UNIVERSIDADE FEDERAL DE VIÇOSA

DANIEL ALTHOFF

MODELAGEM DA DISPONIBILIDADE HÍDRICA EM BACIAS HIDROGRÁFICAS DO CERRADO

VIÇOSA – MINAS GEIRAS
2021

DANIEL ALTHOF

MODELAGEM DA DISPONIBILIDADE HÍDRICA EM BACIAS HIDROGRÁFICAS DO CERRADO

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Engenharia Agrícola, para obtenção do título de *Doctor Scientiae*.

Orientador: Lineu Neiva Rodrigues

Coorientador: Demetrius David da Silva

VIÇOSA – MINAS GERAIS
2021

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
A467m
2021

Althoff, Daniel, 1991-
Modelagem da disponibilidade hídrica em bacias hidrográficas do Cerrado / Daniel Althoff. – Viçosa, MG, 2021.
86 f. : il. (algumas color.) ; 29 cm.

Orientador: Lineu Neiva Rodrigues.
Tese (doutorado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 74-86.

1. Desenvolvimento de recursos hídricos - Modelos matemáticos. 2. Regionalização. 3. Redes neurais (Computação). 4. Bacias hidrográficas. I. Universidade Federal de Viçosa. Departamento de Engenharia Agrícola. Programa de Pós-Graduação em Egenharia Agrícola. II. Título.

CDD 22. ed. 630.25148

DANIEL ALTHOF

MODELAGEM DA DISPONIBILIDADE HÍDRICA EM BACIAS HIDROGRÁFICAS DO CERRADO

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Engenharia Agrícola, para obtenção do título de *Doctor Scientiae*.

APROVADA: 5 de março de 2021.

Assentimento:

Daniel Althoff
Autor

Lineu Neiva Rodrigues
Orientador

AGRADECIMENTOS

Primeiramente à minha família, pelo incentivo, confiança e amor.

À minha esposa Helizani, por todo apoio, amor e orientação.

À Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Engenharia Agrícola, pela oportunidade de realização do curso.

Ao Lineu Neiva Rodrigues, pela orientação, confiança, amizade e apoio.

Ao professor Demetrius David da Silva, pela coorientação e ensinamentos.

A todos os meus amigos que se fizeram presente ao longo dessa jornada.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil (CNPQ) – Processo 142273/2019-8.

“It's the inspired student that continues to learn on their own. That's what separates the real achievers in the world from those who pedal along, finishing assignments.”

(Neil deGrasse Tyson)

RESUMO

ALTHOFF, Daniel, D.Sc., Universidade Federal de Viçosa, março de 2021. **Modelagem da disponibilidade hídrica em bacias hidrográficas do Cerrado.** Orientador: Lineu Neiva Rodrigues. Coorientador: Demetrius David da Silva.

Diversas regiões do Cerrado não dispõem de informações hidrológicas básicas e confiáveis para o desenvolvimento da adequada gestão e manejo de recursos hídricos. Informações relacionadas à quantidade e período de disponibilidade da água são fundamentais para garantir que a alocação do recurso assegure o bem-estar social e desenvolvimento sustentável local. A estimativa da disponibilidade hídrica em regiões não monitoradas, ou com escassez de dados, é um desafio de longa data para a hidrologia. A forma mais comum de se estimar a disponibilidade hídrica em regiões não monitoradas se dá por meio de técnicas de regionalização. A regionalização se baseia na transferência de informações de regiões monitoradas para regiões não monitoradas. Isso pode ser feito por meio da doação de parâmetros de modelos hidrológicos calibrados em regiões monitoradas considerando tanto a proximidade espacial quanto similaridade física da região/bacia não monitorada. Uma alternativa aos métodos de regionalização é se trabalhar com modelos hidrológicos de larga escala, como modelos regionais. Neste caso, os modelos baseados em aprendizado de máquina têm se popularizado, uma vez que conseguem aprender a partir da diversidade de dados que lhes são fornecidos. Portanto, há a oportunidade de se desenvolver um modelo regional capaz de aprender frente à diversidade de cenários e melhor predizer os efeitos de mudanças climáticas e de uso da terra, e fornecendo estimavas mais adequadas da disponibilidade hídrica de bacias hidrográficas. O objetivo geral desta tese, estruturada em forma de artigos, foi avaliar a disponibilidade hídrica em bacias hidrográficas do Cerrado brasileiro. Para isso, primeiramente se calibrou um modelo hidrológico para 411 bacias hidrográficas monitoradas individualmente, cujos limites estão dentro ou interseccionam o Cerrado. Dentre as técnicas de regionalização avaliadas, a doação de parâmetros por proximidade espacial apresentou o melhor desempenho e foi utilizada para se simular séries de vazão em ottobacias de nível 5 em todo o Cerrado. Essas séries de vazão foram então usadas para caracterizar a disponibilidade hídrica, períodos de maior pressão sobre o recurso, e identificar regiões onde o uso sustentável de água pode ser aprimorado por meio de técnicas de armazenamento de água excedente do período de chuvas. Também se desenvolveu um modelo regional, baseado em redes neurais recorrentes, que levou em consideração informações compiladas para as 411 bacias hidrográficas inicialmente delimitadas dentro do

Cerrado, como características topográficas, de clima, do uso e cobertura da terra e do solo. Este modelo se mostrou robusto às mudanças de uso e cobertura da terra, principalmente no que se refere às predições realizadas para períodos de vazões mínimas. Para este modelo regional, foi possível explorar quais as variáveis de maior influência na predição de vazão, e como elas se relacionam com a disponibilidade hídrica.

Palavras-chave: Regionalização. Modelagem hidrológica. Redes neurais artificiais. Ottobacias.

ABSTRACT

ALTHOFF, Daniel, D.Sc., Universidade Federal de Viçosa, March, 2021. **Hydrological modeling for water availability in hydrographic basins of the Brazilian Cerrado biome.** Adviser: Lineu Neiva Rodrigues. Co-adviser: Demetrius David da Silva.

Many regions of the Brazilian Cerrado biome do not have basic and reliable information for the adequate development of water resources planning and management. Information concerning the amount and timing of water availability is fundamental for guaranteeing the proper allocation of the resource to secure the social well-being and the local sustainable development. Predictions in ungauged basins (PUB), or in data-scarce regions, have long been a challenge for hydrology. PUB has been addressed mainly by regionalization techniques. Regionalization is based on the transfer of information obtained from gauged regions to ungauged regions. This may be achieved with the donation of hydrological model parameters calibrated in gauged basins by the spatial proximity or the physical similarity to the ungauged basin. Another alternative to regionalization is using hydrological models of large scales as regional models. In this case, models based on machine learning have gained popularity, since they can learn from the diversity of data provided to them. Thus, there is an opportunity to develop a regional model capable of learning from the diversity of scenarios and better predict the effects of climate change and changes in land use and land cover, providing more adequate estimates of the water availability in basins. The main objective of this thesis, which was structured in articles, is to assess the water availability in hydrographic basins of the Cerrado. To reach this goal, we first calibrated a hydrological model for 411 gauged basins individually, which boundaries are within or intersect the Cerrado. Among the regionalization techniques, the donation of parameters by spatial proximity showed the best performance and was used to simulate runoff series in all level 5 ottobasins in the Cerrado. These runoff series were then used to characterize the water availability, periods of highest pressure on the resource, and identify regions where sustainable water use can increase, e.g., by rainwater harvesting techniques. A single regional model based on recurrent neural networks was developed considering information compiled for 411 gauged basins within the Cerrado concerning characteristics of topography, climate, land use and land cover, and soil. This model was robust to environmental changes, especially concerning predictions made under low-flow conditions.

This model was also explored with respect to which variable showed the highest importance in streamflow prediction, and how they related to water availability.

Keywords: Regionalization. Hydrological modeling. Artificial neural networks. Ottobasins.

SUMÁRIO

1. Introdução geral

O Cerrado é o segundo maior bioma do Brasil e ocupa aproximadamente 24% do território brasileiro. A importância do Cerrado é reconhecida tanto por sua biodiversidade (LAMBERS *et al.*, 2020), quanto pela importância de sua agricultura na econômica nacional (LEMES; DE ANDRADE; LOYOLA, 2020). A agricultura na região é altamente tecnificada e com alto potencial produtivo, sendo responsável por mais da metade da produção de carne bovina e soja do país (KLINK, 2014). O sucesso do bioma na agricultura advém, principalmente, de grandes investimentos em tecnologia, práticas de manejo e o uso de insumos agrícolas (HOSONO; ROCHA; HONGO, 2016; LAMBERS *et al.*, 2020), sendo esses fatores fundamentais para o desenvolvimento e destaque da agricultura brasileira no cenário mundial (CREMAQ, 2010; RADA, 2013).

Outro fator que favoreceu o desenvolvimento da agricultura no Cerrado é o clima, caracterizado em sua grande maioria como tropical de savana, apresentando estações chuvosa e seca bem definidas (ALVARES *et al.*, 2013). Em geral, a estação chuvosa dura de outubro a abril, concentrando 90% da precipitação anual. A longa estação chuvosa permite, em algumas de suas regiões, o plantio de duas safras em uma única estação (SPANGLER; LYNCH; SPERA, 2017). Além disso, o excesso de chuva da estação chuvosa converte, em grande parte, em disponibilidade hídrica ao longo dos cursos d'água das principais regiões hidrográficas do país que interseccionam o Cerrado (LIMA, 2011). Isso confere a algumas de suas regiões um elevado potencial para a expansão de áreas irrigadas (ALTHOFF; RODRIGUES, 2019; FEALQ, 2014). Segundo uma análise territorial para o desenvolvimento da agricultura irrigada (FEALQ, 2014), o Brasil apresenta potencial de área adicional irrigável de 75,2 Mha, dos quais 26,5 Mha se encontram no Cerrado. A Agência Nacional de Águas (ANA) aponta que, da área com potencial para expansão da irrigação, apenas 21,8 Mha se caracterizam como de alta aptidão solo-relevo e 11,2 Mha como de potencial efetivo, isso é, além da aptidão solo-relevo, boa qualidade logística, exclusão de áreas de proteção ambiental e presença da agricultura irrigada já estabelecida (ANA, 2017a). Dentre essas áreas, boa parte se concentra no Cerrado.

Apesar do sucesso do Cerrado ser considerado por alguns como milagroso (RADA, 2013), esse sucesso também é controverso. A expansão da agricultura foi responsável, em diversas regiões, pela fragmentação de ecossistemas, perda de biodiversidade, erosão do solo e poluição dos recursos hídricos (LAMBERS *et al.*, 2020; LEMES; DE ANDRADE; LOYOLA, 2020; LIMA; PERSSON, 2020; RADA, 2013; SOUZA *et al.*, 2020). Além disso, estudos baseados em projeções climáticas estimam o aumento da temperatura, o aumento da duração da estação seca e a redução na disponibilidade hídrica, colocando em cheque a

agricultura na região, especialmente para a agricultura de sequeiro (CHOU *et al.*, 2014; OLIVEIRA *et al.*, 2017; PIRES *et al.*, 2016). Essas projeções têm implicações ainda mais relevantes para regiões que já apresentam conflitos em relação ao uso e alocação da água (POUSA *et al.*, 2019; RODRIGUES; SILVA; FREITAS, 2014).

Dentre as alternativas destacadas para aumentar a resiliência frente às mudanças climáticas e reduzir os conflitos pelo uso da água estão a conservação das partes remanescentes do Cerrado (COSTA; PIRES, 2010; PIRES *et al.*, 2016), intensificação de práticas agrícolas nas regiões já agricultadas (SPERA, 2017), como em pastos degradados, e a melhoria da gestão e planejamento do recurso hídrico local baseado em informações hidrológicas básicas e confiáveis (POUSA *et al.*, 2019). Em regiões onde os conflitos pelo uso da água já são existentes, é importante o aprimoramento do manejo de irrigação através de técnicas de déficit controlado (COMAS *et al.*, 2019; FABEIRO; OLALLA; JUAN, 2001; GEERTS; RAES, 2009) e avaliação da possibilidade de armazenamento de água excedente da estação chuvosa, como em pequenas barragens, para melhoria da sustentabilidade hídrica local (MALVEIRA; ARAÚJO; GÜNTNER, 2012; RODRIGUES *et al.*, 2012). Quando os recursos hídricos estão disponíveis em quantidades adequadas, a intensificação da agricultura pode se dar por meio da expansão de áreas irrigadas, possibilitando a otimização do processo produtivo em uma mesma área plantada, e evitando a expansão sobre ecossistemas naturais. A expansão da área irrigada em grande escala é apresentada em diversos estudos como responsável pela redução de temperatura e anomalias relacionadas à mesma. Isso foi estudado para o centro-oeste americano, onde o clima é predominantemente úmido, com precipitação bem distribuída ao longo do ano (classificação de Köppen: Dfa) (MUELLER *et al.*, 2016; NOCCO; SMAIL; KUCHARIK, 2019), e para o estado da Califórnia, onde o clima é caracterizado como mediterrâneo de verão quente/fresco. Outros autores também discutem o possível incremento de ciclagem da precipitação causado pelo aumento da irrigação durante a estação chuvosa (LATHUILLIÈRE; COE; JOHNSON, 2016).

Contudo, diversas regiões do Cerrado não dispõem de informações confiáveis para que a gestão e o manejo de recursos hídricos possam ser feitos de forma adequada, o que tem dificultado um planejamento de alocação do recurso que assegure o bem-estar social e desenvolvimento sustentável local. A estimativa da quantidade e do período de disponibilidade de água em regiões não monitoradas ou com escassez de dados é um desafio de longa data para a hidrologia. A forma mais comum de se estimar a disponibilidade hídrica em regiões não monitoradas se dá por meio de técnicas de regionalização (GUO *et al.*, 2021; QI *et al.*, 2020).

A regionalização se baseia na transferência de informações de regiões monitoradas para regiões não monitoradas (RAZAVI; COULIBALY, 2013a). Dentre elas podemos citar as regressões, baseadas na distância física ou em atributos físicos de bacias hidrográficas, utilizadas para se estimar assinaturas hidrológicas, como, por exemplo a vazão média ou a vazão que é igualada ou excedida a uma dada frequência. Outra técnica popular é a "doação" de parâmetros de modelos hidrológicos calibrados em bacias monitoradas (ARSENAULT; BRISSETTE, 2016; POOL; VIVIROLI; SEIBERT, 2019; QI *et al.*, 2020). A doação desses parâmetros também pode se dar pela distância espacial ou similaridade física entre as bacias. A vantagem é que com os parâmetros de modelos hidrológicos se pode trabalhar com estimativa de disponibilidade hídrica em curto e médio prazos e estimativas de impactos relacionados a mudanças de clima, uso e cobertura do solo, e práticas conservacionistas (ABDULKAREEM *et al.*, 2019; GASHAW *et al.*, 2018; WANG *et al.*, 2018).

Uma alternativa aos métodos de regionalização é se trabalhar com modelos hidrológicos de larga escala, ou seja, modelos regionais, continentais ou globais. Neste sentido, modelos de aprendizado de máquina tem se popularizado nos anos recentes. Esses modelos se destacam por desempenhar melhor que modelos baseados em processos na modelagem da vazão a partir de dados meteorológicos (BAI; LIU; XIE, 2021; FAN *et al.*, 2020; KRATZERT *et al.*, 2018, 2019b). A flexibilidade desses modelos também permite que um único modelo regional/global possa aprender com as mais diversas fontes de dados (KRATZERT *et al.*, 2019c, 2020). Logo, há a oportunidade para o desenvolvimento de modelos regionais robustos, com capacidade de representar de forma adequada os efeitos das mudanças climáticas e ambientais, uma vez que esses modelos podem aprender a partir de um banco de dados com amplitude no que se refere à diversidade observada de clima e de mudanças de uso e cobertura da terra. Com a popularização e crescente número de técnicas de "inteligência artificial explicável" (MAKSYMIUK; GOSIEWSKA; BIECEK, 2020; SUNDARARAJAN; TALY; YAN, 2017), há também uma oportunidade para explorar o que esses modelos de fato aprendem, aprofundando a compreensão dos processos hidrológicos (KRATZERT *et al.*, 2019a) e aumentando a confiabilidade em seu uso.

A tese, cujo objetivo geral foi avaliar a disponibilidade hídrica em bacias hidrográficas do Cerrado brasileiro, foi estruturada na forma de artigos. No primeiro artigo são ajustados modelos hidrológicos às bacias monitoradas e avaliados métodos de regionalização. A contribuição científica deste artigo está relacionada a aprimoração das técnicas de regionalização geralmente utilizadas no Brasil. Estudos prévios focaram na regionalização de assinaturas hidrológicas e, já a estimativa de séries de vazões, permite uma análise mais dinâmica dos processos e condições hidrológicas no bioma. No segundo artigo

se fez a caracterização do Cerrado quanto às características hidroclimáticas, vulnerabilidade da disponibilidade hídrica e potencial para melhoria do uso sustentável de água. Este artigo apresenta uma contribuição no que se refere à gestão integrada de recursos hídricos no bioma e aponta regiões que requerem atenção ou podem melhor se beneficiar da alocação de recursos e investimento do governo. No terceiro artigo são exploradas técnicas de aprendizado de máquina para o desenvolvimento de um modelo hidrológico regional único e mais bem adaptado às mudanças de uso e cobertura da terra ao longo do tempo. Este artigo avança cientificamente no que se refere à capacidade de um modelo em incluir dados de uso e cobertura da terra para treinar modelos mais assertivos. A partir de técnicas explanatórias de modelos baseados em inteligência artificial, pode-se também avaliar como os diversos atributos de bacias hidrográficas contribuem nas séries de vazões.

2. Artigos científicos

2.1. Artigo 1

Predicting Runoff Series in Ungauged Basins of the Brazilian Cerrado Biome

Abstract

Information concerning water availability in a basin can be key to trustworthy and robust decisions, as well as to reduce disputes over water among its multiple users. The Cerrado region, however, lacks basic hydrological information with respect to water availability in many of its basins. Streamflow simulations and predictions in ungauged basins remain a crucial challenge in water resource planning and management and have been addressed by regionalization approaches. In this study, two regionalization frameworks for predicting runoff series in ungauged basins of the Cerrado biome were assessed. The regionalization frameworks are based on donating parameter sets of the GR5J hydrological model calibrated in gauged catchments by spatial proximity or attributes proximity (physical similarity). The performance of these approaches was assessed using a leave-one-out cross-validation for the gauged catchments. Using a multi-objective function considering both the Kling-Gupta efficiency (*KGE*) and relative Nash-Sutcliffe efficiency (*rNSE*) to calibrate the GR5J model on gauged catchments led to better regionalization performance when compared to using only the KGE as the objective function. Spatial proximity led to *KGE* and *rNSE* of 0.58 and 0.56, while attributes proximity led to *KGE* and *rNSE* of 0.56 and 0.45, respectively. Concerning the benchmark series (mean monthly streamflow from calibration period), the regionalization by spatial proximity showed an increase in median *KGE* by 0.24. Finally, the runoff time series were simulated for 4531 level 5 ottobasins in the Cerrado biome by donating parameter sets by spatial proximity. A regional-sample data set for the Cerrado (HydroCerrado) was also made available with information compiled for the 411 catchments used in this study.

Keywords: regionalization, GR5J, tropical watersheds, savanna.

1. Introduction

With an extension of approximately 200 million hectares and occupying 24% of the Brazilian territory, the Cerrado is Brazil's second largest biome. Despite its acidic soils, which are poor in organic matter and nutrients (KLINK, 2014; RADA, 2013), the Cerrado was fundamental for the ascension of the Brazilian agriculture in recent decades (CREMAQ, 2010; RADA, 2013). The region presents highly technified farms with potential for high productivity, as seen in its latest agriculture frontier, the MATOPIBA (an acronym for the states Maranhão, Tocantins, Piauí, and Bahia).

One of the reasons for the agriculture success in the Cerrado is its long wet season, which lasts from October to April, enabling double-cropping within the same season (SPANGLER; LYNCH; SPERA, 2017). In addition to climate, water availability plays a key factor in the region. The Cerrado has a high water yield, contributing to the availability of surface water in many of the main hydrographic regions of Brazil it encompasses (LIMA, 2011) and with a high potential for the expansion of irrigated areas (ALTHOFF; RODRIGUES, 2019; FEALQ, 2014).

However, over the last few years, there has been an increasing number of conflicts over water use in parts of the Cerrado where irrigated agriculture expanded with poor planning and lack of hydrological information (POUSA *et al.*, 2019). Irrigation represents alone ~67% of all water consumption in the country (ANA, 2017b). Besides, studies considering different projections of climate change have highlighted several risks for sustainable socio-economic development in the region. For instance, Chou *et al.* (2014) predicted an increase in temperature and decrease in rainfall, while others predict an increase in the length of the dry season (PIRES *et al.*, 2016) and a decrease in river discharge (OLIVEIRA *et al.*, 2017).

To guarantee social well-being and sustainable development in the region it is crucial to have enough technical information to support the planning and management of water resources. Water allocation among multiple users usually comes with conflicting interests (RODRIGUES; SILVA; FREITAS, 2014), requiring knowledge of the historic and current water availability and water demand within a basin. Pousa *et al.* (2019) suggested that the knowledge of water availability in a basin, which serves as a management unit, can be key to trustworthy and robust decisions and can contribute to the reduction of conflicts during the dry season. The Cerrado region, however, lacks basic hydrological information that can be used to develop reliable estimates of water availability in specific regions of its basins.

There are many ways to determine water availability in hydrographic basins where streamflow measurements or other hydrometric data are lacking. For example, the complex

interaction between physiographic factors such as climate, land use and land cover, topography, and geology, make the application of mathematical and computational models the best alternative to estimate target variables in ungauged basins. Such models can be valuable tools in the planning and management of water resources because they can simulate the impact of different factors on the basin's water availability. A large number of models have already been developed to assess water availability (BECK *et al.*, 2017; DEVIA; GANASRI; DWARAKISH, 2015; DOUGLAS-MANKIN; SRINIVASAN; ARNOLD, 2010; PERRIN; MICHEL; ANDRÉASSIAN, 2003; STEENHUIS *et al.*, 2009). These models should preferably be simple and easy to implement, making use of recent advances of different data sources such as satellite or remote sensing-derived products including evapotranspiration, net radiation, snow cover, or soil moisture data (CHEN; WANG, 2018; JIANG; WANG, 2019; TRAN; NIEL; WILLEMS, 2018).

The use of gridded products has excelled in hydroclimatic modeling in the past 20 years for both conceptual (lumped/bucket-style) and physically-based (distributed) models, especially in regions with data scarcity (JIANG; WANG, 2019; TRAN; NIEL; WILLEMS, 2018). However, runoff simulations and predictions in ungauged basins (PUB) remain a crucial challenge in water resource planning and management (GUO *et al.*, 2021; QI *et al.*, 2020). PUB has been addressed mainly by regionalization approaches. These approaches usually consist of "donating" hydrological model parameters calibrated at gauged catchments to predict runoff time series or hydrological signatures in ungauged catchments (GUO *et al.*, 2021). In a large sample study, Qi *et al.* (2020) assessed several regionalization approaches for 2277 gauged catchments worldwide and found that donating hydrological parameters based on spatial proximity generally outperformed donation based on physical similarity ("attributes proximity"), whereas both outperformed regression techniques. Considering that previous efforts to regionalize hydrologic information in parts of the Cerrado have mainly been addressed by regression techniques and for hydrologic signatures (LOPES *et al.*, 2017; MORAIS *et al.*, 2020; PRUSKI *et al.*, 2013), exploring the donation of hydrological parameters can be considered an advance since obtaining runoff series allows for more dynamic analyses to be carried out.

Based on the current state of hydrological data availability in the Cerrado and its importance for Brazil's agriculture, there is a need to improve information on water availability in the many unmonitored parts of the biome to help stakeholders make decisions based on scientific knowledge (POUSA *et al.*, 2019). Thus, the objectives of the present study were to (i) assess two frameworks for regionalization of hydrological model's parameters and (ii) to develop a runoff time series data set for all (4531) of the Cerrado level

5 ottobasins. Ottobasins are basins and interbasins hierarchically classified according to the topological system proposed by Otto Pfafstetter (see Furnans and Olivera, 2001). The ottoclassification is adopted in Brazil by the National System of Water Resources Information (in Portuguese: Sistema Nacional de Informações sobre Recursos Hídricos – SNIRH) for the hierarchical classification and coding of hydrologic units. The development of this data set enables the analysis of different hydrological signatures and water availability assessment across the Cerrado biome.

2. Material and Methods

2.1. Study area

The Cerrado biome comprises the entire Federal District and the state of Goiás and encompasses areas of the states of Tocantins, Maranhão, Piauí, Bahia, Minas Gerais, São Paulo, Mato Grosso do Sul, Mato Grosso and, in smaller proportions, the states of Rondônia, Pará, and Paraná (Figure 1). Ribeiro and Walter (1998) also documented disjoint areas of the Cerrado biome in the northern region of the states Amapá, Amazonas, Pará, and Roraima.

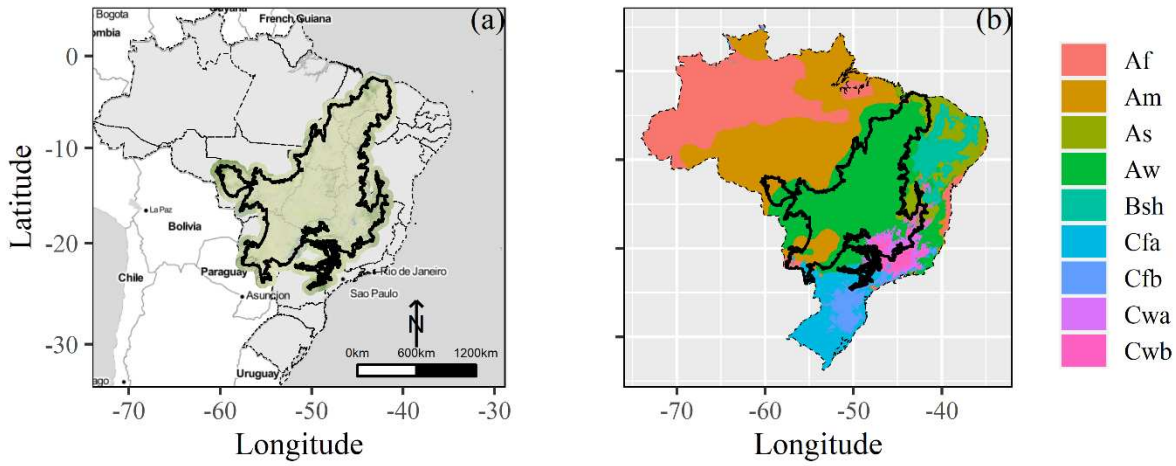

Figure 1. (a) Cerrado in relation to Brazil and **(b)** in relation to the Köppen climate classification for Brazil (data source: Alvares et al., 2013).

The climate in the Cerrado biome is predominantly classified as a tropical savanna climate (Aw) on the Köppen climate classification (ALVARES *et al.*, 2013) (Figure 1b). The Aw climate is characterized by mean air temperature above 18°C in every month of the year where the precipitation of the driest month is less than 60 mm. The tropical savanna climate is also known as a wet and dry climate because it has two well-defined seasons. In the Cerrado, about 90% of the total annual precipitation falls in the wet season, which typically lasts from October to April, while the dry season lasts from May to September (KLINK, 2014; RODRIGUES *et al.*, 2012).

The Cerrado is one of the main producers of surface water for Brazil's major hydrographic regions. For example, the Cerrado accounts for most of the surface water

production for the Paraná, Tocantins-Araguaia, São Francisco, Parnaíba, and Paraguai hydrographic regions (LIMA, 2011). Besides, the largest hydroelectric facilities in the country are placed along rivers in the Cerrado, playing a key role in Brazil's energy matrix (OLIVEIRA *et al.*, 2014).

2.2. Hydrological model

The model chosen for the case study is the GR5J model (LE MOINE, 2008), a five-parameter, daily lumped rainfall-runoff hydrological model (Figure 2). The model estimates the water balance of a basin using a soil moisture accounting reservoir and a conceptual water exchange function (*F*). The routing structure consists of two unit hydrographs and a non-linear store, providing two flow components. The water exchange function *F* allows simulating the exchange of water with deep aquifers or surrounding reservoirs (PUSHPALATHA *et al.*, 2011).

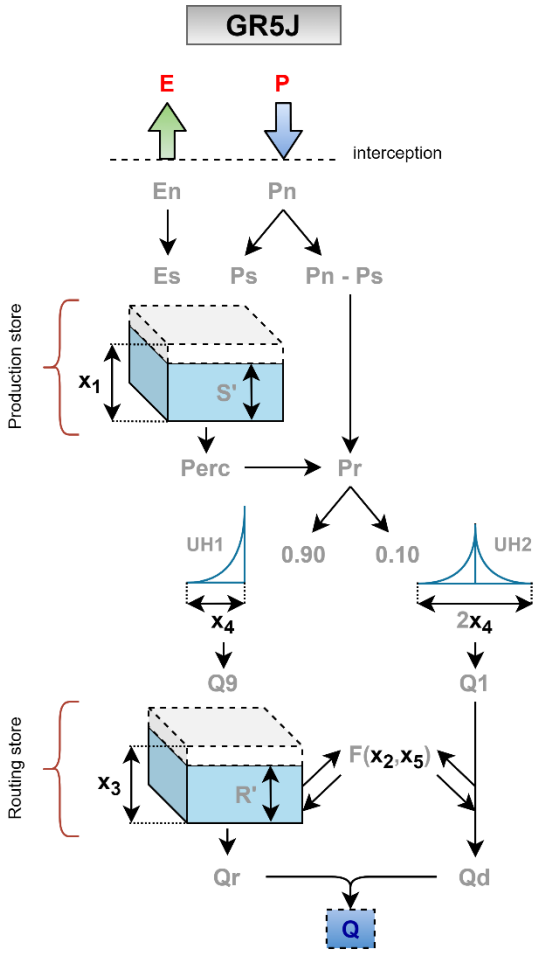

Figure 2. The GR5J model structure, adapted from Pushpalatha *et al.* (2011).

There are five free parameters in the GR5J model that can be optimized. x1 and x3 are the production and routing store capacity, x2 and x5 regulate the groundwater exchange function, and x4 is the time base of the unit hydrograph. The only physical inputs required

by the model are potential evaporation and precipitation. The model is coded using the R environment and programming language (R CORE TEAM, 2020) and the airGR package (CORON *et al.*, 2017, 2020).

2.3. Data sources

This section describes the data sources used for the development of the hydrological models in the Cerrado. The study period of 14.5 years, starting June 1, 2000, and ending December 31, 2014, was chosen based on the best available temporal coverage of all data sets. Streamflow only presents consisted data, i.e., data derived from calibrated rating-curves, up to 31 December 2014, while potential evaporation and rainfall data are available from 1 June 2000.

2.3.1. Observed streamflow and catchment selection

The initial database consisted of 819 stations within the Cerrado biome and a buffer area of 1.5° (~166 km) around it. Daily streamflow data for the gauges were downloaded from the HidroWeb portal (http://www.snirh.gov.br/hidroweb/ - ANA, 2020). HidroWeb is part of the National System of Water Resources Information (in Portuguese: Sistema Nacional de Informações sobre Recursos Hídricos – SNIRH). The SNIRH offers access to the database containing information collected by the National Hydrometeorological Network (in Portuguese: Rede Hidrometeorológica Nacional – RHN). Catchment boundaries were derived using the TauDEM toolset (TARBOTON, 2008) and the SRTM Void Filled 3 arc-seconds (~90 meters) spatial resolution Digitial Elevation Model (DEM) data (EARTH RESOURCES OBSERVATION AND SCIENCE CENTER, 2017).

Unsuitable stations/catchments were discarded using the following criteria. First, catchment boundaries that did not intersect the Cerrado biome were discarded. Second, stations or periods with potentially erroneous streamflow or routing effects due to large dams (identified through visual screening of the stations' streamflow time series) were discarded. Third, catchments with less than 5 years of observed streamflow data during the calibration and validation period (2003-2014, see section 3.4) were discarded. The final databased used in this study comprised 411 gauge stations/catchment boundaries (Figure 3).

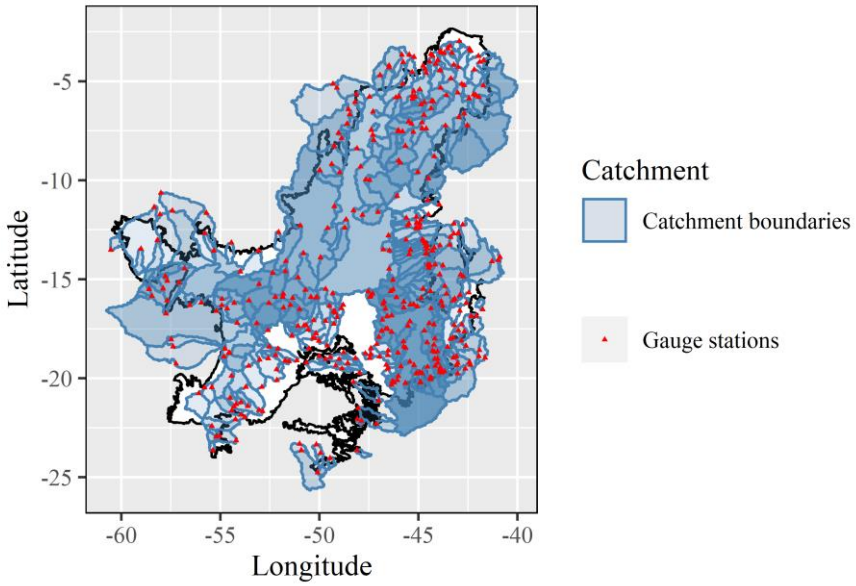

Figure 3. Gauge stations from the Brazilian National Hydrometeorological Network and associated catchment boundaries in the Cerrado region.

2.3.2. Meteorological forcing

To adequately represent precipitation and potential evaporation across the watersheds, gridded data products at daily resolution were used. Daily precipitation data from the Integrated Multi-Satellite Retrievals for the Global Precipitation Measurement (GPM) mission (IMERG) (HUFFMAN *et al.*, 2019b, 2019a) were used, while the ETo-Brazil product (ALTHOFF *et al.*, 2020a, 2020b) was used to represent the potential evaporation over the watersheds. Both IMERG and ETo-Brazil have a spatial resolution of 0.10° x 0.10° (~10 km x 10 km).

IMERG data has been widely tested and is considered one of the most reliable gridded precipitation products available for Brazil (GADELHA *et al.*, 2019). The ETo-Brazil, on the other hand, is a gridded database developed for the Penman-Monteith (ALLEN *et al.*, 1998) evapotranspiration using machine learning techniques, and matching the IMERG spatial and temporal resolution. In addition, both data sets cover areas within the Cerrado where low weather station density limits data availability.

2.4. Model calibration and optimization algorithm

The full record data set was split into three periods. The first 2.5 years were used for the hydrological model warm-up (2000/06/01 to 2002/12/31). The following 9 years (2003/01/01 to 2011/12/31) were used for the calibration of model parameters, and the latest 3 years (2012/01/01 to 2014/12/31) for model validation. Despite the model run beginning in a period of lower soil moisture for its initial conditions (June), 2.5 years is a long enough warm-up period (KIM; KWON; HAN, 2018). A period of 9 years is also considered enough

for the calibration of conceptual hydrological models (AYZEL; HEISTERMANN, 2021; BAI; LIU; XIE, 2021). For gauge stations with an incomplete set of observations, the first 75% of the available observations from 2003 to 2014 were used for the calibration of model parameters, and the remaining 25% for the model validation.

The optimization of the model's parameters was carried out using two approaches. The first is a single-objective optimization for the Kling-Gupta efficiency (*KGE*) (GUPTA *et al.*, 2009) and the second is a multi-objective optimization for the *KGE* and the relative Nash-Sutcliffe efficiency (*rNSE*) (KRAUSE; BOYLE; BÄSE, 2005; NASH; SUTCLIFFE, 1970). These criteria are calculated as follows:

$$rNSE = 1 - \frac{\sum_{i=1}^{n}\left|\frac{\hat{x}_i - x_i}{x_i}\right|^2}{\sum_{i=1}^{n}\left|\frac{x_i - \bar{x}}{\bar{x}}\right|^2} \quad , \quad -\infty < rNSE \leq 1 \tag{1}$$

$$KGE = 1 - ED \quad , \quad -\infty < KGE \leq 1 \tag{2}$$

with: $ED = \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2}$

where x_i and $\hat{x}_i$ are observed and predicted streamflow at time-step i, respectively, $\bar{x}$ is the mean observed streamflow, n is the number of observations, *ED* is the Euclidian distance, r is the linear correlation coefficient between $\hat{x}_i$ and x_i, α is the variability ratio or ratio between the standard deviation of predicted values and standard deviation of observed values ($\sigma_{\hat{x}}/\sigma_x$), and β is the ratio between the mean predicted and mean observed values ($\mu_{\hat{x}}/\mu_x$).

The *KGE* works as a multi-objective function that is based on the decomposition of errors in terms of the mean flow, flow variability, and flow dynamics (GUPTA *et al.*, 2009). The *rNSE* was used in the multi-objective optimization to evaluate if a robust trade-off that emphasizes low-flow conditions can be found. The *rNSE* criterium was chosen considering the importance of the simulations during the dry season for assessing water availability in the Cerrado. The "best" parameter set was chosen from a Pareto front by considering which set minimized the Euclidean distance to the ideal point (*KGE* = 1, *rNSE* = 1) (Eq. 3).

$$Best\ set = min\left[\sqrt{(KGE - 1)^2 + (rNSE - 1)^2}\right] \tag{3}$$

where the *Best set* is the parameter set that minimizes the distance to the ideal point (*KGE* = 1, *rNSE* = 1).

The model was optimized in R (R CORE TEAM, 2020) using the Particle Swarm Optimization (PSO) technique (CLERC, 2012; PARSOPOULOS; VRAHATIS, 2002; RAQUEL; NAVAL, 2005). The multi-objective PSO was implemented using the mopsocd package in R (NAVAL, 2013).

2.5. Regionalization approaches

Streamflow series were simulated at ungauged basins using regionalization of the parameters of the GR5J hydrological model calibrated at gauged basins. The regionalization was based on the best performing method among the two regionalization methods assessed. Both methods were based on donor catchments (i.e., gauged catchments for which the GR5J hydrological model was calibrated) selected by different definitions of homogeneous regions. In this study, homogeneous regions are defined either as (i) regions with spatially close catchments or (ii) regions with catchments having similar attributes in land use, soil type, climate, or catchment area (POOL; VIVIROLI; SEIBERT, 2019). The first regionalization method was based on five donor catchments from spatially close catchments (M1); the second method was also based on five donor catchments, but from catchments having similar attributes (M2). According to Qi *et al.* (2020), the optimal number of donor catchments lies within 4 and 6 gauged catchments, and the optimal threshold to qualify a gauged as an information donor is $KGE > 0.50$. However, since the difference between setting a threshold of $KGE > 0.50$ and $KGE > 0.70$ was minimal for the study by the authors, we adopted stations that scored over the *KGE* threshold of 0.70 during the calibration period.

Spatial proximity was measured by the Euclidean distance between the coordinates of the centroids of given catchments, while attribute similarity was measured by the Euclidean distance in attribute space. The attribute space consisted of fourteen selected catchment characteristics based on commonly used attributes in other regionalization studies (ADDOR *et al.*, 2018; BECK *et al.*, 2020; KRATZERT *et al.*, 2019c; POOL; VIVIROLI; SEIBERT, 2019; RAZAVI; COULIBALY, 2013b). The attributes include the log-transformed catchment area, mean annual precipitation, precipitation seasonality index, mean annual reference evapotranspiration, aridity index, percentage of soil clay content, percentage of soil sand content, soil organic carbon content, soil bulk density, percentage of forest area, percentage of crop area, percentage of grasslands and pasture area, mean elevation, and mean slope. All attributes were standardized before computing the attribute similarity. The catchment attributes description, range values, and sources are described in Table 1.

Table 1. Catchment attributes, sources, and range values.

Cerrado_topo	**Topography**					
Description	***Unit***	***Min.***	***Median***	***Max.***	***Data source***	***References***
Catchment area	km^2	51	4280	746000	Derived from	
Catchment mean elevation	m	64	676	1125	the SRTM at	EROS (2017)
Catchment mean slope	m/m	0.01	0.07	0.25	3 arc-seconds	
Cerrado_clim	**Climate indices - computed over the period 2000/06/01 to 2014/12/31**					

Description	*Unit*	*Min.*	*Median*	*Max.*	*Data source*	*References*
Mean annual precipitation - P	mm/year	726	1463	2012	IMERG	Huffman *et al.* (2019b)
Mean annual potential evapotranspiration (estimated using the Penman-Monteith evapotranspiration) - PET	mm/year	1021	1481	1930	ETo-Brazil	Althoff *et al.* (2020a, 2020b)
Seasonality index or timing of precipitation (estimated using sine curves to represent the annual evapotranspiration and rainfall cycles – positive [negative] value means precipitation peaks during summer [winter])	-	-1.12	0.43	1.21		Eq. 14 in Woods (2009)
Aridity index (PET/P)	-	0.60	1.01	2.63		
Cerrado_soil	**Soil characteristics – only covers the top 2.0 m**					
Description	*Unit*	*Min.*	*Median*	*Max.*	*Data source*	*References*
Clay	%	17.6	28.8	48.1	SoilGrids	Hengl *et al.* (2017)
Sand	%	32.6	55.2	70.6		
Soil organic carbon	g/kg	8.1	14.4	44.7		
Bulk density	kg/m^3	1.25	1.38	1.48		
Cerrado_vege	**Land cover characteristics – averaged from period 2000 to 2014**					
Description	*Unit*	*Min.*	*Median*	*Max.*	*Data source*	*References*
Forest fraction (forest formation, savanna formation, forest plantation)	%	2.2	40.3	99.6	MapBiomas v4.1	MapBiomas (2020)
Crop fraction (agriculture, annual crop, perennial crop, semi-perennial crop, mosaic of agriculture and pasture)	%	0.0	3.5	78.6		
Grassland fraction (grassland and pasture areas)	%	0.1	45.9	96.1		

For each regionalization method, each donor catchment provided a parameter set that was used to compute a hydrograph at the ungauged catchment. The 5 streamflow simulations were aggregated into an ensemble mean hydrograph (Eq. 8).

$$\bar{Q}_i = \frac{1}{N}\sum_{D=1}^{N} Q_{i,D} \quad (8)$$

where $\bar{Q}_i$ is the mean streamflow simulation at time-step i, $Q_{i,D}$ is the streamflow simulation for the parameter set D and at time-step i, N is the total number of parameter sets from donor catchments, D refers to each of the total N parameter sets.

The regionalization was evaluated for the validation period using a leave-one-out cross-validation, i.e., each gauged basin was treated as an ungauged catchment at a time and

its streamflow simulated using the information from donor catchments. The performance of the regionalization approaches was evaluated during the validation period using *KGE* and *rNSE* scores. The scores of the regionalization approaches were also compared to an "upper limit", which are the scores achieved by the local calibration for the validation period, and a "lower limit", which are the scores from benchmark series for the validation period. The benchmark series consisted of the mean monthly streamflow from data observed during the calibration period and are similar to making naïve predictions. For highly seasonal watersheds, as are tropical ones, the mean monthly streamflow from a known period is a reasonable baseline (lower limit) for the regionalization to show improvements on (ALTHOFF; RODRIGUES; BAZAME, 2021).

2.6. Runoff time series simulation at ungauged basins

The streamflow series were simulated at all (4531) level 5 ottobasins considering the regionalization approach that performed best. The ottoclassification is adopted in Brazil by the SNIRH for the hierarchical classification of hydrologic units consisting of basins and interbasins (FURNANS; OLIVERA, 2001). This classification system allows the identification of topologically related watersheds and can be used anywhere in the world. Although some ottobasins are only adjacent interbasins, simulating "streamflow" series is a good measure of their runoff contribution in the watersheds they belong to.

2.7. Data availability

A regional-sample data set was compiled for the 411 catchments in the Cerrado biome used in this study and made available at the following online repository: https://github.com/daniel-althoff/HydroCerrado/. The data set was named HydroCerrado and includes information on the gauge station geolocation, the catchment boundaries, daily time series of observed streamflow (section 2.3.1) and meteorological forcings (section 2.3.2), and catchment attributes (Table 1). The simulated runoff time series for all level 5 ottobasins in the Cerrado were also made available in this data set. Higher levels in ottocoding mean that basins/hydrologic units are finer representations of the land surface into smaller watersheds, i.e., sub-watersheds of lower level ottobasins (FURNANS; OLIVERA, 2001).

3. Results and discussion

3.1. Catchment attributes

Figure 4 presents an overview of the catchment attributes used in the study. Concerning topographic characteristics, watersheds in the southeastern regions of the Cerrado are in general smaller in area, higher in altitude, and present higher slopes. For

climate indices, annual rainfall is higher in western Cerrado and lower in eastern Cerrado. Potential evaporation is higher in the eastern and northern Cerrado, leading to higher aridity indexes in these regions. Most of Cerrado present similar cycles of rainfall and evapotranspiration, with an exception to northern Cerrado which has opposite cycles. For land cover characteristics, higher forest fractions are seen in northern Cerrado, while higher crop and grassland fractions concentrate in the central-southern Cerrado. For soil characteristics, the eastern Cerrado have watersheds with higher clay content, organic carbon concentration, and bulk density.

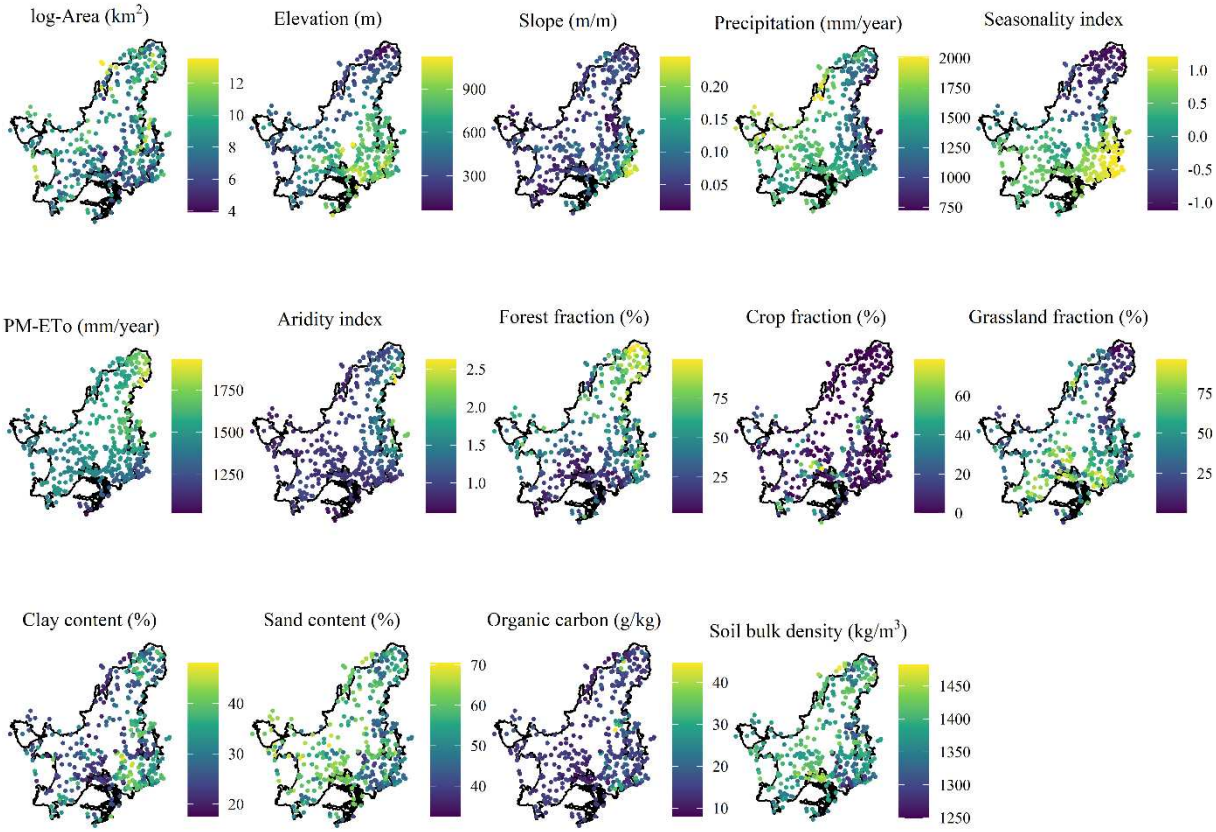

Figure 4. Overview of catchments centroid and corresponding attributes. Description and sources of attributes are shown in Table 1.

The catchments' attributes were considered static in time, despite some may considerably change over time. For example, the average land cover fraction of the entire period was used, although there has been an increase/decrease of more than 40% in land cover fractions for some catchments (Figure 5). Crops have replaced both grasslands and forests in the western region of the Cerrado, while primarily grasslands in the southern. Pastures replaced forests mostly in northern Cerrado.

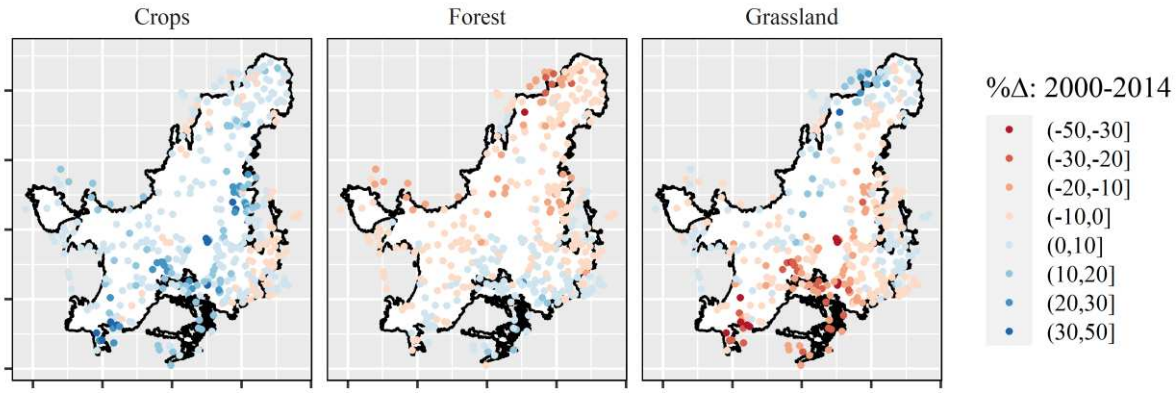

Figure 5. The absolute difference in land cover characteristics from 2000 to 2014 for all catchments. Description and sources of land cover characteristics are shown in Table 1.

As substantial as these changes are, they may be the reason for non-stationarity and a possible lack of power of the hydrological models in simulating streamflow during both the calibration and validation period (GUPTA *et al.*, 2009). For example, the regionalization study by Beck *et al.* (2016) discarded catchments with forest gain or loss > 20% based on the review of Bosch and Hewlett (1982). However, considering the Cerrado current scenario, we would end up with only a few catchments and poor spatial coverage by discarding catchments with significant land cover changes. In contrast, catchments with significant land cover changes can present a degradation in performance during the validation period. Nevertheless, since most of the Cerrado has faced severe land cover changes in recent years, it is better to assess the validation performance for such conditions than to deny their occurrence. Therefore, all catchments were considered regardless.

3.2. Regionalization performance

The performance of the regionalization approaches is shown in Figure 6 along with the upper and lower performance limits, i.e., local calibration and benchmark series. The calibration function using both *KGE* and *rNSE* as objective function led to robust runoff simulations, as seen by equivalent *KGE* values for the local calibration and regionalization by spatial and attributes proximity, but median *rNSE* increase of 0.09, 0.04, and 0.03, respectively, when compared to using only *KGE* as the objective function. This highlights the potential of multi-objective functions for the calibration of hydrological models when specific goals are well-established. In this case, the *rNSE* was chosen to improve simulations under low flow conditions and all further discussion considers only the outcome from using *KGE* & *rNSE* as the objective function.

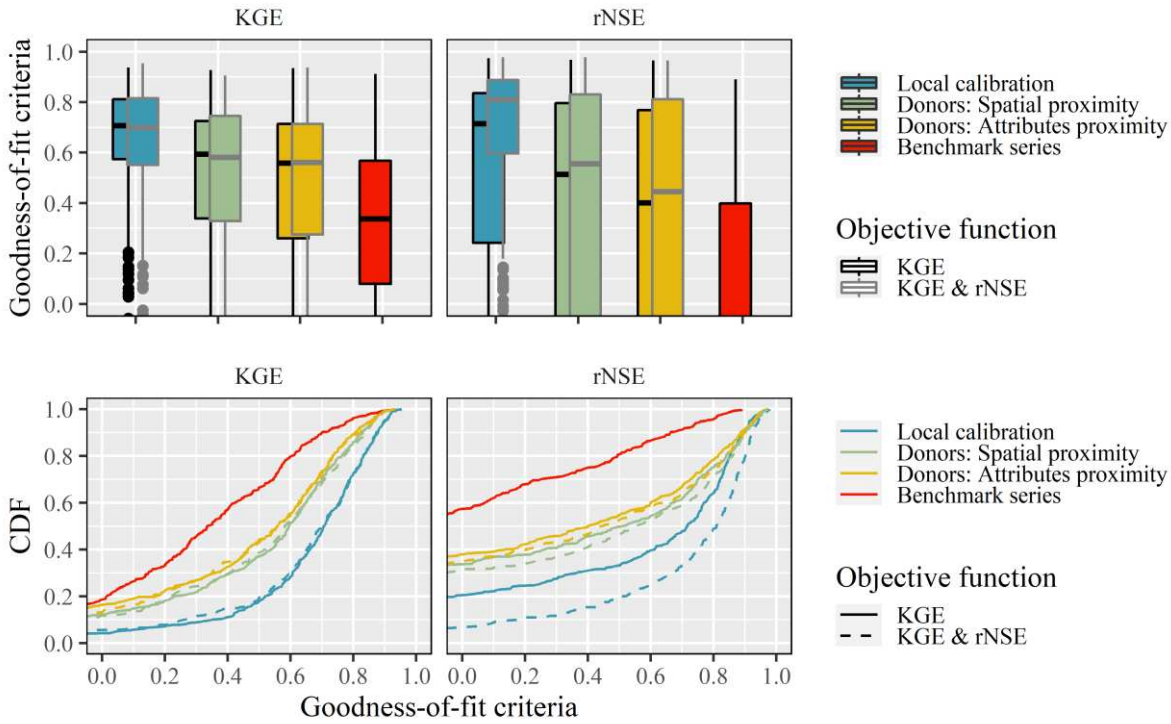

Figure 6. Boxplot of *KGE* and *rNSE* scores for the validation period obtained using local calibration, two regionalization approaches, and benchmark series for all catchments. The regionalization scores represent the average hydrograph obtained using parameter sets from donor catchments and the benchmark series scores represent the mean monthly streamflow from the calibration period. The scores are also grouped in terms of the objective functions used for optimization: single-objective function (*KGE*) and multi-objective function (*KGE & rNSE*).

Among the regionalization approaches, donating parameter sets by spatial proximity resulted in median *KGE* and *rNSE* of 0.58 and 0.56, respectively, which are 0.02 and 0.11 higher than median scores obtained when donating parameter sets by attributes proximity. The substantially higher *rNSE* obtained by spatial proximity regionalization is an important matter for the Cerrado region. This is due to the importance of adequately predicting water availability under low flow conditions for sustainable socio-economic development in many regions. The lower performance achieved by donors with attributes proximity may come from giving all catchment attributes the same weight when computing their physical similarity. Despite physically similar catchments showing close ranges for the attributes, it is unclear which attributes are more meaningful and should be given more weight. There is significant heterogeneity of the land surface conditions, soil profiles, and space-time variability of climatic forcing which is lumped for the catchments and makes it difficult to unveil such relationships (GUO *et al.*, 2021; SIVAPALAN, 2003).

As for the regionalization based on spatial proximity, other aspects also deserve attention, which is the number of catchments used as donors and their distance to target

catchments. In this study, we used the threshold of *KGE* > 0.70 and only 5 donor catchments. Despite following the recommendations by Qi *et al.* (2020), better results may be found using different numbers of donor catchments, either based on the maximum allowed distance or different *KGE* thresholds. The trade-off of using higher *KGE* thresholds is that a lesser number of donor catchments will be available or only in longer distances, resulting in lower performance for predictions in ungauged catchments or even parameter discontinuity for some regions. The *KGE* threshold of 0.70 seemed reasonable compared to the results obtained by Qi *et al.* (2020). This threshold resulted in the availability of 82% (n = 339) of the stations as donor catchments and the average distance between the donor and target catchment of 71 km (median = 60 km, max. = 359 km).

Only the regionalization approach based on spatial proximity was used for further analysis since it performed better than the regionalization based on attributes proximity. Qi *et al.* (2020) showed in a large-scale study that regionalization by spatial proximity is generally better than regionalization by physical similarity. However, this may not be the case for every region (ARSENAULT; BRISSETTE, 2016; GUO *et al.*, 2021). In comparison to the "upper limit", the local calibration (*KGE* = 0.70), the spatial proximity presented a decrease in median *KGE* score of 0.12 and approximately 13% of the catchments presented substantially low *KGE* scores (< 0) as opposed to 6% when local calibration is used. In comparison to the lower limit, the benchmark series (*KGE* = 0.34), donating parameters by spatial proximity showed an increase in median *KGE* score of 0.24.

Figures 7a and 7b present the scatterplots comparing *KGE* scores obtained for the catchments using parameter sets donated by spatial proximity to scores obtained using the local calibration and benchmark series. Although the local calibration can be considered an "upper limit", the regionalization based on spatial proximity showed robustness, with improvements in approximately 34% of the catchments. This is because neighbor catchments can be hydrologically similar and, therefore, calibrate equivalent parameter sets. Ensembles obtained with multiple "equivalent" parameter sets can function as multiple calibration runs and be advantageous for the runoff predictions.

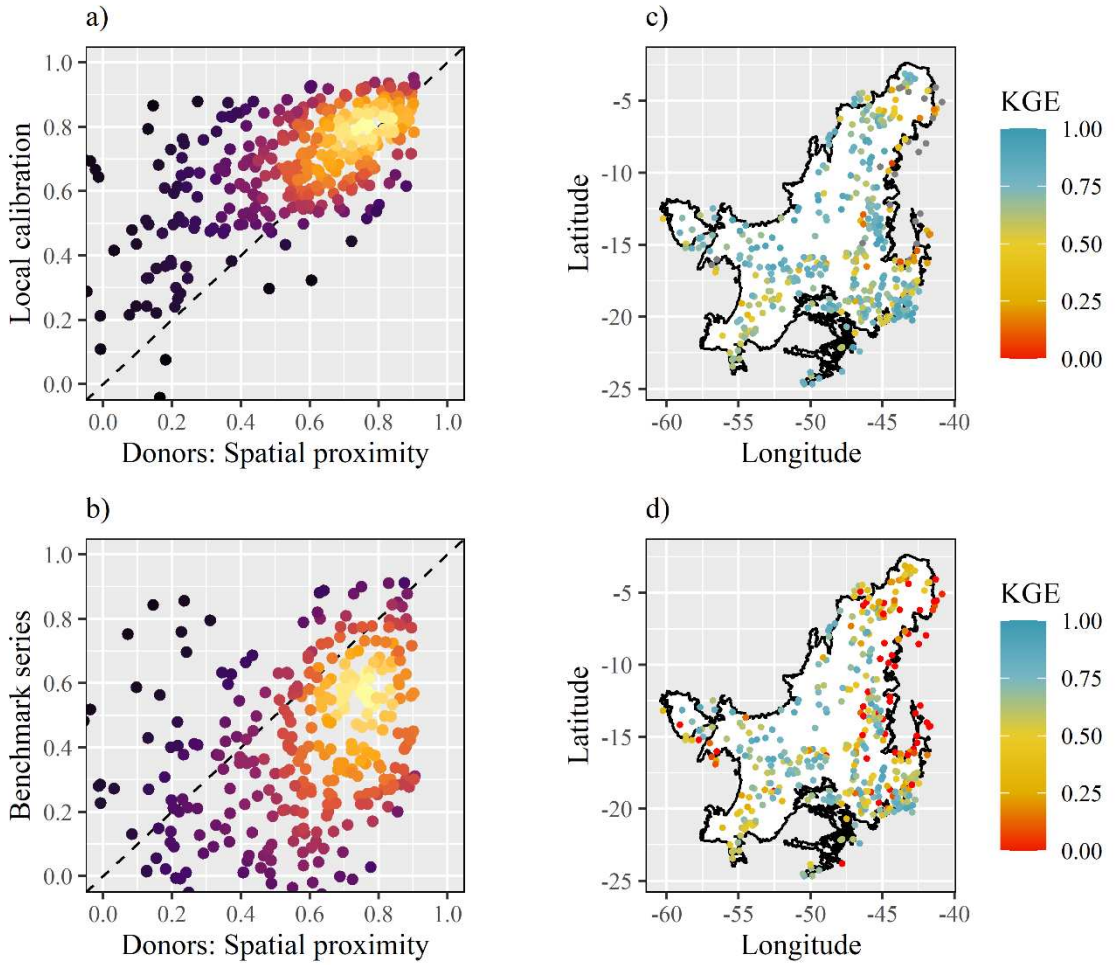

Figure 7. (a-b) *KGE* scores obtained using parameter sets donated by spatial proximity against *KGE* scores obtained using **(a)** local calibration and **(b)** benchmark series. **(c-d)** *KGE* scores obtained for all catchments using **(c)** local calibration and **(d)** parameter sets donated by spatial proximity close. Red dots are catchments with negative KGE scores.

Figures 7c e 7d show maps of the *KGE* scores obtained for the catchments using local calibration and parameter sets donated by spatial proximity, respectively. The performance at ungauged catchments deteriorated more in the eastern and northern regions of the Cerrado, which matches with regions showing higher aridity indexes (Figure 4). A review study by Guo *et al.* (2021) highlights that many authors have also found lower performance scores for regionalization methods when aridity increases. For example, the global regionalization studies by Beck *et al.* (2020) and Qi *et al.* (2020) show lower performance for validation catchments in regions with low humidity index (inverse of aridity index) and in arid regions, respectively.

3.3. Runoff time series in ottobasins

The parameter sets from donor catchments were used to estimate runoff time series for all 4531 level 5 ottobasins (median size = 185 km^2) for the period from 2003 to 2014 (12 years). It is important to highlight that the threshold of *KGE* > 0.70 resulted in distances

between donor catchments and ottobasins similar to observed during the cross-validation (mean = 97 km, median = 90 km, max = 410 km). This means that the threshold allowed several donor catchments with similar distances to replicate the experimental conditions of the cross-validation. For a brief overview of the data set generated for the ottobasins, two important hydrological signatures were computed, the long-term mean runoff, termed here mean runoff, ($\overline{Q}$) (Figure 8b) and minimum runoff with an exceedance probability of 95% (Q_{95}) (Figure 8d). $\overline{Q}$ provides us some information on the water balance and availability in the catchments or ottobasins, while Q_{95} is the reference used by the Brazilian National Water Agency for granting water withdrawal permits (maximum allowed: 70% of Q_{95}).

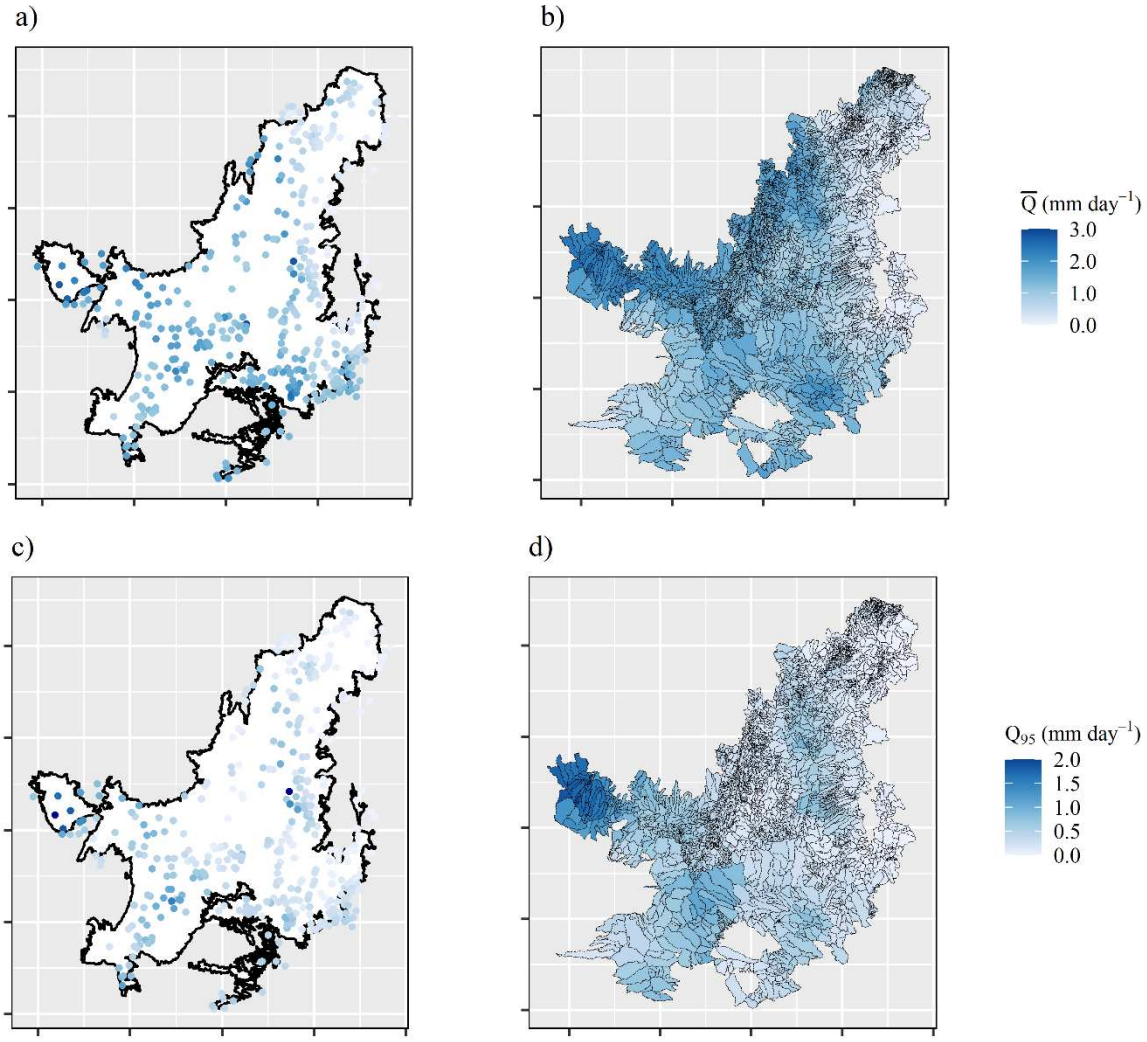

Figure 8. (a, b) Mean runoff ($\overline{Q}$) and **(c, d)** flow at exceedance probability of 95% (Q_{95}) for **(a, c)** streamflow series in gauged catchments and **(b, d)** simulated series for level 5 ottobasins.

There is a good spatial agreement between hydrological signatures computed for observed and simulated runoff series (Figure 8). Overall, there is remarkably higher water availability, both in $\overline{Q}$ and Q_{95}, at the transition between the Cerrado and the Amazon biome (western Cerrado), where the tropical climate registers higher rainfall rates (Figure 4). There

is also considerable water availability in southern Cerrado, which is closer to regions with a temperate climate. The eastern and northern Cerrado are marked by lower water availability due to their proximity to the transition to semiarid regions.

The correlation between the catchment attributes and hydrological signatures reinforces some already known relationships and unveils other aspects regarding the hydrological processes. For example, precipitation correlates to higher water availability (correlation to $\overline{Q}$: $\rho = 0.78$, and Q_{95}: $\rho = 0.42$), whereas higher aridity correlates with lower water availability ($\overline{Q}$: $\rho = -0.75$, and Q_{95}: $\rho = -0.41$). The mean slope shows a negative correlation with Q_{95} ($\rho = -0.20$), which is likely due to better drainage and/or increased direct runoff in the wet season. The soil characteristics presented a lower correlation with the hydrological signatures. In contrast, lower water availability ($\overline{Q}$) is seen in watersheds with higher forest fractions ($\overline{Q}$: $\rho = -0.50$). In comparison to grasslands and crops, forests help intercept rainfall and increase perspiration, which can reduce runoff. In relation to base flow, the effect of forests is context-dependent. Although they can improve infiltration, increasing recharge and therefore the base flow during the dry season, the increase in perspiration for water-limited regions may aggravate the water availability issue (LELE, 2009). In this study, higher forest fractions also negatively correlate to Q_{95} ($\rho = -0.32$). The overall negative correlation between forest fractions and these hydrological signatures can also be attributed to higher forest fractions being seen in the northern Cerrado, where the climate is more arid (Figure 4). It may also be the case that crop first expanded over forested areas in the regions with higher water availability, i.e., higher chances of success for the agribusiness activity. Thus, the effects of increasing or decreasing forest fractions would be better assessed along with the trends in base flow for these areas.

These results are important and practical, giving water managers more confidence in water resource planning and management. For instance, let us consider that we were able to compile only 411 observed streamflow series for the Cerrado (median catchment area = 4280 km^2). Thus, hydrological signatures are generally obtained for large catchments and water availability can only be assumed for smaller river segments under high uncertainty. Furthermore, many areas do not fall within larger catchments and are completely unmonitored (Figure 3). The simulated runoff series data set comprises 4531 ottobasins (median area = 185 km^2) of which roughly 50% represent independent river segments and not interbasins.

Despite the cross-validation being used to approximate the regionalization performance, the average area of ottobasins is generally lower than that of the gauged basins. Thus, there may exist biases in predicting hydrological responses for ottobasins which are

difficult to measure. In conclusion, despite imperfect, the application and assessment of regionalization techniques provide us with valuable information at an expected performance (cross-validation). Finally, the data set for the ottobasins is also provided with some uncertainty quantification, i.e., the prediction intervals consisting of quantiles (5% and 95%) derived from the hydrographs simulated using donated parameters.

4. Conclusions

Hydrological models calibrated considering both the Kling-Gupta efficiency (*KGE*) and relative Nash-Sutcliffe efficiency (*rNSE*) in a multi-objective function resulted in better performance for the regionalization as opposed to using the *KGE* in an objective function alone. The multi-objective function resulted in a median *rNSE* increase of 0.10 for the local calibration, and 0.03 and 0.04 for regionalization by attributes proximity and spatial proximity, respectively, while maintaining equivalent *KGE* scores.

The regionalization by spatial proximity outperformed that by attributes proximity (physical similarity), especially concerning *rNSE* scores. Spatial proximity led to *KGE* and *rNSE* of 0.58 and 0.56, while attributes proximity led to *KGE* and *rNSE* of 0.56 and 0.45, respectively. When compared to an upper-performance limit, which is the performance achieved with local calibration, the regionalization by spatial proximity showed a decrease in median *KGE* by 0.12, but showed robustness and was able to improve the performance on approximately 34% of the catchments. Concerning the benchmark series (mean monthly streamflow from calibration period), the regionalization by spatial proximity showed an increase in median *KGE* by 0.24.

The regionalization of parameter sets across all the Cerrado provides fundamental information for improving knowledge about water availability. The comparison between observed and simulated hydrologic signatures showed good spatial agreement. Finally, a regional-sample data set was made available with information compiled for the 411 catchments used in this study, and all runoff time series simulated for the level 5 ottobasins.

2.2. Artigo 2

Assessment of Water Availability Vulnerability in the Cerrado

Abstract

Many basins in the Cerrado biome already face increases in disputes over water caused by the agriculture expansions without long-term planning and monitoring of natural resources. The amount and timing of water availability are crucial for the socio-economic development of agricultural activities and increasing climate change resilience in the Cerrado. This study uses a data set developed for the Cerrado to characterize its climate and current status concerning water availability. Different climate indices and hydrological signatures were used to describe the energy and water budget, climate seasonality, flow magnitude, flow dynamics, as well as the average duration of low flow events. A risk analysis was carried out to classify the monthly risk severity and identify periods and regions under higher pressure for water availability in the Cerrado. Overall, the water availability is higher in the western and southern regions, and lower in the eastern and northern regions which transition to a semiarid climate. The water resources in the Cerrado are under the highest pressure in September and October. We also identified regions with potential for increasing sustainable water use. These regions show a low ratio between the flow that is equaled to or exceeded 95% of the time and the mean flow (Q_{95}/Q), which means that sustainable water use can increase by techniques related to flow regularization and rainwater harvesting.

Keywords: climate indices; hydrological signatures; sustainable water use; water scarcity.

1. Introduction

The importance of the Cerrado biome (Brazilian tropical woodland savanna) is recognized both for being a biodiversity hotspot, with the richest flora among all of the world's savannas (LAMBERS *et al.*, 2020), and for its notable presence in the national economy (LEMES; DE ANDRADE; LOYOLA, 2020). For many years has the Cerrado been a focal area for investments and expansion of agriculture (LEMES; DE ANDRADE; LOYOLA, 2020; SPERA, 2017). Cerrado agriculture was responsible for turning Brazil, from a net importer of grain in the 1980s to one of the world's major grain producers (HOSONO; ROCHA; HONGO, 2016). The biome itself is responsible for more than half of the country's soybean and beef production (IBGE, 2017; KLINK, 2014).

The Cerrado's soils present good physical structure but are acidic and infertile. The agricultural success of the biome comes mainly from improvements in management practices and high technological and fertilization inputs (HOSONO; ROCHA; HONGO, 2016; LAMBERS *et al.*, 2020). Another key factor that is responsible for the agriculture success in the Cerrado is its climate. The Cerrado climate is mainly classified as tropical savanna with dry winter (ALVARES *et al.*, 2013). This characterizes a warm season with plentiful rainfall and sunlight to permit abundant plant growth. It is common for some regions to benefit from a long wet season and adopt double-cropping to increase profitability (SPANGLER; LYNCH; SPERA, 2017).

Despite the Cerrado agriculture success being considered by some as miraculous (RADA, 2013), it is also controversial. The agriculture expansion brought landscape fragmentation, loss of biodiversity, soil erosion, and pollution of water resources(LAMBERS *et al.*, 2020; LEMES; DE ANDRADE; LOYOLA, 2020; LIMA; PERSSON, 2020; RADA, 2013; SOUZA *et al.*, 2020). Moreover, climate change projections have estimated warming trends, decreases in consecutive wet days, increases in the dry season length (CHOU *et al.*, 2014; PIRES *et al.*, 2016), and decrease in water availability in river streams (OLIVEIRA *et al.*, 2017). These projections jeopardize the success of agriculture, especially the rainfed. Different studies show that the conservancy of remnants of the Cerrado can be fundamental in mitigating climate change (COSTA; PIRES, 2010; PIRES *et al.*, 2016). Using a climate model and deforestation scenarios for the Cerrado, Costa and Pires (2010) show that rainfall decreases in most months and the dry season becomes longer. Alternatively, the intensification of agriculture practices over already degraded pastures may be a way to provide economic development while avoiding expansion over natural ecosystems (SPERA, 2017).

Possible beneficial effects have been reported for agricultural intensification in different parts of the world. Although not the objective of agriculture intensification and/or irrigation expansion, Mueller *et al.* (2016) argue that its side-effects include the increase in the potential evapotranspiration on hotter days, which contributes to precipitation recycling and decrease in temperature. Lathuillière, Coe, and Johnson (2016) also report the possible increase in precipitation recycling by irrigation expansion in already cultivated areas. Kueppers, Snyder, and Sloan (2007) discuss the irrigation cooling effects on a regional scale for California. Other strategies for agricultural intensification include rainwater harvesting, which can increase water availability during drought periods for irrigation purposes (PINHATI; RODRIGUES; AIRES DE SOUZA, 2020; RODRIGUES *et al.*, 2012). Fortunately, it has been reported for the Cerrado alone a potential to expand its irrigated areas in approximately 26.5 Mha (FEALQ, 2014), while the country's current irrigated area is still below 8 Mha.

In Brazil, water resource management and planning have usually developed in regions or hydrographic basins where conflicts arose from restrictions related to low water availability, e.g. the Brazilian semiarid, or excess in demand, e.g. large cities (ANA, 2014). Managers should be aware, however, that the poor water resource management and planning and the lack of proper information in data-scarce regions may do more harm than benefit, especially concerning irrigation strategies (HE *et al.*, 2017). Thus, improved knowledge about water availability is critical not only to support better public policies (e.g. irrigation, energy production, and water supply) but also to contribute to more effective water management instruments (e.g. water permits). Despite the considerable water availability in the Cerrado for irrigation expansion (FEALQ, 2014), there are regions where the expansion occurred with poor planning and hydrological information aggravating conflicts over water use (POUSA *et al.*, 2019). Particularly in the Cerrado biome, many basins already face increased disputes over water use because of the accelerated development of agriculture disassociated of long-term planning and monitoring of natural resources uses (RODRIGUES, 2018).

The amount and timing of water availability are crucial for the socio-economic development of agricultural activities in the Cerrado, as well as for increasing its resilience to climate change. Furthermore, the dry season and drought periods, along with projections for increased hydrological droughts put the development in the Cerrado in check (RODRIGUES *et al.*, 2020). Thus, the objective of this study was to characterize the water availability for the Cerrado in terms of surface water sources, define the periods of higher pressure on the resource, and identify regions where sustainable water use can increase by

local storage of excess runoff and regularization. Such information is important for improving integrated water resource management and to help guide which regions should have higher priorities or would better benefit from investment allocation by the government.

2. Material and Methods

This section first presents the study area and the streamflow and meteorological forcings used in this study. These data sets are used to give context to the water availability situation of the Cerrado by deriving the climate indices and hydrological signatures described in the following subsections. In this study, water availability is discussed in terms of water available from surface sources and mainly characterized by the hydrological signatures related to flow distribution. The last two subsections describe the risk classification proposed herein and how to identify regions with potential for improving sustainable water use.

2.1. Study area

The Cerrado occupies approximately 23% of the Brazilian territory (Figure 1a) and is responsible for yielding water for many of the country's main hydrographic regions intersecting it (LIMA, 2011). The Cerrado hydrography is shown in Figure 1b at the spatial resolution of 1:250.000 (ANA, 2020b), showing its distribution of river flow to many bordering regions. Its climate is mainly characterized as tropical savanna with dry winter (ALVARES *et al.*, 2013), i.e., two well-defined seasons with mean monthly above 18°C for all months and precipitation of the driest month, in winter, below 60 mm. The climate transitions to a hot semi-arid climate bordering the Caatinga, tropical monsoon climate bordering the Amazon and Pantanal, and temperate climate bordering the Atlantic Forest. These transition areas are important in understanding the climate variation across the Cerrado, especially concerning irrigated agriculture because of the precipitation distribution over the year.

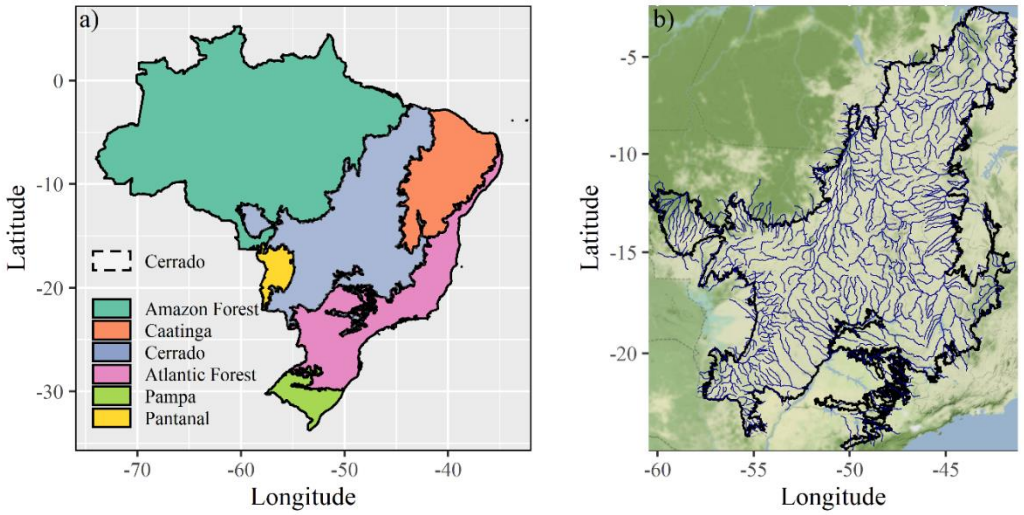

Figure 1. (a) The Brazilian biomes (data source: IBGE, 2019) and **(b)** the Cerrado hydrography (data source: ANA, 2020).

2.2. Streamflow and meteorological forcings

The HydroCerrado is a data set for catchments in the Cerrado biome (ALTHOFF *et al.,* 2021, Artigo 1). Aside from the information compiled for gauged catchments, the HydroCerrado also includes the boundaries (Figure 2a), meteorological forcings (Figure 2b), and streamflow series (Figure 2b) estimated for all (4531) level 5 ottobasins. Ottobasins consist of hydrologic units (basins and interbasins) classified according to the Pfafstetter hierarchical classification (FURNANS; OLIVERA, 2001). The ottobasins streamflow series were simulated based on the regionalization of the parameters calibrated for the GR4J rainfall-runoff hydrological model (PERRIN; MICHEL; ANDRÉASSIAN, 2003). Although there can be a lot of uncertainty in streamflow series obtained via regionalization, this is the best data set available for the hydrologic units in this resolution. Furthermore, some hydrological signatures derived from the streamflow series obtained for the ottobasins showed good spatial agreement to observed streamflow series in the study by Althoff *et al.* (2021, Artigo 1).

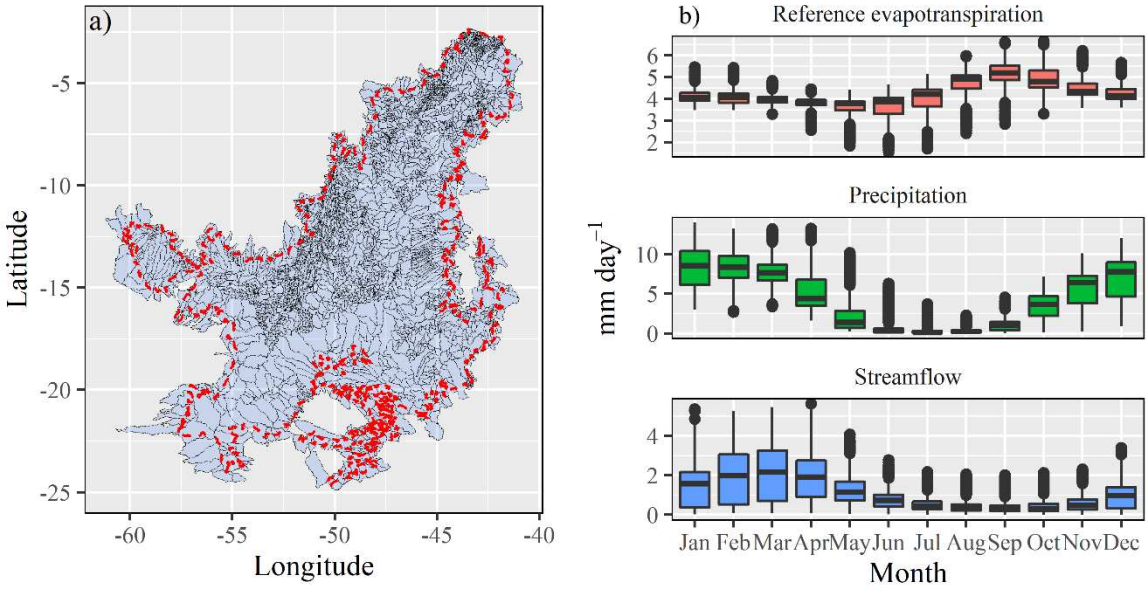

Figure 2. (a) Boundaries of all level 5 ottobasins in the Cerrado and **(b)** their respective mean monthly reference evapotranspiration, precipitation, and streamflow.

The data set covers the period from 2003 to 2014. The streamflow series are based on the regionalization of hydrological models and gridded data that are available from 2000-06-1 to the near-present. However, the observed streamflow data used to calibrate the models used for regionalization only had a quality check assured by the Brazilian National Water Agency (ANA) up until 2014. The period from 2000-06-01 to 2002-12-31 was then used as a warm-up period for the model. The meteorological forcings, i.e., daily time series for precipitation (P) and Penman-Monteith reference evapotranspiration (ETo), were derived from the Integrated Multi-satellite Retrievals for Global Precipitation Measurement (IMERG) (HUFFMAN *et al.*, 2019b, 2019a) and the gridded reference evapotranspiration for Brazil based on machine learning (ETo-Brazil) (ALTHOFF *et al.*, 2020a, 2020b), respectively. It is important to highlight that because the climate indices and hydrological signatures (see the following section) are derived from relatively short time series (12 years), it may be the case where some regions faced drier or rainier years than the historical average. This could result in indices that do not accurately reflect a given region's climate and water availability.

2.3. Assessment of water availability

2.3.1. Climate indices

The climate is, in a large part, responsible for the processes near the surface concerning precipitation and evaporation. Whether water remains on land or returns to the atmosphere depends on the water and energy limitations (KNOBEN; WOODS; FREER, 2018). Thus, four climate indices were used to characterize hydrologically relevant aspects

of the ottobasins: (i) the average annual moisture index (I_m) (KNOBEN; WOODS; FREER, 2018), (ii) the moisture index seasonality ($I_{m,r}$) (KNOBEN; WOODS; FREER, 2018), (iii) the precipitation seasonality index (SI), (iv) and precipitation timing index (TI).

The average annual moisture index I_m (Eq. 2) ranges from -1 to 1, where -1 indicates the aridest conditions, i.e., water-limited, and 1 indicates the most humid conditions, i.e., energy-limited. The moisture index seasonality $I_{m,r}$ (Eq. 3) ranges from 0 to 2, where 0 indicates no seasonal changes in the water and energy budgets, whereas 2 indicates a climate shift between fully arid (I_m = -1) to fully humid (I_m = 1) within a year.

$$MI(t) = \begin{cases} 1 - \frac{E_m(t)}{P_m(t)} & , P_m(t) > E_m(t) \\ 0 & , P_m(t) = E_m(t) \\ \frac{P_m(t)}{E_m(t)} - 1 & , P_m(t) < E_m(t) \end{cases} \quad (1)$$

$$I_m = \frac{1}{12} \sum_{t=1}^{t=12} MI(t) \quad (2)$$

$$I_{m,r} = \max[MI(1, 2, \dots 12)] - \min [MI(1, 2, \dots 12] \quad (3)$$

where *MI(t)* is the monthly moisture index (dimensionless), $E_m(t)$ is the mean monthly potential evaporation, $P_m(t)$ is the mean monthly precipitation, *t* denotes the months from January to December (1 to 12), I_m denotes the average annual moisture index, and $I_{m,r}$ denotes the moisture index seasonality. Although I_m and $I_{m,r}$ rely on potential evaporation data, the reference evapotranspiration is close related to it and, since it is readily available in the HydroCerrado data set, it was used instead.

The precipitation seasonality and timing indices SI and TI were derived from sine curves (Eqs. 4-5) used to represent the annual temperature and precipitation cycles (WOODS, 2009). SI (Eq. 6) is a positive integer and indicates the amplitude of the precipitation sine curve concerning the mean precipitation; a SI equal 0 indicates no seasonality, whereas higher values indicate stronger seasonality. TI (Eq. 7) ranges from -1 to 1 where -1 indicates a complete mismatch between precipitation and temperature peaks (~6 months), i.e., precipitation peaks in the cold season, while 1 indicates that they peak together, i.e., precipitation peaks in the warm season. The variables summarize the precipitation amplitude and its phase in relation to temperature into a single variable. The reference evapotranspiration was also used instead of temperature in these equations. This adaptation to the precipitation timing index should be favorable because the precipitation peak is then truly compared with the peak in atmospheric demands.

$$P(t) = \bar{P}[1 + \delta_P \sin(2\pi(t - s_P)/\tau)] \quad (4)$$

$$E(t) = \bar{E}[1 + \delta_E \sin(2\pi(t - s_E)/\tau)] \quad (5)$$

$$SI = \delta_P \quad (6)$$

$$TI = sign(\delta_P)sign(\delta_E)\cos(2\pi(s_P - s_E)/\tau) \quad (7)$$

where *P(t)* and *E(t)* are the daily time series for precipitation and reference evapotranspiration (mm day^{-1}), respectively, *t* is time (in units of days), $\bar{P}$ and $\bar{E}$ are the time-averaged precipitation and reference evapotranspiration, respectively, δ_P and δ_E are the seasonal amplitude in the respective sin curves, τ is the duration of the seasonal cycle (365 days), and s_P and s_E are the precipitation and reference evapotranspiration phase shift (days), respectively.

2.3.2. Hydrological signatures

Hydrological signatures have commonly been used to classify basins and gain insights into their hydrological processes (ADDOR *et al.*, 2018; MCMILLAN, 2020). This study uses 4 signatures (Table 1) selected mainly to characterize the water availability concerning flow distribution, event frequency, and duration, and flow dynamics. Because basins with different areas can present streamflow at very different magnitudes, all flow-related variables are expressed in flow units relative to the respective basin area, i.e., mm day^{-1} (L m^{-2} day^{-1}) instead of, for example, m^3 day^{-1}. Using mm day^{-1} as a flow unit makes the water availability comparable among basins with different areas. It is also used as the standard flow unit for these hydrological signatures in many studies (ADDOR *et al.*, 2018; ALMAGRO *et al.*, 2020; KNOBEN; WOODS; FREER, 2018).

Table 1. Overview of the hydrological signatures selected for this study.

Signature	Abb.	Unit	Description
Flow distribution			
Mean flow	$\overline{Q}$	mm day^{-1}	Mean flow for the period
Flow percentiles	Q_{95}	mm day^{-1}	Flow exceedance percentiles from FDC
Event frequency and duration			
Low flow duration	Q_{LD}	days	The average duration of consecutive daily *low flow events (WESTERBERG; MCMILLAN, 2015)
Flow dynamics			
Baseflow index	BFI	-	Contribution of baseflow to total streamflow (GUSTARD; BULLOCK; DIXON, 1992)

In Westerbeg and McMillan (2015), low event flows are days when the flow is lower than 0.2$\overline{Q}$. Here, low-flow events were considered days when the flow is lower Q_{80}, i.e., 80th percentile of daily flow.

The hydrological signatures related to flow distribution were chosen to measure average water availability and availability under low-flow conditions. Additionally, the Q_{95} is used as a reference by ANA to grant water withdrawal permits (maximum allowed: 70%

of Q_{95} from which existing water permits are discounted), although ANA uses the term Q_{95} to refer to streamflow in volumetric flow unit, i.e., in $m^3\ s^{-1}$ instead mm day^{-1}. The flow dynamics signature, BFI [0, 1], measures the proportion of the streamflow that derives from subsurface sources. Higher BFI indicates sustained streamflow during drought periods (GUSTARD; BULLOCK; DIXON, 1992). The signatures related to event frequency and duration were also selected to provide insights on low-flow conditions. These signatures were calculated per year and the average taken for the yearly signatures.

2.4. Risk classification

Considering the importance of the Q_{95} for granting water withdrawal permits, a risk classification analysis was proposed herein to categorize the severity of pressure on water availability for each month. The main contribution of this risk classification is assessing streamflow series regarding its distribution, i.e., probability that the streamflow is less than or equal to a flow percentile. First, daily flows were categorized according to pressure severity on water availability in river streams (Table 2). The risk classification was then averaged for each month. This analysis was important to detect the regions and periods with the highest pressure on the surface water resource. Although some definitions of risk and water security can often be described as the potential to imposing harmful states for human beings or the environment (HALL; BORGOMEO, 2013), this study considered the risk as to the susceptibility to low flow conditions. This is because information concerning water demand is scarce and the hydrological signature (Q_{95}) used as a reference for granting water is also based on flow distribution.

Table 2. Classes of risk severity according to water availability.

Risk classification	Water availability
1 - Very low	$Q > Q_{80}$
2 - Low	$Q_{90} < Q \leq Q_{80}$
3 - Moderate	$Q_{95} < Q \leq Q_{90}$
4 - High	$Q \leq Q_{95}$

2.5. Potential for increasing the sustainable water use

The ratio between Q_{95} and $\overline{Q}$ ($Q_{95}/\overline{Q}$) is proposed here to be used as a proxy for the potential in increasing sustainable water use. $\overline{Q}$ provides information on the average water availability in ottobasins, whereas Q_{95} is the reference signature for granting water withdrawal permits and the flow likely seen during drought periods. Thus, the $Q_{95}/\overline{Q}$ gives us information on the regularization capacity of the ottobasins and may be seen as the relative water availability. Lower ratios mean that there is greater potential for storing water locally,

e.g. using on-farm dams or large dams to accumulate excess runoff, and making water available in periods of drought, i.e., increasing effective water availability.

3. Results and discussion

3.1. Hydroclimatic characterization

To better showcase how the climate indices were obtained and characterize the different regions across the Cerrado, Figure 3 shows the time series of the meteorological forcings and plots that summarize the climate indices for an arbitrary ottobasin (ID = 74929, longitude = -45.07, latitude = -9.09). The sine curves are approximations to the meteorological forcings seasonality and show the mismatch between precipitation and reference evapotranspiration peaks. The timing index (TI = -0.544) indicates a mismatch of approximately 125 days between the precipitation and reference evapotranspiration peaks. The seasonality index (SI = 1.178) indicates a strong seasonality, with an amplitude higher than the mean precipitation (3.14 mm day^{-1}). The negative moisture index (I_m = -0.316) indicates a condition of aridity where monthly potential evapotranspiration is generally higher than precipitation. The moisture index seasonality above 1 ($I_{m,r}$ = 1.412) indicates that monthly climate conditions range from energy-limited to water-limited over the year.

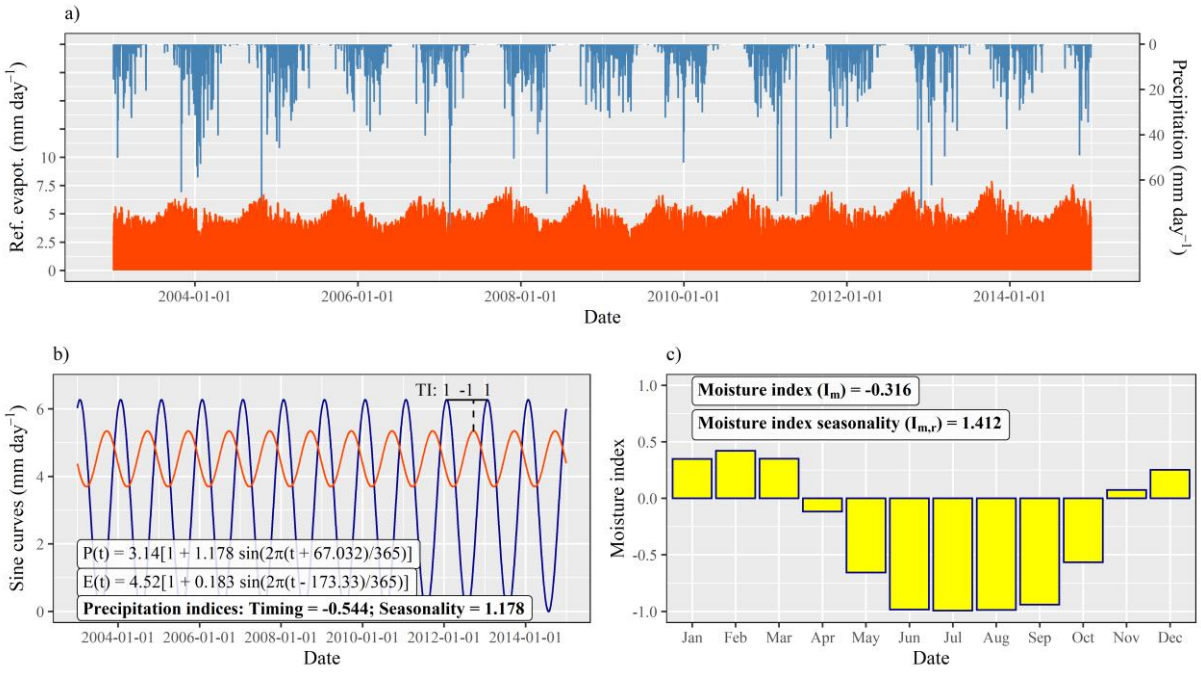

Figure 3. Time series of **(a)** precipitation and reference evapotranspiration, **(b)** their respective sine curves, and **(c)** monthly moisture index for an arbitrary ottobasin (ID = 74929).

The context of the climate indices for all of the Cerrado's level 5 ottobasins is shown in Figure 4. The Cerrado ottobasins are predominantly water-limited (negative I_m) and their seasonality is remarkably high, as seen by $I_{m,r}$ values higher than 1.4 for most of the Cerrado.

This is an outcome of its high energy input over the year and two well-defined seasons. The climate indices show evidence, however, of the transitions to different climates. For example, the eastern Cerrado presents the aridest conditions (lower I_m) and low seasonality (low $I_{m,r}$), characterizing the transition to semiarid climate. The southern Cerrado presents ottobasins with lower water limitation or even some energy limitation and, with lower moisture index seasonality, characterize the transition to humid subtropical zones. The western Cerrado borders the Amazon and characterizes the transition to a tropical monsoon climate.

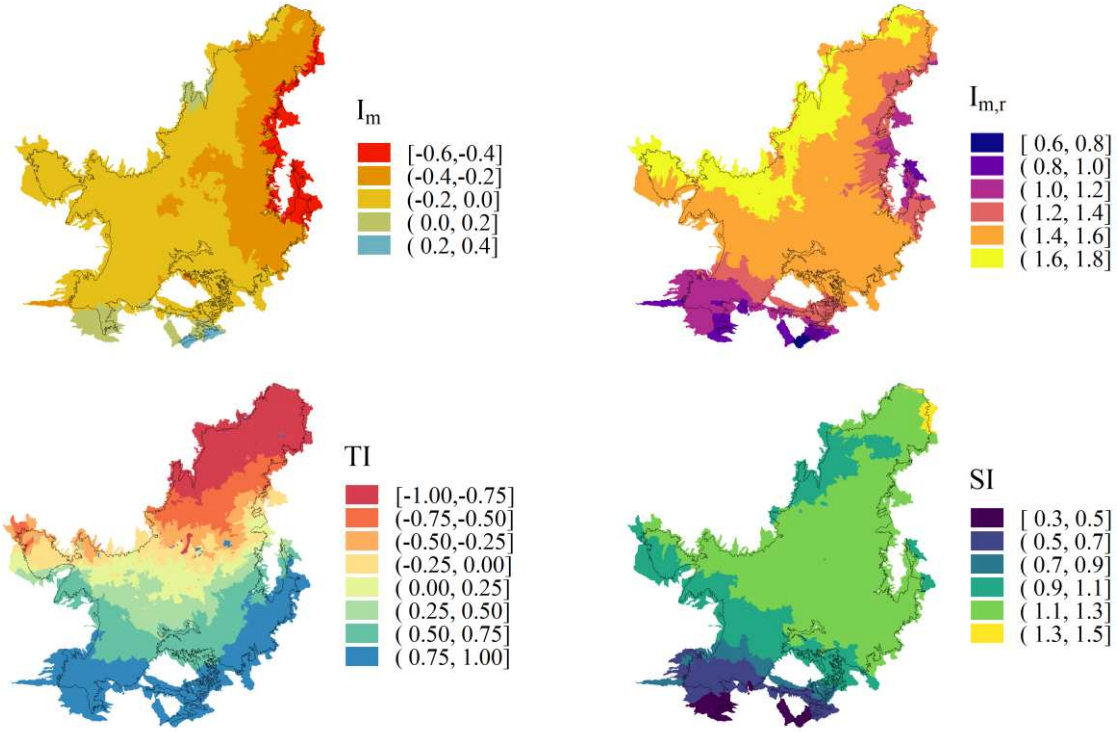

Figure 4. Climate indices for the level 5 ottobasins. I_m = moisture index, $I_{m,r}$ = moisture index seasonality, SI = precipitation seasonality index, and TI = precipitation timing index.

TI shows precipitation to peak along with the atmospheric demand for water in the south/southeastern regions (positive values), while the precipitation peaks in opposite periods than does evapotranspiration in the northern (negative values). The central Cerrado shows a smaller mismatch between precipitation and evapotranspiration peaks. SI shows a strong seasonality for precipitation (SI > 1.0) for most of the Cerrado, whereas the southern Cerrado shows values closer to 0 in the transition to humid subtropical regions. The lower SI values in the southern region mean that rainfall is better distributed over the year in these ottobasins.

Figure 5 summarizes the hydrological signatures for the same arbitrary ottobasin showcased in Figure 3. A significant portion of streamflow comes from shallow or subsurface flow (BFI = 0.872), meaning that most of the precipitation infiltrates and does

not contribute to surface runoff. The mean flow ($\overline{Q}$ = 0.313 mm day^{-1}) is approximately 80% higher than the 95th percentile of daily flow (Q_{95} = 0.173 mm day^{-1}) which is used as a reference for granting water withdrawal permits. This means that there are opportunities for increasing sustainable water use and increasing availability in the drought periods through storage and rainwater harvesting techniques (further discussed in section 3.3). The flow considered as a daily low flow event (Q_{80} = 0.202 mm day^{-1}) is approximately 17% higher than the Q_{95}. Despite not registering low flow events every year, e.g., from 2004 to 2007, its average duration is 57.7 days and has ranged up to 138 and 141 days in 2010 and 2014, respectively.

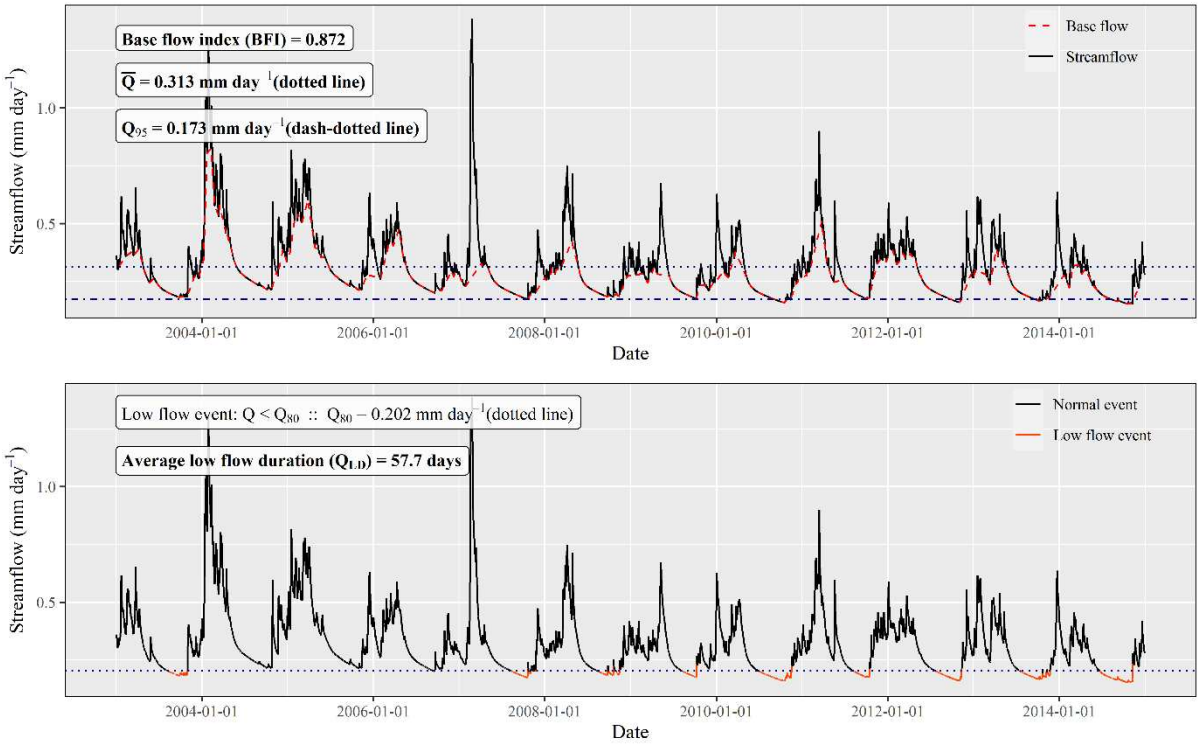

Figure 5. Time series of streamflow and respective hydrological signatures for an arbitrary ottobasin (ID = 74929). $\overline{Q}$ = mean flow, Q_{95} = 95th percentile of daily flow, BFI = baseflow index, and Q_{LD} = low flow event ($Q < Q_{80}$) duration.

The hydrological signatures (Figure 6) show a remarkably higher water availability, in terms of $\overline{Q}$, at the transition between the Cerrado and the Amazon biome (western Cerrado); however, it does not necessarily mean higher Q_{95}. The ottobasins in this region are characterized by their low water-limitation and by their higher annual precipitation. There is also considerably higher water availability in southern Cerrado, especially with respect to Q_{95}. Due to their proximity to the transition to semiarid regions, the eastern and northern Cerrado are marked by lower overall water availability.

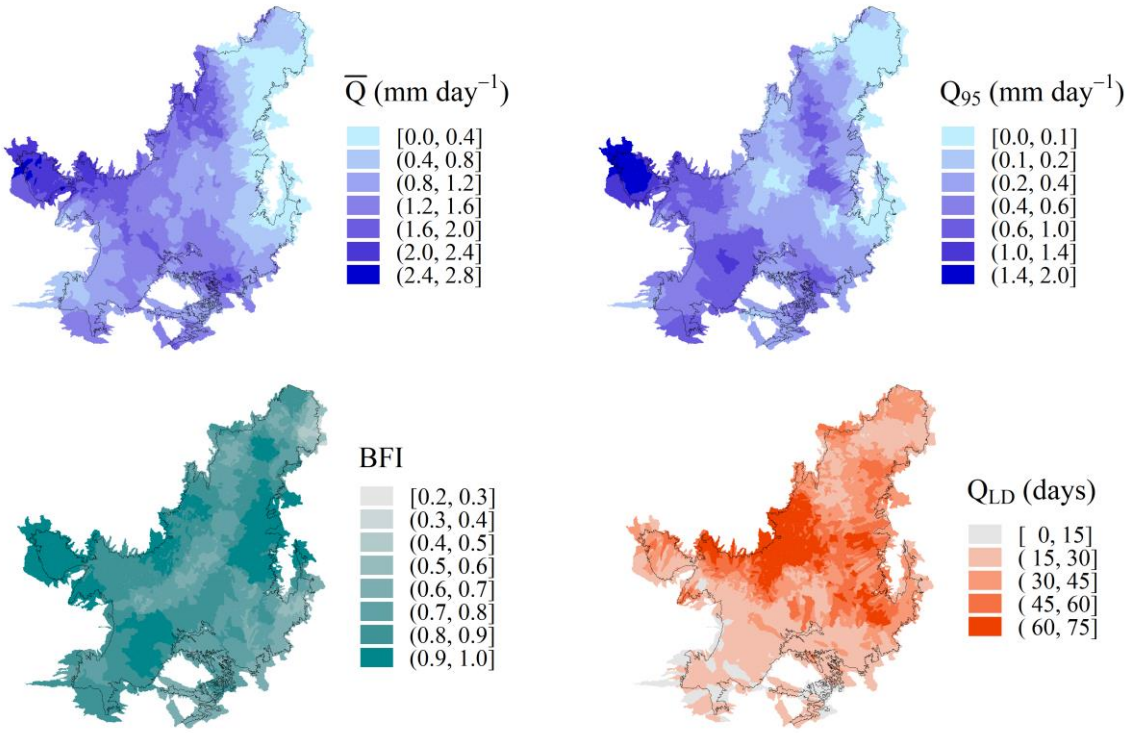

Figure 6. Hydrological signatures for the level 5 ottobasins. $\overline{Q}$ = mean flow, Q_{95} = 95th percentile of daily flow, BFI = baseflow index, and Q_{LD} = low flow event ($Q < Q_{80}$) duration.

Representing the flow dynamics, BFI registered higher values for regions presenting higher Q_{95}. This is because higher BFI means that most of the precipitation infiltrates and runoff is composed mainly of subsurface runoff and baseflow, i.e., flow is better sustained during drought periods. As for event frequency, to consider events where the daily flow was lower than Q_{80} would lead to an annual frequency of approximately 20% of daily flow observations, i.e., 73 days. However, it is important to assess the duration of these events and whether they concentrate in the same period. This information is important to highlight regions where drier years can apply further pressure on water availability and result in restrictions over long periods. For example, there are several ottobasins in the middle portion of the Cerrado, especially in the western region, with considerable duration for low flow events. In these regions, the periods when daily flows are below Q_{80} last for nearly or over two months. This can be attributed to the high seasonality in the region's climate ($I_{m,r,}$ and SI in Figure 4). In contrast, despite the frequency of ~73 days of low flow events per year, most of the northern and southern regions show that such events last between 15 to 30 days on average. In the southern region, this can be attributed to better rainfall distribution over the year, which can result in closer values for the lower tail of the daily flow cumulative distribution function (CDF), e.g., close values for the 80th and 95th percentile of daily flows. In contrast, the northern region is under more arid conditions, which can also result in similar values for the lower tail of the CDF. In this case, however, values closer to 0. When these

values are closer, it is easy for small perturbations in streamflow to interrupt the dry spell, i.e., the permanence of low flow conditions.

3.2. Risk assessment

The risk assessment is shown in Figure 7 for the same arbitrary ottobasin (ID = 74929), where daily flows are classified according to their risk of imposing pressure on water availability (Table 2). The percentage of occurrence plot is based on the risk classifications for flow events in a given month for the entire time series, e.g., the percentage of time flow was below Q_{95} for all the Septembers from 2003 to 2014. For this particular ottobasin, the occurrence of events characterizing risks higher than "very low" increases from June until September, which registered a high occurrence of "low" (29.4%), "moderate" (25.0%), and "high" (20.6%) risk events. Only on rare occasions did June register events of low risk, i.e., characterizing a low pressure on water availability.

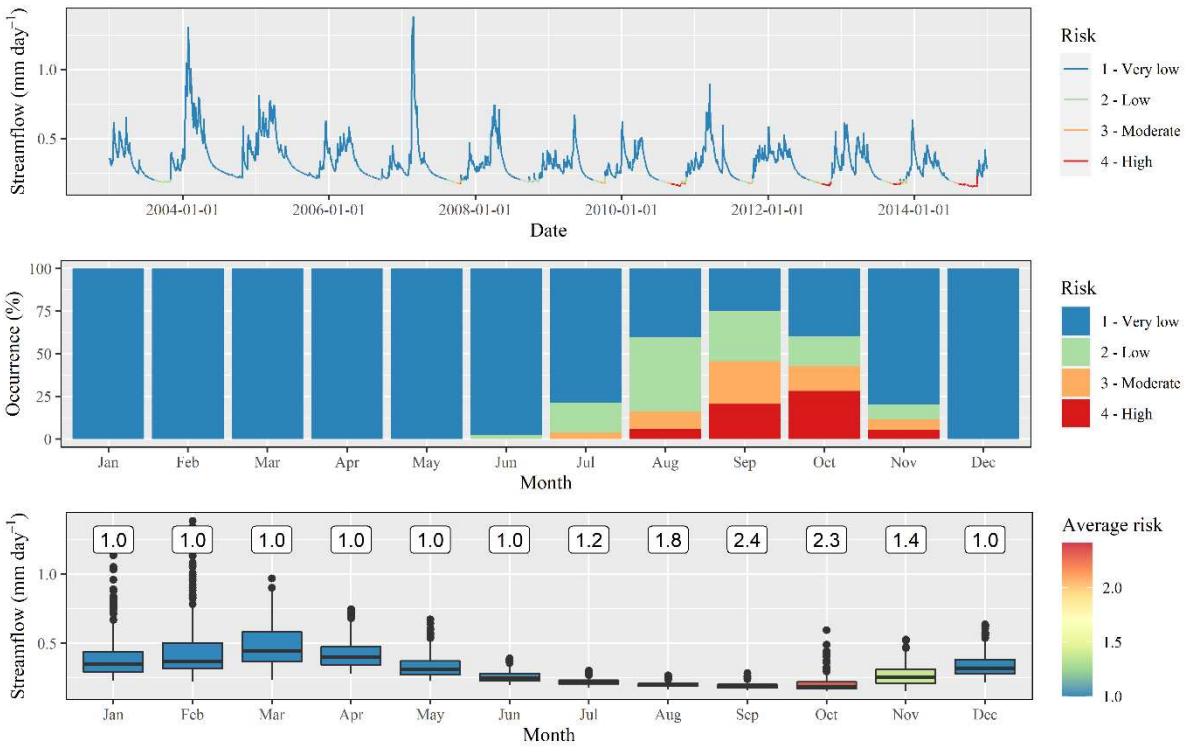

Figure 7. Risk assessment of pressure severity on water availability for an arbitrary ottobasin (ID = 74929).

The final assessment is based on averaging the risks for each month over the entire study period (2003 – 2014), which is labeled for each month above the boxplot and is also represented in the boxplot color. The average risk of pressure on water availability in September can be considered from low to moderate (avg. risk = 2.4). Despite showing a lower average risk (2.3), October presented the highest occurrence of high risk (28.2%) of pressure on water availability. The occurrence of "high" risk in October can be attributed to uncertainty in rainfall occurrence in the transition between the dry and wet seasons. When

the dry season lasts until late in October, the flow in river streams continues to decrease, and the pressure on the water resource increases.

The risk assessment is summarized for the entire Cerrado in Figure 8 by reporting the month classified with the highest risk, i.e., highest average risk on water availability, and the risk associated with it. As part of the southern hemisphere, the Cerrado has higher energy input from August to December, whereas the wet season begins in October and lasts until March or early April. This justifies September and October being the months of highest risk for most of the Cerrado. On the other hand, the wet season mainly concentrates from February to April in the northern Cerrado, mismatching the period of high energy input (negative TI in Figure 4). Thus, the period from November to January characterizes as the period of highest risk for the northern region. Overall, the period of highest risk generally characterizes the beginning of the wet season. This means that agriculture activities are put in jeopardy not only by the uncertain precipitation distribution in time and space but even for irrigators by the lower availability in river streams.

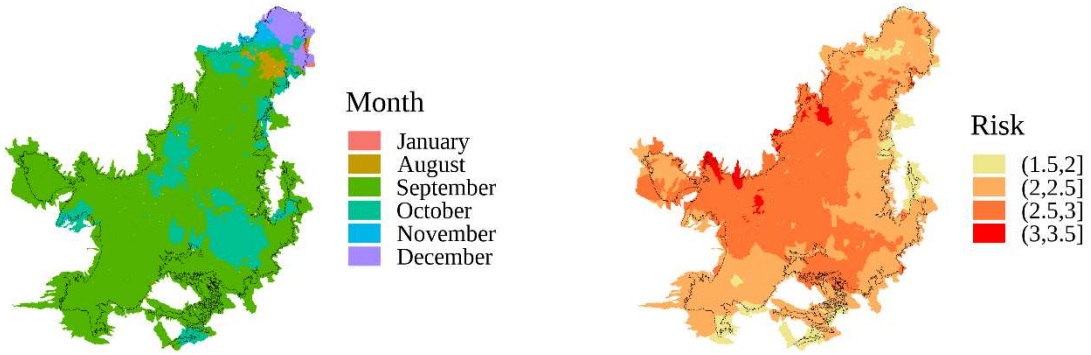

Figure 8. For each ottobasin, the month presenting the highest average risk of pressure on water availability and the average risk associated with it.

The risk is reportedly higher in the central region (from 2.5 to 3.5), matching regions with the highest moisture index seasonality (Figure 4). Some fragments show average risk to be above 3, i.e., from moderate to high risk, whereas most of Cerrado presents the average risk in these months to be from low to moderate (2.5 to 3.0) or at least low in the optimistic cases (2.0 to 2.5). Only a few regions reported risk to be between very low and low (1.5 to 2.0). The severity of the pressure on water availability can be explained mainly by the strong seasonality in the moisture index (Figure 4). The high moisture index seasonality means that one of the seasons has a much higher energy input than water, resulting in long drought periods and allowing for a continuous decrease in streamflow. Some of these regions also match the region where low flow events last longer (Figure 6).

3.3. Can sustainable water use increase?

It is evident that some regions show higher average water availability than others (mean flow in Figure 6). Yet, due to their seasonality, the hydrological signature used as a reference for granting water withdrawal permits (Q_{95}) can be much lower than the mean water availability ($\overline{Q}$). Thus, the $Q_{95}/\overline{Q}$ ratio (Figure 9) can be used to assess the time distribution of flows and the potential for increasing local storage (i.e. the lower the ratio, the poorer the flow distribution in time, which indicates a higher potential to use local storage as an adaptation measure to increase water availability). Lower ratios mean higher chances of success for water resource management strategies as small in situ storages. Such techniques have been widely documented to increase regional water sustainability, increase resilience to climate change, and decrease conflicts over water use (BOELEE *et al.*, 2013; IGLESIAS; GARROTE, 2015; MALVEIRA; ARAÚJO; GÜNTNER, 2012; RODRIGUES *et al.*, 2012). In summary, lower $Q_{95}/\overline{Q}$ ratios also mean that local storage has a higher potential to support sustainable irrigation expansion, i.e., expansion of irrigation activities without depleting freshwater stocks or loss of environmental flow (ROSA *et al.*, 2020).

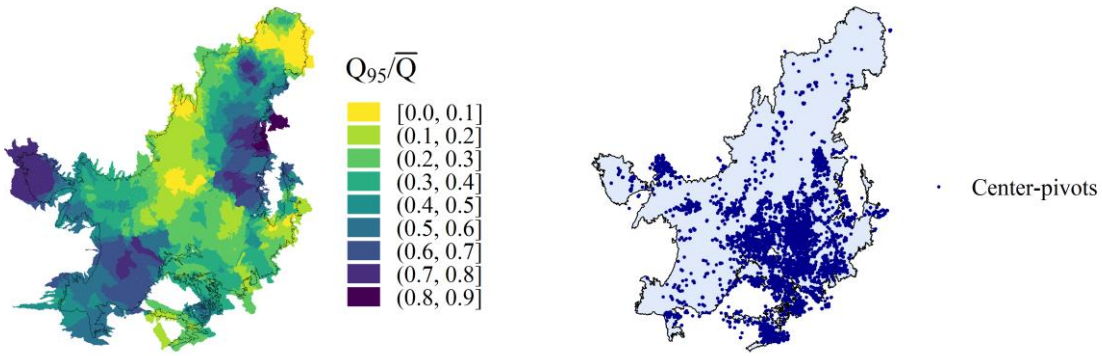

Figure 9. The ratio between hydrological signatures Q_{95} and $\overline{Q}$ for each ottobasin and the center-pivots (data source: ANA, 2019) spatial distribution in the Cerrado. $\overline{Q}$ = mean flow, and $Q_{95} = 95^{th}$ percentile of daily flow. The center-pivots are used here to represent irrigation water demands.

There is an important dynamic, in terms of energy- or water-limited watersheds, between moisture index, its seasonality, the timing of precipitation index, and the $Q_{95}/\overline{Q}$ (Figure 4 and 9). In general, both very water-limited ($I_m < -0.4$) and energy-limited (positive I_m) watersheds tend to have lower climate seasonality, i.e., lower $I_{m,r}$ values and, therefore, higher $Q_{95}/\overline{Q}$. For example, soil moisture increases in the eastern Cerrado (very water-limited) correlate better to/result in higher evaporation rates instead of runoff. This means that precipitation in eastern Cerrado generates smaller responses in runoff (high BFI, Figure 6). In contrast, the western and southwestern Cerrado, marked by lower energy limitation, has better precipitation distribution over the year (TI closer to 0), resulting in higher

infiltration, soil moisture, and runoff year-round (high BFI). This results in runoff better correlating to soil moisture than to precipitation. Such correlations between soil moisture, precipitation, evaporation, and runoff are well documented in the studies by Ghajarnia et al. (2020a) and Ghajarnia et al. (2020b). Thus, $Q_{95}/\overline{Q}$ ratios were generally positively correlated to BFI and negatively correlated to climate seasonality ($I_{m,r}$). The eastern Cerrado, despite showing high $I_{m,r}$, showed higher $Q_{95}/\overline{Q}$ ratios. This can be attributed to high BFI and, therefore, lower direct attribution of precipitation on runoff. In between, the central Cerrado, which presents lower degrees of limitations concerning either water or energy, shows higher $I_{m,r}$, and lower $Q_{95}/\overline{Q}$ ratios. As for the northern Cerrado, it presents very water-limited ottobasins, but the wet period opposes the period of higher evaporative demands (negative timing index, Figure 4). This increases the opportunity for water infiltrating the soil to reach river streams during the wet season and results in increased mean flow. However, since it characterizes as very water-limited, it presents much lower availability during the dry season (low Q_{95}) and, therefore, low $Q_{95}/\overline{Q}$ ratios.

As opposed to small farm reservoirs, the construction of large dams can regularize flow for periods longer than years. However, both small and large dams can significantly increase water losses due to evaporation (ALTHOFF; RODRIGUES; SILVA, 2020; PEREIRA *et al.*, 2009). More technified alternatives could help prevent such losses and at the same time produce energy. For example, Cavusoglu et al. (2017) reported water evaporation as a reliable source of renewable energy that can help reduce surface water evaporation. Solar panels have also been used to cover open-water reservoirs and irrigation canals, reporting decreased evaporative losses and increased efficiency for the photovoltaic systems (YOUSSEF; KHODZINSKAYA, 2019). The northern Cerrado, which is under a more arid climate, should require a stronger effort to improve the sustainable use of surface water. Farm-level reservoirs should preferably be deeper than wider (ALTHOFF; RODRIGUES; SILVA, 2020). Alternatively, underground reservoirs have been built to restrict subsurface runoff and extend the period when soil moisture is higher.

Regions with higher relative water availability ($Q_{95}/\overline{Q}$) should be assessed carefully. For example, the western of Bahia state (eastern Cerrado) presents high $Q_{95}/\overline{Q}$ but low water availability, requiring due attention. The low water availability ($\overline{Q}$ and Q_{95}) and high demand, which is represented here by the center-pivots, resulted in conflicts over water use which have been widely documented (POUSA *et al.*, 2019). Alternatively, farmers began exploiting other options than surface water. The increasing agribusiness activity in the region

led, however, to intensified groundwater withdrawal and unsustainable exploitation as documented by severe drawdowns in groundwater levels (MARQUES *et al.*, 2020).

Overall, the central Cerrado shows more opportunities to improve sustainable surface water exploitation through the local storage of water, especially in its central area (low $Q_{95}/\overline{Q}$). The western and southern areas, on the other hand, already present higher water availability and a higher baseflow index for some regions. Figure 6 also presents the center-pivots irrigation systems in the Cerrado as of 2017 (ALTHOFF; RODRIGUES, 2019; ANA, 2019). The high concentration of center-pivots in the central-southern region can be partially attributed to favorable conditions during cropping periods, i.e., positive timing of precipitation index. Higher precipitation in periods when the climate is favorable for crop development means higher productivity and profitability. The profitability often translates into technification, e.g., use of irrigation systems. Althoff and Rodrigues (2019) also highlight that there has been a trend for smaller center-pivots (< 50 ha) in the last decade. This likely means that smallholders are also becoming more technified and benefiting from irrigation systems to guarantee productivity/profitability when precipitation is uncertain. Moreover, there is still a large area where irrigation can expand in the Cerrado with lower risk due to the existing potential to increase water availability through local small scale storage, i.e., regions with low relative water availability (Figure 9) and high mean flow (Figure 6). In these regions, the improvement of the water resource management can support agricultural development when crop water requirements are higher than the natural supply (green water scarcity) (ROSA *et al.*, 2020). In regions already facing conflicts (blue water scarcity), the best alternative can be the adoption of deficit irrigation techniques in years of lower availability (GEERTS; RAES, 2009).

Moreover, we would like to highlight one important limitation of this study. The study was developed considering the regionalization of the parameters from a hydrological model calibrated using a relatively short and recent streamflow series. This means that the hydrological signatures were not obtained from natural streamflow series but series considering the already existent water withdrawals upstream. It may be the case for some regions where an amount of water larger than the Q_{95} presented in this study can be withdrawn without drying out the river. Thus, it is important to also consider the ongoing water grants for each ottobasin to make more informed decisions. Besides, if anomalies were registered during the short period used in this study, it can also imply lower reliability for the derived indices. Lastly, the risk assessment could be improved in future studies by considering current water grants and demands.

4. Conclusions

The high seasonality of the moisture index agrees with former climate classifications which define Cerrado's climate as a tropical savannah climate, i.e., two well-defined seasons. The transitions between the Cerrado and the other biomes showed to be determining factors in its climate characteristics, e.g., lower moisture index in the northeastern Cerrado, bordering the Caatinga, or higher moisture index in the southern Cerrado, bordering the Atlantic Rainforest. The water availability (mean flow) is generally higher in the western and southern regions. The higher baseflow indices highlight, for some regions, the rainfall low contribution to runoff, whereas regions with a lower baseflow index tend to be more responsive to rainfall.

The hydrological signature related to low flow event duration emphasizes the central to eastern regions as more prone to long-lasting dry periods. In contrast, low flow events tend to have shorter durations in the southern Cerrado. Overall, the water resource in the Cerrado is under the highest pressure from September to October, whereas the pressure shifts to November, December, and January in the northern Cerrado. The average risk is also higher in the same regions that showed high moisture index seasonality. These regions show a mean flow much higher than the minimum flow used as a reference for granting water and, given the high climate seasonality in the Cerrado, these regions can benefit from, for example, local storage of excess runoff and from water resources management strategies, such as shared water resources management.

2.3. Artigo 3

Towards Robust Hydrological Modeling to Assess River Discharge in Watersheds Under Severe Environmental Change

Abstract

The impacts of land cover change on hydrological processes have long been studied. Such impacts have traditionally been assessed with a priori knowledge, e.g., using methods based on the curve number, or by calibrating hydrological models over different time periods. However, how hydrological processes respond to such changes is extremely context dependent. Thus, there is an opportunity for the development of hydrological models that are able to learn from large hydrological data sets under the context of severe environmental changes. In this study, a single regional hydrological model is developed based on long short-term memory (LSTM) neural networks using different input configurations. One model considers only meteorological forcings as inputs (I1), another model considers meteorological forcings and static catchment attributes (I2), and a third model also considers meteorological forcings and catchment attributes but where the land cover characteristics are dynamic (I3). The models are trained using information from 411 catchments in the Cerrado biome. The Cerrado biome is a good candidate for this study because it has faced severe land cover changes in recent decades. The data set includes, for each catchment, the daily streamflow observations (target), daily precipitation and reference evapotranspiration (meteorological forcings), and 21 catchment attributes. Considering catchment attributes increases the performance of the LSTM model (I2 and I3 median *KGE*: 0.69) when compared to using only meteorological forcings (I1 median *KGE*: 0.40). Additionally, considering the land use cover characteristics as dynamic improves the predictions under low-flow conditions (I3 median *rNSE*: 0.62) when compared to the model considering such characteristics as static (I2 median *rNSE*: 0.53). This study also uses the integrated gradients technique to explore the contribution of the catchment characteristics to streamflow and the number of time steps of influence for the network in different regions.

Keywords: Cerrado; data-driven; explainable artificial intelligence; regional hydrological model.

1. Introduction

A long-standing challenge in hydrological modeling is extrapolating hydrological information from gauged/monitored areas to areas lacking information (KRATZERT *et al.*, 2019c). Inaccurate knowledge of water availability can affect local water resource planning and management and severely harm social well-being and economic development. Because river discharge is the major water resource for agricultural, industrial, and urban water use (GUO *et al.*, 2021), adequately extrapolating such information is crucial. Thus, ungauged regions often depend on regionalization approaches to transfer information from hydrological model-calibrated gauged catchments (GUO *et al.*, 2021; RAZAVI; COULIBALY, 2013a).

The most common techniques used to transfer information are approaches based on the calibration of hydrological models on individual gauged catchments and (i) donating parameters to ungauged catchments by physical proximity and/or similarity (ARSENAULT; BRISSETTE, 2016; BECK *et al.*, 2016; POOL; VIVIROLI; SEIBERT, 2019; QI *et al.*, 2020) or (ii) building regressions to derive parameters at ungauged catchments based on catchment attributes (ARSENAULT; BRISSETTE, 2016; BECK *et al.*, 2016, 2020). However, regression approaches are usually limited by the equifinality issue of parameter sets, whereas parameter donation can be limited by the gauge station network density, regional heterogeneity with respect to topography, climate, and geology, and low robustness of parameters to environmental and climate changes.

When working at larger scales, i.e., regional or continental scales, models can be calibrated for an entire region or subset of regions (KRATZERT *et al.*, 2018; PAGLIERO *et al.*, 2019). Macroscale models are usually calibrated to regions clustered by physical similarities and spatial proximity so that the model is aware of characteristics such as vegetation, topography, and pedology. Another alternative, as shown by Beck *et al.* (2016), is to produce parameter maps by regionalizing parameter sets of donor catchments. This approach is also based on catchment attributes in the hope that the model adequately learns different hydrological behaviors.

Data-driven methods have recently gained much popularity by succeeding over process-based models in mapping streamflows from meteorological forcing data (BAI; LIU; XIE, 2021; FAN *et al.*, 2020; KRATZERT *et al.*, 2018, 2019c). The main advantage of these methods is that they are free from predetermined conceptual structures and can learn how to represent hydrological processes themselves. Furthermore, a recent study by Kratzert *et al.* (2019b) showed that ancillary data from catchment attributes were able to improve streamflow estimates of a single long short-term memory neural network (LSTM)

(HOCHREITER; SCHMIDHUBER, 1997b) hydrological model trained on data from 531 catchments in the United States of America. In fact, the model proposed by the authors outperformed many popular conceptual hydrological models calibrated catchment-by-catchment or regionally.

As ancillary data, Kratzert *et al.* (2019b) used 27 catchment attributes as static input features. They recognized that treating catchment attributes as static is a strong assumption, especially over long time periods. Moreover, several studies have reported on the impacts of hydrological process dynamics inherent to land cover changes (ABDULKAREEM *et al.*, 2019; SRIWONGSITANON; TAESOMBAT, 2011; ZHOU *et al.*, 2015). Traditionally, the impacts of anthropogenic land cover changes have been simulated with methodologies based on the curve number (ABDULKAREEM *et al.*, 2019; GASHAW *et al.*, 2018; OLIVEIRA *et al.*, 2016; SILVA *et al.*, 2018; WANG *et al.*, 2018) or calibrations over different periods (LIU *et al.*, 2009). However, the use of the curve number can be limited by the many uncertainties related to land cover, e.g., degraded pastures and grasslands, density of forests, cultivated crop varieties, and soil and water conservation strategies. Thus, there is an opportunity for the development of a data-driven hydrological model that can benefit and learn from inputting land cover as a dynamic feature.

Given the importance of land cover changes on hydrological responses, the objective of this study is to develop a hydrological model that considers land cover characteristics as dynamic and assess whether including such information can improve the predictions made by the model. The main contribution of this study is considering the continuous changes in land use in a deep learning model and using explainable artificial intelligence techniques not only to investigate what the model learns from the regional-sample data set but to also compare it to previous knowledge concerning the catchment attributes effects on hydrological processes in the biome.

2. Material and Methods

The study area in this paper is the Brazilian Cerrado biome. Its watersheds are characterized by a tropical climate and strong seasonality. The Cerrado biome is a good candidate for this study because significant land cover changes have been reported for it in recent decades (ALTHOFF *et al.*, 2020c; MAPBIOMAS, 2020; ALTHOFF *et al.*, 2021, **Artigo 1**). It is considered the most important biome with respect to agriculture and has many areas classified as high potential for expanding irrigation activities (ALTHOFF; RODRIGUES, 2019; FEALQ, 2014). However, many regions of the Cerrado biome have

also reported conflicts over water use (POUSA *et al.*, 2019). Thus, it is important that hydrological models developed for the region are able to leverage land cover changes.

The following subsections describe the data-driven hydrological model, the input configurations tested, the streamflow and features data sets, the criteria used to assess performance, the benchmarking strategies, and the techniques used for model interpretation.

2.1. Hydrological model

The hydrological model used for this study is a type of recurrent neural network (RNN), specifically, the long short-term memory (LSTM) neural network (HOCHREITER; SCHMIDHUBER, 1997b). In particular, the LSTM configuration does not suffer from exploding/vanishing gradients and is suited for learning long-term dependencies (HOCHREITER; SCHMIDHUBER, 1997a). Recurrent layers consist of memory cells that are able to store information over sequential time steps similar to state vectors (KRATZERT *et al.*, 2019c). The LSTM is, therefore, able to model watershed processes, such as soil moisture and/or snow accumulation, comparably to process-based models (KRATZERT *et al.*, 2019a). The following subsections describe the operations behind the LSTM network (**Section 2.1.1**) and the model and training setups chosen for the study (**Section 2.1.2**).

2.1.1. Long short-term neural network

LSTM layers consist of a number of consecutive memory cells that are responsible for transferring/updating two states over consecutive time steps: a cell state and a hidden state. The cell state characterizes the memory of the system, and the hidden state characterizes what the layer will eventually output at the last time step of the sequence. In a memory cell, the hidden state of the previous time step (output of the previous memory cell) and the input vector at the current time step pass through three gate units that control which information is added or removed from the cell and hidden states. These gate units are the input gate, the forget gate, and the output gate. The forget gate (f_t, Eq. 1) deletes information from the cell state, while the input gate (i_t, Eq. 2) and cell update (g_t, Eq. 3) add new information to it (Eq. 4). The input gate controls which components of the cell state are updated, and the cell updates what information is added. Last, the output gate (o_t, Eq. 5) controls which information is updated to the hidden state (h_t, Eq. 6). Given an input sequence $x = [x_1, \ldots, x_T]$ with T time steps, where x_t is a vector of input features at time step t ($1 \leq t \leq T$), the forward pass through the LSTM layer can be described by the following equations:

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \quad (1)$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (2)$$

$$g_t = \tanh\left(W_g x_t + U_g h_{t-1} + b_g\right) \quad (3)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (4)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (5)$$

$$h_t = o_t \odot \tanh(c_t) \quad (6)$$

where f_t, i_t, and o_t are the forget gate, input gate, and output gate, respectively, and g_t is the cell update at time step t ($1 \leq t \leq T$). h_{t-1} and c_{t-1} are the hidden state and cell state at the previous time step ($t - 1$), respectively. W and U are weight matrices, and b is a weight vector representing the learnable parameters for each gate. The subscripts indicate their respective gates. $\sigma(.)$ is the sigmoid function (outputs between 0 and 1), tanh(.) is the hyperbolic tangent function (outputs between -1 and 1), and $\odot$ is the elementwise multiplication.

2.1.2. Hyperparameters and training setups

The structure chosen for the study consisted of an input layer, a single hidden layer (LSTM), and an output layer. The input layer received an input array of two dimensions, one dimension representing the number of input features and the other dimension representing the number of time steps (lookback series). The number of input features depended on the input configuration (see **Section 2.2**), whereas the lookback series was fixed at 270 time steps. A dropout rate of 0.4 was applied to the LSTM layer during the training phase, which consisted of 256 memory cell units. The dropout rate was adopted as a regularization technique, i.e., to reduce the chances of overfitting the model (SRIVASTAVA *et al.*, 2014). The number of memory cells and the dropout rate were chosen based on the study by Kratzert *et al.* (2019b). The LSTM was run in a sequence-to-value mode to output a single prediction: the daily discharge. The models were trained considering the Kling-Gupta efficiency (*KGE*) (GUPTA *et al.*, 2009) as an objective function.

For training the model, the batch size was set to 1024 and the epochs were set to 100. The data set (see **Section 2.3. The HydroCerrado data set**) was split into training and testing periods. The training period consisted of the first 75% of available observations of the time series for each catchment, while the testing period consisted of the remaining 25% of available observations. A validation set, with 20% of the observations in the training period, was randomly subset from the training set. The validation set was used to track performance improvements over each epoch and avoid choosing weights from an overfitted model.

The models were built and trained using the R programming language and environment (R CORE TEAM, 2020), and the following R packages were used: dplyr (WICKHAM *et al.*, 2020), tensorflow (FALBEL *et al.*, 2020b), keras (FALBEL *et al.*, 2020a), and kerasgenerator (HERLAMBANG, 2019). The Keras callback “model

checkpoint" was used to save the model's weights after each epoch in the training phase. The weights that performed best for the validation set and without signs of overfitting were chosen as the final set of weights for the models.

2.2. Input configurations

Instead of the traditional station-by-station calibration of multiple models, this study focuses on the ability of a single regional model to capture hydrologic behaviors in different catchments. Therefore, three input configurations are used to train different LSTM hydrological models:

- *LSTM with no attributes*. A single LSTM hydrological model is trained on combined calibration data of all catchments using meteorological features.
- *LSTM with static attributes*. A single LSTM hydrological model is trained on combined calibration data of all catchments using time series of meteorological features and 21 static catchment attributes.
- *LSTM with static and dynamic attributes*. A single LSTM hydrological model is trained on combined calibration data of all catchments using time series of meteorological features and 12 static and 9 dynamic catchment attributes.

The meteorological features are time series of the daily precipitation (P) and daily Penman-Monteith reference evapotranspiration (E) averaged over each catchment, i.e., the pixels from daily grids within a given catchment boundary are averaged as a single value for each day. While only precipitation and evapotranspiration are considered for the first input configuration, 21 catchment attributes are considered for the second and third input configurations. These attributes relate to catchment topography (log-transformed area, mean elevation, and mean slope), climate indices (mean annual precipitation, mean annual evapotranspiration, timing of precipitation index, and aridity index), soil characteristics (clay content, silt content, sand content, organic carbon content, and bulk density), and land cover characteristics (fractions of forest formation, forest plantation, savanna formation, grassland, pasture, crops, urban infrastructure, river and lake, and wetland). All sources of input features are described in **Section 2.3**.

Although all the catchment attributes are considered static for the second input configuration, e.g. by assuming the average land cover characteristics over the study period, substantial changes in land cover characteristics have been documented in many regions of the Cerrado biome over the last few decades (MAPBIOMAS, 2020; ALTHOFF *et al.*, 2021, **Artigo 1**). Therefore, the third input configuration assesses whether accounting for environmental changes, represented by yearly information on land cover characteristics

(fractions), can improve streamflow predictions. While the second input configuration considers these variables static, i.e., the same values for the entire time series for each land cover fraction, the third input configuration updates the values of land cover fractions for different years.

2.3. The HydroCerrado data set

The HydroCerrado data set consists of a regional sample data set for catchments in the Cerrado biome compiled by Althoff *et al.* (2021, **Artigo 1**). The data set includes information on catchment boundaries, daily time series of observed streamflow, daily time series of meteorological forcing (precipitation and Penman-Monteith evapotranspiration), and catchment attributes. A brief description/overview is given in the following subsections for the target variable of the LSTM hydrological model, the daily streamflow, and the input features.

2.3.1. Daily streamflow

The streamflow database consists of daily streamflow observations for 411 catchments within or intersecting the Cerrado biome. The observations in the database cover the period from 2000-06-01 to 2014-12-31, where catchments have at least 5 years of equivalent observed data (median = 14.6 years, mean = 14.0 years). Catchment areas range from 51 to 746,000 km^2, and the mean daily streamflow ranges from 0.005 to 2.844 mm day^{-1}. More information on the data quality assessment can be found in Althoff *et al.* (2021, **Artigo 1**).

According to the Pfafstetter hierarchical classification of basins (FURNANS; OLIVERA, 2001; PFAFSTETTER, 1989), the Cerrado catchments are distributed over 6 of Brazil's 8 major basins (Figure 1a). The locations of the gauge stations and respective catchment boundaries are shown in Figure 1b.

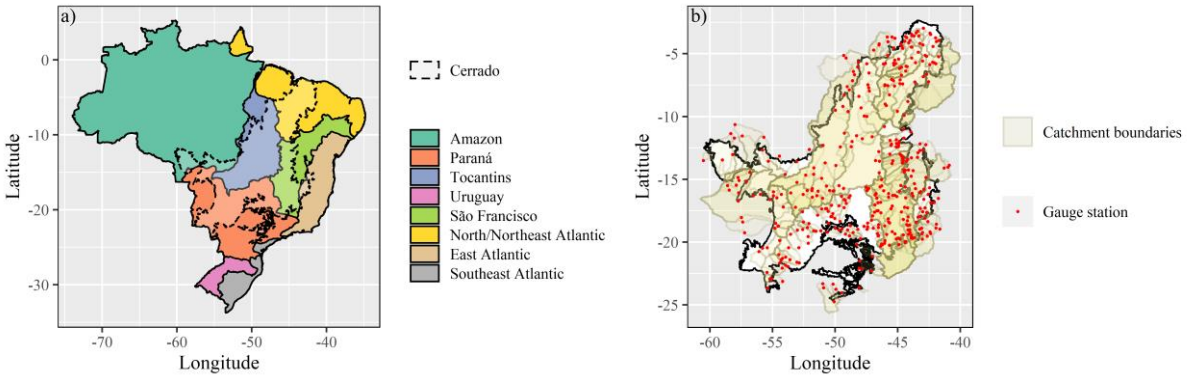

Figure 1. (a) The Cerrado biome in relation to Brazil's major basins (according to the Pfafstetter hierarchical classification) and **(b)** the catchment boundaries and gauge stations used in this study.

2.3.2. Input features

P and E daily time series were derived from the Integrated Multi-SatellitE Retrievals for Global Precipitation Measurement (IMERG) (HUFFMAN *et al.*, 2019b, 2019a) and the gridded reference evapotranspiration for Brazil based on machine learning (ETo-Brazil) (ALTHOFF *et al.*, 2020a, 2020b), respectively, and averaged over catchment areas. The log-transform area, mean altitude, and mean slope are derived from the Shuttle Radar Topography Mission (SRTM) digital elevation model at 3 arc-seconds (EARTH RESOURCES OBSERVATION AND SCIENCE CENTER, 2017). The mean annual precipitation ($\overline{P}$) and the mean annual evapotranspiration ($\overline{E}$) are derived from the P and E time series. The timing of the precipitation index is estimated using sine curves to represent the annual E and P cycles (Eq. 14 in Woods, 2009). The aridity index is the ratio between mean annual evapotranspiration and rainfall ($\overline{E}/\overline{P}$). Soil characteristics are derived from the SoilGrids data sets (HENGL *et al.*, 2017) for the top 2.0 m soil layer. Land use characteristics are derived from the MapBiomas data sets (MAPBIOMAS, 2020). An overview of several catchment attributes is given in Figure 2.

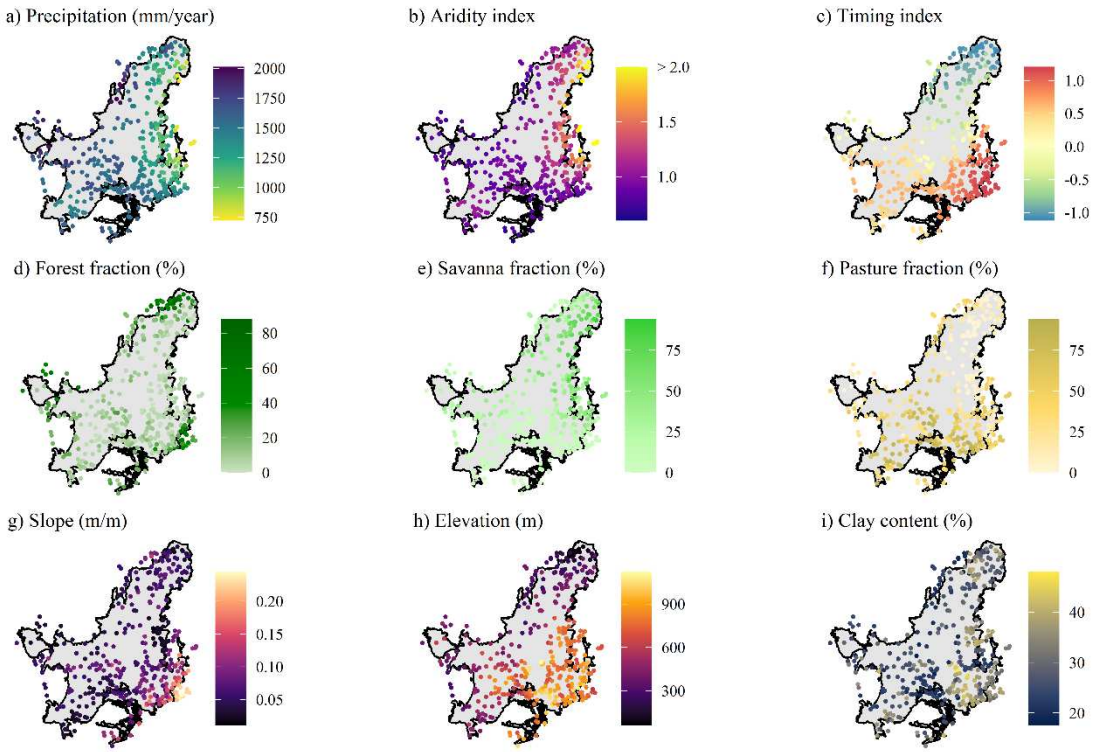

Figure 2. Overview of 9 catchment attributes: **(a)** mean annual precipitation, **(b)** aridity index, **(c)** timing of precipitation index, **(d)** forest formation fraction, **(e)** savanna formation fraction, **(e)** pasture fraction, **(f)** mean slope, **(g)** mean elevation, and **(h)** mean clay content.

Before training, all features are standardized to their respective mean and standard deviation of the training period. We recognize that, as derived gridded products, there is a general uncertainty in the catchment attributes. For example, taking the mean elevation or

annual rainfall from different digital elevation models or different precipitation gridded products would lead to different values. However, using multiple sources of the same feature, e.g., different gridded rainfall products, has been shown to improve performance in rainfall-runoff modeling with deep learning (KRATZERT *et al.*, 2020).

2.4. Assessment of model performance

The performance of the LSTM hydrological models was assessed for each catchment using the following criteria: Kling-Gupta efficiency (*KGE*) (GUPTA *et al.*, 2009), Nash-Sutcliffe efficiency (*NSE*) (NASH; SUTCLIFFE, 1970), and relative *NSE* (*rNSE*) (KRAUSE; BOYLE; BÄSE, 2005). These criteria are calculated as follows:

$$KGE = 1 - ED \quad , \quad -\infty < KGE \leq 1 \tag{7}$$

$$with: ED = \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2}$$

$$NSE = 1 - \frac{\sum_{i=1}^{n} |\hat{x}_i - x_i|^2}{\sum_{i=1}^{n} |x_i - \bar{x}|^2} \quad , \quad -\infty < NSE \leq 1 \tag{8}$$

$$rNSE = 1 - \frac{\sum_{i=1}^{n} \left| \frac{\hat{x}_i - x_i}{x_i} \right|^2}{\sum_{i=1}^{n} \left| \frac{x_i - \bar{x}}{\bar{x}} \right|^2} \quad , \quad -\infty < rNSE \leq 1 \tag{9}$$

where x_i and $\hat{x}_i$ are the observed and predicted streamflow at time step i, respectively, $\bar{x}$ is the mean observed streamflow, n is the number of observations, *ED* is the Euclidian distance, r is the linear correlation coefficient between $\hat{x}_i$ and x_i, α is the variability ratio or the ratio between the standard deviation of predicted values and the standard deviation of observed values ($\sigma_{\hat{x}}/\sigma_x$), and β is the ratio between the mean predicted and mean observed values ($\mu_{\hat{x}}/\mu_x$).

The *NSE* is one of the most traditional goodness-of-fit criteria used in water resources and is reported in most applicable studies. Due to some well-reported *NSE* limitations (GUPTA *et al.*, 2009; KRAUSE; BOYLE; BÄSE, 2005; LEGATES; MCCABE, 1999; SCHAEFLI; GUPTA, 2007), the *KGE* is chosen as the main performance indicator and objective function for training the models. The *KGE* decomposes errors in terms of mean flow, flow variability, and flow dynamics (GUPTA *et al.*, 2009). Since the *rNSE* gives more emphasis to low-flow conditions, it is chosen considering the importance of water availability in the Cerrado biome during the dry season. The paired Wilcoxon test is used to compare scores found for different models/input configurations, where $p < 0.05$ means that there is a statistical difference between scores.

The adoption of intervals and thresholds with respect to performance criteria can be viewed as a subjective means to assess the performance of hydrological models, especially in large-scale hydrology. This is because the criteria are not intercomparable among basins

under different contexts, i.e., different seasonality and rainfall-runoff dynamics (GUPTA *et al.*, 2009). Thus, the performance of the hydrological models is compared to the performance of benchmarks (**Section 2.5**). If a model performs better than the lower limit benchmark, the model can be considered useful.

2.5. Benchmarking

The intention in using benchmarks is to assess if the models result in reasonable simulations. Therefore, we compare the performance achieved by the LSTM hydrological models against a high- and a low-quality benchmark:

- The *benchmark model – "high-quality benchmark"* is a conceptual ("bucket-style") hydrological model calibrated to individual catchments.
- The *benchmark series – "low-quality benchmark"* is a mean monthly streamflow series observed for each catchment in the training period.

The high-quality benchmark represents an upper limit of expected performance for lumped hydrological models. In this case, we adopt the performance obtained for the testing period in a previous study (Althoff et al. 2021, **Artigo 1**) by running the GR5J model ("benchmark model") for the same catchments and on the same meteorological forcing. This is considered a high-quality benchmark since models were calibrated for each catchment individually in an attempt to better adjust to catchment uniqueness. The mean monthly streamflow series is considered a low-quality benchmark because it simply repeats the mean monthly streamflow for all days in a corresponding month. Thus, it has no complexity and considers no processes, despite being documented to provide high performance scores in some cases (ALTHOFF; RODRIGUES; BAZAME, 2021).

2.6. LSTM model interpretation

To interpret the LSTM hydrological models, the integrated gradients technique (SUNDARARAJAN; TALY; YAN, 2017) is adopted. Gradients are analogous to model coefficients and measure the attribution of each input feature to the output. Since neural networks are completely differentiable, it is possible to calculate the exact gradient with respect to the input features (KRATZERT *et al.*, 2019c). Therefore, the gradients can be integrated along the path of the input dimension (lookback series for LSTMs) and a baseline vector (sequence of zeros), measuring the attribution of each feature in the final prediction. In this study, the gradient integral was approximated using the trapezoidal rule. More details on the procedures for this technique are found in Kratzert *et al.* (2019a) and Sundararajan *et*

al. (2017). The following subsections describe how this technique is used to rank catchment attribute importance and compute the number of time steps of influence for the model.

2.6.1. Importance of features

To estimate the importance of the features used to train the models, i.e., catchment attributes, a global sensitivity analysis was performed. Gradients were obtained for all features and for each observation in the testing period. The global sensitivity measure of individual features was derived for each catchment by taking their average absolute gradient (KRATZERT *et al.*, 2019c). Because the magnitude of the streamflow varies for different catchments, for each catchment, we normalized the feature sensitivity with respect to the sum of sensitivities obtained for the catchment attributes.

2.6.2. Time steps of influence

The time steps of influence (TSOI) is a measure of how many past days actually influence the streamflow prediction (KRATZERT *et al.*, 2019a). To compute the TSOI, the procedure presented by Kratzert *et al.* (2019a) was used. First, integrated gradients were summed across features for each time step. Then, the difference from time step to time step was computed to find the time step t ($1 \leq t \leq T$) when the difference exceeded a threshold of 0.002. This threshold was empirically set by Kratzert *et al.* (2019a) to ignore noise in the integrated gradient signal. For each observation in the testing period, the TSOI was calculated by the following:

$$TSOI = T - t \tag{10}$$

where T is the total length of the sequence, i.e., 270, and t is the time step when the difference between consecutive time steps exceeds the threshold.

3. Results and discussion

3.1. Performance obtained by different input configurations

This section presents the comparison between three different input configurations used for the LSTM hydrological models discussed in **Section 2.2**. Figure 3 shows the cumulative density functions for the performance scores obtained for each catchment in the testing period. When considering only the meteorological forcing, the model obtained median *KGE*, *NSE*, and *rNSE* values of 0.40, 0.16, and 0.17, respectively. Approximately 25% of the stations gave a *KGE* score greater than 0.59. This input configuration for the LSTM hydrological model was able to improve the median *KGE* and median *rNSE* by 0.04 and 0.51, respectively, in relation to the benchmark series. On the other hand, by not considering catchment attributes, the median *KGE* and *rNSE* achieved by the model were

0.31 and 0.55 lower than those obtained by the high-quality benchmark, i.e., the performance obtained using GR5J models calibrated for each catchment.

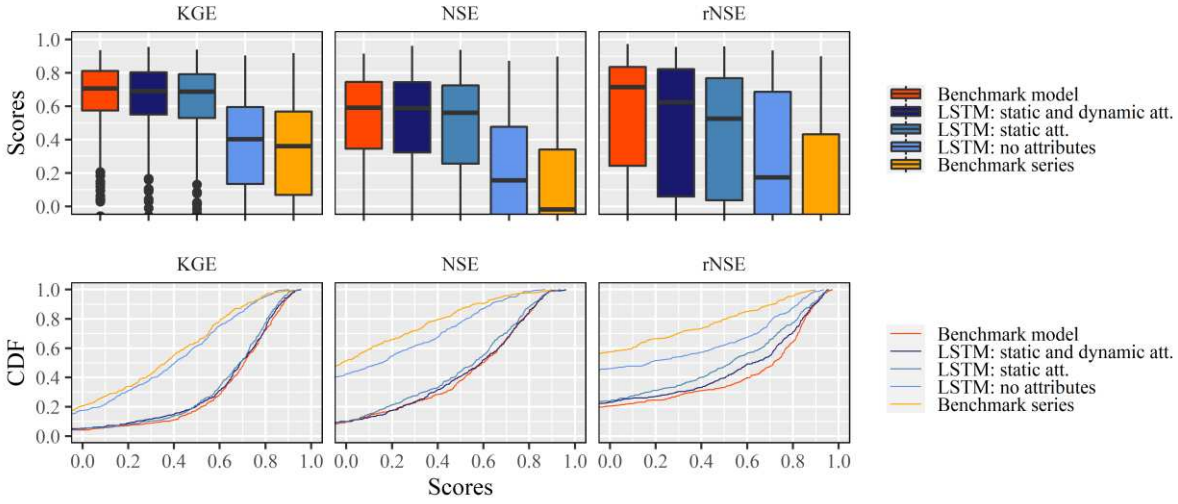

Figure 3. Boxplots and cumulative distribution functions of performance scores obtained for 411 catchments in the Cerrado biome.

By considering catchment attributes, the LSTM model showed significant improvement in the overall performance for the catchments, which was expected. When considering all the catchment attributes as static, the LSTM model scored median *KGE*, *NSE*, and *rNSE* values of 0.69, 0.56, and 0.53, respectively. These data indicate a *KGE* that is 0.33 times higher than that of the benchmark series and 0.29 times higher than that of the LSTM model considering no attributes. In contrast, the LSTM model trained on both static and dynamic catchment attributes, i.e., land characteristics as dynamic in time, resulted in median *KGE*, *NSE*, and *rNSE* values of 0.69, 0.59, and 0.62, respectively. Overall, both models that considered catchment attributes scored closer to the high-quality benchmark (*KGE* = 0.71 and *rNSE* = 0.72). However, the median *rNSE* score was higher by 0.09 for the LSTM model considering changes in land cover characteristics as opposed to the LSTM model considering such characteristics as static. This means that by considering the environmental changes over time, the model was able to give a better prediction under low flow conditions. Kratzert *et al.* (2019b) also highlight that randomness in the training procedures may be further mitigated by running model ensembles and improving final predictions.

The advantage of calibrating a model for each catchment is that the model is able to better capture the catchment uniqueness (BEVEN, 2020). Such uniqueness may be difficult for a regional model to capture since it may be poorly reflected in the available/considered attributes. In contrast, the conceptual structure of a hydrological model can also limit the performance for particular catchments. In addition, it is difficult for conceptual models to

extend their calibrated parameters to consider changes in land cover. These models are less robust to changes in both the environment and climate. The regional model, on the other hand, learns from a very diverse number of scenarios. Overall, the LSTM model with both static and dynamic attributes has lower *KGE* and *rNSE* scores in 55% (paired Wilcoxon test: $p < 0.001$) and 54% ($p < 0.001$) of the catchments, respectively, when compared to the benchmark model (high-quality benchmark) (Figure 4a). Although with similar performance, the Wilcoxon test shows that the performance of the LSTM model is still lacking when compared to a model calibrated on individual catchments. Considering land cover as dynamic also resulted in *KGE* and *rNSE* scores higher than 52% ($p = 0.52$) and 60% ($p < 0.001$) of the catchments, respectively, when compared to assuming these characteristics as static. This means that considering land cover changes significantly improved the predictions under low-flow conditions (higher *rNSE*).

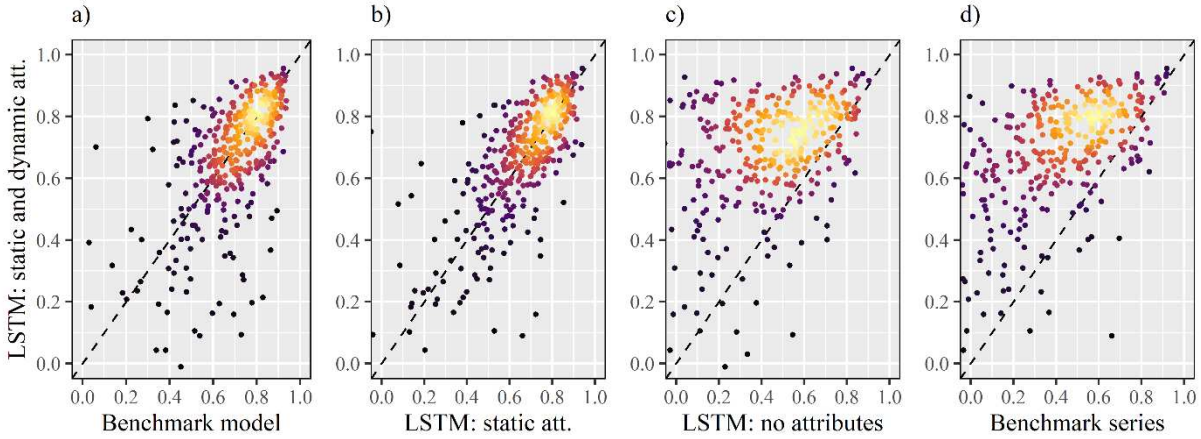

Figure 4. Comparison between *KGE* scores obtained by the LSTM hydrological model considering static and dynamic catchment attributes to those obtained by the **(a)** benchmark model, **(b)** LSTM model using all catchment attributes as static, **(c)** LSTM model using only meteorological forcing, and **(d)** benchmark series.

There are a few catchments that perform better by considering only the meteorological forcing, which is due to the general simulations of the LSTM that considers catchment attributes failing to represent the unique hydrological responses of particular catchments. In fact, the structure of the LSTM model has also failed to map hydrological processes for some catchments, since the lower limit benchmark, the time series based on mean monthly streamflow observed in the training period, is able to outperform the LSTM model in 10% of the catchments. Nevertheless, the model is useful for 90% of the catchments, outperforming the lower limit benchmark. Perhaps by considering different model structures and hyperparameters, a better overall performance could be reached for the LSTM model. Another alternative would be to mitigate randomness in predictions by running model ensembles (KRATZERT *et al.*, 2019c). However, in both cases, running

multiple calibrations on such a large data set would take a considerable amount of time for an ordinary computer.

3.2. Feature ranking

After confirming that catchment attributes increased the ability of the LSTM hydrological model to simulate streamflow, we assessed the overall importance of such features for predictions made in the testing period by collecting their individual integrated gradients. Figure 5 presents the average sensitivity of each catchment attribute for the 411 catchments in the testing period. The feature importance was ranked considering the LSTM model with both static and dynamic attributes as inputs. Among the catchment attributes, the land cover characteristics, which are the dynamic attributes, were the ones that showed the highest sensitivities, e.g., pasture (#1), grassland (#2), savanna formation (#3), crop (#5), and forest formation (#7) fractions. Pastures occupy most of the land cover in the Cerrado catchments, followed by savanna formation, forest formation, crops, and grassland. Despite its large relative areas in the Cerrado biome, the lesser attribution of streamflow to forest may be because they are generally less prone to runoff generation than the other aforementioned land covers (FALCÃO *et al.*, 2020). This is because forests are better at intercepting and infiltrating rainfall and, at the same time, make use of water available in deeper soil layers for evapotranspiration. The remaining land cover characteristics occupy relatively small portions of areas in the Cerrado biome and show lower sensitivity in the model.

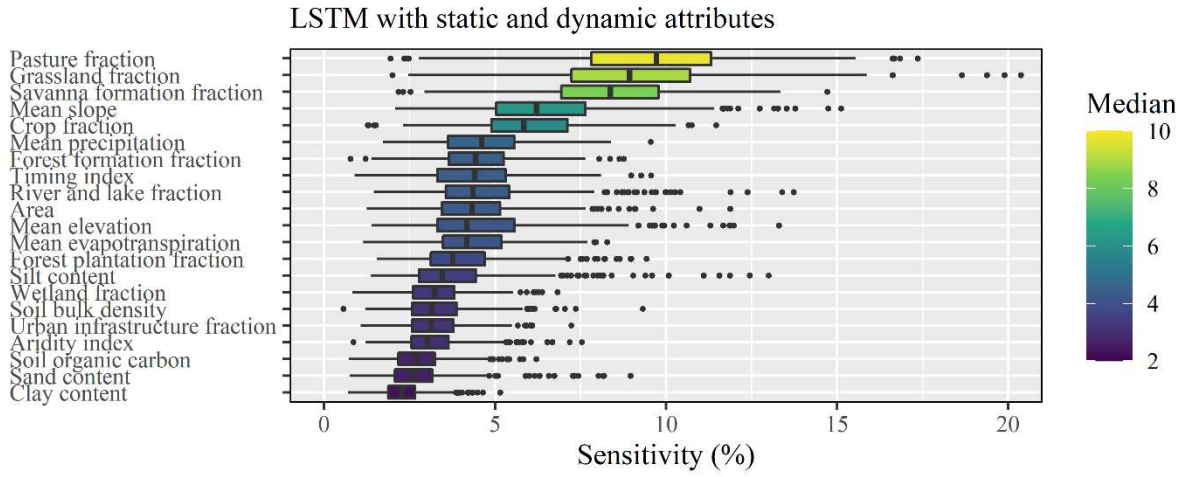

Figure 5. Catchment attributes and their relative contribution to streamflow (sensitivity) for 411 catchments in the Cerrado biome.

The topographic characteristic that showed the highest sensitivity was the catchment mean slope (#4), followed by area and (#10) mean elevation (#11). The effect of slope on runoff and infiltration has long been studied (NASSIF; WILSON, 1975), where an increased

slope generally results in increased runoff and lower infiltration, i.e., higher (or lower) flows during the wet (or dry) season. To confirm whether the model correctly learns these effects, the median gradients by catchment, attribute, and year are plotted against the catchment attributes (Figure 6). The gradients and attribute values are standardized for each plot to better assess their relationship regardless of their magnitude. Indeed, Figure 6 shows a positive slope for the catchment mean slope, while the area and mean elevation show slopes close to 0.

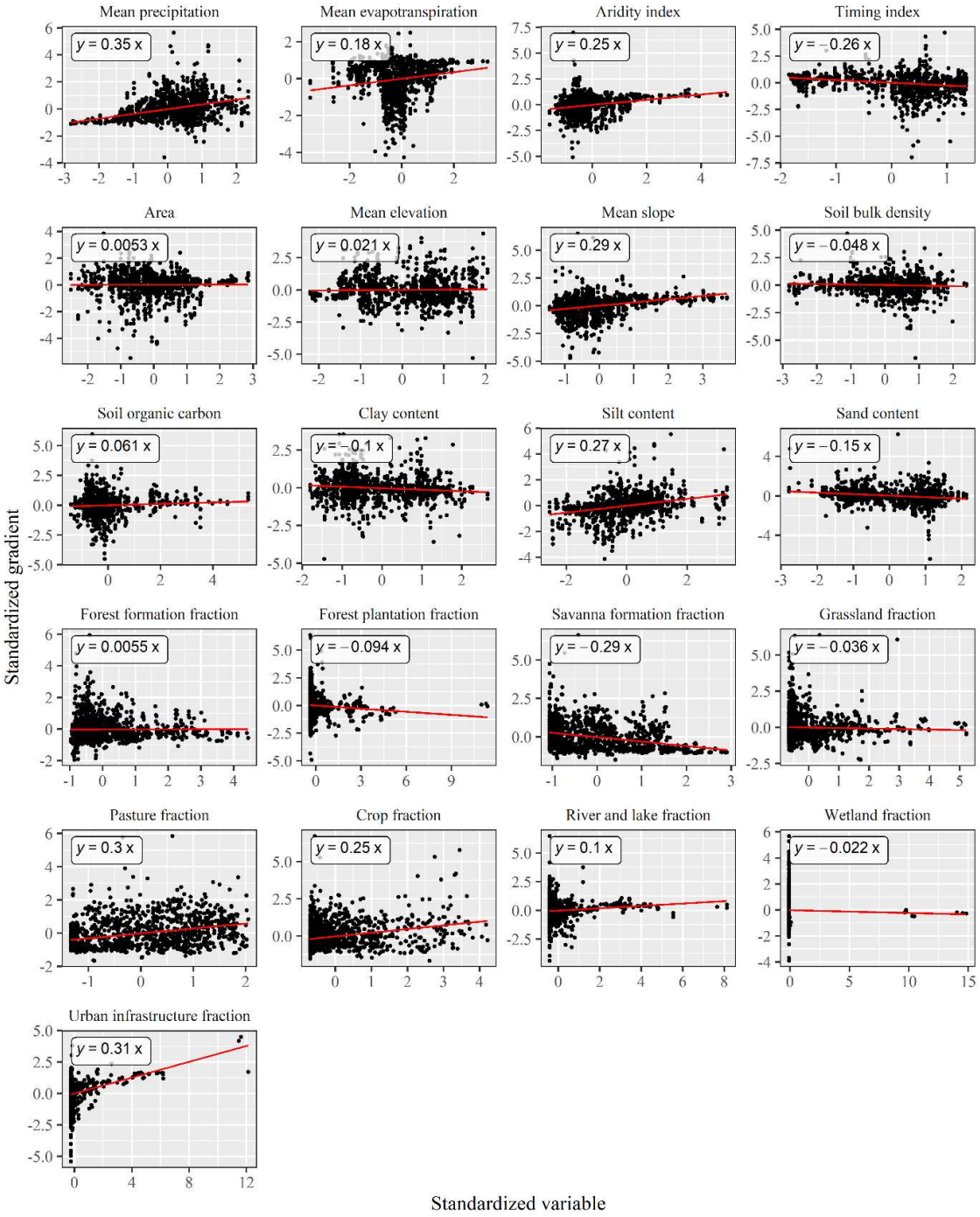

Figure 6. Scatterplot between median gradients plotted against the static catchment attributes for 411 catchments. Variables were standardized (mean = 0, standard deviation = 1) for each plot.

For climate characteristics, the mean precipitation (#6) and timing index (#8) show higher sensitivities than the mean evapotranspiration (#12) and aridity index (#18) (Figure 5). The mean precipitation sensitivity is higher, likely because it is the main driver of runoff generation and shows a positive correlation with the gradient (Figure 6). The precipitation timing index, on the other hand, indicates whether rainfall occurs in the warmer (positive index) or colder (negative index) season. Positive timing index values are related to lower attribution to streamflow since water is made available for evaporation in a more arid environment. This relationship is also confirmed by the negative slope in Figure 6. In contrast, mean evapotranspiration and the aridity index show a positive slope (Figure 6), which is not expected. Generally, higher evapotranspiration and more arid environments are related to lower streamflow, and a negative correlation to gradients is expected. The gradient decreases as the mean evapotranspiration increases until a given point but increases for higher mean evapotranspiration values. As the final streamflow prediction depends on the sum of all feature gradients, it is possible that the relationships of gradients to higher mean evapotranspiration and higher aridity became weak, thus presenting gradients closer to 0 (nonstandardized gradient values) for higher values of these variables. It is also important to consider that this technique attempts to zoom in on specific features, but the hydrological responses are a product of the complex interactions between the features.

Soil characteristics show the least importance to the model (Figure 5) and, therefore, the least attribution to streamflow. Generally, sand and organic carbon show a positive correlation with infiltration rates, whereas clay, silt, and bulk density show a negative correlation with infiltration rates (PATLE *et al.*, 2019). Figure 6 shows a negative (or positive) slope for sand (or silt), which confirms our previous knowledge. Sand (or silt) with higher (or lower) infiltration rates results in less (or more) direct runoff generation. Normally, a positive slope would be expected for clay. However, the soils in the Cerrado biome are older and well structured, providing soil with larger macropore volumes (SILVA *et al.*, 2003). Therefore, higher infiltration rates occur in the Cerrado biome despite loamy soils, reflecting a negative slope between clay and the gradient. Bulk density and soil organic matter, on the other hand, have slopes close to 0. Although we recognize that the uncertainties and biases related to using a global gridded soil product can reflect poor learning of the model, the soil characteristics provide some useful information to the model that agrees with previous hydrological knowledge in the region.

For land cover characteristics, Figure 6 shows a negative correlation with forest plantations and natural formations, with the exception of forest formation fraction, which shows slope close to zero. In general, this means that greater natural formation fractions

result in a decreased attribution to runoff, likely because of increased infiltration. In contrast, pasture and crop fractions show positive slopes, i.e., less infiltration and increased attribution to runoff. Rivers and lakes and urban infrastructure also show positive slopes because they usually directly translate rainfall into direct runoff, with the exception of reservoirs or other structures engineered to retain runoff/harvest rainfall. To provide a closer look at the land cover attribution to streamflow throughout the year, the median gradient is taken for each day of the year and catchment, and the interquartile ranges for all catchments are highlighted in Figure 7. For instance, vegetated land cover, with the exception of wetlands, show similar behavior. Although at different magnitudes, it has a higher positive attribution (gradient) to streamflow in the wet season and a lower attribution in the dry season.

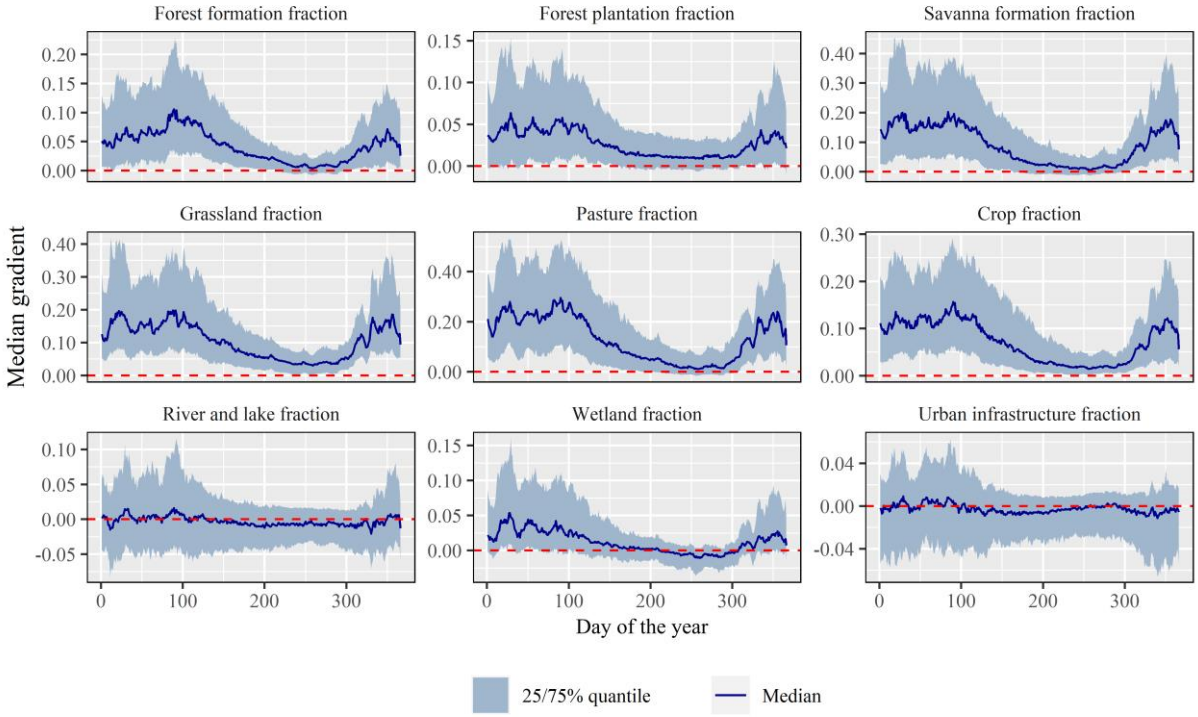

Figure 7. Median gradient by catchment, land cover characteristics, and day of the year. Only catchment medians and interquartile ranges are shown.

The median attribution (gradient) of urban infrastructure and river and lake fractions to streamflow generation is close to zero, with interquartile ranges showing both positive and negative values. The positive attribution in the wet season can be related to rainfall translating directly to streamflow, which would result in a negative attribution in the dry season, i.e., lower base flow. In contrast, the negative attribution in the wet season can be related to the interception of excess runoff in reservoirs or rain harvesting structures and could result in a positive attribution in the dry season (higher base flow). Wetland fractions likely show lower importance because of their small representativeness in terms of area and variance close to 0 (Figure 5); however, their hydrological behavior is accurately captured.

This result shows a positive attribution to streamflow in the wet season because precipitation falls in saturated areas. In contrast, lower infiltration results in a negative attribution in the dry season.

Overall, there is also much information with respect to the land cover characteristics that influence hydrological responses that are not available. For instance, how dense are the forests? How old are the forest plantations? Are pastures and grasslands degraded? Which crops are planted in which period of the year? What tillage systems are used? However, even if this information was available, there would likely still be many uncertainties and biases. Thus, it is not guaranteed that the model would be able to adequately learn these interactions, especially in a lumped model.

3.3. Time steps of influence

In this study, we followed the findings of a previous large-scale study (KRATZERT *et al.*, 2019c) to set the model hyperparameters. However, one important aspect that can be different for different regions and climates and that should be considered when building the LSTM models is the length of the lookback series. For example, using an overly extensive lookback series takes longer to train and can add unnecessary noise to the model. In contrast, a short lookback series may not be able to fully capture the effects related to the water balance and soil moisture accounting. Figure 8 presents the time steps of influence (TSOI) observed within each of the major hydrographic basins intersecting the Cerrado biome. The catchments within the North/Northeast Atlantic, East Atlantic, and São Francisco basins show a lower TSOI than the catchments within the Amazon, Tocantins, and Parana basins. The maximum median TSOIs of the North/Northeast Atlantic, East Atlantic, and São Francisco are 101, 115, and 114, respectively. The TSOI begins increasing at the beginning of the wet season and starts decreasing in the transition between the wet and dry seasons. As precipitation inputs decrease during the dry season, so does TSOI, which shows less influence from distant time steps.

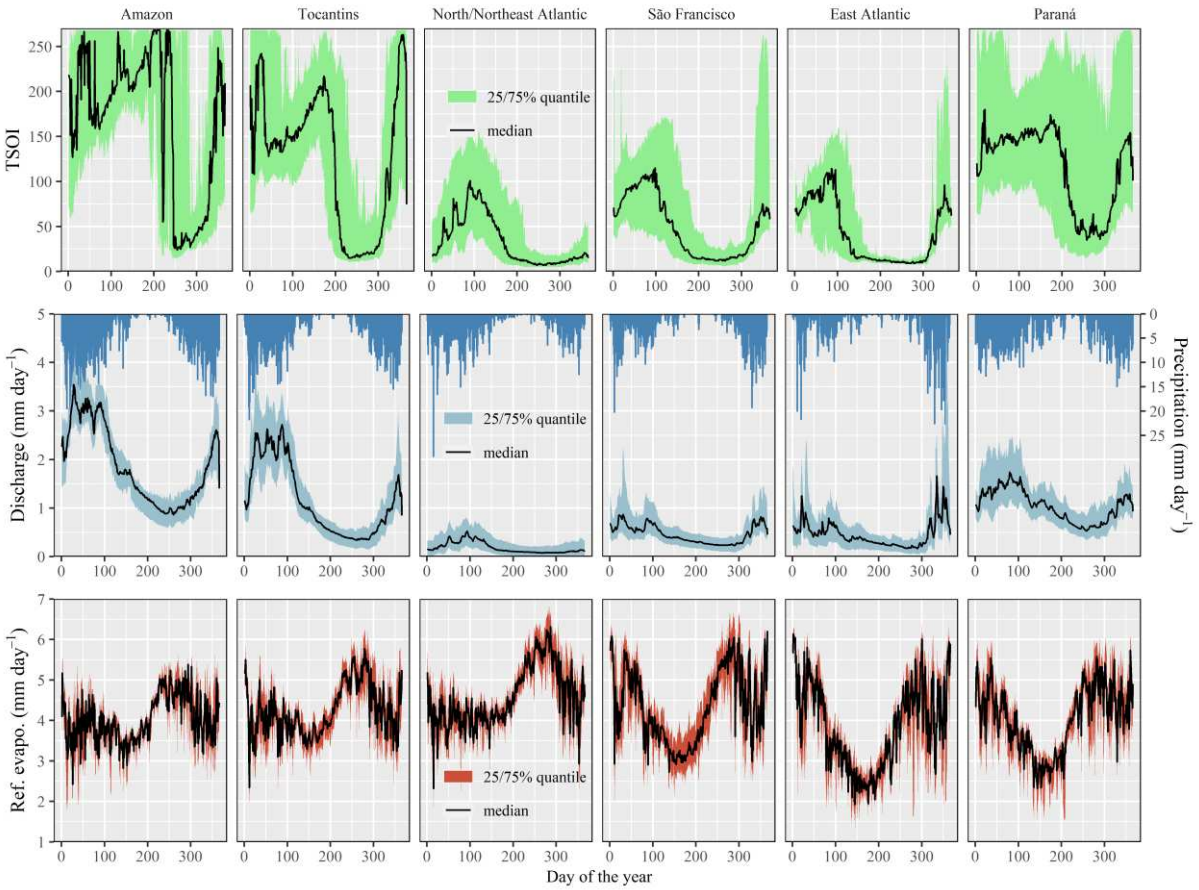

Figure 8. Time steps of influence, streamflow, precipitation, and reference evapotranspiration for the major hydrographic basins intersecting the Cerrado biome.

The maximum median TSOIs of the Amazon, Tocantins, and Paraná are 268, 263, and 180, respectively. Catchments with higher streamflow and precipitation magnitudes tend to have a higher dependency on the lookback series because soil moisture is a key factor in predicting runoff (GHAJARNIA *et al.*, 2020). Thus, longer series are required for the LSTM model to adequately represent the effects of antecedent soil moisture on rainfall-runoff events and model the processes as soil moisture accumulation. However, it seems that the 270 time step lookback series may not have been enough for some catchments, especially in the Amazon and Tocantins basins where maximum TSOI was frequently equal to the lookback series length. Even if we find that some catchments benefit from longer lookback series, we are building a single regional model. Thus, further tests are required to assess whether using a much longer lookback series can truly affect predictions for catchments influenced by shorter series. The alternative would be to find the best trade-off for the lookback series length across all catchments via hyperparameter tuning or using multiple regional models that cover smaller regions.

Because the threshold used to determine the TSOI (**Section 2.6.2**) was empirically derived by Kratzert *et al.* (2019a) considering only two catchments, there may be

inconsistencies in the TSOI analysis. In our empirical observations, the threshold may actually differ for different catchments and periods of the year. Thus, there is a degree of uncertainty in using a single threshold value. Perhaps more adequate methodologies can be developed, for example, the time step along the sum of integrated gradients across features that surpasses a small percentage of the sum of integrated gradients in the last step.

4. Conclusions

In this study, a single regional hydrological model is developed for the Cerrado biome based on long short-term memory neural networks using different input configurations. Three input configurations are assessed. One considers only the meteorological forcing (I1), another considers meteorological forcings and static catchment attributes (I2), and a third also considers meteorological forcings and catchment attributes but where the land cover characteristics are dynamic (I3).

Considering catchment attributes significantly improved the performance of the models (median *KGE*: 0.69) when compared to using only meteorological forcings (median *KGE*: 0.40). In addition, considering the land cover changes resulted in significant improvement for predictions made under low-flow conditions (median *rNSE*: 0.62) as opposed to assuming these attributes as static (median *rNSE*: 0.53). The models were benchmarked against a conceptual hydrological model calibrated on each basin individually (median *KGE* = 0.71) and performed closely. Despite not performing as well as the benchmark model calibrated on individual basins, we expect the regional LSTM model to be more robust to environmental and climate changes, since it has learned from a diverse number of examples in different climates and under significant land cover changes. This hypothesis should be tested in a future study by separating catchments during cross-validation and benchmarking against traditional streamflow regionalization techniques.

Among the catchment attributes, pasture, grassland, and savanna formation fractions showed the highest importance for streamflow prediction, followed by the catchments' mean slope and crop fractions. Previous hydrological knowledge of the region was compared to the features' overall attribution to streamflow to confirm that the LSTM hydrological model adequately learned from these attributes. The number of time steps of influence for the model was also explored in the context of the major hydrographic basins intersecting the Cerrado biome. We can conclude that longer lookback series, perhaps longer than 270 days, are required to account for soil moisture accumulation in regions with higher precipitation and streamflow throughout the year. However, shorter lookback series (< 200 days) seem sufficient in regions with shorter wet seasons and more arid conditions.

3. Conclusões gerais

Os métodos de regionalização tradicionalmente se baseiam na doação de parâmetros de modelos hidrológicos calibrados em regiões monitoradas para a estimativa de séries de vazão em regiões não monitoradas. A doação de informações a partir da proximidade espacial apresentou melhor desempenho que a doação a partir de similaridade de características físicas. Os modelos apresentaram desempenho satisfatório, superando, em geral, o desempenho da série de vazões utilizada como referência ("benchmark series"). No primeiro artigo também foi disponibilizado um banco de dados (HydroCerrado) com todas as informações compiladas para regiões monitoradas do Cerrado (411 bacias) e com séries de vazão simuladas para as ottobacias de nível 5 por meio da regionalização em todo o bioma Cerrado.

No segundo artigo, o clima do Cerrado foi caracterizado no que diz respeito às condições de aridez/umidade e a disponibilidade hídrica média avaliada por meio da vazão média e da vazão mínima de referência associada à permanência de 95% no tempo utilizada para concessão de outorgas de uso de água em nível federal no Brasil. O Cerrado apresenta maior disponibilidade hídrica nas regiões oeste e sul. Sua região central apresenta a maior vulnerabilidade quanto à disponibilidade hídrica, ou seja, elevada frequência de dias com vazão próximo ou abaixo da vazão mínima de referência para fins de outorga. O período mais vulnerável vai de agosto a outubro em quase todo o Cerrado, enquanto a região norte apresenta maior vulnerabilidade de novembro a janeiro. Também se deu destaque a regiões onde o uso sustentável de água pode melhorar a partir de técnicas de retenção de água da chuva. As implicações dessas técnicas foram discutidas quanto ao contexto climático de cada região.

O modelo hidrológico único desenvolvido para o Cerrado, permitiu a obtenção de melhores estimativas de vazão, em especial para condições de baixa vazão, quando as variáveis relacionadas ao uso e cobertura do solo são consideradas dinâmicas no tempo. Por meio de técnicas que permitem a interpretação de modelos baseados em inteligência artificial foi possível comparar o que o modelo aprendeu com base em informações hidrológicas prévias da região. O modelo mostrou que, em relação às características das bacias hidrográficas, os usos e cobertura da terra, representado por frações de pastagem, formação campestre, e de formação de savana, apresentaram a maior importância para o modelo. Também foi possível explorar a influência da série histórica nas predições de vazões para as bacias. Em geral, regiões com estações de precipitação longas requerem séries mais longas para a predição de vazão ($\geq$ 270 dias), enquanto regiões mais áridas requerem uma série mais curta ($<$ 200 dias). Isso se deve ao fato de as informações antecedentes estarem relacionadas

ao processo de acumulação de umidade no solo. De forma similar, as predições na estação seca geralmente dependem de uma série mais curta de dados. Uma vez que este modelo regional aprende a partir do fornecimento de informações de diversas bacias, sob diferentes contextos e dinâmicas de uso e cobertura do solo, se espera que apresente uma maior resiliência para fazer estimativas em regiões não monitoradas ou frente às mudanças contínuas do clima e/ou ambiente. Esta hipótese deve, contudo, ser mais bem avaliada em estudos futuros.

Referências

ABDULKAREEM, J. H.; PRADHAN, B.; SULAIMAN, W. N. A.; JAMIL, N. R. Long-term runoff dynamics assessment measured through land use/cover (LULC) changes in a tropical complex catchment. **Environment Systems and Decisions**, v. 39, n. 1, p. 16–33, 2019. Disponível em: https://doi.org/10.1007/s10669-018-9696-3

ADDOR, N.; NEARING, G.; PRIETO, C.; NEWMAN, A. J.; VINE, N. L.; CLARK, M. P. Selection of hydrological signatures for large-sample hydrology. 2018. Disponível em: https://eartharxiv.org/repository/view/1444/. Acesso em: 7 dez. 2020.

ALLEN, R. G.; PEREIRA, L. S.; RAES, D.; SMITH, M. **Crop evapotranspiration - Guidelines for computing crop water requirements - FAO Irrigation and drainage paper 56**. 9. ed. Food and Agriculture Organization of the United Nations, Rome: *[s. n.]*, 1998.

ALMAGRO, A.; OLIVEIRA, P. T. S.; MEIRA NETO, A. A.; ROY, T.; TROCH, P. CABra: a novel large-sample dataset for Brazilian catchments. **Hydrology and Earth System Sciences Discussions**, p. 1–40, 2020. Disponível em: https://doi.org/10.5194/hess-2020-521

ALTHOFF, D.; DIAS, S. H. B.; FILGUEIRAS, R.; RODRIGUES, L. N. ETo-Brazil: A Daily Gridded Reference Evapotranspiration Data Set for Brazil - Repository. **Mendeley Data**, v. V3, 2020 a. Disponível em: https://doi.org/10.17632/sstjw74ryh.3

ALTHOFF, D.; DIAS, S. H. B.; FILGUEIRAS, R.; RODRIGUES, L. N. ETo-Brazil: A Daily Gridded Reference Evapotranspiration Data Set for Brazil (2000–2018). **Water Resources Research**, v. 56, n. 7, 2020 b. Disponível em: https://doi.org/10.1029/2020WR027562. Acesso em: 20 jul. 2020.

ALTHOFF, D.; RODRIGUES, L. N. The expansion of center-pivot irrigation in the Cerrado biome. **IRRIGA**, v. 1, n. 1, Inovagri Special Edition, p. 56–61, 2019. Disponível em: https://doi.org/10.15809/irriga.2019v1n1p56-61

ALTHOFF, D.; RODRIGUES, L. N.; BAZAME, H. C. Uncertainty quantification for hydrological models based on neural networks: the dropout ensemble. **Stochastic Environmental Research and Risk Assessment**, 2021. Disponível em: https://doi.org/10.1007/s00477-021-01980-8. Acesso em: 5 fev. 2021.

ALTHOFF, D.; RODRIGUES, L. N.; SILVA, D. D. da. Impacts of climate change on the evaporation and availability of water in small reservoirs in the Brazilian savannah. **Climatic Change**, v. 159, n. 2, p. 215–232, 2020. Disponível em: https://doi.org/10.1007/s10584-020-02656-y

ALVARES, C. A.; STAPE, J. L.; SENTELHAS, P. C.; DE MORAES GONÇALVES, J. L.; SPAROVEK, G. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v. 22, n. 6, p. 711–728, 2013. Disponível em: https://doi.org/10.1127/0941-2948/2013/0507

ANA. **Subsídios para a discussão da compatibilização da geração de energia hidrelétrica com expansão da agricultura irrigada na bacia do rio São Marcos**. Brasília - DF, Brazil: Agência Nacional de Águas, 2014.

ANA. **Atlas Irrigação**: Uso de água na agricultura irrigada. Brasília - DF, Brazil: Agência Nacional de Águas, 2017 a.

ANA. **Conjuntura dos recursos Hídricos no Brasil 2017**: Relatório pleno. Brasília - DF, Brazil: Agência Nacional de Águas, 2017 b.

ANA. **Levantamento da agricultura irrigada por pivôs centrais no Brasil (1985-2017)**. Brasília - DF, Brazil: Agência Nacional de Águas, Embrapa Milho e Sorgo, 2019.

ANA. **HidroWeb: Sistema de Informações Hidrológicas**. Versão 3.1.1. *[S. l.]*: Agência Nacional de Águas (ANA), 2020 a. *E-book*. Disponível em: http://www.snirh.gov.br/hidroweb/. Acesso em: 10 abr. 2020.

ANA. **Portal de Metadados Geoespaciais**. *[S. l.]*: Agência Nacional de Águas (ANA), 2020 b. *E-book*. Disponível em: https://metadados.snirh.gov.br/geonetwork/. Acesso em: 10 abr. 2020.

ARSENAULT, R.; BRISSETTE, F. Analysis of continuous streamflow regionalization methods within a virtual setting. **Hydrological Sciences Journal**, v. 61, n. 15, p. 2680–2693, 2016. Disponível em: https://doi.org/10.1080/02626667.2016.1154557

AYZEL, G.; HEISTERMANN, M. The effect of calibration data length on the performance of a conceptual hydrological model versus LSTM and GRU: A case study for six basins from the CAMELS dataset. **Computers & Geosciences**, p. 104708, 2021. Disponível em: https://doi.org/10.1016/j.cageo.2021.104708

BAI, P.; LIU, X.; XIE, J. Simulating runoff under changing climatic conditions: A comparison of the long short-term memory network with two conceptual hydrologic models. **Journal of Hydrology**, v. 592, p. 125779, 2021. Disponível em: https://doi.org/10.1016/j.jhydrol.2020.125779

BECK, H. E.; DIJK, A. I. J. M. van; ROO, A. de; DUTRA, E.; FINK, G.; ORTH, R.; SCHELLEKENS, J. Global evaluation of runoff from 10 state-of-the-art hydrological models. **Hydrology and Earth System Sciences**, v. 21, n. 6, p. 2881–2903, 2017. Disponível em: https://doi.org/10.5194/hess-21-2881-2017

BECK, H. E.; DIJK, A. I. J. M. van; ROO, A. de; MIRALLES, D. G.; MCVICAR, T. R.; SCHELLEKENS, J.; BRUIJNZEEL, L. A. Global-scale regionalization of hydrologic model parameters. **Water Resources Research**, v. 52, n. 5, p. 3599–3622, 2016. Disponível em: https://doi.org/10.1002/2015WR018247

BECK, H. E.; PAN, M.; LIN, P.; SEIBERT, J.; DIJK, A. I. J. M. van; WOOD, E. F. Global Fully Distributed Parameter Regionalization Based on Observed Streamflow From 4,229 Headwater Catchments. **Journal of Geophysical Research: Atmospheres**, v. 125, n. 17, p. e2019JD031485, 2020. Disponível em: https://doi.org/10.1029/2019JD031485

BEVEN, K. Deep learning, hydrological processes and the uniqueness of place. **Hydrological Processes**, v. 34, n. 16, p. 3608–3613, 2020. Disponível em: https://doi.org/10.1002/hyp.13805

BOELEE, E.; YOHANNES, M.; PODA, J.-N.; MCCARTNEY, M.; CECCHI, P.; KIBRET, S.; HAGOS, F.; LAAMRANI, H. Options for water storage and rainwater harvesting to improve health and resilience against climate change in Africa. **Regional Environmental Change**, v. 13, n. 3, p. 509–519, 2013. Disponível em: https://doi.org/10.1007/s10113-012-0287-4

BOSCH, J. M.; HEWLETT, J. D. A review of catchment experiments to determine the effect of vegetation changes on water yield and evapotranspiration. **Journal of Hydrology**, v. 55, n. 1, p. 3–23, 1982. Disponível em: https://doi.org/10.1016/0022-1694(82)90117-2

CAVUSOGLU, A.-H.; CHEN, X.; GENTINE, P.; SAHIN, O. Potential for natural evaporation as a reliable renewable energy resource. **Nature Communications**, v. 8, n. 1, p. 1–9, 2017. Disponível em: https://doi.org/10.1038/s41467-017-00581-w

CHEN, L.; WANG, L. Recent advance in earth observation big data for hydrology. **Big Earth Data**, v. 2, n. 1, p. 86–107, 2018. Disponível em: https://doi.org/10.1080/20964471.2018.1435072

CHOU, S. C. *et al.* Assessment of climate change over South America under RCP 4.5 and 8.5 downscaling scenarios. **American Journal of Climate Change**, v. 03, n. 05, p. 512–527, 2014. Disponível em: https://doi.org/10.4236/ajcc.2014.35043

CLERC, M. Beyond standard particle swarm optimisation. *In*: **Innovations and Developments of Swarm Intelligence Applications**. *[S. l.]*: IGI Global, 2012. p. 1–19.

COMAS, L. H.; TROUT, T. J.; DEJONGE, K. C.; ZHANG, H.; GLEASON, S. M. Water productivity under strategic growth stage-based deficit irrigation in maize. **Agricultural Water Management**, v. 212, p. 433–440, 2019. Disponível em: https://doi.org/10.1016/j.agwat.2018.07.015

CORON, L. *et al.* **airGR: Suite of GR Hydrological Models for Precipitation-Runoff Modelling**. Versão 1.4.3.65. *[S. l.: s. n.]*. *E-book*. Disponível em: https://CRAN.R-project.org/package=airGR. Acesso em: 21 abr. 2020.

CORON, L.; THIREL, G.; DELAIGUE, O.; PERRIN, C.; ANDRÉASSIAN, V. The suite of lumped GR hydrological models in an R package. **Environmental Modelling & Software**, v. 94, p. 166–171, 2017. Disponível em: https://doi.org/10.1016/j.envsoft.2017.05.002

COSTA, M. H.; PIRES, G. F. Effects of Amazon and Central Brazil deforestation scenarios on the duration of the dry season in the arc of deforestation. **International Journal of Climatology**, v. 30, n. 13, p. 1970–1979, 2010. Disponível em: https://doi.org/10.1002/joc.2048

CREMAQ, P. The miracle of the cerrado. **The Economist**, v. 26, p. 08–10, 2010.

DEVIA, G. K.; GANASRI, B. P.; DWARAKISH, G. S. A review on hydrological models. **Aquatic Procedia**, v. 4, INTERNATIONAL CONFERENCE ON WATER RESOURCES, COASTAL AND OCEAN ENGINEERING (ICWRCOE'15), p. 1001–1007, 2015. Disponível em: https://doi.org/10.1016/j.aqpro.2015.02.126

DOUGLAS-MANKIN, K. R.; SRINIVASAN, R.; ARNOLD, J. G. Soil and Water Assessment Tool (SWAT) Model: Current Developments and Applications. **Transactions of the ASABE**, v. 53, n. 5, p. 1423–1431, 2010. Disponível em: https://doi.org/10.13031/2013.34915

EARTH RESOURCES OBSERVATION AND SCIENCE CENTER. Shuttle Radar Topography Mission (SRTM) Void Filled. U.S. Geological Survey, , 2017. Disponível em: https://doi.org/10.5066/f7f76b1x. Acesso em: 29 maio. 2019.

FABEIRO, C.; OLALLA, F. M. S.; JUAN, J. A. Yield and size of deficit irrigated potatoes. **Agricultural Water Management**, v. 48, n. 3, p. 255–266, 2001. Disponível em: https://doi.org/10.1016/S0378-3774(00)00129-3

FALBEL, D.; ALLAIRE, J. J.; CHOLLET, F.; RSTUDIO; GOOGLE; TANG, Y.; BIJL, W. V. D.; STUDER, M.; KEYDANA, S. **keras: R Interface to "Keras"**. Versão 2.3.0.0. *[S. l.: s. n.]*. *E-book*. Disponível em: https://CRAN.R-project.org/package=keras. Acesso em: 23 jul. 2020a.

FALBEL, D.; ALLAIRE, J. J.; RSTUDIO; TANG, Y.; EDDELBUETTEL, D.; GOLDING, N.; KALINOWSKI, T.; GOOGLE. **tensorflow: R Interface to "TensorFlow"**. Versão 2.2.0. *[S. l.: s. n.]*. *E-book*. Disponível em: https://CRAN.R-project.org/package=tensorflow. Acesso em: 4 jan. 2021b.

FALCÃO, K. dos S.; PANACHUKI, E.; MONTEIRO, F. das N.; DA SILVA MENEZES, R.; RODRIGUES, D. B. B.; SONE, J. S.; OLIVEIRA, P. T. S. Surface runoff and soil erosion in a natural regeneration area of the Brazilian Cerrado. **International Soil and Water Conservation Research**, p. S2095633920300241, 2020. Disponível em: https://doi.org/10.1016/j.iswcr.2020.04.004

FAN, H.; JIANG, M.; XU, L.; ZHU, H.; CHENG, J.; JIANG, J. Comparison of Long Short Term Memory Networks and the Hydrological Model in Runoff Simulation. **Water**, v. 12, n. 1, p. 175, 2020. Disponível em: https://doi.org/10.3390/w12010175

FEALQ. **Análise territorial para o desenvolvimento da agricultura irrigada no Brasil**. Piracicaba, SP, Brazil: Fundação de Estudos Agrários Luiz de Queiroz, 2014.

FURNANS, J.; OLIVERA, F. Watershed Topology: The Pfafstetter System. *In*: 21ST ANNUAL ESRI INTERNATIONAL USER CONFERENCE2001, **2001 User Conference Proceedings**. *[S. l.: s. n.]* Disponível em: https://proceedings.esri.com/library/userconf/proc01/professional/papers/pap1008/p1008.htm. Acesso em: 22 dez. 2020.

GADELHA, A. N.; COELHO, V. H. R.; XAVIER, A. C.; BARBOSA, L. R.; MELO, D. C. D.; XUAN, Y.; HUFFMAN, G. J.; PETERSEN, W. A.; ALMEIDA, C. das N. Grid box-level evaluation of IMERG over Brazil at various space and time scales. **Atmospheric Research**, v. 218, p. 231–244, 2019. Disponível em: https://doi.org/10.1016/j.atmosres.2018.12.001

GASHAW, T.; TULU, T.; ARGAW, M.; WORQLUL, A. W. Modeling the hydrological impacts of land use/land cover changes in the Andassa watershed, Blue Nile Basin, Ethiopia. **Science of The Total Environment**, v. 619–620, p. 1394–1408, 2018. Disponível em: https://doi.org/10.1016/j.scitotenv.2017.11.191

GEERTS, S.; RAES, D. Deficit irrigation as an on-farm strategy to maximize crop water productivity in dry areas. **Agricultural Water Management**, v. 96, n. 9, p. 1275–1284, 2009. Disponível em: https://doi.org/10.1016/j.agwat.2009.04.009

GHAJARNIA, N.; KALANTARI, Z.; DESTOUNI, G. **Data-driven worldwide quantification of large-scale hydroclimatic co-variation patterns and comparison with reanalysis and Earth System modeling**. *[S. l.]*: Hydrology, 2020. preprint. Disponível em: https://doi.org/10.1002/essoar.10505258.1. Acesso em: 24 dez. 2020.

GHAJARNIA, N.; KALANTARI, Z.; ORTH, R.; DESTOUNI, G. Close co-variation between soil moisture and runoff emerging from multi-catchment data across Europe. **Scientific Reports**, v. 10, n. 1, p. 4817, 2020. Disponível em: https://doi.org/10.1038/s41598-020-61621-y

GUO, Y.; ZHANG, Y.; ZHANG, L.; WANG, Z. Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: A comprehensive review. **WIREs Water**, v. 8, n. 1, p. e1487, 2021. Disponível em: https://doi.org/10.1002/wat2.1487

GUPTA, H. V.; KLING, H.; YILMAZ, K. K.; MARTINEZ, G. F. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. **Journal of Hydrology**, v. 377, n. 1, p. 80–91, 2009. Disponível em: https://doi.org/10.1016/j.jhydrol.2009.08.003

GUSTARD, A.; BULLOCK, A.; DIXON, J. M. **Low flow estimation in the United Kingdom**. *[S. l.]*: Institute of Hydrology, 1992.

HALL, J.; BORGOMEO, E. Risk-based principles for defining and managing water security. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, v. 371, n. 2002, p. 20120407, 2013. Disponível em: https://doi.org/10.1098/rsta.2012.0407

HE, X.; WADA, Y.; WANDERS, N.; SHEFFIELD, J. Intensification of hydrological drought in California by human water management. **Geophysical Research Letters**, v. 44, n. 4, p. 1777–1785, 2017. Disponível em: https://doi.org/10.1002/2016GL071665

HENGL, T. *et al.* SoilGrids250m: Global gridded soil information based on machine learning. **PLoS One**, v. 12, n. 2, p. e0169748, 2017. Disponível em: https://doi.org/10.1371/journal.pone.0169748

HERLAMBANG, R. D. B. **kerasgenerator: Data generator for deep learning using Keras interface to R**. Versão 0.0.0.9000. *[S. l.: s. n.]*. *E-book*. Disponível em: https://github.com/bagasbgy/kerasgenerator/. Acesso em: 4 jan. 2021.

HOCHREITER, S.; SCHMIDHUBER, J. LSTM can Solve Hard Long Time Lag Problems. *In*: MOZER, M. C.; JORDAN, M. I.; PETSCHE, T. (org.). **Advances in Neural Information Processing Systems 9**. *[S. l.]*: MIT Press, 1997 a. p. 473–479. *E-book*. Disponível em: http://papers.nips.cc/paper/1215-lstm-can-solve-hard-long-time-lag-problems.pdf. Acesso em: 23 jul. 2020.

HOCHREITER, S.; SCHMIDHUBER, J. Long Short-Term Memory. **Neural Computation**, v. 9, n. 8, p. 1735–1780, 1997 b. Disponível em: https://doi.org/10.1162/neco.1997.9.8.1735

HOSONO, A.; ROCHA, C. M. C. da; HONGO, Y. **Development for Sustainable Agriculture: The Brazilian Cerrado**. *[S. l.]*: Springer, 2016.

HUFFMAN, G. J.; BOLVIN, D. T.; BRAITHWAITE, D.; HSU, K.; JOYCE, R.; KIDD, C.; NELKIN, E. J.; SOROOSHIAN, S.; TAN, J.; XIE, P. **Integrated Multi-satellitE Retrievals for GPM (IMERG)**: Algorithm Theoretical Basis Document. *[S. l.]*: National Aeronautics and Space Administration, 2019 a. Disponível em: https://docserver.gesdisc.eosdis.nasa.gov/public/project/GPM/IMERG_ATBD_V06.pdf.

HUFFMAN, G. J.; STOCKER, E. F.; BOLVIN, D. T.; NELKIN, E. J.; TAN, J. **GPM IMERG Final Precipitation L3 1 day 0.1 degree x 0.1 degree V06**. Greenbelt, MD: Goddard Earth Sciences Data and Information Services Center (GES DISC), 2019 b. Disponível em: 10.5067/GPM/IMERGDF/DAY/06.

IBGE. **Censo Agropecuário 2017 [In English: Brazilian Agricultural Census 2017]**. *[S. l.]*: Instituto Brasileiro de Geografia e Estatística, 2017.

IBGE. **Biomas e Sistema Costeiro-Marinho do Brasil - 1:250 000**. *[S. l.]*: Instituto Brasileiro de Geografia e Estatística, 2019. *E-book*. Disponível em: https://www.ibge.gov.br/geociencias/informacoes-ambientais/15842-biomas.html. Acesso em: 10 abr. 2020.

IGLESIAS, A.; GARROTE, L. Adaptation strategies for agricultural water management under climate change in Europe. **Agricultural Water Management**, v. 155, p. 113–124, 2015. Disponível em: https://doi.org/10.1016/j.agwat.2015.03.014

JIANG, D.; WANG, K. The role of satellite-based remote sensing in improving simulated streamflow: A review. **Water**, v. 11, n. 8, p. 1615, 2019. Disponível em: https://doi.org/10.3390/w11081615

KIM, K. B.; KWON, H.-H.; HAN, D. Exploration of warm-up period in conceptual hydrological modelling. **Journal of Hydrology**, v. 556, p. 194–210, 2018. Disponível em: https://doi.org/10.1016/j.jhydrol.2017.11.015

KLINK, C. A. Policy intervention in the Cerrado savannas of Brazil: Changes in the land use and effects on conservation. **A. Consorte-McCrea, & E. Ferraz Santos, Ecology and Conservation of the Maned Wolf: Multidisciplinary Perspectives**, p. 293–308, 2014.

KNOBEN, W. J. M.; WOODS, R. A.; FREER, J. E. A Quantitative Hydrological Climate Classification Evaluated With Independent Streamflow Data. **Water Resources Research**, v. 54, n. 7, p. 5088–5109, 2018. Disponível em: https://doi.org/10.1029/2018WR022913

KRATZERT, F.; HERRNEGGER, M.; KLOTZ, D.; HOCHREITER, S.; KLAMBAUER, G. NeuralHydrology – Interpreting LSTMs in Hydrology. *In*: SAMEK, W.; MONTAVON, G.; VEDALDI, A.; HANSEN, L. K.; MÜLLER, K.-R. (org.). **Explainable AI: Interpreting, Explaining and Visualizing Deep Learning**. Cham: Springer International Publishing, 2019 a. (Lecture Notes in Computer Science).p. 347–362. *E-book*. Disponível em: https://doi.org/10.1007/978-3-030-28954-6_19. Acesso em: 27 nov. 2020.

KRATZERT, F.; KLOTZ, D.; BRENNER, C.; SCHULZ, K.; HERRNEGGER, M. Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. **Hydrology and Earth System Sciences**, v. 22, n. 11, p. 6005–6022, 2018. Disponível em: https://doi.org/10.5194/hess-22-6005-2018

KRATZERT, F.; KLOTZ, D.; HERRNEGGER, M.; SAMPSON, A. K.; HOCHREITER, S.; NEARING, G. S. Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. **Water Resources Research**, v. 55, n. 12, p. 11344–11354, 2019 b. Disponível em: https://doi.org/10.1029/2019WR026065

KRATZERT, F.; KLOTZ, D.; HOCHREITER, S.; NEARING, G. S. **A note on leveraging synergy in multiple meteorological datasets with deep learning for rainfall-runoff**

modeling. *[S. l.]*: Global hydrology/Modelling approaches, 2020. preprint. Disponível em: https://doi.org/10.5194/hess-2020-221. Acesso em: 7 jan. 2021.

KRATZERT, F.; KLOTZ, D.; SHALEV, G.; KLAMBAUER, G.; HOCHREITER, S.; NEARING, G. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. **Hydrology and Earth System Sciences**, v. 23, n. 12, p. 5089–5110, 2019 c. Disponível em: https://doi.org/10.5194/hess-23-5089-2019

KRAUSE, P.; BOYLE, D. P.; BÄSE, F. Comparison of different efficiency criteria for hydrological model assessment. **Advances in Geosciences**, v. 5, p. 89–97, 2005. Disponível em: https://doi.org/10.5194/adgeo-5-89-2005

KUEPPERS, L. M.; SNYDER, M. A.; SLOAN, L. C. Irrigation cooling effect: Regional climate forcing by land-use change. **Geophysical Research Letters**, v. 34, n. 3, 2007. Disponível em: https://doi.org/10.1029/2006GL028679. Acesso em: 31 jan. 2021.

LAMBERS, H.; DE BRITTO COSTA, P.; OLIVEIRA, R. S.; SILVEIRA, F. A. O. Towards more sustainable cropping systems: lessons from native Cerrado species. **Theoretical and Experimental Plant Physiology**, v. 32, n. 3, p. 175–194, 2020. Disponível em: https://doi.org/10.1007/s40626-020-00180-z

LATHUILLIÈRE, M. J.; COE, M. T.; JOHNSON, M. S. A review of green- and blue-water resources and their trade-offs for future agricultural production in the Amazon Basin: what could irrigated agriculture mean for Amazonia? **Hydrology and Earth System Sciences**, v. 20, n. 6, p. 2179–2194, 2016. Disponível em: https://doi.org/10.5194/hess-20-2179-2016

LE MOINE, N. **Le bassin versant de surface vu par le souterrain: une voie d'amélioration des performances et du réalisme des modèles pluie-débit**. 2008. Ph.D. thesis - Université Pierre et Marie Curie, Paris, 2008. Disponível em: https://hal.inrae.fr/tel-02591478/document

LEGATES, D. R.; MCCABE, G. J. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. **Water Resources Research**, v. 35, n. 1, p. 233–241, 1999. Disponível em: https://doi.org/10.1029/1998WR900018

LELE, S. Watershed services of tropical forests: from hydrology to economic valuation to integrated analysis. **Current Opinion in Environmental Sustainability**, v. 1, n. 2, p. 148–155, 2009. Disponível em: https://doi.org/10.1016/j.cosust.2009.10.007

LEMES, L.; DE ANDRADE, A. F. A.; LOYOLA, R. Spatial priorities for agricultural development in the Brazilian Cerrado: may economy and conservation coexist? **Biodiversity and Conservation**, v. 29, n. 5, p. 1683–1700, 2020. Disponível em: https://doi.org/10.1007/s10531-019-01719-6

LIMA, J. E. F. W. Situação e perspectivas sobre as águas do cerrado. **Ciência e Cultura**, v. 63, n. 3, p. 27–29, 2011.

LIMA, M. G. B.; PERSSON, U. M. Commodity-Centric Landscape Governance as a Double-Edged Sword: The Case of Soy and the Cerrado Working Group in Brazil. **Frontiers in Forests and Global Change**, v. 3, 2020. Disponível em: https://doi.org/10.3389/ffgc.2020.00027. Acesso em: 19 mar. 2020.

LIU, X.; REN, L.; YUAN, F.; SINGH, V. P.; FANG, X.; YU, Z.; ZHANG, W. Quantifying the effect of land use and land cover changes on green water and blue water in northern part of China. **Hydrology and Earth System Sciences**, v. 13, n. 6, p. 735–747, 2009. Disponível em: https://doi.org/10.5194/hess-13-735-2009

LOPES, T. R.; ZOLIN, C. A.; PRADO, G. do; PAULINO, J.; ALMEIDA, F. T. de; LOPES, T. R.; ZOLIN, C. A.; PRADO, G. do; PAULINO, J.; ALMEIDA, F. T. de. Regionalization of maximum and minimum flow in the Teles Pires basin, Brazil. **Engenharia Agrícola**, v. 37, n. 1, p. 54–63, 2017. Disponível em: https://doi.org/10.1590/1809-4430-eng.agric.v37n1p54-63/2017

MAKSYMIUK, S.; GOSIEWSKA, A.; BIECEK, P. Landscape of R packages for eXplainable Artificial Intelligence. **arXiv:2009.13248 [cs, stat]**, 2020. Disponível em: http://arxiv.org/abs/2009.13248. Acesso em: 12 nov. 2020.

MALVEIRA, V. T. C.; ARAÚJO, J. C. de; GÜNTNER, A. Hydrological Impact of a High-Density Reservoir Network in Semiarid Northeastern Brazil. **Journal of Hydrologic Engineering**, v. 17, n. 1, p. 109–117, 2012. Disponível em: https://doi.org/10.1061/(ASCE)HE.1943-5584.0000404

MAPBIOMAS. **Projeto de Mapeamento Anual da Cobertura e Uso do Solo do Brasil [In English: Brazilian Annual Land Use and Land Cover Mapping Project]**. *[s. l.]*, 2020. Disponível em: http://mapbiomas.org/.

MARQUES, E. A. G.; SILVA JUNIOR, G. C.; EGER, G. Z. S.; ILAMBWETSI, A. M.; RAPHAEL, P.; GENEROSO, T. N.; OLIVEIRA, J.; JÚNIOR, J. N. Analysis of groundwater and river stage fluctuations and their relationship with water use and climate variation effects on Alto Grande watershed, Northeastern Brazil. **Journal of South American Earth Sciences**, v. 103, p. 102723, 2020. Disponível em: https://doi.org/10.1016/j.jsames.2020.102723

MCMILLAN, H. K. Linking hydrologic signatures to hydrologic processes: A review. **Hydrological Processes**, v. 34, n. 6, p. 1393–1409, 2020. Disponível em: https://doi.org/10.1002/hyp.13632

MORAIS, M. A. V.; VIOLA, M. R.; MELLO, C. R. de; RODRIGUES, J. A. M.; OLIVEIRA, V. A. de. Regionalization of reference streamflows for the Araguaia River basin in Brazil. **Semina: Ciências Agrárias**, v. 41, n. 3, p. 829, 2020. Disponível em: https://doi.org/10.5433/1679-0359.2020v41n3p829

MUELLER, N. D.; BUTLER, E. E.; MCKINNON, K. A.; RHINES, A.; TINGLEY, M.; HOLBROOK, N. M.; HUYBERS, P. Cooling of US Midwest summer temperature extremes from cropland intensification. **Nature Climate Change**, v. 6, n. 3, p. 317–322, 2016. Disponível em: https://doi.org/10.1038/nclimate2825

NASH, J. E.; SUTCLIFFE, J. V. River flow forecasting through conceptual models part I — A discussion of principles. **Journal of Hydrology**, v. 10, n. 3, p. 282–290, 1970. Disponível em: https://doi.org/10.1016/0022-1694(70)90255-6

NASSIF, S. H.; WILSON, E. M. The influence of slope and rain intensity on runoff and infiltration. **Hydrological Sciences Bulletin**, v. 20, n. 4, p. 539–553, 1975. Disponível em: https://doi.org/10.1080/02626667509491586

NAVAL, P. **mopsocd: Multi-objective Particle Swarm Optimization with Crowding Distance**. Versão 0.5.1. *[S. l.: s. n.]*. *E-book*. Disponível em: https://CRAN.R-project.org/package=mopsocd. Acesso em: 10 dez. 2020.

NOCCO, M. A.; SMAIL, R. A.; KUCHARIK, C. J. Observation of irrigation-induced climate change in the Midwest United States. **Global Change Biology**, v. 25, n. 10, p. 3472–3484, 2019. Disponível em: https://doi.org/10.1111/gcb.14725

OLIVEIRA, V. A. de; MELLO, C. R. de; VIOLA, M. R.; SRINIVASAN, R. Assessment of climate change impacts on streamflow and hydropower potential in the headwater region of the Grande river basin, Southeastern Brazil. **International Journal of Climatology**, v. 37, n. 15, p. 5005–5023, 2017. Disponível em: https://doi.org/10.1002/joc.5138

OLIVEIRA, P. T. S.; NEARING, M. A.; HAWKINS, R. H.; STONE, J. J.; RODRIGUES, D. B. B.; PANACHUKI, E.; WENDLAND, E. Curve number estimation from Brazilian Cerrado rainfall and runoff data. **Journal of Soil and Water Conservation**, v. 71, n. 5, p. 420–429, 2016. Disponível em: https://doi.org/10.2489/jswc.71.5.420

OLIVEIRA, P. T. S.; NEARING, M. A.; MORAN, M. S.; GOODRICH, D. C.; WENDLAND, E.; GUPTA, H. V. Trends in water balance components across the Brazilian Cerrado. **Water Resources Research**, v. 50, n. 9, p. 7100–7114, 2014. Disponível em: https://doi.org/10.1002/2013WR015202

PAGLIERO, L.; BOURAOUI, F.; DIELS, J.; WILLEMS, P.; MCINTYRE, N. Investigating regionalization techniques for large-scale hydrological modelling. **Journal of Hydrology**, v. 570, p. 220–235, 2019. Disponível em: https://doi.org/10.1016/j.jhydrol.2018.12.071

PARSOPOULOS, K. E.; VRAHATIS, M. N. Recent approaches to global optimization problems through Particle Swarm Optimization. **Natural Computing**, v. 1, n. 2, p. 235–306, 2002. Disponível em: https://doi.org/10.1023/A:1016568309421

PATLE, G. T.; SIKAR, T. T.; RAWAT, K. S.; SINGH, S. K. Estimation of infiltration rate from soil properties using regression model for cultivated land. **Geology, Ecology, and Landscapes**, v. 3, n. 1, p. 1–13, 2019. Disponível em: https://doi.org/10.1080/24749508.2018.1481633

PEREIRA, S. B.; PRUSKI, F. F.; SILVA, D. D.; RAMOS, M. M. Evaporação líquida no lago de Sobradinho e impactos no escoamento devido à construção do reservatório. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v. 13, n. 3, p. 346–352, 2009. Disponível em: https://doi.org/10.1590/S1415-43662009000300018

PERRIN, C.; MICHEL, C.; ANDRÉASSIAN, V. Improvement of a parsimonious model for streamflow simulation. **Journal of Hydrology**, v. 279, n. 1, p. 275–289, 2003. Disponível em: https://doi.org/10.1016/S0022-1694(03)00225-7

PFAFSTETTER, O. Classification of hydrographic basins: coding methodology. **unpublished manuscript, Departamento Nacional de Obras de Saneamento,** v. 18, n. 1989, p. 1–2, 1989.

PINHATI, F. S. C.; RODRIGUES, L. N.; AIRES DE SOUZA, S. Modelling the impact of on-farm reservoirs on dry season water availability in an agricultural catchment area of the Brazilian savannah. **Agricultural Water Management**, v. 241, p. 106296, 2020. Disponível em: https://doi.org/10.1016/j.agwat.2020.106296

PIRES, G. F.; ABRAHÃO, G. M.; BRUMATTI, L. M.; OLIVEIRA, L. J.; COSTA, M. H.; LIDDICOAT, S.; KATO, E.; LADLE, R. J. Increased climate risk in Brazilian double cropping agriculture systems: Implications for land use in Northern Brazil. **Agricultural and Forest Meteorology**, v. 228, p. 286–298, 2016. Disponível em: https://doi.org/10.1016/j.agrformet.2016.07.005

POOL, S.; VIVIROLI, D.; SEIBERT, J. Value of a Limited Number of Discharge Observations for Improving Regionalization: A Large-Sample Study Across the United States. **Water Resources Research**, v. 55, n. 1, p. 363–377, 2019. Disponível em: https://doi.org/10.1029/2018WR023855

POUSA, R.; COSTA, M. H.; PIMENTA, F. M.; FONTES, V. C.; BRITO, V. F. A. de; CASTRO, M. Climate change and intense irrigation growth in western Bahia, Brazil: The urgent need for hydroclimatic monitoring. **Water**, v. 11, n. 5, p. 933, 2019. Disponível em: https://doi.org/10.3390/w11050933

PRUSKI, F. F.; NUNES, A. de A.; PRUSKI, P. L.; RODRIGUEZ, R. del G. Improved regionalization of streamflow by use of the streamflow equivalent of precipitation as an explanatory variable. **Journal of Hydrology**, v. 476, p. 52–71, 2013. Disponível em: https://doi.org/10.1016/j.jhydrol.2012.10.005

PUSHPALATHA, R.; PERRIN, C.; LE MOINE, N.; MATHEVET, T.; ANDRÉASSIAN, V. A downward structural sensitivity analysis of hydrological models to improve low-flow simulation. **Journal of Hydrology**, v. 411, n. 1–2, p. 66–76, 2011. Disponível em: https://doi.org/10.1016/j.jhydrol.2011.09.034

QI, W.; CHEN, J.; LI, L.; XU, C.; LI, J.; XIANG, Y.; ZHANG, S. **A framework to regionalize conceptual model parameters for global hydrological modeling**. *[S. l.]*: Global hydrology/Modelling approaches, 2020. preprint. Disponível em: https://doi.org/10.5194/hess-2020-127. Acesso em: 22 dez. 2020.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria: R Foundation for Statistical Computing, 2020. *E-book*. Disponível em: https://www.R-project.org/

RADA, N. Assessing Brazil's Cerrado agricultural miracle. **Food Policy**, v. 38, n. Supplement C, p. 146–155, 2013. Disponível em: https://doi.org/10.1016/j.foodpol.2012.11.002

RAQUEL, C. R.; NAVAL, P. C. An effective use of crowding distance in multiobjective particle swarm optimization. *In*: 2005, New York, NY, USA. **Proceedings of the 7th annual conference on Genetic and evolutionary computation**. New York, NY, USA: Association for Computing Machinery, 2005. p. 257–264. Disponível em: https://doi.org/10.1145/1068009.1068047. Acesso em: 10 dez. 2020.

RAZAVI, T.; COULIBALY, P. Streamflow Prediction in Ungauged Basins: Review of Regionalization Methods. **Journal of Hydrologic Engineering**, v. 18, n. 8, p. 958–975, 2013 a. Disponível em: https://doi.org/10.1061/(ASCE)HE.1943-5584.0000690

RAZAVI, T.; COULIBALY, P. Streamflow Prediction in Ungauged Basins: Review of Regionalization Methods. **Journal of Hydrologic Engineering**, v. 18, n. 8, p. 958–975, 2013 b. Disponível em: https://doi.org/10.1061/(ASCE)HE.1943-5584.0000690

RIBEIRO, J. F.; WALTER, B. M. T. Fitofisionomias do bioma Cerrado. *In*: **SANO, S. M.; ALMEIDA, S. P. de (Ed.). Cerrado: ambiente e flora**. Planaltina: Embrapa-CPAC, 1998. p. 87–167.

RODRIGUES, J. A. M.; VIOLA, M. R.; ALVARENGA, L. A.; MELLO, C. R. de; CHOU, S. C.; OLIVEIRA, V. A. de; UDDAMERI, V.; MORAIS, M. A. V. Climate change impacts under representative concentration pathway scenarios on streamflow and droughts of basins in the Brazilian Cerrado biome. **International Journal of Climatology**, v. 40, n. 5, p. 2511–2526, 2020. Disponível em: https://doi.org/10.1002/joc.6347

RODRIGUES, L. N. Plataforma tecnológica para o gerenciamento compartilhado de água. **Empresa Brasileira de Pesquisa Agropecuária**, 2018. Disponível em: https://www.embrapa.br/busca-de-noticias/-/noticia/32706192/artigo---plataforma-tecnologica-para-o-gerenciamento-compartilhado-de-agua

RODRIGUES, L. N.; SANO, E. E.; STEENHUIS, T. S.; PASSO, D. P. Estimation of Small Reservoir Storage Capacities with Remote Sensing in the Brazilian Savannah Region. **Water Resources Management**, v. 26, n. 4, p. 873–882, 2012. Disponível em: https://doi.org/10.1007/s11269-011-9941-8

RODRIGUES, L. N.; SILVA, L. M. C.; FREITAS, M. A. S. Reservação: Planejamento e gerenciamento da água com vistas à redução de conflitos. **ITEM. Irrigação e Tecnologia Moderna**, v. 100, p. 34–38, 2014.

ROSA, L.; CHIARELLI, D. D.; RULLI, M. C.; DELL'ANGELO, J.; D'ODORICO, P. Global agricultural economic water scarcity. **Science Advances**, v. 6, n. 18, p. eaaz6031, 2020. Disponível em: https://doi.org/10.1126/sciadv.aaz6031

SCHAEFLI, B.; GUPTA, H. V. Do Nash values have value? **Hydrological Processes**, v. 21, n. 15, p. 2075–2080, 2007. Disponível em: https://doi.org/10.1002/hyp.6825

SILVA, V. de P. R. da; SILVA, M. T.; SINGH, V. P.; DE SOUZA, E. P.; BRAGA, C. C.; DE HOLANDA, R. M.; ALMEIDA, R. S. R.; DE ASSIS SALVIANO DE SOUSA, F.; BRAGA, A. C. R. Simulation of stream flow and hydrological response to land-cover changes in a tropical river basin. **CATENA**, v. 162, p. 166–176, 2018. Disponível em: https://doi.org/10.1016/j.catena.2017.11.024

SILVA, E. M.; AZEVEDO, J. A. de; RAUBER, J. C.; REATTO, A. **Caracterização físico-hídrica e hidráulica de solos do bioma Cerrado submetidos a diferentes sistemas de preparo.**: Boletim de pesquisa e desenvolvimento 101. Planaltina, DF: Empresa Brasileira de Pesquisa Agropecuária / Embrapa Cerrados, 2003.

SIVAPALAN, M. Prediction in ungauged basins: a grand challenge for theoretical hydrology. **Hydrological Processes**, v. 17, n. 15, p. 3163–3170, 2003. Disponível em: https://doi.org/10.1002/hyp.5155

SOUZA, C. M. *et al.* Reconstructing Three Decades of Land Use and Land Cover Changes in Brazilian Biomes with Landsat Archive and Earth Engine. **Remote Sensing**, v. 12, n. 17, p. 2735, 2020. Disponível em: https://doi.org/10.3390/rs12172735

SPANGLER, K. R.; LYNCH, A. H.; SPERA, S. A. Precipitation drivers of cropping frequency in the Brazilian Cerrado: Evidence and implications for decision-making.

Weather, Climate, and Society, v. 9, n. 2, p. 201–213, 2017. Disponível em: https://doi.org/10.1175/WCAS-D-16-0024.1

SPERA, S. Agricultural Intensification Can Preserve the Brazilian Cerrado: Applying Lessons From Mato Grosso and Goiás to Brazil's Last Agricultural Frontier. **Tropical Conservation Science**, v. 10, p. 1940082917720662, 2017. Disponível em: https://doi.org/10.1177/1940082917720662

SRIVASTAVA, N.; HINTON, G.; KRIZHEVSKY, A.; SUTSKEVER, I.; SALAKHUTDINOV, R. Dropout: a simple way to prevent neural networks from overfitting. **The Journal of Machine Learning Research**, v. 15, n. 1, p. 1929–1958, 2014.

SRIWONGSITANON, N.; TAESOMBAT, W. Effects of land cover on runoff coefficient. **Journal of Hydrology**, v. 410, n. 3, p. 226–238, 2011. Disponível em: https://doi.org/10.1016/j.jhydrol.2011.09.021

STEENHUIS, T. S.; COLLICK, A. S.; EASTON, Z. M.; LEGGESSE, E. S.; BAYABIL, H. K.; WHITE, E. D.; AWULACHEW, S. B.; ADGO, E.; AHMED, A. A. Predicting discharge and sediment for the Abay (Blue Nile) with a simple model. **Hydrological Processes**, v. 23, n. 26, p. 3728–3737, 2009. Disponível em: https://doi.org/10.1002/hyp.7513

SUNDARARAJAN, M.; TALY, A.; YAN, Q. Axiomatic Attribution for Deep Networks. **arXiv:1703.01365 [cs]**, 2017. Disponível em: http://arxiv.org/abs/1703.01365. Acesso em: 2 jan. 2021.

TARBOTON, D. G. **Terrain analysis using digital elevation models (TauDEM)**. *[s. l.]*, 2008. Disponível em: http://hydrology.usu.edu/taudem/.

TRAN, Q. Q.; NIEL, J. D.; WILLEMS, P. Spatially Distributed Conceptual Hydrological Model Building: A Generic Top-Down Approach Starting From Lumped Models. **Water Resources Research**, v. 54, n. 10, p. 8064–8085, 2018. Disponível em: https://doi.org/10.1029/2018WR023566

WANG, Q.; XU, Y.; XU, Y.; WU, L.; WANG, Y.; HAN, L. Spatial hydrological responses to land use and land cover changes in a typical catchment of the Yangtze River Delta region. **CATENA**, v. 170, p. 305–315, 2018. Disponível em: https://doi.org/10.1016/j.catena.2018.06.022

WESTERBERG, I. K.; MCMILLAN, H. K. Uncertainty in hydrological signatures. **Hydrology and Earth System Sciences**, v. 19, n. 9, p. 3951–3968, 2015. Disponível em: https://doi.org/10.5194/hess-19-3951-2015

WICKHAM, H.; FRANÇOIS, R.; HENRY, L.; MÜLLER, K.; RSTUDIO. **dplyr: A Grammar of Data Manipulation**. Versão 1.0.2. *[S. l.: s. n.]. E-book*. Disponível em: https://CRAN.R-project.org/package=dplyr. Acesso em: 4 jan. 2021.

WOODS, R. A. Analytical model of seasonal climate impacts on snow hydrology: Continuous snowpacks. **Advances in Water Resources**, v. 32, n. 10, p. 1465–1481, 2009. Disponível em: https://doi.org/10.1016/j.advwatres.2009.06.011

YOUSSEF, Y. W.; KHODZINSKAYA, A. A Review of Evaporation Reduction Methods from Water Surfaces. **E3S Web of Conferences**, v. 97, p. 05044, 2019. Disponível em: https://doi.org/10.1051/e3sconf/20199705044

ZHOU, G.; WEI, X.; CHEN, X.; ZHOU, P.; LIU, X.; XIAO, Y.; SUN, G.; SCOTT, D. F.; ZHOU, S.; HAN, L.; SU, Y. Global pattern for the effect of climate and land cover on water yield. **Nature Communications**, v. 6, n. 1, p. 5918, 2015. Disponível em: https://doi.org/10.1038/ncomms6918