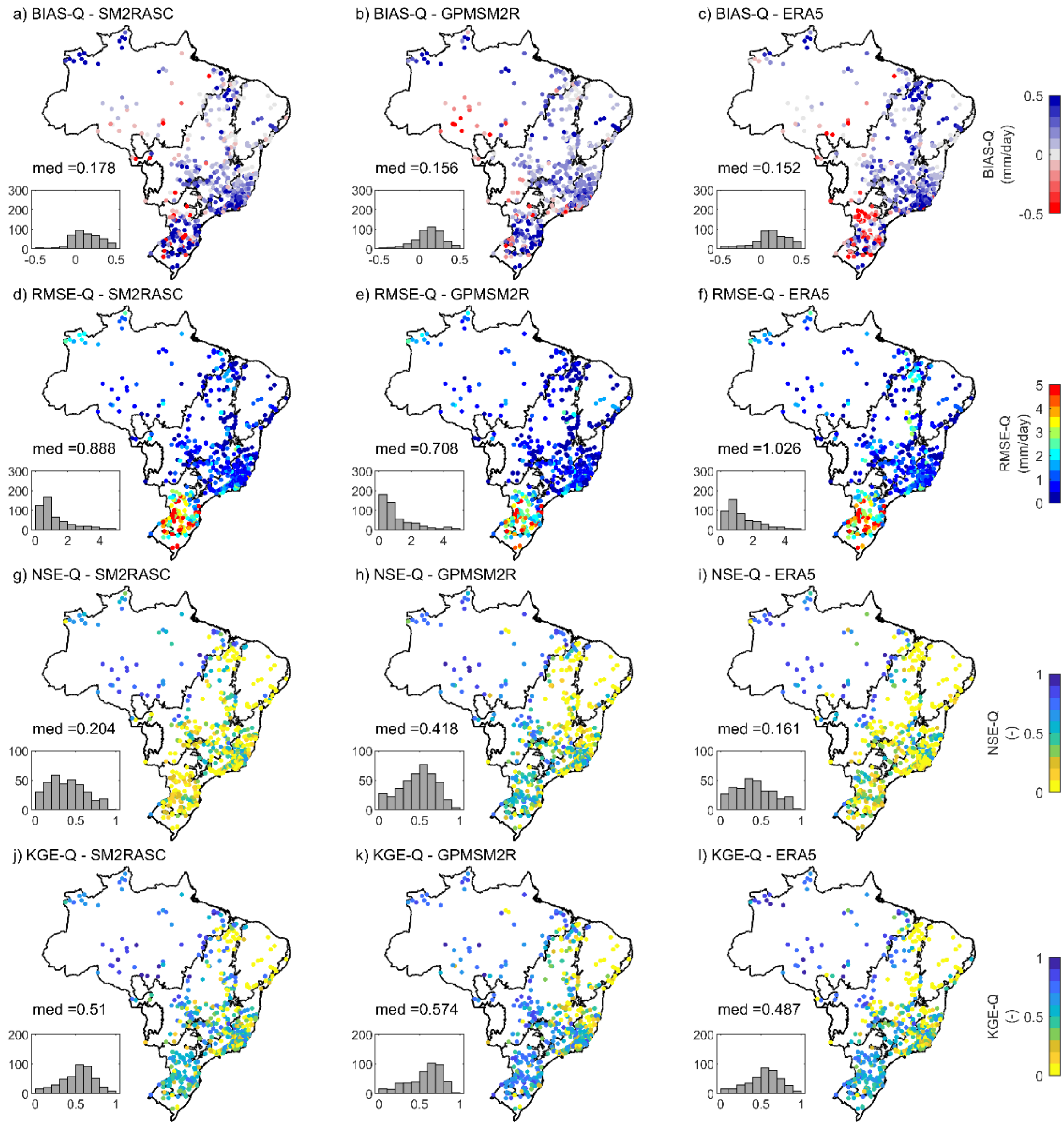

Figure 5: Performance scores for the three different rainfall products in simulating daily river discharge in Brazilian catchments in the validation phase. a), b) and c) are the mean bias (BIAS-Q) for SM2RASC, GPMSM2R, and ERA5, respectively. d), e) and f) are the root mean squared error (RMSE-Q) for SM2RASC, GPMSM2R, and ERA5, respectively. g), h) and i) are the Nash-Sutcliffe Efficiency (NSE-Q) for SM2RASC, GPMSM2R, and ERA5, respectively. j), k) and l) are the Kling-Gupta Efficiency (KGE-Q) for SM2RASC, GPMSM2R, and ERA5, respectively.

As expected, the general performance scores of the hydrological modeling in simulating the river discharge in Brazilian catchments decrease from the calibration to the validation period, as illustrated in Figure 6. The median BIAS-Q of the ERA5 product slightly increases, while the satellite rainfall products present a decrease. It is also clear to note that the range of the BIAS-Q is considerably increased in the validation period. The range of RMSE-Q is slightly increased in the validation period for all products analyzed, while the median of RMSE-Q (mRMSE-Q) is diminished for ERA5 and SM2RASC and improved for GPMSM2R. NSE-Q and KGE-Q present a significant reduction of their values, indicating that hydrological modeling has a great dependency on the observed data. Important to note that we used a relatively short series for calibration and validation, and the results might be better with longer series. Nonetheless, median values of KGE-Q are mostly larger than 0.5.

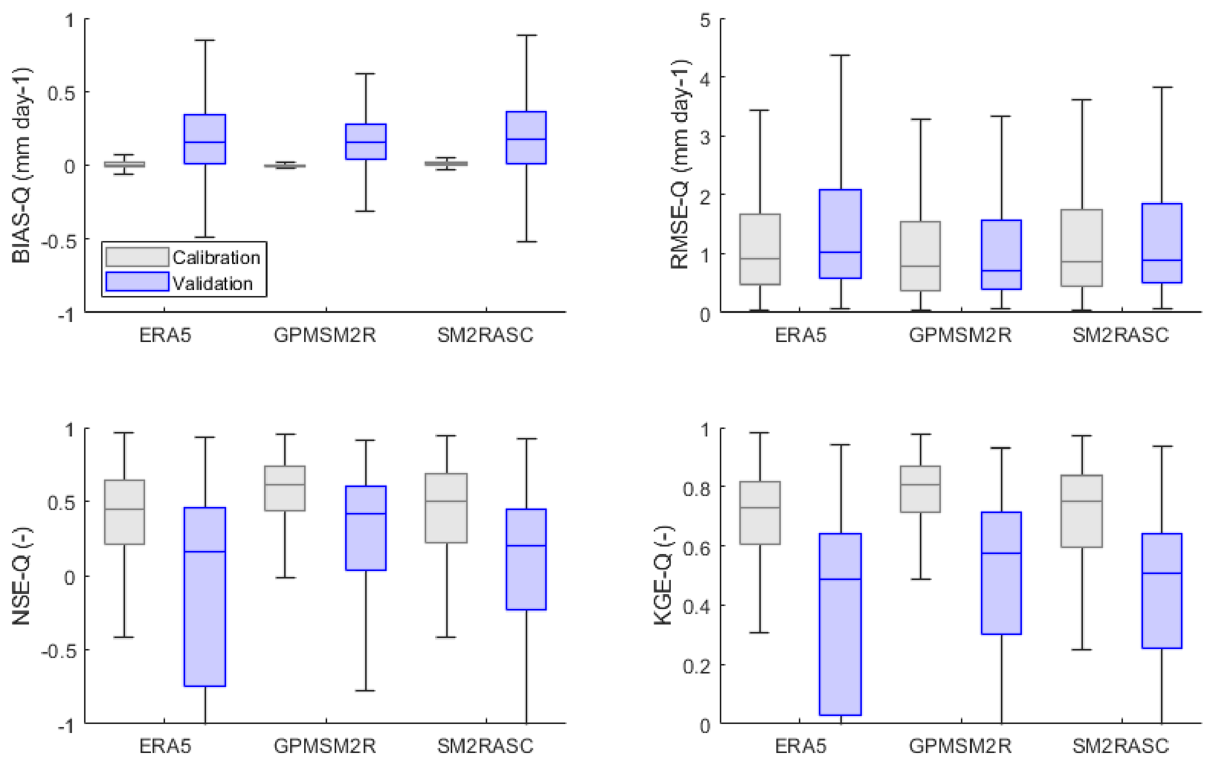

Figure 6: Box charts for comparison between calibration and validation phases of hydrological modeling using the MISDc model and the three different rainfall products. a) BIAS-Q results for hydrological modeling; b) RMSE-Q results for hydrological modeling; c) NSE-Q results for hydrological modeling; d) KGE-Q results for hydrological modeling.

Even though the performance of the daily river discharge modeling is not satisfactory for all the catchments analyzed here, the satellite rainfall products showed their ability to be a reliable tool in different hydroclimatic conditions of Brazil. In Figure 7 we show the KGE as a function of the catchment size through our set o 520 catchments. For the subset of catchments with areas smaller than 1,000 km, many of them perform very well (KGE-Q > 0.5), with a small group presenting negative values. The same pattern is noted for the catchments with areas varying between 10,000 km² and 100,000 km², while for those with a range of 1,000 km² - 10,000 km² some catchments are presenting extremely poor performance (KGE-Q < 0). For the group of large catchments (>100,000 km²), we can note that a larger percentage presents negative values of KGE, indicating that the model was not capable of capturing the main features of streamflow generation in those catchments. Even though, most of these catchments present good performance. As in Jiang and Bauer-Gottwein (2019), we also did not find a clear pattern between catchment size and model/product performance.

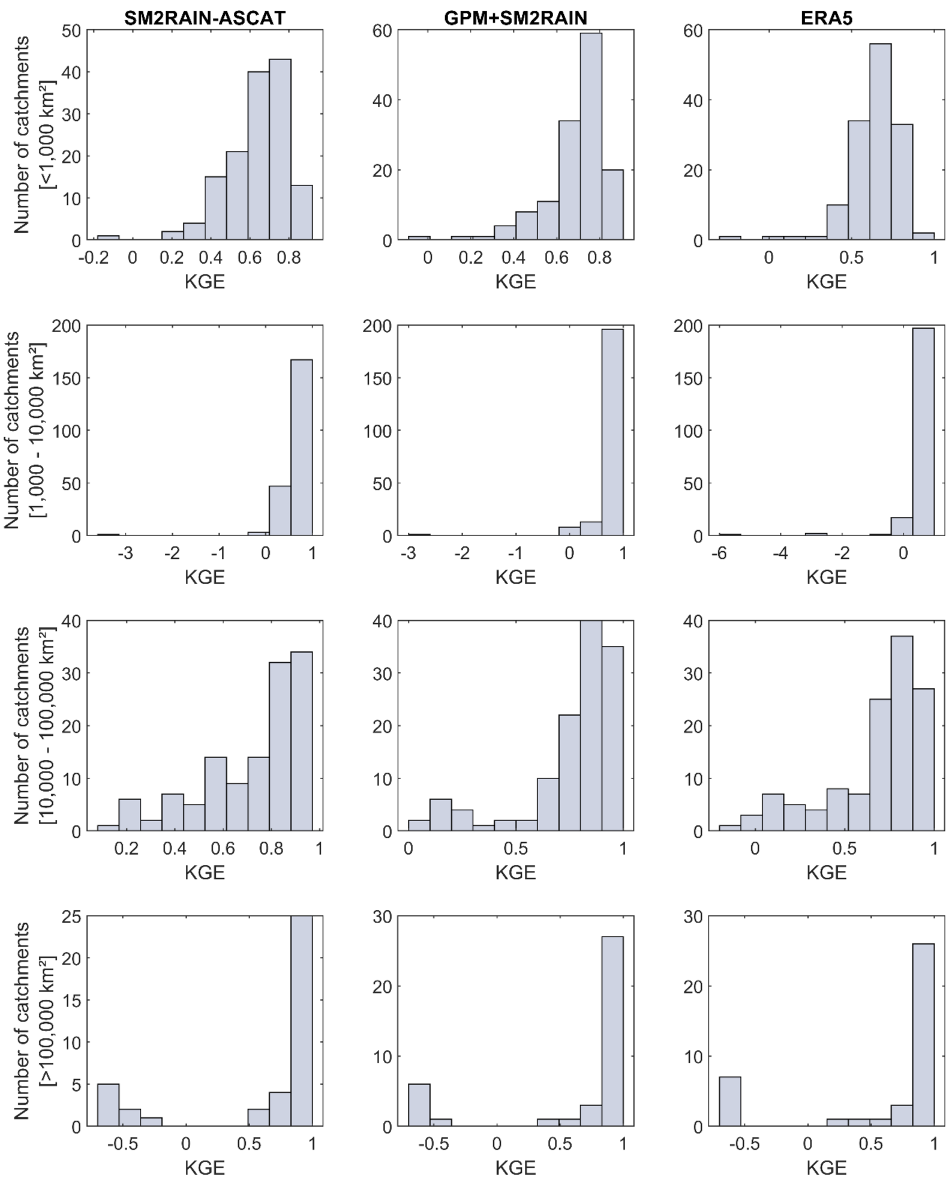

Figure 7: Performance scores, in terms of KGE, during the calibration period for the SM2RAIN-ASCAT, GPM+SM2RAIN, and ERA5, with varying areas of the catchment. The histograms are representing the number of catchments with <1,000 km², 1,000 – 10,000 km², 10,000 – 100,000 km², and >100,000 km².

Four specific examples of catchments with distinct characteristics of rainfall, temperature, and river discharge regimes are shown in Figure 8. Figures 8a and 8b are related to the “Óbidos” streamflow monitoring, a catchment with 4,800,000 km^2 in the Amazon river basin, with large amounts of annual precipitation and low interannual variability of precipitation. As can be seen, the MISDc model showed a good agreement with all the products (KGE-Q > 0.88), and this was carried for the validation period, with KGE-Q > 0.77. Modeling results for the “Araguantins” station are shown in Figures 8c and 8d. This station is in the Tocantins-Araguaia river basin, in the Cerrado biome, which has a great impact on evapotranspiration and groundwater withdraws for irrigations purposes. Our results demonstrate the high reliability of both satellite rainfall products and hydrological model to simulate daily river discharge under these conditions, with KGE > 0.5 for calibration and validation periods. Figures 8e and 8f, in turn, are related to the “Vila Urucuia” monitoring station, in the Cerrado biome, which has a well-defined rainy/dry season, with long periods of streamflow recession. In this catchment, GPMSM2R performs better than SM2RASC and ERA5 in the calibration, with KGE-Q = 0.90, while SM2RASC, with KGE-Q = 0.77, performs better in the validation. The last catchment results are shown in Figures 8g and 8h for the “Campos (Ponte Municipal)” streamflow gauge, related to the Paraíba do Sul river basin. This catchment is especially important due to its role as a water supplier for the most populous regions of Brazil, including the metropolitan areas of São Paulo and Rio de Janeiro, providing water for more than 14 million people. The calibration period generates much better values of KGE-Q than validation, but it is still possible to estimate the river discharge with reliability in the validation period with remote sensing products (KGE-Q > 0.56). Simulations with SM2RASC and GPMSM2R are also able to capture the 2014/2015 megadrought in São Paulo (Escobar, 2015), with signs of decreasing river discharge since previous years. This fact reinforces the importance of hydrologic modeling using remote sensing products to support water management agents.

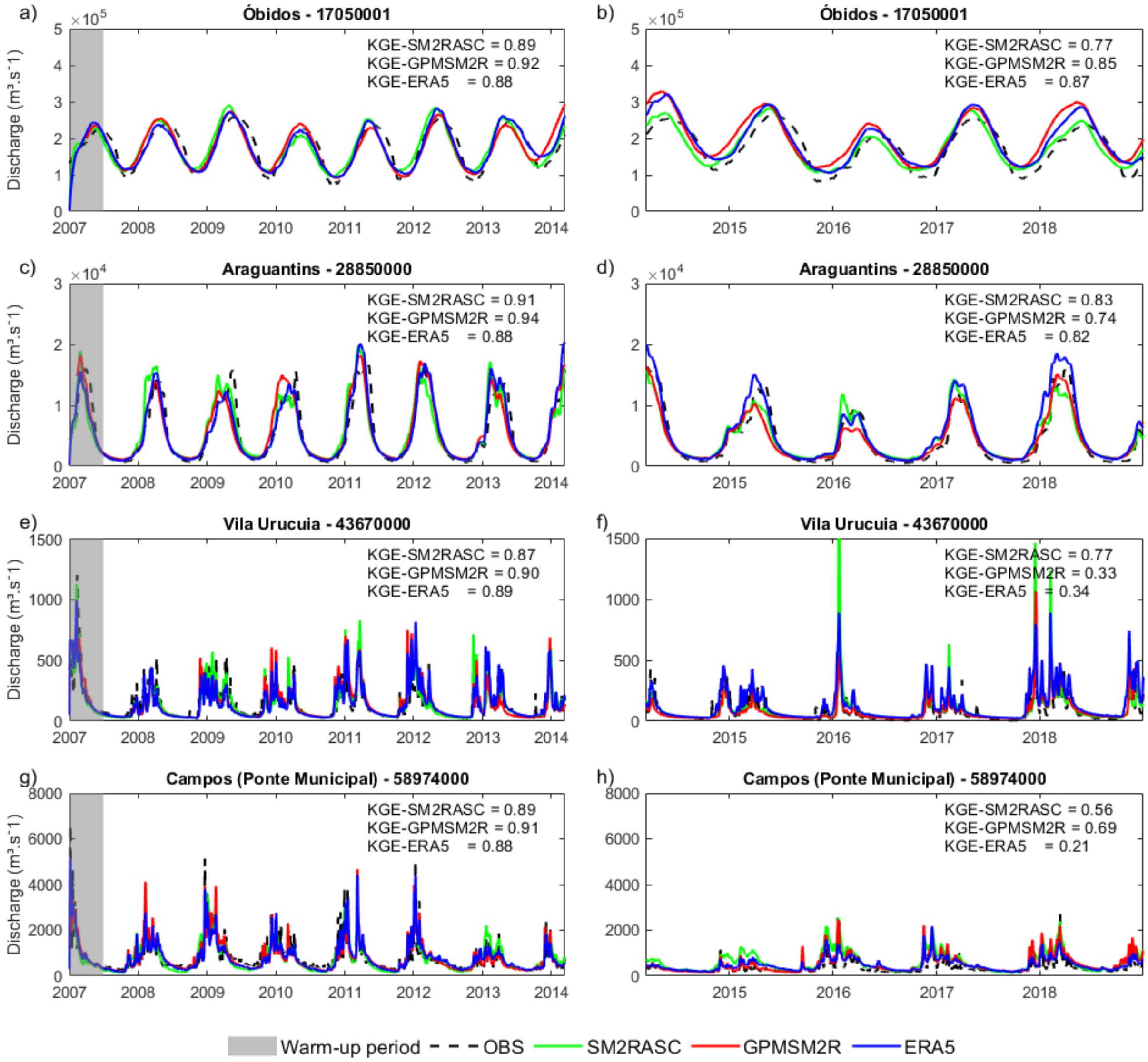

Figure 8: Daily river discharge simulated during calibration and validation phases by MISDc with SM2RASC (red line), GPMSM2R (blue line), and ERA5 (green line) rainfall products as inputs. The Black dashed line is the observed daily river discharge. a) and b) Daily river discharge for "Óbidos" streamflow gauge. c) and d) Daily river discharge for "Araguantins" streamflow gauge. e) and f) Daily river discharge for the "Vila Urucuia" streamflow gauge. g) and h) Daily river discharge for "Campos (Ponte Municipal)" streamflow gauge.

Our results are in accordance with those found by Camici et al. (2020), for European basins. Also using the MISDc and the SM2RASC as the precipitation input, they found a mKGE-Q = 0.722 and mKGE-Q = 0.569 for a subset of 1318 catchments, in the calibration and validations periods, respectively. In turn, our results for calibration and validation are mKGE-Q = 0.748 and mKGE-Q = 0.510 for the SM2RASC as input in hydrological modeling. Even though we cannot directly compare the performance score results due to methodological differences, it is possible to correlate the good performance of the GPMSM2R in this study and in that conducted for a subset of African basins (Brocca et al., 2020), in which this rainfall product outperformed other rainfall products in estimating daily river flow discharge, with mKGE-Q > 0.8. In mNSE-Q terms, our results of SM2RASC performance are slightly worse than those found by Camici et al. (2018) for Italy, but we have used a larger subset of catchments and a longer time-series during calibration and validation. Although there is a clear advantage of forcing the hydrological model with the ground-based rainfall product against satellite and reanalysis rainfall products in catchments in southern and southeastern Brazil, it is clear the added value of the satellite rainfall products in simulating/estimating the daily river in regions with low density of rain gauge stations, such as northern Brazil (Amazon), discharge by observing the skill scores results, presented in Figure 9. When comparing only the non-ground-based data, bottom-up satellite rainfall products present more reliability on estimating river discharge than reanalysis data, as we can see in Figure 8d, wherein 383 catchments is better to use SM2RAIN-ASCAT (260 catchments) or GPM+SM2RAIN (123 catchments) products than the ERA5, and 137 catchments in which is preferable to employ the ERA5 data. Aside from our results, it is possible to find robust performance of these products over different areas of the globe, with distinct climatic and geophysical conditions (such as Brazil, Africa and Europe), allowing a reliable and continuous river discharge monitoring.

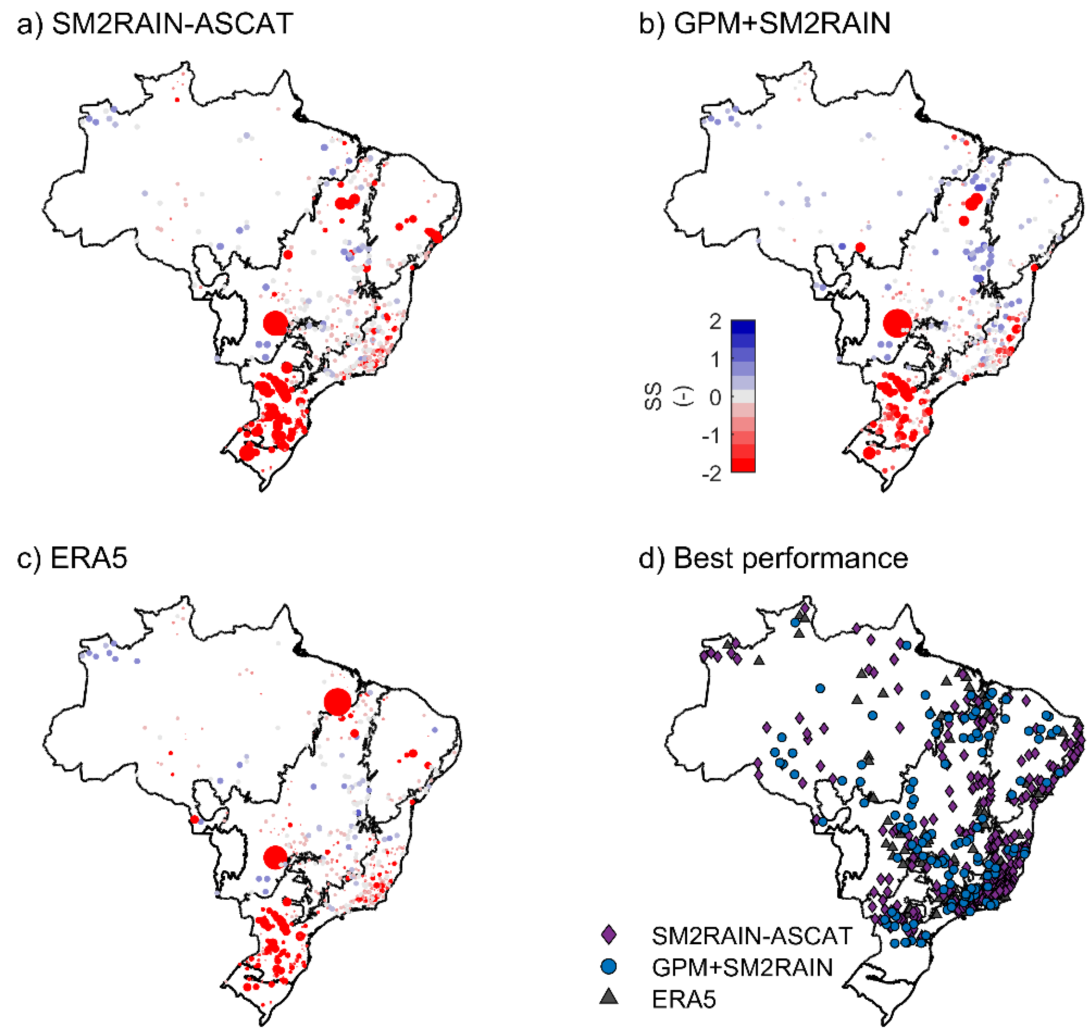

Figure 9: Performance of the satellite and reanalysis rainfall products against ground-based rainfall products during the calibration period. a) to c) are the results of the Skill Score (SS), dimensionless, of the SM2RAIN-ASCAT, GPM+SM2RAIN, and ERA5 rainfall products, respectively, against Xavier et al. (2016) dataset. d) shows the best performance between SM2RAIN-ASCAT, GPM+SM2RAIN, and ERA5 in estimating daily river discharge in Brazilian catchments.

5.3. Reliability of remote sensing products on hydrologic signatures

The reliability of the ground-based, SM2RASC, GPMSM2R, and ERA5 rainfall products to simulate hydrologic signatures in Brazilian biomes in terms of coefficient of determination is shown in Figure 10, in terms of determination coefficient between observed and estimated signatures for all catchments in each biome. Additionally, we present a qualitative analysis based on a performance ranking, on the right side. We highlight here the good ability of products to estimate the hydrologic signatures (Mean Q, Q5, and Q95) related to the magnitude of river discharge. Frequency and duration of low-flows and high-flows are poorly simulated by rainfall products through BFI and HFD are well represented, while the FDC slope and 7-day minimum Q are not correctly represented. From the knowledge that high/low flows are, in some cases, more important than normal flows, it is important to note that the model estimations did not perform satisfactorily for all the catchments and the employment of the MISDc model combined with the rainfall products (ground-based, satellite, or reanalysis) should be carefully taken. This is probably due to the lack of streamflow generation process by the model, which was developed for different physical and geomorphological conditions of Brazilian catchments. Overall, the satellite-based products SM2RASC and GPMSM2R are better predictors for hydrologic signatures than counterpart ground-based for the Atlantic Forest, Cerrado, and Caatinga biomes. On the other hand, ground-based rainfall product performs better in the Pampa, where there is no well-defined seasonal cycle. The performance between products was similar in the Amazon biome, with the satellite-based rainfall products being better predictors for mean and low-flows signatures and the ground-based rainfall product being a better predictor for high-flows, baseflow, and timing of river discharge. The ERA5 rainfall product presented a satisfactory performance, achieving better results than ground-based products in Atlantic Forest, Cerrado and Caatinga biomes.

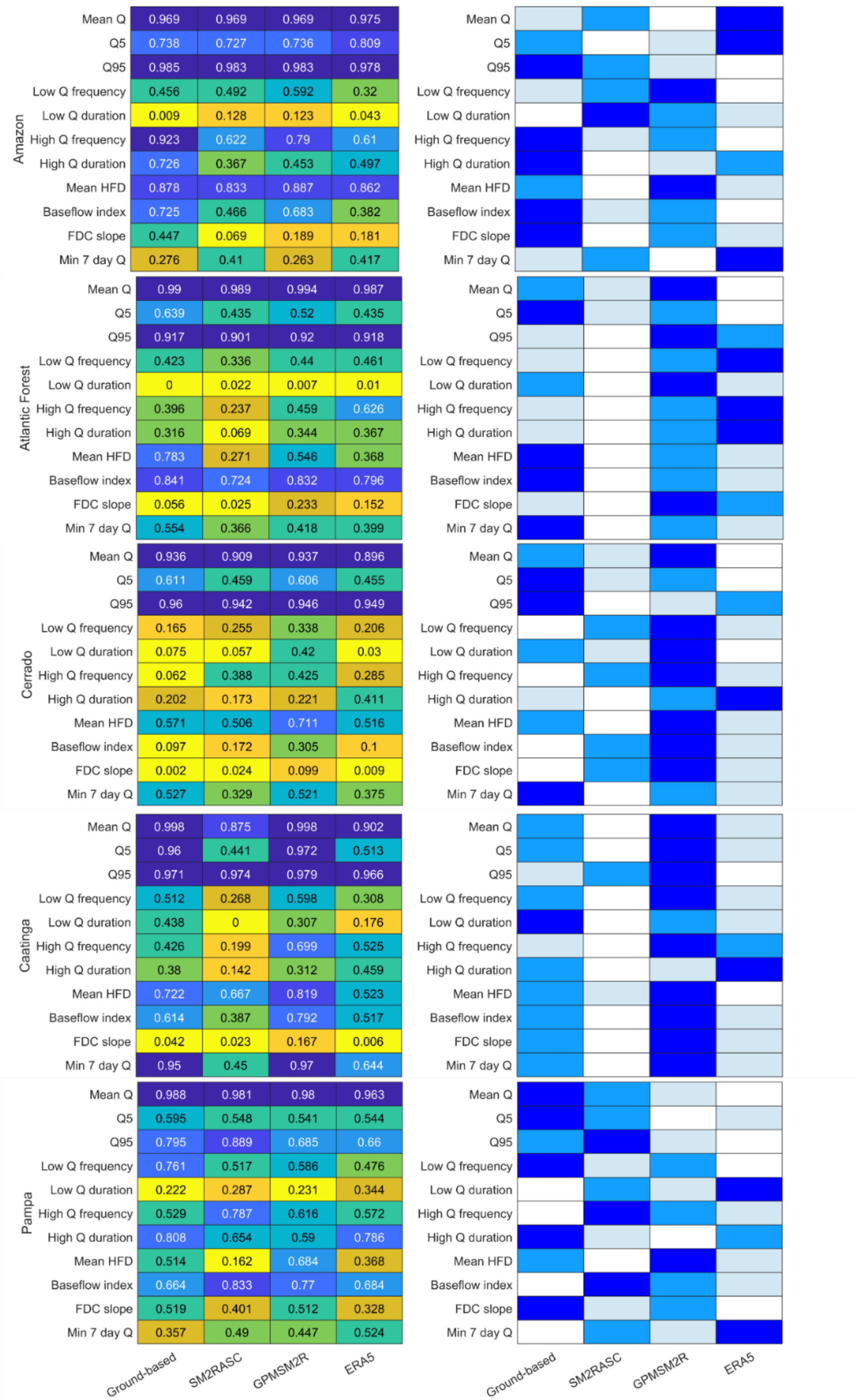

Figure 10: Hydrologic signatures from daily river discharge simulated by MISDc with ground-based, SM2RASC, GPMSM2R, and ERA5 rainfall products as inputs. The panel on the left side represents the ability of the rainfall product to represent each hydrologic signature, in terms of coefficient of determination, for a given biome. The panel on the right side represents a rainfall product ranking of ability to represent the hydrologic signature in each biome.

6. Conclusions

In this study, we evaluated the performance of three rainfall products in estimating daily precipitation over Brazil: the satellite-based "bottom-up" SM2RAIN-ASCAT, the satellite-based "bottom-up" plus "top-down" GPM+SM2RAIN, and the reanalysis-based ERA5. Additionally, we assessed the suitability of these products in simulating the daily river discharge and hydrologic signatures in Brazilian catchments. We used a subset of 520 catchments from the CABra dataset and the hydrologic modeling was carried out using the MISDc hydrologic model. The performance evaluation was executed by the employment of statistical scores such as the bias, root mean squared error, Nash-Sutcliffe efficiency, and Kling-Gupta efficiency.

The performance of the satellite-based products in simulating daily precipitation over Brazil was better than ERA5 for five of six biomes. SM2RAIN-ASCAT presented the lowest bias values, while GPM+SM2RAIN presented the lowest values of RMSE. Daily river discharge was also better modeled with SM2RAIN-ASCAT and GPM+SM2RAIN in terms of KGE during calibration. All products presented low values of bias and more than 82% of catchments have median values of KGE greater than 0.5. Performance scores were slightly worsened in the validation but are still able to estimate daily river discharge with great accuracy (mKGE > 0.5). Finally, satellite rainfall products performed well in estimating hydrologic signatures of Brazilian catchments, being better predictors than ground-based observations for Atlantic Forest, Cerrado, and Caatinga biomes, aside from being as good as ground-based rainfall product in the Amazon.

Our study shows that there is an added value when using products like SM2RAIN-ASCAT and GPM+SM2RAIN in tropical catchments like those in Brazil. The employment of near real-time satellite-based rainfall products in hydrologic extremes analysis, such as floods and droughts, improves the capacity of impact mitigation of these extremes. Irrigation on agricultural lands could be done with sustainable use of water resources, social and economic impacts caused by floods could be mitigated, and a better understanding of water fluxes could be achieved, on multiple spatial and temporal scales. Definitively, satellite-based rainfall products are a valuable tool for data-scarce regions, due to their low-latency and global-land range, allowing continuous and high-quality water resources monitoring.

7. Acknowledgments

This study was supported by grants from the Ministry of Science, Technology, and Innovation – MCTI and National Council for Scientific and Technological Development – CNPq [grants numbers 441289/2017-7 and 309752/2020-5]. This study was also financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and CAPES Print. The authors gratefully acknowledge support from EUMETSAT through the Global SM2RAIN project (contract no. EUM/CO/17/4600001981/BBo) and the "Satellite Application Facility on Support to Operational Hydrology and Water Management (H SAF)" CDOP 3 (grant no. EUM/C/85/16/DOC/15) and from the European Space Agency through the project SMOS+Rainfall (contract no. 4000114738/15/I-SBo).

8. References

Almagro, A., Oliveira, P.T.S., Meira Neto, A.A., Roy, T., Troch, P., 2021. CABra: a novel large-sample dataset for Brazilian catchments. Hydrol. Earth Syst. Sci. 25, 3105–3135. https://doi.org/10.5194/hess-25-3105-2021

Almagro, A., Oliveira, P.T.S., Nearing, M.A., Hagemann, S., 2017. Projected climate change impacts in rainfall erosivity over Brazil. Sci. Rep. 7, 1–12. https://doi.org/10.1038/s41598-017-08298-y

Almagro, A., Oliveira, P.T.S., Rosolem, R., Hagemann, S., 2020. Performance evaluation of Eta/HadGEM2-ES and Eta/MIROC5 precipitation simulations over Brazil. Atmos. Res. 244, 105053.

Avila-Diaz, A., Benezoli, V., Justino, F., Torres, R., Wilson, A., 2020. Assessing current and future trends of climate extremes across Brazil based on reanalyses and earth system model projections. Clim. Dyn. 55, 1403–1426. https://doi.org/10.1007/s00382-020-05333-z

Barrett, E.C., Beaumont, M.J., 1994. Satellite rainfall monitoring: an overview. Remote Sens. Rev. 11, 23–48. https://doi.org/10.1080/02757259409532257

Bishop, C.H., Abramowitz, G., 2013. Climate model dependence and the replicate Earth paradigm. Clim. Dyn. 41, 885–900. https://doi.org/10.1007/s00382-012-1610-y

Brocca, L., Ciabatta, L., Massari, C., Moramarco, T., Hahn, S., Hasenauer, S., Kidd, R., Dorigo, W., Wagner, W., Levizzani, V., 2014. Soil as a natural rain gauge: Estimating global rainfall from satellite soil moisture data. J. Geophys. Res. Atmos. 119, 5128–5141. https://doi.org/10.1002/2014JD021489

Brocca, L., Filippucci, P., Hahn, S., Ciabatta, L., Massari, C., Camici, S., Schüller, L., Bojkov, B., Wagner, W., 2019. SM2RAIN-ASCAT (2007-2018): Global daily satellite rainfall data from ASCAT soil moisture observations. Earth Syst. Sci. Data 11, 1583–1601. https://doi.org/10.5194/essd-11-1583-2019

Brocca, L., Massari, C., Pellarin, T., Filippucci, P., Ciabatta, L., Camici, S., Kerr, Y.H., Fernández-Prieto, D., 2020. River flow prediction in data scarce regions: soil moisture integrated satellite rainfall products outperform rain gauge observations in West Africa. Sci. Rep. 10, 1–14. https://doi.org/10.1038/s41598-020-69343-x

Brocca, L., Melone, F., Moramarco, T., 2011. Distributed rainfall-runoff modelling for flood frequency estimation and flood forecasting. Hydrol. Process. 25, 2801–2813. https://doi.org/10.1002/hyp.8042

Brunetti, M.T., Melillo, M., Gariano, S.L., Ciabatta, L., Brocca, L., Peruccacci, S., 2021. Satellite rainfall products outperform ground observations for landslide forecasting in India. Hydrol. Earth Syst. Sci. Discuss. https://doi.org/https://doi.org/10.5194/hess-2021-42

Camici, S., Ciabatta, L., Massari, C., Brocca, L., 2018. How reliable are satellite precipitation estimates for driving hydrological models: A verification study over the Mediterranean area. J. Hydrol. 563, 950–961. https://doi.org/10.1016/j.jhydrol.2018.06.067

Camici, S., Massari, C., Ciabatta, L., Marchesini, I., Brocca, L., 2020. Which rainfall score is more informative about the performance in river discharge simulation? A comprehensive assessment on 1318 basins over Europe. Hydrol. Earth Syst. Sci. 24, 4869–4885. https://doi.org/10.5194/hess-24-4869-2020

Crow, W.T., Huffman, G.J., Bindlish, R., Jackson, T.J., 2009. Improving satellite-based rainfall accumulation estimates using spaceborne surface soil moisture retrievals. J. Hydrometeorol. 10, 199–212. https://doi.org/10.1175/2008JHM986.1

Crow, W.T., Ryu, D., 2009. A new data assimilation approach for improving runoff prediction using remotely-sensed soil moisture retrievals. Hydrol. Earth Syst. Sci. 13, 1–16. https://doi.org/10.5194/hess-13-1-2009

Crow, W.T., Zhan, X., 2007. Continental-scale evaluation of remotely sensed soil moisture products. IEEE Geosci. Remote Sens. Lett. 4, 451–455. https://doi.org/10.1109/LGRS.2007.896533

Escobar, H., 2015. Drouth triggers alarms in Brazil’s biggest metropolis. Science (80-.). 347, 812.

Gnann, S.J., Coxon, G., Woods, R.A., Howden, N.J.K., McMillan, H.K., 2021. TOSSH: A Toolbox for Streamflow Signatures in Hydrology. Environ. Model. Softw. 138, 104983. https://doi.org/10.1016/j.envsoft.2021.104983

Gupta, H. V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377, 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., Thépaut, J.N., 2020. The ERA5 global reanalysis. Q. J. R. Meteorol. Soc. 146, 1999–2049. https://doi.org/10.1002/qj.3803

Huffman, G.J., Adler, R.F., Bolvin, D.T., Gu, G., Nelkin, E.J., Bowman, 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

Jiang, L., Bauer-Gottwein, P., 2019. How do GPM IMERG precipitation estimates perform as hydrological model forcing? Evaluation for 300 catchments across Mainland China. J. Hydrol. 572, 486–500. https://doi.org/10.1016/j.jhydrol.2019.03.042

Lucas, M.C., Kublik, N., Rodrigues, D.B.B., Meira Neto, A.A., Almagro, A., Melo, D. de C.D., Zipper, S.C., Oliveira, P.T.S., 2021. Significant baseflow reduction in the São Francisco river basin. Water (Switzerland). https://doi.org/10.3390/w13010002

Maggioni, V., Massari, C., 2018. On the performance of satellite precipitation products in riverine flood modeling: A review. J. Hydrol. 558, 214–224.

https://doi.org/10.1016/j.jhydrol.2018.01.039

Massari, C., Brocca, L., Pellarin, T., Abramowitz, G., Filippucci, P., Ciabatta, L., Maggioni, V., Kerr, Y., Fernandez Prieto, D., 2020. A daily 25 km short-latency rainfall product for data-scarce regions based on the integration of the Global Precipitation Measurement mission rainfall and multiple-satellite soil moisture products. Hydrol. Earth Syst. Sci. 24, 2687–2710. https://doi.org/10.5194/hess-24-2687-2020

Melo, D. de C.D., Scanlon, B.R., Zhang, Z., Wendland, E., Yin, L., 2016. Reservoir storage and hydrologic responses to droughts in the Paraná River basin, south-eastern Brazil. Hydrol. Earth Syst. Sci. 20, 4673–4688. https://doi.org/10.5194/hess-20-4673-2016

Melo, D.D.C.D., Xavier, A.C., Bianchi, T., Oliveira, P.T.S., 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

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. J. Hydrol. 10, 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Paredes-Trejo, F., Barbosa, H., 2017. Evaluation of the SMOS-derived soil water deficit index as agricultural drought index in Northeast of Brazil. Water 9. https://doi.org/10.3390/w9060377

Paredes-Trejo, F., Barbosa, H., dos Santos, C.A.C., 2019. Evaluation of the performance of SM2RAIN-derived rainfall products over Brazil. Remote Sens. 11, 1–28. https://doi.org/10.3390/rs11091113

Paredes-Trejo, F., Barbosa, H.A., Spatafora, L.R., 2018. Assessment of SM2RAIN-derived and state-of-the-art satellite rainfall products over Northeastern Brazil. Remote Sens. 10. https://doi.org/10.3390/rs10071093

Paredes-trejo, F.J., Barbosa, H.A., Kumar, T.V.L., 2017. Validating CHIRPS-based satellite precipitation estimates in Northeast Brazil. J. Arid Environ. 139, 26–40. https://doi.org/10.1016/j.jaridenv.2016.12.009

Pellarin, T., Louvet, S., Gruhier, C., Quantin, G., Legout, C., 2013. A simple and effective method for correcting soil moisture and precipitation estimates using AMSR-E measurements. Remote Sens. Environ. 136, 28–36. https://doi.org/10.1016/j.rse.2013.04.011

Pellarin, T., Tran, T., Cohard, J.M., Galle, S., Laurent, J.P., De Rosnay, P., Vischel, T., 2009. Soil moisture mapping over West Africa with a 30-min temporal resolution using AMSR-E observations and a satellite-based rainfall product. Hydrol. Earth Syst. Sci. 13, 1887–1896. https://doi.org/10.5194/hess-13-1887-2009

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P.A., Carrillo, G., 2011. Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA. Hydrol. Earth Syst. Sci. 15, 2895–2911. https://doi.org/10.5194/hess-15-2895-2011

Souza, R., Feng, X., Antonino, A., Montenegro, S., Souza, E., Porporato, A., 2016. Vegetation response to rainfall seasonality and interannual variability in tropical dry forests. Hydrol. Process. 30, 3583–3595. https://doi.org/10.1002/hyp.10953

Wagner, P.D., Fiener, P., Wilken, F., Kumar, S., Schneider, K., 2012. Comparison and evaluation of spatial interpolation schemes for daily rainfall in data scarce regions. J. Hydrol. 464–465, 388–400. https://doi.org/10.1016/j.jhydrol.2012.07.026

Westerberg, I.K., McMillan, H.K., 2015. Uncertainty in hydrological signatures. Hydrol. Earth Syst. Sci. 19, 3951–3968. https://doi.org/10.5194/hess-19-3951-2015

Xavier, A.C., King, C.W., Scanlon, B.R., 2016. Daily gridded meteorological variables in Brazil (1980-2013). Int. J. Climatol. 2659, 2644–2659. https://doi.org/10.1002/joc.4518

CHAPTER 4

THE DOMINANT ATTRIBUTES OF STREAMFLOW VARIABILITY ON BRAZILIAN CATCHMENTS

Almagro, A., Oliveira, P. T. S., Meira Neto, A. A., Roy, T., and Troch, P. A.: Dominant attributes of streamflow variability on Brazilian catchments [in preparation to the Special Issue "Processes and Patterns in Tropical Hydrology" of Hydrological Processes]. (Impact Factor 2021: 3.565)

Abstract

Although almost 20% of global freshwater flows in rivers, the catchment-scale relationships between drivers and flow generation are still unknown in many parts of the world. Large-sample catchment-scale hydrological behaviors allow us to identify similarities, trends, and dominant processes. Here, we seek to understand which attributes control the streamflow variability in Brazilian catchments, how the catchments group based on hydrologic behavior similarities, and what are the dominant hydrological processes over the different groups. We used hydrological signatures and attributes of 735 catchments from the Catchment Attributes for Brazil (CABra) large-sample dataset. By employing machine learning algorithms, we grouped the catchments based on hydrological similarities and investigated the dominant processes and their driving attributes. Our results identified six groups of similar catchments: "non-seasonal", "dry", "rainforest", "savannah", "extremely-dry", and "extremely-wet" catchments. Aridity index is the main driver of streamflow for four groups, highlighting the influence and importance of land-atmosphere water flux in tropical catchments. Deep and porous soils, combined with low precipitation, led catchments from the "extremely-dry" group to not reach saturation and streamflow generation most of the year, resulting in ephemeral catchments. In contrast, high soil storage capacity in "non-seasonal" and "rainforest" catchments associated with high precipitation led to high streamflow discharge all year due to the subsurface fluxes' contribution. Our study contributes to improving the streamflow predictability and hydrological behavior identification by further understanding the hydrological similarities and their signatures due to catchment landscape characteristics.

Keywords: catchment functioning; hydrological similarity; clustering analysis; streamflow.

1. Introduction

Catchments are considered complex poorly defined systems that present great variability in space, time, and processes (Beven, 2000; McDonnell and Woods, 2004). This variability and its consequences are described as the uniqueness of place (Beven, 2000), which considers that a catchment is unique due to its landscape characteristics (such as topography, soils, geology, vegetation, and anthropogenic modification), in a way that is difficult to regionalize hydrological behaviors. Aside from the above-mentioned, we can consider that catchments present at least some level of self-organization, in where its geomorphology, soils, and vegetation are adaptive to (and a result of) the landscape co-evolution (Dooge, 1986; Blöschl and Sivapalan, 1995; Sivapalan, 2005; Troch *et al.*, 2013).

The identification and classification of the catchment's hydrological behavior is the first step towards a better understanding of many complex levels involved in catchment hydrological processes (McDonnell and Woods, 2004). One of its main goals is to provide insights into hydrological behavior and functioning similarities (or dissimilarities), connections, and relationships between different catchments (Wagener *et al.*, 2007). Deciphering meaningful patterns in observations inevitably relies on a catchment classification capable of predicting the dominant controls on the water fluxes (Sivapalan, 2005). Sivakumar *et al.* (2013) well-stated the fundamental idea of the catchment classification is to group catchments by their salient and emblematic characteristics and develop from them suitable methodologies for catchment-related purposes, such as prediction, extremes hazard assessments, and theory development. Moreover, it provides structure to hydrology science (Wagener *et al.*, 2008).

Although there is not a commonly agreed-upon catchment classification or grouping system by the hydrologic science society (McDonnell and Woods, 2004; Wagener *et al.*, 2007), some directions and starting points have been proposed by Wagener et al. (2007). The classification must (i) map the catchment geophysical conditions across spatiotemporal scales; catchment functions related to the partition, storage, and release of the water entering the catchment must be considered; (ii) it must be reproducible, the catchment functions should initially be based on the streamflow; and (iii) it must weigh the uncertainty of the metrics adopted for classification. Based on this, it is possible to achieve a reliable catchment classification system capable of limiting all the internal complexity and spatial variability within determined classes (McDonnell and Woods, 2004).

To classify and group catchments to further investigate their hydrological behavior, data collection is the key (Beven, 2000). The more we can collect, organize, and process hydrological data, the further we reach in hydrological science. Therefore, integrated assessment of large-sample datasets is essential to a better understand catchment functioning and classification (Almagro *et al*., 2021). In recent years, a set of large-sample datasets emerged from the need for a large amount of data to construct general inferences that would be impossible by analyzing individual or small groups of catchments (Wagener *et al.*, 2007; Lyon and Troch, 2010; Addor *et al.*, 2017). The forerunner of the large-sample datasets is the MOPEX (Duan *et al.*, 2006) for the U.S. catchments, which have been extensively used in the hydrology field. Later, based on the MOPEX dataset, Addor *et al.* (2017) developed the widely-used CAMELS dataset, relying on a set of 673 catchments in the U.S. In turn, the CAMELS dataset was the inspiration and base for other large-sample datasets across the globe, such as in Chile (Alvarez-Garreton *et al.*, 2018), Great Britain (Coxon *et al.*, 2020), Australia (Fowler *et al.*, 2021), China (Hao *et al.*, 2021), and Brazil (Chagas *et al.*, 2020; Almagro *et al.*, 2021). Recently, Almagro *et al.* (2021) made available a set of more than 100 catchment attributes for 735 catchments in Brazil, the Catchment Attributes for Brazil database (CABra), paving the way for the development of a catchment classification approach and identification of the main controls of catchment behavior.

Although there is still a lack of a commonly agreed-upon framework for catchment classification, there is a consensus that mathematical techniques to find patterns and distinguish similar catchments are an essential component of the system (Wagener *et al.*, 2008). In this context, the clustering method appears as a valuable tool, as discussed in Sivapalan (2005). Several studies across the globe adopted the clustering methods to achieve clusters of similar catchments, from the fuzzy partitioning method (Sawicz *et al.*, 2011, 2014) to hierarchical clustering (Chaney *et al.*, 2020; Jehn *et al.*, 2020).

To our knowledge and date, no study has performed such catchment classification at Brazilian nor tropical catchments. Due to the extreme importance of Brazilian water resources for the global water cycle and also international, regional, and local water trading, we aim to fill this gap by (i) proposing an approach to classify the Brazilian catchments by their similarities on hydrological behavior, and (ii) by determining the landscape characteristics that control the streamflow variability. The catchment classification used a clustering technique to group 735 catchments from the CABra dataset based on 15 hydrological signatures. To determine the controls on streamflow variability, we applied a random forest algorithm to identify and quantify the main drivers of hydrological processes over the catchment groups. By determining the dominant hydrological processes throughout the Brazilian catchments, we contribute to the development of a common classification framework in the hydrology field. By (i) and (ii) we provide insights into the two pathways, with our classification leading to a better understanding of the dominant hydrological processes and them, this may contributing to a common framework formation (Sivakumar, 2008). Finally, we extensively discuss our results, which can lead to an improved understanding of the Brazilian catchments' functioning and behavior.

2. Material and methods

2.1. Study area and dataset

Our study uses data from a set of catchments in Brazil, a tropical country with continental dimensions (~8,500,000 km²). Brazil has significant importance on the global hydrologic cycle, with almost 20% of globe's freshwaters flowing in its rivers (Rodrigues *et al.*, 2015), and as well as global food production, being the largest producer and exporter of grain and beef in the world (Oliveira *et al.*, 2019). To classify the Brazilian catchments by their hydrological similarity and to assess the main controls of the streamflow variability, we used data from the Catchment Attributes for Brazil – CABra database (Almagro *et al.*, 2021), which is a large-sample dataset for 735 catchments in Brazil. The set of catchments spans all Brazilian biomes and hydrologic regions, and it comprises more than 100 attributes related to topography, climate, streamflow, soils, geology, groundwater, land-use, and hydrologic disturbance. Along with the catchment attributes, the CABra dataset also presents daily time-series of climate variables and river discharge, which can be accessed at: https://doi.org/10.5281/zenodo.4070146.

To generate the regime curves for each group we have used precipitation (P), actual evaporation (E), streamflow (Q), and changes in storage (ΔS). The daily values are derived from a gridded ground-based precipitation (Xavier et al., 2016), from the GLEAM product, from streamflow gauges from the Brazilian Water Agency and the relationship $\Delta S = P-E-Q$, respectively.

2.2. Hydrologic signatures

To verify whether or not catchments are similar, we considered a set of 15 hydrologic signatures, obtained from the CABra dataset (Almagro *et al.*, 2021), which used observed data from the Brazilian Water Agency (ANA) to calculate them. These signatures quantify the streamflow and runoff characteristics of each catchment in response to climate forcings and landscape characteristics, providing numerical values that can be compared and tested against other catchments, allowing us to group them into similar classes of hydrological behavior. (Sawicz *et al.*, 2011; Westerberg and McMillan, 2015). The 15 hydrological signatures used are listed in Table 1.

Table 1: Hydrological signatures chosen for this study.

Attribute	Long name	Unit
Mean Q	Mean daily streamflow	$mm.day^{-1}$
Q1	Streamflow 1st quantile	$mm.day^{-1}$
Q5	Streamflow 5th quantile	$mm.day^{-1}$
Q95	Streamflow 95th quantile	$mm.day^{-1}$
Q99	Streamflow 99th quantile	$mm.day^{-1}$
Low Q frequency	Min streamflow frequency	$day.y^{-1}$
Low Q duration	Min streamflow duration	days
High Q frequency	Max streamflow frequency	$day.y^{-1}$
High Q duration	Max streamflow duration	days
HFD	Mean half-flow date	day of the year
Zero-Q frequency	Frequency of zero-flow	$day.y^{-1}$
Elasticity	Streamflow elasticity to precipitation	%
BFI	Baseflow index	-
Runoff coeff.	Coefficient of runoff	-
Coef. of variation	Streamflow coefficient of variation	-

2.3. Cluster analysis to group catchments by streamflow similarities

To perform the classification, categorizing, and grouping of CABra catchments, we used a cluster analysis methodology, which is widely used in hydrological studies (e.g., Ali et al., 2012; Rao e Srinivas, 2006; Sawicz et al., 2011; Sivakumar et al., 2013), along with a Principal Component Analysis (PCA), which enable us to not only group the catchments based on their hydrological similarity but also to identify the most remarkable signatures for each group. Cluster analysis consists of the process of grouping units according to their similarity in previously defined measures. In this study, the catchments are the units, and the similarity measures are the hydrological signatures.

Among several clustering methods described in the literature, the most popular ones are hierarchical agglomerative, Fuzzy partitioning, *k*-means, principal component analysis, and affinity propagation (Milligan and Cooper, 1987; García-Escudero *et al.*, 2010). Here, we implemented the *k*-means method optimized by the Elbow approach (Nainggolan *et al.*, 2019). *k*-means is a unsupervised partitional clustering algorithm to classify multivariate observations based on the Euclidean distance method (Equation 1) between data samples, proposed by MacQueen (1967), and is widely used for clustering large datasets (Jehn *et al.*, 2020). This algorithm divides data into *k* sections (user choice), defines an initial centroid value for it, and randomly selects and assigns the samples (e.g., catchments) to one of the *k* clusters. Then, the distance between each sample to the centroid of each cluster is computed and, after a given number of continuous iterations, when all samples are within one of the k groups, it results in an optimal cluster solution, converging to a local minimum of the criteria parameters (Shi *et al.*, 2010; Marutho *et al.*, 2018; Syakur *et al.*, 2018; Nainggolan *et al.*, 2019).

$$d_{(x_i,y_i)} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad 1$$

As mentioned, the number of clusters needs to be defined a priori. In some cases, when the user has expertise in the field, it is acceptable to define a fixed number of clusters. In our study, however, we automatically determined the number of clusters using the Elbow approach, which minimizes the human interference in the clustering analysis. In this context, the Elbow method is one of the most popular approaches to determine the optimal number of clusters (*k*) for a given dataset. It based on the notion that adding the value k of clusters beyond an optimal k, will not significantly contribute to the analysis. Starting from 1, the value of k is added one by one, looking for the percentage of variance explained as a function of k. During the process, the Sum Squared Error (SSE – Equation 2) – the sum of the average Euclidean distance of each sample to the centroid – is recorded for each step, and when it drops significantly, forming a small angle, the value of k is achieved.

$$SSE_i = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^j - c_k \right\|^2 \qquad 2$$

where x is the data in each cluster; C_k is the k-th cluster.

To decrease the uncertainties and to produce reliable groups, the parameters for the initialization of the cluster analysis were set in a way that the catchments of a given group need to explain at least 85% of the hydrologic signatures' internal variance through 10 individual repeats by the employment of the Elbow method.

2.4. Identifying dominant attributes to streamflow variability by random forest analysis

The most influential attributes to the streamflow variability were also assessed in our analysis. Among the >100 catchment attributes of the CABra dataset, a subset of 18 attributes was selected (Table 2) to perform a regression analysis using the random forest technique. The attributes were selected based on their representativeness of the attribute class and, as reported in the literature (Addor *et al.*, 2018), some of the attributes presented in the CABra dataset were excluded to avoid redundant information through the analyses.

Table 2: Catchment attributes chosen for dominant attributes analysis.

Class	Attribute	Long name	Unit	Source
Topography	Area	Area of the catchment	km^2	Almagro *et al.*, 2021
	Elevation	Mean elevation of the catchment	m	Almagro *et al.*, 2021
	Slope	Mean slope of the catchment	%	Almagro *et al.*, 2021
Climate	Precipitation	Mean precipitation	$mm.d^{-1}$	Xavier *et al.*, 2016
	Aridity index	Aridity index	-	Almagro *et al.*, 2021
	Prec. seasonality	Seasonality of precipitation cycle	-	Almagro *et al.*, 2021
Soil	Soil sand	Sand fraction in the soil	%	Hengl *et al.*, 2017
	Bulk density	Bulk density in the soil	-	Hengl *et al.*, 2017
	Soil depth	Depth of the soil to the bedrock	m	Hengl *et al.*, 2017
Geology	HAND	Height above the nearest drainage	-	Nobre *et al.*, 2011
	Porosity	Subsurface soil porosity	-	Gleeson *et al.*, 2014
	Permeability	Subsurface soil permeability	-	Gleeson *et al.*, 2014
Land-use	Forest	Forest fraction of cover	%	Buchhorn *et al.*, 2019
	Crops	Crop’s fraction of cover	%	Buchhorn *et al.*, 2019
	Grass	Grass fraction of cover	%	Buchhorn *et al.*, 2019
	Summer NDVI	NDVI in DJF (rainy-season)	-	Buchhorn *et al.*, 2019
Anthropogenic	Disturbance	Index of hydrological disturbance	-	Almagro *et al.*, 2021
	Reservoirs	Number of reservoirs	-	ANA, 2020

Random forests are a machine-learning algorithm that relies on numerous regression trees to generate an ensemble of predictions (Breiman, 2001; Wang *et al.*, 2015), as illustrated in Figure 1. Random forest regressor algorithms, for instance, relate predictors to a response variable. Here we consider the catchment attributes as the predictors and the hydrological signatures as the response variables. Once the regression tree is grown, each predictor (catchment attribute) is randomly shuffled through the regression, and the prediction is performed. Them, prediction accuracy is assessed removing one-by-one the predictors, indicating how much the given catchment attribute is important to the hydrological signature prediction, in a methodology called Recursive Feature Elimination (RFE) which indicates the most relevant and redundant predictors. For each predictor removed, the mean squared error (MSE) of the prediction is assessed. The more the error (expressed in MSE) in hydrological signatures increases, the more important and influential is the catchment attribute to it. This way, we can assess the most influential attributes on the hydrological signatures of each catchment group, making it possible to identify the main drivers of the streamflow variability.

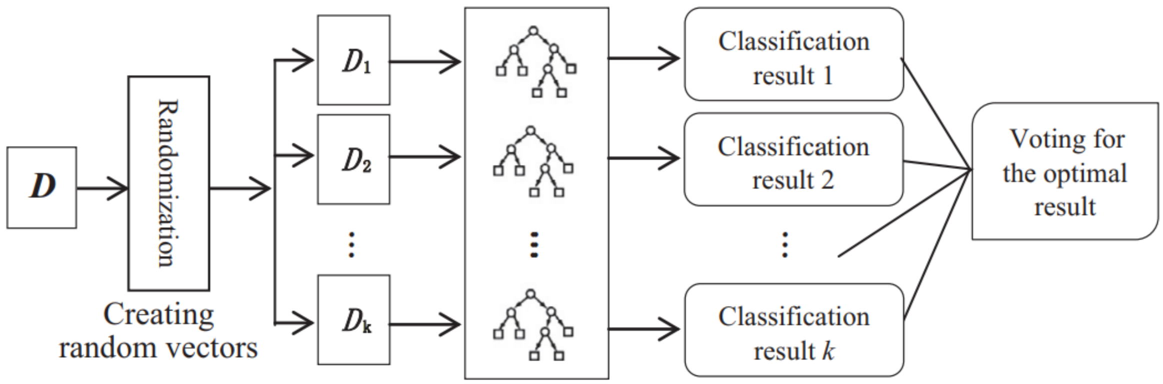

Figure 1: Random forest basic principles and functioning (Wang et al., 2015).

There are a series of advantages on using random forest algorithms, as very well described in Addor *et al.* (2018) and Tyralis *et al.* (2019), which can be summarized in four main aspects: random forests are adapted to capture non-linear relationships between the hydrological signatures and catchment attributes; random forests are not limited by any physical principles, making possible to identify unknown relationships which would be impossible by traditional hydrological modeling; as random forests use an ensemble of regression trees, there is a lowered chance of data overfitting when using big training data; and last (and maybe more important), random forests are interpretable.

The random forest is commonly used to predict hydrological responses based on the climatic variables or catchment attributes. It was already employed in Addor *et al.* (2018) for creating a ranking of hydrological signatures based on their predictability, Booker and Snelder (2012) for frequency of high-flows, and Booker and Woods (2014), to also predict hydrological signatures. As mentioned before, we will employ the random forest to assess the most influential catchment attributes to the hydrological signatures based on the predictions of the random forest algorithms. To do so, we assumed 100 trees to ensure convergence, as indicated in Addor *et al.* (2018).

3. Results and discussion

3.1. Groups with similar hydrologic behavior

Considering the hydrological signatures behavior and similarity, we classified the Brazilian catchments into six groups, as shown in Figure 2. The climatological and hydrological behavior of each one of the groups will be discussed along this section, considering the long-term water balance through the Budyko space, the relationship between the long-term means of precipitation and streamflow, and, additionally, a principal component analysis to show which are the characteristic signatures of each class of catchments.

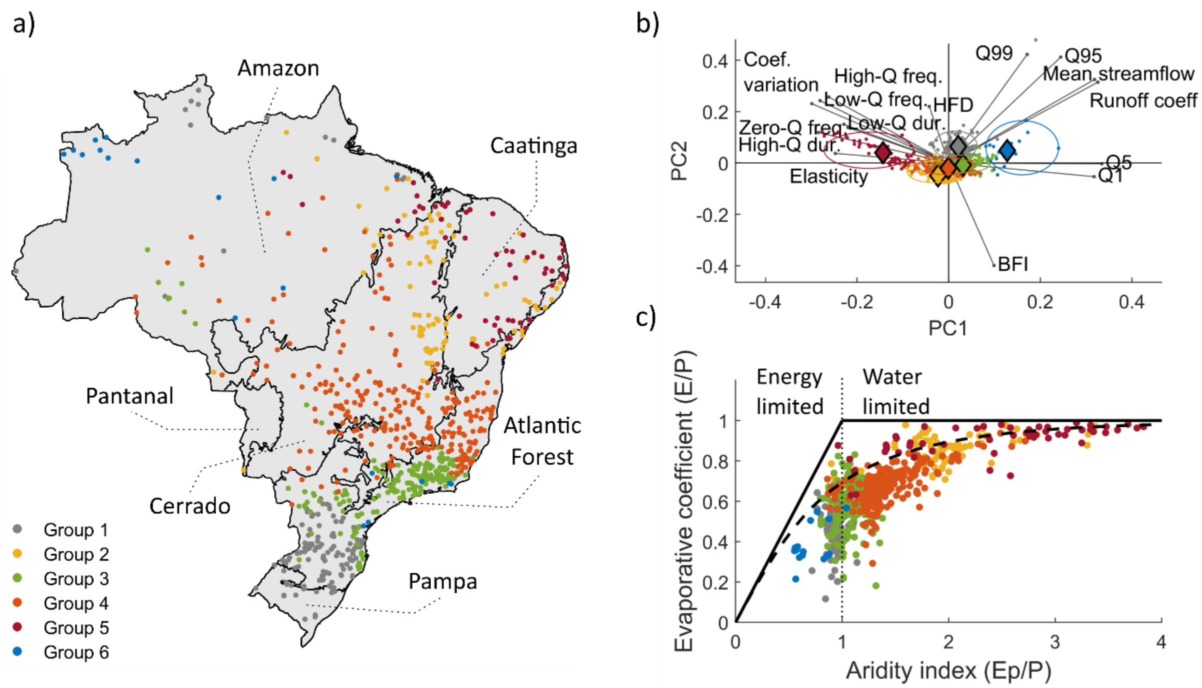

Figure 2: Groups with similar hydrological behavior, based on a catchment classification of their hydrological signatures. A) is the spatial distribution of the six groups over Brazilian biomes: Amazon, Atlantic Forest, Cerrado, Caatinga, Pampa, and Pantanal; b) is the biplot of the principal component analysis of the hydrological signatures from the six hydrological groups; c) is the distribution of the hydrological groups in the space of Budyko.

The spatial distribution of the catchment groups through the Budyko space (Budyko, 1948) can give some insights into the long-term annual water balance (Figure 2c). Group 1 of catchments is mainly energy limited, as shown in Budyko space, falling below the Budyko curve, showing that these catchments are not primarily driven by the climate components of the Budyko equation. The opposite behavior can be noted in catchments within Group 2, which present high values of aridity index (Ep/P), making them largely climate dependent. An extreme climatic condition of Group 2 is Group 5, which presents the highest values for aridity index among all groups, and catchments are mostly located in the extreme of the Budyko space. Group 3 presents an aridity index varying slightly close to 1, and mostly below the Budyko curve, indicating the absence of water stress in the catchments. Group 6 can be considered the extreme conditions of wetness from Group 3, with all the catchments presenting the aridity index lower than one, indicating that those catchments are energy limited, with most of the water precipitated turned into streamflow (surface and/or subsurface). Finally, Group 4 presents its catchments with an aridity index ranging between one and two, and this group can be considered a transition between the energy-limited catchments from Group 3 and the water-limited catchments from Group 2. We can also note that this transitional behavior is also observed through the Brazilian territory, where catchment grouping varies through the latitudinal gradient, from Group 3 to Group 2, with Group 4 in the middle of the road. As reported in Sawicz *et al.* (2011), we also noted the important role of catchment proximity for the catchment similarity through Brazilian catchments, even if the same is not true in other regions (Jehn *et al.*, 2020).

When looking at the Q vs P diagram (Figure 2d), we note that Brazilian catchments present, in average, a 2 mm of initial abstraction of precipitation before the streamflow generation, and that for some catchments, especially from Group 5, this means that for most of the time, there will not be any streamflow. Points deviating above the 1:1 line indicate that there is more streamflow than precipitation for those catchments, which can be a result of groundwater exploration, irrigation, or another kind of water import to the catchment.

According to the principal component analysis (Figure 2b), Groups 1, 3, and 6 hydrologic signatures are mainly subjected to distribution characteristics of streamflow (mean streamflow and quantiles) aside from the runoff coefficient. This occurs mainly due to the high values of streamflow through the year, even during the dry season (which does not occur only for Group 1). The principal component analysis for Groups 2 and 5 show the characteristic of higher streamflow elasticity, and extremes flow (both in frequency and duration). This is expected for dry catchments due to their characteristics of low mean streamflow, making storm events generate high flows. As mentioned before, Group 5 presents more accentuated signatures than Group 2 but with the same behavior.

The regime curves for precipitation, actual evapotranspiration, streamflow, and storage for the six groups are shown in Figure 3, as well their hydrological signatures and attributes from climate and landscape, which are presented in Figure 4 and Figure 5. The integrated analysis of catchments' location, hydrological signatures, and landscape attributes enable us to identify and understand the main features of climate and landscape (attributes) and their relationship with the hydrological behavior (hydrological signatures), and if there is any kind of spatial predictability on the catchment classification.

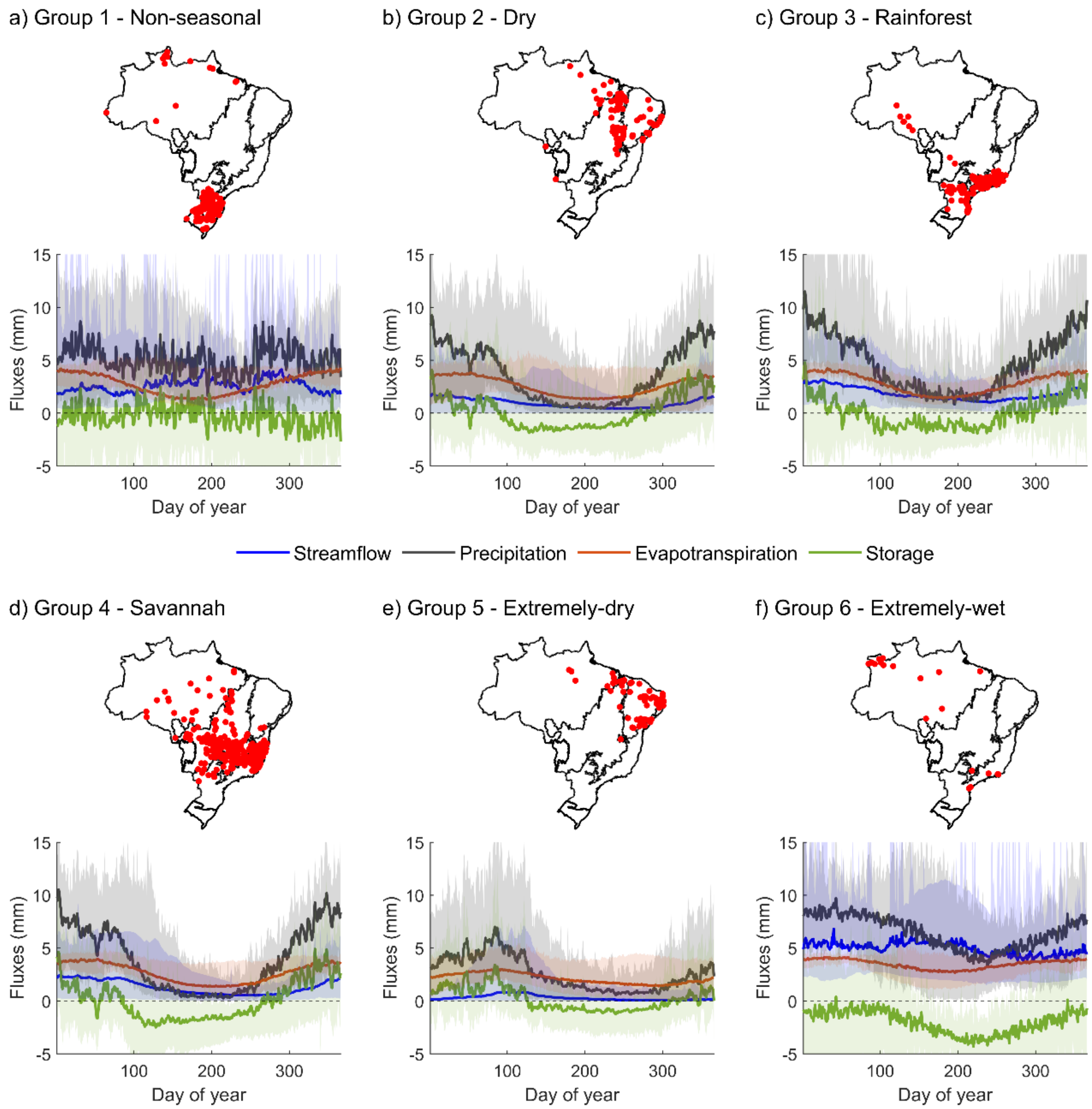

Figure 3: Spatial distribution (upper panels) and daily regime curves (lower panels) from water balance components (precipitation, actual evapotranspiration, streamflow, and changes in water storage) for each of the hydrological groups. A) is the Non-seasonal group; b) is the Dry group; c) is the Rainforest group; d) is the Savannah group; e) is the Extremely-dry group; and f) is the Extremely-wet group all the units are in mm.

The typical signatures from Group 1 are the high values of mean streamflow (in median ~2.5 mm.day^{-1}), runoff coefficient, and 95th and 99th percentiles of streamflow, and consequently present a great value of high-flows frequency (Figure 4). This group is composed of 127 catchments mainly located in the southern portion of Brazil, in the Pampa and Atlantic Forest biomes, but there are many catchments also located in the Amazon. As can be seen in Figure 3a, there is no well-defined rainy or dry season in this group, which led us to name it "Non-seasonal catchments". Catchments in this group are mostly covered by forest and grassland, but there is a considerable percentage (17%) of croplands (most likely from the Pampa biome). On a median, the area of these catchments is 1,670 km², located at 668 m of altitude, with a 12% of slope (Figure 5). Soil can be classified as "clay", with the highest organic carbon content among the catchment groups, and low bulk density. The climate variable of these catchments that make their characteristic attribute is the precipitation seasonality, which is close to zero in median terms, with a small range of values, indicating that there is not a clear relationship between the precipitation and temperature annual cycles, with no well-defined rainy and dry-seasons.

Group 2 is composed of 89 catchments especially in the northeastern portion of Brazil, comprising all Brazilian biomes (predominantly in Cerrado), except the Pampa. These catchments have median area, altitude, and mean slope of 6,890 km², 289 m, and 6%, respectively (Figure 5). Land-cover is mainly comprised of forest, grass, and shrub formations, while the climate is represented by well-defined rainy and dry seasons, with the highest temperatures of all groups, ranging from 20ºC to 32ºC. The aridity index for Group 2's catchments shows a wide range, from 1 – in coastal catchments – to 3.5 – in semi-arid catchments. Streamflow is a direct result of low amounts of precipitation and high values of evapotranspiration, with values up to 2 mm.day^{-1} (Figure 4), in mean for all the catchments within the group, which led us to name it as "Dry catchments". Soil is a "sandy loam", with intermediate values of organic carbon content and bulk density.

Group 3 is comprised of 174 catchments mainly in the Atlantic Forest biome, but also located in Cerrado and Amazon. This group represents the "Rainforest catchments" in Brazil. It is mostly covered by forest and grass with 50% and 35% of the area, respectively (Figure 5). There is a well-defined rainy season, which covers the October-March period, with high amounts of precipitation, that can reach up to 12 mm.day^{-1}, in mean. Temperatures range from

15ºC to 27ºC, and it is in-phase with the precipitation cycle. Streamflow, in turn, presents a short lag to the precipitation cycle, with values of daily streamflow ranging from 1 to 3 mm.day^{-1}, in cluster average. In median terms, Group 3 presents the smallest catchment area, with 1,380 km². Elevation (829 m) and mean slope (17%) are the highest among groups though. Soil is classified as "sandy clay", with high values of organic carbon content and low bulk density. Catchments from Group 3 present the baseflow and low-quantiles of streamflow as main hydrologic signatures, which explain most of the variance across the signatures (Figure 4). The aridity index for these catchments ranges from 0.8 to 1.5, showing a relatively non-water-stressed condition.

Group 4 is comprised of 269 catchments, mainly located in the Brazilian Cerrado, a typical savannah biome (Merten and Minella, 2013; Oliveira *et al.*, 2015), but also cover catchments in Atlantic Forest and Amazon. Due to this characteristic, we call this group the "Savannah catchments". As well as groups 2 and 3, there are well-defined rainy and dry seasons, with very similar regime curves of precipitation and evapotranspiration. The mean daily streamflow is slightly larger than 1 mm.day^{-1}, and we can consider this group as a transitional group between the "Dry catchments" to the "Rainforest catchments". This behavior is likewise transferred to the hydrological signatures, which present an intermediate condition between groups 2 and 3 for the 8 out of 15 signatures (Figure 4). Catchments within this group are mainly covered by grass and forests, present a mean area of 6,420 km², with 696 m of mean elevation and 7% of mean slope (Figure 5). Soil can be classified as "sandy loam", with intermediate values of soil organic carbon and bulk density. The aridity index ranges from 1 to 2 (with a median value of 1.5) and the precipitation through the catchments are in-phase (precipitation seasonality close to 1) with temperature.

Catchments within Group 5 presents similar conditions to the "Dry catchments" groups but are non-perennial streams. It comprises 57 catchments in the Caatinga biome, with well-defined rainy and dry seasons, high temperatures of at least 25ºC (during winter), and the highest values of evapotranspiration throughout Brazilian catchments. One of the most notorious differences between this Group 5 and Group 2 is the land-cover characteristics, being covered mostly by shrub and grass formation. These catchments are mostly flat (6% of slope), in medium altitude (534 m), with 3440 km² of area, in median terms (Figure 5). The soil of the catchments is mainly composed of sand (62%), being classified as "sandy loam". They also present higher values of bulk density and lower values of organic carbon content. The catchments in this group are mainly subjected to hydrologic signatures based on the frequency and duration, and dynamics of streamflow, such as streamflow elasticity, low-frequency, low-duration, high-frequency, high-duration, and zero-flow frequency (Figure 4). This group of catchments presents the highest values of aridity index between all Brazilian catchments ($Ep/P \approx 3$), showing an extremely water scarcity, being the only group to present zero-flow frequency higher than zero, meaning that their catchments are subjected to present any flow for at least one day per year.

Group 6 represents the “Extremely-wet catchments” in Brazil. It comprises 19 catchments in the Amazon and Atlantic Forest. This group presents the highest amounts of daily precipitation of Brazilian catchments as well the highest runoff coefficient. Temperatures range from 15ºC (in southern catchments) to 27ºC (in northern catchments), and the streamflow is, in mean, 5 mm day^{-1}. Catchments are very flat (4% of mean slope) and large (24,000 km^2), located in the lowest mean altitude of all groups (223 m) (Figure 5). They are predominantly covered by forest (>90% of the area), showing the high influence of the vegetation in the catchment’s streamflow. Soil is classified as “sandy clay loam”, with intermediate values of organic carbon content and very low values of bulk density. Group 6 can be considered an extreme wet condition of Groups 1 and 3, which are also mainly covered by forests, and the hydrologic behavior is mostly figured by mean streamflow, and, as a direct consequence, the runoff coefficient (Figure 4).

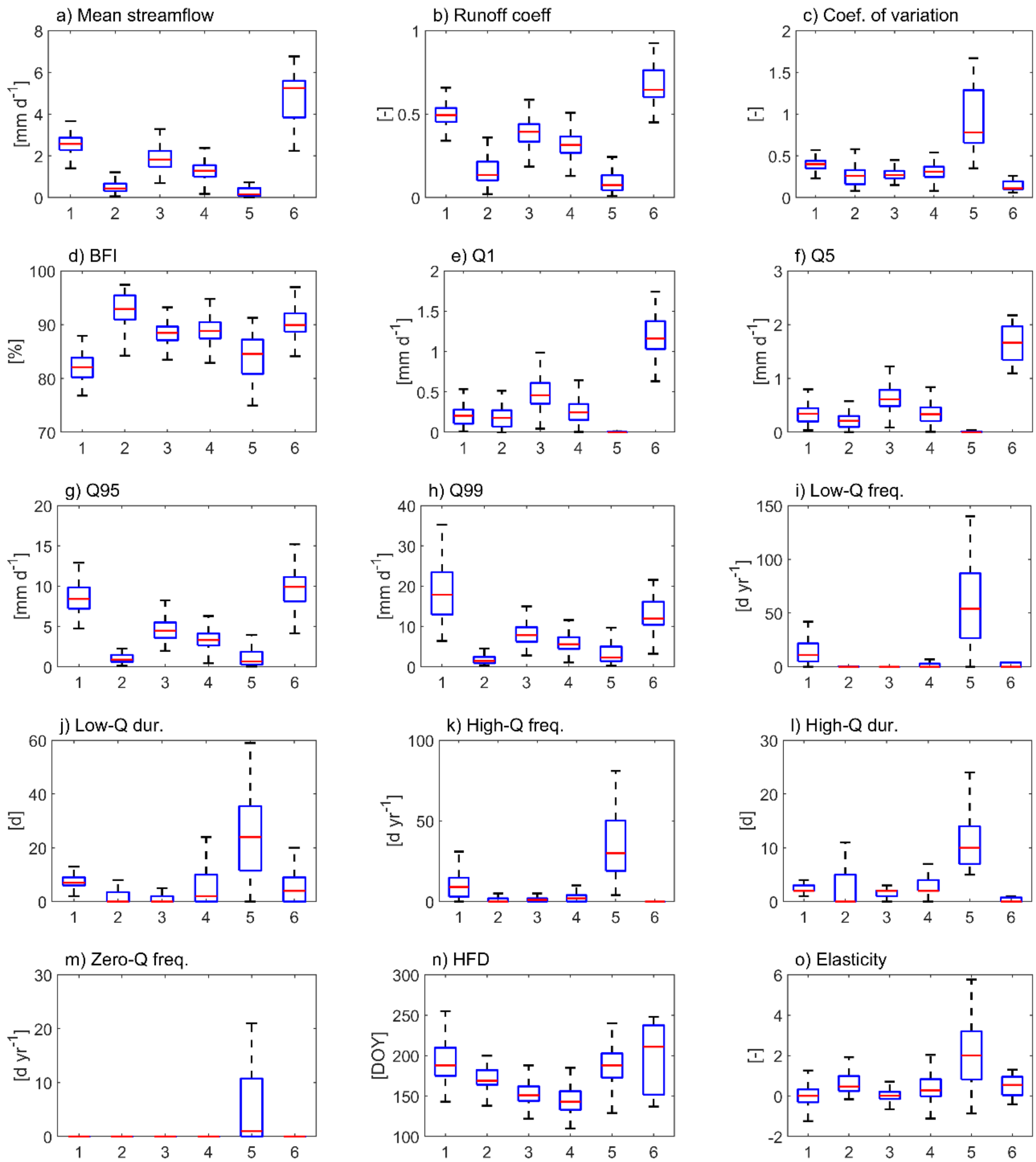

Figure 4: Boxplots of the hydrological signatures characterizing the catchments within each of the six hydrological groups. A) is the mean streamflow, in mm.day-1; b) is the runoff coefficient, dimensionless; c) is the coefficient of variation of the streamflow, dimensionless; d) is the baseflow index, dimensionless; e) to h) are the percentiles of daily streamflow, in mm.day-1; i) and j) are the frequency and duration of low-flows, respectively; k) and l) are the frequency and duration of the high-flows, respectively; m) is the frequency of zero-flow; n) is the half-flow day; and o) is the streamflow elasticity to the precipitation, dimensionless.

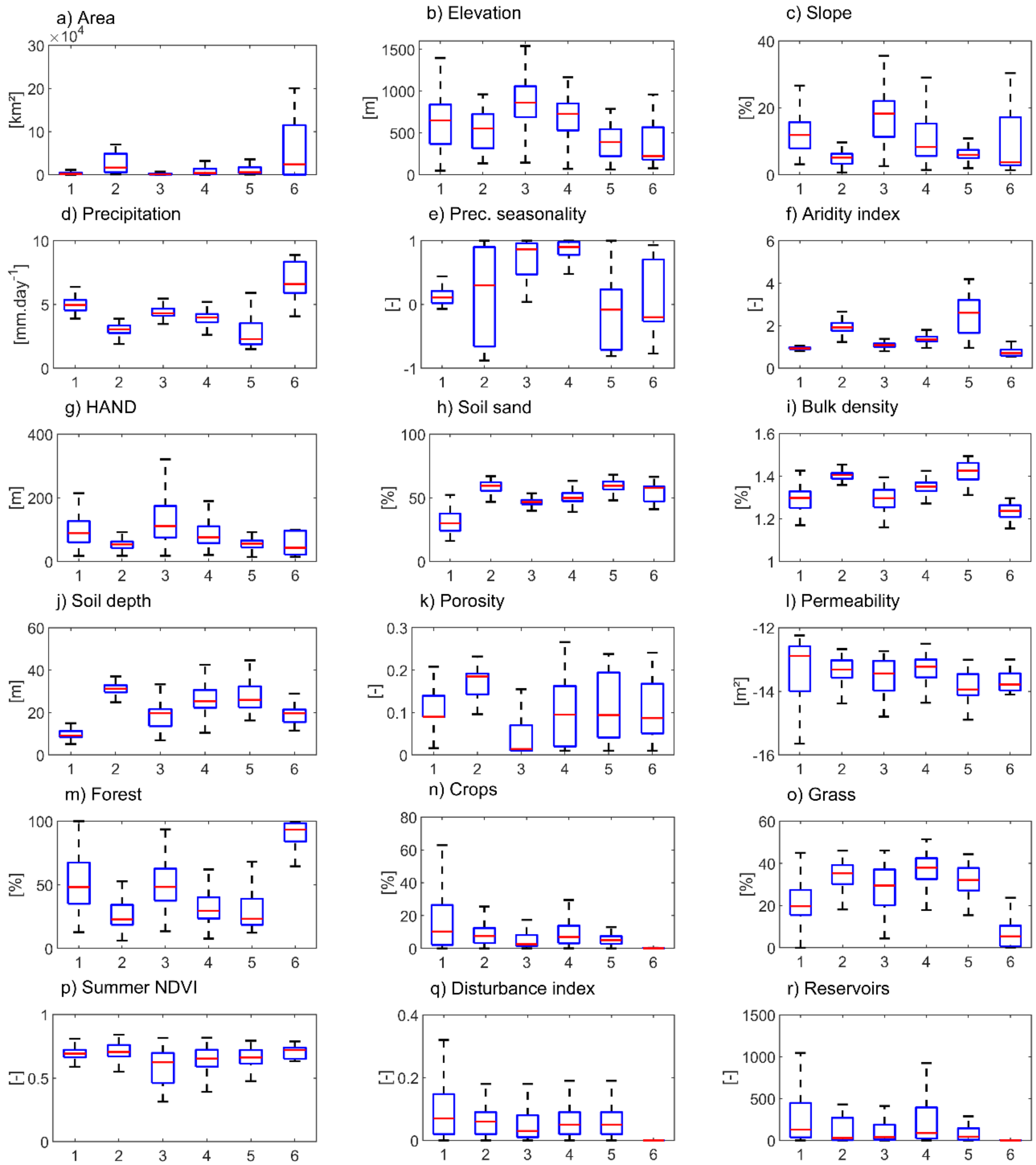

Figure 5: Boxplots of catchment attributes signatures characterizing the catchments within each of the six hydrological groups. A) to c) are the attributes related to the topography: the area, in km², the mean elevation, in m, and the slope, in %; d) to f) are the attributes related to the climate: mean precipitation, in mm.day-1, the precipitation seasonality, dimensionless, and the aridity index, dimensionless; g) to i) are the attributes related to the soil: fraction of sand, in %, the bulk density, in %, and the soil depth to the bedrock, in m; j) to l) are the attributes related to the geology: height above nearest drainage, in m, subsurface porosity, dimensionless, and the subsurface permeability, in m²; m) to p) are the attributes related to the land-cover: the fractions of forest cover, crops cover, and grass cover, respectively, in %, and the NDVI during the summer season, dimensionless; and q) to r) are the attributes related to the anthropogenic modification: the hydrological disturbance index, dimensionless, and the number of reservoirs in the catchment.

3.2. Dominant attributes to streamflow variability in Brazilian catchments

The results from the random forest analysis followed by the recursive feature elimination, showing the attributes importance grouped by attribute classes, are shown in Figure 6. Through the competition of the hydrological processes to the overall catchment functioning (L'vovich, 1979), we found that different attributes classes dominate hydrological behavior across the different groups. In summary, the hydrological behavior in the "Non-seasonal" catchments is mainly controlled by the landscape cover; the "Dry catchments" are driven by the climate with the land-use also being influential; "Rainforest catchments" presents, in general, a similar dependency on the land-use and climate of the catchment; the "Savannah catchments" are mainly dominated by the influence of the climate; the climate is even more important for catchments classified as "Extremely-dry", being the class with more influence of climate along with all groups, with significant human impact due to its high number of reservoirs; finally, but not less important, the "Extremely-wet catchments" hydrological behavior is ruled mainly by the climate of the catchment, followed by topography and soil attributes. As will be comprehensively explored in the next sections, our results reveal that large-scale catchments are more likely to be more impacted by attributes related to the landscape features, such as topography, soils, and geology. Small catchments, in turn, due to the more homogeneous landscape and faster response to changes in precipitation and temperature, are subjected to be climate-induced.

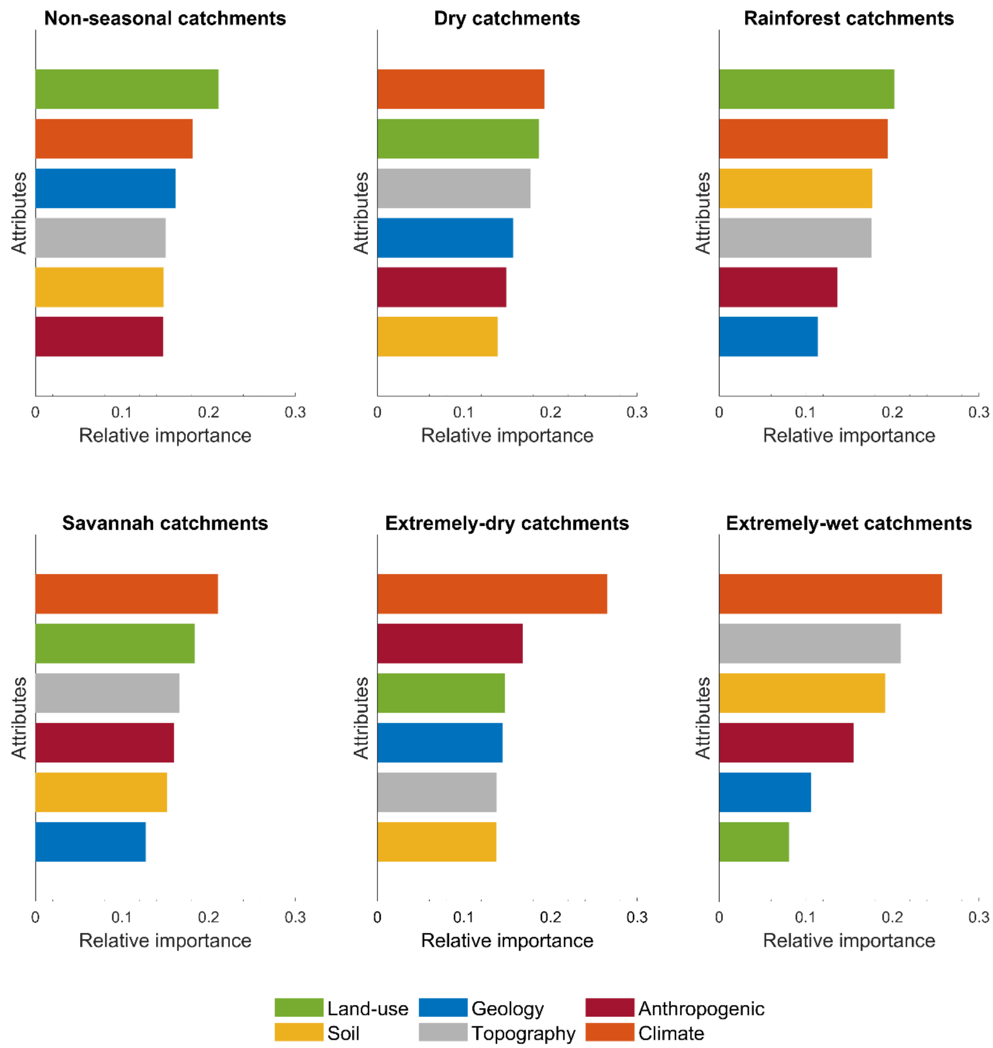

Figure 6: Bar plots of the classes of attributes controlling the streamflow variability through the hydrological groups in Brazilian catchments.

3.2.1. Group 1 – Non-seasonal catchments

The diagram relating the catchment attributes importance to the hydrological signatures of the "Non-seasonal catchments" is presented in Figure 7. The main driving factor to the mean daily streamflow of Group 1 is the aridity index, being most of its catchments in an energy-limited condition. Aside from the aridity index, soil and geology attributes also act as important factors to the mean daily streamflow, but with lower influence than the climate. There is also a strong inverse relationship between the percentage of grass and the 1st and 5th percentiles of daily streamflow, and baseflow index (BFI) in Group 1's catchments. This could be due to cattle ranching or croplands, which are common in the southern portion of Brazil. Compaction by animal or mechanical grazing can modify soil structure and hydrological behavior since it can reduce infiltration, increasing overland flow (Trimble and Mendel, 1995; Singleton *et al.*, 2000; Pietola *et al.*, 2005). The reduced water infiltration decreases the water availability at to the deeper soil layers, lowering the BFI, which, in turn, shows a significant contribution to the streamflow at the reference low percentiles (Q1 and Q5). Consequently, we note a positive influence on low-flows frequency and duration. The integrated analysis of Figure 7 along with the regime curves of Group 1, presented in Section 3.1, indicates how to identify the hydrological behavior of the group. It is possible to note that all precipitation variability is damped by the changes in water storage, and the streamflow is closely related to the actual evaporation. As this group does not present a well-defined dry season, once the evapotranspiration decrease, the streamflow automatically increases, in the same magnitude, reinforcing the idea of a climate-driven group.

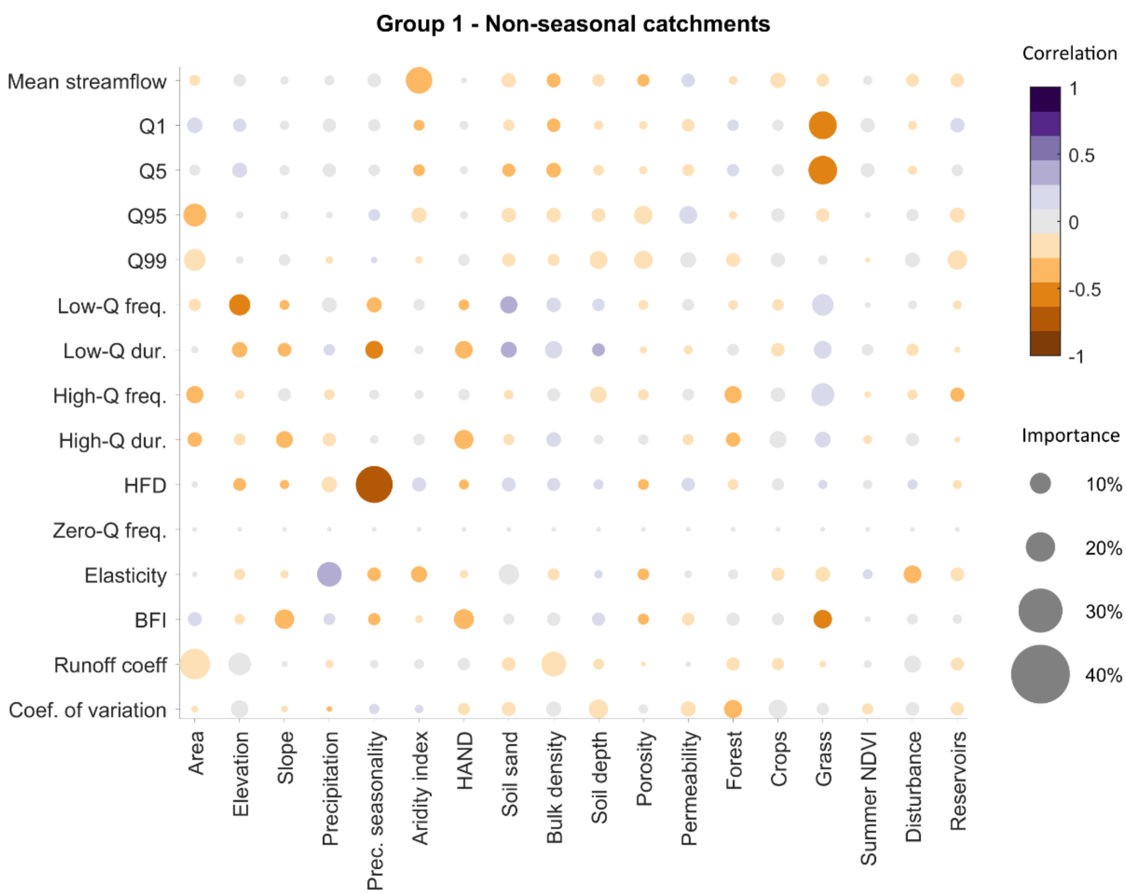

Figure 7: Diagram of the main controls of hydrological behavior of Group 1 – Non-seasonal catchments. The size of each circle indicates the importance of a given catchment attribute to a given hydrological signature (the bigger the circle, the bigger its importance). The color of each circle indicates the correlation strength between a given catchment attribute to a given hydrological signature.

3.2.2. Group 2 – Dry catchments

The mean streamflow from the "Dry catchments" is largely dominated by the climate, as seen through high importance showed by the precipitation seasonality (with positive correlation) and aridity index (with negative correlation), as indicated in Figure 8. The catchments within this group fit on the class of arid (or semi-arid) catchments, during a significant part of the year, specifically in the dry season, there is more bottom-up water movement through evapotranspiration than precipitation because most of the storage is held on the unsaturated zone as capillary water, which is then evaporated in the interstorm period (Wagener *et al.*, 2007), leading to very low values of streamflow. Attributes related to the topography of the catchments are mainly important for the frequency and duration of extreme events (high- and low-flows), being the slope one of the main drivers of low-flows frequency and duration. Aside from the mean streamflow, the aridity index is the main control of most of the hydrological signatures from the "Dry catchments", showing the importance of climate to this kind of catchments. Soil properties also play an important role in the dry catchments due to its characteristics of large depth to the bedrock, bulk density, and porosity, leading to storage in deepest soil layers, disfavoring the direct overland flow, but increasing the BFI.

Although our analysis shows that forest cover delivers more streamflow in the rivers, there is an interesting positive correlated importance between the forest cover fraction and the streamflow percentiles (mainly the lower bounds) and the low-flows frequency and duration. The forest fraction is inversely proportional to the lower bounds of percentiles and proportional to the frequency and duration of low-flow events. This can be explained by the high values of actual evapotranspiration in these catchments. Since tall trees (which characterize the forests) present a developed and efficient root system to achieve deep soil moisture to their survival (Harper *et al.*, 2014), they can reach deep soil layers, which will be further evaporated by transpiration, lowering the soil water available to the streamflow.

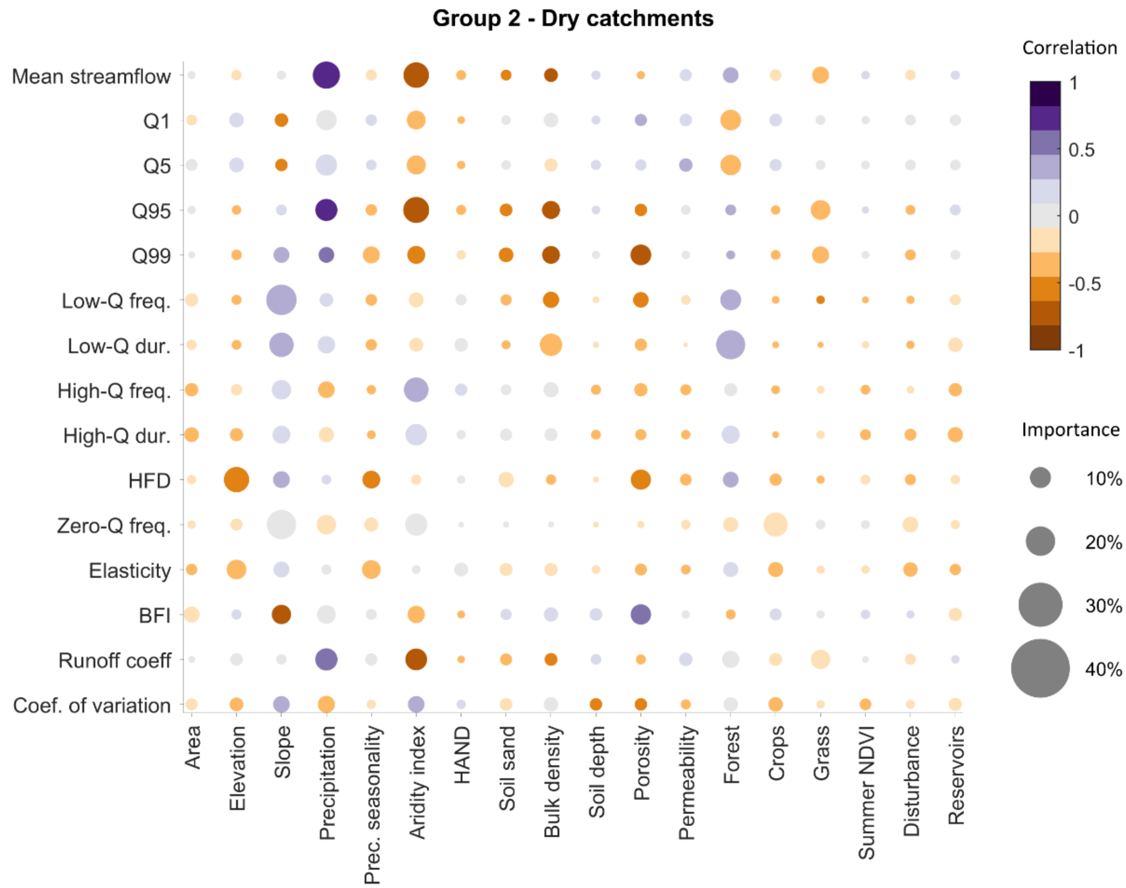

Figure 8: Diagram of the main controls of hydrological behavior of Group 2 – Dry catchments. The size of each circle indicates the importance of a given catchment attribute to a given hydrological signature (the bigger the circle, the bigger its importance). The color of each circle indicates the correlation between a given catchment attribute to a given hydrological signature.

3.2.3. Group 3 – Rainforest catchments

The influence of catchment attributes on the hydrological signatures from the "Rainforest catchments" is presented in Figure 9. The main attribute controlling the mean streamflow of its rivers is the precipitation, rationed by the aridity index. The influence presented here by the aridity index is not so relevant to the streamflow as in other groups (like Groups 2 and 4) mainly because the available energy to evaporate the water precipitated is lower than the precipitation all over the hydrological year. Consequently, the streamflow is a direct response to the precipitation, following the same pattern of the annual cycle. Throughout the "Rainforest catchments", there is the highest influence of the attributes related to the anthropogenic intervention of all groups, here represented by the hydrological disturbance index and the number of reservoirs. Those attributes mainly affect the mean and percentiles of streamflow, with a negative correlation. Soil and geological properties also present themselves as an important factor to hydrological signatures such as the high-flows frequency, half-flow day, streamflow elasticity, baseflow index, and streamflow coefficient of variation. This characteristic reinforces the importance of the surface and subsurface layers on controlling and attenuating the quick variations on water input (precipitation) by storing the water through the soil layer.

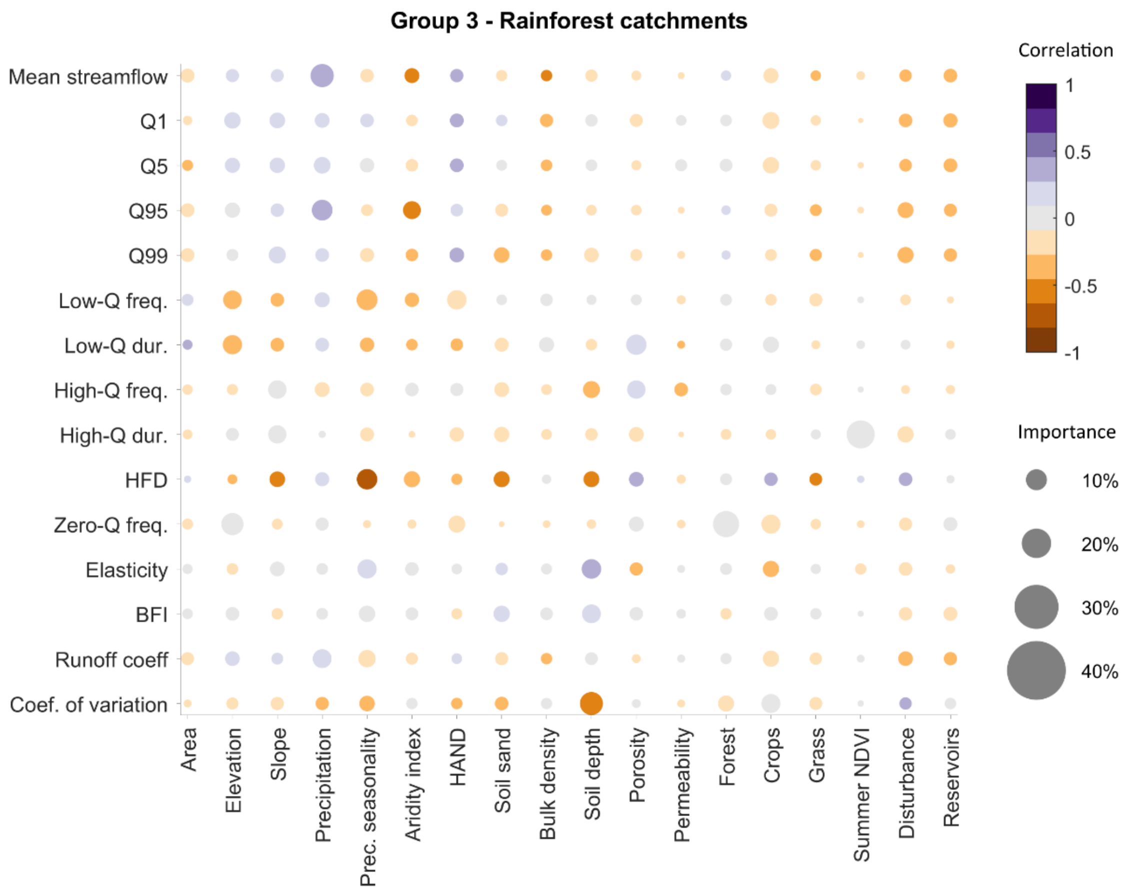

Figure 9: Diagram of the main controls of hydrological behavior of Group 3 – Rainforest catchments. The size of each circle indicates the importance of a given catchment attribute to a given hydrological signature (the bigger the circle, the bigger its importance). The color of each circle indicates the correlation between a given catchment attribute to a given hydrological signature.

3.2.4. Group 4 – Savannah catchments

The hydrological behavior of the "Savannah catchments" is mainly subjected to the climate, as illustrated in Figure 10. There is a clear influence of the climatic attributes, here represented by the mean precipitation, precipitation seasonality, and aridity index through the streamflow variability, here symbolized by the hydrological signatures, which can reach up to

50% for the mean streamflow. There is also an important influence of the slope on some signatures, such as the low percentiles of streamflow (1st and 5th), half-flow day, baseflow index, and coefficient of variation. Porosity also acts as an important factor (with a small positive correlation) to the elevation of the low percentiles of streamflow, likely because the storage of water in a more porous soil can lead to higher values of mean streamflow by subsurface contribution.

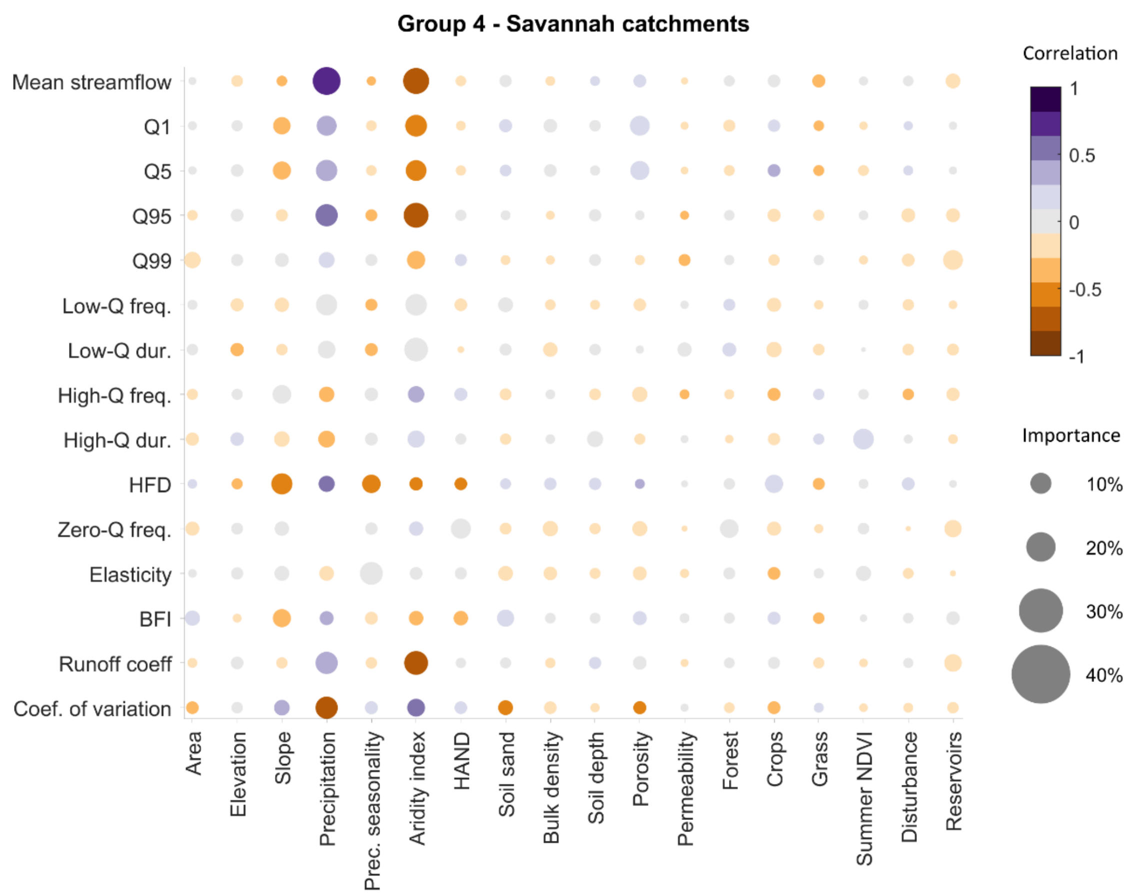

Figure 10: Diagram of the main controls of hydrological behavior of Group 4 – Savannah catchments. The size of each circle indicates the importance of a given catchment attribute to a given hydrological signature (the bigger the circle, the bigger its importance). The color of each circle indicates the correlation between a given catchment attribute to a given hydrological signature.

3.2.5. Group 5 – Extremely-dry catchments

Through the group of "Extremely-dry catchments" there is more complexity involved in the most influential attributes than other groups here presented (Figure 11). In this group of catchments, most of the time, as well Group 2's catchments, the precipitated water remains on the most superficial layers of the soil, in the unsaturated zone, being evaporated, due to the high amounts of energy available. Through the hydrological processes competition mediated by the climate and landscape, the climate limits the hydrological behavior of this group is a great part of the year: there is no water available to generate runoff. Moreover, there is an important influence (inversely correlated) of the topography with the mean streamflow, with the increase in the altitude and slope, making this group of catchments an exporter of water (i.e., leaky catchments). Soil properties like bulk density and soil depth are also important to the low percentiles of streamflow. Aside from being the main controls of the streamflow elasticity and coefficient of variation, the fraction of forests is also important for the mean streamflow and percentiles, with higher values of forest generating more streamflow. This is due to the presence of vegetation with a larger leaf area than grass, for example. Some studies suggest that the plant stomata closes during very dry conditions in the air, not allowing the evapotranspiration by the plant (Werth and Avissar, 2004), increasing the water available in the soil for subsurface flow.

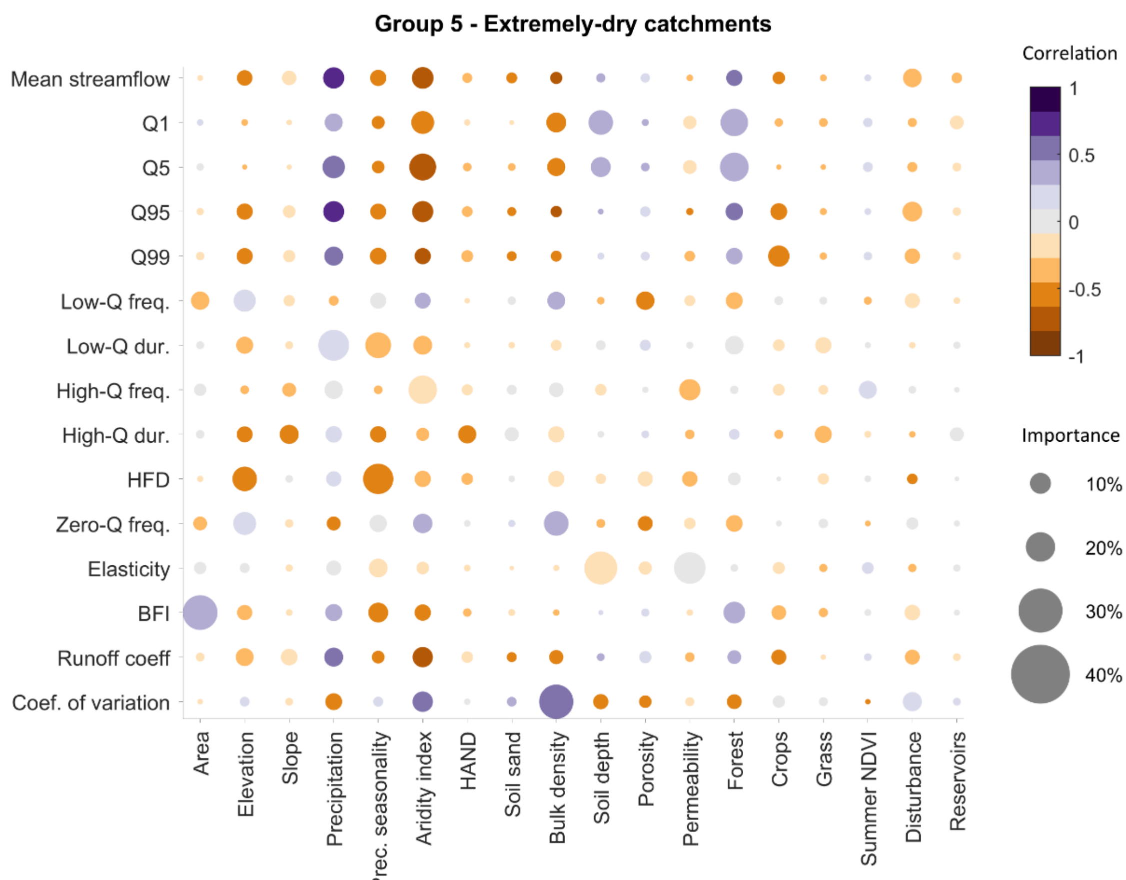

Figure 11: Diagram of the main controls of hydrological behavior of Group 5 – Extremely-dry catchments. The size of each circle indicates the importance of a given catchment attribute to a given hydrological signature (the bigger the circle, the bigger its importance). The color of each circle indicates the correlation between a given catchment attribute to a given hydrological signature.

3.2.6. Group 6 – Extremely-wet catchments

The influence and correlation between the catchment attributes and the hydrological signatures from the "Extremely-wet catchments" are shown in Figure 12. Although we consider this group as an extreme wetter hydrological condition of Groups 1 and 3, there are differences in the main drivers of the streamflow variability in this group. As the water is available in abundance and there is not enough energy throughout the year, as can be seen in the regime curves, the climate plays the main role in the hydrological processes occurring across these catchments. Landscape attributes play an important role in controlling the hydrological behavior of the catchments in this group. While the soil properties (soil depth and porosity) are the main controls of the mean and percentiles of streamflow, the mean area and mean slope are the drivers of the baseflow index, runoff coefficient (along with soil porosity), coefficient of variation, and high-flow frequency and duration. Still, concerning the frequency and duration of high-flows events, there is a large positive correlated influence of the height above nearest drainage (HAND), indicating that these events are also influenced by the subsurface flow since the HAND is closely related to the water table depth (Nobre *et al.*, 2011). Our results reinforce the idea that landscape features (such as topography and soils) are of main importance in controlling the hydrological behavior of large-scale catchments (Group 6 present the highest mean area of all groups), due to its slow response to the climatological inputs.

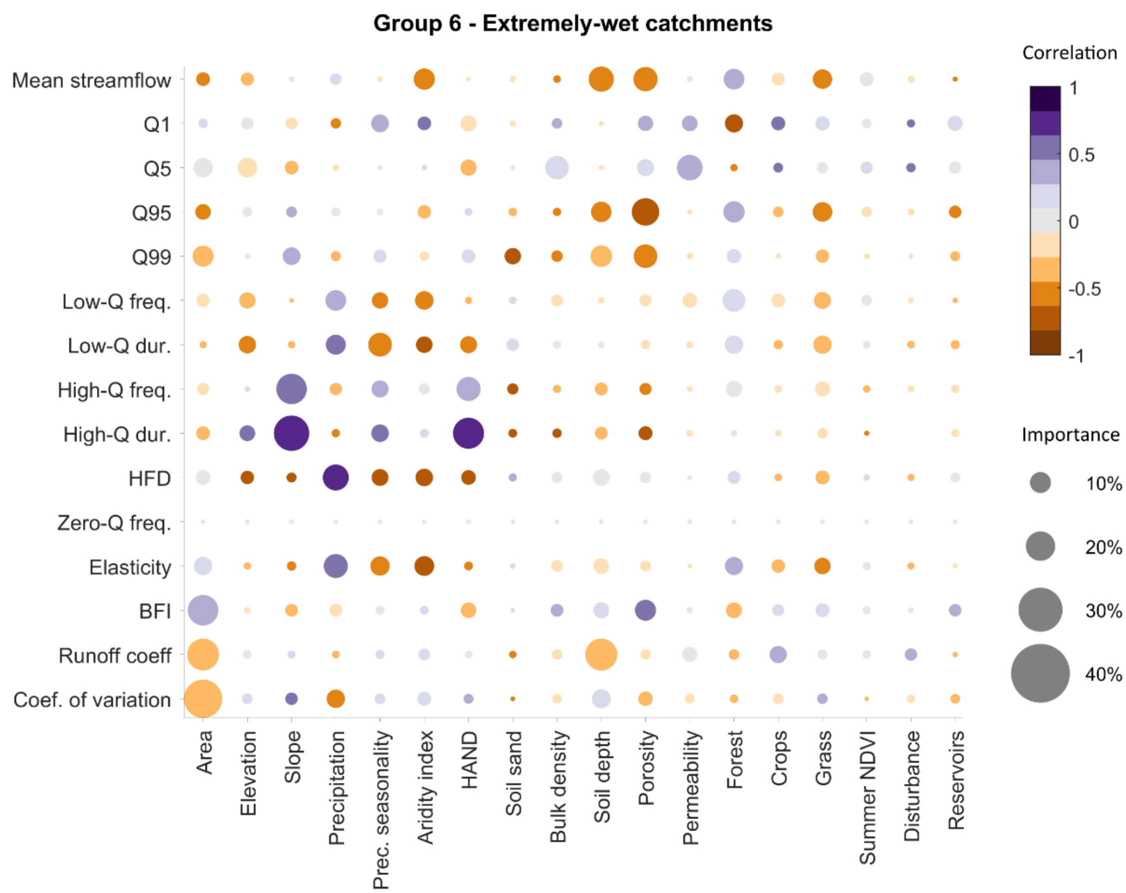

Figure 12: Diagram of the main controls of hydrological behavior of Group 6 – Extremely-wet catchments. The size of each circle indicates the importance of a given catchment attribute to a given hydrological signature (the bigger the circle, the bigger its importance). The color of each circle indicates the correlation between a given catchment attribute to a given hydrological signature.

4. Conclusions

In this study, we conducted the first catchment classification of the Brazilian catchments. To do so, we applied a clustering method to assess the hydrological behavior similarity between 735 catchments from the CABra large-sample dataset. This approach considered a set of hydrological signatures, which represents (in a simple way) the various complexes hydrological processes that occur within a catchment. Additionally, we made a comprehensive exploration of the main drivers of the streamflow variability through the catchment groups. We employed a random forest regressor algorithm to identify and quantify the importance of climate and landscape attributes –that most interfere with the hydrological signatures.

Our catchment classification grouped the Brazilian catchments into six groups. Group 1

(“Non-seasonal catchments”) are mainly located in the south, are in an energy-limited condition, and do not present rainy/dry season. Group 2 (“Dry-catchments”) are mostly located northeast, with high evapotranspiration through the year, sometimes larger than precipitation, leading to a strong dry season. Group 3 (“Rainforest catchments”) are distributed primarily in mountainous areas of the Atlantic Forest, with high values of precipitation and streamflow. Group 4 (“Savannah catchments”) are mainly located in the Brazilian savannah, the Cerrado, and presents well-defined rainy and dry seasons, with streamflow being controlled by the aridity index. Group 5 (“Extremely-dry catchments”) is the extreme condition of Group 2, with high values of evapotranspiration and low amounts of precipitation through the year, leading to an ephemeral condition for its catchments. Finally, Group 5 (“Extremely-wet catchments”) is the extreme condition of Groups 1 and 3, presenting high amounts of precipitation, large areas, and being mostly controlled by climate and topography features.

In some way, even with the striking dissimilarities between the “Extremely-dry” and “Extremely-wet” catchments, we noted that there is a transitional behavior from the first to the last through the other catchment groups (in order: “Extremely-dry” to “Dry” to “Savannah” to “Rainforest” to “Extremely-wet”). The only exception, which does not mean that we cannot identify any similarities with other groups, is the “Non-seasonal” catchments, which present a very particular hydrological behavior, mainly due to its climatological condition of non-seasonality between water and energy availability.

Even a single employment of clustering catchments – as done in this work – cannot create a general framework for a classification system because of the kind of data and subjectivity insert (Sawicz *et al.*, 2011), catchment classification is considered a first step to overcome the great challenges – such as flow predictions in ungauged basins – and reveal unknown processes on hydrology fields (McDonnell and Woods, 2004), and its intend to be the first step towards a better understanding of catchment hydrology in Brazil (and also tropical hydrology). More specifically, our results may be useful for the understanding of the hydrological processes involved in the catchment groups, making it possible to further regionalize hydrological information (McDonnell and Woods, 2004), generalize hypothesis (Wagener *et al.*, 2007), and better predict how responses to climate change are likely to be.

5. Acknowledgments

This study was supported by grants from the Ministry of Science, Technology, and Innovation – MCTI and National Council for Scientific and Technological Development – CNPq [grants numbers 441289/2017-7 and 309752/2020-5]. This study was also financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 and CAPES Print. The authors gratefully acknowledge Dr. Murugesu Sivapalan and Dr. Ross A. Woods for the informal meetings and talking about catchment hydrology, catchment classification, and hydrological processes, which were very insightful and important to the development of this work.

6. References

Addor N, Nearing G, Prieto C, Newman AJ, Le Vine N, Clark MP. 2018. A Ranking of Hydrological Signatures Based on Their Predictability in Space. *Water Resources Research* **54** (11): 8792–8812 DOI: 10.1029/2018WR022606

Addor N, Newman AJ, Mizukami N, Clark MP. 2017. The CAMELS data set: catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences* **21**: 5293–5313 DOI: 10.5194/hess-21-5293-2017

Ali G, Tetzlaff D, Soulsby C, McDonnell JJ, Capell R. 2012. A comparison of similarity indices for catchment classification using a cross-regional dataset. *Advances in Water Resources* **40**: 11–22 DOI: 10.1016/j.advwatres.2012.01.008

Almagro A, Oliveira PTS, Meira Neto AA, Roy T, Troch P. 2021. CABra: a novel large-sample dataset for Brazilian catchments. *Hydrology and Earth System Sciences* **25** (6): 3105–3135 DOI: 10.5194/hess-25-3105-2021

Alvarez-Garreton C, Mendoza PA, Pablo Boisier J, Addor N, Galleguillos M, Zambrano-Bigiarini M, Lara A, Puelma C, Cortes G, Garreaud R, et al. 2018. The CAMELS-CL dataset: Catchment attributes and meteorology for large sample studies-Chile dataset. *Hydrology and Earth System Sciences* **22** (11): 5817–5846 DOI: 10.5194/hess-22-5817-2018

ANA. 2020. Technical Note N. 52/2020/SPR. Brasília.

Beven KJ. 2000. Uniqueness of place and process representations in hydrological modelling. *Hydrology and Earth System Sciences* **4** (2): 203–213 DOI: 10.5194/hess-4-203-2000

Blöschl G, Sivapalan M. 1995. Scale issues in hydrological modelling: A review. *Hydrological Processes* **9** (3–4): 251–290 DOI: 10.1002/hyp.3360090305

Booker DJ, Snelder TH. 2012. Comparing methods for estimating flow duration curves at ungauged sites. *Journal of Hydrology* **434–435**: 78–94 DOI: 10.1016/j.jhydrol.2012.02.031

Booker DJ, Woods RA. 2014. Comparing and combining physically-based and empirically-based approaches for estimating the hydrology of ungauged catchments. *Journal of Hydrology* **508**: 227–239 DOI: 10.1016/j.jhydrol.2013.11.007

Breiman L. 2001. Random forests. *Machine Learning* **1** (45): 5–32 DOI: https://doi.org/10.1023/A:1010933404324

Buchhorn M, Smets B, Bertels L, Lesiv M, Tsendbazar N-E, Herold M, Fritz S. 2019. Copernicus Global Land Service: Land Cover 100m: epoch 2015: Globe DOI: 10.5281/ZENODO.3243509

Budyko MI. 1948. *Evaporation under natural conditions*. Israel Program for Scientific Translations: Jerusalem.

Chagas VBP, Chaffe PLB, Addor N, Fan FM, Fleischmann AS, Paiva RCD, Siqueira VA. 2020. CAMELS-BR: hydrometeorological time series and landscape attributes for 897 catchments in Brazil. *Earth System Science Data* **12** (3): 2075–2096 DOI: 10.5194/essd-12-2075-2020

Chaney N, Torres-Rojas L, Vergopolan N, Fisher C. 2020. Two-way coupling between the sub-grid land surface and river networks in Earth system models. *Geoscientific Model Development Discussions* (October): 1–31 DOI: 10.5194/gmd-2020-291

Coxon G, Addor N, Bloomfield J, Freer J, Fry M, Hannaford J, Howden N, Lane R, Lewis M, Robinson E, et al. 2020. CAMELS-GB: Hydrometeorological time series and landscape attributes for 671 catchments in Great Britain. *Earth System Science Data Discussions* (April): 1–34 DOI: 10.5194/essd-2020-49

Dooge JCI. 1986. Looking for hydrologic laws. *Water Resources Research* **22** (9S): 46S-58S DOI: 10.1029/WR022i09Sp0046S

Duan Q, Schaake J, Andréassian V, Franks S, Goteti G, Gupta H V., Gusev YM, Habets F, Hall a., Hay L, et al. 2006. Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. *Journal of Hydrology* **320** (1–2): 3–17 DOI: 10.1016/j.jhydrol.2005.07.031

Fowler KJA, Acharya SC, Addor N, Chou C, Peel MC. 2021. CAMELS-AUS: Hydrometeorological time series and landscape attributes for 222 catchments in Australia. *Earth System Science Data Discussions* (January) DOI: https://doi.org/10.5194/essd-2020-228

García-Escudero LA, Gordaliza A, Matrán C, Mayo-Iscar A. 2010. A review of robust clustering methods. *Advances in Data Analysis and Classification* **4** (2): 89–109 DOI: 10.1007/s11634-010-0064-5

Gleeson T, Moosdorf N, Hartmann J, van Beek LPH. 2014. A glimpse beneath earth's surface: GLobal HYdrogeology MaPS (GLHYMPS) of permeability and porosity. *Geophysical Research Letters* **41** (11): 3891–3898 DOI: https://doi.org/10.1002/2014GL059856

Hao Z, Jin J, Xia R, Tian S, Yang W, Liu Q, Zhu M, Jing C. 2021. Catchment attributes and meteorology for large sample study in contiguous China. *Earth System Science Data Discussions* (April) DOI: https://doi.org/10.5194/essd-2021-71

Harper A, Baker IT, Denning AS, Randall DA, Dazlich D, Branson M. 2014. Impact of evapotranspiration on dry season climate in the Amazon forest. *Journal of Climate* **27** (2): 574–591 DOI: 10.1175/JCLI-D-13-00074.1

Hengl T, De Jesus JM, Heuvelink GBM, Gonzalez MR, Kilibarda M, Blagotić A, Shangguan W, Wright MN, Geng X, Bauer-Marschallinger B, et al. 2017. *SoilGrids250m: Global gridded soil information based on machine learning*. DOI: 10.1371/journal.pone.0169748

Jehn FU, Bestian K, Breuer L, Kraft P, Houska T. 2020. Using hydrological and climatic catchment clusters to explore drivers of catchment behavior. *Hydrology and Earth System Sciences* **24** (3): 1081–1100 DOI: 10.5194/hess-24-1081-2020

L'vovich MIL. 1979. *World Water Resources and their Future*. AGU: Washington, D.C.

Lyon SW, Troch PA. 2010. Development and application of a catchment similarity index for subsurface flow. *Water Resources Research* **46** (3): 1–13 DOI: 10.1029/2009WR008500

MacQueen J. 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* **1** (14): 281–297 Available at: http://books.google.de/books?hl=de&lr=&id=IC4Ku_7dBFUC&oi=fnd&pg=PA281&dq=Mac

Queen+some+methods+for+classification&ots=nNTcK1IdoQ&sig=fHzdVcbvmYJ-lTNHu1HncmOFOkM#v=onepage&q=MacQueen some methods for classification&f=false

Marutho D, Hendra Handaka S, Wijaya E, Muljono. 2018. The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News. *Proceedings - 2018 International Seminar on Application for Technology of Information and Communication: Creative Technology for Human Life, iSemantic 2018*: 533–538 DOI: 10.1109/ISEMANTIC.2018.8549751

McDonnell JJ, Woods R. 2004. On the need for catchment classification. *Journal of Hydrology* **299** (1–2): 2–3 DOI: 10.1016/j.jhydrol.2004.09.003

Merten GH, Minella JPG. 2013. The expansion of Brazilian agriculture: Soil erosion scenarios. *International Soil and Water Conservation Research* **1** (3): 37–48 DOI: 10.1016/S2095-6339(15)30029-0

Milligan GW, Cooper MC. 1987. Methodology Review: Clustering Methods. *Applied Psychological Measurement* **11** (4): 329–354 DOI: 10.1177/014662168701100401

Nainggolan R, Perangin-Angin R, Simarmata E, Tarigan AF. 2019. Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method. *Journal of Physics: Conference Series* **1361** (1) DOI: 10.1088/1742-6596/1361/1/012015

Nobre AD, Cuartas LA, Hodnett M, Rennó CD, Rodrigues G, Silveira A, Waterloo M, Saleska S. 2011. Height Above the Nearest Drainage - a hydrologically relevant new terrain model. *Journal of Hydrology* **404** (1–2): 13–29 DOI: 10.1016/j.jhydrol.2011.03.051

Oliveira PTS, Almagro A, Colman CB, Kobayashi ANA, Meira Neto AA, Rodrigues DBB, Gupta H V. 2019. NEXUS of Water-Food-Energy-Ecosystem Services in the Brazilian Cerrado. In *Water and Climate: Modeling in Large Basins*, Silva RCV da, , Tucci CEM, , Scott CA (eds).ABRHidro; 7–35.

Oliveira PTS, Nearing MA, Wendland E. 2015. Orders of magnitude increase in soil erosion associated with land use change from native to cultivated vegetation in a Brazilian savannah environment. *Earth Surface Processes and Landforms* **40** (11): 1524–1532 DOI: 10.1002/esp.3738

Pietola L, Horn R, Yli-Halla M. 2005. Effects of trampling by cattle on the hydraulic and mechanical properties of soil. *Soil and Tillage Research* **82** (1): 99–108 DOI:

10.1016/j.still.2004.08.004

Rao AR, Srinivas V V. 2006. Regionalization of watersheds by fuzzy cluster analysis. *Journal of Hydrology* **318** (1–4): 57–79 DOI: 10.1016/j.jhydrol.2005.06.004

Rodrigues DBB, Gupta H V, Mediondo EM, Oliveira PTS. 2015. Assessing uncertainties in surface water security: An empirical multimodel approach. *Water Resources Research* **51** (11): 9013–9028 DOI: 10.1002/2014WR016691.Received

Sawicz K, Wagener T, Sivapalan M, Troch PA, Carrillo G. 2011. Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA. *Hydrology and Earth System Sciences* **15**: 2895–2911 DOI: 10.5194/hess-15-2895-2011

Sawicz KA, Kelleher C, Wagener T, Troch P, Sivapalan M, Carrillo G. 2014. Characterizing hydrologic change through catchment classification. *Hydrology and Earth System Sciences* **18** (1): 273–285 DOI: 10.5194/hess-18-273-2014

Shi N, Liu X, Guan Y. 2010. Research on k-means clustering algorithm: An improved k-means clustering algorithm. *3rd International Symposium on Intelligent Information Technology and Security Informatics, IITSI 2010*: 63–67 DOI: 10.1109/IITSI.2010.74

Singleton PL, Boyes M, Addison B. 2000. Effect of treading by dairy cattle on topsoil physical conditions for six contrasting soil types in Waikato and Northland, New Zealand, with implications for monitoring. *New Zealand Journal of Agricultural Research* **43** (4): 559–567 DOI: 10.1080/00288233.2000.9513453

Sivakumar B. 2008. Dominant processes concept, model simplification and classification framework in catchment hydrology. *Stochastic Environmental Research and Risk Assessment* **22** (6): 737–748 DOI: 10.1007/s00477-007-0183-5

Sivakumar B, Singh VP, Berndtsson R, Khan SK. 2013. Catchment Classification Framework in Hydrology: Challenges and Directions. *Journal of Hydrologic Engineering* **20** (1): A4014002 DOI: 10.1061/(asce)he.1943-5584.0000837

Sivapalan M. 2005. Pattern, Process and Function: Elements of a Unified Theory of Hydrology at the Catchment Scale. In *Encyclopedia of Hydrological Sciences*, Anderson M (ed.).John Wiley: London; 193–219. DOI: 10.1002/0470848944.hsa012

Syakur MA, Khotimah BK, Rochman EMS, Satoto BD. 2018. Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster.

IOP Conference Series: Materials Science and Engineering **336** (1) DOI: 10.1088/1757-899X/336/1/012017

Trimble SW, Mendel AC. 1995. The cow as a geomorphic agent — A critical review. *Geomorphology* **13** (1–4): 233–253 DOI: 10.1016/0169-555X(95)00028-4

Troch PA, Carrillo G, Sivapalan M, Wagener T, Sawicz K. 2013. Climate-vegetation-soil interactions and long-term hydrologic partitioning: signatures of catchment co-evolution. *Hydrology and Earth System Sciences* **17** (6): 2209–2217 DOI: 10.5194/hess-17-2209-2013

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 DOI: 10.3390/w11050910

Wagener T, Sivapalan M, McGlynn B. 2008. Catchment Classification and Services-Toward a New Paradigm for Catchment Hydrology Driven by Societal Needs. In *Encyclopedia of Hydrological Sciences*1–12. DOI: 10.1002/0470848944.hsa320

Wagener T, Sivapalan M, Troch P, Woods R. 2007. Catchment Classification and Hydrologic Similarity. *Geography Compass* **1** (4): 901–931

Wang Z, Lai C, Chen X, Yang B, Zhao S, Bai X. 2015. Flood hazard risk assessment model based on random forest. *Journal of Hydrology* **527**: 1130–1141 DOI: 10.1016/j.jhydrol.2015.06.008

Werth D, Avissar R. 2004. The Regional Evapotranspiration of the Amazon. *Journal of Hydrometeorology* **5** (1): 100–109 DOI: 10.1175/1525-7541(2004)005<0100:TREOTA>2.0.CO;2

Westerberg IK, McMillan HK. 2015. Uncertainty in hydrological signatures. *Hydrology and Earth System Sciences* **19**: 3951–3968 DOI: 10.5194/hess-19-3951-2015

Xavier AC, King CW, Scanlon BR. 2016. Daily gridded meteorological variables in Brazil (1980-2013). *International Journal of Climatology* **2659** (October 2015): 2644–2659 DOI: 10.1002/joc.4518

GENERAL CONCLUSIONS

The results of **Chapter 1** show that for the long-term mean monthly analysis, HadGEM2-ES and Eta/HadGEM2-ES well simulated rainy and dry seasons in the Amazon, Atlantic Forest, and Cerrado biomes, expressing the suitability of the GCMs to simulate mean fields of precipitation in large areas, with HadGEM2-ES presenting itself as a viable alternative for larger Brazilian biomes. In turn, Eta/MIROC5 showed great improvements when compared to its driving-GCM MIROC5. The downscaling brought the simulated means close to the observational means in most biomes. The long-term mean seasonal analysis showed that the Eta RCM modifies the range of precipitation, with less reliability of models to simulate means in the dry season (JJA and SON). It is recommended the following model for each biome: HadGEM2-ES for the Amazon, Eta/HadGEM2-ES for the Atlantic Forest, the Cerrado, and the Pampa, and Eta/MIROC5 for the Caatinga and the Pantanal. The presented results show that downscaled models have several biases, which can be originated from the driving GCMs, introduced by the downscaling RCM, and related to uncertainties in the observational data. To not impact later hydrological applications, the biases in climate change and projections must be addressed. Aside, results of this chapter shows that further hydrological investigations using climate change projections are possible to be performed with accuracy.

The CABra large-sample dataset is introduced in **Chapter 2**. It comprises more than 100 topography, climate, streamflow, groundwater, soil, geology, land-use, and land cover, and hydrologic disturbance attributes for 735 catchments in Brazil derived from several sources, such as observed time series, observed and modeled gridded data, remote sensing data, and reanalysis data. The development of a novel and comprehensive large-sample dataset opens up several opportunities to test and develop a hypothesis in a unique environment like Brazil. The results found here along with the comprehensive dataset made available, play a key role in advancing the scientific understanding of climate-landscape-hydrology interactions, aside from aiding catchment classification efforts that will ultimately unravel the underlying dominant controls of Brazilian regional hydrology across space and time.

The results of **Chapter 3** showed the added value of satellite rainfall products in hydrological modeling of daily river discharge. The performance of the satellite-based products in simulating daily precipitation over Brazil was better than ERA5 for five of six biomes.

SM2RAIN-ASCAT presented the lowest bias values, while GPM+SM2RAIN presented the lowest values of RMSE. Daily river discharge was also better modeled with SM2RAIN-ASCAT and GPM+SM2RAIN in terms of KGE during calibration. All products presented low values of bias and more than 82% of catchments have median values of KGE greater than 0.5. Satellite rainfall products performed well in estimating hydrologic signatures, being better predictors than ground-based observations for Atlantic Forest, Cerrado, and Caatinga biomes, aside from being as good as ground-based rainfall products in the Amazon. As shown in this chapter, satellite-based rainfall products are a valuable tool for data-scarce regions, due to their low-latency and global-land range, allowing continuous and high-quality water resources monitoring, which improves the capacity of impact mitigation of extremes events.

Finally, the catchment classification, presented in **Chapter 4**, grouped the Brazilian catchments into six groups. Group 1 ("Non-seasonal catchments") are mainly located in the south, are in an energy-limited condition, and do not present rainy/dry season. Group 2 ("Dry-catchments") are mostly located northeast, with high evapotranspiration through the year, sometimes larger than precipitation, leading to a strong dry season. Group 3 ("Rainforest catchments") are distributed primarily in mountainous areas of the Atlantic Forest, with high values of precipitation and streamflow. Group 4 ("Savannah catchments") are mainly located in the Cerrado, and presents well-defined rainy and dry seasons, with streamflow being controlled by the aridity index. Group 5 ("Extremely-dry catchments") is the extreme condition of Group 2, with high values of evapotranspiration and low amounts of precipitation through the year, leading to an ephemeral condition for its catchments. Finally, Group 5 ("Extremely-wet catchments") is the extreme condition of Groups 1 and 3, with very high amounts of precipitation and large areas, being mostly controlled by landscape features. There was also noted a transitional behavior from one group to another (in order: "Extremely-dry" to "Dry" to "Savannah" to "Rainforest" to "Extremely-wet"). The results of this chapter are the first stage towards a better understanding of catchment hydrology in Brazil (and also tropical hydrology). More specifically, results may be useful for the understanding of the hydrological processes involved in the catchment groups, making it possible to further regionalize hydrological information and better predict how responses to climate change are likely to be.

Three of the chapters of this doctoral thesis are already published in *Atmospheric Research*, *Hydrology and Earth System Sciences*, and *Journal of Hydrology* (Chapters 1, 2, and

3, respectively), while one is being prepared for submission in a special issue of *Hydrological Processes* (Chapter 4). The development and results of the chapters pave the way for a better understanding of different hydrologic behavior and their drivers related to climate, landscape, and hydrology in Brazilian catchments. Additionally, this work is an open discussion about catchment hydrology in Brazil, in which the structure and core were built by authors, co-authors, anonymous and non-anonymous reviewers, journal editors, and collaborators.

APPENDIX

Supplementary figures and tables related to the first chapter “PERFORMANCE EVALUATION OF ETA/HADGEM2-ES AND ETA/MIROC5 PRECIPITATION SIMULATIONS OVER BRAZIL”.

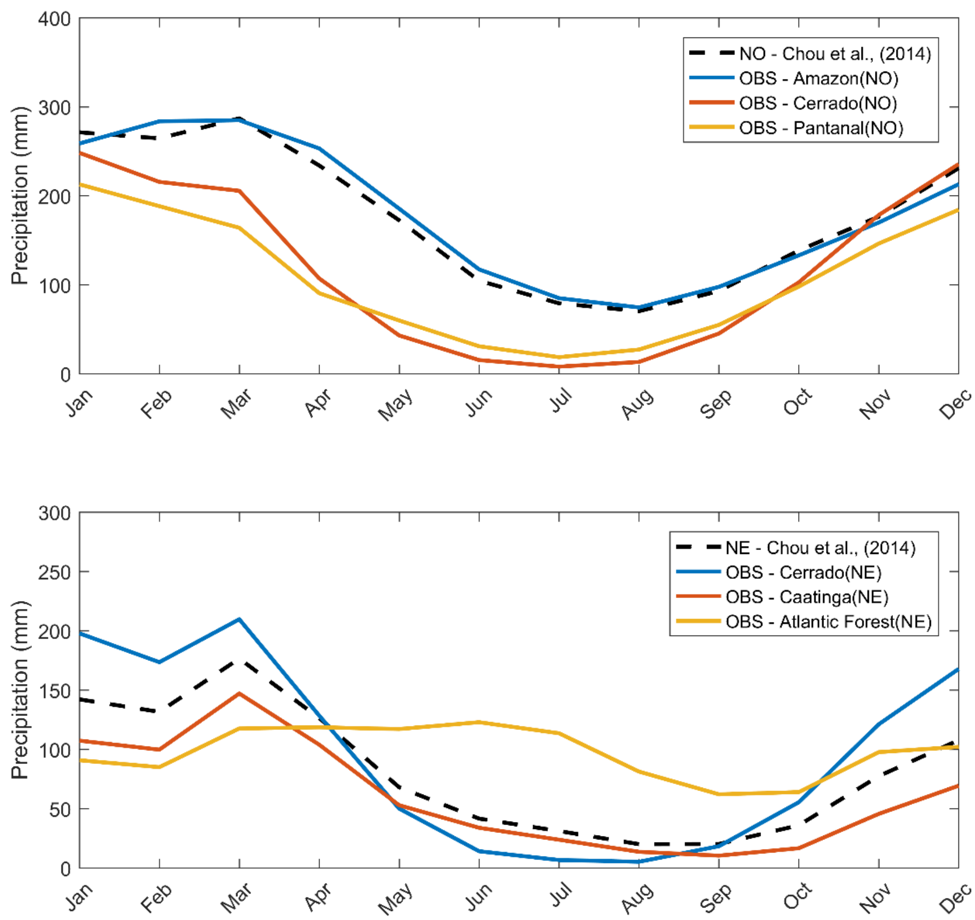

Figure S1. Long-term mean monthly rainfall for the 1980-2005 period for observations using the division adopted by Chou et al., (2014) and the portion of Brazilian biomes within the division. We can see that the area-averaged value using administrative regions in Chou et al., (2014) is not capable to reproduce the internal variability of the biomes. NO and NE are the regions adopted in Chou et al., (2014) and are represented by dashed lines. Solid lines are the mean monthly values for the biome within NO and NE.

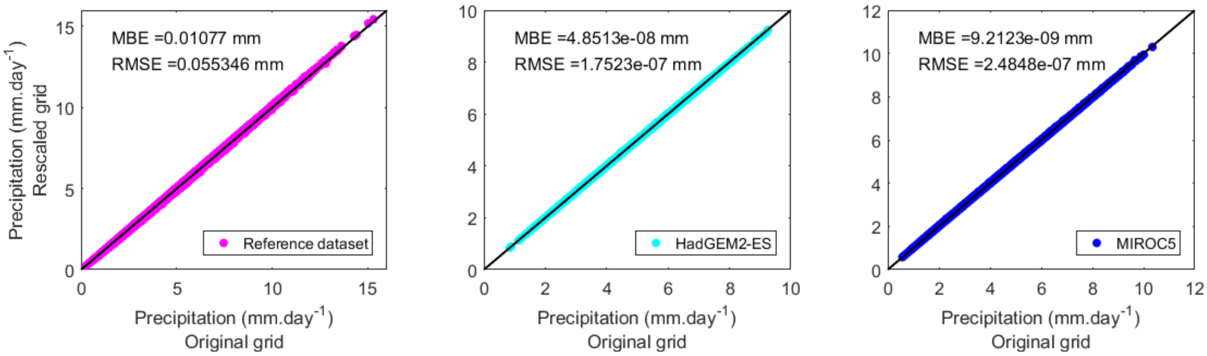

Figure S2. Scatter plot showing the relationship between the reference (purple), HadGEM2-ES (cyan), and MIROC5 (blue) in your original and rescaled grids. The proximity of data distribution to the 1:1 line shows that no significant error was added to the precipitation simulations.

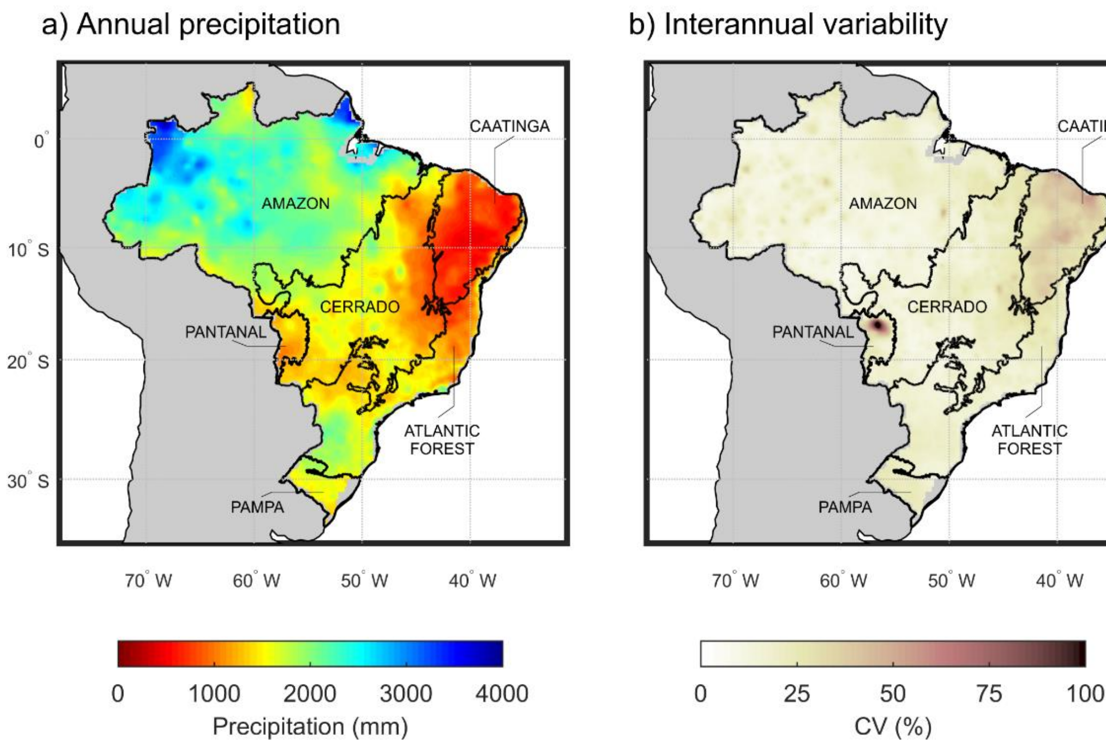

Figure S3. Spatial distribution of the a) mean annual precipitation and b) annual coefficient of variation over the Brazilian biomes from 1980 to 2005. Here we can note that the lowest values are in the northeastern portion, on Caatinga biome, while the highest values were found in the northern portion, on Amazon biome. Due the lowest values in northeastern portion, there we found the high interannual variability. Pantanal also show high values of annual variability.

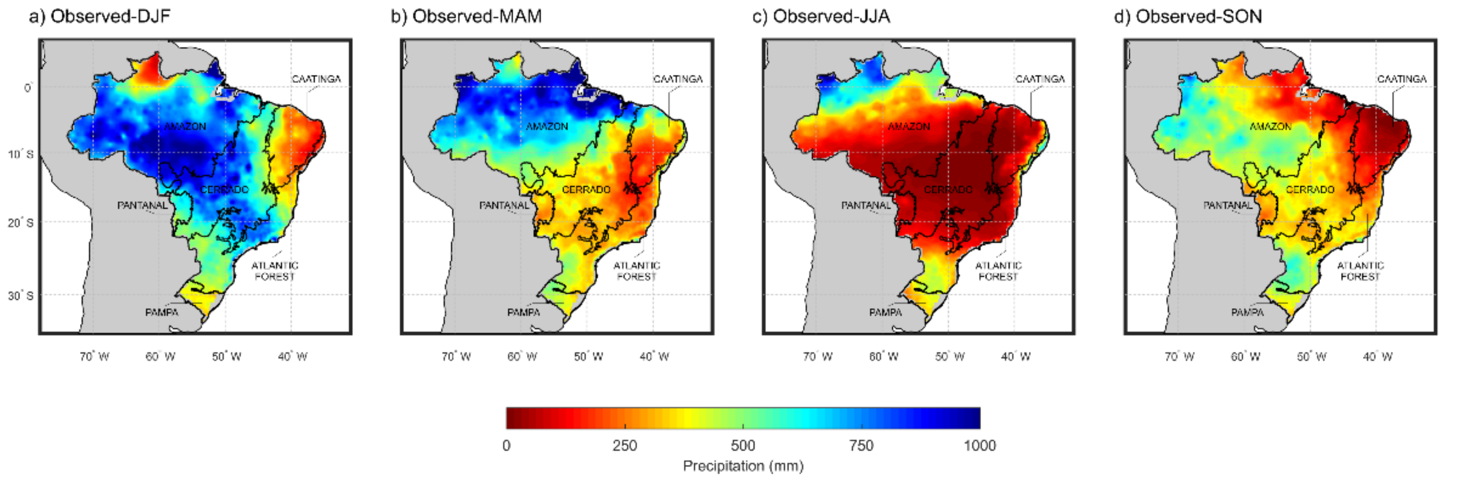

Figure S4. Spatial distribution of mean seasonal precipitation over Brazil from 1980 to 2005. DJF season (a) is the rainy season and is very influenced by the activity of the South Atlantic Convergence Zone (SACZ), while the Intercontinental Convergence Zone (ITCZ) is responsible by the rains in the MAM season (b) in Brazil. The JJA season (c) is defined as the dry season and presents the lowest amounts of rain of all season, which could reach up to 10 mm per month. In the SON season (d), the SACZ back to acting in the rain regime throughout Brazil, starting the monsoon period.

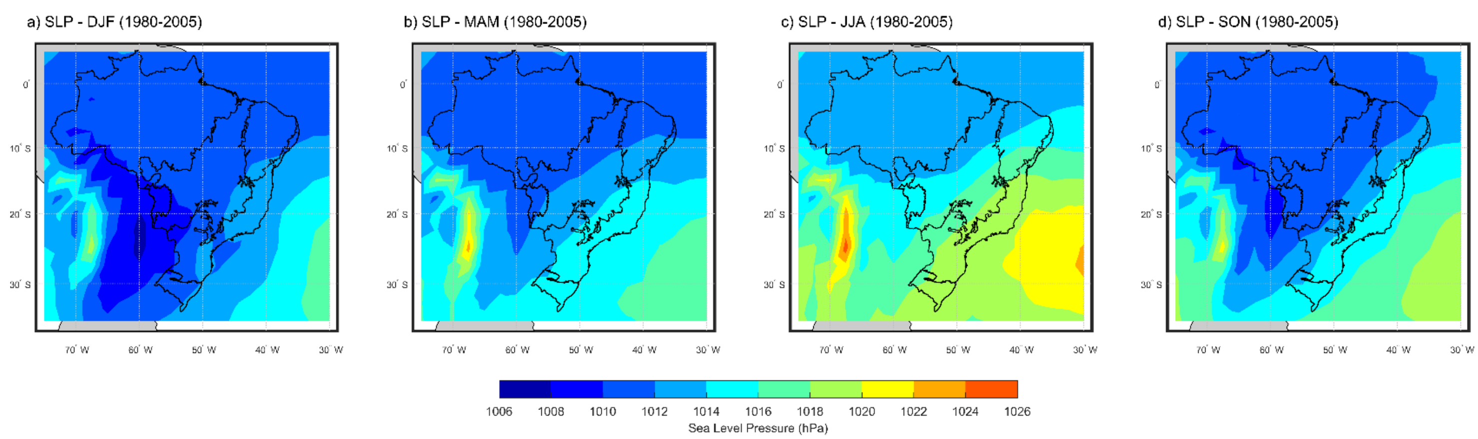

Figure S5. Spatial distribution of mean sea level pressure (SLP) over Brazil from 1980 to 2005. DJF season (a) is the rainy season and is very influenced by the activity of the South Atlantic Convergence Zone (SACZ), while the Intercontinental Convergence Zone (ITCZ) is responsible by the rains in the MAM season (b) in Brazil, both influenced by the low-pressure systems which provide the convective activity. The JJA season (c) is defined as the dry season and presents the lowest amounts of rain of all season, which could reach up to 10 mm per month, and it is strongly influenced by the high-pressure systems that dominate Brazil in this period. In the SON season (d), the SACZ back to acting in the rain regime throughout Brazil, starting the monsoon period.

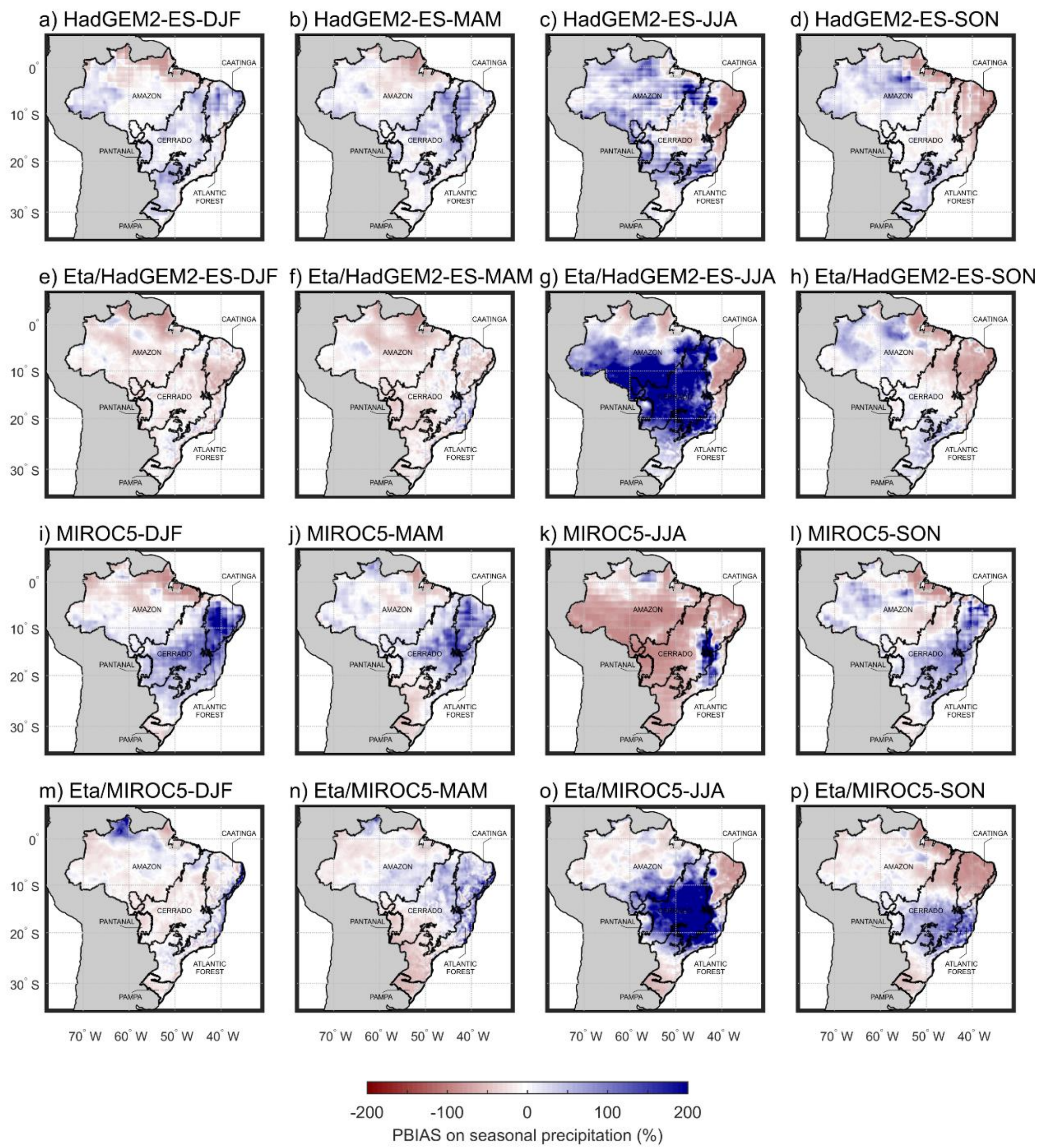

Figure S6. Relative biases (PBIAS) in each season (DJF, MAM, JJA and SON) simulated precipitation in Brazilian biomes. a) to d) represent the PBIAS for the HadGEM2-ES for all seasons; e) to h) represent the PBIAS for the Eta/HadGEM2-ES for all seasons; i) to l) represent the PBIAS for the MIROC5 for all seasons; and m) to p) represent the PBIAS for the Eta/MIROC5 for all seasons. Shades of blue indicate a positive PBIAS while shades of red indicate a negative PBIAS.

Table S1. Long-term mean monthly precipitation for the period from 1980 to 2005 for Brazilian biomes. From these mean monthly values, we generated the aggregated values of mean seasonal (DJF, MAM, JJA and SON) and annual precipitation.

Biome	Model	J	F	M	A	M	J	J	A	S	O	N	D
Amazon	Observed	259	284	285	253	186	117	85	75	98	133	170	213
	HadGEM2-ES	255	243	240	223	198	156	109	98	113	144	188	237
	Eta/HadGEM2-ES	199	209	203	187	152	144	117	99	116	143	171	189
	MIROC5	216	220	285	260	161	85	45	40	67	137	201	211
	Eta/MIROC5	245	242	273	251	168	112	83	69	90	113	149	217
Atlantic Forest	Observed	189	160	135	104	96	81	70	59	99	126	148	179
	HadGEM2-ES	214	200	156	109	82	69	71	67	106	131	172	201
	Eta/HadGEM2-ES	169	154	136	99	85	89	97	98	125	144	153	168
	MIROC5	257	211	185	96	50	37	34	41	89	163	198	252
	Eta/MIROC5	209	169	156	112	78	69	73	92	131	171	185	213
Cerrado	Observed	248	216	206	107	43	15	8	13	45	103	178	236
	HadGEM2-ES	277	279	245	142	47	21	17	14	46	96	177	251
	Eta/HadGEM2-ES	195	200	170	95	41	27	35	38	60	98	140	208
	MIROC5	353	313	298	139	24	7	4	4	26	130	258	375
	Eta/MIROC5	213	203	206	127	42	24	23	40	96	149	175	217
Caatinga	Observed	108	107	144	104	53	35	24	14	12	19	52	75
	HadGEM2-ES	106	196	189	157	63	13	5	3	4	13	32	84
	Eta/HadGEM2-ES	56	119	127	94	44	15	10	5	5	8	20	41
	MIROC5	210	215	267	196	49	15	11	7	6	27	94	193
	Eta/MIROC5	96	149	190	157	72	20	9	4	8	11	25	48
Pampa	Observed	123	150	122	171	133	132	130	98	141	150	127	110
	HadGEM2-ES	157	146	122	144	134	113	132	119	157	183	170	138
	Eta/HadGEM2-ES	117	101	85	121	136	140	130	123	137	168	153	120
	MIROC5	108	81	82	101	76	56	46	53	91	128	115	97
	Eta/MIROC5	113	97	70	75	68	50	66	67	79	91	84	116
Pantanal	Observed	213	188	164	91	60	31	19	27	55	98	146	184
	HadGEM2-ES	269	234	210	122	65	56	48	26	54	96	158	211
	Eta/HadGEM2-ES	153	136	89	67	56	50	58	51	75	95	111	152
	MIROC5	283	251	216	85	9	3	3	6	37	124	182	256
	Eta/MIROC5	164	126	111	66	25	30	33	45	73	115	141	160

Table S2. Long-term mean seasonal and annual precipitation for the period from 1980 to 2005 for Brazilian biomes.

Biome	Model	DJF	MAM	JJA	SON	Annual
Amazon	Observed	755	724	277	401	2156
	HadGEM2-ES	736	661	362	445	2204
	Eta/HadGEM2-ES	597	542	360	431	1929
	MIROC5	648	706	170	405	1929
	Eta/MIROC5	705	691	264	352	2012
Atlantic Forest	Observed	527	335	211	373	1446
	HadGEM2-ES	615	347	208	409	1580
	Eta/HadGEM2-ES	491	320	284	422	1517
	MIROC5	720	332	112	450	1614
	Eta/MIROC5	591	347	234	487	1659
Cerrado	Observed	699	356	37	326	1418
	HadGEM2-ES	807	435	52	319	1613
	Eta/HadGEM2-ES	604	305	101	298	1307
	MIROC5	1041	460	14	413	1928
	Eta/MIROC5	632	376	87	421	1516
Caatinga	Observed	291	302	74	83	750
	HadGEM2-ES	386	408	21	48	864
	Eta/HadGEM2-ES	215	266	30	33	544
	MIROC5	619	512	33	128	1291
	Eta/MIROC5	293	419	33	44	789
Pampa	Observed	383	426	360	418	1587
	HadGEM2-ES	441	401	365	510	1716
	Eta/HadGEM2-ES	339	342	393	458	1532
	MIROC5	286	260	155	334	1035
	Eta/MIROC5	326	213	183	254	975
Pantanal	Observed	585	314	77	299	1276
	HadGEM2-ES	715	396	129	309	1550
	Eta/HadGEM2-ES	441	212	159	281	1093
	MIROC5	791	310	12	343	1456
	Eta/MIROC5	450	203	108	328	1089

Table S3. Rank of error ($\in_j$) defined by Equation 5 and Table 1 for all rainy and dry periods, seasonal cycle and annual cycle properties for all the Brazilian biomes. $\in_j$ is the rank error between the observed, "*obs*", value and simulated by the model "*j*".

Biome	Model	Rainy/dry period			Seasonal cycle					Annual cycle					Overall
		DM		$\in_j$	*RMSE*		*CC*		$\in_j$	$\bar{P}$		*CV*		$\in_j$	$\in_j$
		obs	*j*		*obs*	*j*	*obs*	*j*		*obs*	*j*	*obs*	*j*		
Amazon	HadGEM2-ES	3	3	0,00	-	175	1	0,81	0,26	2156	2208	0,15	0,1	0,08	0,14
	Eta/HadGEM2-ES	3	3	0,00	-	202	1	0,70	0,30	2156	1929	0,15	0,14	0,34	0,23
	MIROC5	3	3	0,00	-	180	1	0,82	0,27	2156	1929	0,15	0,13	0,34	0,21
	Eta/MIROC5	3	3	0,00	-	122	1	0,90	0,18	2156	2012	0,15	0,12	0,30	0,14
Atlantic Forest	HadGEM2-ES	5	5	0,00	-	118	1	0,84	0,20	1446	1580	0,19	0,2	0,25	0,16
	Eta/HadGEM2-ES	5	5	0,00	-	129	1	0,74	0,21	1446	1517	0,19	0,24	0,10	0,12
	MIROC5	5	5	0,00	-	165	1	0,78	0,27	1446	1614	0,19	0,18	0,32	0,21
	Eta/MIROC5	5	5	0,00	-	193	1	0,55	0,32	1446	1659	0,19	0,2	0,34	0,23
Cerrado	HadGEM2-ES	5	5	0,00	-	107	1	0,90	0,23	1418	1613	0,19	0,15	0,20	0,15
	Eta/HadGEM2-ES	5	5	0,00	-	72	1	0,89	0,15	1418	1307	0,19	0,2	0,12	0,10
	MIROC5	5	5	0,00	-	197	1	0,85	0,42	1418	1928	0,19	0,18	0,56	0,40
	Eta/MIROC5	5	5	0,00	-	98	1	0,88	0,21	1418	1516	0,19	0,19	0,11	0,10
Caatinga	HadGEM2-ES	8	8	0,00	-	86	1	0,96	0,17	750	864	0,34	0,37	0,12	0,10
	Eta/HadGEM2-ES	8	9	0,50	-	94	1	0,94	0,18	750	544	0,34	0,43	0,24	0,16
	MIROC5	8	7	0,50	-	204	1	0,91	0,40	750	1291	0,34	0,24	0,61	0,41
	Eta/MIROC5	8	8	0,00	-	128	1	0,87	0,25	750	789	0,34	0,27	0,04	0,09
Pampa	HadGEM2-ES	1	0	0,06	-	68	1	0,59	0,23	1587	1716	0,22	0,16	0,12	0,11
	Eta/HadGEM2-ES	1	1	0,00	-	68	1	0,44	0,23	1587	1532	0,22	0,17	0,01	0,04
	MIROC5	1	9	0,44	-	69	1	0,46	0,24	1587	1035	0,22	0,18	0,40	0,33
	Eta/MIROC5	1	10	0,50	-	85	1	0,08	0,30	1587	975	0,22	0,25	0,41	0,38
Pantanal	HadGEM2-ES	7	6	0,33	-	55	1	0,97	0,15	1276	1550	0,31	0,14	0,33	0,23
	Eta/HadGEM2-ES	7	8	0,33	-	96	1	0,89	0,26	1276	1093	0,31	0,16	0,22	0,17
	MIROC5	7	6	0,33	-	129	1	0,96	0,35	1276	1456	0,31	0,19	0,24	0,19
	Eta/MIROC5	7	7	0,00	-	89	1	0,89	0,24	1276	1089	0,31	0,16	0,22	0,17