

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Escola de Engenharia
Programa de Pós-Graduação em Saneamento, Meio Ambiente e Recursos Hídricos

Gustavo de Oliveira Corrêa

**ESTRUTURA DE MODELOS HIDROLÓGICOS E SUA INTER-RELAÇÃO COM
ATRIBUTOS FÍSICOS E ASSINATURAS HIDROLÓGICAS**

Belo Horizonte
2023

Gustavo de Oliveira Corrêa

**ESTRUTURA DE MODELOS HIDROLÓGICOS E SUA INTER-RELAÇÃO COM
ATRIBUTOS FÍSICOS E ASSINATURAS HIDROLÓGICAS**

Dissertação apresentada ao Programa de Pós-graduação em Saneamento, Meio Ambiente e Recursos Hídricos da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Saneamento, Meio Ambiente e Recursos Hídricos.

Área de concentração: Hidráulica e Recursos Hídricos

Linha de pesquisa: Modelagem de Processos Hidrológicos

Orientador: Francisco Eustáquio Oliveira e Silva

Belo Horizonte
2023

C824e

Corrêa, Gustavo de Oliveira.

Estrutura de modelos hidrológicos e sua inter-relação com atributos físicos e assinaturas hidrológicas [recurso eletrônico] Gustavo de Oliveira Corrêa. – 2023.

1 recurso online (170 f.: il., color.) : pdf.

Orientador: Francisco Eustáquio Oliveira e Silva.

Dissertação (mestrado) - Universidade Federal de Minas Gerais, Escola de Engenharia.

Bibliografia: f. 148-170.

Exigências do sistema: Adobe Acrobat Reader.

1. Engenharia sanitária - Teses. 2. Recursos hídricos - Desenvolvimento - Teses. 3. Hidrologia - Modelos - Teses. 4. Aprendizado do computador - Teses. 5. Hidrologia - Teses. I. Silva, Francisco Eustáquio Oliveira e. II. Universidade Federal de Minas Gerais. Escola de Engenharia. III. Título.

CDU: 628(043)



UNIVERSIDADE FEDERAL DE MINAS GERAIS
[ESCOLA DE ENGENHARIA]
COLEGIADO DO CURSO DE GRADUAÇÃO / PÓS-GRADUAÇÃO EM [SANEAMENTO, MEIO AMBIENTE E RECURSOS
HÍDRICOS]

FOLHA DE APROVAÇÃO

"Estrutura de Modelos Hidrológicos e Sua Inter-relação Com Atributos Físicos e Assinaturas Hidrológicas"

GUSTAVO DE OLIVEIRA CORREA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

Prof. Francisco Eustáquio Oliveira e Silva

Prof. Veber Afonso Figueiredo Costa

Prof. Dirceu Silveira Reis Júnior

Prof. Pedro Luiz Borges Chaffe

Aprovada pelo Colegiado do PG SMARH

Versão Final aprovada por

Profa. Priscilla Macedo Moura

Prof. Francisco Eustáquio Oliveira e Silva

Belo Horizonte, 18 de agosto de 2023.



Documento assinado eletronicamente por **Francisco Eustaquio Oliveira e Silva, Professor do Magistério Superior**, em 18/08/2023, às 11:34, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Pedro Luiz Borges Chaffe, Usuário Externo**, em 18/08/2023, às 11:34, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Dirceu Silveira Reis Junior, Usuário Externo**, em 18/08/2023, às 11:35, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Veber Afonso Figueiredo Costa, Professor do Magistério Superior**, em 18/08/2023, às 11:35, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2548765** e o código CRC **627FED79**.

*À memória de minha querida avó Geralda,
cuja sabedoria e luz continuam a me guiar.*

AGRADECIMENTOS

Preciso agradecer a Deus ou a qualquer movimento maior que tenha contribuído na finalização deste estudo. Com certeza foi necessário.

Aos meus pais, agradeço por serem minha base sólida, por todo o amor, ensinamentos e apoio incondicional que sempre me deram. Agradeço a toda a minha família pelo incentivo constante, pelas palavras motivadoras e pelo aconchego que sempre encontro.

Agradeço aos meus amigos, que, mesmo nas horas mais atribuladas, me trouxeram alegria, descontração e fizeram a jornada mais leve e enriquecedora. Agradeço também a compreensão (ou não) sobre minha reclusão nos últimos meses.

A meu orientador expressei um profundo e sincero agradecimento. Seu direcionamento certeiro iluminou os nebulosos caminhos da pesquisa. Sua paciência e disponibilidade inabaláveis foram fundamentais nos momentos mais árduos, guiando-me com sabedoria e serenidade.

Finalmente, agradeço a todos meus amigos da equipe de recursos hídricos da Tetra Tech. A troca de experiências e o ambiente colaborativo em que estamos imersos foram cruciais para o meu redirecionamento para a Hidrologia. Em especial: Marcão, Vivi e Osmar.

“Escolhemos pela emoção e justificamos pela razão.”

Uma adaptação dos Pensamentos de Blaise Pascal

RESUMO

Há tempos a literatura hidrológica busca relações entre as assinaturas hidrológicas, atributos físicos das bacias e os modelos hidrológicos. A compreensão dessa inter-relação é fundamental para compreender e representar o comportamento das bacias hidrográficas, especialmente em face das limitações associadas ao monitoramento disponível. Dentre as diversas classificações de modelos hidrológicos, uma delas classifica-os em baseados em princípios físicos e orientados a dados. No meio deste cenário, cresce a discussão na comparação destas abordagens e o quanto as assinaturas hidrológicas e as características físicas das bacias impactam na eficácia dos modelos, utilizando amplas bases de dados hidrometeorológicas públicas. Neste aspecto, as redes neurais recorrentes do tipo LSTM têm ganhado notoriedade nos estudos hidrológicos pelo seu desempenho frente aos modelos conceituais. Devido a diversas opções de modelos conceituais, alguns estudos buscam, por meio de *frameworks* multimodelos, identificar e analisar a adequabilidade da estrutura do modelo conceitual às diferentes bacias, assumindo que os modelos hidrológicos estão intrinsecamente conectados às propriedades físicas das bacias, dadas as complexas interações entre clima, solos, vegetação e topografia. Algo ainda não explorado é comparar as redes LSTM frente a melhor estrutura de modelo conceitual calibrada individualmente para cada bacia.

Este trabalho tem como objetivo identificar e analisar a inter-relação entre as assinaturas hidrológicas, as características físicas das bacias hidrográficas e a eficácia das diferentes estruturas de modelos hidrológicos, seja baseado em processo físicos ou orientados a dados. Para tal, utiliza as redes LSTM e, numa extensão aos estudos consultados, faz uso de um *framework* de estruturas conceituais flexíveis (FUSE) para avaliação dessa adequabilidade. Uma extensa base de bacias pública foi utilizada, avaliando desde a calibração multiobjetivo de modelos conceituais utilizando assinaturas hidrológicas, a estimativa de assinaturas pelas características físicas, a caracterização de grupos homogêneos, e a relação das assinaturas, atributos físicos e os modelos, seja conceitual ou orientado a dados.

O estudo reitera as conclusões de outros pesquisadores ao indicar que a calibração apoiada nas assinaturas hidrológicas é adequada e estende a interpretabilidade física dos modelos conceituais. Os resultados também indicam que as características físicas podem ser empregadas para inferir o comportamento das assinaturas hidrológicas em bacias sem monitoramento hidrométrico e propõe cinco agrupamentos por similaridade das métricas hidrológicas, com consistência espacial. Os resultados reforçam a dificuldade de se estabelecer uma relação entre as assinaturas hidrológicas, características físicas e a estrutura do modelo conceitual, mas obtêm uma série de regras de associações, capazes de fundamentar a seleção de estruturas apropriadas a representação dos fenômenos. Por fim, o estudo destaca a utilização de assinaturas hidrológicas como entrada para as redes LSTM, validando a hipótese que a utilização de assinaturas hidrológicas como camada estática, no lugar dos atributos físicos, permite o alcance de resultados semelhantes ou melhores, com vantagens já demonstradas em estudos fundamentados na análise de modelos baseados em processos físicos.

Palavras-chave: modelagem hidrológica; fuse; aprendizado de máquina; lstm.

ABSTRACT

For some time, the hydrological literature has been seeking a relationship between hydrological signatures, physical attributes of basins, and hydrological models. Understanding this interrelation is fundamental to understanding and representing the behavior of the basins, especially given the limitations associated with available monitoring. Among the various classifications of hydrological models, one classifies them based on physical principles and data-driven approaches. Against this backdrop, there is growing discussion about comparing these approaches and how much the hydrological signatures and physical characteristics of the basins impact on the effectiveness of the models, using extensive public hydrometeorological databases. In this regard, Long Short-Term Memory (LSTM) recurrent neural networks have gained prominence in hydrological studies due to their performance compared to conceptual models. Given the myriad options for conceptual models, some studies aim, through multi-model frameworks, to identify and analyze the appropriateness of the conceptual model structure to different watersheds, assuming that hydrological models are intrinsically connected to the physical properties of watersheds given the complex interactions between climate, soils, vegetation, and topography. Something that has not yet been explored is comparing LSTM networks against the best conceptual model structure calibrated individually for each basin.

This research aims to identify and analyze the interrelation between hydrological signatures, the physical characteristics of the basins and the effectiveness of different hydrological model structures, whether physical process-based or data-driven. To this end, it employs LSTM networks and, extending beyond the referenced studies, uses a Flexible Model Structure (FUSE) framework to evaluate this suitability. An extensive database of public basins was used, evaluating the multi-object calibration of conceptual models using hydrological signatures, the estimation of signatures by physical characteristics, the characterization of homogeneous groups, and the relationship between signatures, physical attributes and models, whether conceptual or data-driven.

The study reiterates the conclusions of other researchers by indicating that calibration based on hydrological signatures is appropriate and extends the physical interpretability of conceptual models. The results also indicate that physical characteristics can be used to infer the behavior of hydrological signatures in basins without hydrometric monitoring and propose five groupings by similarity of hydrological metrics, with spatial consistency. The results reinforce the difficulty of establishing a relationship between hydrological signatures, physical characteristics and the structure of the conceptual model, but provide a series of association rules capable of supporting the selection of appropriate structures to represent the phenomena. Finally, the study highlights the use of hydrological signatures as input to LSTM networks, validating the hypothesis that the use of hydrological signatures as a static layer, instead of physical attributes, allows similar or better results to be achieved, with advantages already demonstrated in studies based on the analysis of models based on physical processes.

Keywords: hydrologic modelling; fuse; machine learning; lstm.

LISTA DE FIGURAS

Figura 1 – Adaptações do SCE-UA para a abordagem multiobjetivo.....	35
Figura 2 – Distância da multidão.....	36
Figura 3 – Pontos de referência no caso de 3 objetivos.....	37
Figura 4 – Funções de pontuação.....	39
Figura 5 – Bacias utilizadas.....	41
Figura 6 – Número de estruturas hidrológicamente consistentes encontradas na calibração.....	42
Figura 7 – Número de estruturas hidrológicamente por função de agregação.....	44
Figura 8 – Soma das estruturas hidrológicamente consistentes em todas as bacias para cada algoritmo.....	45
Figura 9 – Soma das estruturas hidrológicamente consistentes e com NSE acima de 0,7 em todas as bacias para cada algoritmo.....	46
Figura 10 – Histograma da métrica de desempenho para as soluções hidrológicamente consistentes calibradas somente no domínio das assinaturas.....	47
Figura 11 – Média das regressões para todas as assinaturas.....	66
Figura 12 – Média dos coeficientes de determinação para os percentis da curva de permanência.....	67
Figura 13 – Maiores coeficientes de determinação obtidos na predição das assinaturas hidrológicas.....	68
Figura 14 – Valores do SHAP para o escoamento de base na estação ANA 13880000.....	69
Figura 15 – Valores do SHAP para a taxa de mudança na estação ANA 13880000.....	69
Figura 16 – Valores do SHAP para as principais características para a predição do índice do escoamento de base, Q75, coeficiente do escoamento superficial e taxa de subida.....	70
Figura 17 – Principais atributos na predição de todas as assinaturas.....	71
Figura 18 – Correlação entre as assinaturas hidrológicas utilizadas no agrupamento.....	72
Figura 19 – Proporção da variância total explicada pelas componentes principais identificadas.....	72

Figura 20 – Importância relativa das assinaturas hidrológicas por componente identificada.	73
Figura 21 – Análise do número ótimo <i>clusters</i> a partir do coeficiente de Silhueta. ...	73
Figura 22 – Análise do número ótimo de <i>clusters</i> a partir do Método de Elbow.....	74
Figura 23 – Análise do número ótimo de <i>clusters</i> a partir do Índice Calinski-Harabasz.	74
Figura 24 – Agrupamentos em relação as componentes principais.	75
Figura 25 – Representação da importância média das assinaturas hidrológicas em cada <i>cluster</i>	75
Figura 26 – Gráfico de coordenadas paralelas das assinaturas hidrológicas de cada bacia coloridas por cluster.	76
Figura 27 – Boxplot com as assinaturas hidrológicas em cada agrupamento.	77
Figura 28 – Agrupamentos obtidos em relação aos biomas brasileiros.	78
Figura 29 – Agrupamentos obtidos em relação as zonas climáticas brasileiras.....	80
Figura 30 – Agrupamentos obtidos em relação as zonas de umidade.	80
Figura 31 – Agrupamentos obtidos em relação a krigagem dos dados de chuva disponíveis no Hidroweb entre os anos hidrológicos de 1980 e 2010.	81
Figura 32 – Agrupamentos obtidos em relação aos principais aquíferos.	81
Figura 33 - Diagrama de representação dos fluxos em uma bacia	87
Figura 34 – Diagramas simplificados dos quatro modelos principais que se baseia o FUSE.....	89
Figura 35 - Bacias utilizadas neste estudo	93
Figura 36 – Frequência relativa dos mecanismos nas estruturas selecionadas.....	96
Figura 37 – Mapa dos mecanismos utilizados para modelar a camada superior do solo	98
Figura 38 – Regras de associação obtidas para a camada superior do solo	99
Figura 39 – Mapa dos mecanismos utilizados para modelar a camada inferior do solo	100
Figura 40 – Regras de associação obtidas para a camada inferior do solo	103
Figura 41 – Mapa dos mecanismos utilizados para modelar o escoamento superficial.	105
Figura 42 – Regras de associação obtidas para o escoamento superficial.....	106
Figura 43 – Mapa dos mecanismos utilizados para modelar a percolação	109
Figura 44 – Regras de associação obtidas para a percolação.....	110

Figura 45 – Mapa dos mecanismos utilizados para modelar a evapotranspiração .	111
Figura 46 – Associações que levaram aos mecanismos de Evapotranspiração	113
Figura 47 – Mapa dos mecanismos utilizados para modelar o fluxo intermediário .	114
Figura 48 – Associações que levaram ao mecanismo de fluxo intermediário	115
Figura 49 – Mapa dos mecanismos utilizados para modelar a propagação.....	116
Figura 50 – Associações que levaram ao mecanismo de propagação	117
Figura 51 – Mapa das bacias não calibradas,	118
Figura 52 – Associações que levaram a bacias que não foram representadas em nenhum modelo.....	119
Figura 53 – Principais associações sumariadas.....	120
Figura 54 – Associação dos mecanismos em relação a declividade.....	120
Figura 55 – Comparação dos agrupamentos obtidos com os mecanismos e os por similaridade de assinaturas hidrológicas	122
Figura 56 – Visualização de uma camada de rede LSTM.....	132
Figura 57 – Função de densidade de probabilidade não paramétrica do NSE	136
Figura 58 – Frequência cumulativa do NSE	136
Figura 59 – Sensibilidade dos resultados das redes LSTM em função do número de passos de tempo anteriores fornecidos e o número de neurônios	137
Figura 60 – NSE de cada bacia no FUSE e na rede LSTM com assinaturas como camada estática	138
Figura 61 – Comparação entre o NSE obtido no melhor modelo do FUSE e a rede LSTM com assinaturas hidrológicas como camada estática.....	139
Figura 62 – Distribuição espacial da diferença relativa entre as Q90 simuladas e observadas para as redes LSTM e a melhor estrutura do FUSE	139

LISTA DE TABELAS

Tabela 1 – Configurações testadas.....	40
Tabela 2 – Características das bacias utilizadas.....	41
Tabela 3 – Número de estruturas por cada algoritmo de otimização em cada bacia.	42
Tabela 4 – Assinaturas hidrológicas utilizadas.....	61
Tabela 5 – Processos hidrológicos e mecanismos no FUSE	90
Tabela 6 – Associações relevantes obtidas como consequência a camada superior do solo.....	97
Tabela 7 – Associações relevantes obtidas como consequência a camada inferior do solo.....	102
Tabela 8 – Associações obtidas como consequência o escoamento superficial.....	107
Tabela 9 – Associações obtidas como consequência a percolação.....	107
Tabela 10 – Associações obtidas como consequência a evapotranspiração.....	112
Tabela 11 – Associações obtidas como consequência o fluxo intermediário	115
Tabela 12 – Associações obtidas como consequência a propagação	117
Tabela 13 - Entradas da rede LSTM	134
Tabela 14 – Desempenho das redes treinadas em agrupamentos	140
Tabela 15 – Variação do NSE para as arquiteturas otimizadas das redes LSTM e do melhor modelo conceitual em cada bacia no FUSE	140

LISTA DE ABREVIATURAS E SIGLAS

ABC - *Artificial Bee Colony*

ANA - Agência Nacional de Água

BC - Box-Cox

CABra - *Catchment Attributes for Brazil*

CAMELS - *Catchment attributes and meteorology for large-sample studies*

CatBoost - *Categorical Boosting*

DE – Differential Evolution

DL - *Deep Learning*

FUSE - *Framework for Understanding Structural Errors*

GA – *Genetic Algorithms*

IBGE - Instituto Brasileiro de Geografia e Estatística

KGE - Coeficiente de Kling-Gupta *Efficiency*

LightGBM - *Light Gradient Boosting Machine*

LSTM - *Long Short-Term Memory*

ML - *Machine Learning*

MOCOM - *Multi-Objective Complex Evolution*

MOSCDE - *Multi-Objective Shuffled Complex Differential Evolution*

MPSO - *Multiobjective Particle Swarm Optimization*

NSE Nash-Sutcliffe *efficiency*

PCA - *Principal Components Analysis*

PSO - *Particle Swarm Optimization*

RF - *Random Forest*

RNAs - Redes Neurais Artificiais

RNN - *Recurrent Neural Network*

SCE-UA - *Shuffled Complex Evolution - University of Arizona*

SVM - *Support Vector Machine*

SVR - *Support Vector Regression*

XGB - *eXtreme Gradient Boosting*

SUMÁRIO

1	INTRODUÇÃO	17
2	OBJETIVOS	20
2.1	Objetivo Geral	20
2.2	Objetivos Específicos	20
3	CALIBRAÇÃO MULTIOBJETIVO UTILIZANDO ASSINATURAS HIDROLÓGICAS	21
3.1	Introdução	21
3.2	Desenvolvimento teórico	24
3.2.1	Calibração multiobjetivo	24
3.2.2	NSGA-II e III	27
3.2.3	SCE-UA	28
3.3	Metodologia	29
3.3.1	Métricas	31
3.3.2	SCE Multiobjetivo	32
3.3.3	Ordenação da Frente de Pareto	35
3.3.4	Funções de agregação	38
3.3.5	Área de estudo	40
3.4	Resultados	41
3.5	Discussão	47
3.6	Conclusões	50
4	PREDIÇÃO DE ASSINATURAS HIDROLÓGICAS E CLASSIFICAÇÃO DE BACIAS TROPICAIS	52
4.1	Introdução	52
4.2	Referencial teórico	55
4.2.1	Regressão	56
4.2.2	Agrupamentos	58
4.3	Metodologia	59
4.3.1	Experimento 1	60
4.3.2	Experimento 2	63
4.4	Resultados e discussão	65
4.4.1	Estimativa das assinaturas hidrológicas a partir de atributos físicos via técnicas de aprendizado de máquina	65
4.4.2	Inferência de agrupamentos de um conjunto de assinaturas hidrológicas	71
4.5	Conclusões	82
5	SOBRE A ESTRUTURA DO MODELO CONCEITUAL NA MODELAGEM HIDROLÓGICA	84

5.1	Introdução	84
5.2	Desenvolvimento teórico	86
5.2.1	Mecanismos e processos hidrológicos	86
5.2.2	O FUSE.....	88
5.2.3	Regras de associação: Apriori	91
5.3	Metodologia.....	92
5.4	Resultados	95
5.4.1	Camada superior do solo	96
5.4.2	Camada inferior do solo	99
5.4.3	Escoamento superficial	104
5.4.4	Percolação	107
5.4.5	Evapotranspiração	111
5.4.6	Fluxo intermediário	114
5.4.7	Propagação em bacias	116
5.4.8	Bacias sem estruturas adequadas.....	117
5.4.9	Principais características e assinaturas	119
5.4.10	Agrupamentos	121
5.5	Conclusão	122
6	REDES NEURAIS LSTM TREINADAS UTILIZANDO ASSINATURAS HIDROLÓGICAS	126
6.1	Introdução	126
6.2	Redes neurais e LSTM.....	129
6.3	Metodologia.....	132
6.4	Resultados e discussão	135
6.5	Conclusão	141
7	CONCLUSÃO GERAL E RECOMENDAÇÕES.....	142
	REFERÊNCIAS.....	147

1 INTRODUÇÃO

Há diversas razões pelas quais é necessário modelar o comportamento das bacias. A principal, talvez, seja referente às limitações das medições: não é possível monitorar precisamente todos os processos físicos em escala de bacia. Ademais, é preciso extrapolar as medições disponíveis, no espaço e no tempo, para bacias não monitoradas, e avaliar cenários em que as condições atuais sejam alteradas. Modelos fornecem um meio de extrapolar ou prever a resposta hidrológica, auxiliando a tomada de decisão em problemas de engenharia de recursos hídricos, como no planejamento e gerenciamento dos recursos hídricos, em inundações urbanas, na avaliação da qualidade da água de mananciais etc. (CLARK et al., 2015; HRACHOWITZ; CLARK, 2017).

Ao longo dos anos diversas teorias para cada um dos processos do ciclo hidrológico foram desenvolvidas. Enquanto modelos de base física tem sido discutidos e aprimorados há diversas décadas, os modelos orientados a dados, na hidrologia, só ganharam relevância em anos recentes (NEARING et al., 2021). Por esse motivo, há centenas de estruturas de modelos conceituais na literatura, assim como uma variedade de modelos com estruturas flexíveis que permitem explorar as alternativas de estruturas e identificar representações apropriadas (BROWN; CRAIG, 2020; DAL MOLIN; KAVETSKI; FENICIA, 2021). Dentre os modelos orientados a dados, os algoritmos de aprendizado por máquina, sobretudo as redes com memória de longo e curto prazo (LSTM, do inglês *Long Short-Term Memory*), têm sido largamente utilizadas, com êxito, em estudos hidrológicos recentes (HUNT et al., 2022; LEY; BORMANN; CASPER, 2023).

Para representar o comportamento da bacia em um modelo é adotado o processo de calibração, que envolve a identificação de parâmetros do modelo e seus valores otimizados (CHILKOTI; BOLISSETTI; BALACHANDAR, 2019). A utilização de outras métricas hidrológicas na calibração pode torná-la mais robusta (KAVETSKI et al., 2018). Quando essas métricas representam o comportamento dos dados observados elas têm sido chamadas de assinaturas hidrológicas. Em geral, as abordagens baseadas em assinaturas hidrológicas têm fornecido melhores previsões das métricas pré-selecionadas (HALLOUIN; BRUEN; O'LOUGHLIN, 2020).

A calibração baseada em assinaturas pode auxiliar ainda em situações com escassez de informações hidrométricas. Em bacias não monitoradas, algumas assinaturas podem ser estimadas por regionalização (GUO et al., 2021b) e usadas para calibração (KITTEL et al., 2020). Quando se atinge uma representação adequada das assinaturas hidrológicas de interesse, o modelo tem sido denominado como hidrologicamente consistente (FENICIA et al., 2018; SHAFII; TOLSON, 2015). Entretanto, a inclusão de outras métricas no processo da busca otimizada pelos parâmetros do modelo aumentam a complexidade do problema, o que exige que esse tema seja melhor explorado (HERNANDEZ-SUAREZ; NEJADHASHEMI; DEB, 2021; HUO; LIU; ZHANG, 2018; SHAFII; TOLSON, 2015).

Há tempos há um consenso entre hidrólogos que existe uma relação entre modelos e características físicas das bacias. As interações entre clima, solos, vegetação e topografia permite, em tese, a identificação de padrões em bacias hidrográficas “naturais”, os quais refletem o funcionamento desses sistemas. Na realidade, as bacias hidrográficas são altamente heterogêneas, dinâmicas e encontram-se em contínua evolução, em relação à vegetação, solo e morfologia, respondendo dinamicamente às forçantes climáticas, também variáveis no tempo e no espaço (BERGHUIJS et al., 2016; BERGHUIJS; WOODS; HRACHOWITZ, 2014).

Entretanto, a estimativa de assinaturas hidrológicas a partir dos atributos físicos é um problema ainda em aberto (CHOUBIN et al., 2019). Dentre as técnicas atuais, os algoritmos de regressão baseados em aprendizado de máquina têm se destacado (WU et al., 2022). A identificação de regiões homogêneas é um objetivo usual em estudos de regionalização, que leva ao problema de classificação de bacias. Embora não se tenha um consenso sobre a classificação de bacias, identificar e categorizar as funções dominantes lançaria luz às razões da similaridade entre elas, quantificando o grau de semelhança que pode existir (DOOGE, 1986; LYON; TROCH, 2010).

É comum que hidrólogos assumam como verdadeira a hipótese de que a similaridade entre atributos físicos se estenda à resposta hidrológica e permita a definição de unidades. Com efeito, quando se dispõe de monitoramento, é possível verificar essa hipótese por meio da análise de séries temporais. A vazão é usualmente utilizada para isso, por ser considerada uma resposta integrada a todos os processos físicos

ocorridos na bacia. Assim, parece plausível admitir que uma avaliação de modelos hidrológicos se apoie nas observações derivadas da relação entre assinaturas hidrológicas e características físicas.

Este estudo se orienta pela avaliação dessas inter-relações. O Capítulo 5 busca interpretar essas relações à luz das estruturas de modelos conceituais, empregando uma extensa base de dados de bacias públicas denominada *Catchment Attributes for Brazil* (CABra; ver ALMAGRO et al., 2021) e um *framework* multimodelos que permite a representação de até 624 estruturas, o *Framework for Understanding Structural Errors* (FUSE; ver CLARK et al., 2008). Mas antes, o Capítulo 3 explora a calibração multiobjetivo desses modelos conceituais, com o objetivo de identificar a melhor estratégia de incorporar as assinaturas hidrológicas no algoritmo de otimização.

O Capítulo 4 aborda a relação entre características físicas e assinaturas hidrológicas, e classifica as bacias brasileiras por semelhança de suas assinaturas hidrológicas. Por fim, o Capítulo 6 avalia os resultados e as interpretações derivadas de modelos baseados em dados, por meio das redes LSTM, utilizando as assinaturas hidrológicas, as características físicas e os grupos de bacias similares.

Desta forma, não se buscou caracterizar ou prescrever explicitamente a homogeneidade das bacias ou modelos a fim de descrever a complexidade dos processos envolvidos, mas sim explorar as frágeis conclusões subjacentes à aparente simplicidade assumida durante a avaliação e interpretação de resultados derivados de modelos hidrológicos. O que este estudo de propôs foi identificar e analisar as inter-relações entre as características físicas, assinaturas hidrológicas e a estrutura do modelo, seja ele conceitual ou orientado a dados.

Se essas relações existirem e forem estabelecidas, admite-se que as predições das respostas hidrológicas e as interpretações físicas delas inferidas sejam suficientemente verossímeis para a construção de modelos úteis, cuja estrutura permita sua extrapolação espacial e temporal. Por isso, se reconhece que o interesse não se resume ao estabelecimento de padrões, mas a caracterização crível de diferenças entre bacias hidrográficas.

2 OBJETIVOS

2.1 Objetivo Geral

O objetivo geral do estudo é investigar as inter-relações entre os atributos físicos das bacias, suas assinaturas hidrológicas e a estrutura de modelos, conceituais por meio do *framework* flexível FUSE, e baseados em dados a partir de redes neurais do tipo LSTM.

2.2 Objetivos Específicos

Cada capítulo desse estudo busca os seguintes objetivos específicos:

- Identificar a abordagem de otimização mais adequada para obter conjuntos de parâmetros hidrológicamente consistentes;
- Estabelecer relações entre atributos físicos e assinaturas hidrológicas para caracterização de comportamentos hidrológicos similares;
- Caracterizar relações entre atributos físicos, assinaturas hidrológicas e a estrutura de modelos conceituais; e
- Avaliar o desempenho e a interpretação física dos modelos orientados a dados do tipo redes neurais LSTM em relação aos modelos conceituais identificados.

3 CALIBRAÇÃO MULTIOBJETIVO UTILIZANDO ASSINATURAS HIDROLÓGICAS

“Catchment hydrology is at a cross roads.”

(SEIBERT; MCDONNELL, 2002)

3.1 Introdução

O desenvolvimento de modelos hidrológicos confiáveis é um componente vital no processo de tomada de decisão na gestão de recursos hídricos. Os modelos conceituais são abstrações físicas, em que seus parâmetros representam as propriedades das bacias. O desempenho da predição dos modelos conceituais depende, portanto, de uma adequada estimativa dos seus parâmetros (BOMHOF; TOLSON; KOUWEN, 2019; MOSTAFAIE et al., 2018).

Para tanto, geralmente é adotado o processo de calibração, que envolve a identificação de parâmetros do modelo e seus valores otimizados. Segue-se à comparação das variáveis observadas e simuladas e a quantificação do desempenho por meio de uma função objetivo ou verossimilhança que mede o ajuste entre as variáveis. (CHILKOTI; BOLISSETTI; BALACHANDAR, 2019).

Em uma abordagem mais simplista, as funções objetivo são formuladas com base em métricas de desempenho tradicionais, com diferentes transformações, em uma tentativa de enfatizar ou equilibrar a importância de diferentes condições da variável de interesse (HERNANDEZ-SUAREZ; NEJADHASHEMI; DEB, 2021). Tal abordagem é por vezes referida como baseada em desempenho ou no domínio do tempo (FENICIA et al., 2018).

A vazão observada no exutório da bacia hidrográfica, normalmente utilizada na calibração, é consequência de vários subprocessos, como infiltração, fluxo lateral, evaporação, escoamento de base e percolação (ARNOLD et al., 2012). As funções objetivo empregadas para mensurar a similaridade entre variáveis modeladas e observadas podem não ser capazes de caracterizar a relevâncias desses processos (GUPTA; WAGENER; LIU, 2008). Dessa forma, métricas como a soma ao quadrado das diferenças entre vazões observadas e modeladas ou a medida de eficiência de Nash-Sutcliffe (NSE) (NASH; SUTCLIFFE, 1970) podem não ser suficientes para

estimar parâmetros hidrológicamente consistentes. Da mesma forma, a Máxima Verossimilhança pode não garantir a identificação de estruturas e/ou parâmetros do modelo que reproduzam características hidrológicas desejadas dos processos sob investigação (MARTINEZ; GUPTA, 2011).

Para resolver esse problema, tem-se explicitamente incorporado métricas de interesse nas simulações, avaliando os modelos não apenas com base em medidas estatísticas, mas também utilizando outros indicadores hidrológicos (HERNANDEZ-SUAREZ; NEJADHASHEMI; DEB, 2021). Vários estudos empregam assinaturas hidrológicas, que refletem o comportamento da bacia hidrográfica, desejáveis de serem reproduzidas pelo modelo hidrológico (por exemplo, DAGGUPATI et al., 2015; KIESEL et al., 2020; SAHRAEI; ASADZADEH; UNDUICHE, 2020; ZHANG et al., 2016). Fenicia et al. (2018) chamam essa abordagem de “calibração no domínio da assinatura”, em contraste com a “calibração no domínio do tempo”.

Em geral, as abordagens baseadas em assinaturas hidrológicas têm fornecido melhores previsões das métricas pré-selecionadas em comparação com as abordagens baseadas somente em desempenho (HALLOUIN; BRUEN; O'LOUGHLIN, 2020). Considerar as características do regime em vez da magnitude da vazão, durante a calibração, têm potencial para aliviar os problemas causados por incertezas dos parâmetros e equifinalidade, melhorando a confiança no realismo do modelo. Como resultado, processos hidrológicos mais fidedignos seriam capturados, forçando diretamente o modelo a reproduzir características importantes do hidrograma, e o desempenho do modelo seria melhorado em uma calibração multiobjetivo (GUPTA; WAGENER; LIU, 2008; ZHANG et al., 2016).

A utilização das assinaturas pode também tornar a calibração mais robusta a deficiências das métricas de desempenho. Ao se utilizar assinaturas menos sensíveis a erros no tempo e magnitude dos picos, pode-se estimar parâmetros mais adequados do que na calibração baseada na minimização de resíduos das séries temporais (FENICIA et al., 2018). Da mesma forma, assinaturas podem filtrar o ruído e reduzir a sensibilidade da função objetivo à incerteza dos dados (SADEGH et al., 2015; WESTERBERG; MCMILLAN, 2015).

A calibração baseada em assinaturas pode auxiliar ainda em situações com escassez de informações hidrométricas. Por exemplo, quando a precipitação observada e as séries temporais de vazão não são concomitantes, a calibração no domínio de assinatura pode assumi-las aproximadamente constantes no tempo (SADEGH et al., 2015). Em bacias não monitoradas, algumas assinaturas podem ser estimadas por regionalização (GUO et al., 2021b) e usadas para calibração (KITTEL et al., 2020).

As principais abordagens para calibração de diversos objetivos tem sido: (i) agregá-los em um único objetivo (KHU; MADSEN, 2005; ZHANG et al., 2016); (ii) transformar os objetivos em restrições (GOBEYN; GOETHALS, 2019; SHAFII; TOLSON, 2015); ou (iii) uma frente de Pareto (FENICIA et al., 2018; HERNANDEZ-SUAREZ; NEJADHASHEMI; DEB, 2021). A maioria desses estudos relatam uma grande dificuldade para a frente de Pareto com o aumento no número de objetivos.

O problema central da calibração utilizando a frente de Pareto é que ele se torna um problema de tomada de decisão, i.e., de quais conjuntos de calibração selecionar dentre as muitas possibilidades. Por outro lado, os objetivos da métrica de desempenho e as assinaturas hidrológicas não são necessariamente conflitantes, como os problemas de otimização clássicos da literatura.

Diversos estudos atuais tecem conclusões sobre a robustez da calibração multiobjetivo na consistência e equifinalidade dos modelos hidrológicos (ver ALTHOFF; RODRIGUES, 2021; HER; SEONG, 2018; HERNANDEZ-SUAREZ; NEJADHASHEMI; DEB, 2021; TUO et al., 2018), ou não (ZHANG et al., 2018b). Contudo, ao se empregar uma metaheurística na busca de uma solução otimizada no espaço de soluções, não se pode assumir previamente que a heurística ou a estrutura do modelo são adequadas. Primeiro, é necessário verificar se que tal abordagem é aplicável ao modelo hidrológico.

Por essa razão, é importante certificar-se quando a utilização das estratégias de ordenação da frente de Pareto em um contexto hidrológico, assim como verificar a aplicabilidade do conceito de dominância, avaliando um grande conjunto de estruturas de modelos em várias bacias. Outras abordagens para tratar problemas multiobjetivos, como funções agregadoras e utilizar as assinaturas como restrições,

precisam ser avaliadas. Pode-se ainda valer-se da lógica fuzzy, o que permitiria abstrair a imprecisão dos dados hidrométricos (SEIBERT; MCDONNELL, 2002)

O que este estudo se propõe é investigar estratégias de frente de Pareto e de agregação das funções objetivo na calibração de um grande conjunto de estruturas de modelo, avaliando qual metaheurística conseguiria calibrar mais parâmetros das estruturas de modelos dentro de limites para considerá-los hidrológicamente consistentes. As questões que se pretende responder são:

- Qual o desempenho das estratégias multiobjetivos com assinaturas hidrológicas frente à calibração de objetivo único no domínio do tempo?
- Qual a melhor abordagem multiobjetivo para otimizar tanto as assinaturas hidrológicas quanto a métrica de desempenho (NSE)?
- Qual o comportamento do NSE quando somente as assinaturas hidrológicas são empregadas na calibração?

3.2 Desenvolvimento teórico

3.2.1 Calibração multiobjetivo

A calibração de modelo hidrológicos pode ser abordado como um problema de otimização. Os parâmetros do modelo precisam ser calibrados, para que o modelo possa representar adequadamente os processos hidrológicos. Portanto, a seleção de parâmetros tem um impacto crucial no desempenho geral do modelo e nos resultados da predição hidrológica. Entretanto, devido à complexidade, o elevado número de parâmetros, à alta dimensionalidade e à grande quantidade de dados a serem processado pelo modelo, o problema de otimização de parâmetros do modelo hidrológico é um problema NP-difícil que precisa ser resolvido por uma heurística (HUO; LIU; ZHANG, 2018).

A calibração multiobjetivo de um modelo envolve ajustar as estimativas a duas ou mais funções objetivo que quantificam sua consistência com a dos dados medidos da resposta do sistema (BOMHOF; TOLSON; KOUWEN, 2019). Ela poderia ser usada, a exemplo de Tuo et al. (2018), para calibrar os parâmetros utilizando tanto a vazão

quando a profundidade da camada neve observadas. Transformações na métrica do tempo podem ser realizadas para calibrar os parâmetros utilizando tanto as altas como as baixas vazões (YANG et al., 2022) ou a utilização conjunta dos percentis da curva de permanência (CHILKOTI; BOLISETTI; BALACHANDAR, 2018), ainda que este não apresente os resultados esperados (GARCIA; FOLTON; OUDIN, 2017).

A característica primordial de um problema multiobjetivo é que a solução não será, em geral, única. De fato, é comum ter várias soluções com a propriedade de que a passagem de uma solução para outra resulta na melhoria de um objetivo enquanto causa uma deterioração no valor de pelo menos outro. Sem definir prioridades não é possível determinar objetivamente, dentre tais soluções, qual é a melhor. Estas soluções são chamadas de conjunto ótimo de Pareto ou não dominadas (YAPO; GUPTA; SOROOSHIAN, 1998). Todas as outras soluções que não estão na frente de Pareto ótimas são referidas como soluções dominadas porque são de qualidade igual ou pior em todos os objetivos em comparação com pelo menos uma solução não dominada (BOMHOF; TOLSON; KOUWEN, 2019).

Uma alternativa à frente de Pareto ótima é definir um único objetivo e adicionar qualquer objetivo adicional como uma restrição (GOBEYN; GOETHALS, 2019; QIN et al., 2015). Com o aumento do número de objetivos, o número de soluções não dominadas geradas pelos algoritmos torna-se muito grande, levando à “resistência à dominância” (PURSHOUSE; FLEMING, 2007). Essa resistência refere-se às dificuldades em produzir novas soluções candidatas que dominarão as soluções atuais de desempenho ruim, mas localmente não dominadas. Shafii e Tolson (2015), em estudo envolvendo 15 objetivos, também destacaram essa questão, e esse foi o motivo para converter os objetivos contínuos em objetivos binários, o que na verdade combina a abordagem de frente de Pareto com a de transformar os objetivos em restrições. Quando a diferença relativa de uma assinatura hidrológica é menor que um determinado limiar ela recebe uma pontuação de 1, caso contrário a pontuação para essa assinatura será zero.

Complementarmente, a alternativa mais tradicional ao se tratar de problemas multiobjetivos é agregar todos em uma única função (BOMHOF; TOLSON; KOUWEN, 2019), como por exemplo a soma ponderada usada por Madsen (2003) e Garcia et al.

(2017), em que as diferentes funções objetivo foram transformadas e ponderadas para compensar as diferenças nas magnitudes da vazão. Já Madsen (2000) e Liong al. (2001) usaram uma função de distância. Adicionalmente, Hernandez-Suarez et al. (2021) compararam uma técnica denominada programação de compromisso (ZELENY, 2011) e um método de pseudo-pesos (DEB, 2011), sem obter diferença entre os métodos. Por fim, Fenicia et al. (2018) agregaram os objetivos calculados pela diferença relativa entre as assinaturas hidrológicas simuladas e observadas naquela de maior valor.

Por outro lado, Efstratiadis e Koutsoyiannis (2010) listaram as principais deficiências da abordagem tradicional de agregar múltiplos objetivos em um único escalar na calibração de parâmetros de modelos hidrológicos. Eles observam que a agregação de objetivos aumenta a complexidade da superfície da função objetivo, o que pode fazer com que até mesmo os melhores otimizadores de objetivo único retornem soluções dominadas ou soluções com um compromisso ruim entre os objetivos (por exemplo, região extrema da frente de Pareto).

Independente da abordagem para se solucionar um problema de otimização multiobjetivo, a questão de um problema NP-difícil é como percorrer o espaço de soluções em busca de uma solução otimizada. Hanus et al. (2021), por exemplo, utilizaram o método de Monte Carlo com três milhões de simulações em 6 bacias na Áustria para calibrar os parâmetros utilizando oito objetivos baseados em assinaturas hidrológicas. Smith et al. (2019) utilizaram o Hiper Cubo Latino para calibrar os parâmetros do modelo GR4J em 303 bacias do Reino Unido, utilizando seis objetivos e cem mil simulações.

Entretanto tais soluções exigem um poder computacional considerável, levando a metaheurísticas para encontrar soluções otimizadas. Diversas técnicas vêm sendo empregadas na literatura hidrológica (ver EFSTRATIADIS; KOUTSOYIANNIS, 2010). As próximas seções apresentam os algoritmos NSGA-II e III, amplamente utilizados em otimizações multiobjetivos e o algoritmo SCE-UA, demasiadamente utilizado em estudos hidrológicos, especialmente de objetivo único.

3.2.2 NSGA-II e III

O *Nondominated Sorting Genetic Algorithm II* (NSGA-II; DEB et al., 2002) é um algoritmo evolutivo multiobjetivo rápido e elitista usado para propagar uma população de soluções candidatas para uma frente de Pareto definida por um conjunto de objetivos. O algoritmo faz uso de conceitos como como mutação, cruzamento e seleção para encontrar soluções otimizadas para a frente de Pareto (DEB, 2011). As soluções são codificadas em cromossomos. O algoritmo atribui valores de aptidão a esses cromossomos com base em uma abordagem de classificação não dominada. A abordagem de ordenação não dominada visa preservar a diversidade entre as soluções não dominadas, fazendo uso de um conceito chamado de distância da multidão. A distância da multidão é uma medida da proximidade com as demais soluções que é usada na seleção, privilegiando soluções mais distantes das demais (GOBEYN; GOETHALS, 2019; WANG et al., 2018).

O NSGA-II tem sido utilizado extensivamente na literatura hidrológica. Por exemplo, Ercan e Goodall (2016) propuseram uma biblioteca para calibrar os parâmetros do modelo SWAT utilizando-o. Newland et al. (2018) empregaram-no para calibrar os parâmetros de um modelo celular autômato de mudanças no uso do solo. Yang et al. (2014) utilizaram para calibrar os parâmetros do modelo distribuído MOBIDIC. Mas, na comparação com outros algoritmos o desempenho do NSGA-II não é um destaque. Mostafaie et al. (2018) compararam cinco diferentes técnicas de calibração no modelo hidrológico concentrado GR4J, dentre elas o NSGA-II, sendo a Otimização de Enxame de Partículas Multiobjetivo (MPSO, REDDY; NAGESH KUMAR, 2007) a melhor em termos de precisão. Em Huo e Liu (2019) o NSGA-II também teve desempenho inferior a outros três algoritmos, dentre eles o Otimização por Enxame de Partículas (PSO, do inglês *Particle Swarm Optimization*) e a Colônia Artificial de Abelhas (ABC, do inglês *Artificial Bee Colony*) multiobjetivos na calibração do modelo Xinanjiang.

Reed e Devireddy (2004) propuseram uma versão aprimorada do algoritmo NSGA-II, chamada ϵ -NSGA-II, na qual eles empregam o conceito de dominância ϵ , permitindo que o usuário especifique a precisão com a qual deseja quantificar cada objetivo a ser otimizado. Sahrei et al. (2019) concluíram que, em modelos hidrológicos, os algoritmos de otimização multiobjetivos precisam ser aprimorados com estratégias, como a dominância ϵ , para evitar um aumento exponencial no número de soluções

não dominadas. Adicionalmente, Monteil et al. (2020) combinaram a evolução multiobjetivo utilizando simplex com o ϵ -NSGA-II.

O NSGA-III é baseado na estrutura NSGA-II, também usando duas estratégias de calibração multiobjetivo. A principal diferença é o procedimento para manter a diversidade entre as soluções. Enquanto o NSGA-II usa distâncias da multidão, o NSGA-III é baseado em direções de referência. Uma direção de referência é uma linha que cruza a origem e um ponto de referência fornecido no espaço objetivo (DEB; JAIN, 2014; JAIN; DEB, 2014). Ele foi utilizado por Hernandez-Suarez et al. (2018) na calibração dos parâmetros do SWAT utilizando 167 assinaturas hidrológicas. Já Swathi et al. (2019) utilizaram-no na calibração dos parâmetros do SWMM.

3.2.3 SCE-UA

O método *Shuffled Complex Evolution* (SCE-UA), desenvolvido na Universidade do Arizona, é baseado na ideia de partes independentes buscando soluções para um problema e depois trabalhando juntas para compartilhar suas descobertas (DUAN; SOROOSHIAN; GUPTA, 1992).

O SCE-UA é um algoritmo de otimização global baseado em quatro conceitos: a combinação de abordagens determinísticas e probabilísticas, evolução sistemática de “complexos” de pontos que abrangem o espaço de parâmetros, na direção do ótimo global, evolução competitiva e embaralhamento complexo (DUAN; SOROOSHIAN; GUPTA, 1994).

Estudos têm mostrado que o algoritmo SCE é eficaz e eficiente para a calibração dos parâmetros de modelos hidrológicos (por exemplo, KAN et al., 2018a, 2018b; SEONG; HER; BENHAM, 2015), podendo ser utilizado, inclusive, na busca multiobjetivo agregando todos em um único (GAN et al., 2018; KITTEL et al., 2020).

O algoritmo SCE-UA original foi desenvolvido para resolver problemas de otimização de objetivo único. Posteriormente, Yapo et al. (1998) estenderam o algoritmo para resolver problemas de otimização multiobjetivo, desenvolvendo o algoritmo *Multi-Objective Complex Evolution* (MOCOM). O MOCOM modifica o algoritmo SCE-UA para empregar a classificação por frente Pareto, ordenando as soluções pelo critério de dominância.

Posteriormente, Guo et al. (2013) desenvolveram o *Multi-Objective Shuffled Complex Differential Evolution* (MOSCDE). Com intuito de utilizar as informações dos indivíduos na população evolutiva de forma mais completa e melhorar a capacidade de pesquisa, o algoritmo MOSCDE substitui a pesquisa *simplex* do SCE pelo algoritmo de evolução diferencial (DE, do inglês *Differential Evolution*) e estende o algoritmo de objetivo único do SCE para a estrutura de multiobjetivo utilizando a frente de Pareto e a distância da multidão. Além disso, o operador de mutação de Cauchy é introduzido para evitar o problema de convergência prematura da evolução diferencial. O MOSCDE demonstrou ser mais eficiente do que NSGA-II em cinco *benchmarks* de referência.

Outra alternativa multiobjetivo é o MOSCEM, uma versão estendida do método SCEM-UA para avaliação de incerteza (VRUGT et al., 2003) que mescla a vantagem do embaralhamento complexo com a estratégia de pesquisa do algoritmo Metropolis e o procedimento de atribuição de aptidão empregado dentro do algoritmo SPEA (ZITZLER; THIELE, 1999). O MOSCEM foi utilizado por Zhang et al. (2018) em duas métricas de desempenho no modelo FLEX-Topo em uma bacia na China.

Por fim, Adeyri et al. (2020) utilizaram uma variante multiobjetivo do SCE (SCE-R), superando 5 técnicas dentre elas algoritmos baseados no PSO e DE multiobjetivos na calibração dos parâmetros do modelo GR5J.

3.3 Metodologia

O presente estudo compara diversos algoritmos para calibração multiobjetivo dos parâmetros de 624 estruturas de modelo do FUSE em 9 bacias brasileiras, valendo-se de um objetivo no domínio do tempo e 5 assinaturas hidrológicas.

O FUSE é uma estrutura de modelagem modular que permite a geração de centenas de modelos hidrológicos conceituais pela recombinação de elementos de modelos comumente usados, permitindo escolher vários mecanismos para representar cada processo hidrológico no modelo (CLARK et al., 2008). Mais detalhes são apresentados no Capítulo **Erro! Fonte de referência não encontrada.**

A metodologia pode ser dividida em duas etapas:

1. Avaliação de variações do conceito de dominância, ordenação da frente de Pareto e evolução das soluções em adaptações multiobjetivo do SCE-UA; e
2. Comparação das abordagens de frente de Pareto, restrições e função agregadora variando as condições de como um objetivo é atingido ou agregado.

A partir do algoritmo SCE-UA, implementado no *framework* multimodelos do FUSE, foram adaptadas duas heurísticas para evolução dos complexos. A primeira é baseada no MOCOM (YAPO; GUPTA; SOROOSHIAN, 1998), sendo chamada nesse trabalho de MOCOM* pelas incertezas na descrição do algoritmo. A segunda é uma implementação do MOSCDE (GUO et al., 2013).

Em relação à formação da frente de Pareto, alterou-se o conceito da dominância, para a seguinte proposição: uma solução domina uma segunda caso a supere na maioria dos objetivos. Dessa forma, a frente de Pareto ótima é formada por aquelas soluções que empatam na quantidade de objetivos superados e não são dominadas por nenhuma outra. Havendo 6 objetivos, por exemplo, a solução A e B estão na mesma frente de Pareto se A é superior a B em 3 objetivos, e B é superior a A nos outros 3. O desempate para ordenação na evolução dos complexos é feito pelo objetivo no domínio do tempo. Caso A seja superior a C em 2 objetivos e C seja superior nos outros 4, nessa nova abordagem, C domina A.

Essa abordagem foi comparada com a formação da frente de Pareto com o conceito de dominância tradicional, utilizando a ordenação das soluções conforme o NSGA-II e NSGA-III.

Uma vez definida a melhor estratégia para geração da frente de Pareto testou-se variações na forma de se considerar cada objetivo, agregando as assinaturas hidrológicas em um objetivo único por meio de diferentes funções.

Tendo o FUSE 624 estruturas de modelos disponíveis, todas as comparações foram no âmbito de quantas estruturas, hidrológicamente consistentes, cada heurística consegue calibrar. Definiu-se como consistentes aquelas que apresentam erro relativo máximo de 5% em cada uma das assinaturas hidrológicas (SHAFII; TOLSON, 2015). Para cada algoritmo realizou-se 10.000 simulações.

Utilizou-se 9 bacias da base do CABra aleatoriamente distribuídas nas regiões brasileiras, usando os 15 primeiros anos das séries temporais, correspondente ao período de 01/10/1980 à 30/09/1995. Os primeiros 6 meses foram utilizados como aquecimento do modelo. Não foi o foco desse estudo avaliar a validação (SHEN; TOLSON; MAI, 2022).

3.3.1 Métricas

A função objetivo utilizada para calibrar os parâmetros do modelo hidrológico quantifica a qualidade do ajuste entre o modelo e os dados observados. Por exemplo, o NSE tem sido muito utilizado quando o interesse é principalmente os picos de vazão. Transformações de raiz quadrada e logarítmica são frequentemente aplicadas as vazões com o intuito que o NSE represente melhor as vazões mais baixas (PUSHPALATHA et al., 2012). A eficiência de Kling-Gupta, ou KGE (GUPTA et al., 2009), é uma função objetivo mais recente baseada em uma combinação ponderada da variabilidade relativa, correlação e viés, e tem se tornado bastante popular (MIZUKAMI et al., 2019).

Para obter previsões probabilísticas de alta qualidade, o modelo de resíduos deve fornecer uma descrição realista das propriedades estatísticas dos erros nas previsões determinísticas. Segundo McInerney *et al.* (2017) os modelos de resíduos que encontram um bom desempenho na escala diária incluem aqueles que transformam o escoamento usando a transformação Box-Cox (BC). Já Hunter et al. (2021) conclui que o modelo de erro com a transformação de Box-Cox, considerando $\lambda = 0,2$ (BC-02), fornece uma descrição adequada de erros residuais quando o modelo hidrológico é calibrado usando a função objetivo do NSE sendo aplicado a mesma transformação nas vazões (BC-02). Sendo assim, esse estudo emprega o NSE com a transformação de BC-02 como métrica de desempenho, embora o foco não seja descrever a incerteza preditiva.

A literatura utiliza uma variedade de assinaturas para avaliação de modelos, com um amplo uso da curva de permanência (CHILKOTI; BOLISSETTI; BALACHANDAR, 2018; SADEGH et al., 2015; WESTERBERG; MCMILLAN, 2015). Fenicia et al. (2018) sugerem que complementar os quantis da curva de permanência com a taxa de mudança melhora tanto a confiabilidade quanto a precisão das vazões preditas, que

se tornam semelhantes às obtidas a partir da calibração no domínio do tempo. Por isso, neste trabalho utilizou-se valores da curva de permanência (Q2, Q10, Q90 e Q98) e a taxa de mudança, sugeridas em Matos (2021) por apresentarem potencial para melhorar a estimativa dos parâmetros do modelo.

A diferença relativa de cada assinatura (ρ_i) foi calculada segundo a Equação (1) empregada em Kavetski et al. (2018):

$$\rho_i = \left| \frac{g_i - \bar{g}_i}{\bar{g}_i} \right| \quad (1)$$

Em que g_i é a assinatura hidrológica simulada e \bar{g}_i é a assinatura hidrológica observada.

3.3.2 SCE Multiobjetivo

O SCE-UA começa com uma geração aleatória de uma população inicial de soluções no espaço de parâmetros viável confinado pelos limites inferior e superior dos valores dos parâmetros. Cada solução individual é avaliada pela função objetivo que descreve a correspondência entre uma variável de saída do modelo e os valores observados.

Após a inicialização, os conjuntos de parâmetros são particionados em várias subamostras, chamadas de complexos. As soluções em cada complexo são evoluídas de acordo com o algoritmo genético CCE, baseado no método de busca *simplex* (NELDER; MEAD, 1965), na tentativa de substituir as piores soluções de menor aptidão por outras melhores. Nesta fase, cada complexo evolui independentemente por um certo número de gerações. Por fim, todas as soluções individuais dos complexos são embaralhadas em uma nova população, a partir da qual novos complexos são formados e evoluídos como antes (GUO et al., 2013).

O MOCOM-UA, proposto por Yapo et al. (1998), utiliza a frente de Pareto para ordenar as soluções e substitui o CCE por uma adaptação chamada MOSIM. Nessa adaptação, a etapa de reflexão só é utilizada se a pior solução é dominada pelas demais, caso contrário é utilizada a etapa de contração. Entretanto, a frente de Pareto formada pelo conceito de dominância ordena soluções em uma mesma posição. Não fica claro como MOCOM ordena essas soluções na formação dos complexos. Por esse motivo, a implementação utilizada nesse estudo será chamada de MOCOM*. O

presente estudo também avalia estratégias para ordenação das soluções de uma mesma frente de Pareto. Ou seja, dado um conjunto de soluções não dominadas, como atribuir uma ordem de preferência.

Conforme anteriormente descrito, o MOSCDE, proposto por (GUO et al., 2013), substitui a pesquisa *simplex* no SCE pelo algoritmo de evolução diferencial (DE) adicionando a mutação de Cauchy para evitar o problema de convergência prematura da evolução diferencial. O MOSCDE utiliza um complexo separado com todas as soluções não dominadas. Esse complexo é sempre copiado para o complexo que se pretende evoluir.

A evolução diferencial utiliza uma população (nesse caso cada complexo) contendo N elementos com vetores de n parâmetros x_i^g em uma geração g . Cada novo vetor de parâmetros v_i^{g+1} é gerado segundo a Equação (2)

$$v_i^{g+1} = x_{r_1}^g + F(x_{r_2}^g - x_{r_3}^g), \quad i = 1, 2, \dots, N \quad (2)$$

em os inteiros r_1, r_2 e r_3 são escolhidos aleatoriamente dentre as N soluções no complexo. O parâmetro de mutação F é um número real entre $[0,2]$, sendo razoável utilizar o valor de 0,5 segundo Storn e Price (1997). A implementação desse algoritmo realizada nesse estudo, inspirada nas etapas de contração e reflexão do SCE-UA, utilizou o valor de $F = 0,5$ em uma primeira tentativa de evolução e caso a solução não supere a pior do complexo se utiliza $F = 2$, conforme tradicionalmente realizado no SCE-UA.

A pior solução no complexo, $x_{r_4}^g$, é selecionada para ser substituída, sendo o conjunto novo de parâmetros formado conforme A Equação (3):

$$u_{i,j}^{g+1} = \begin{cases} v_{i,j}^{g+1} & \text{se } \text{aleatório}(0,1) \leq CR \text{ ou } j = \text{aleatório}(1,n) \\ x_{r_4,j}^g & \text{caso contrário} \end{cases} \quad (3)$$

sendo:

- j : cada um dos parâmetros;

- aleatório(1,n): um inteiro escolhido entre o número de parâmetros, garantindo que pelo menos um dos parâmetros serão do vetor $v_{i,j}^{g+1}$; e
- CR: o parâmetro de mutação sendo adotado nesse estudo o valor de 0,30 (QIN et al., 2010).

O novo conjunto de parâmetros $u_{i,j}^{g+1}$ é avaliado e, se dominar a solução $x_{r_4,j}^g$, a substitui no complexo.

A cada 5 iterações o MOSCDE aplica mutação de Cauchy, no complexo das soluções não dominadas, com o objetivo de evitar máximos locais e a convergência prematura eventual ocasionada pelo DE, conforme a Equação (4).

$$x_{i,j}^g = x_{i,j}^g + (1 + \eta C(0,1)), \quad \text{se diversidade}(j) < \varepsilon \text{ e aleatorio}(0,1) < 0,5 \quad (4)$$

sendo:

- $x_{i,j}^g$: o valor do parâmetro j no individuo i ;
- $C(0,1)$: a distribuição de Cauchy padrão;
- η : o coeficiente da mutação de Cauchy, nesse caso definido como $\eta = 0,5 - 0,4 * \frac{\text{iteração}}{\text{máxIterações}}$ sendo iteração e máxIterações a iteração atual e o número máximo de interações (conforme QIN et al., 2010), respectivamente; e
- ε : é o limite da diversidade sendo definido como 0,1 (QIN et al., 2010).

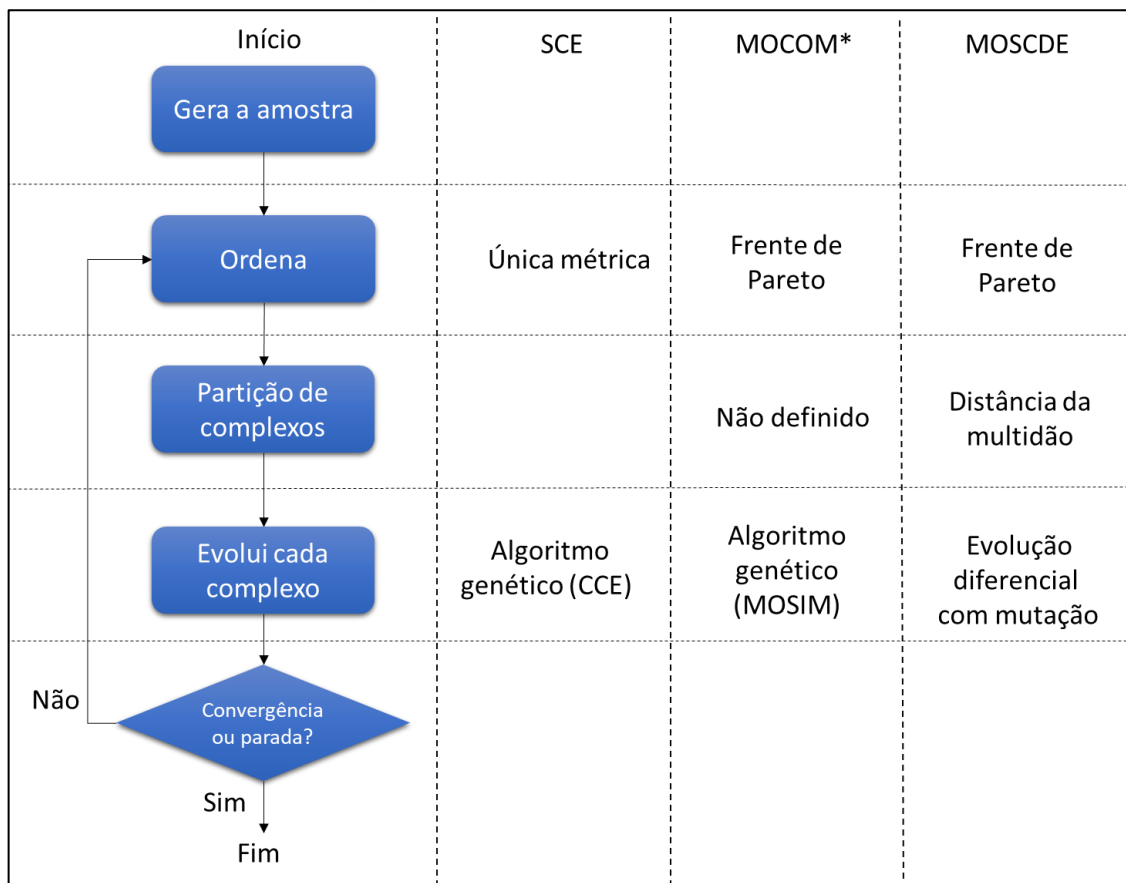
A diversidade é então calculada pela Equação (5):

$$\text{diversidade}(j) = \sqrt{\frac{1}{N} \sum_{i=1}^N \left(\frac{x_{i,j}^g - \overline{x_j^g}}{u_j - l_j} \right)^2} \quad (5)$$

sendo $\overline{x_j^g}$ a média dos parâmetros j das soluções no complexo e u_j e l_j os limites superiores e inferiores do parâmetro j , respectivamente.

As adaptações do SCE-UA para atender multiobjetivos são sumariadas na Figura 1. A partir dos dois algoritmos, MOCOM* e MOSCDE, foram avaliadas estratégias para formações da frente de Pareto.

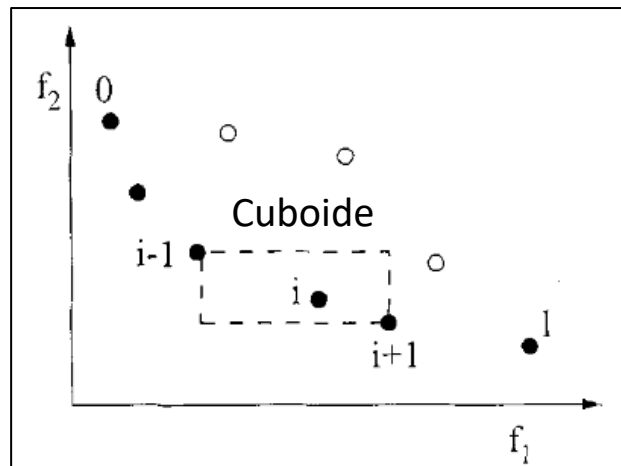
Figura 1 – Adaptações do SCE-UA para a abordagem multiobjetivo.



3.3.3 Ordenação da Frente de Pareto

Para ordenação das soluções, na formação dos complexos, alguma métrica precisa ser estabelecida para preferenciar as soluções de uma mesma frente de Pareto. Por exemplo, pode-se obter uma estimativa da densidade das soluções em torno de uma determinada solução, calculando a distância média da solução mais próxima de cada lado do ponto da solução, ao longo de cada um dos objetivos. Essa medida serve como uma estimativa do perímetro do cuboide formado usando os vizinhos mais próximos como vértices, também chamado de distância da multidão (DEB et al., 2002).

Figura 2 – Distância da multidão.



Fonte: (DEB et al., 2002)

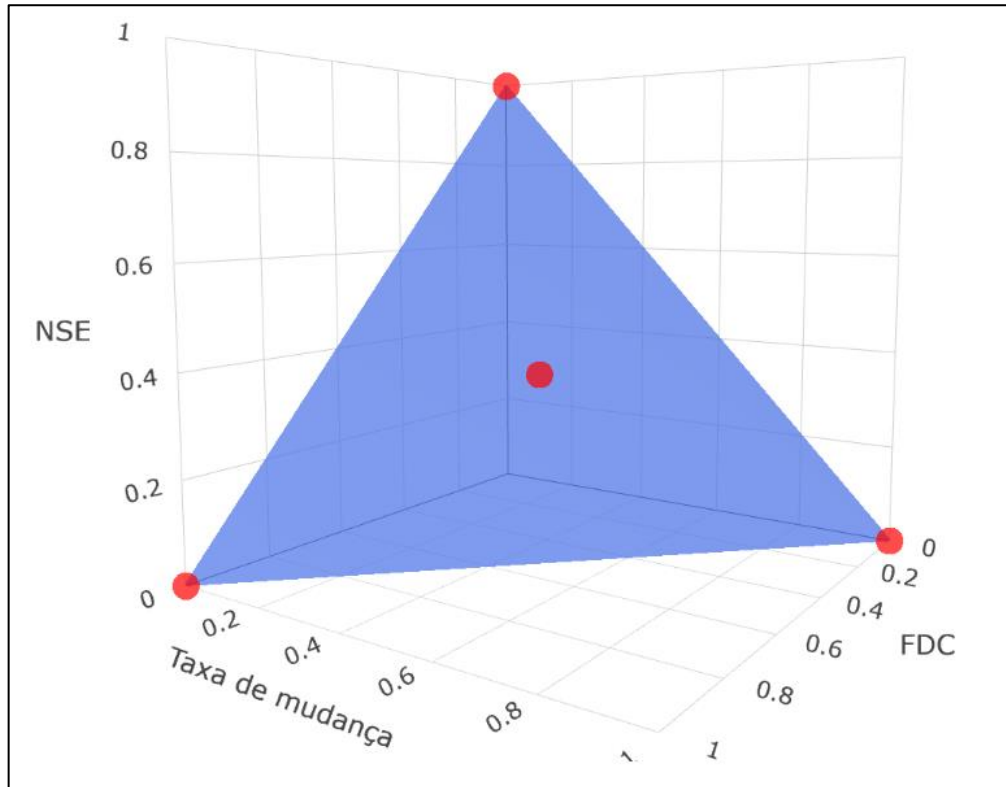
Na Figura 2 a distância da multidão da i -ésima solução em sua frente de Pareto (marcada com círculos sólidos) é o comprimento médio do lado do retângulo (mostrado com uma caixa tracejada). O cálculo da distância da multidão requer a classificação da população de acordo com cada valor da função objetivo em ordem crescente de magnitude. As soluções de contorno (soluções com os menores e maiores valores da função) recebem um valor de distância infinita. Todas as outras soluções intermediárias recebem um valor de distância igual à diferença normalizada absoluta nos valores das soluções adjacentes. Este cálculo é realizado para todas as funções objetivo. O valor global da distância da multidão é calculado como a soma dos valores de distância individuais correspondentes a cada objetivo.

É dessa forma que o MOSCDE ordena as soluções dentro das frentes de Pareto para a formação dos complexos. Tal artifício foi proposto no NSGA-II (DEB et al., 2002), com o objetivo de privilegiar soluções mais distantes das demais, mantendo uma boa densidade das soluções e evitando ótimos locais.

Conforme relatado, é atribuído uma distância infinita para ambos os extremos de um determinado objetivo, tanto o melhor quanto o pior valor. Neste sentido, o NSGA-III (DEB; JAIN, 2014) utiliza a abordagem de Das e Dennis (1998) para determinar um conjunto de pontos de referência a serem usados em cada geração. Esta abordagem define pontos de referência em um hiperplano normalizado, que é igualmente inclinado interceptando todos os eixos dos objetivos, além de pontos de referência uniformemente distribuídos em todo o hiperplano (YANNIBELLI et al., 2020). Por exemplo, a Figura 3 apresenta o hiperplano formado, com quatro pontos de referência,

caso os objetivos fossem maximizar o NSE, e minimizar as diferenças relativas em relação a percentis da curva de permanência e a taxa de mudança, além de uma referência que melhor equilibrasse os três objetivos.

Figura 3 – Pontos de referência no caso de 3 objetivos.



O NSGA-III, além de priorizar soluções não dominadas, também enfatiza os membros da população que estão, de alguma forma, associados a cada um desses pontos de referência. Uma vez que os pontos de referência criados acima são amplamente distribuídos em todo o hiperplano normalizado, as soluções obtidas provavelmente também serão amplamente distribuídas na frente de Pareto ótima ou próximos a ela.

O algoritmo sugere que esses pontos podem ser fornecidos pelo usuário. Dessa forma este trabalho utilizou $N+1$ pontos de referência, sendo N o número de objetivos: um para cada vértice de cada objetivo, capturando a melhor solução desse objetivo e um ponto no centro do hiperplano, privilegiando a solução que balanceia igualmente os objetivos.

Em síntese os algoritmos MOCOM* e MOSCDE foram avaliados:

1. com uma adaptação do conceito de dominância;

2. ordenando a frente de Pareto utilizando a distância da multidão; e
3. os pontos de referência no hiperplano;

Conforme anteriormente relatado, a adaptação do conceito de dominância considera na mesma frente de Pareto soluções que dominam umas às outras no mesmo número de objetivos.

3.3.4 Funções de agregação

Após avaliar as adaptações multiobjetivos do SCE, testou-se diferentes transformações sobre as assinaturas hidrológicas, alterando a forma de considerá-las na otimização único ou multiobjetivo. Para cada diferença relativa das assinaturas hidrológicas, calculou-se uma pontuação (P_i). As funções de agregação alteram a forma de se calcular P_i .

A primeira função é baseada no estudo de Shafii e Tolson (2015), valendo-se de uma pontuação binária para cada diferença relativa das assinaturas hidrológicas, penalizando as soluções com erro superior a 5%, conforme a Equação (6).

$$P_i = bin(\rho_i) = \begin{cases} 0 & \text{se } \rho_i \leq 0,05 \\ 1 & \text{caso contrário} \end{cases} \quad (6)$$

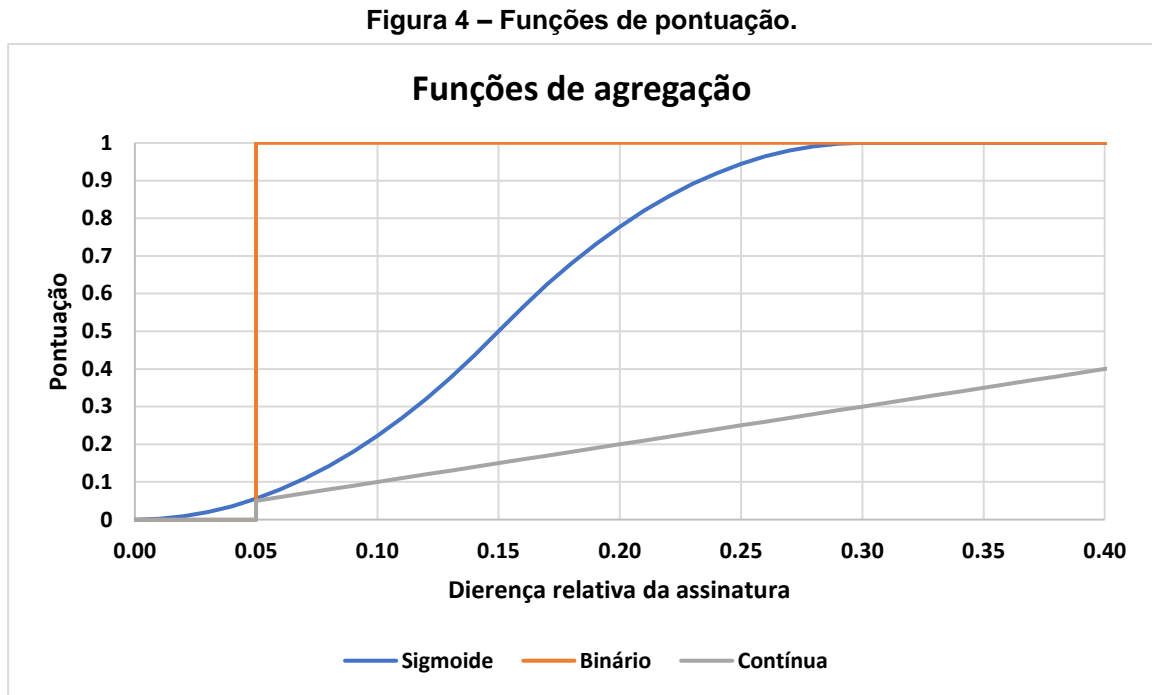
Adicionalmente, avaliou-se uma função contínua até atingir o limite requerido, apresentado na Equação (7):

$$P_i = cont(\rho_i) = \begin{cases} 0 & \text{se } \rho_i \leq 0,05 \\ \rho_i & \text{caso contrário} \end{cases} \quad (7)$$

A terceira alternativa foi uma pontuação contínua definida por uma função sigmoideal, pontuando diferenças relativas de até 30%, conforme Equação (8).

$$P_i = sig(\rho_i) = \begin{cases} 0 & \text{se } \rho_i = 0 \\ 2 * \left(\frac{\rho_i}{0,3}\right)^2 & \text{se } \rho_i \leq 0,15 \\ 1 - 2 * \left(\frac{0,3 - \rho_i}{0,3}\right)^2 & \text{se } \rho_i \leq 0,30 \\ 1 & \text{se } \rho_i \geq 0,30 \end{cases} \quad (8)$$

Todas essas funções foram traçadas na Figura 4.



A última função é a maior diferença relativa entre as assinaturas hidrológicas, proposto em Fenicia et al. (2018):

$$P_i = \max(\rho_{FDC}, \rho_{FI}) \quad (9)$$

Nesse caso os autores sugerem uma média das diferenças relativas dos quantis da curva de permanência ($\rho_{FDC} = \text{média}(\rho_{Q2}, \rho_{Q10}, \rho_{Q90}, \rho_{Q98})$) impedindo-os de dominar a métrica geral o que diminuiria a importância da taxa de mudança (ρ_{FI}).

Para o SCE de objetivo único a pontuação final foi a soma de todas as assinaturas e a componente do NSE (conforme SHAFII; TOLSON, 2015), apresentado na Equação (10):

$$P_f = (1 - NSE) + \sum P_i \quad (10)$$

Também se combinou a aplicação da função sigmoide sobre a assinatura mais distante do valor ótimo (Equações (8) e (9)), para o caso do objetivo único. Todas as

combinações de número de objetivos com as funções agregadoras estão sintetizadas na Tabela 1.

Tabela 1 – Configurações testadas

Id	Nº objetivos	Função sobre as assinaturas
0	1	Sem assinaturas
1	1	$\sum bin(\rho_i)$
2	1	$\sum sig(\rho_i)$
3	1	$sig(max(\rho_{FDC}, \rho_{FI}))$
4	1+5	-
5	1+1	$max(\rho_{FDC}, \rho_{FI})$
6	1+5	$bin(\rho_i)$
7	1+5	$cont(\rho_i)$
8	1+5	$sig(\rho_i)$

3.3.5 Área de estudo

As bacias utilizadas neste estudo foram escolhidas aleatoriamente na base do CABra, de modo a representar as 5 regiões do país, regimes hidrológicos e biomas distintos. Elas são apresentadas na Figura 5, enquanto suas características encontram-se sumarizadas na Tabela 2.

Figura 5 – Bacias utilizadas.

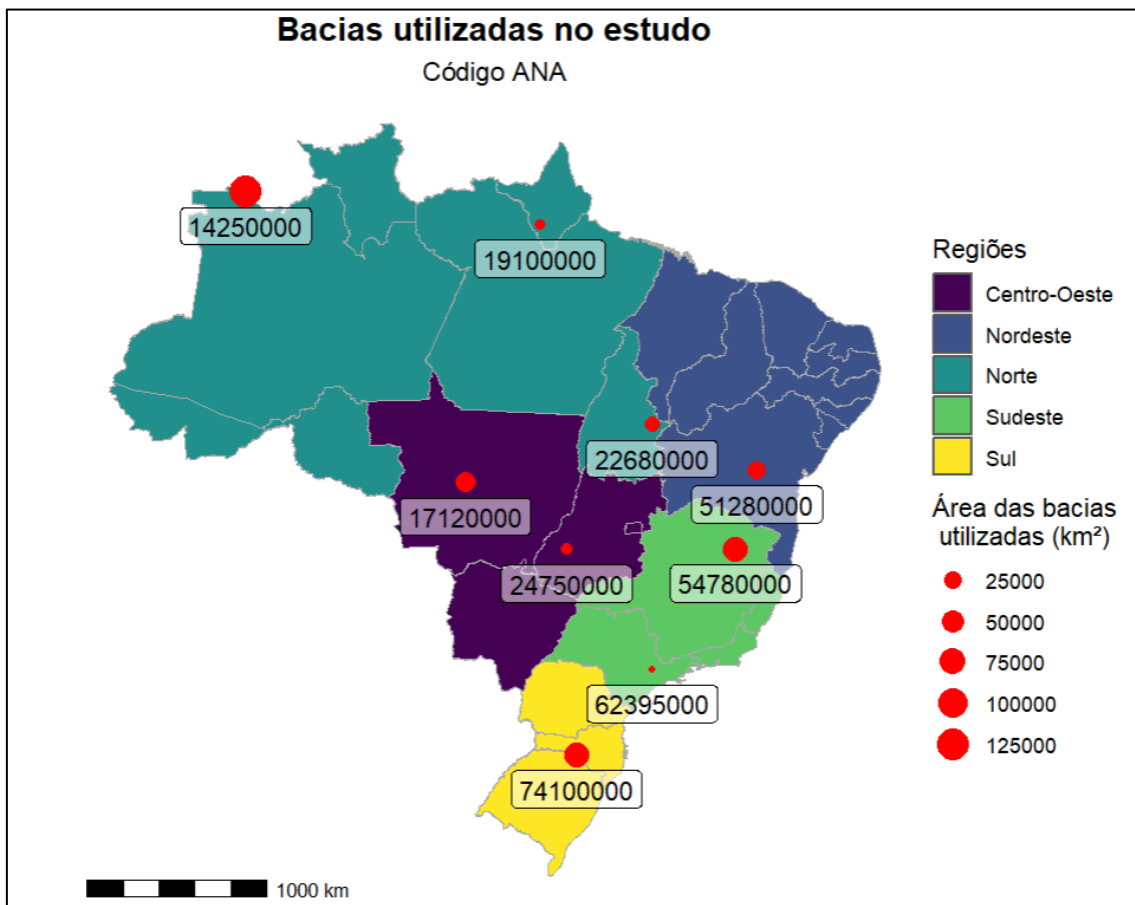


Tabela 2 – Características das bacias utilizadas.

Código ANA	Área (km ²)	Taxa de mudança	Q98 (mm/dia)	Q90 (mm/dia)	Q10 (mm/dia)	Q2 (mm/dia)
14250000	127541	0,018	1,250	2,257	9,834	11,40
17120000	37021	0,016	0,802	0,943	2,774	3,370
19100000	4443	0,053	0,150	0,291	2,840	4,555
22680000	17074	0,100	0,702	0,791	2,301	3,837
24750000	6429	0,186	0,247	0,332	3,827	7,866
51280000	23335	0,173	0,012	0,024	0,472	1,354
54780000	63676	0,149	0,055	0,095	0,926	2,553
62395000	617	0,308	0,373	0,523	2,473	5,586
74100000	62088	0,224	0,373	0,597	5,196	12,28

3.4 Resultados

Inicialmente são apresentados os resultados da comparação entre o MOCOM* e MOSCDE, variando a forma de compor e ordenar a frente de Pareto, da seguinte forma:

1. Frente de Pareto formada pelas soluções que dominam as demais em maior número de objetivos;
2. Conceito de dominância formal com ordenação pela distância da multidão;
3. Conceito de dominância formal com ordenação pelas soluções mais próximas de cada ponto de referência;

Incorporou-se os resultados do algoritmo SCE-UA de único objetivo, no qual este otimiza somente a métrica do NSE. O número de estruturas no qual se obteve um conjunto de parâmetros que resultou em um erro máximo de 5% em relação as assinaturas hidrológicas são apresentadas na Figura 6, para cada um dos algoritmos e em função do número de simulações. O número total de estruturas selecionadas, por bacia, é sumariado na Tabela 3. O Algoritmo de melhor desempenho em cada bacia foi destacada em negrito nesta tabela.

Figura 6 – Número de estruturas hidrológicamente consistentes encontradas na calibração.

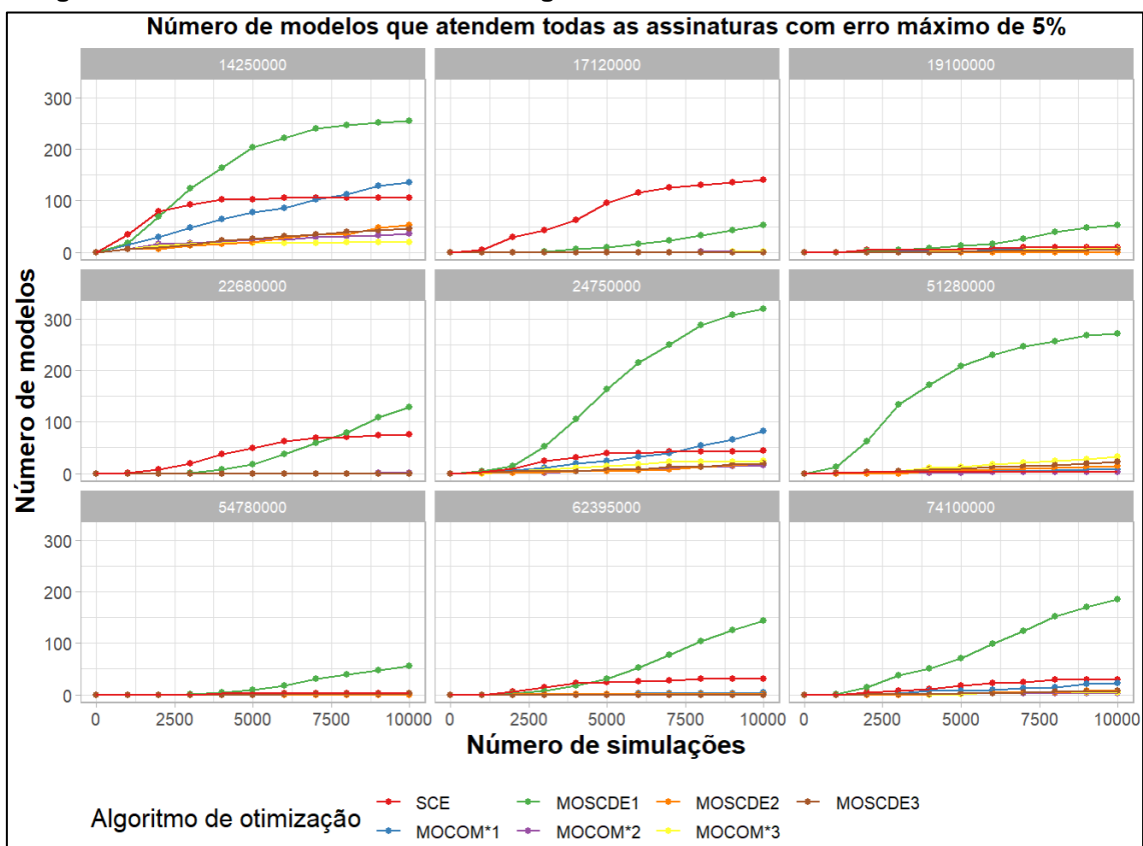


Tabela 3 – Número de estruturas por cada algoritmo de otimização em cada bacia.

Algoritmo	Bacias								
	14250000	17120000	19100000	22680000	24750000	51280000	54780000	62395000	74100000
SCE	105	140	10	76	44	3	2	30	29
MOCOM*1	135	0	6	0	83	7	0	5	23

Algoritmo	Bacias								
	14250000	17120000	19100000	22680000	24750000	51280000	54780000	62395000	74100000
MOCOM*2	36	1	0	1	16	2	1	0	3
MOCOM*3	20	1	6	0	25	32	0	1	4
MOSCDE1	254	53	52	129	320	272	55	143	185
MOSCDE2	52	0	0	0	20	14	0	1	7
MOSCDE3	46	0	5	0	20	22	2	0	7

Notavelmente, a maioria dos algoritmos de otimização multiobjetivo apresentam um desempenho inferior ao do SCE de objetivo único, mesmo que esse não busque a consistência das assinaturas. Os resultados só passam a ser aceitáveis quando se altera o conceito de dominância de uma solução sobre a outra (MOCOM*1 e MOSCDE1).

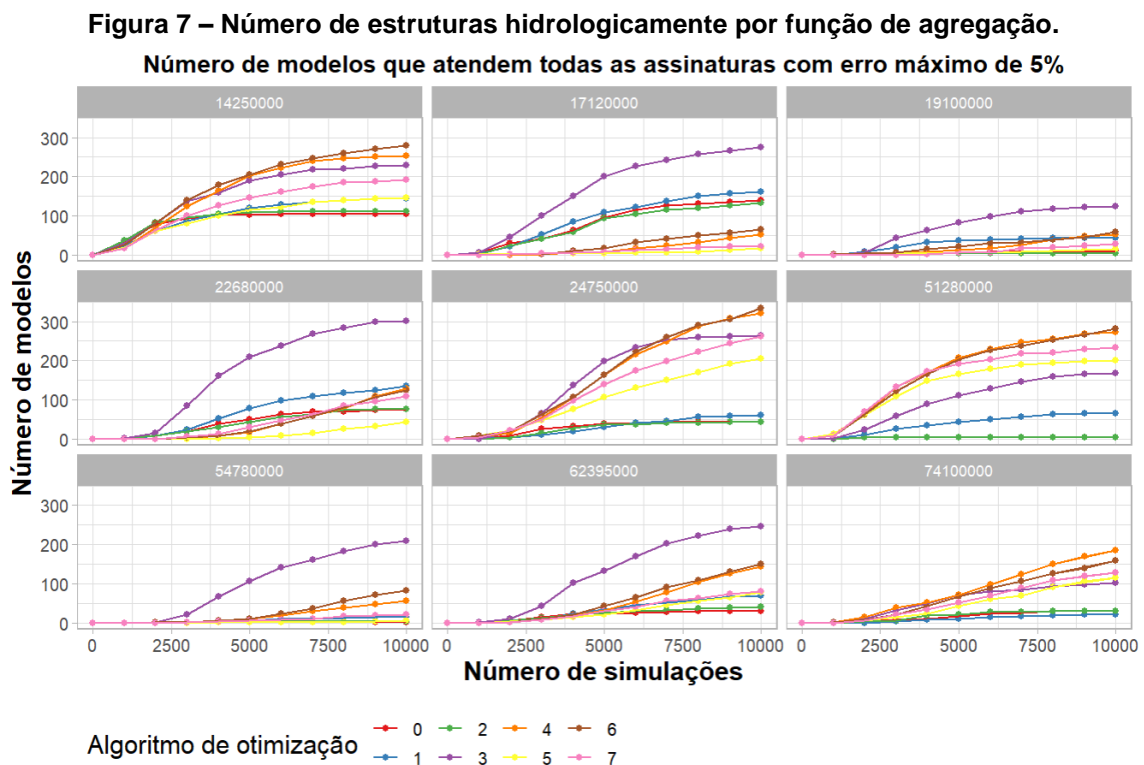
Nessas 9 bacias, o conceito de dominância formal não se aplica em uma calibração multiobjetivo de modelos hidrológicos envolvendo assinaturas hidrológicas e métricas temporais (MOCOM*2, MOCOM*3, MOSCDE2 e MOSCDE3). Mesmo alterando a ordenação nas frentes de Pareto, esses algoritmos são inferiores ao SCE de objetivo único na maioria dos casos. Nesses algoritmos, ao encontrar alguma solução muito boa para um determinado objetivo, por exemplo o Q2, ela tende a se perpetuar na frente de Pareto ótima mesmo que todos os outros objetivos estejam muito distantes do desejado. Uma solução que simule sempre a mesma vazão (Q2, por exemplo), atinge esse objetivo de forma ótima, enquanto os demais tem desempenho ruim. Mesmo assim dificilmente essa solução irá sair da frente de Pareto ótima, deteriorando a evolução dos complexos.

Isso sinaliza que a métrica de desempenho não deve ser tratada como conflitante em relação as assinaturas hidrológicas e sim de maneira colaborativa. Quando a estratégia é alterada para privilegiar soluções que superem as outras em um maior número de objetivos (MOCOM*1), os resultados começam a ser comparáveis com o SCE de objetivo único. Mas é só quando os complexos são evoluídos por evolução diferencial (MOSCDE1) é que os resultados são, em sua maioria, superiores ao SCE. A evolução diferencial se mostrou uma melhor abordagem nesse tipo de problema multiobjetivo. Vale a pena investigar seu desempenho ao empregá-la no SCE de objetivo único.

Dado o desempenho superior do MOSCDE1 ele foi empregado adiante no estudo. As alternativas testadas, sumariadas anteriormente na Tabela 1, estão detalhadas a seguir:

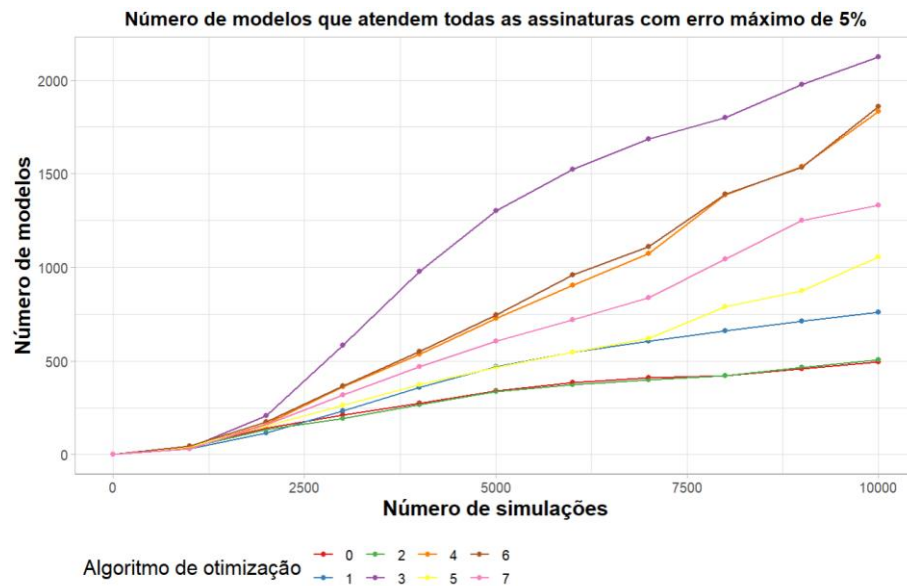
0. SCE-UA: NSE;
1. SCE-UA: pontuação binária nas assinaturas somado na NSE;
2. SCE-UA: somatório da pontuação sigmoide em cada assinatura somado na NSE;
3. SCE-UA: pontuação sigmoide sobre a pior assinatura simulada somado na NSE;
4. MOSCDE1- 6 objetivos: NSE e as 5 assinaturas sem transformação;
5. MOSCDE1- 2 objetivos: NSE e máxima diferença relativa das assinaturas
6. MOSCDE1- 6 objetivos NSE e pontuação binária sobre as 5 assinaturas;
7. MOSCDE1- 6 objetivos: NSE e pontuação contínua sobre as 5 assinaturas; e

Os resultados, em função do número de simulações e de cada algoritmo utilizado são apresentados nos gráficos da Figura 7, por bacia avaliada.



Por sua vez, a soma de estruturas, dentro do limite estabelecido, pode ser visualizada no gráfico da Figura 8. Nesse caso, todos os algoritmos tem desempenho superior ao SCE-UA otimizando somente o NSE, exceto a alternativa que otimiza o NSE mais o somatório da função sigmoide sobre a diferenças relativas das 5 assinaturas (alternativa 2), no qual houve um empate.

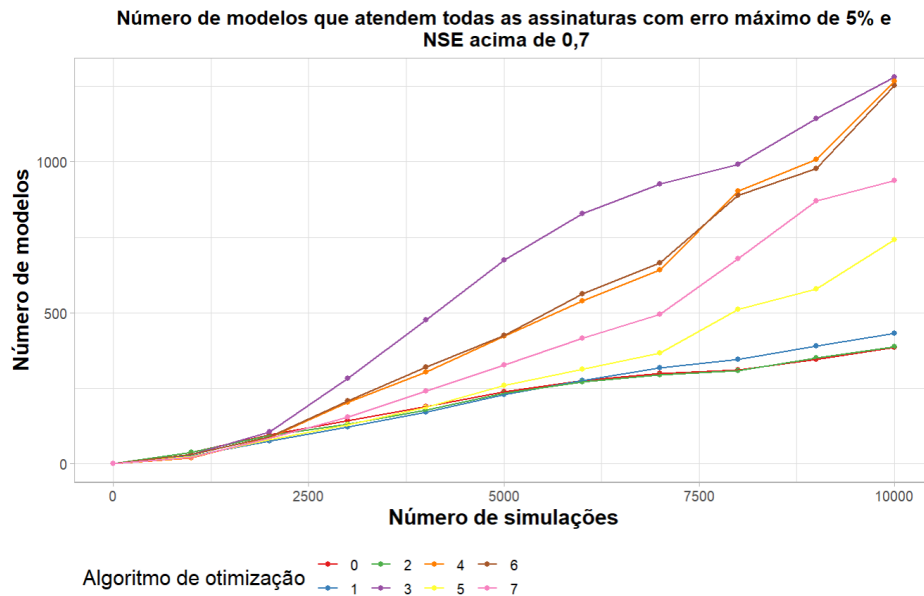
Figura 8 – Soma das estruturas hidrológicamente consistentes em todas as bacias para cada algoritmo.



Considerar cada assinatura como uma função binária somado ao NSE tem desempenho levemente superior, equivalente à abordagem com dois objetivos (NSE e máxima diferença relativa das assinaturas). Na abordagem multiobjetivo, não parece ter influência utilizar uma função binária ou não sobre as assinaturas, mas utilizar a função contínua reduziu o desempenho. Conclui-se que, nesse caso, o MOSCDE1 não obteve vantagens de qualquer função que agregue ou restrinja os objetivos das assinaturas.

Por outro lado, utilizar uma otimização de objetivo único somando o NSE com a máxima diferença relativa das assinaturas teve um desempenho consideravelmente superior aos demais, especialmente até 5.000 simulações. Essa superioridade sobre as demais alternativas é menos expressiva quando se considera somente as estruturas que, além de erro máximo de 5% nas assinaturas, tenham NSE acima de 0,7, mas ainda relevante, conforme pode ser observado na Figura 9.

Figura 9 – Soma das estruturas hidrológicamente consistentes e com NSE acima de 0,7 em todas as bacias para cada algoritmo



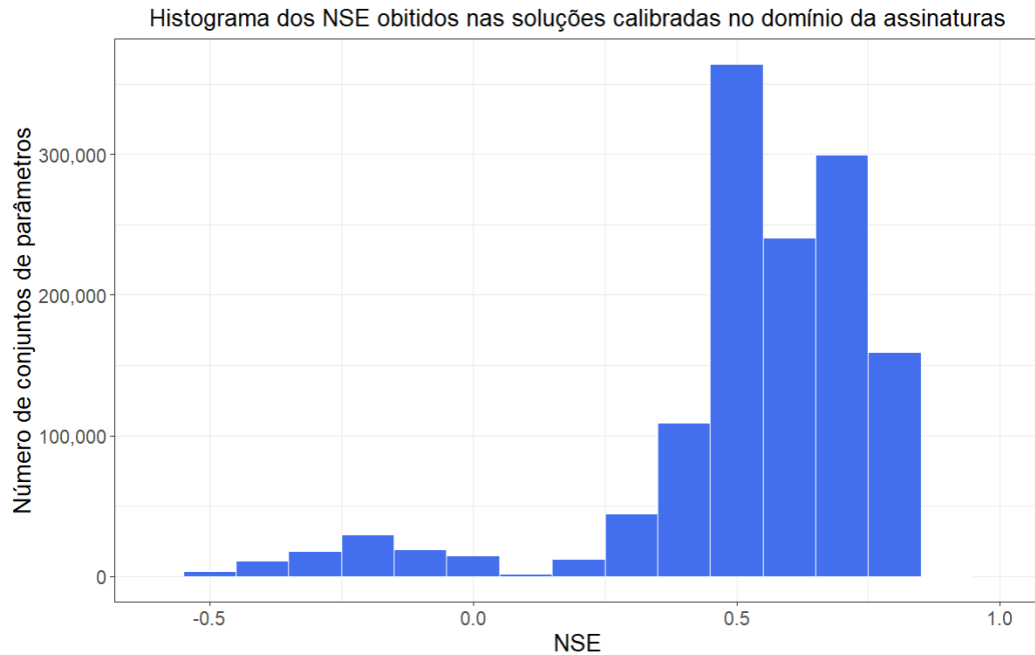
Esses resultados contradizem Efstratiadis e Koutsoyiannis (2010), uma vez que uma função agregadora em um único objetivo foi àquela que conseguiu o melhor compromisso entre os objetivos, enquanto a abordagem de dominância teve que ser adaptada para evitar que o algoritmo multiobjetivo se perdesse em uma região extrema da frente de Pareto.

Por outro lado, o uso de uma função contínua em detrimento de uma binária diminuiu o desempenho, conforme relatado por Shafii e Tolson (2015), mesmo que os limites da função tenham sido diferentes.

Por fim, utilizou-se o algoritmo de melhor desempenho (de número 3), para calibrar os parâmetros das 624 estruturas somente no domínio das assinaturas. Calibrou-se os parâmetros nas 9 bacias somente utilizando as assinaturas hidrológicas, com a finalidade de analisar a métrica do tempo (NSE). Ou seja, fornecendo ao algoritmo de otimização somente as diferenças relativas das assinaturas hidrológicas, qual seria o desempenho do NSE.

Ao todo, em torno de 7% das soluções simuladas no algoritmo de calibração estavam dentro dos limites para considerá-las hidrológicamente consistentes. O histograma de NSE dessas soluções é apresentado na Figura 10.

Figura 10 – Histograma da métrica de desempenho para as soluções hidrológicamente consistentes calibradas somente no domínio das assinaturas



É possível perceber que a maioria das soluções hidrológicamente consistentes apresentam um bom desempenho na métrica do tempo, conforme Fenicia et al., (2018).

3.5 Discussão

A seguir as perguntas apresentadas na introdução são exploradas.

1. Qual o desempenho das estratégias multiobjetivos com assinaturas hidrológicas frente a calibração de objetivo único no domínio do tempo?

A primeira parte do estudo mostra que, a depender da abordagem, transformar o problema em multiobjetivo pode deteriorar até mesmo as métricas que não são consideradas em uma otimização objetivo único. A ordenação da frente de Pareto na forma empregada pelos algoritmos NSGA-II e III não se apresentaram adequadas na calibração dos parâmetros dos modelos hidrológicos. O SCE-UA, na sua forma tradicional, conseguiu melhores resultados na consistência das assinaturas, mesmo sem levar em consideração essas métricas, ao se comparar com essas abordagens multiobjetivo.

As assinaturas hidrológicas não podem ser tratadas como independentes. Em problemas multiobjetivos tradicionais, que envolvam, por exemplo, qualidade e custo,

encontrar uma frente de Pareto ótima em termos de custo, benefício e custo-benefício, é relevante para a tomada de decisão. Em um problema multiobjetivo de calibração dos parâmetros de modelos hidrológicos, como o deste estudo, encontrar soluções que somente atende uma assinatura e erra demasiadamente as demais é pouco relevante.

Ao se adaptar a forma de escolher as melhores soluções, alterando a forma que os melhores indivíduos (ou soluções ou conjunto de parâmetros) de uma geração são levados diretamente para a próxima geração, garantiu-se que a qualidade da solução não degrade ao longo das iterações. Em síntese, a ideia de preservar as melhores soluções encontradas, garantindo que a busca não perca as melhores soluções durante o processo de otimização, não deve levar em conta objetivos individuais, e sim uma combinação deles.

É válido destacar também o ganho de se evoluir as soluções utilizando a evolução diferencial (DE) em detrimento do algoritmo genético (GA). A mutação do GA é estocástica e não direcionada, enquanto na DE é baseada na diferença entre os vetores das soluções, tornando-a direcionada e potencialmente mais eficaz na exploração do espaço de solução no âmbito deste estudo. A recombinação em DE tende a ter um maior efeito global, considerando a diferença entre pares aleatórios de soluções, enquanto o GA é mais susceptível a ótimos locais, operando em pontos específicos das soluções.

2. Qual a melhor abordagem multiobjetivo para otimizar tanto as assinaturas hidrológicas quanto a métrica de desempenho (NSE)?

Agregar todos os objetivos em um único pode ser a melhor abordagem na calibração multiobjetivo dos parâmetros de modelos hidrológicos, mas a forma que essa combinação é realizada altera significativamente o desempenho do algoritmo. O sucesso dessa abordagem é promovido pela simplificação dos objetivos. Permitir que a busca seja realizada somente utilizando o NSE e a pior assinatura evitou que o algoritmo se perdesse no espaço de soluções, em contraste quando se tenta aglutinar todos os 6 objetivos.

De certa forma, parte dos resultados são coerentes com Efstratiadis e Koutsoyiannis (2010). Agregar todos os objetivos em um único escalar apresentaram praticamente o mesmo resultado que o SCE-UA de único objetivo. A solução agregadora que se destaca é aquela que ignora a maioria dos objetivos, e se concentra no NSE e na pior assinatura hidrológica da solução. Esse o ponto que até então não havia sido avaliado em nenhum estudo até então consultado.

Se por um lado dar um foco ao algoritmo de busca foi benéfico, essas funções de combinação pioraram o desempenho ao se utilizar uma abordagem multiobjetivo com a frente de Pareto adaptada. A alteração na estratégia da formação da frente de Pareto já foi suficiente para direcionar o algoritmo no espaço de soluções.

Essas percepções reforçam a necessidade de se avaliar os algoritmos e heurísticas de otimização na calibração de modelos hidrológicos, especialmente multiobjetivos, além de entender os motivos que levam ou sucesso (ou insucesso) em um modelo hidrológico. O desempenho ruim na calibração de um modelo pode estar relacionado com uma deficiência no algoritmo de calibração.

3. Como são os resultados do NSE quando somente as assinaturas hidrológicas são utilizadas na calibração?

Em um contexto de bacias não monitoradas, não se teria dados observados para se otimizar a métrica de desempenho (o NSE neste caso). Por outro lado, as assinaturas hidrológicas poderiam ser regionalizadas. Assumindo que essas assinaturas poderiam ser previamente determinadas, o que se pretendeu investigar foi se a calibração somente as utilizando seria suficiente para uma aderência com os dados observados.

Fornecer as expectativas de baixa (Q98 e Q90) e alta (Q2 e Q10) vazão, além de um indicador do comportamento esperado entre a transição dos patamares (Taxa de Mudança) foi suficiente para que a calibração multiobjetivo encontrasse conjuntos de parâmetros adequados. Quando erro máximo de 5% nessas assinaturas é alcançado, 60% das soluções atingem o NSE acima de 0,5, e quase 90% são acima de 0,4.

Estudos futuros podem explorar melhor essas percepções, no sentido de se avaliar assinaturas regionalizadas e quais conseguiriam aprimorar a métrica de desempenho.

3.6 Conclusões

Sem adaptar o algoritmo de otimização para o contexto de calibração dos parâmetros de modelos hidrológicos, os algoritmos multiobjetivos tem desempenho pior que um algoritmo de único objetivo no domínio do tempo. Em síntese, mesmo considerando somente o NSE, o SCE-UA consegue atingir mais facilmente erros relativos menores das assinaturas hidrológicas, em relação aos algoritmos multiobjetivo que também levam em consideração essas outras métricas em conjunto. É necessário adaptar o conceito de dominância de uma solução sobre a outra para que uma abordagem multiobjetivo apresente melhores resultados. Os resultados mostraram que não se deve considerar a métrica do tempo e das assinaturas hidrológicas de forma independente, como é realizado comumente na abordagem de muitos objetivos (ALTHOFF; RODRIGUES, 2021; HERNANDEZ-SUAREZ et al., 2018; SHAFII; TOLSON, 2015).

Há de se destacar também a utilização da evolução diferencial junto ao SCE-UA. A evolução diferencial tem a capacidade de evitar mínimos locais e continuar a procurar por melhores soluções. A operação de mutação é baseada na diferença de dois indivíduos selecionados aleatoriamente, enquanto no algoritmo genético, a mutação é um pequeno ajuste aleatório. Essa diferença faz com que DE seja mais direcionado do que GA.

Na mesma direção, o problema da calibração dos parâmetros de modelos hidrológicos multiobjetivo com assinaturas hidrológicas pode ser simplificado em um problema de objetivo único que agregue todos os objetivos em um, desde que se dê um foco nos objetivos. É possível que outros algoritmos, além do SCE-UA, tenham bons resultados ao se aglutinar vários objetivos em um único.

Calibrar somente no domínio das assinaturas conseguiria apresentar métricas aceitáveis no domínio do tempo. Esse resultado pode ser de grande valia em um contexto de bacias sem dados observados. As assinaturas hidrológicas podem ser

regionalizadas e modelos poderiam ser calibrados em bacias com monitoramento insuficiente ou inexistente.

Por fim, este estudo direciona a calibração multiobjetivo na calibração dos parâmetros de modelos hidrológicos, ao concluir que a melhor estratégia seria avaliar a pior assinatura hidrológica de interesse em conjunto com a métrica de desempenho, no caso o NSE. Muitos algoritmos de otimização são projetados para problemas de único objetivo. Dessa forma, usar uma função de pontuação agregadora com foco na maior diferença relativa das assinaturas pode permitir que se utilize uma ampla variedade de ferramentas e técnicas existentes. Além disso, essa solução produz uma única “melhor” solução em vez de uma série de soluções na frente de Pareto, o que somente adiará a decisão final de qual solução utilizar.

Ademais, definiu-se as prioridades na busca das soluções, de forma que estudos futuros não precisam definir pesos para diferentes objetivos. Utilizar a máxima diferença relativa das assinaturas hidrológicas se mostrou muito eficaz, contribuindo também para melhores métricas do NSE.

4 PREDIÇÃO DE ASSINATURAS HIDROLÓGICAS E CLASSIFICAÇÃO DE BACIAS TROPICAIS

“[...] the single best recommendation on the issue of rainfall-runoff modeling of ungauged catchments may be to install a stream gauge.”

(BLÖSCHL, 2006)

4.1 Introdução

A maioria das bacias do mundo são mal ou não monitoradas (SINGH; DEVI, 2022). Um dos grandes desafios na hidrologia diz respeito a como extrapolar informações hidrológicas de uma área monitorada a outra não monitorada (KRATZERT et al., 2019a). A transferência do conhecimento adquirido em uma região a outra pode ser denominada regionalização (WANG; CAO; FENG, 2021). Apesar de existirem diversas classificações dessas técnicas (GUO et al., 2021a; POOL; VIS; SEIBERT, 2021; QI et al., 2020; TEGEGNE; KIM, 2018), pode-se categorizá-las como baseadas em similaridade, física e/ou decorrente da proximidade espacial, ou em regressão, em que se busca expressá-la matematicamente.

Independentemente da categoria, assume-se como verdadeira a hipótese de que exista uma relação entre atributos de bacias similares e suas características hidrológicas. A similaridade hidrológica é mensurada a partir da resposta da bacia à diferentes forçantes meteorológicas. A variabilidade espaço-temporal dessas forçantes e a complexidade em monitorá-las dificulta o estabelecimento dessas relações. Como diferentes características hidrológicas podem ser descritas por meio de assinaturas, caso selecionado um conjunto adequado parece plausível admitir que ele seja capaz de expressar o grau de similaridade entre bacias. (CHOUBIN et al., 2019; GUO et al., 2021a).

Assim, sob a conjectura da similaridade hidrológica, uma série de estudos em que parâmetros de modelos hidrológicos foram estimados empregaram assinaturas hidrológicas para, por exemplo, calibrá-los exclusivamente a partir de um conjunto de

assinaturas (KAVETSKI et al., 2018) ou, no caso de bacias sem monitoramento, identificar bacias doadoras desses parâmetros (TEGEGNE; KIM, 2018) ou condicionar sua estimativa (WESTERBERG et al., 2014; WU et al., 2022)). Ainda em relação a bacias sem monitoramento, destacam-se estudos em que a proposta foi regionalizar as próprias assinaturas hidrológicas, seja para inferência do comportamento hidrológico (WORLAND et al., 2019; ZHANG et al., 2018c) ou para estimativa dos parâmetros e incertezas a eles associados (DAL MOLIN et al., 2023; MATOS, 2021).

A distinção em se utilizar assinaturas hidrológicas na modelagem é que elas permitem destacar diferentes características hidrológicas, inclusive em escalas temporais distintas, serem regionalizadas mais facilmente se comparadas com a regionalização de atributos físicos (GUO et al., 2021a) e a possibilidade de serem estimadas por produtos de sensoriamento remoto (KUNNATH-POOVAKKA et al., 2018). A princípio, regionalizá-las para fins de estimativa de parâmetros a partir de atributos físicos parece um caminho mais promissor do que buscar relações entre esses atributos e os parâmetros desses modelos, já que a presunção de similaridade independe das estruturas conceituais admitidas. A seleção de assinaturas adequadas a esse propósito, no entanto, é uma tarefa desafiadora (ADDOR et al., 2018; MCMILLAN, 2021). A integração de estimativas oriundas de diferentes bases de dados públicas também pode ser conflitante, visto que há variações de cálculo dessas assinaturas descritas na literatura (BECK et al., 2016).

A estimativa de assinaturas hidrológicas a partir dos atributos físicos é um problema ainda em aberto (CHOUBIN et al., 2019). Não obstante, podem-se distinguir três abordagens: (i) interpolação (ZHANG et al., 2014); (ii) regressão (TYRALIS et al., 2021); e (iii) modelagem hidrológica (WESTERBERG et al., 2016). Dentre essas categorias, os algoritmos de regressão baseados em aprendizado de máquina têm se destacado (WU et al., 2022), em especial a técnica *Random Forest* (ADDOR et al., 2018; DESAI; OUARDA, 2021; PRIETO et al., 2019; ZHANG et al., 2018c). Em consonância com estudos em outras áreas do conhecimento (ALJAMAAN; ALAZBA, 2020; ANAND; VELU; WHIG, 2022; GUNTURI; SARKAR, 2021) seria pertinente estender essa comparação aos algoritmos baseados em *Gradient Boosting* (BENTÉJAC; CSÖRGŐ; MARTÍNEZ-MUÑOZ, 2021) e avaliar um conjunto maior de

assinaturas, dada a prevalência de estudos baseados em bacias localizadas em climas não tropicais.

Nesse contexto, o primeiro experimento do presente estudo se propõe a avaliar diversos algoritmos de regressão para estimativa de uma gama de assinaturas hidrológicas a partir de atributos físicos em 735 bacias presentes na base CABra (ALMAGRO et al., 2021). As perguntas que guiaram o experimento são:

- Quão bem as assinaturas hidrológicas podem ser preditas utilizando-se atributos físicos em bacias tropicais?
- Quais são as mais bem estimadas?
- Quais atributos físicos da bacia são mais relevantes para esse propósito? e
- Algum algoritmo se destaca?

Já o segundo experimento vale-se somente das assinaturas mais bem preditas e busca identificar agrupamentos que sintetizem o comportamento espacial conjunto dessas assinaturas hidrológicas.

Isso posto, é pertinente destacar que a identificação de regiões homogêneas é um objetivo usual em estudos de regionalização. Admite-se que bacias contidas em uma mesma região apresentam comportamento hidrológico similar em relação a variável regionalizada. Caso essas regiões sejam construídas a partir da análise de múltiplas variáveis é imediata a noção de se classificar bacias (HE; BÁRDOSSY; ZEHE, 2011; TAREK et al., 2021).

Não há um consenso quanto a um sistema único de classificação de bacias hidrográficas (WAGENER et al., 2007). Identificar e categorizar as funções dominantes reveladas por meio de um conjunto de características da resposta hidrológica é uma estratégia para quantificar o grau de semelhança que pode existir entre as bacias (LYON; TROCH, 2010). Compreender como e por que determinado comportamento ocorre em uma bacia, em última análise, lançaria luz às razões da similaridade (ou dissimilaridade) entre elas (DOOGE, 1986).

Quando não se sabe a classe a qual cada bacia deve pertencer, o problema de classificação se torna um problema de agrupamento (LI; SHEPPERD; GUO, 2020).

Ademais, não se tem qualquer informação prévia sobre o número de agrupamentos (ou *clusters*) de bacias desejáveis, o que é um problema comum ao lidar com conjuntos de dados do mundo real (LIU; WU; SHEN, 2011). Em estudos que o avaliaram como, por exemplo, Smith et al. (2015), Kuentz et al. (2017), Jehn et al. (2020) e Jaffrés et al. (2022): (i) não houve uma congruência em relação a um número ótimo de clusters; (ii) o número de agrupamentos variou entre 5 e 88 e a redução do número de agrupamentos implicava num aumento do número de bacias não classificadas; e (iii) um dos grupos criados reunia um conjunto heterogêneo de bacias. Nesses estudos assinaturas hidrológicas e atributos físicos foram empregados para construção dos agrupamentos.

Diante desse quadro, o segundo experimento norteou-se pela estimativa e interpretação dos agrupamentos obtidos a partir da análise integrada das assinaturas hidrológicas descritas na base de dados CABra. Sua relação com o experimento anterior, portanto, limita-se ao emprego das variáveis mais bem preditas pelos algoritmos avaliados. As perguntas que o orientaram são:

- A variação no número de agrupamentos obtidos a partir de diferentes técnicas é significativa no presente caso?
- Agrupamentos baseados somente nas assinaturas hidrológicas são espacialmente consistentes? e
- É evidente algum controle físico ou climático sobre os agrupamentos obtidos?

4.2 Referencial teórico

A Mineração de Dados refere-se ao processo de extrair padrões, relações ou informações de interesse de grandes volumes de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). A cerne da mineração de dados não reside apenas na análise estatística ou computação, mas na transformação desses extensos conjuntos de dados em conhecimento compreensível e utilizável, auxiliando a tomada de decisões em variados setores, desde negócios até ciências da saúde (HAND; MANNILA; SMYTH, 2001).

Dentre os principais problemas endereçados pela mineração de dados, três se destacam pela sua frequência e importância: classificação/regressão, agrupamento e

associação. A classificação, no contexto da mineração, objetiva categorizar entradas desconhecidas em classes pré-definidas, como determinar se uma transação é fraudulenta com base em um histórico (WITTEN; FRANK; HALL, 2011). A regressão busca prever valores contínuos a partir de variáveis independentes, como a previsão de vendas de um produto com base em tendências anteriores (HASTIE; TIBSHIRANI; FRIEDMAN, 2016). Diferentemente desses, o agrupamento, ou *clustering*, trabalha sem categorias pré-definidas, visando identificar e agrupar pontos de dados semelhantes com base em suas características intrínsecas, ajudando a revelar estruturas ou segmentos não evidentes em conjuntos de dados (JAIN, 2010). Os algoritmos de associação buscam encontrar regras que podem indicar a coocorrência de itens em um determinado conjunto (AGRAWAL; IMIELIŃSKI; SWAMI, 1993).

O Aprendizado de Máquina (ML, do inglês *Machine Learning*) concentra-se na construção de sistemas que podem aprender a partir de dados. Em vez de serem explicitamente programados para realizar uma tarefa, esses sistemas são treinados usando grandes conjuntos de dados e algoritmos que dão a eles a capacidade de melhorar seu desempenho por meio da experiência. O foco em ML geralmente está em criar modelos de previsão, enquanto a mineração de dados se concentra na descoberta de padrões e percepções nos dados. Ainda assim, a distinção entre ML e mineração de dados nem sempre é nítida, e há muita sobreposição entre os dois campos, não sendo o foco deste estudo aprofundar na teoria sobre essa temática. Nesse sentido, o trabalho de Hastie et al. (2016) e o livro de Witten et al. (2011) conseguem fornecer um aspecto teórico e prático das duas áreas. Esses autores também são os desenvolvedores do WEKA (*Waikato Environment for Knowledge Analysis*).

A apresentação dos algoritmos a seguir limita-se a uma descrição sumária e sua seleção ocorreu a partir de sua relação com estudos correlatos.

4.2.1 Regressão

Na mineração de dados, os algoritmos de regressão são ferramentas que visam prever um valor de uma variável contínua com base em variáveis independentes. Esses algoritmos, foram adaptados e aprimorados com a evolução das técnicas de aprendizado de máquina (WITTEN; FRANK; HALL, 2011).

Um dos métodos mais bem-sucedidos nesse aspecto é o *Random Forest* (RF; BREIMAN, 2001). Ele pode ser utilizado tanto para classificação quanto para regressão. É uma combinação de árvores de decisão, treinadas com método de "bagging" (*Bootstrap Aggregating*). A ideia principal por trás do RF é criar várias árvores de decisão durante o treinamento e produzir a seleção da classe majoritária no caso da classificação, ou uma média das saídas, no caso da regressão, de cada árvore (BREIMAN, 1996).

Em hidrologia o RF foi utilizado, por exemplo, em Talukdar et al. (2023) e Araza et al. (2020), para prever vazões, em Addor et al. (2018), para inferir assinaturas hidrológicas, e em Desai e Ouarda (2021) e Prieto et al. (2019), para estimar os parâmetros do modelo hidrológico utilizando as assinaturas.

Os algoritmos de *boosting* são uma família de métodos de ML que visam melhorar o desempenho de um modelo por meio da combinação ponderada de modelos ligeiramente melhores que aleatórios. O conceito central do *boosting* é construir sequencialmente esses modelos, em que cada novo modelo tenta corrigir os erros cometidos pelos modelos anteriores (BISHOP, 2006; FREUND; SCHAPIRE, 1999).

Um exemplo dessa classe de algoritmos é o *Gradient Boosting*. Enquanto o RF constrói cada árvore de forma independente, o *Gradient Boosting* constrói árvores sequencialmente, em que cada árvore nova tenta corrigir os erros da árvore anterior. A ideia do "boosting" é aumentar a performance de modelos fracos ao construí-los em sequência, focando nas instâncias que foram mal classificadas ou mal previstas nas etapas anteriores (FRIEDMAN, 2001).

A partir desse método, vários outros vem surgindo como: o *eXtreme Gradient Boosting* (XGB; CHEN; GUESTRIN, 2016), o *Light Gradient Boosting Machine* (LightGBM; KE et al., 2017) e o *Categorical Boosting* (CatBoost; PROKHORENKOVA et al., 2018), com desempenho notável (BENTÉJAC; CSÖRGŐ; MARTÍNEZ-MUÑOZ, 2021). Esses métodos propõem variantes ao algoritmo *Gradient Boosting* para melhorar a velocidade de treinamento e sua generalidade.

Ji et al. (2021) compararam o XGB com redes de memória de longo e curto prazo (LSTM, do inglês *Long Short-Term Memory*) e regressão de vetor suporte (SVR, do

inglês *Support Vector Regression*) na predição de vazões em bacias hidrográficas glaciais com dados escassos e observaram que, apesar de uma ligeira superioridade das redes LSTM, os resultados obtidos pelo XGB foram bastantes similares. O XGB também superou o SVR e o RF em um vasto estudo envolvendo 2.257 bacias em três continentes (LE et al., 2022). Em Zhang et al. (2019) o XGB também foi superior na predição de vazão comparado aos mesmo algoritmos de aprendizado de máquina (RF e SVR) e quatro modelos conceituais concentrados (SimHyd, Sacramento, SMAR e Tank). Em Szczepanek (2022) comparou-se os principais algoritmos baseados em *Gradient Boosting* (XGB, LightGBM e CatBoost) na estimativa de vazões em bacias montanhosas nos Cárpatos Poloneses, com superioridade do CatBoost e do LightGBM.

Além da predição de vazões, o CatBoost foi avaliado, por exemplo, para estimar a evapotranspiração de referência (HUANG et al., 2019; ZHANG; ZHAO; ZHENG, 2020) enquanto o XGB para estimativa da evaporação (EL BILALI et al., 2023)

4.2.2 Agrupamentos

Os algoritmos de agrupamento são uma das principais técnicas de aprendizado não supervisionada. Eles agrupam um conjunto de dados em grupos (ou *clusters*) almejando que os dados dentro de um mesmo cluster sejam mais semelhantes entre si do que aqueles em outros clusters (HASTIE; TIBSHIRANI; FRIEDMAN, 2016).

As técnicas de análise de agrupamento podem ser divididas em dois grupos principais: hierárquicos e particionais. Os algoritmos hierárquicos encontram recursivamente os agrupamentos a partir de duas estratégias: (i) começando com cada dado como seu próprio agrupamento e mesclando pares semelhantes, formando uma hierarquia de agrupamentos; ou (ii) começando com todos os pontos de dados em um único grupo e recursivamente dividindo cada cluster em grupos menores. Comparado aos algoritmos de agrupamento hierárquico, os algoritmos de agrupamento particionais encontram todos os agrupamentos simultaneamente como uma partição dos dados e não impõem uma estrutura hierárquica (JAIN, 2010).

Estudos recentes têm buscado agrupar as bacias hidrográficas a partir de suas assinaturas e atributos físicos. Por exemplo, Kuentz et al. (2017) usaram mais de

35.000 bacias hidrográficas na Europa e as classificaram, com base em 16 assinaturas hidrológicas e 35 atributos físicos, utilizando um algoritmo de agrupamento hierárquico baseado no *K-Means*. Eles foram capazes de encontrar dez agrupamentos distintos. No entanto, os autores utilizaram algumas das assinaturas identificadas como de baixa previsibilidade espacial por Addor et al. (2018). Além disso, um terço de suas bacias hidrográficas foi agregada em uma grande classe sem atributos distinguíveis.

Jehn et al. (2020) utilizaram o mesmo algoritmo para gerar 10 agrupamentos em 671 bacias do CAMELS (ADDOR et al., 2017) utilizando uma série de atributos da bacia e 6 assinaturas hidrológicas identificadas como de melhor previsibilidade espacial (ADDOR et al., 2018). Os resultados mostraram uma menor importância do solo e da geologia, o que pode decorrer da ausência de uma assinatura que representasse as baixas vazões (YAEGER et al., 2012). Houve também um número considerável de bacias não agrupadas. Ademais, o estudo indicou que uma classificação de bacias baseada apenas em atributos físicos só funcionaria em regiões onde o principal fator não fosse o clima. Algo ainda aparentemente inexplorado é agrupar as bacias com similaridade somente em suas assinaturas hidrológicas, como aqui proposto.

Complementarmente, Jaffrés et al. (2022) aplicaram o Fuzzy C-Means em 749 bacias na Austrália utilizando 40 atributos, sendo 16 assinaturas hidrológicas. Os resultados sugeriram 5, 8 ou 9 *clusters*. Houve uma semelhança entre a classificação climática modificada de Köppen para a Austrália e a classificação de 8 *clusters*, demonstrando que o clima é um controle importante nessa região.

4.3 Metodologia

Conforme mencionado, o estudo foi dividido em dois experimentos:

- No primeiro, avaliou-se a capacidade dos principais algoritmos de regressão de aprendizado por máquina em estimarem 163 indicadores hidrológicos com base em 52 atributos físicos das bacias; e
- Na sequência, tomou-se como base as assinaturas mais bem estimadas em cada categoria, de maneira a cobrir as principais características hidrológicas, e

agrupou-se as bacias avaliando-se técnicas/métricas para determinação do número de clusters.

4.3.1 Experimento 1

Todos os atributos físicos do banco de dados do CABra foram utilizados na predição das assinaturas hidrológicas. As 18 assinaturas já disponíveis nessa base de dados foram complementadas, resultando nas assinaturas descritas na Tabela 4. Alguns índices foram recalculados e substituídos, sendo apresentados a seguir.

No CABra, o índice do escoamento de base foi calculado usando um filtro digital recursivo proposto por Lyne e Hollick (1979). Neste trabalho utilizou-se também o filtro digital proposto por Arnold et al. (1995), no qual o escoamento direto (Q_D) em um determinado tempo t é calculado pela Equação (11):

$$Q_{D_t} = aQ_{D_{t-1}} + \frac{1+a}{2}(Q_t - Q_{t-1}) \quad (11)$$

em que Q é a vazão diária e a foi fixado em 0,925 com base no estudo realizado por Eckhardt (2008).

A elasticidade (e) da vazão em relação a precipitação e a evapotranspiração potencial foi calculada pela proposta de Sankarasubramanian et al. (2001), por meio da Equação (12).

$$e = \text{mediana} \left(\frac{Q_t - \bar{Q}}{P_t - \bar{P}} * \frac{\bar{Q}}{\bar{P}} \right) \quad (12)$$

em que Q_t é a vazão em um determinado passo de tempo, \bar{Q} a vazão média daquele passo de tempo e P_t pode ser a evapotranspiração potencial ou a precipitação, para escala diária, mensal e anual. Embora a base de dados empregada já forneça os valores da elasticidade da precipitação para diferentes bacias, em uma breve verificação não houve correspondência em relação aos resultados obtidos. A elasticidade calculada pela Equação (12) não apresenta boa previsibilidade, mas uma

simples inversão nos fatores resulta em uma melhora significativa de seu desempenho no experimento. Assim, apesar de não se ter identificado correspondência na literatura, optou-se pela manutenção da nomenclatura e o cômputo dos índices a partir da Equação (13).

$$e = \text{mediana} \left(\frac{Q_t - \bar{Q}}{P_t - \bar{P}} * \frac{\bar{P}}{\bar{Q}} \right) \quad (13)$$

Para cada um dos índices os seguintes algoritmos foram avaliados:

- *CatBoost*;
- *Gradient Boosting*;
- XGB;
- *LightGBM*;
- *Random Forest* (RF);
- SRV;
- *K* vizinhos mais próximos (KNN);
- Regressão linear múltipla (RLM).

Tabela 4 – Assinaturas hidrológicas utilizadas.

Índice	Categoria	Assinatura	Descrição
1-4	Distribuição	Estatísticas descritivas	Média ¹ , Mediana ¹ , Assimetria ¹ e Curtose ¹
5-103		Percentis	Percentis da curva de permanência entre 99 à 1%
104-115	Magnitude	Média Mensal	Média mensal de janeiro até dezembro ²
116		Média anual	Vazão média anual ¹
117-118		Média do verão e inverno	Vazões médias durante as estações do verão e inverno ³
119		Coefficiente de escoamento superficial	Razão entre a vazão média diária e a precipitação média diária ¹
120-122		Taxas de mudança	Taxa de subida, taxa de descida e taxa de mudança ¹
123-128		Elasticidade	Elasticidade da precipitação e da evapotranspiração, diária mensal e anual ⁴
129-133	Dinâmica	Razão entre quantis	Razão entre os quantis 90/10, 80/20, 75/25, 70/30 e 66/33 da curva de permanência ¹
134-138		Inclinação da curva de permanência	Inclinação da curva de permanência entre os quantis 10/90, 20/80, 25/75, 30/70 e 33/66 ¹
139-140		Escoamento de base	Índices do escoamento de base ^{5 e 6}
141		CV da vazão diária	Coefficiente de variação da vazão diária ⁶

Índice	Categoria	Assinatura	Descrição
142		CV das vazões baixas	Coefficiente de variação das vazões baixas ($0,2 * média$) ⁶
143		CV das vazões altas	Coefficiente de variação das vazões altas ($9 * média$) ⁶
144	Frequência	Frequência das vazões baixas	Frequência das vazões baixas ($0,2 * média$) ⁶
145		Frequência das vazões altas	Frequência das vazões altas ($9 * média$) ⁶
146		Frequência dos dias sem vazão	Frequência dos dias sem vazão ⁶
147-153	Duração	Mínimo anual	Mínimo anual das médias móveis de 1,3,7,15,30,60 e 90 dias ¹
154-160		Máximo anual	Máximo anual das médias móveis de 1,3,7,15,30,60 e 90 dias ¹
161		Duração das vazões baixas	Duração das vazões baixas ($0,2 * média$) ⁶
162		Duração das vazões altas	Duração das vazões altas ($9 * média$) ⁶
163	Tempo	Dia da vazão média	Dia em que a vazão acumulada atinge a metade da vazão anual ⁶

¹: (OLDEN; POFF, 2003)

²: (ARCHFIELD et al., 2014)

³: (ADDOR et al., 2018)

⁴: (SANKARASUBRAMANIAN; VOGEL; LIMBRUNNER, 2001)

⁵: Arnold et al. (1995)

⁶: (ALMAGRO et al., 2021)

Os cálculos foram avaliados considerando o ano hidrológico iniciando em primeiro de outubro. As médias móveis consideraram janelas rolantes, ou seja, a média dos últimos n dias.

Para validar os resultados foi utilizado a técnica de validação cruzada *k-fold* (KOHAVI, 1995). O *k-fold* consiste em dividir a base de dados de forma aleatória em k subconjuntos, sendo o k definido previamente, com aproximadamente a mesma quantidade de amostras em cada um deles. A cada iteração, treino e validação, um conjunto formado por k-1 subconjuntos são utilizados para treinamento e o subconjunto restante para validação, gerando um resultado da métrica para avaliação. Esse processo garante que cada subconjunto será utilizado para teste na avaliação do modelo, obtendo-se a métrica de avaliação para toda a base de dados. Para critérios de comparação utilizou-se k=10, conforme Addor et al. (2018), assim como a predição das assinaturas foi avaliada utilizando o coeficiente R².

Para atenuar qualquer efeito dos sorteios aleatórios, repetiu-se todo o procedimento 5 vezes. Para cada algoritmo e cada assinatura calculou-se a média dos R² obtidos em cada subconjunto do *k-fold* em cada um dos 5 procedimentos.

A importância de cada variável, em regressões baseadas em árvore, normalmente é baseada em quantas vezes um determinado modelo seleciona uma variável para divisão de um nó e o quanto ela é melhorada por causa da divisão (FRIEDMAN; MEULMAN, 2003). No entanto, Lundberg et al. (2019) apontaram que a importância de uma determinada variável pode diminuir mesmo que um modelo mude para depender mais dessa variável. Por isso, eles propuseram o SHAPley Additive exPlanation (SHAP) para calcular a importância de cada variável (LUNDBERG; LEE, 2017).

O SHAP se origina da teoria dos jogos e constrói um modelo de interpretação aditivo (ŠTRUMBELJ; KONONENKO, 2014), e todos os recursos (ou as variáveis) são considerados "contribuintes". O valor de Shapley é uma alocação de lucro justa entre muitas partes interessadas, dependendo de sua contribuição e foi derivado do nome do economista que o introduziu (SHAPLEY; ROTH, 1988).

A resposta da saída de um modelo de aprendizado de máquina à entrada pode ser obtida pelo cálculo do valor SHAP (ver LUNDBERG; LEE, 2017). No estudo da análise de características, o SHAP tem sido amplamente utilizado recentemente (por exemplo CHOWDHURY et al., 2022; LU et al., 2023; NOHARA et al., 2022; VEGA GARCÍA; AZNARTE, 2020). Na hidrologia o SHAP foi utilizado para entender os modelos de aprendizado por máquina na estimativa de vazões em Mushtaq et al. (2022) e Wang et al. (2022).

4.3.2 Experimento 2

As assinaturas mais bem estimadas dentro de cada categoria foram selecionadas para serem agrupadas. Entretanto, optou-se por remover aquelas que poderiam influenciar o agrupamento de bacias espacialmente próximas, como as vazões médias mensais ou o dia da vazão média. Com estações muito bem definidas e diferentes no norte e sul do país, o uso dessas assinaturas poderia conduzir a agrupamentos decorrentes exclusivamente da proximidade espacial. Embora esse seja um comportamento desejado espera-se que ele seja uma consequência do agrupamento de bacias por similaridade das assinaturas hidrológicas.

Ao se considerar um elevado número de assinaturas, mesmo limitando-as, um elevado número de assinaturas correlacionadas e, portanto, com informações redundantes, pode ser obtido. Com isso em mente, desenvolveu-se uma análise de componentes principais (PCA, do inglês *Principal Components Analysis*; JOLLIFFE, 2006) para reduzir a dimensionalidade do problema, por meio da determinação de componentes que são combinações lineares das variáveis originais. As componentes são escolhidas para representar a fração máxima possível da variabilidade contida nos dados originais (WILKS, 2019). O PCA foi utilizado, por exemplo, em Jaffés et al. (2022), Jehn et al. (2020) e Kuentz et al. (2017) para agrupar bacias similares e em Berhanu et al. (2015) na classificação dos regimes de cursos d'água.

Os componentes principais foram agrupados utilizando o algoritmo *K-Means*. Um dos grandes dilemas desse algoritmo é determinar o número de clusters (K) (LASHENG; YUQIANG, 2017). Para tentar resolver esse problema, as seguintes técnicas/métricas foram avaliadas para determinação do número de *clusters*:

- Método do Elbow (THORNDIKE, 1953): é baseado na observação de que, com o aumento do número de *clusters*, a variação total dentro do *cluster* (ou seja, a soma dos quadrados das distâncias dentro do *cluster*) diminui. A variação reduz mais incisivamente no início e então desacelera, resultando em um gráfico com forma de "cotovelo". O número de clusters é escolhido no "cotovelo" do gráfico;
- Coeficiente de Silhueta: esse método, proposto por Rousseeuw (1987), mede o quão bem um ponto de dados se encaixa no seu *cluster* comparado aos outros *clusters*. O coeficiente de silhueta varia de -1 a 1, em que valores elevados indicam que o registro foi bem agrupado. O número de *clusters* é escolhido como aquele que maximiza a média do coeficiente de silhueta para todos os registros do banco de dados;
- Índice Calinski-Harabasz: Este índice, proposto por Calinski e Harabasz (1974), é a razão entre a dispersão entre *clusters* e a dispersão dentro dos *clusters*, multiplicada por um fator dependente do número de agrupamentos e observações. Um valor maior do índice indica uma melhor separação dos *clusters*.

Definido o número de *clusters* eles foram analisados em termos das assinaturas hidrológicas e dos padrões espaciais climáticos, de umidade, biomas, aquíferos e

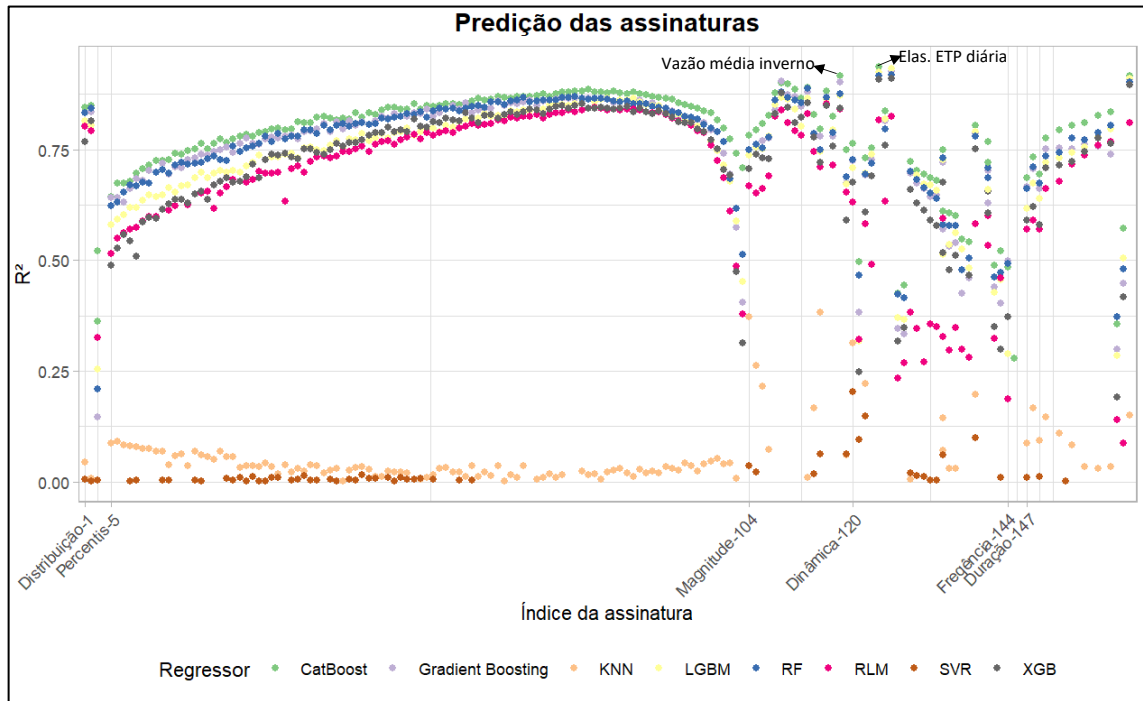
precipitação. Essas informações foram disponibilizadas pelo Instituto Brasileiro de Geografia e Estatística (IBGE), em resolução 1:250.000 para os biomas (IBGE, 2019) e 1:5.000.000 para as zonas climáticas e de umidade (IBGE, 2002). O sistema de aquíferos foi disponibilizado pela Agência Nacional de Água (ANA), em escala de 1:1.000.000 (ANA, 2022), enquanto os dados precipitação foram construídos pela krigagem dos dados disponibilizados pela ANA no HidroWeb entre os anos hidrológicos de 1980 e 2010, intervalo no qual os dados de vazão estão disponíveis no CABra.

4.4 Resultados e discussão

4.4.1 Estimativa das assinaturas hidrológicas a partir de atributos físicos via técnicas de aprendizado de máquina

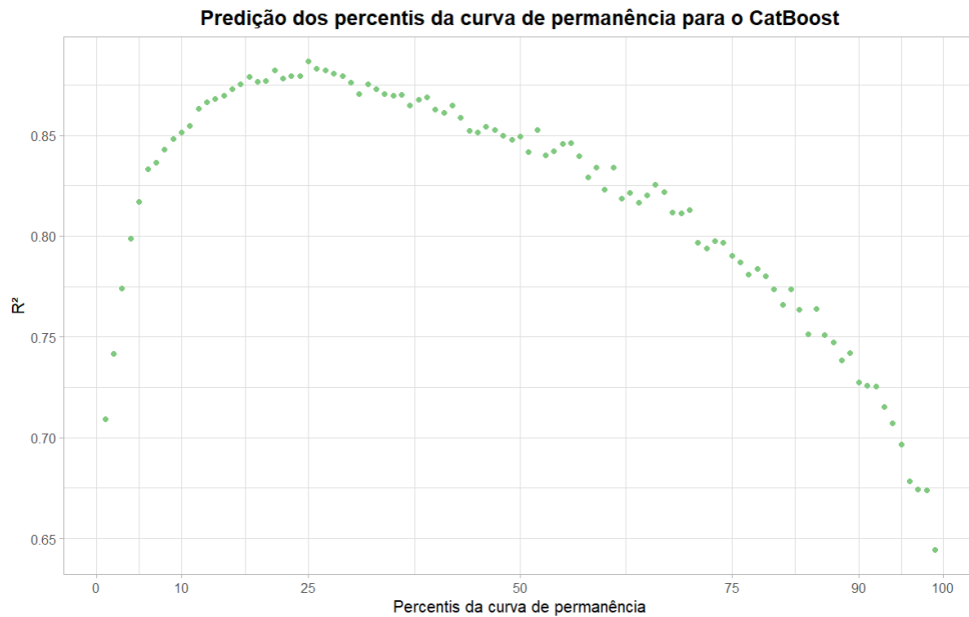
As médias de R^2 , em cada um dos 10 *folds*, nas 5 repetições para cada algoritmo, está apresentação na Figura 11, em função do identificador da assinatura indicado na Tabela 4. Os divisores da escala representam o início de cada categoria. De maneira geral o *CatBoost* apresenta as melhores predições. Esse resultado é similar ao Szczepanek (2022), mesmo que o contexto seja outro. Os demais algoritmos baseados em árvore de decisão (RF, Gradient Boosting, LGBM e XGB) têm um desempenho muito similar. Até mesmo a regressão linear múltipla consegue prever satisfatoriamente grande parte das assinaturas. Os únicos algoritmos que não obtiveram bons resultados, de uma maneira geral, são o KNN e o SVR.

Figura 11 – Média das regressões para todas as assinaturas.



A Figura 12 restringe o resultado para os percentis da curva de permanência. As baixas vazões apresentam piores resultados, em consonância com Addor et al. (2018). A partir que a permanência diminui, a métrica de desempenho da predição aumenta, até o atingir o ápice no Q25 ($R^2=0,89$), piorando na sequência com o aumento da vazão (ou a redução da permanência). Em permanências menores que 6 ($R^2=0,83$), a redução de R^2 é abrupta. Há de se destacar o resultado para o Q85 ($R^2=0,76$), empregado para representação das vazões baixas, e Q10 ($R^2=0,85$), para a vazões altas.

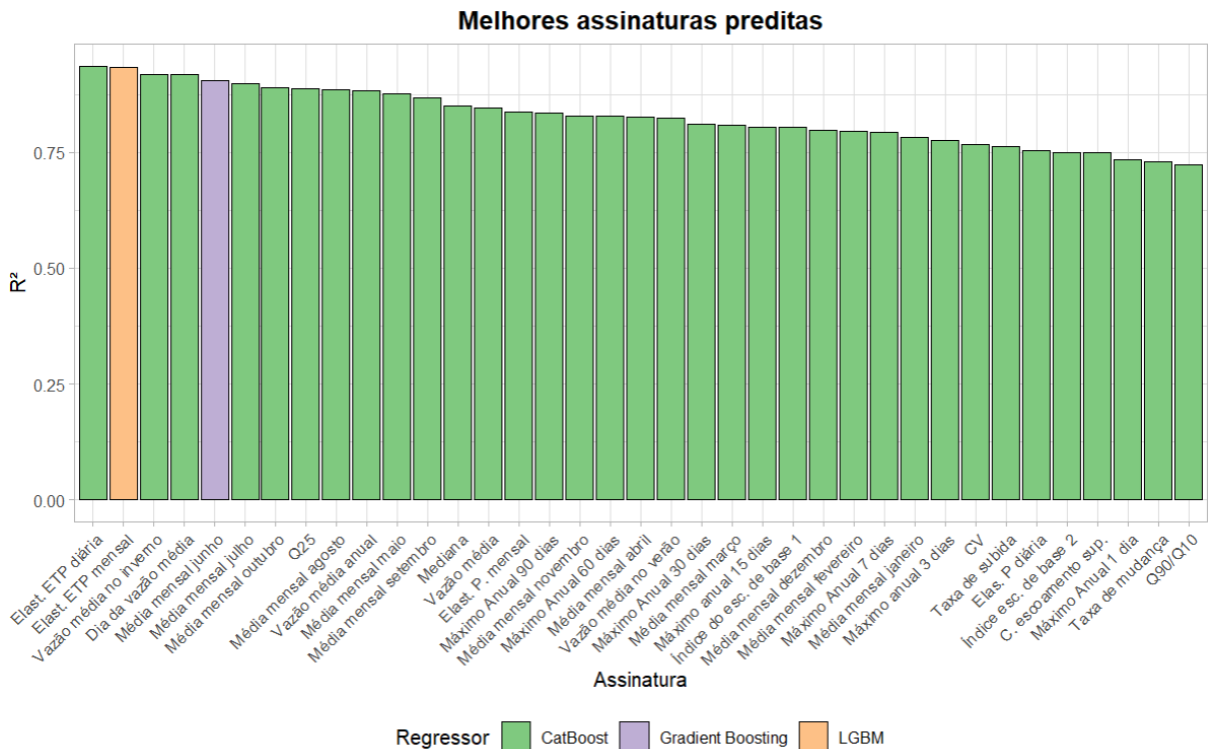
Figura 12 – Média dos coeficientes de determinação para os percentis da curva de permanência.



Admitiu-se que as assinaturas mais bem estimadas foram aquelas cujo coeficiente de determinação obtida a partir da regressão superou de 0,71. Elas estão apresentadas no gráfico da Figura 13, no qual, para facilitar a visualização, o Q25 foi o único percentil da curva de permanência apresentado. A elasticidade da evapotranspiração diária ($R^2=0,94$), seguida da mensal ($R^2=0,93$), são as assinaturas com melhores coeficientes. A vazão média no inverno ($R^2=0,92$) e o dia da vazão média ($R^2=0,92$) vêm na sequência, e também foram muito bem preditas em Addor et al. (2018), sendo o inverno com melhor resultado em relação ao verão. Os meses da vazão média mais bem estimados são aqueles do início do inverno brasileiro (junho e julho).

A mediana das vazões ($R^2=0,85$) tem a estimativa equivalente a média ($R^2=0,85$). O melhor máximo anual foi o na janela de 90 dias ($R^2=0,83$), e o melhor mínimo o de 3 dias de janela ($R^2=0,69$ e, portanto, não apresentado no gráfico). O índice de escoamento de base ($R^2=0,80$), calculado conforme Arnold et al. (1995), teve resultados superiores ($R^2=0,75$) ao de Lyne e Hollick (1979) (fornecido na base do CABra). A elasticidade da precipitação mensal ($R^2=0,84$) também foi bem estimada, assim como o CV ($R^2=0,77$), a taxa de subida ($R^2=0,76$), coeficiente do escoamento superficial ($R^2=0,75$) e a razão entre Q90 e Q10 ($R^2=0,72$).

Figura 13 – Maiores coeficientes de determinação obtidos na predição das assinaturas hidrológicas



As inclinações da curva de permanência não são bem previstas, sendo a inclinação usual (entre os percentis 33 e 67) a mais bem estimada ($R^2=0,68$). De maneira geral, os modelos têm desempenhos ruins para as vazões mínimas, as durações das vazões altas e baixas, frequências e elasticidades anuais.

Os resultados obtidos são similares ao de Addor et al. (2018) para as assinaturas comuns, com métricas aparentemente equivalentes, à exceção do índice do escoamento de base, que não havia sido bem estimado, diferente do presente trabalho. Isso reforça o impacto da escolha na forma de se calcular as assinaturas em relação aos resultados obtidos (BECK et al., 2016).

Em relação à importância das características na predição das assinaturas, a Figura 14 e Figura 15 ilustram os valores do SHAP na predição do índice do escoamento de base e taxa de mudança, respectivamente, na bacia cujo código ANA é 13880000. Em ambos os casos, parte-se de um valor esperado e cada característica colabora com o incremento ou decremento desse valor. Para o índice do escoamento de base, a baixa declividade e área de contribuição elevada contribuíram para o incremento do índice (valor estimado = 0,913, valor observado = 0,937). Por outro lado, as mesmas

características reduziram significativamente a taxa de subida (valor estimado = 0,150, valor observado = 0,153). Enquanto a área extensa e a baixa declividade permitiram um maior armazenamento para geração do escoamento de base, elas também estão associadas a uma menor variabilidade da vazão.

Figura 14 – Valores do SHAP para o escoamento de base na estação ANA 13880000.

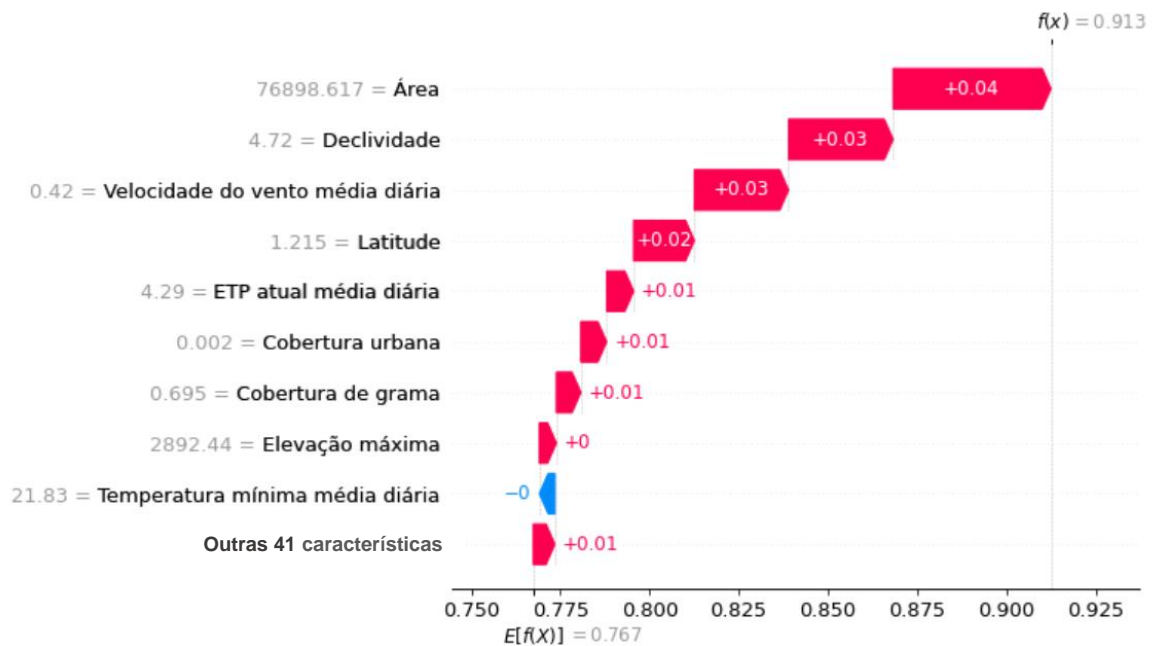
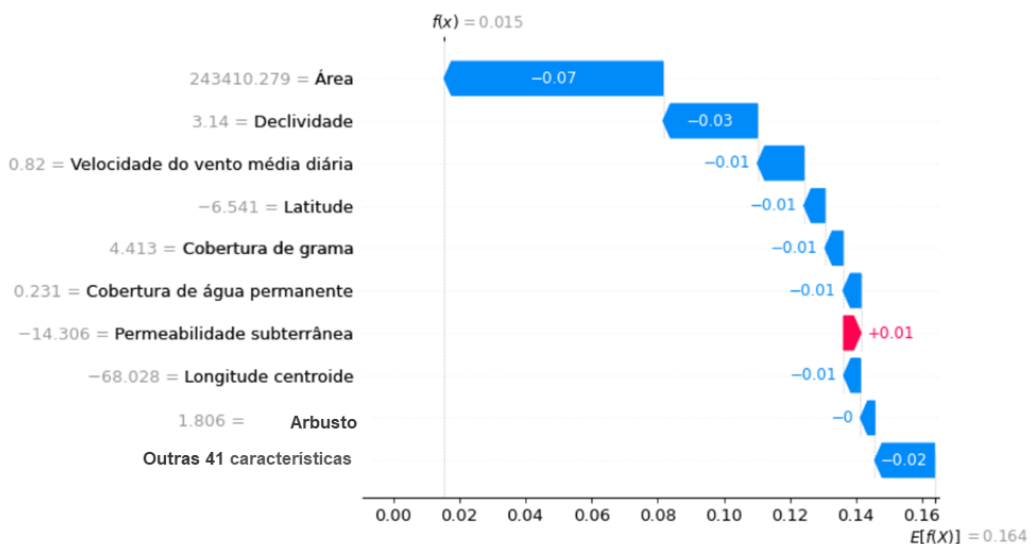


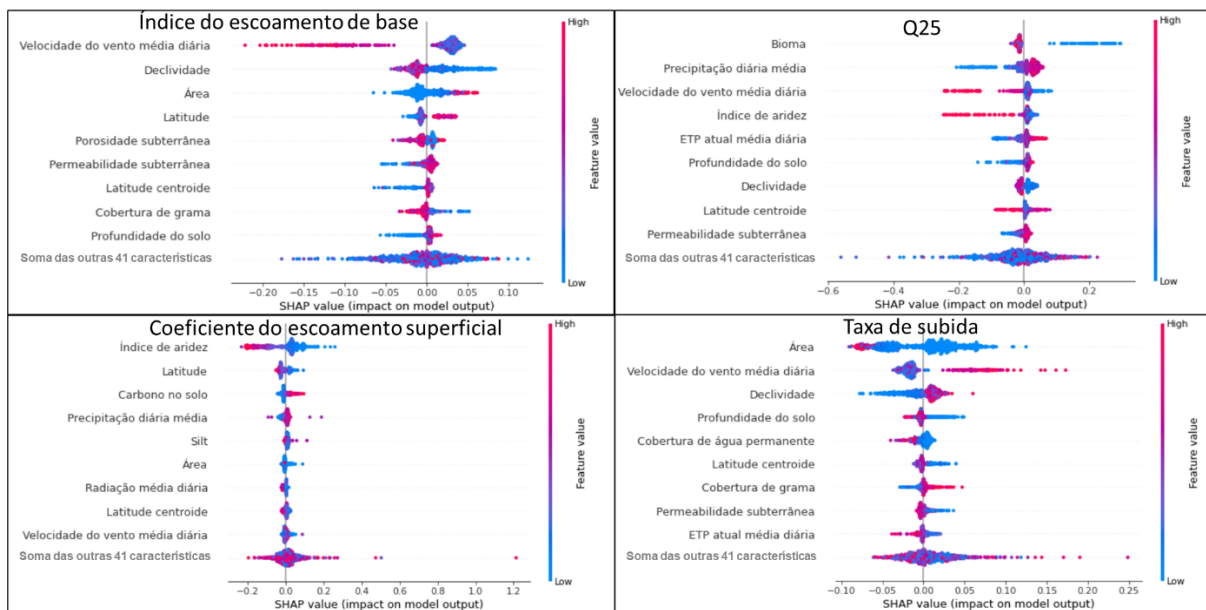
Figura 15 – Valores do SHAP para a taxa de mudança na estação ANA 13880000.



A Figura 16 apresenta o gráfico do tipo enxame de abelhas do valor do SHAP para as principais características na predição do índice do escoamento de base, da Q25, do coeficiente de escoamento superficial e da taxa de subida. Nesses gráficos, o eixo vertical esquerdo denota as características das bacias ordenados com base na

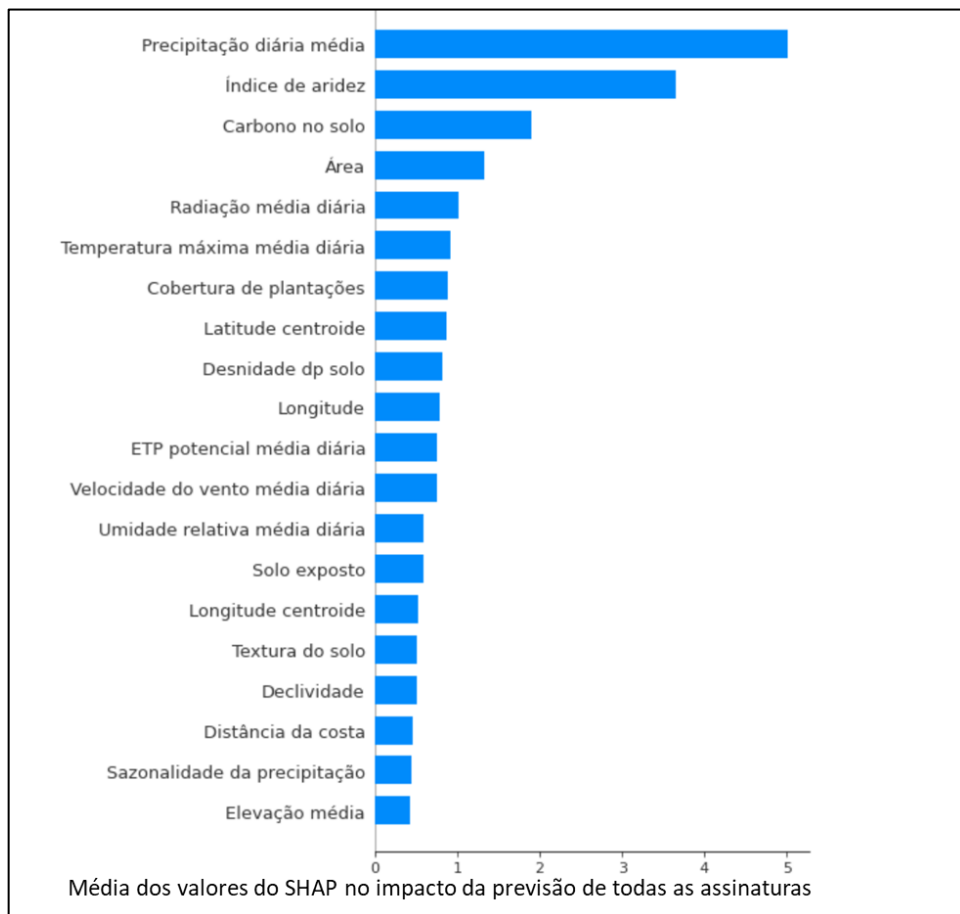
importância de cima para baixo. O eixo horizontal representa a magnitude dos valores SHAP para as previsões. O eixo vertical direito a magnitude da importância da característica colorida segundo sua importância no incremento ou decremento do valor da assinatura. Vê-se, a exemplo, que valores mais elevados de precipitação e de permeabilidade subterrânea estão associados a valores mais elevados de Q25, índices de aridez menores ao aumento do coeficiente de escoamento superficial, e a material orgânica no solo (carbono no solo) contribui para diminuir esse coeficiente.

Figura 16 – Valores do SHAP para as principais características para a predição do índice do escoamento de base, Q75, coeficiente do escoamento superficial e taxa de subida



Por fim, a Figura 17 resume a contribuição dos principais atributos físicos para predição do conjunto de assinaturas. De uma forma geral, a precipitação diária média é a variável mais importante, seguida do índice de aridez. O carbono do solo e área de contribuição integram o grupo das 4 características mais relevantes. Como a vazão é uma resposta à forçantes meteorológicas, esperava-se que as assinaturas hidrológicas que as envolvessem fossem fortemente influenciadas por indicadores que caracterizassem a precipitação e evapotranspiração. O destaque é a quantidade de matéria orgânica no solo, estimada a partir de sensoriamento remoto, que contribui de forma significativa para explicar o comportamento das bacias.

Figura 17 – Principais atributos na previsão de todas as assinaturas.

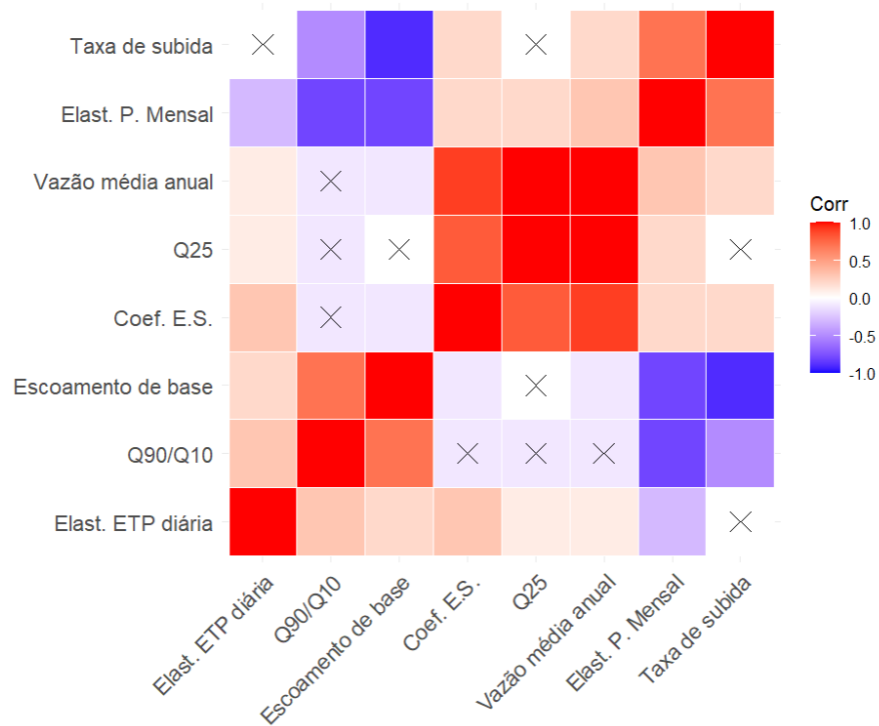


4.4.2 Inferência de agrupamentos de um conjunto de assinaturas hidrológicas

Dentre as assinaturas mais bem estimadas, escolheu-se aquelas que melhor representassem todas as categorias (Tabela 4). Entretanto, assinaturas que pudessem incorporar padrões sazonais decorrentes de sua localização foram excluídas (ex.: as vazões médias mensais e o dia de sua ocorrência). Conforme anteriormente indicado, o intuito é que a coerência espacial decorra do comportamento hidrológico similar de bacias próximas e não do compartilhamento de um mesmo regime de chuvas, por exemplo.

As assinaturas empregadas e a correlação entre elas podem ser visualizadas na Figura 18. Há uma correlação positiva entre a vazão média anual, Q25 e o coeficiente de escoamento superficial, assim como o índice do escoamento de base e a razão entre os quantis 90 e 10. Esses dois últimos se relacionam inversamente com a taxa de subida e a elasticidade de precipitação mensal. A assinatura que aparentemente menos se relaciona com as demais é a elasticidade da evapotranspiração.

Figura 18 – Correlação entre as assinaturas hidrológicas utilizadas no agrupamento.



A partir desse quadro, procedeu-se à análise de componentes principais. Conforme indicado nas Figura 19 e Figura 20, três componentes explicam 90% da variância das assinaturas. A primeira componente explica bem a razão entre a Q90 e a Q10, o índice do escoamento de base, a elasticidade da precipitação e a taxa de subida. A segunda explica a Q25, o coeficiente do escoamento superficial e a vazão anual média. Por fim, a terceira a elasticidade da evapotranspiração.

Figura 19 – Proporção da variância total explicada pelas componentes principais identificadas.

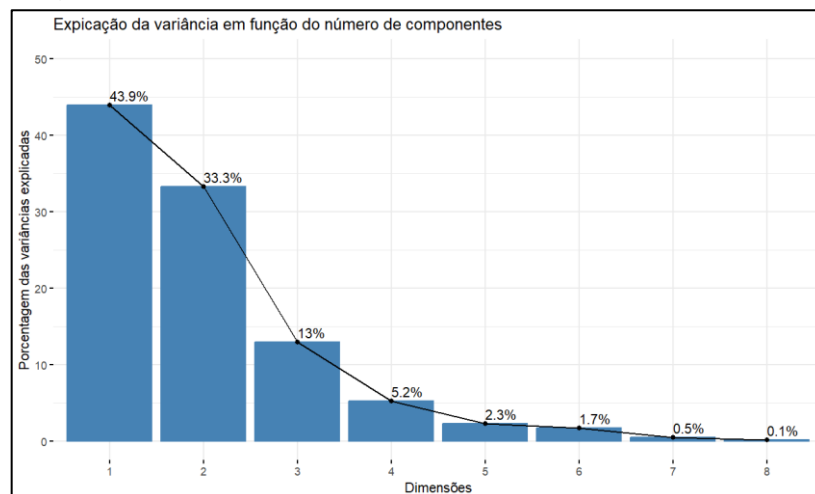
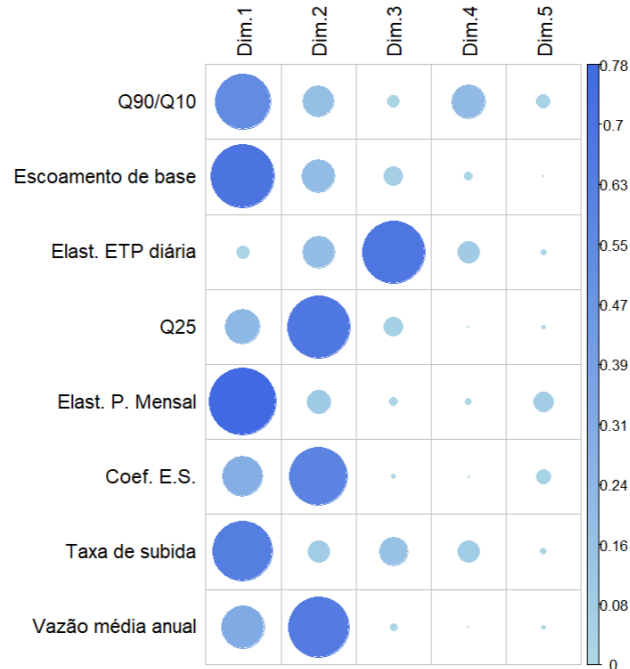
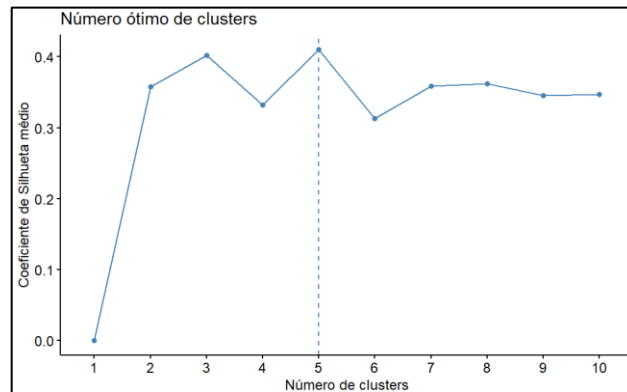


Figura 20 – Importância relativa das assinaturas hidrológicas por componente identificada.



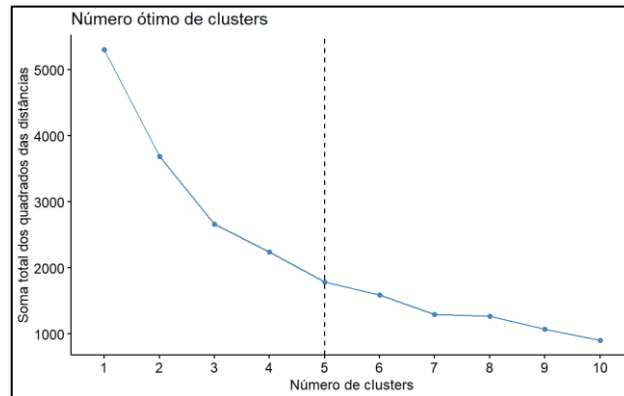
A partir das três primeiras componentes, avaliou-se o número de *clusters*. A Figura 21 apresenta o gráfico do coeficiente de silhueta médio em função do número de *clusters*. Nesse caso o número ideal é de 5 agrupamentos.

Figura 21 – Análise do número ótimo *clusters* a partir do coeficiente de Silhueta.



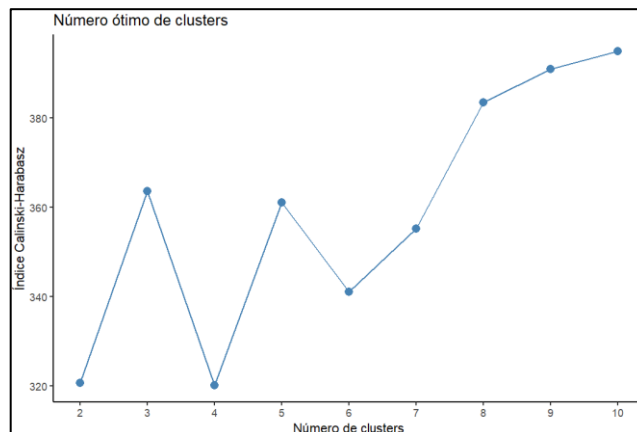
A Figura 22 ilustra o Método do Elbow. Por ser um método visual, diferentes interpretações podem alterar o número de clusters. O resultado sugere 3, 5 ou 7 agrupamentos.

Figura 22 – Análise do número ótimo de *clusters* a partir do Método de Elbow.



Por outro lado, o Índice Calinski-Harabasz, apresentado na Figura 23, sugere acima de 8 clusters, mas com valores representativos para 3, 5 e 7 clusters.

Figura 23 – Análise do número ótimo de *clusters* a partir do Índice Calinski-Harabasz.



Frente aos resultados, prosseguiu-se com 5 agrupamentos e as 3 componentes principais como entrada para o *K-Means*. A Figura 24 apresenta cada *cluster* encontrado em relação as principais componentes. Há um *cluster* mais centralizado e outros 4 influenciados pelas regiões mais positivas e negativas das componentes 1 e 2. Em complemento, os gráficos das Figura 25, Figura 26 e Figura 27 apresentam o comportamento de cada assinatura em cada um dos *clusters*.

Figura 24 – Agrupamentos em relação as componentes principais.

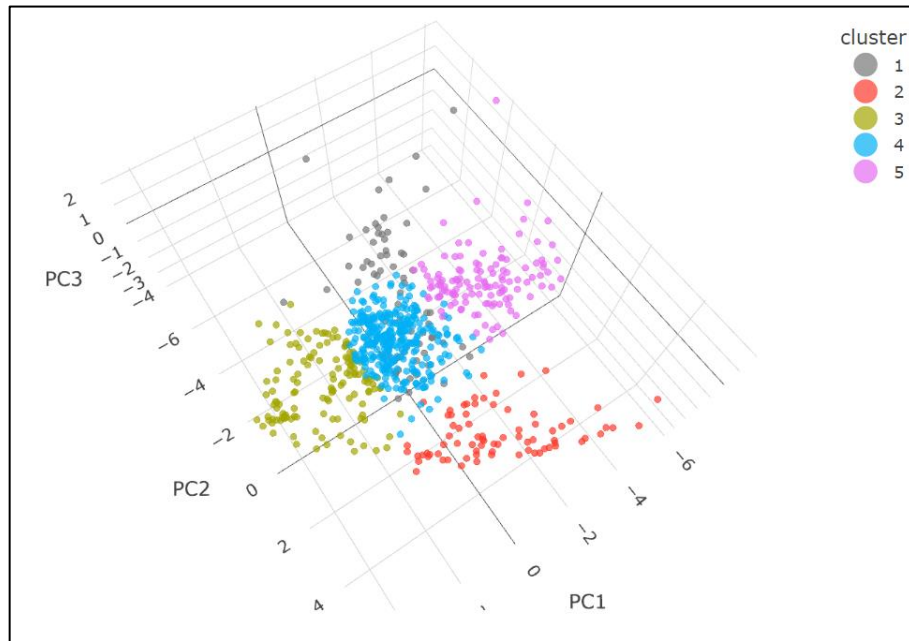
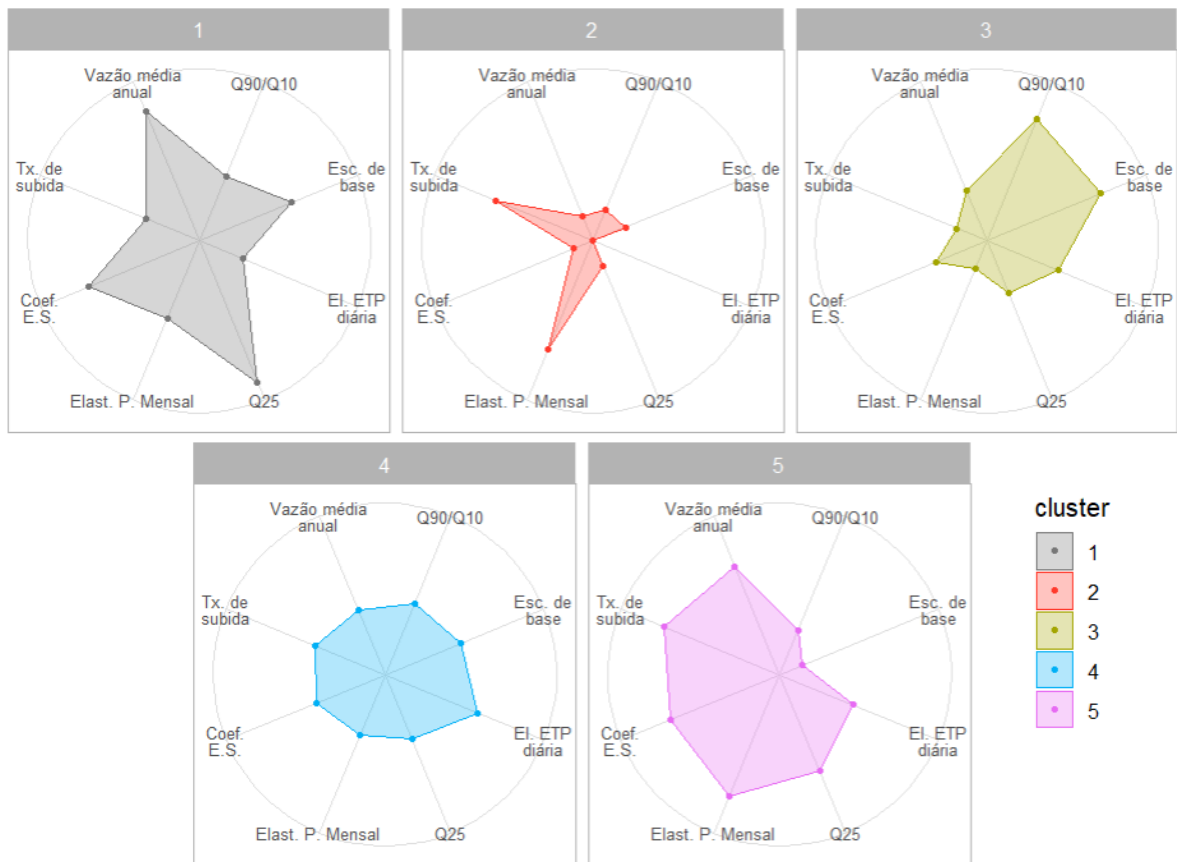


Figura 25 – Representação da importância média das assinaturas hidrológicas em cada *cluster*.

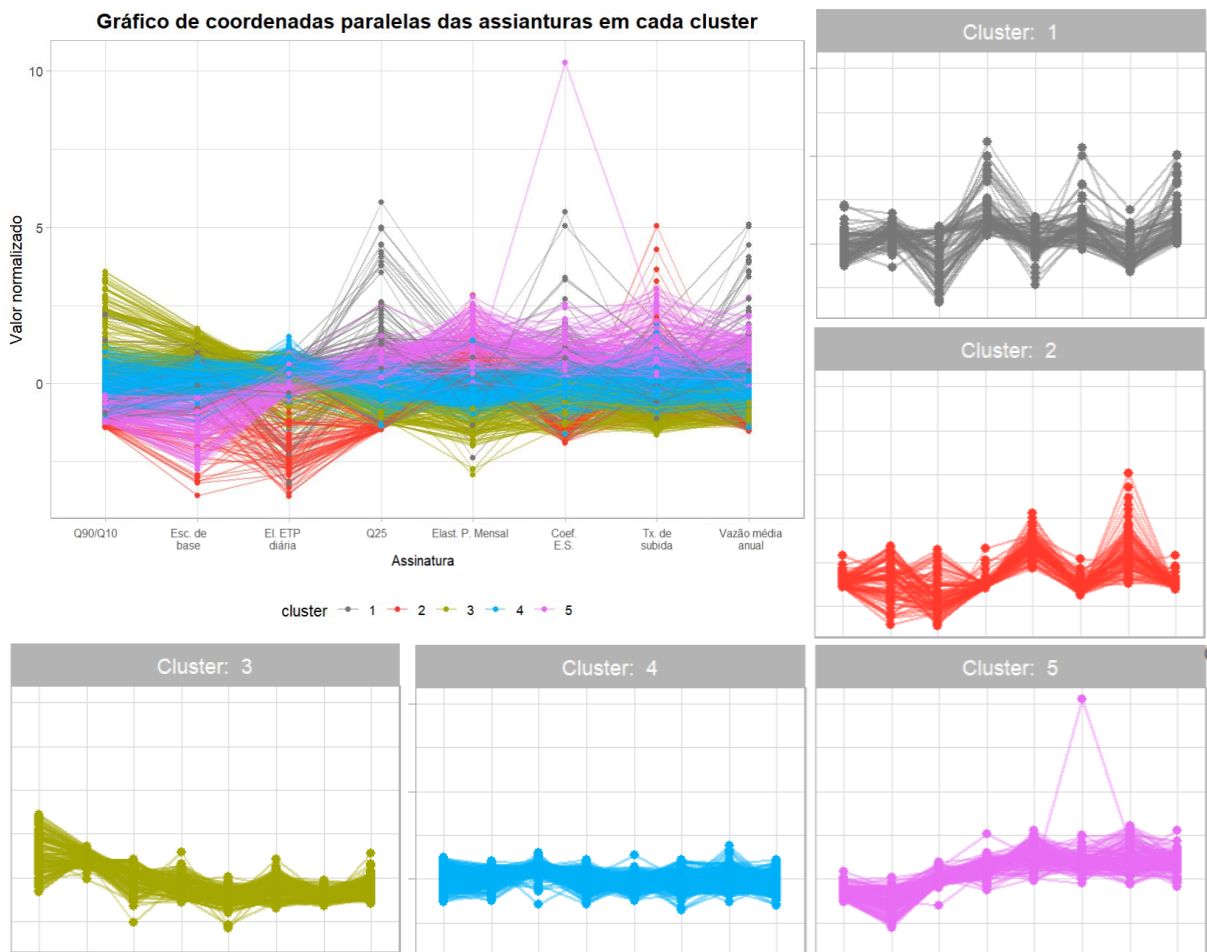
Assinaturas hidrológicas de cada cluster



No gráfico de radar é possível perceber que o *cluster* 1 representa bacias com vazões médias anuais elevadas; o *cluster* 2 bacias com vazões médias anuais baixas e

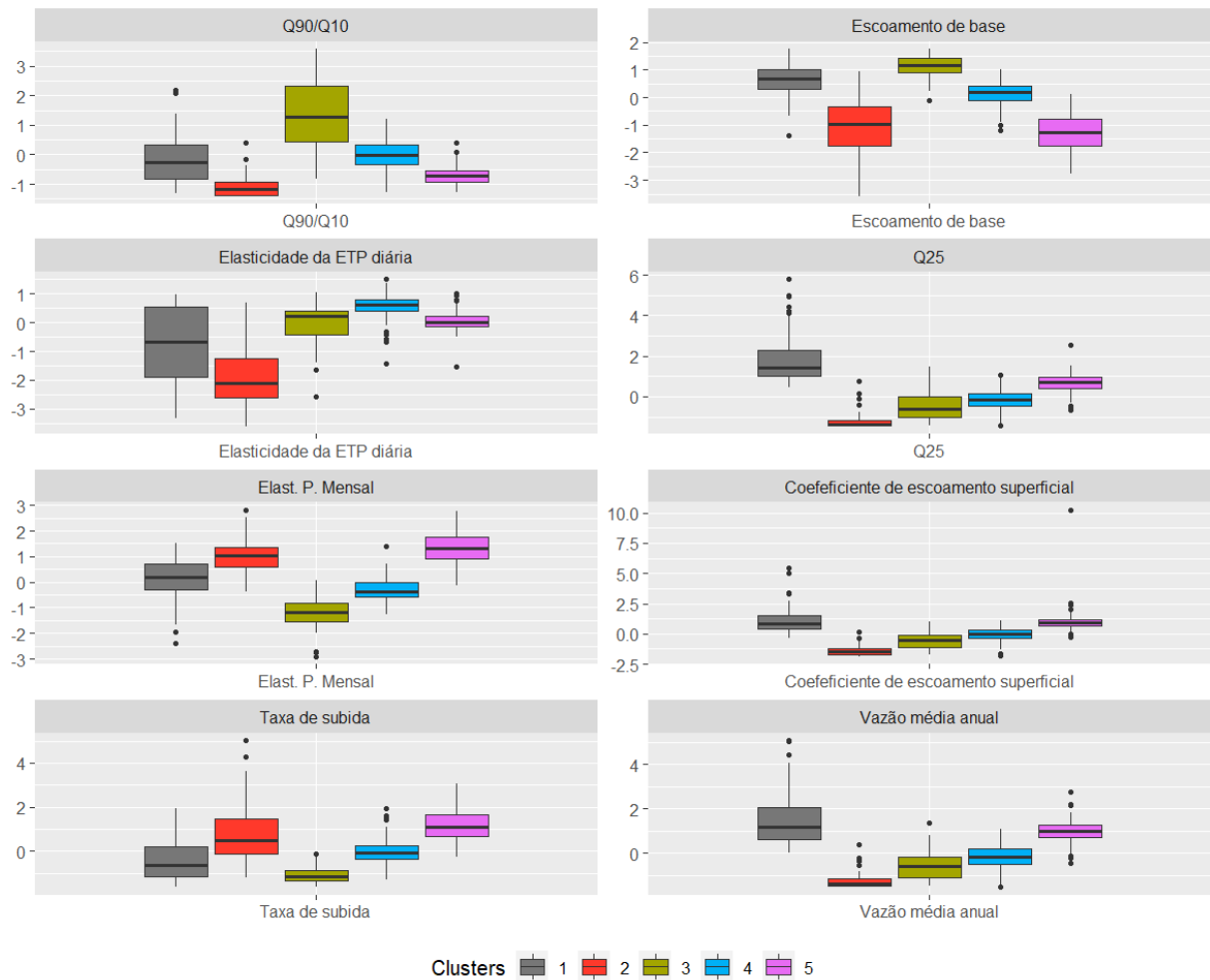
elevada taxa de subida; o *cluster* 3 também agrupa bacias com vazões médias anuais inferiores à média global mas com elevado índice de escoamento de base; já o *cluster* 4 tem todas as assinaturas em torno de um valor médio; e, por fim, o *cluster* 5 bacias em que o escoamento de base não é expressivo e as demais assinaturas (exceto o Q90/Q10) se comportam acima da média. No gráfico de coordenadas paralelas, é possível perceber uma boa segmentação dos *clusters*, com o *cluster* 3 sendo o oposto do 5.

Figura 26 – Gráfico de coordenadas paralelas das assinaturas hidrológicas de cada bacia coloridas por cluster.



O gráfico *boxplot* reforça a homogeneidade de cada agrupamento, com exceção do *cluster* 1 que apresentou uma dispersão maior na maioria das assinaturas.

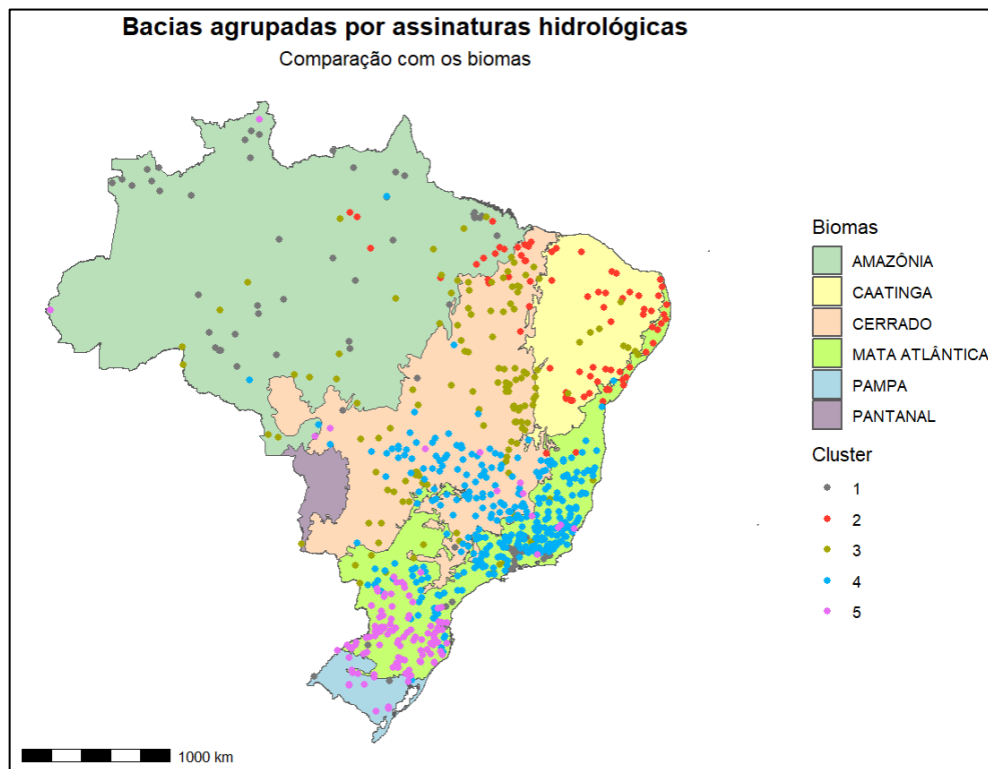
Figura 27 – Boxplot com as assinaturas hidrológicas em cada agrupamento.
Comportamento das assinaturas hidrológicas em cada cluster



4.4.2.1 Caracterização dos agrupamentos

Especialmente, houve uma tendência de bacias próximas estarem no mesmo agrupamento, assim como houve uma boa relação com os biomas brasileiros (Figura 28), as zonas climáticas (Figura 29), zonas de umidade (Figura 30), isoietas de precipitação anual média (Figura 31) e principais aquíferos (Figura 32).

Figura 28 – Agrupamentos obtidos em relação aos biomas brasileiros.



- Agrupamento 1: esse *cluster* conta com 71 bacias e quase 60% delas no bioma amazônico. O restante é majoritariamente pertencente à mata atlântica. A região amazônica é dominada pelo clima equatorial, que é caracterizado por altas temperaturas durante todo o ano (geralmente acima de 25 graus Celsius) e precipitação abundante (média anual superior a 2.000 mm). Nessas bacias, destaca-se a elevada área de contribuição. Fora dessa região, as bacias se concentraram na região na divisa dos estados de Minas Gerais, Rio de Janeiro e São Paulo, com uma elevação acima da média para a região da Mata Atlântica, e, por isso, apresentam índices pluviométricos mais elevados. As bacias do agrupamento 1 se destacam por uma elevada Q25 e vazão média anual, alta geração de escoamento superficial e elevado índice de escoamento de base. Comparadas com outras bacias úmidas, as bacias desse grupo apresentam uma menor taxa de subida.
- Agrupamento 2: as 76 bacias pertencentes a esse grupo, em sua maioria, localizam-se no bioma da Caatinga e sua fronteira com a Mata Atlântica no nordeste brasileiro. O clima predominante é o Tropical de Zona Equatorial e o Tropical Nordeste Oriental, definindo a região do semiárido brasileiro, que é caracterizado por baixas precipitações anuais (menos de 800 mm por ano) e

temperatura média anual variando de 25 a 29 graus Celsius. Uma das características do semiárido é a intermitência dos cursos d'água, resultando em uma elevada taxa de subida e elasticidade da precipitação. As vazões são baixas, com coeficiente de escoamento superficial reduzido, assim como o índice do escoamento de base.

- Agrupamento 3: o terceiro *cluster*, com 136 bacias, abrange o Cerrado brasileiro, especialmente na região norte e nordeste, nos limites com a Caatinga. Apesar de apresentar vazões médias anuais mais baixas, assim como as bacias no semiárido, esse grupo de bacias apresenta elevados índices de escoamento de base, em comparação com as vazões altas (Q90/Q10), provavelmente em função do elevado índice de escoamento de base. Nesse sentido, cabe destacar a influência dos aquíferos Bauru (pertencente ao Guarani) e Urucuia no comportamento dessas bacias.
- Agrupamento 4: por estar na região mais bem monitorada no território brasileiro, com 328 bacias, esse é o maior grupo. Predominantemente localizado no sudeste brasileiro, em clima Tropical Brasil Central, variando de bacias úmidas e semiúmidas. Esse grupo de bacias tende a acompanhar a isoieta de 1500 mm de precipitação anual, um valor médio para a região brasileira. Esse pode ser um indicativo da razão das assinaturas hidrológicas também apresentarem valores médios, exceto em relação a elasticidade da evapotranspiração, sendo esse grupo marcado pela maior sensibilidade a esse fator.
- Agrupamento 5: O *cluster* 5, com 124 bacias, localizado no sul do Brasil é o mais bem definido. Claramente influenciado pelo clima temperado e pela chuva anual média em torno de 2.000 mm anuais, esse grupo de bacias apresenta alto coeficiente de escoamento superficial e altas vazões, que podem variar significativamente em função das variações na precipitação. As bacias desse grupo, apresentam o menor índice de escoamento de base de todos os agrupamentos, razão pela qual a Q90 é muito menor que a Q10. Esse grupo de bacias apresenta índices hidrológicos opostos ao do terceiro agrupamento

Figura 29 – Agrupamentos obtidos em relação as zonas climáticas brasileiras.



Figura 30 – Agrupamentos obtidos em relação as zonas de umidade.

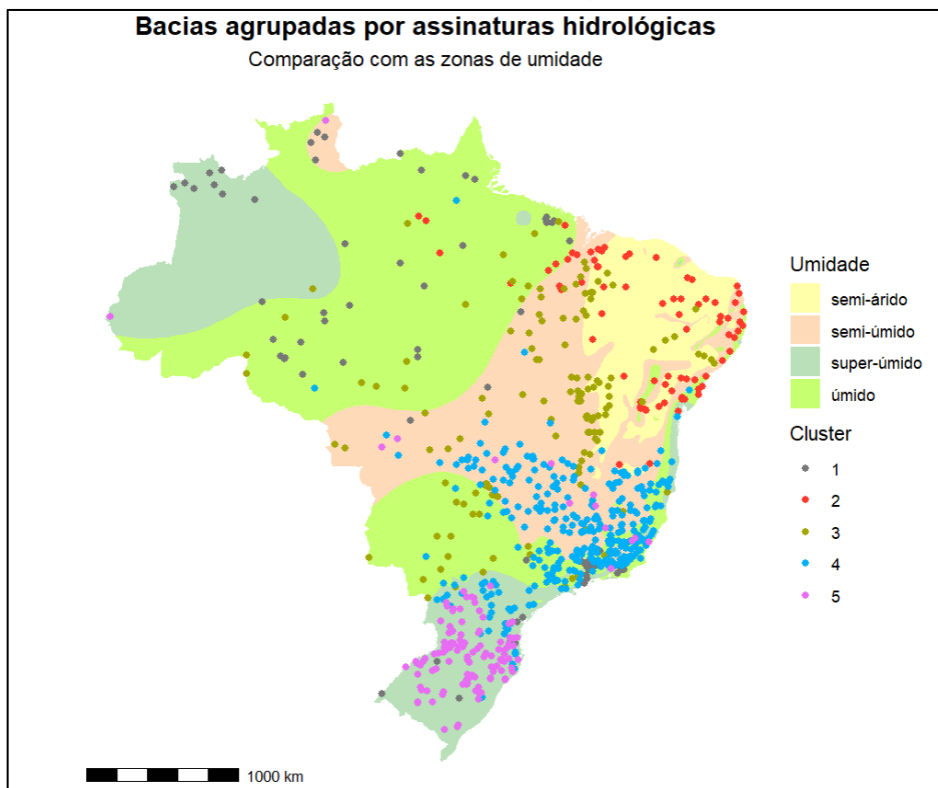


Figura 31 – Agrupamentos obtidos em relação a krigagem dos dados de chuva disponíveis no Hidroweb entre os anos hidrológicos de 1980 e 2010.

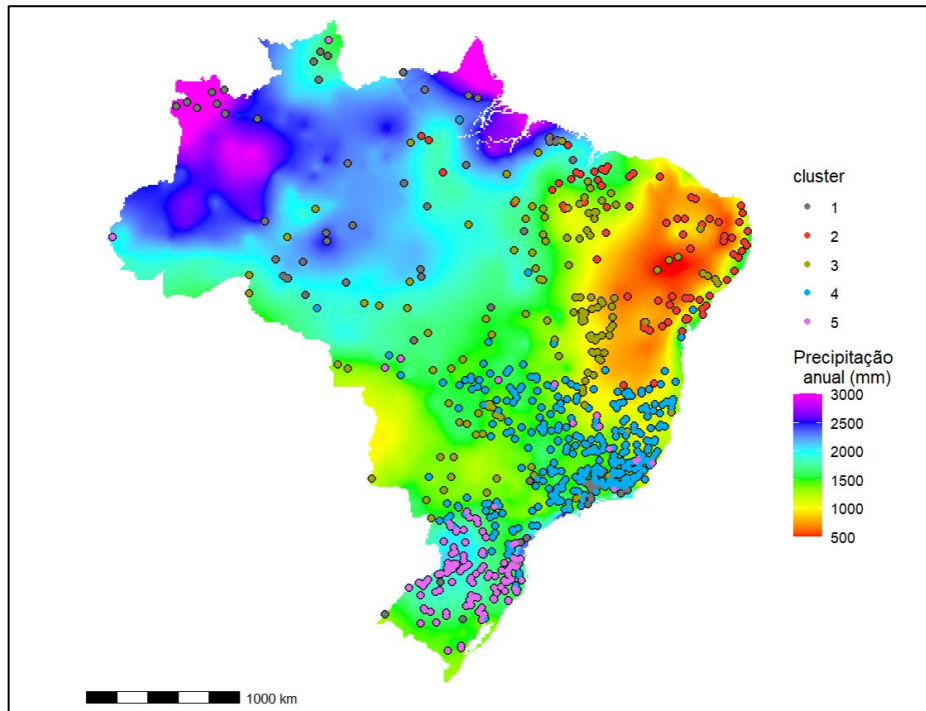
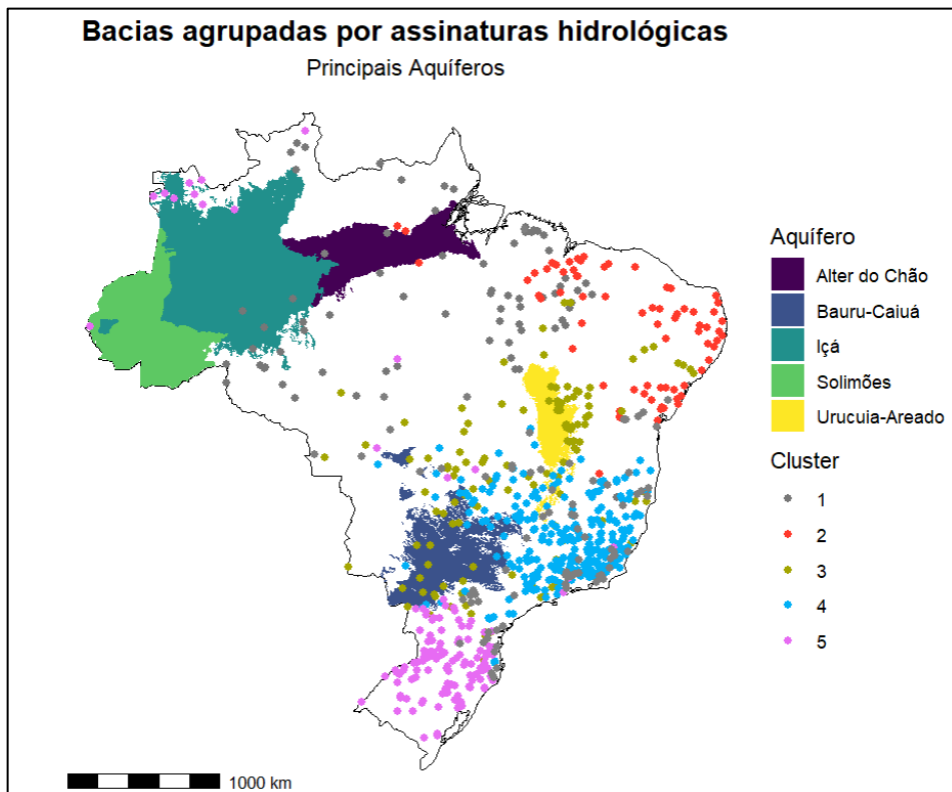


Figura 32 – Agrupamentos obtidos em relação aos principais aquíferos.



4.5 Conclusões

Ao todo, 163 índices hidrológicos foram estimados utilizando atributos físicos das bacias por meio de regressores baseados em aprendizado por máquina. Mais de 131 assinaturas foram bem estimadas ($R^2 > 0,7$) e os resultados são coerentes com outros estudos da literatura, tanto quanto à eficiência dos algoritmos (SZCZEPANEK, 2022), quanto a correspondência das assinaturas bem estimadas (ADDOR et al., 2018). Além disso, identificou-se os atributos mais relevantes para estimativas das assinaturas, o que pode ser útil em futuras análises para se entender os fatores que levam a certos comportamentos dominantes em cada bacia. Entre os algoritmos avaliados, se destaca o *CatBoost*.

Diferente dos estudos consultados, o agrupamento das bacias foi desenvolvido exclusivamente a partir das assinaturas hidrológicas. Dentre as assinaturas mais bem estimadas, escolheu-se 8 para agrupar as 735 bacias do CABra. As técnicas analisadas indicaram que 5 *clusters* era um número adequado para esse propósito. Os agrupamentos gerados tiveram uma boa consistência espacial, sendo influenciados pelo clima, umidade, bioma e precipitação média, mostrando que a combinação desses fatores desencadeia comportamentos dominantes distintos.

O primeiro *cluster*, predominante na região amazônica é caracterizado por vazões elevadas o ano todo, influenciado por um alto índice de escoamento de base. O segundo *cluster*, na região do semiárido, é definido por vazões mais baixas e mais sensibilidade quanto à precipitação. O terceiro também apresenta baixas vazões, mas a influência dos aquíferos particulariza a influência do escoamento de base. O quarto *cluster*, na região sudeste, apresenta índices hidrológicos médios, com uma maior influência em variações da evapotranspiração. E o quinto *cluster*, marcado pelo clima temperado, também apresenta altas vazões, mas com pouca geração de escoamento de base.

A maior dispersão das assinaturas para os clusters das regiões norte e centro-oeste, assim como uma menor consistência espacial, pode ser um reflexo da menor representatividade na base de dados do CABra, por se tratarem de regiões menos monitoradas.

Trabalhos futuros podem utilizar as assinaturas preditas na calibração de modelos hidrológicos (em consonância com o que foi realizado no Capítulo 3), além de explorar os valores de SHAP e aprofundar sobre os fatores que podem levar a certos comportamentos hidrológicos. Ao reunir mais dados sobre a região norte do país, o agrupamento 1 pode ser segmentado, diminuindo sua dispersão e melhorando sua compreensão e caracterização. Outros algoritmos de agrupamento devem ser avaliados, a respeito do Fuzzy C-Means, identificando bacias que podem pertencer a mais de um *cluster*.

5 SOBRE A ESTRUTURA DO MODELO CONCEITUAL NA MODELAGEM HIDROLÓGICA

“As far as we can tell, these models are dead on arrival.”

(NEARING et al., 2021)

5.1 Introdução

Um dos dilemas da modelagem hidrológica envolve a seleção do modelo hidrológico. Idealmente o modelo deveria ser adequado para a bacia que se pretende modelar. No entanto, os modelos hidrológicos são frequentemente selecionados por praticidade, conveniência, experiência e hábito (ADDOR; MELSEN, 2019).

Para que um modelo hidrológico represente adequadamente a função da bacia, ele deve representar apropriadamente os processos hidrológicos subjacentes (PRIETO et al., 2021). Esses processos, como a infiltração da água no solo ou fluxo intermediário, geralmente são muito complexos para parametrizar ou dificilmente compreendidos em escalas de interesse para serem representados em todos os detalhes em modelos matemáticos. Por esse motivo são normalmente abstraídos em modelos conceituais (MAI; CRAIG; TOLSON, 2020).

As descrições do modelo não são únicas, uma vez que dependem das simplificações e interpretações durante a conceituação do fenômeno. A disponibilidade de diferentes equações leva a uma variedade de opções para descrever os processos hidrológicos. O desenvolvimento do modelo é, portanto, específico do problema ou mesmo da bacia no qual ele foi testado e validado (FENICIA et al., 2016; HÖGE; WÖHLING; NOWAK, 2018; MAI; CRAIG; TOLSON, 2020).

Considerando a diversidade das propriedades das bacias hidrográficas, como vegetação, geologia e clima, é improvável que um único modelo possa ser aplicado adequadamente em uma ampla gama de bacias hidrográficas (CHADALAWADA; HERATH; BABOVIC, 2020). Por exemplo, os parâmetros do modelo GR5J foram calibrados para 179 bacias do cerrado brasileiro com bons resultados, mas apresenta métricas com baixo desempenho em algumas bacias até mesmo no período de calibração (ALTHOFF; RODRIGUES, 2021). Uma das maiores fontes da incerteza da

modelagem hidrológica é frequentemente atribuída à fase de conceituação do modelo (FENICIA et al., 2018; POULIN et al., 2011; SEILLER; ANCTIL; ROY, 2017).

Por outro lado, há uma variedade de modelos com estruturas flexíveis na literatura (por exemplo CLARK et al., 2008; CRAIG et al., 2020; KAVETSKI; FENICIA, 2011). Estruturas de modelagem flexíveis e modulares oferecem uma oportunidade para explorar impactos das alternativas para a estrutura do processo e identificar configurações de modelo apropriadas (BROWN; CRAIG, 2020; DAL MOLIN; KAVETSKI; FENICIA, 2021). Modelos flexíveis permitem comparar diferentes hipóteses sobre o funcionamento das bacias hidrográficas, alterando ou acrescentando componentes ao modelo (por exemplo, DAVID et al., 2022; FENICIA et al., 2016; PRIETO et al., 2021).

As comparações entre modelos devem ser baseadas em um número significativo de bacias hidrográficas para fornecer conclusões que não sejam o resultado do acaso. Os princípios hidrológicos gerais devem ser derivados de relações estatisticamente significativas, as quais não podem ser obtidas com dados de apenas algumas bacias hidrográficas (ADDOR et al., 2020; ANDRÉASSIAN et al., 2006; GUPTA et al., 2014).

Padrões emergentes de grandes amostras podem fornecer informações únicas sobre como os modelos funcionam em uma variedade de diferentes tipos de bacias hidrográficas e informar a compreensão da similaridade hidrológica entre diferentes lugares (HRACHOWITZ; CLARK, 2017). Os benefícios de tais estudos incluem uma compreensão mais completa dos locais onde um determinado modelo hidrológico pode ser usado com alguma medida de confiança e a capacidade de modelar bacias não monitoradas por regionalização (KNOBEN et al., 2020).

Não há na literatura uma diretriz inequívoca que permita estabelecer, *a priori*, uma relação entre a estrutura do modelo conceitual e as características físicas da bacia, como clima, geologia, topografia, solo e cobertura vegetal. Não obstante, há estudos que indicam que essa relação seja presumível caso se considere assinaturas hidrológicas, como o índice de aridez, o índice do escoamento de base e a curva de permanência (por exemplo, DAVID et al., 2022; KNOBEN et al., 2020; LEY et al., 2015; MASSMANN, 2020) sem, no entanto, estabelecê-las.

Diante desse quadro, o presente estudo se propõe a investigar a relação entre estruturas de modelos conceituais e assinaturas hidrológicas. Para tal, 351 bacias hidrográficas da base pública do CABra (ALMAGRO et al., 2021) foram modeladas para cada uma das 624 estruturas de modelos conceituais presentes no *Framework for Understanding Structural Errors* (FUSE; CLARK et al., 2008). As assinaturas hidrológicas foram empregadas durante a calibração dos parâmetros, sob o propósito de garantir a consistência hidrológica das estruturas consideradas adequadas à representação dos processos hidrológicos em cada bacia (HRACHOWITZ et al., 2014). As questões científicas que conduziram o desenvolvimento foram:

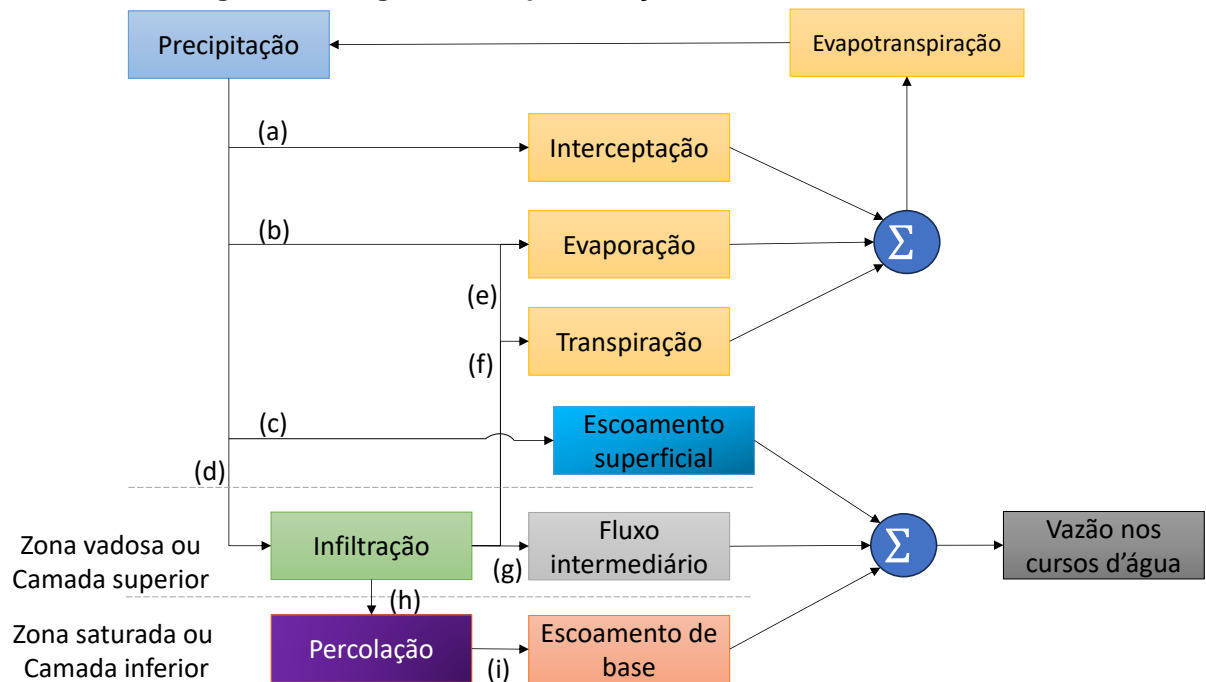
- Há uma relação explícita entre as estruturas dos modelos e as propriedades físicas das bacias e/ou as assinaturas hidrológicas?
- Essas relações permitem caracterizar os processos hidrológicos dominantes?
- Bacias agrupadas por similaridade de assinaturas hidrológicas compartilham os mesmos componentes do modelo hidrológico?

5.2 Desenvolvimento teórico

5.2.1 Mecanismos e processos hidrológicos

A complexidade dos processos em sistemas hidrológicos é, em seu nível fundamental, caracterizada por dois pontos principais de partição que controlam como a água é armazenada e liberada das bacias hidrográficas por meio de fluxos ascendentes, descendentes ou laterais (CLARK et al., 2015; SAVENIJE; HRACHOWITZ, 2017). Próximo à superfície terrestre, a precipitação é dividida em (a) evaporação e (b) interceptação da superfície do solo, bem como de corpos de água abertos, (c) escoamento superficial e (d) a infiltração na zona radicular. A água que entra na zona vadosa é ainda dividida em (e) evaporação do solo, (f) transpiração das plantas, (g) fluxo intermediário lateral, (h) percolação para a zona saturada e (i) escoamento de base. Todos os fluxos (a-i) estão presentes em praticamente qualquer bacia hidrográfica, embora com diferentes importâncias relativas em diferentes ambientes (CHOW et al., 1988; HRACHOWITZ; CLARK, 2017). Essa divisão de fluxos é ilustrada na Figura 33.

Figura 33 - Diagrama de representação dos fluxos em uma bacia



Adaptado de Chow et al. (1988) e Hrachowitz e Clark (2017)

Esses fenômenos físicos que ocorrem na escala da bacia, como a geração de escoamento superficial ou a percolação, são usualmente identificados como processos hidrológicos. Para representá-los os modelos utilizam conjuntos de equações ou “mecanismos hidrológicos”. Um único modelo hidrológico representa uma combinação de mecanismos, um por processo hidrológico. Para que um modelo hidrológico represente adequadamente uma bacia, ele deve representar adequadamente os processos hidrológicos que ocorrem na mesma (PRIETO et al., 2021).

A identificação confiável de mecanismos hidrológicos é importante para vários propósitos, incluindo melhorar o desempenho na previsão, obter informações sobre o funcionamento interno das bacias, e para que o modelo obtenha boas métricas pelas razões corretas (KIRCHNER, 2006). Espera-se também que a seleção de mecanismos hidrológicos respeite o princípio da parcimônia, que favorece representações de modelos mais simples (JAKEMAN; HORNBERGER, 1993).

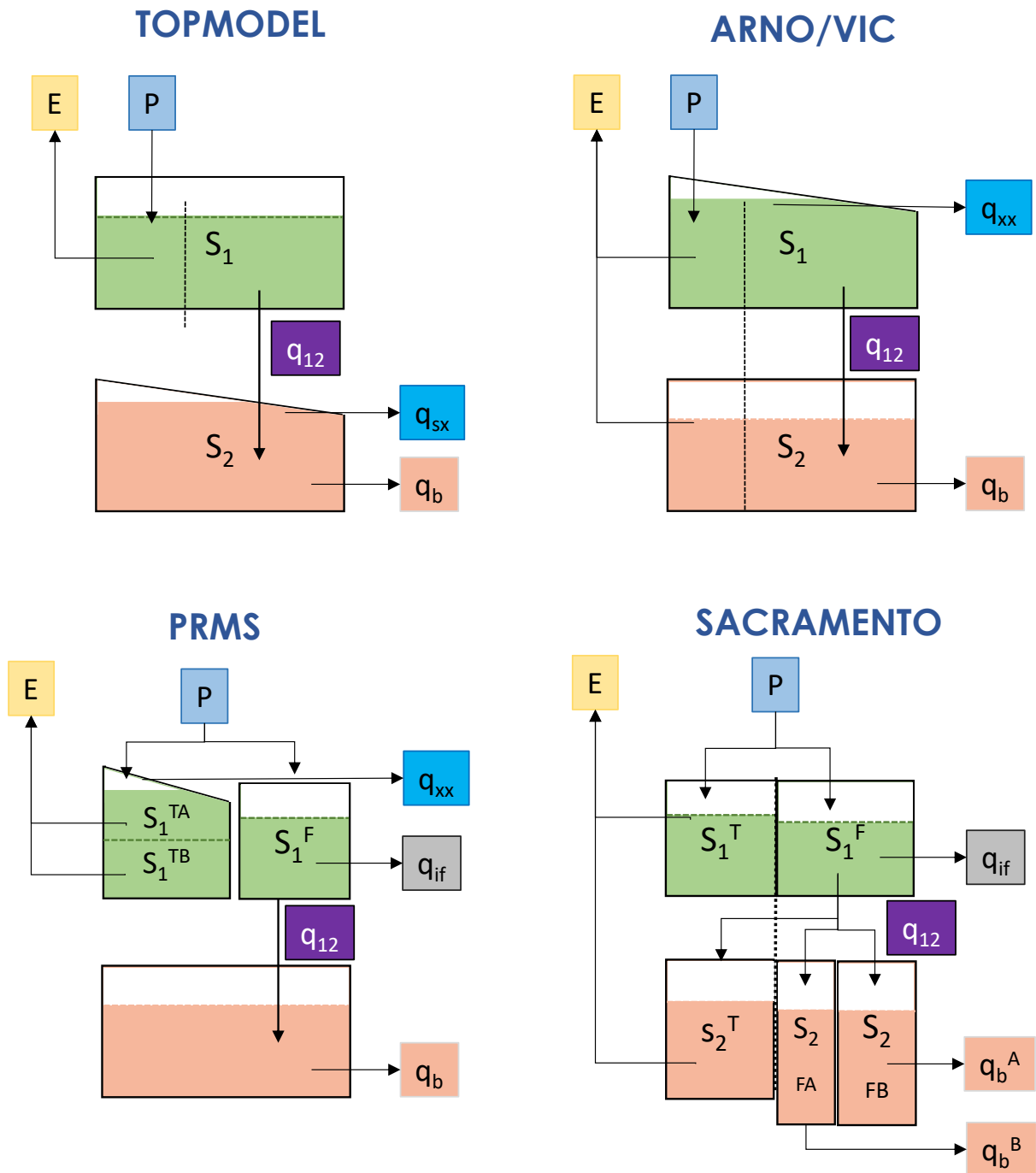
5.2.2 O FUSE

O FUSE é uma estrutura de modelagem modular que permite a geração de centenas de modelos hidrológicos conceituais pela recombinação de elementos de modelos comumente usados, permitindo escolher vários mecanismos para representar cada processo hidrológico no modelo (CLARK et al., 2008). Um total de sete processos são considerados:

1. Arquitetura da camada superior do solo para o armazenamento de água que ocorre na zona não saturada;
2. Arquitetura da camada inferior do solo para o armazenamento da água que ocorre na zona saturada;
3. Geração de escoamento superficial;
4. Percolação, movimento vertical da água da zona não saturada (camada superior do solo) para a zona saturada (camada inferior do solo);
5. Evapotranspiração;
6. Fluxo intermediário para o movimento lateral da água no solo; e
7. Concentração do hidrograma de escoamento superficial (forma e tempo) à medida que a água se move ao longo do curso d'água.

Os 19 mecanismos no FUSE estão detalhados na Tabela 5. Eles são representados por componentes de 4 modelos existentes: PRMS (LEAVESLEY et al., 1983), SACRAMENTO (SINGH, 2012), TOPMODEL (BEVEN; KIRKBY, 1979) e ARNO/VIC (LIANG et al., 1994), todos de complexidade semelhantes e sumariados na Figura 34. Cada estrutura é uma combinação dos 7 mecanismos hidrológicos, com um único mecanismo especificado para cada processo hidrológico. Um total de 624 estruturas modelo determinísticas podem ser consideradas, desconsiderando o modelo de erros. As entradas no FUSE são as séries temporais de precipitação média observada na bacia e evapotranspiração potencial, e as saídas são as séries temporais de vazão simulada (PRIETO et al., 2021).

Figura 34 – Diagramas simplificados dos quatro modelos principais que se baseia o FUSE.



A camada superior do solo (S_1) pode ser definida por um único reservatório ou em duas zonas (PRMS). A camada inferior (S_2) também pode ser definida com um único reservatório incidindo (TOPMODEL e PRMS) ou não (ARNO/VIC) a evapotranspiração, ou ainda em dois reservatórios em paralelo (SACRAMENTO). A evaporação pode incidir de maneira ponderada sobre as duas camadas do solo, ou primeiro na camada superior seguido da inferior, se necessário.

O escoamento superficial (q_{sx}) é calculado usando a chuva sobre uma área saturada que pode ser função de S_1 ou S_2 (TOPMODEL). Quando algum dos reservatórios excede sua capacidade esse transbordamento também se torna escoamento superficial. A propagação, quando utilizada, calcula a fração do escoamento superficial de um determinado passo de tempo que será contabilizada no próximo passo de tempo.

Tabela 5 – Processos hidrológicos e mecanismos no FUSE

Processo	Mecanismo
1 - Camada superior	m_{11} : camada superior definida por uma única variável de estado (TOPMODEL, ARNO/VIC)
	m_{12} : camada superior dividida em reservatórios acima e abaixo da capacidade de campo (SACRAMENTO)
	m_{13} : armazenamento abaixo da capacidade de campo subdividido em recarga e excesso (PRMS)
2 - Camada inferior e escoamento de base	m_{21} : armazenamento do escoamento de base com um reservatório de tamanho fixo não linear (ARNO/VIC)
	m_{22} : um reservatório abaixo da capacidade de campo e mais dois reservatórios paralelos (SACRAMENTO)
	m_{23} : reservatório do escoamento de base de tamanho ilimitado linear (PRMS)
	m_{24} : reservatório do escoamento de base de tamanho ilimitado parametrizado com uma potência (TOPMODEL)
3 - Escoamento superficial	m_{31} : a área saturada está relacionada ao armazenamento na zona não saturada via distribuição de Pareto (ARNO/ VIC)
	m_{32} : área saturada é uma função linear do armazenamento abaixo da capacidade de campo na zona não saturada (variante PRMS)
	m_{33} : a área saturada está relacionada ao armazenamento na camada inferior por meio da distribuição do índice topográfico (válido apenas para TOPMODEL)
4 - Percolação	m_{41} : a água da capacidade de campo até a saturação está disponível para percolação (PRMS)
	m_{42} : percolação definida pelo teor de umidade na camada inferior, a percolação será mais rápida quando a zona inferior estiver seca (SACRAMENTO)
	m_{43} : controle de zona saturada: a água desde o esvaziamento até a saturação está disponível para percolação (VIC)
5 - Evapotranspiração	m_{51} : ponderação: a evapotranspiração em cada camada do solo depende da fração relativa das camadas superior e inferior do solo
	m_{52} : sequencial: a evapotranspiração potencial é satisfeita primeiro pela evaporação da camada superior do solo, e qualquer demanda residual é satisfeita pela evaporação da camada inferior do solo.
6 -Fluxo intermediário	m_{61} : Sem fluxo intermediário (TOPMODEL e ARNO/VIC)
	m_{62} : Com fluxo intermediário. Função linear de armazenamento livre na camada superior (SACRAMENTO E PRMS)
7 - Propagação	m_{71} : sem propagação
	m_{72} : propagação por meio de uma distribuição Gama com parâmetro de forma = 3

A percolação (q_{12}) representa o movimento vertical da água entre as camadas superior e inferior do solo, e, em todos os casos, é representada com dois parâmetros. O fluxo

intermediário (q_{if}) ocorre na zona não saturada do solo e pode não ser modelado, como no caso dos modelos TOPMODEL e ARNO/VIC, ou como uma constante que multiplica a saturação da camada superior.

A propagação em bacias refere-se ao cálculo dos fluxos no tempo e no espaço dentro da bacia. O objetivo da propagação da bacia hidrográfica é transformar a chuva efetiva em vazão, semelhante aos métodos de propagação em reservatório e canal (PONCE, 1994).

O FUSE foi utilizado, por exemplo, em Lane et al. (2019) para calibrar parâmetros de 1.000 bacias na Grã-Bretanha utilizando 4 estruturas; por Saavedra et al. (2022), para calibrar parâmetros de 3 bacias no Peru usando 78 estruturas; por Prietro et al. (2021), em uma bacia na Espanha utilizando as 624 estruturas; em Konapala et al. (2020), para calibrar parâmetros de 4 estruturas em 671 bacias do CAMELS-US; em Clark et al. (2011), para estudar 5 estruturas em uma bacia na Nova Zelândia; e em 24 bacias no Reino Unido utilizando 78 modelos (COXON et al., 2014).

5.2.3 Regras de associação: Apriori

Existem diversas técnicas de mineração de dados disponíveis na literatura (CHEN; HAN; YU, 1996; CHEUNG et al., 1996). Uma dessas técnicas são as regras de associação, destacando-se o algoritmo Apriori (AGRAWAL; SRIKANT, 1994). Em grandes bancos de dados, um dos desafios é descobrir associações importantes entre os atributos, tal que a presença de alguns deles implique na presença de outros. O objetivo, então, é encontrar todas as regras de associação relevantes entre os itens, do tipo X (antecedente) $\Rightarrow Y$ (consequente).

Para tratar desta questão Agrawal et al. (1993) propuseram um modelo matemático no qual as regras de associação geradas devem atender a um suporte e confiança mínimos especificados pelo decisor. O suporte corresponde à frequência com que ocorrem os padrões em toda a base de dados, ou seja, é a porcentagem dos itens da base que possuem X e Y , enquanto a confiança é a proporção dos registros que incluem X e que também incluem Y . Assim, se a confiança é de 60%, a regra pode ser lida como: em 60% das vezes que acontece X , também acontece Y .

Já muito aplicado em sistemas de recomendação (RAHMAN; DASH, 2017; WU et al., 2020), nos últimos anos o algoritmo Apriori tem encontrado aplicação nas ciências naturais. Por exemplo, o algoritmo foi utilizado na análise de risco de inundações no estado do Arizona, onde os índices de chuva acumulados em 3 e 6 horas foram associados com o risco de inundações (ZHONG et al., 2020). Em Li et al. (2022), os pesquisadores propuseram uma versão otimizada do algoritmo Apriori para encontrar fatores que levam aos deslizamentos de terra. De maneira similar, Chen et al. (2023) e Li et al. (2023) identificaram fatores que levam a desastres geológicos.

5.3 Metodologia

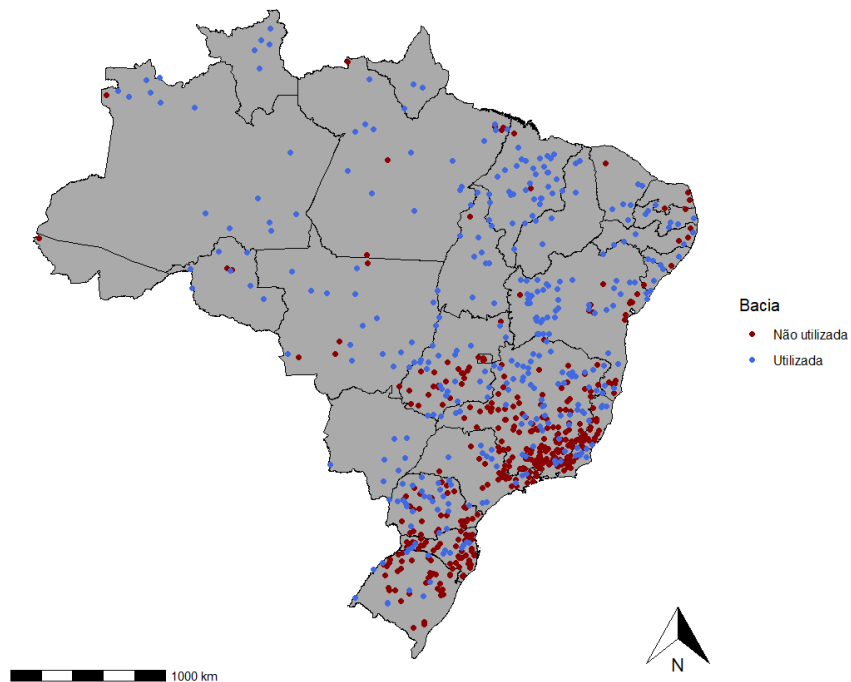
As estimativas de vazão diárias são realizadas no Brasil por meio da média dos níveis de água instantâneos registrados às 7 e 17 horas. Por isso, para bacias com tempos de concentração significativamente inferior a 24 horas, pode-se observar durante a estação chuvosa oscilações expressivas dos níveis de água entre registros temporais subsequentes. Esse fato tem potencial para enviesar a análise dos modelos, visto que a inferência dos volumes escoados frente aos registros de chuva diária acumulada não será representativa do balanço hídrico real.

Por esse motivo, buscou-se bacias em que o tempo de concentração superasse 24 horas, em uma tentativa de mitigar os erros na representação do volume. A estimativa do tempo de concentração se deu por meio da Equação de Kirpich, utilizando relações derivadas da geomorfologia fluvial para estimar a declividade média e o comprimento do curso d'água principal (lei de Hack; WILLEMIN, 2000), a partir da Equação (4):

$$A = \left(\frac{tc}{0,39} \right)^{1,56} * 4,9 \quad (14)$$

Na qual A é a área da bacia em km² e tc o tempo de concentração em horas. Ao especificar um tempo de concentração superior a 24 horas, optou-se pela seleção de bacias com área de drenagem igual ou superior a 3.250 km². Tal restrição resultou em 351 bacias, apresentadas na Figura 35.

Figura 35 - Bacias utilizadas neste estudo
Bacias disponíveis e utilizadas no CABra



Esse conjunto de bacias foi utilizado para calibrar os parâmetros no *framework* flexível do FUSE, em suas 624 estruturas, por meio dos critérios multiobjetivos apresentados no Capítulo **Erro! Fonte de referência não encontrada.**, buscando-se parâmetros do modelo que otimizem a métrica de desempenho e atendam as 5 assinaturas hidrológicas com um erro relativo máximo de 5%, assumindo que tal restrição permita encontrar conjunto de parâmetros hidrológicamente consistentes, ou seja, em que as respostas otimizadas decorram da presunção dos motivos corretos (GUPTA; WAGENER; LIU, 2008).

Empregou-se a função objetivo que agrega a componente do NSE e a função sigmoide das máximas diferenças relativas das assinaturas. Cada bacia foi utilizada para calibrar os parâmetros em cada uma das 624 estruturas em 5.000 simulações.

Os modelos foram treinados nos 15 primeiros anos hidrológicos (de 01/10/1980 a 30/09/1995) e validados nos 15 anos seguintes (de 01/10/1995 a 30/09/2010). Os 6 primeiros meses foram desconsiderados no cálculo das métricas (período de aquecimento do modelo). Calibrada todos os parâmetros de cada estrutura, buscou-se por aquelas capazes de modelar ou que melhor conseguissem representar cada uma delas. Para cada bacia as seguintes etapas foram desenvolvidas:

1. Buscou-se todas as estruturas que atendiam um NSE mínimo de 0,7 e o erro relativo das assinaturas hidrológicas de no máximo 5%;
2. Não havendo estruturas, relaxou-se o erro máximo para cada assinatura hidrológica para valores de 10%, 15%, 20% e 25% aumentando o erro máximo, seguindo essa faixa de valores, até que se encontre pelo menos uma estrutura.
3. Não havendo estruturas, relaxou-se em conjunto o erro máximo para cada assinatura hidrológica para os valores de 10%, 15%, 20%, 25% e o NSE mínimo para valores de 0,65, 0,60, 0,55 e 0,5, sempre em pares, ou seja, o relaxamento do erro máximo relativo das assinaturas para 15% acompanha um NSE mínimo de 0,60;
4. Para todas as estruturas que atingiram a métrica mínima, utilizou-se a mesma lógica para avaliar a validação, estendendo o erro máximo das assinaturas para 30% e o NSE mínimo para 0,45;
5. Persistindo a ausência de estruturas que atendam os critérios mínimos essa bacia passou novamente pelo SCE-UA multiobjetivo, mas agora com 15.000 simulações, voltando ao passo 1; e
6. Ao final, caso haja mais de uma estrutura que atenda aos critérios, escolheu-se a(s) estrutura(s) com menor número de parâmetros.

Com isso gerou-se um banco de dados com as estruturas de modelos conceituais do FUSE que conseguiram modelar cada uma dessas bacias. A partir disso, buscou encontrar associações entre as características físicas e as assinaturas hidrológicas com os mecanismos utilizados em cada estrutura por meio do algoritmo Apriori. Para tal, toda a base de dados foi categorizada em 5 classes de igual frequência, sendo os valores classificadas em: Muito Baixo, Baixo, Médio, Alto e Muito Alto.

Por fim, se agrupou as estruturas (categóricas) adequadas para cada bacia, em 5 grupos, utilizando o *K-Modes* (CHATURVEDI; GREEN; CAROLL, 2001). Esse algoritmo é uma variante do *K-Means* projetado para lidar com dados categóricos. No *K-Means*, a similaridade entre pontos de dados é medida usando a distância euclidiana, a qual é adequada para dados contínuos ou numéricos. Em contraste, o *K-Modes* usa uma medida de similaridade diferente, adequada a dados categóricos. A medida de similaridade mais comumente usada no *K-Modes* é a distância de Hamming (NOROUZI; FLEET; SALAKHUTDINOV, 2012), que conta o número de categorias que são diferentes entre dois pontos de dados. O "centroide" de um *cluster*

no *K-Modes* é o valor modal de cada categoria nos pontos de dados atribuídos ao *cluster*, ou seja, a categoria mais comum.

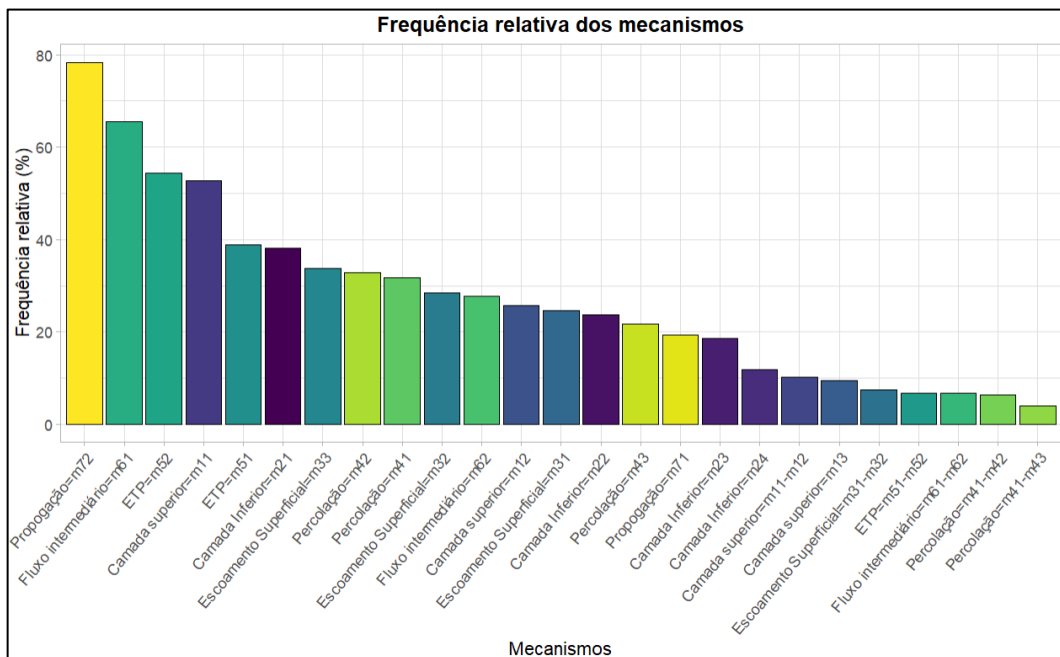
5.4 Resultados

Das 351 bacias calibradas, 26 (7,4%) não atingiram o desempenho mínimo nas métricas em nenhuma das 624 estruturas. Dessas 325, 28 não obtiveram o desempenho mínimo na fase de validação, ou seja, em 15,4% das bacias não foi possível encontrar uma estrutura conceitual no FUSE capaz de representá-la. Essa relação de bacias, no qual obteve-se uma estrutura que atendesse métricas mínimas é superior, por exemplo, a Lane et al. (2019) para Grã-Bretanha.

A estrutura mais frequente, presente em 18 bacias, apresenta a configuração da forma $(m_{12}, m_{21}, m_{33}, m_{42}, m_{51}, m_{62}, m_{72})$, a segunda estrutura, com 17 bacias tem os mecanismos $(m_{11}, m_{21}, m_{31}, m_{43}, m_{51}, m_{62}, m_{72})$. A frequência relativa dos mecanismos, nas bacias que puderam se calibradas, é apresentada no gráfico da Figura 36. Desses dados é possível analisar que:

- Utilizar a propagação (m_{72}) foi necessária em 78% das bacias. Uma vez que se limitou a base de bacias àquelas de maior área, esse era um comportamento, de certa forma, esperado;
- Por outro lado, em 66% das estruturas não se considerou a camada de fluxo intermediário (m_{61}). Adicioná-la eleva a complexidade do modelo (o número de parâmetros) sem que se observe um ganho significativo;
- Há uma preferência para o mecanismo mais simples de evapotranspiração sequencial (m_{52} ; 54%). Mas adotar uma ponderação entre os reservatórios, adicionando um novo parâmetro no modelo (m_{51}), fez-se necessário em 39% das bacias; e
- Modelar a camada superior como um reservatório com uma única camada de estado (m_{11}) atendeu a 53% das bacias, mostrando que parâmetros adicionais nesse mecanismo só foram necessários em uma porção menor de bacias.

Figura 36 – Frequência relativa dos mecanismos nas estruturas selecionadas.



A seguir são apresentadas as associações obtidas para cada mecanismo. A comparação com os *clusters* utiliza aqueles obtidos no Capítulo **Erro! Fonte de referência não encontrada.** Para cada processo será apresentado o mapa com os mecanismos em cada bacia, assim como as regras de associação em forma de tabela e de diagrama. Na tabela, são indicadas as métricas de confiança e suporte, e neste caso é indicado também o número de bacias. A tabela sumaria as regras mais relevantes, sob o propósito de se identificar uma explicação física para a associação. Cada regra da tabela apresenta seu índice em cada figura do diagrama. A tabela também discrimina o intervalo de cada classe em que há uma associação. No diagrama a cor da regra está relacionado com a confiança daquela regra dentro das associações obtidas no mecanismo, de maneira que quanto mais vermelho maior a confiança.

5.4.1 Camada superior do solo

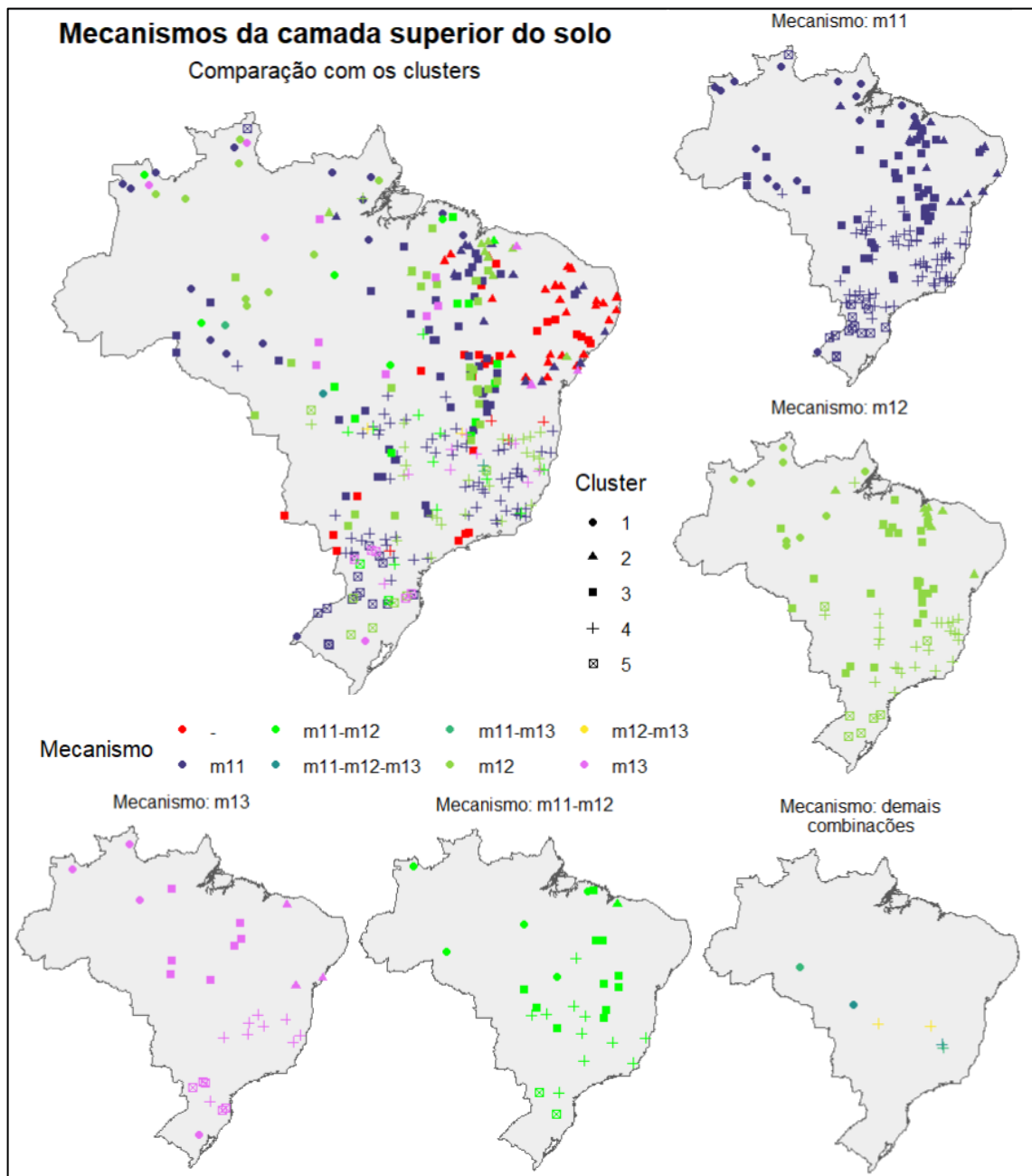
A Figura 37, a seguir, apresenta o mapa com a localização das bacias modeladas por cada um dos mecanismos da camada superior do solo. O formato do símbolo destaca o agrupamento no qual cada bacia pertence. Das 296 bacias em que se obteve alguma estrutura, 91% delas podem ser modeladas com os mecanismos m_{11} ou m_{12} . A complexidade que o PRMS impõe (um parâmetro a mais) a essa camada só se torna relevante a um pequeno número de bacias (28).

As regras obtidas para esses mecanismos podem ser visualizadas na Figura 37 e Tabela 6. As associações com o mecanismo m_{11} (TOPMODEL, ARNO/VIC) se deram em bacias com elevado teor de argila no solo, elevada declividade da bacia, taxas de mudança e subida altas e reduzida porosidade do solo. Esses fatores indicam bacias com mais dificuldade para infiltração, e, portanto, poderiam ser representadas como uma estrutura mais simples com somente uma variável de estado. O m_{12} teve associações envolvendo a porcentagem de argila no solo alta e baixa porosidade do subsolo, apontando situações um pouco melhores para a infiltração do que o m_{11} . Outras associações relevantes do m_{12} envolvem cobertura urbana muito baixa e elevada cobertura de arbustos, situações que sugerem solos mais propícios a retenção de água. É interessante notar a relação com NDVI alto para o inverno, ou seja, a vegetação permanece saudável mesmo no período seco. Essas situações levaram a representar a camada superior dividida em reservatórios acima e abaixo da capacidade de campo (m_{12} ; SACRAMENTO).

Tabela 6 – Associações relevantes obtidas como consequência a camada superior do solo

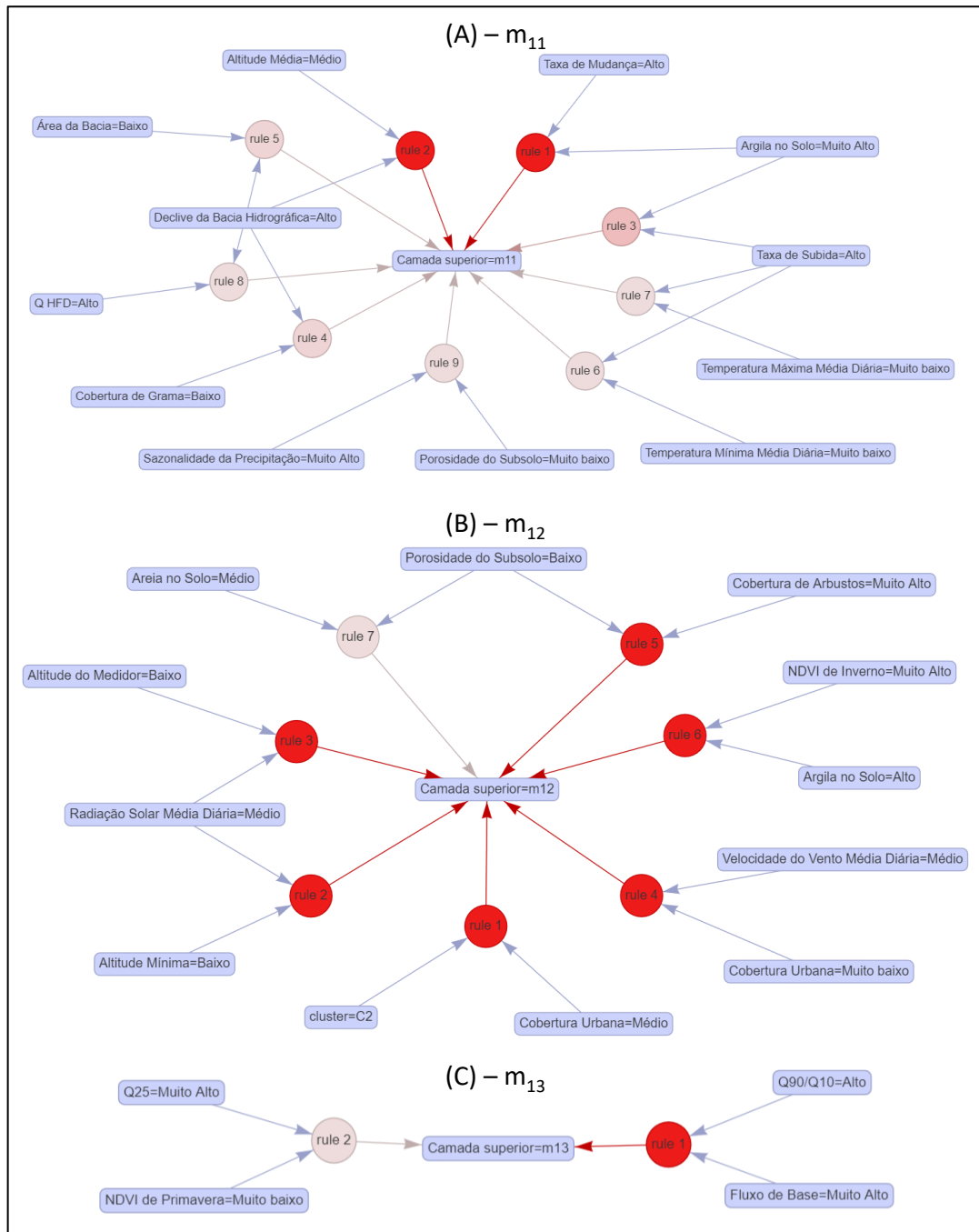
Regra	Antecedente	Precedente	Suporte	Confiança
1-A	Argila do solo (%) muito alta $\geq 35,7$ e Taxa de mudança alta = [0,108;0,168)	m_{11}	0,05 (16)	1,00
2-A	Cobertura de grama (%) baixa = [19,9;30,2) e Declividade média da bacia (%) alta = [8,2;13,1)	m_{11}	0,05 (16)	0,94
9-A	Sazonalidade da precipitação (%) muito alta $> 0,96$ e Porosidade do subsolo muito baixa $< 0,05$	m_{11}	0,05 (15)	0,94
4-B	Vel. do vento média diária (m/s) média = [1,15;1,24) e Cobertura urbana (%) muito baixa $< 0,05$	m_{12}	0,02 (6)	0,86
5-B	Porosidade do subsolo baixa = [0,05;0,11) e Cobertura de arbustos (%) muito alta $> 24,9$	m_{12}	0,02 (6)	0,86
6-B	NDVI no inverno muito alto $> 0,76$ e Argila no solo (%) alta = [31,8;35,7)	m_{12}	0,02 (6)	0,86
1-C	Índice do esc. de base muito alto $> 0,90$ e Q90/Q10 alto = [0,17;0,28)	m_{13}	0,01 (4)	0,80
2-C	NDVI de primavera muito baixo $< 0,66$ e Q25 muito alto $> 0,17$	m_{13}	0,02 (5)	0,71

Figura 37 – Mapa dos mecanismos utilizados para modelar a camada superior do solo



O mecanismo m_{13} (PRMS) foi necessário em somente 28 bacias. Por esse motivo, há uma maior dificuldade em se encontrar associações. Ainda assim, houve uma associação com bacias de elevado índice do escoamento de base. Para essas bacias foi relevante dividir o reservatório da capacidade de campo em recarga e excesso.

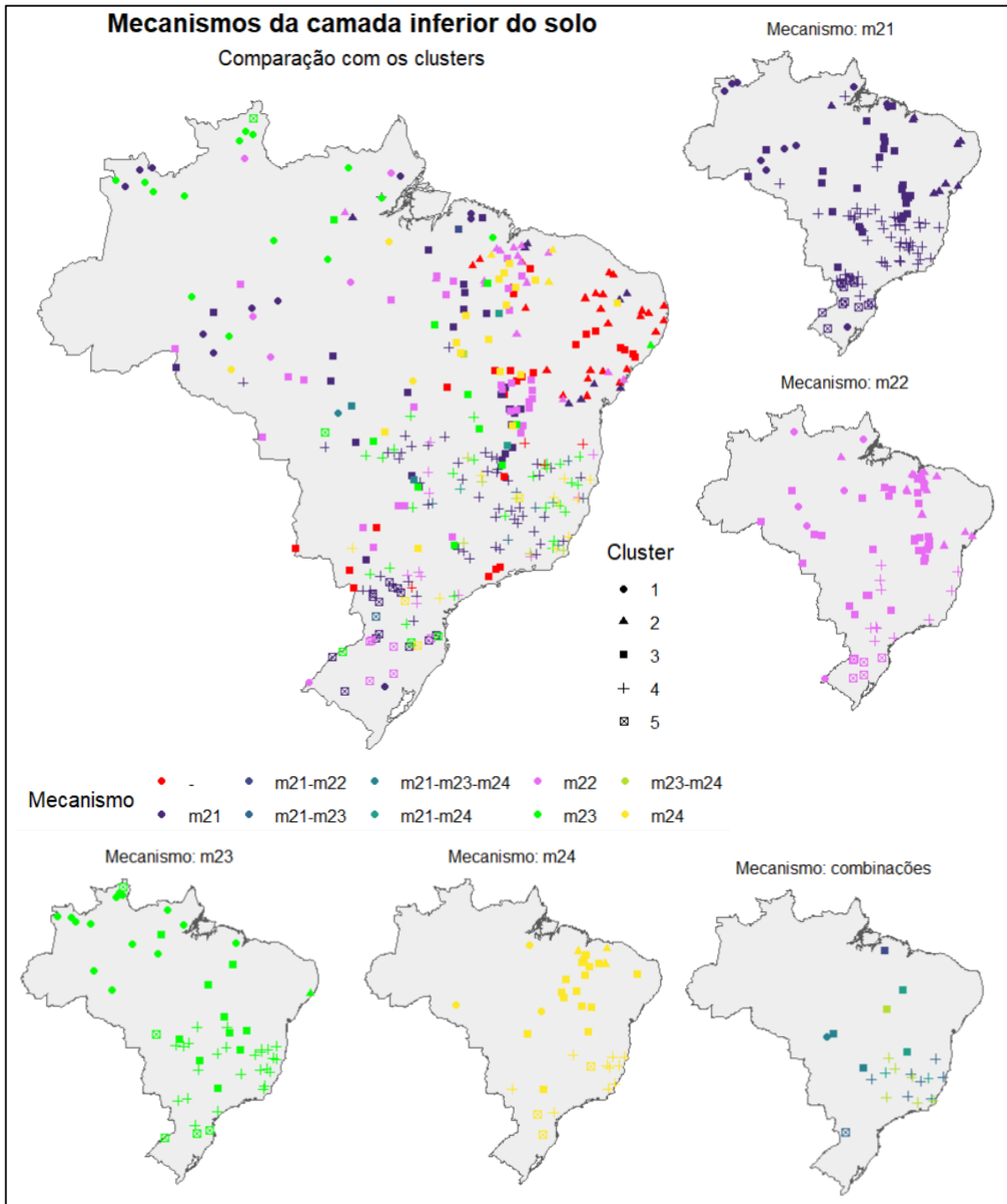
Figura 38 – Regras de associação obtidas para a camada superior do solo



5.4.2 Camada inferior do solo

A Figura 39 apresenta o mapa com a localização das bacias modeladas por cada um dos mecanismos da camada inferior do solo. Novamente, o formato do símbolo destaca o agrupamento no qual cada bacia pertence.

Figura 39 – Mapa dos mecanismos utilizados para modelar a camada inferior do solo



Ao todo, 38% das bacias foram modeladas com o mecanismo m_{21} , 24% com o mecanismo m_{22} e 19% e 12% com os mecanismos m_{23} e m_{24} , respectivamente. Em 8% mais de um mecanismo foi capaz de representar o processo. A utilização do m_{22} impõe ao modelo mais 3 parâmetros.

As regras obtidas para esses mecanismos podem ser visualizadas na Figura 40 e Tabela 3. A regra 6-A, por exemplo, teve um suporte de 0,05, ou seja, os antecedentes

que se associam ao mecanismo m_{21} aconteceram em 15 bacias. Em 88% das bacias em que a cobertura urbana é alta e a cobertura de grama é baixa o mecanismo utilizado é o m_{21} .

O mecanismo m_{21} é o mais restritivo em termos de capacidade de geração de escoamento de base. Talvez por esse motivo esse mecanismo relacionou-se com solos duros, menos porosos, com baixa cobertura de grama e alta cobertura urbana. Tais regras indicam que, em bacias menos permeáveis, o processo da camada inferior e a geração do escoamento de base tendem a ser mais restritivos, limitando o reservatório para o escoamento de base.

O m_{22} , por outro lado, é o mecanismo mais complexo dos 4. Ele propõe três reservatórios para representar essa camada: um abaixo da capacidade de campo e outros dois em paralelo. As associações nesse caso foram com bacias com solo profundo, alto índice de escoamento de base e reduzidas taxas de mudança e subida. A profundidade do solo elevada pode sugerir camadas de solo diferentes, levando à representação com reservatórios diferentes. Taxas de mudança ou subida reduzidas indicam que pode ocorrer uma maior infiltração da bacia, exigindo uma maior complexidade desse mecanismo. Por fim, a radiação solar diária muito alta pode pressupor que há energia suficiente para que a evapotranspiração atinja a camada inferior do solo, própria do mecanismo m_{22} usado no SACRAMENTO. É provável que nessas bacias o processo da camada inferior e o escoamento de base sejam processos dominantes.

Diferente dos outros dois mecanismos supracitados, o m_{23} , baseado no PRMS, não permite a evapotranspiração na camada inferior do solo. Antagonicamente aos outros dois, esses mecanismos apresentaram associações com solos pouco profundos e fofos. Aparentemente, essa situação pode indicar que a evapotranspiração ocorre somente na camada aerada. O elevado valor da vazão, tanto nos meses de junho quando de outubro, além de uma vazão mínima anual muito alta, indica que, nessas bacias, outros processos contribuem para o escoamento em períodos sem chuva. Ao aplicar um fator linear sobre o reservatório da camada inferior na geração do escoamento de base, os mecanismos m_{23} pode manter esse escoamento mais regular durante o tempo.

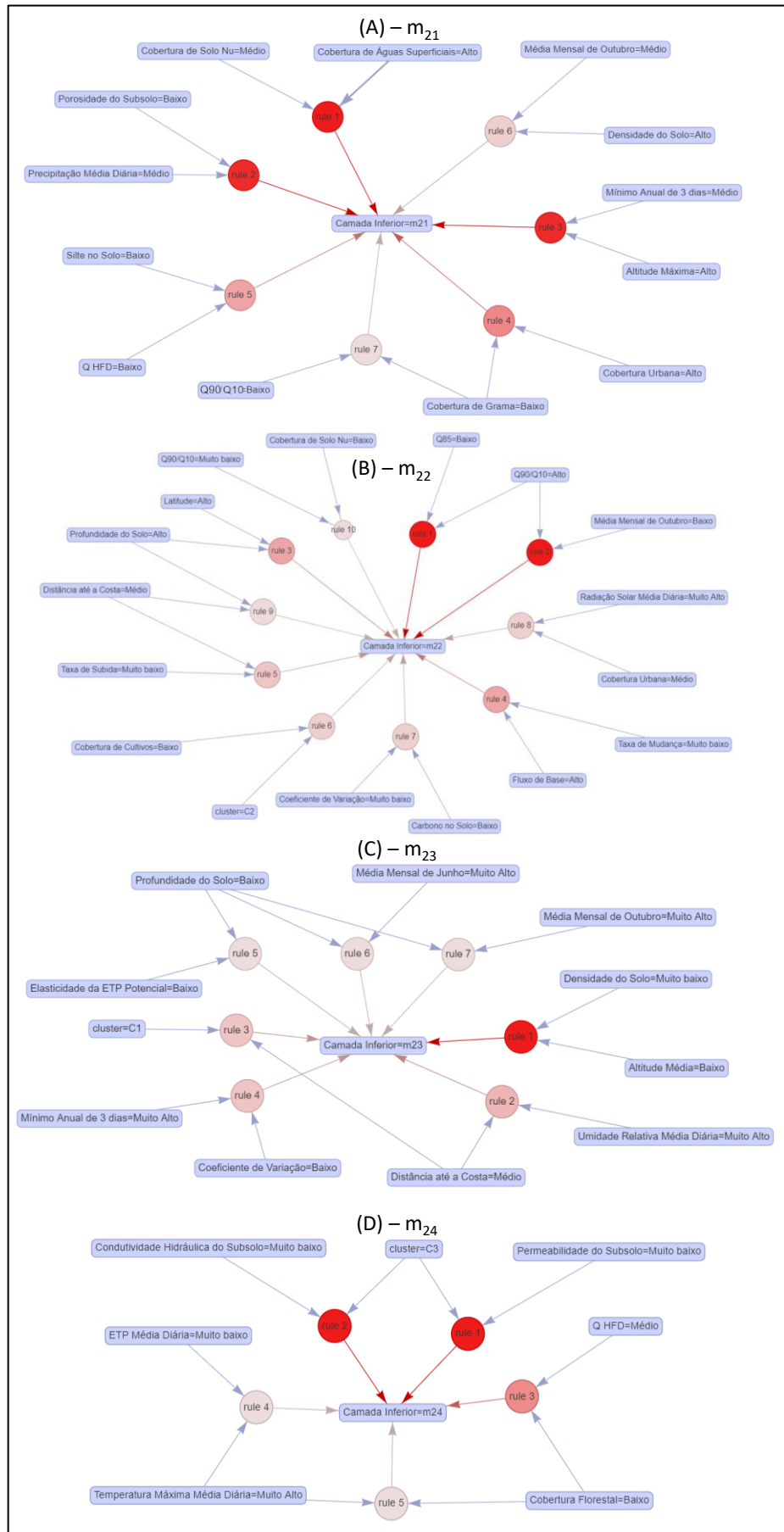
Já o m_{24} , baseado no TOPMODEL, associou-se com solos pouco permeáveis e de baixa condutividade hidráulica. Tais fatores também sinalizam uma maior dificuldade para a evapotranspiração atingir as camadas mais profundas do solo.

Neste processo, o mecanismo m_{23} obteve associações com o *cluster 1*, enquanto o m_{22} com o *cluster 2* e o m_{24} com o *cluster 3*.

Tabela 7 – Associações relevantes obtidas como consequência a camada inferior do solo

Regra	Antecedente	Precedente	Suporte	Confiança
1-A	Cobertura de solo exposto (%) média = [0,19;0,33) e Cobertura de águas superficiais (15) alta = [0,19;0,30)	m_{21}	0,04 (13)	0,93
2-A	Precipitação média diária (mm) média = [3,93;4,28) e Porosidade do subsolo baixa = [0,05;0,11)	m_{21}	0,04 (12)	0,92
6-A	Cobertura de grama (%) baixa = [19,9;30,2) e Cobertura urbana (%) alta = [0,29;0,75)	m_{21}	0,05 (15)	0,88
9-A	Densidade do solo (g/cm^3) alta = [1,36;1,39) e Vazão média de outubro (mm/d) média = [0,5;0,7)	m_{21}	0,05 (13)	0,83
3-B	Latitude alta = [-13;-6,29) e Profundidade do solo (m) alta = [28,4;32,0)	m_{22}	0,03 (8)	0,89
4-B	Taxa de mudança muito baixa < 0,03 e Índice de escoamento de base alto = [0,85;0,90)	m_{22}	0,03 (8)	0,89
8-B	Radiação solar média muito alta > 18,6 e Cobertura urbana (%) média = [0,15;0,29)	m_{22}	0,03 (10)	0,83
1-C	Densidade do solo (g/cm^3) muito baixa < 1,31 e Altitude média da bacia (m) baixa = [334;491)	m_{23}	0,02 (5)	1,00
5-C	Profundidade do solo (m) baixa = [18,2;23,6) e Vazão média de junho (mm/d) muito alta > 1,96	m_{23}	0,03 (8)	0,80
1-D	Permeabilidade do subsolo (m^2) muito baixa < -14 Cluster = 3	m_{24}	0,03 (8)	0,80
2-D	Cond. hidráulica do subsolo (ms^{-1}) muito baixa < -7 e Cluster = 3	m_{24}	0,03 (8)	0,80

Figura 40 – Regras de associação obtidas para a camada inferior do solo



5.4.3 Escoamento superficial

A Figura 41, a seguir, apresenta o mapa com a localização das bacias modeladas por cada um dos mecanismos para o escoamento superficial. A maior parte das bacias (40%) foi modelada com o mecanismo m_{33} (TOPMODEL), no qual o escoamento superficial é gerado somente quando ocorre saturação na camada inferior, como se a camada superior e a inferior funcionassem como dois sistemas em série. Esse mecanismo tem também um menor número de parâmetros, sendo que o m_{31} e m_{32} adicionam um parâmetro a mais no modelo.

A Figura 42 e a Tabela 8 sintetizam as principais regras de associação encontradas para os três mecanismos para modelar o escoamento superficial. Os mecanismos m_{33} e m_{32} apresentaram associações antagônicas. Enquanto o m_{33} se fez presente em bacias com maior vazão e menor permeabilidade e condutividade do subsolo, o mecanismo m_{32} foi associado a bacias com vazões mais baixas e maior permeabilidade e condutividade do solo. O m_{33} (TOPMODEL) ocorreu também em bacias com profundidade do solo baixa e declividade muito elevada. Em bacias com menor aptidão para a infiltração, a camada inferior do solo funciona mais próxima da camada superior, influenciando a geração do escoamento superficial. É importante lembrar que o TOPMODEL usa um índice topográfico para representar a variabilidade espacial da umidade do solo em uma bacia hidrográfica. Esse índice é baseado na hipótese de que a umidade do solo é uma função da topografia local. Dessa forma, esse mecanismo, na forma utilizada no FUSE, parece mais relevante em bacias mais íngremes.

Em um solo mais permeável, a sua capacidade de absorver água pode levar a menos escoamento superficial, e, portanto, vazões altas menores. Dessa maneira, há um indicativo de que o escoamento superficial não é um processo dominante nessas bacias, podendo ser representado como relação linear do quanto o reservatório da camada superior se encontra preenchido (m_{32} ; PRMS). Soma-se a isso a associação desse mecanismo a bacias com elevados índice de escoamento de base.

Por outro lado, o mecanismo m_{31} (ARNO/VIC) se fez presente quando a porosidade do subsolo é muito alta, a razão entre o Q_{90} e Q_{10} é elevada e a evapotranspiração potencial média diária é muito alta, o que sugere que, nessas bacias, outros processos

interferem no escoamento superficial, e sua modelagem exige mais complexidade. Em vez de assumir uma relação linear simples entre a umidade do solo e a área saturada, o m_{31} utiliza a distribuição de Pareto para representar a relação não-linear entre a umidade do solo e a topografia. Isso significa que, em diferentes níveis de umidade, diferentes frações da bacia contribuirão para o escoamento superficial. Isso pode permitir uma representação mais dinâmica e apropriada na forma como a umidade do solo influencia no escoamento superficial.

Figura 41 – Mapa dos mecanismos utilizados para modelar o escoamento superficial.

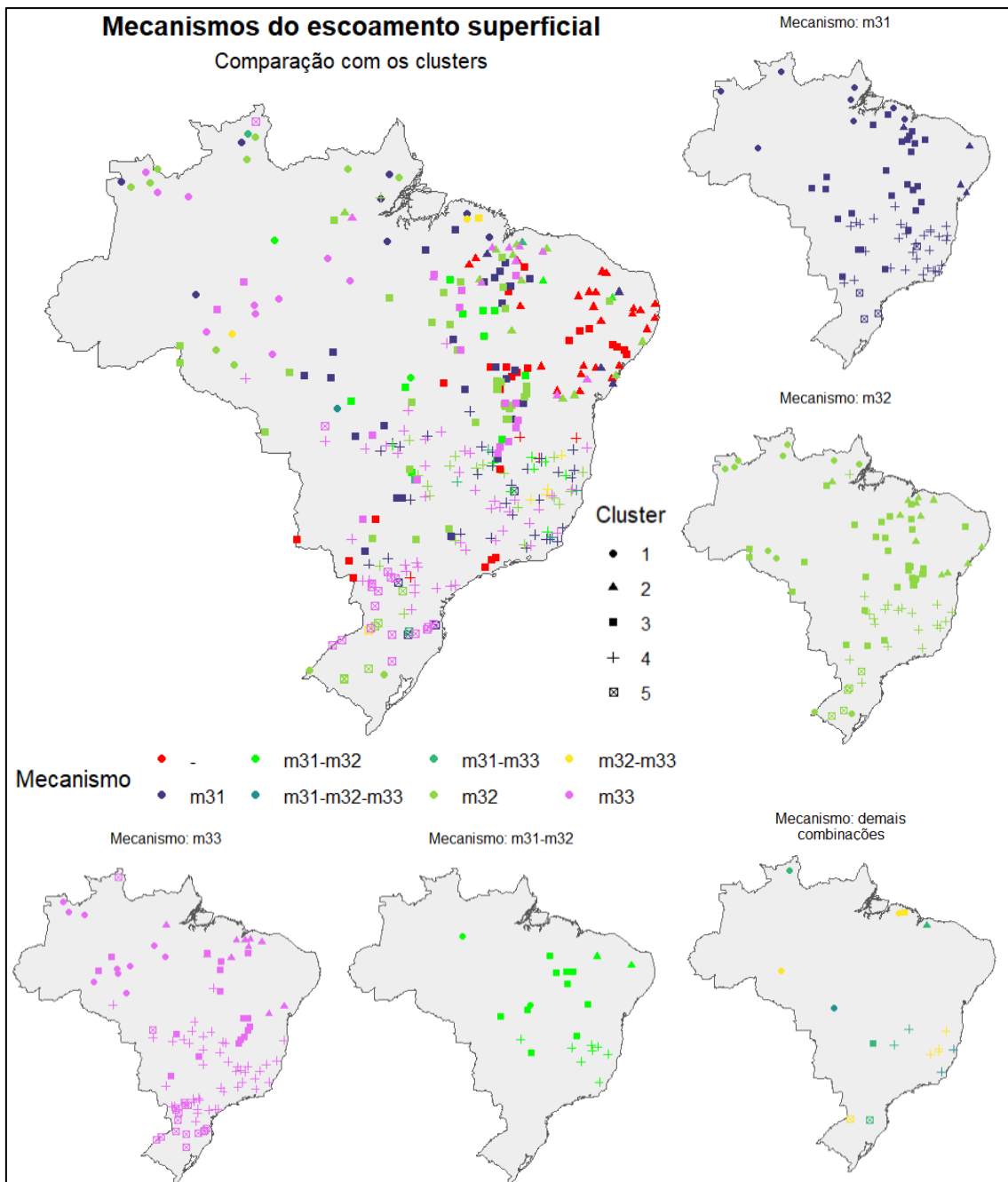


Figura 42 – Regras de associação obtidas para o escoamento superficial

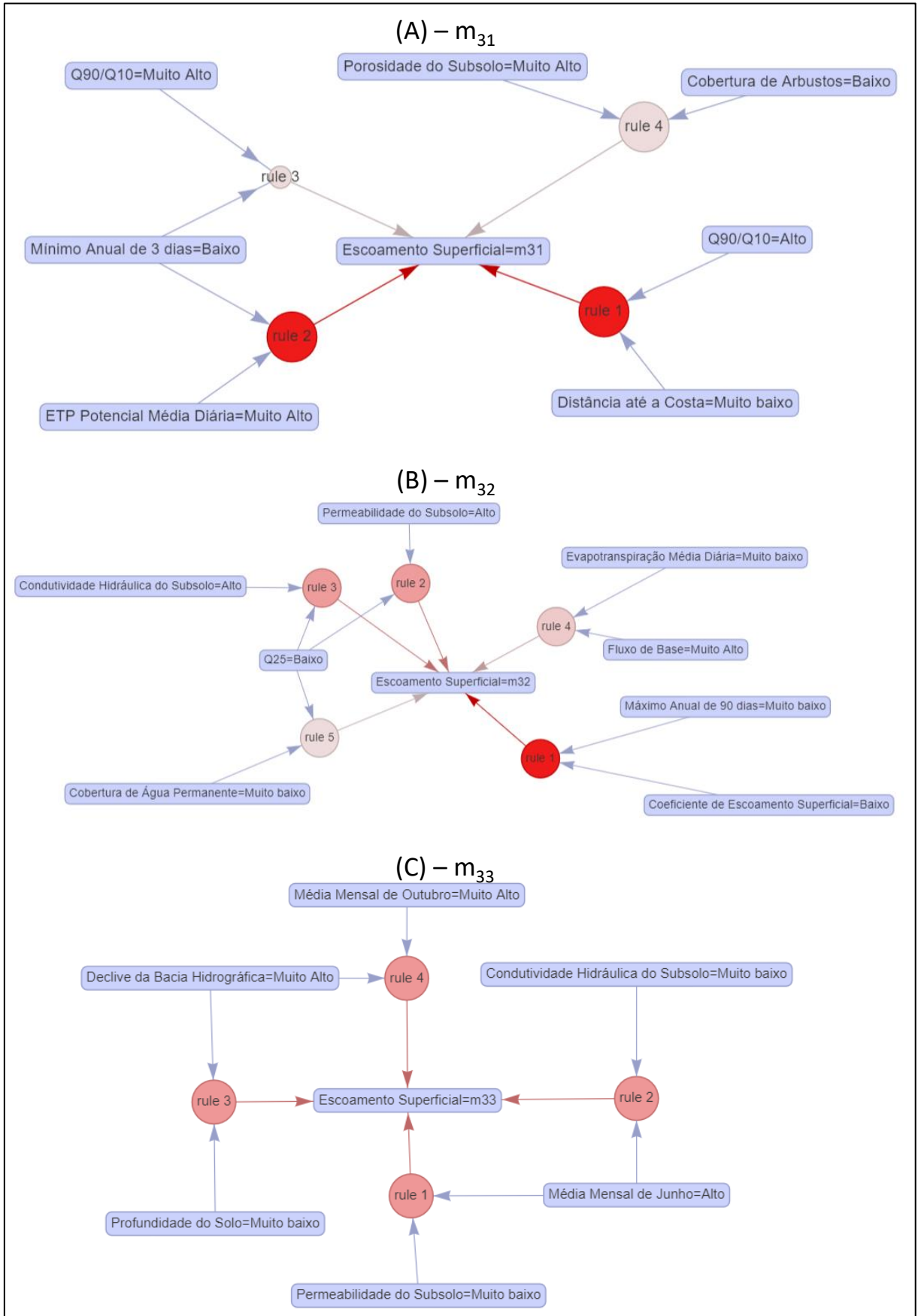


Tabela 8 – Associações obtidas como consequência o escoamento superficial

Regra	Antecedente	Precedente	Suporte	Confiança
3-A	Vazão mín. anual 3 dias (mm/d) baixa = [0,16;0,31) e Q90/Q10 muito alto $\geq 0,27$	m ₃₁	0,02 (6)	0,88
4-A	Porosidade do subsolo muito alta $> 0,19$ e Cobertura de arbustos (%) baixa = [8,47;13)	m ₃₁	0,02 (6)	0,88
2-B	Permeabilidade do subsolo (m ²) alta = [-13,2,-13) e Q25 (mm/dia) baixa = [0,74;1,36)	m ₃₂	0,05 (20)	0,71
3-B	Condut. hidráulica do subsolo alta (ms ⁻¹) = [-6,24;-6,00) e Q25 (mm/dia) baixa = [0,74;1,36)	m ₃₂	0,05 (20)	0,71
5-B	Vazão máx. anual 90 dias (mm/d) muito baixa $\leq 1,1$ e Coef. de esc. superficial baixo = [0,20;0,29)	m ₃₂	0,05 (19)	0,73
1-C	Permeabilidade do subsolo (m ²) muito baixa ≤ -14 e Vazão média de junho (mm/d) alta = [0,99;1,96)	m ₃₃	0,05 (14)	0,88
2-C	Cond. hidráulica do subsolo (ms ⁻¹) muito baixa $< -6,99$ e Vazão média de junho (mm/d) alta = [0,99;1,96)	m ₃₃	0,05 (14)	0,88
3-C	Profundidade do solo (m) muito baixa ≤ 18 e Declividade da bacia (%) muito alta $\geq 13,1$	m ₃₃	0,05 (14)	0,88
4-C	Vazão média de outubro (mm/d) muito alta $\geq 1,31$ e Declividade da bacia (%) muito alta $\geq 13,1$	m ₃₃	0,05 (14)	0,88

5.4.4 Percolação

O processo de percolação foi modelado em 32% das bacias com o mecanismo m₄₁ (PRMS), 33% como m₄₂ (SACRAMENTO), 22% como m₄₃ (VIC) e os 13% restantes como uma combinação dos três mecanismos. Todos os mecanismos possuem o mesmo número de parâmetros, e seus mapas estão apresentados na Figura 44.

Tabela 9 – Associações obtidas como consequência a percolação

Regra	Antecedente	Precedente	Suporte	Confiança
1-A	Carbono no solo (%) muito alto $> 24,7$ e Taxa de subida média = [0,08;0,12)	m ₄₁	0,03 (9)	0,90
2-A	NDVI outono muito baixo $< 0,57$ e Coeficiente de variação da vazão alto = [1,22;1,61)	m ₄₁	0,03 (10)	0,83
3-A	Areia no solo (%) muito baixa $< 46,2$ e Taxa de subida média = [0,08;0,12)	m ₄₁	0,03 (9)	0,81
1-B	NDIVI outono muito alto $> 0,70$ e Taxa de mudança alta = [0,11;0,17)	m ₄₂	0,03 (10)	0,83
2-B	NDIVI outono muito alto $> 0,70$ e Declividade (%) muito baixa $< 4,0$	m ₄₃	0,02 (7)	0,88
1-C	Permeabilidade do subsolo (m ²) muito baixa ≤ -14 e Vazão média de junho (mm/d) alta = [0,99;1,96)	m ₄₃	0,02 (7)	0,88
2-C	Latitude (°) baixa = [-21,6;-17,1) e Porosidade do subsolo muito baixa $< 0,25$	m ₄₃	0,02 (5)	0,83

As regras, apresentadas na Figura 44 e Tabela 9, associam o mecanismo m₄₁ com solos muito arenosos e com elevada matéria orgânica. E assim como o mecanismo m₄₃, há associações com os índices do NDVI no outono e verão com valores muito

baixos, sugerindo que a vegetação tem dificuldade em encontrar água, até mesmo no verão. A diferença entre os mecanismos está na quantidade de água disponível para percolação na camada superior do solo. Enquanto o m_{41} não permite a percolação da capacidade de campo, o m_{43} pode levar a camada superior ao ponto de murcha. Dessa forma, os dois mecanismos permitem a percolação ocorra mais facilmente, mas o m_{43} permite o esvaziamento do reservatório. Coincidentemente, o m_{43} também foi associado a subsolos muito porosos e bacias de muito baixas declividades, fatores que podem aumentar percolação, enquanto a elevada matéria orgânica, associada ao m_{41} , pode ajudar a reter a água e propiciar condições próximas à capacidade de campo.

O mecanismo m_{42} , por outro lado, modelado no SACRAMENTO, parametriza a percolação de maneira não linear, sendo controlada pela camada inferior, ocorrendo mais rápida quando esta camada está seca. Tal artifício pode permitir que mais água permaneça na camada superior, disponível para a vegetação, resultando, nesse caso, em um NDVI no outono agora muito elevado.

Figura 43 – Mapa dos mecanismos utilizados para modelar a percolação

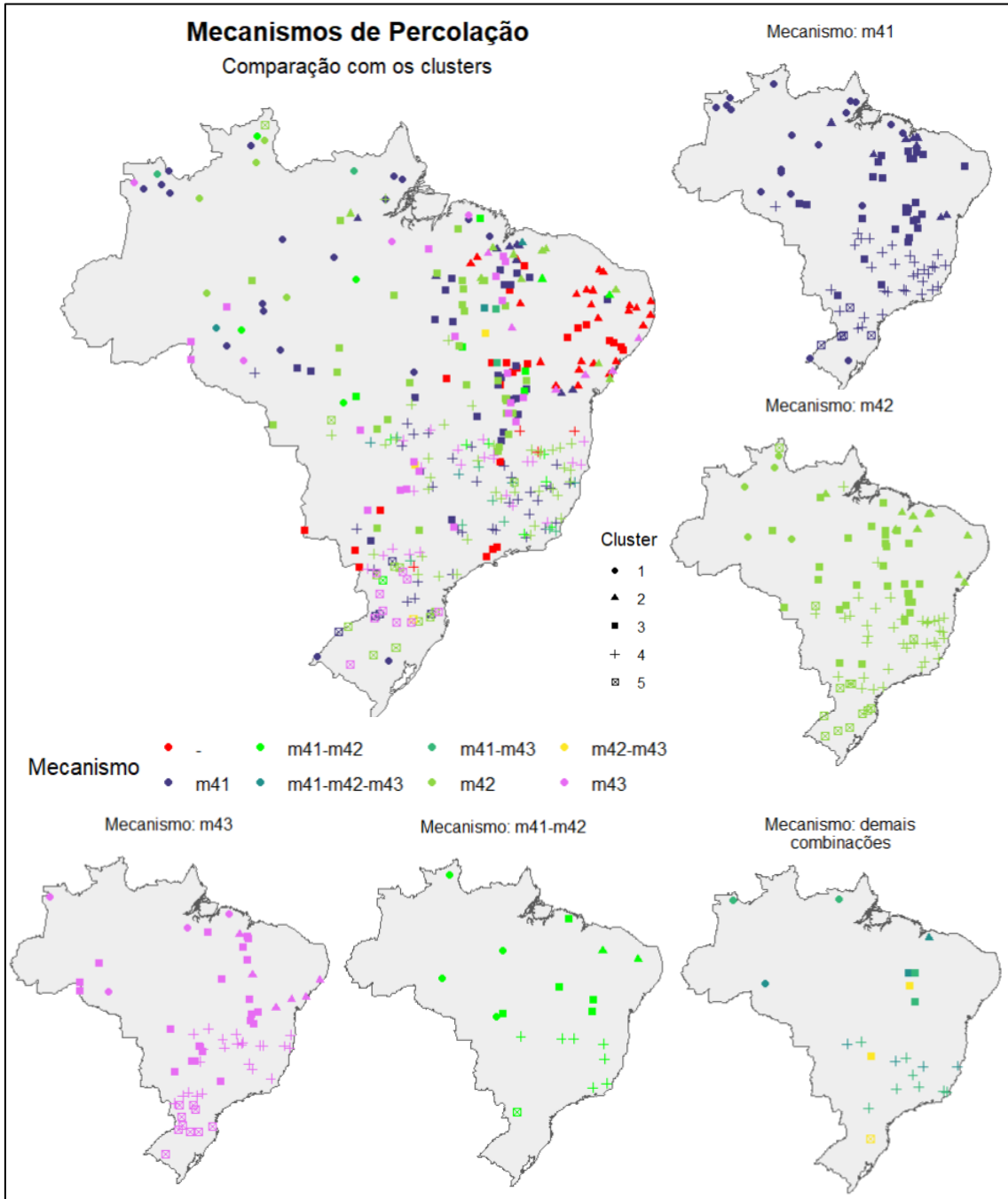
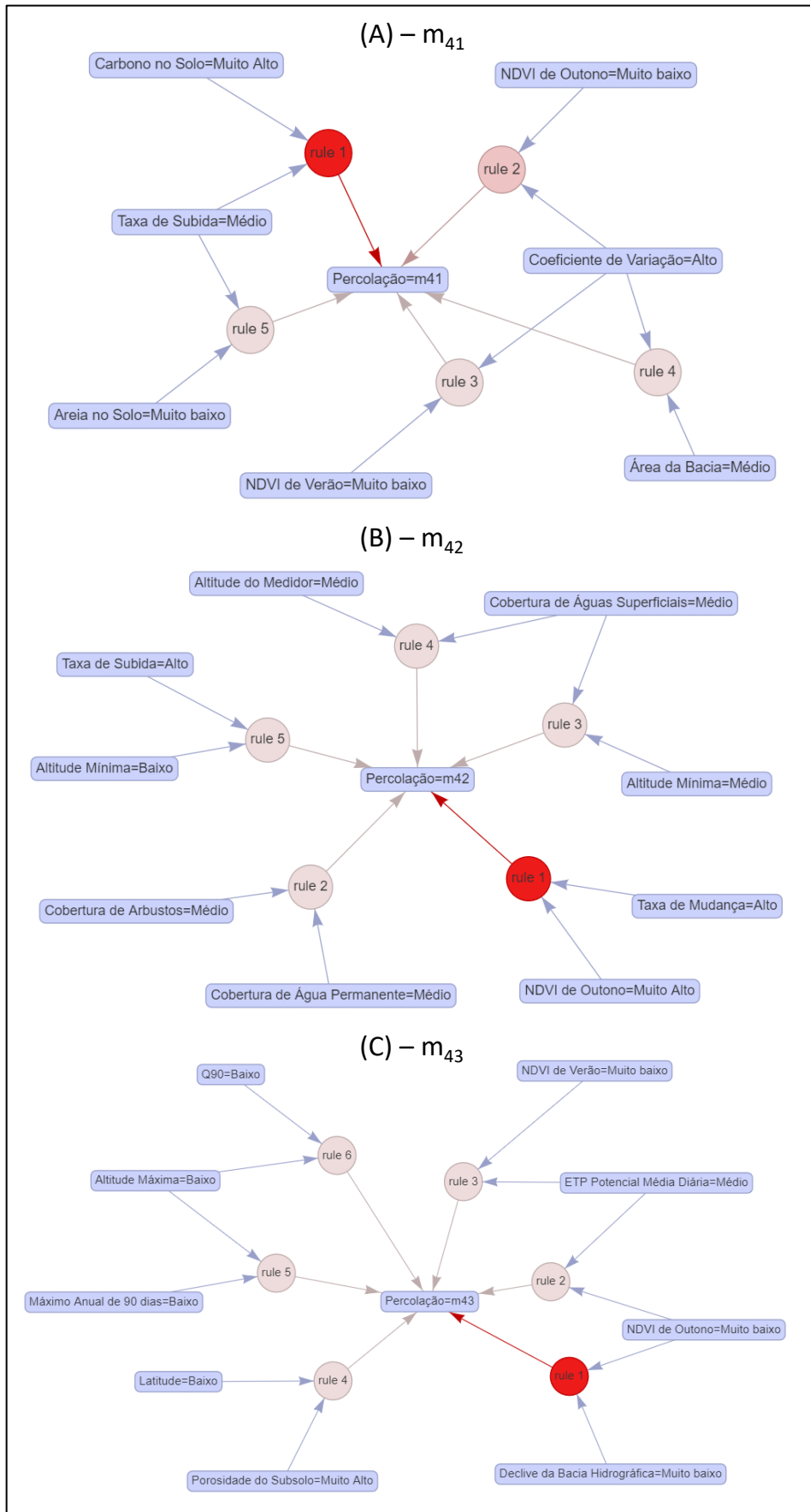
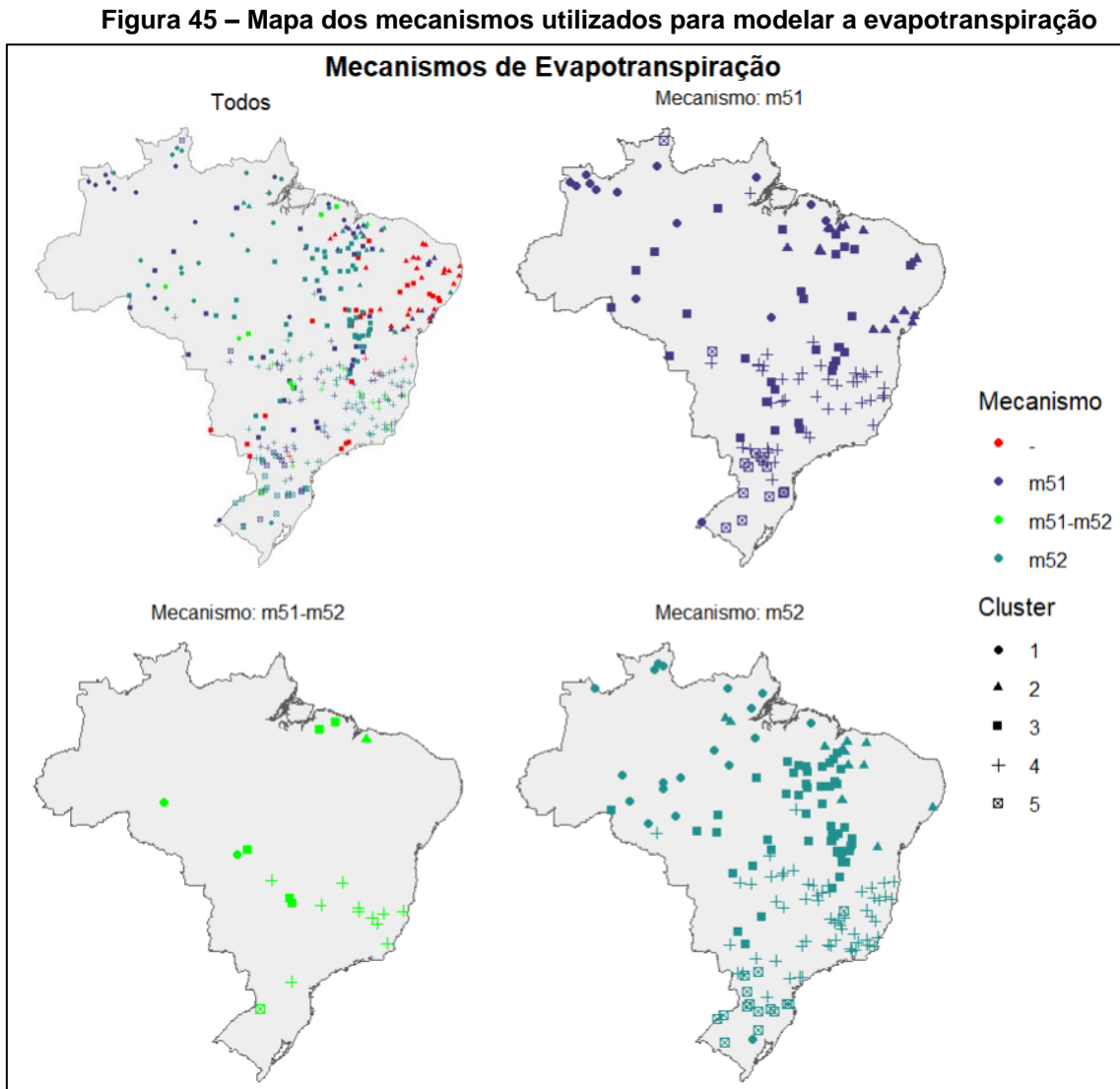


Figura 44 – Regras de associação obtidas para a percolação



5.4.5 Evapotranspiração

Para a evapotranspiração, 55% das bacias foram representadas pelo mecanismo m_{52} , 40% pelo m_{51} e 15% por ambos. A Figura 45, a seguir, apresenta o mapa com a localização desses mecanismos. Há de se destacar que o mecanismo m_{52} apresenta um parâmetro a menos.



As regras de associação são apresentadas na Figura 46. A Tabela 10 sumariza as principais associações. O mecanismo m_{52} teve associações com taxas de mudanças da vazão elevadas, índice de aridez muito alto, muita cobertura urbana, profundidade do solo baixa e porosidade do solo muito baixa. Em contrapartida, o mecanismo m_{51} associou-se a taxas de mudanças reduzidas, cobertura elevada de arbustos, índice de aridez médio, e permeabilidade e condutividade hidráulica do subsolo muito altas.

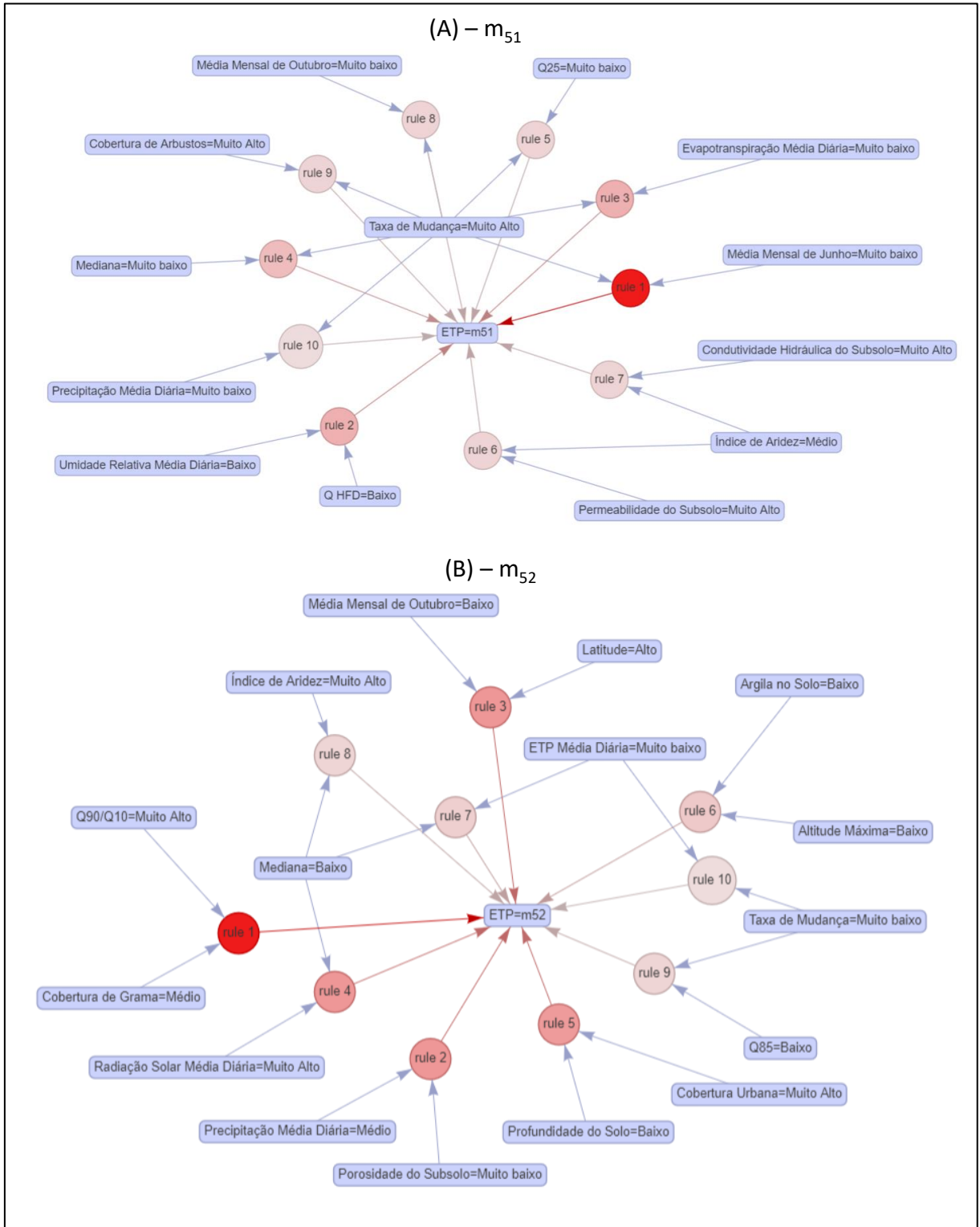
No mecanismo m_{52} a evapotranspiração potencial incide primeiro na camada superior do solo, e na sequência, caso ainda haja potencial, retira da camada inferior. As associações mostram uma preferência para esse mecanismo em bacias mais secas, urbanas, de solo raso e de baixa porosidade. Pode-se deduzir que, em bacias com menor capacidade de armazenamento de água no solo, a evapotranspiração ocorre principalmente perto da superfície.

Por outro lado, o mecanismo m_{51} aplica a evapotranspiração potencial de maneira ponderada sobre as duas camadas do solo. Isso pode indicar uma situação em que a água está mais prontamente disponível e a evapotranspiração ocorre em uma profundidade maior. Consequentemente, esse mecanismo associou-se a taxas de mudanças reduzidas, índice de aridez médio e permeabilidade e condutividade hidráulica do subsolo muito altas.

Tabela 10 – Associações obtidas como consequência a evapotranspiração

Regra	Antecedente	Precedente	Suporte	Confiança
1-A	Taxa de mudança muito alta $> 0,16$ e Q25 (mm/dia) $< 0,74$	m_{51}	0,06 (21)	0,81
6-A	Índice de aridez média = $[1,25;1,39)$ e Cond. hidráulica do subsolo (ms^{-1}) muito alta > -6	m_{51}	0,06 (17)	0,81
7-A	Índice de aridez média = $[1,25;1,39)$ e Permeabilidade do subsolo (m^2) muito alta ≥ -13	m_{51}	0,06 (17)	0,81
2-B	Precipitação média diária (mm) média = $[1,56;2,37)$ e Porosidade do subsolo muito baixa $< 0,047$	m_{52}	0,06 (19)	0,95
5-B	Cobertura urbana (%) muito alta $\geq 0,75\%$ e Profundidade do solo (m) baixa = $[18,2;23,6)$	m_{52}	0,06 (18)	0,95
8-B	Índice de aridez muito alto $\geq 1,76$ Vazão mediana (mm/d) = $[0,45;0,76)$	m_{52}	0,06 (19)	0,95
9-B	Taxa de Mudança $< 0,03$ e Q85 (mm/dia) = $[0,21;0,38)$	m_{52}	0,06 (19)	0,90

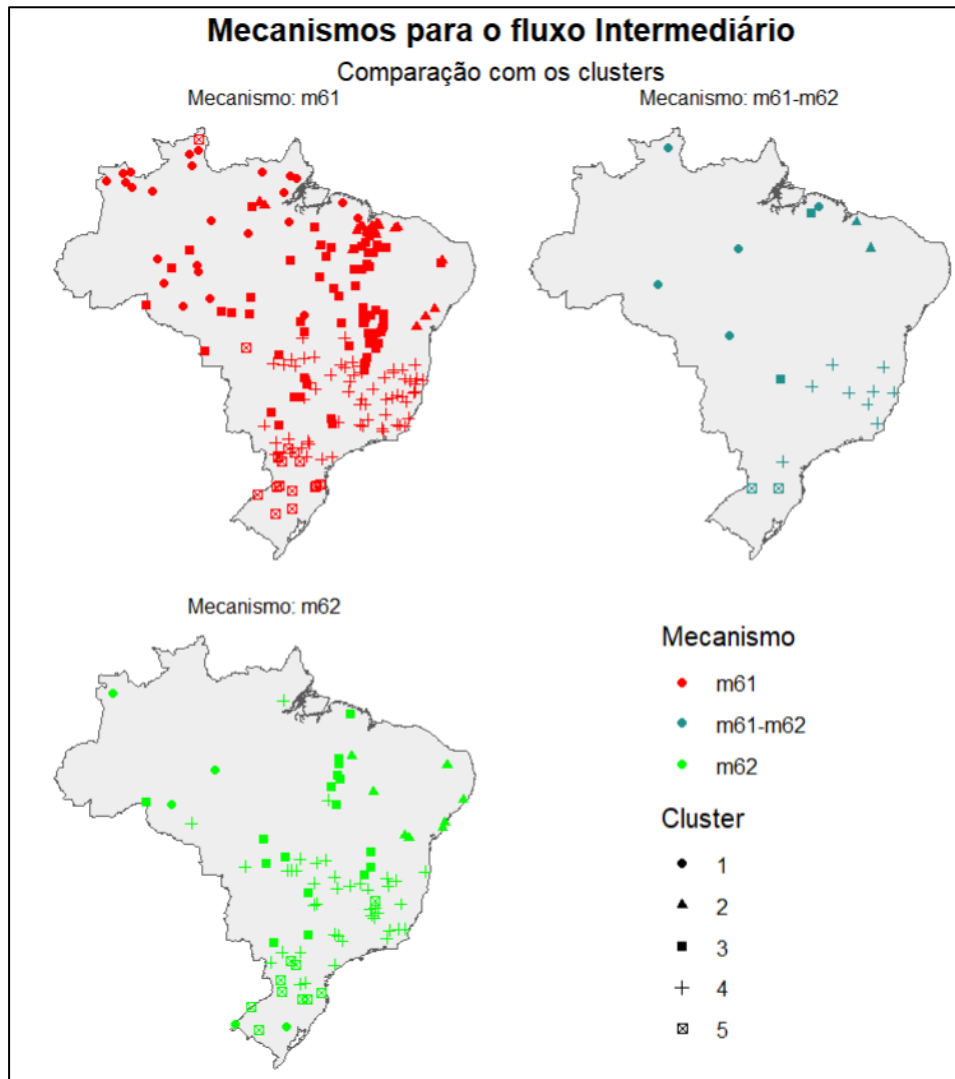
Figura 46 – Associações que levaram aos mecanismos de Evapotranspiração



5.4.6 Fluxo intermediário

O fluxo intermediário foi necessário em 82 bacias. Em 194, adicionar o fluxo intermediário não melhorou o desempenho, e nas outras 20 pode-se utilizar ambas as alternativas. Para esses casos isso significa que o parâmetro adicional que esse processo exige compensa-se simplificando algum dos outros processos.

Figura 47 – Mapa dos mecanismos utilizados para modelar o fluxo intermediário



As regras de associação apresentadas na Tabela 11 e Figura 48 mostram que não modelar o fluxo intermediário (m_{61}) está associado um uma taxa de mudança muito baixa e/ou o escoamento de base muito alto. A taxa de mudança muito baixa associada ao mecanismo m_{61} ocorre em 19% da base de dados (suporte), e 18% quando o índice do escoamento de base é muito alto. Uma vez que a bacia tem taxa de mudança muito baixa, em 93% delas ocorreu o m_{61} (confiança), já para o

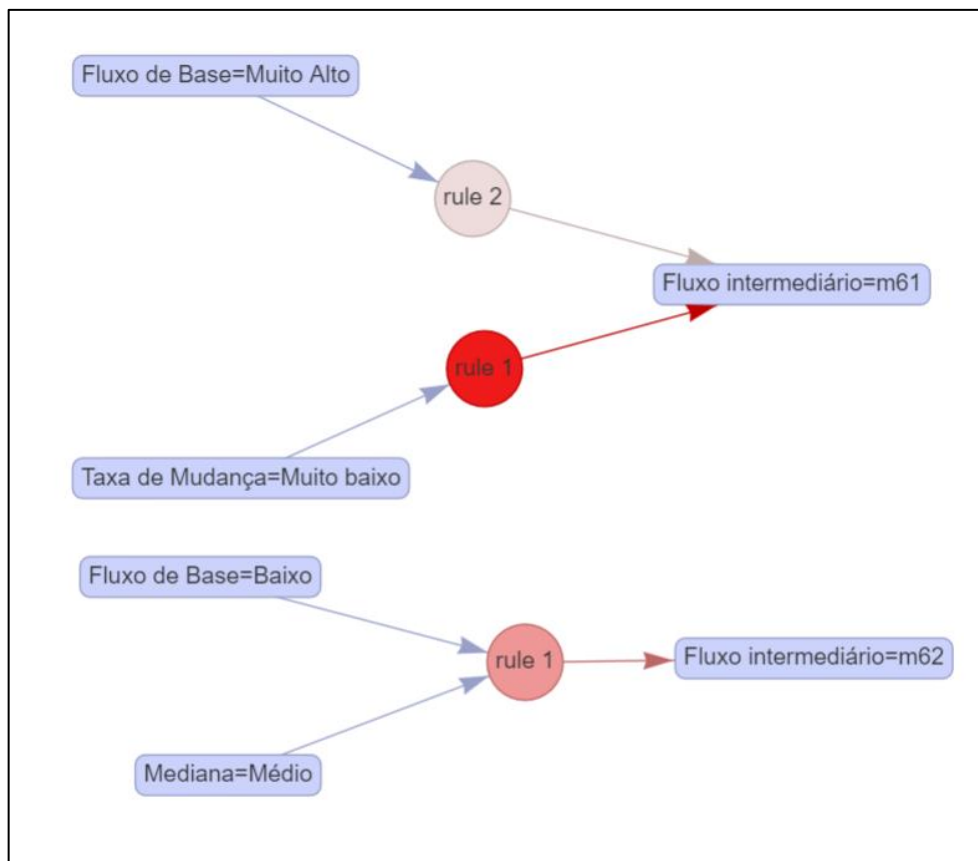
escoamento de base muito alto essa relação é de 90%. Por outro lado, a necessidade de se modelar o fluxo intermediário (m_{62}) parece estar relacionado com baixos índices de escoamento de base e mediana da vazão média anual em valores médios.

Se a taxa de mudança é muito baixa, isso pode indicar que a água pode ter tempo suficiente para infiltrar-se no solo e atingir o lençol freático, contribuindo para o escoamento de base, em vez de permanecer na zona não saturada e contribuir para o fluxo intermediário. Por outro lado, em bacias com baixo índice de escoamento de base, pode ser necessário modelar o fluxo intermediário para obter uma representação precisa do comportamento da bacia.

Tabela 11 – Associações obtidas como consequência o fluxo intermediário

Antecedente	Precedente	Suporte	Confiança
Taxa de mudança $\leq 0,03$	m_{61}	0,19	0,93
Índice do escoamento de base $\geq 0,90$	m_{61}	0,18	0,90
Vazão mediana (mm/dia) = [0,76;1,02) e Índice do escoamento de base = [0,74;0,80)	m_{62}	0,04	0,82

Figura 48 – Associações que levaram ao mecanismo de fluxo intermediário



5.4.7 Propagação em bacias

Em 78% das bacias foi necessário modelar a propagação (m_{72}). Em 19% esse mecanismo não é necessário (m_{71}) O mapa desses mecanismos é apresentado na Figura 49. As associações para esse processo são apresentadas na Figura 50 e Tabela 12.

Em suma, o mecanismo m_{72} dominou quando a sazonalidade da precipitação, taxa de mudança e taxa de subida são altas, a razão da Q90 sobre a Q10 é muito alta e há valores médios pra a Q25 e o NDVI na primavera. A não necessidade de se modelar a propagação foi associada a bacias de baixa declividade e bacias do *cluster 1*.

Figura 49 – Mapa dos mecanismos utilizados para modelar a propagação

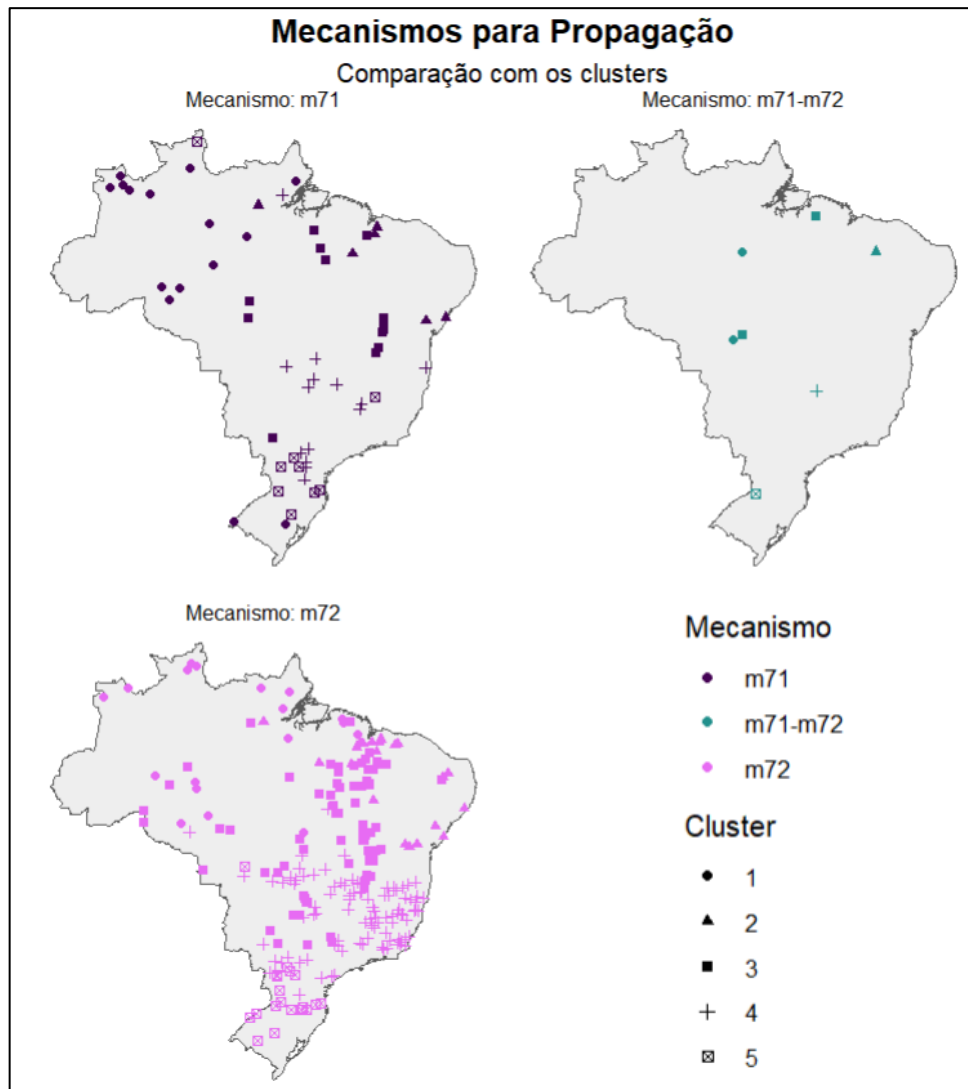


Figura 50 – Associações que levaram ao mecanismo de propagação

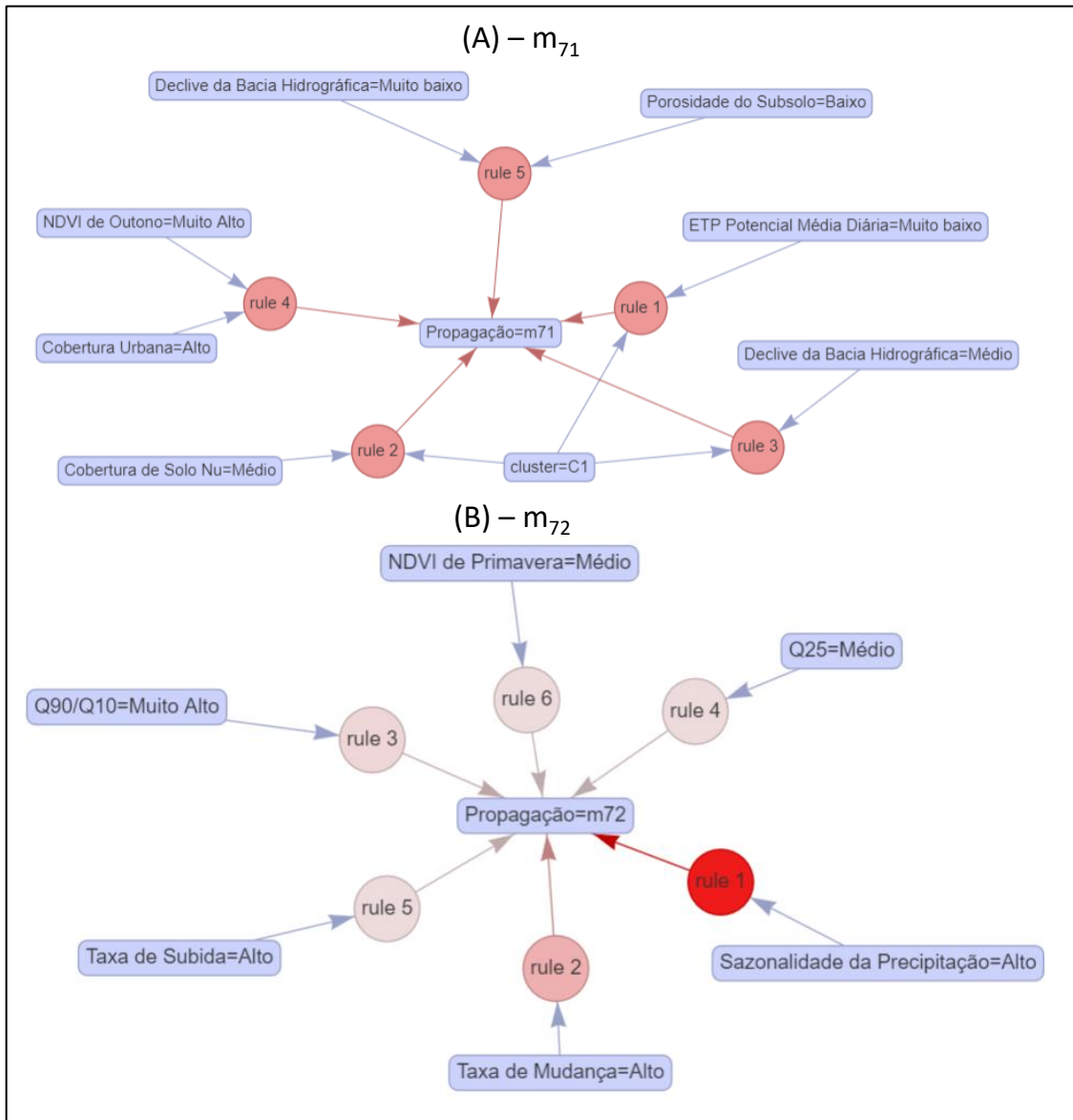


Tabela 12 – Associações obtidas como consequência a propagação

Antecedente	Precedente	Suporte	Confiança
Sazonalidade da precipitação = [0,81;0,96)	m ₇₂	0,19	0,97
Taxa de Mudança = [0,11,0,17)	m ₇₂	0,18	0,93
Q25 = [1,35;1,88)	m ₇₂	0,18	0,92
Taxa de subida = [0,12;1,17)	m ₇₂	0,18	0,92
NDVI de Primavera = [0,69;0,72)	m ₇₂	0,18	0,92
Q90/Q10 >= 0,27	m ₇₂	0,19	0,92

5.4.8 Bacias sem estruturas adequadas

Conforme já mencionado, em 55 bacias não se encontrou uma estrutura de modelo conceitual do FUSE que atingisse métricas mínimas na etapa de validação. Essas

bacias são apresentadas na Figura 51. Nenhuma dessas bacias pertencem ao agrupamento um ou cinco, que são os grupos de bacias com maiores vazões. A maior parte dessas bacias pertencem ao agrupamento 3 ou 4.

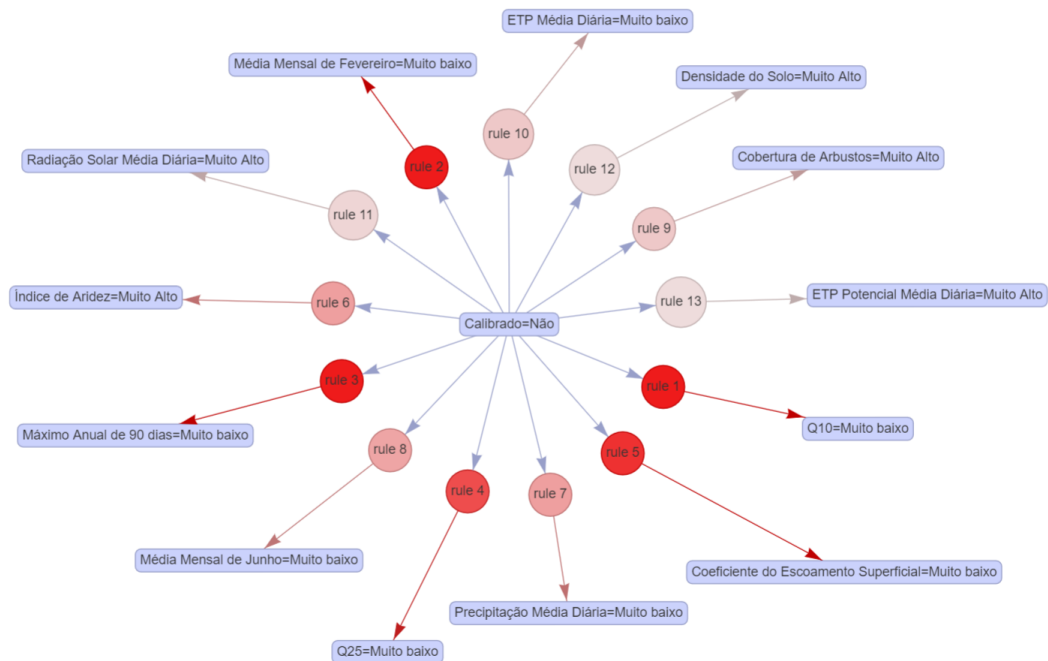
Buscando encontrar as razões para explicar as dificuldades em de modelar essas bacias, buscou-se regras no sentido contrário das anteriormente apresentadas: quais são os precedentes dado que uma bacia não conseguiu ser representada por nenhuma estrutura. As regras de associação obtidas, ilustradas na Figura 52, em sua maioria, caracterizam as bacias do semiárido brasileiro. A dificuldade em se encontrar um modelo para essas bacias parece estar relacionado ao coeficiente de escoamento superficial, vazões e precipitação com valores muito baixos, além de radiação, índice de aridez e ETP potencial com valores elevados.

Em bacias com vazões muito baixas, pequenas variações na precipitação podem ter um grande impacto na saída do modelo. Por outro lado, na maioria dos casos, a evapotranspiração potencial é elevada e a real é muito baixa. Dessa forma, qualquer oscilação das forçantes meteorológicas pode conduzir o modelo a simular vazões significativamente díspares das observadas.

Figura 51 – Mapa das bacias não calibradas,



Figura 52 – Associações que levaram a bacias que não foram representadas em nenhum modelo.



5.4.9 Principais características e assinaturas

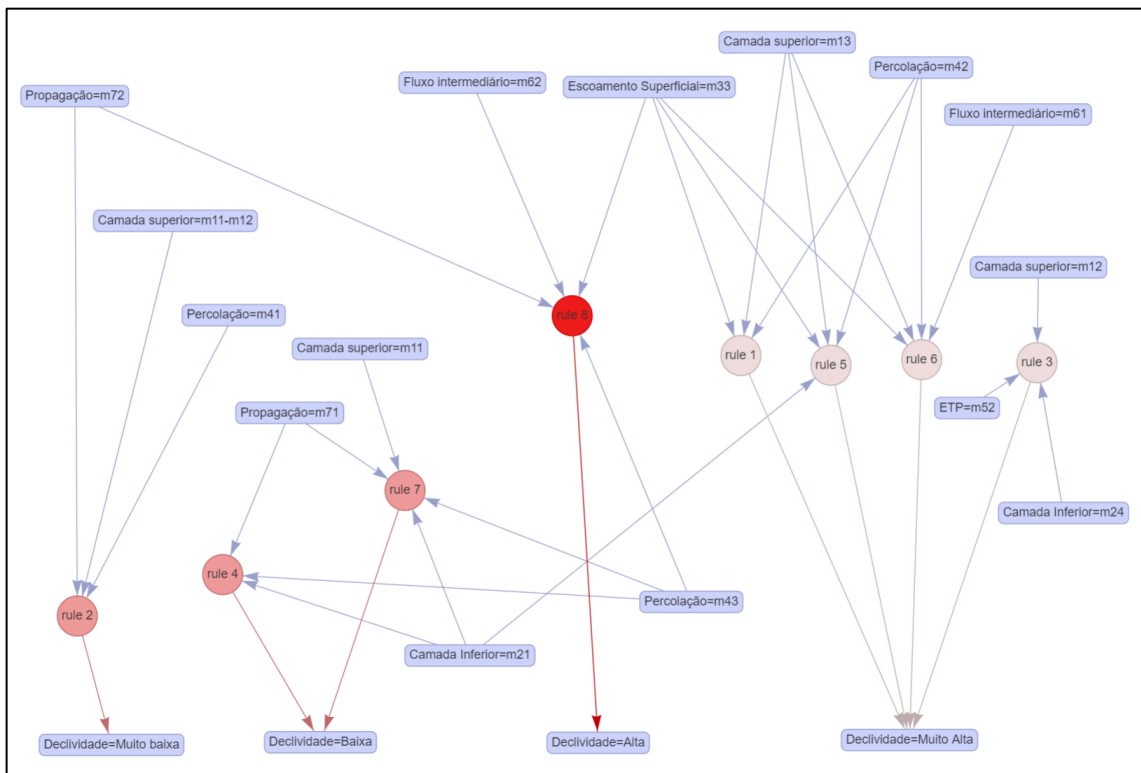
A Figura 53 sumariza as principais características e assinaturas que se associaram a cada um dos mecanismos. As cores representam os mecanismos, conforme a Figura 33.

Já a Figura 54 apresenta as regras de associação entre a combinação de mecanismos e a declividade, que foi a característica que melhor segregou essa combinação. Percebe-se que bacias de menor declividade empregam os mecanismos m_{11} (TOMODEL e ARNO/VIC) ou m_{12} (SACRAMENTO), enquanto o m_{13} (PRMS) associou-se com a declividade elevada. De maneira similar as associações sugeriram o mecanismo de percolação m_{41} (PRMS) para bacias pouco íngremes, o m_{43} (VIC) a partir que se aumenta a declividade e o m_{42} (SACRAMENTO) em bacias muito íngremes.

Figura 53 – Principais associações sumariadas

	Muito baixo	Baixo	Médio	Alto	Muito Alto
Cobertura urbana	m12		m12/m22	m21/m71	
Caborno no solo		m22		m31	m41
Cobertura de arbustos		m31	m42	m24	m12/m51/-
Porosidade do subsolo	m11/m52	m12/m21/m71			m31/m43
Densidade do solo	m23			m21	-
Profundidade do solo	m11/m33	m23		m22	
Permeabilidade do subsolo	m33/m24			m32	m51
Condutividade hidráulica do subsolo	m33/m24			m32	m51
Areia no solo	m41		m12/m22		m24
Q85		m22/m52		m21	
Taxa de mudança	m22/m52/m61			m11/m42/m72	m51
Taxa de subida	m22		m41	m11/m72	
Q90/Q10		m21		m13	m31/m52/m72
Escoamento de base		m62		m22	m13/m32/m61
Mínimo anual 3d		m31	m21		m23
CV	m22	m23		m41/m24	
Q25	m51/sm	m32	m72		m13
Q_HFD		m21	m24	m11	
Área		m11	m41		
Declividade	m43			m11	m33
ETP potencial média	m32/m24	m23			m31/-
Cluster	m23/m71	m12/m22	m24		

Figura 54 – Associação dos mecanismos em relação a declividade



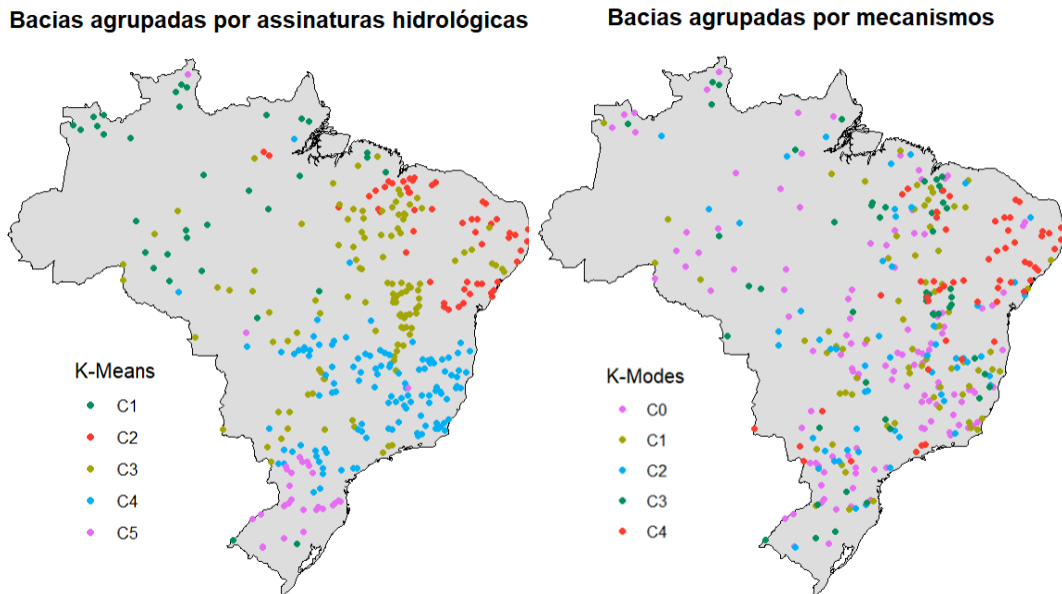
5.4.10 Agrupamentos

Para manter a consistência com o Capítulo **Erro! Fonte de referência não encontrada**, manteve-se o número de agrupamentos desejados em 5 na execução do K-Modes. O *cluster* C0 se destaca pelo bacias no qual há muitas combinações de mecanismos que são capazes de modelá-las. No *cluster* C1 há um predomínio do mecanismo m_{11} . O *cluster* C2 se caracteriza pelos mecanismos m_{42} e m_{51} . No C3 ocorre os mecanismos m_{22} e m_{32} ou m_{33} . Por fim, o C4 agrupou todas as bacias no qual não foi possível encontrar uma estrutura adequada.

A diferença entre o C1 e o C0 é que o C1 tende a usar o mecanismo m_{11} e o C0 agrupa bacias em que mais de um mecanismo na camada superior do solo é possível. Mas o m_{11} acontece em 53% das bacias. Em suma, esses dois *clusters* aglutinam bacias que permitem combinações de mecanismos, mesmo que o C1 tenha preferido um mecanismo mais simples para a camada superior do solo. O C2 pode indicar bacias em que a evapotranspiração é um processo dominante, enquanto o C3 sugere a camada inferior do solo e o escoamento de base como dominantes.

A Figura 55 apresenta o mapa das bacias agrupadas pelos mecanismos e aqueles obtidos no Capítulo **Erro! Fonte de referência não encontrada**, por similaridade de assinaturas hidrológicas. A falta de padrões nas regras de associação reflete na falta de relação entre os agrupamentos e em não se ter uma consistência espacial entre os agrupamentos por mecanismos hidrológicos. Embora haja uma indicação de mecanismos pelas assinaturas hidrológicas e características físicas, elas são sutis. As razões que certos mecanismos do FUSE podem ser preferenciais não podem ser generalizadas, impedindo padrões espaciais.

Figura 55 – Comparação dos agrupamentos obtidos com os mecanismos e os por similaridade de assinaturas hidrológicas



5.5 Conclusão

Este estudo utiliza uma grande base pública de bacias e calibra os parâmetros de centenas de estruturas de modelos conceituais em uma abordagem multiobjetivo utilizando assinaturas hidrológicas. Estudos anteriores precisaram limitar o número de bacias hidrográficas ou o número de modelos (ver, por exemplo, DE BOER-EUSER et al., 2017; KNOBEN et al., 2020; KONAPALA; KAO; ADDOR, 2020; LANE et al., 2019; NIJZINK et al., 2016). As métricas de calibração também receberam pouca atenção em estudos de larga escala. A consideração de vários aspectos hidrográficos permite uma avaliação abrangente do desempenho do modelo (ROJAS-SERNA et al., 2016; YANG et al., 2020), a fim de garantir que as respostas ótimas decorram de razões fisicamente embasadas.

Algumas associações interessantes foram obtidas. Camadas de solo mais complexas (mais parâmetros ou número de reservatórios) está ligada a solos mais profundos ou com mais matéria orgânica. Bacias menos permeáveis podem ter seus processos mais simplificados. O fluxo intermediário, em grande parte das bacias, não precisa ser representado, enquanto a propagação se fez quase sempre necessária. Tais associações podem ajudar a compreender e determinar os processos hidrológicos dominantes.

O estudo também reforça a dificuldade de se calibrar os parâmetros de modelos para bacias áridas, em consonância com a literatura (KNOBEN et al., 2020; KONAPALA; KAO; ADDOR, 2020). Características climáticas extremas são desafiadoras ao modelador. Por exemplo, situações que a evapotranspiração média diária é muito baixa, e evapotranspiração potencial é muito alta. A falta de precipitação e a limitada capacidade de geração de escoamento superficial dificultam a modelagem, porque, na maior parte do tempo, os mecanismos podem estar desativados. A abordagem nesse caso talvez seja de encontrar os gatilhos que levam a ativação de determinados mecanismos e então utilizar modelos que permitam essa oscilação de mecanismos ativos.

Por outro lado, não há um padrão espacial nas estruturas dos modelos ou nas bacias no qual um número elevado ou reduzido de combinações é apropriado. As associações apresentam certas combinações de características e/ou assinaturas que levam a determinado mecanismo, mas que não podem ser generalizadas em relações explícitas. Um número maior de bacias pode ser analisado, e mecanismos pouco significativos podem ser descartados, viabilizando computacionalmente essa busca.

Há ainda a incerteza, não só nos dados hidrométricos, mas nas características físicas das bacias. Esses dados, em grande parte oriundos de produtos de sensoriamento remoto, podem afetar os resultados, em especial a heterogeneidade do solo. Mas outros produtos do sensoriamento remoto podem ser mais bem explorados, não só por possuírem uma melhor resolução espacial, mas por apresentarem também uma variação temporal que pode ser analisada em conjunto. É o caso do NDVI, LAI e umidade do solo. As associações obtidas destacam o efeito sazonal do NDVI e sua obtenção é, por vezes, mais direta do que a quantidade de matéria orgânica no solo, por exemplo. A exploração desses índices pode ser um caminho mais adequado na tentativa de se encontrar padrões nos comportamentos das bacias.

Os estudos hidrológicos têm se esforçado em demasia na tentativa de explicar e encontrar padrões e relações nos comportamentos das bacias. A abordagem mais comum nesse aspecto é regionalizar os parâmetros do modelo (HRACHOWITZ et al., 2013; RAZAVI; COULIBALY, 2013). Mas é severamente criticada tanto usando técnicas de regressão (ARSENAULT; BRISSETTE, 2014; OUDIN et al., 2008) quando

em proximidade espacial (LEBECHEREL; ANDRÉASSIAN; PERRIN, 2016; NERI; PARAJKA; TOTH, 2020) quanto nos padrões espaciais das características físicas (MERZ; TARASOVA; BASSO, 2020). Embora possivelmente existam bacias doadoras potenciais para cada bacia, a abordagem de regionalização depende de cada bacia e não há um padrão espacial (POOL; VIS; SEIBERT, 2021). Dessa forma, encontrar padrões regionais nos parâmetros dos modelos continua sendo um desafio, mesmo que já muito explorado.

De maneira mais incipiente, alguns estudos buscam uma relação entre as características das bacias com a estrutura do modelo conceitual. Embora Knoben et al. (2020) digam que a escolha da parametrização do modelo (ou seja, quais equações são usadas e como os parâmetros são usados dentro delas) determine seus pontos fortes e fracos do modelo, os seus resultados não encontram relações entre as bacias hidrográficas e as características geológicas, topográficas, de cobertura e vegetação. De maneira similar, Spieler et al. (2020) também não encontram qualquer relação entre a estrutura do modelos e as características das 12 bacias por eles estudadas.

A dificuldade de encontrar padrões talvez seja explicada pela alta heterogeneidade das bacias hidrográficas, que são entidades dinâmicas e em evolução, respondendo dinamicamente às forçantes meteorológicas, que também exibem uma significativa variabilidade no espaço e no tempo. As interações entre clima, solos, vegetação e topografia contribuem, assim, para a geração de comportamentos complexos nesses sistemas (SIVAPALAN, 2006).

Há mais de 30 anos vem se questionando se a modelagem hidrológica da vazão com base física é mesmo realista (BEVEN, 1989; GRAYSON; MOORE; MCMAHON, 1992). Os questionamentos se concentram sobre as condições em que as equações são derivadas e a variabilidade espacial da bacia hidrográfica. Nesse sentido, os estudos têm buscado representar a bacia em unidades menores, em teoria mais prováveis de serem homogêneas, por meio de modelos distribuídos (SEMENOVA; BEVEN, 2015). Apesar de sua complexidade e realismo físico, os modelos baseados em processos distribuídos têm um desempenho pior, semelhante ou apenas um pouco melhor do que os modelos conceituais concentrados (por exemplo, KRATZERT et al., 2019b; MENDOZA et al., 2015; REED; MINSKER; GOLDBERG, 2003; SMITH

et al., 2012). Mas a dúvida ainda é se os mecanismos atuais conseguem, de fato, representar/abstrair os processos hidrológicos, e se esses modelos realmente já “nascem mortos”.

Certamente, um fator que complica o problema de avaliar e refinar o comportamento de modelos baseados em processos são as representações fixas da variabilidade espacial, assim como as representações fixas da física do modelo (único conjunto de representações de processo) e valores fixos de parâmetros do modelo. Essas restrições resultam em modelos com capacidade insuficiente para simular adequadamente a heterogeneidade dos processos hidrológicos.

Adicionalmente, a forma tradicional que os processos são abstraídos precisam também ser revistas. Dividir o solo em duas camadas como tem sido feito há décadas não parece ser efetivo. Trocar uma equação por outra, utilizar reservatórios em paralelo (em cima ou embaixo), modelar o fluxo intermediário etc., pode fazer pouca diferença, justamente porque esses mecanismos podem não ser capazes de representar todos os processos dominantes nas bacias nas diferentes escalas temporais associadas a esses processos. Essa lacuna é, por vezes, preenchida adicionando um parâmetro em um mecanismo qualquer.

É preciso aceitar também que nem sempre as entradas e saídas do processo são conhecidas em todas as bacias. A necessidade de se fechar o balanço de massas assume que toda a entrada se dá pela chuva e que toda saída é obrigatoriamente evapotranspiração ou vazão. Sabe-se, no entanto, que o fluxo subterrâneo não respeita os limites topográficos e que as diferentes escalas temporais envolvidas podem constituir-se em um impedimento para verificá-la explicitamente.

Por outro lado, é inegável que essas abstrações dos processos hidrológicos em modelos conceituais funcionem para muitos casos. Mas é oportuno que esses modelos evoluam aprendendo com os dados disponíveis. Talvez a abordagem mais adequada para entender o comportamento das bacias em relação a estrutura do modelo seja a hierárquica progressiva, modelando uma bacia da maneira mais simples possível. Em um processo contínuo de revisão, novos mecanismos, reservatórios e parâmetros são adicionados na busca de um desempenho suficientemente adequado.

6 REDES NEURAIAS LSTM TREINADAS UTILIZANDO ASSINATURAS HIDROLÓGICAS

“All models are wrong...”

(BOX, 1976)

6.1 Introdução

Quase como um contraponto aos modelos hidrológicos baseados em processos, há os modelos orientados a dados. Enquanto os modelos baseados em processos decorrem da conceitualização dos fenômenos físicos constituintes do ciclo hidrológico em diferentes escalas espaçotemporais, os modelos orientados a dados são unicamente baseados em observações e não requerem informações sobre as propriedades físicas dos sistemas em análise. Esses métodos reproduzem as relações empíricas entre entradas e saídas a partir de dados históricos, estabelecendo uma representação global do sistema hidrológico (BEVEN, 2012; DUAN; ULLRICH; SHU, 2020).

O primeiro registro de um modelo hidrológico orientado à dados (MULVANEY, 1851) compreende uma relação linear simples entre a vazão e a precipitação (HUNT et al., 2022). De fato, os modelos baseados em processos dominaram a literatura hidrológica desse então, até que Young e Beven (1994) propuseram uma modelagem mecanicista baseada em dados (DBM), sugerindo que, em vez de impor estruturas ao modelo desde o início, deve-se permitir que os dados sugiram uma estrutura apropriada, e, em seguida, verificar se há uma interpretação mecanicista da estrutura do modelo aprendido (YOUNG, 2003).

As abordagens orientadas a dados só voltaram a ter relevância recente na comunidade científica com os algoritmos de aprendizado por máquina (ML, do inglês *Machine Learning*), sobretudo, devido à disponibilidade de informações, aumento do poder computacional e desempenho notável (LEY; BORMANN; CASPER, 2023; NEARING et al., 2021). Numerosas abordagens de diferentes modelos de ML tem sido aplicadas em hidrologia, usando, por exemplo, Redes Neurais Artificiais (RNAs) (NOORI; KALIN, 2016; PENG et al., 2017), regressores baseados em máquinas de vetor suporte (SVM, do inglês *Support Vector Machine*) (HUANG et al., 2014; LIU;

ZHOU; ZHANG, 2015), *Random Forest* (TALUKDAR et al., 2023) e programação genética (CHADALAWADA; HERATH; BABOVIC, 2020; HERATH; CHADALAWADA; BABOVIC, 2021). O rápido progresso da ML é um dos avanços mais importantes na hidrologia nos últimos anos (SCHMIDT et al., 2020).

A introdução da aprendizagem profunda (DL, do inglês *Deep Learning*) (LECUN; BENGIO; HINTON, 2015) na hidrologia a partir de 2016 (LALOY et al., 2017; SHEN, 2018; TAO et al., 2016) , especialmente das redes com memória de longo e curto prazo (LSTM, do inglês *Long Short-Term Memory*) como uma ferramenta de modelagem dinâmica para umidade do solo e vazão (FANG; PAN; SHEN, 2019; KRATZERT et al., 2018), tornou oportuna a comparação desses modelos aos baseados em processos em grandes bases públicas de bacias em relação as métricas de desempenho.

Para essas grandes bases de bacias as redes LSTM foram testadas com êxito, por exemplo, nos Estados Unidos (KRATZERT et al., 2018) , Grã-Bretanha (LEES et al., 2021), Chile (FUENTE; ANDRÉS, 2021), Dinamarca (KOCH; SCHNEIDER, 2022) e no cerrado brasileiro (ALTHOFF; RODRIGUES; SILVA, 2021). Em geral, o objetivo central desses estudos é comparar as redes em relação a modelos conceituais, se limitando a no máximo quatro estruturas de modelo conceituais.

Estudos recentes mostram que, ao se fornecer como entrada da rede uma série de atributos estáticos, tem-se melhorado seu desempenho, não só em modelagens hidrológicas (KRATZERT et al., 2018), mas também em outras áreas do conhecimento (LEONTJEVA; KUZOVKIN, 2016; MIEBS et al., 2020). Com essa camada estática, redes LSTM podem ter seu conhecimento transferido para outras bacias até mesmo geograficamente distantes (MA et al., 2020; ORUCHE et al., 2021). Visto o êxito do uso de assinaturas hidrológicas em modelos de base física, presume-se que utilizar as assinaturas hidrológicas como camada estática seja uma abordagem adequada, ainda que não avaliada.

Em geral, as pesquisas recentes em hidrologia geralmente se concentram na modelagem de múltiplas bacias (SEIBERT; BERGSTRÖM, 2022). No entanto, os processos de uma bacia são influenciados por fatores específicos, e os modelos tendem a funcionar de forma desigual para zonas de baixa, média e alta produção de

escoamentos (LOGANATHAN; MAHINDRAKAR, 2020). Por exemplo, abordagens específicas podem ser necessárias em bacias onde os efeitos do fluxo de águas subterrâneas são dominantes (WHEATER; PEACH; BINLEY, 2007).

Em relação às redes LSTM, a estratégia utilizada na literatura (KRATZERT et al., 2019a; LEES et al., 2021) foi a incorporação de uma camada estática, combinando os atributos da bacia, oferecendo um nível de interpretabilidade sobre como o modelo aprende a diferenciar aspectos de comportamentos complexos específicos da bacia hidrográfica. O conceito é que essa camada vincule as características da bacia hidrográfica à dinâmica de locais específicos e que o modelo aprenda a partir das múltiplas combinações que compõem as bacias analisadas. Essa adaptação da rede LSTM permite que os atributos estáticos fornecidos controlem quais células de estado serão usadas para uma determinada bacia. Em princípio, a abordagem permite explicitamente o compartilhamento de partes das redes para bacias com comportamento semelhante.

Dessa forma, o que se conjectura é que o modelo consegue sozinho identificar bacias similares e ativar partes específicas da rede, e isso possibilita a modelagem conjunta de diversas bacias (RAZAVI; COULIBALY, 2013). A hipótese deste trabalho é que, assim como acontece com os modelos conceituais (BECK et al., 2016; SAMANIEGO; KUMAR; ATTINGER, 2010), a estrutura da rede LSTM esteja relacionada com as características das bacias, i.e., seus atributos físicos e, principalmente, suas assinaturas hidrológicas. Há uma expectativa que redes treinadas em agrupamentos de bacias semelhantes apresentem resultados superiores àqueles em que o treinamento ocorreu sobre toda a base de dados, assumindo que o processo hidrológico dominante influencie a estrutura da rede. Ainda que isso tenha sido pouco abordado na literatura, outros estudos, como Rigotto (2022), mostram que essa pode ser uma estratégia interessante.

Nos estudos em que uma camada estática de atributos físicos foi utilizada, notou-se que os autores optaram por não restringir o número de atributos fornecidos à rede. Apesar de no âmbito de uma modelagem baseada em dados isso parecer justificável, a aplicabilidade em cenários de escassez de informações, comum em países em desenvolvimento como o Brasil, sugerem que essa estratégia deva ser revista. No

presente estudo, motivado pelos resultados anteriores, se optou por substituir os atributos físicos por assinaturas hidrológicas, sob o propósito de facilitar a transferência do conhecimento a regiões não monitoradas.

Nesse contexto, o objetivo primordial deste capítulo é avaliar o treinamento das redes LSTM a partir das assinaturas hidrológicas, assim como comparar os resultados obtidos desses modelos àqueles de base física, valendo-se da melhor configuração encontrada na estrutura de modelos conceituais flexíveis do FUSE (CLARK et al., 2008), em uma extensão aos estudos consultados. As perguntas que se pretende explorar são:

- Qual desempenho das redes LSTM em todos os biomas nacionais? E em relação ao *framework* multimodelos do FUSE?
- Utilizar assinaturas hidrológicas em detrimento de atributos físicos da bacia como camada estática resulta em algum ganho para a rede LSTM? e
- Treinar redes em grupos de bacias com comportamento hidrológicos similares melhora o seu desempenho sobretudo em bacias onde os resultados têm sido sistematicamente ruins?

6.2 Redes neurais e LSTM

As aplicações das Redes Neurais (RNAs) são baseadas em sua capacidade de imitar a estrutura mental e neural humana para construir uma boa aproximação das relações funcionais entre os valores passados e futuros de uma série temporal. A história das RNAs começou no início da década de 1940, quando McCulloch e Pitts (1943) desenvolveram a primeira representação computacional de um neurônio. Posteriormente, White e Rosenblatt (1963) propuseram a ideia dos perceptrons, em que as entradas alimentam diretamente as saídas (*feed-forward*) por meio de pesos em uma única camada de neurônios, sem a formação de um ciclo. Antagonicamente, a rede neural recorrente (RNN; do inglês *Recurrent Neural Network*), proposta na década de 80, permite conexões entre neurônios dentro das camadas ocultas com um atraso de tempo, sendo capazes de transportar informações do passado (RUMELHART; HINTON; WILLIAMS, 1986).

Os neurônios são organizados em camadas, que aprendem em conjunto a partir de dados de entrada. Em termos gerais, um neurônio recebe várias entradas, multiplica cada entrada por um peso específico (que é aprendido durante o treinamento), soma esses produtos, adiciona um viés e passa a soma resultante por uma função de ativação para produzir sua saída. Essa saída é então passada como entrada para os neurônios na próxima camada. Este processo é repetido para cada camada da rede (GOODFELLOW; BENGIO; COURVILLE, 2016; SCHMIDHUBER, 2015).

A utilização de redes neurais na previsão de vazões se inicia na década de 90 (DANIELL, 1991). Posteriormente, Hsu et al. (1997) compararam uma RNN com uma RNA tradicional. Eles chegaram a um desempenho similar, mas a RNA é sensível ao número de passos de tempos anteriores fornecidos como entrada, enquanto a RNN elimina a necessidade de se buscar o número de entradas defasadas à rede.

Os parâmetros de uma rede neural são os valores que a rede aprende durante o processo de treinamento. Eles incluem pesos e vieses associados a cada neurônio na rede. Durante o treinamento, esses parâmetros são ajustados por meio de algoritmos de otimização (como o Gradiente Descendente; RUDER, 2017) para minimizar a função de perda, que mede o desempenho do modelo (LECUN; BENGIO; HINTON, 2015).

Diferente dos modelos hidrológicos baseados em processos, o gradiente da função de perda em relação a qualquer parâmetro de rede sempre pode ser calculado explicitamente. Essa propriedade é usada na etapa de retropropagação, na qual os parâmetros da rede são adaptados para minimizar a perda geral (GOODFELLOW; BENGIO; COURVILLE, 2016). Entretanto, as redes RNN tendem a desaparecer com o gradiente em relações de longo prazo, acima de 10 passos de tempo (BENGIO; SIMARD; FRASCONI, 1994). As redes com memória de longo e curto prazo (LSTM), propostas por Hochreiter e Schmidhuber (1997), são uma variação das RNNs que são projetadas para evitar o problema do desaparecimento do gradiente, permitindo que elas aprendam dependências de longo prazo. Dividindo a arquitetura em três etapas (ou *gates* na terminologia em inglês) de entrada, saída e esquecimento, a rede LSTM pode receber muitos passos de tempos anteriores e descartar partes deles.

Recentemente, as redes LSTM tem ganhado relevância notória nas ciências naturais. Elas foram utilizadas com sucesso para prever a qualidade do ar (ZHANG et al., 2021), da água (LIANG et al., 2020) e inundações (FANG et al., 2021) na China, para estimar a altura e energia de ondas (SAREEN et al., 2023), a vulnerabilidade a terremotos (JENA et al., 2021) e a evapotranspiração potencial (YAN et al., 2023).

Redes LSTM se mostraram adequadas para gerar previsões dos comportamentos de sistemas hidrológicos dinâmicos complexos (KRATZERT et al., 2019a; QIU et al., 2021; ZHANG et al., 2018a). Isso se deve principalmente à sua capacidade de representar propriedades de Markov por meio das suas células de memória e de aprender as formas funcionais dos *gates* que determinam que tipos de informações são retidas (ou esquecidas) a cada passo de tempo (LEES et al., 2021).

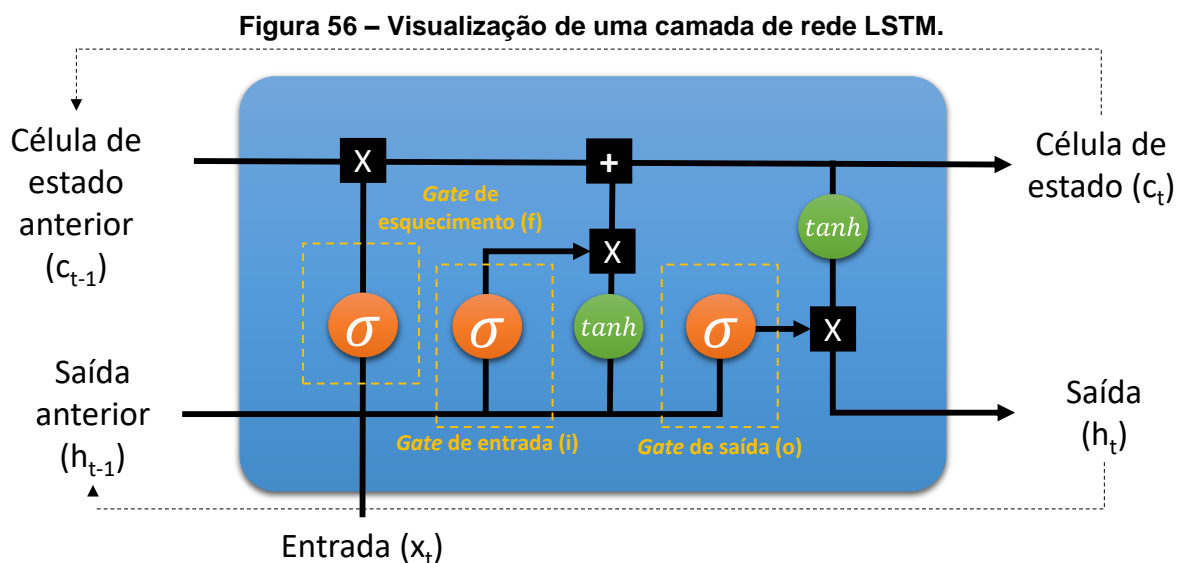
Diferente dos parâmetros, os hiperparâmetros são parâmetros que definem a estrutura da rede (por exemplo, o número de camadas ocultas, o número de neurônios em cada camada) e como a rede é treinada (por exemplo, a taxa de aprendizado, o tamanho do lote, e o número de épocas). Ao contrário dos parâmetros, os hiperparâmetros não são aprendidos durante o treinamento do modelo e devem ser definidos antes do treinamento. O ajuste dos hiperparâmetros é geralmente um processo experimental de tentativa e erro (BENGIO, 2012; BERGSTRA et al., 2011).

O tamanho do lote (*batch size*) refere-se ao número de amostras de treinamento que são usadas para atualizar os pesos da rede de uma só vez durante o treinamento. O tamanho do lote pode ter um grande impacto na convergência do modelo e no desempenho final (KESKAR et al., 2017). A taxa de aprendizado (*learning rate*) controla o tamanho dos ajustes feitos aos pesos da rede durante o treinamento. Taxas de aprendizado muito altas podem fazer com que o treinamento da rede seja instável, enquanto taxas de aprendizado muito baixas podem fazer com que o treinamento da rede seja muito lento (WILSON et al., 2018). O tamanho da camada oculta (*hidden size*) se refere ao número de neurônios na camada homônima de uma rede neural. Um número maior de neurônios pode aumentar a capacidade da rede de capturar padrões complexos nos dados, mas também pode aumentar o risco de sobreajuste (MONTAVON; ORR; MÜLLER, 2012). Por fim, uma época é um ciclo completo de

treinamento no qual o algoritmo de aprendizado viu todos os dados de treinamento uma vez (PRECHELT, 1998).

6.3 Metodologia

O cerne do estudo é avaliar redes LSTM nas mesmas 351 bacias do Capítulo 5, utilizando a base de dados do CABra (ALMAGRO et al., 2021). O conceito central das redes LSTM são as células de estado e seus diversos *gates*. As células de estado transferem informações relevantes ao longo da cadeia de sequências. Elas funcionam como a memória da rede. Assim, mesmo as informações dos primeiros passos de tempo podem chegar aos passos de tempo posteriores, reduzindo os efeitos da memória de curto prazo. A cada passo de tempo, informações são adicionadas ou removidas nas células de estado por meio de *gates*. Os *gates* são redes neurais diferentes, i.e, que aprendem quais informações são relevantes e devem ser mantidas ou esquecidas durante o treinamento. Uma visualização de uma camada da rede LSTM é apresentada na Figura 56.



Cada *gate* recebe as informações da saída anterior (h_{t-1}) e as entradas atuais (x_t) que passam por uma ativação sigmoide (σ), transformando os dados em valores entre 0 e 1, determinando quais valores serão mantidos ou esquecidos. O resultado do *gate* de entrada (i_t ; Equação (15)), é multiplicado ponto a ponto (símbolo X) por uma ativação da tangente hiperbólica (\tanh), criando valores candidatos que podem entrar na célula de estado (\tilde{c}_t ; Equação (16)). A sigmoide do *gate* de esquecimento (f_t ; Equação (17))

é multiplicada ponto a ponto pela célula de estado anterior (c_{t-1}) e esse resultado é somado a \tilde{c}_t , resultando na nova célula de estado (c_t ; Equação (18)). Por fim, o *gate* de saída (o_t ; Equação (19)) ajuda a decidir quais informações da célula de estado atual são conduzidas à saída (h_t ; Equação (20)) dessa camada da rede.

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i) \quad (15)$$

$$\tilde{c}_t = \tanh(W_c * [h_{t-1}, x_t] + b_c) \quad (16)$$

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f) \quad (17)$$

$$c_t = f_t \times c_{t-1} + i_t \times \tilde{c}_t \quad (18)$$

$$o_t = \sigma(W_o * [h_{t-1}, x_t] + b_o) \quad (19)$$

$$h_t = o_t \times \tanh(c_t) \quad (20)$$

A rede de cada *gate* tem seu peso (W) e viés (b). A saída da camada LSTM passa por uma camada densa, conectando todos os valores, em uma função linear, produzindo a variável de interesse, nesse caso a vazão. Forneceu-se como entradas as séries temporais de precipitação, temperatura máxima, temperatura mínima, radiação, e evapotranspiração potencial calculada pelo método de Penman-Monteith (presentes na base CABra), de maneira similar ao realizado por Kratzert et al. (2018).

A literatura hidrológica já mostrou que as redes LSTM se beneficiam ao se fornecer como entrada da rede atributos físicos estáticos junto com as séries temporais (ALTHOFF; RODRIGUES, 2021; KRATZERT et al., 2018; LEES et al., 2021). Nesse sentido, avaliou-se também redes LSTM fornecendo como camada estática assinaturas hidrológicas. Dessa forma, utilizando 351 bacias, treinou-se redes com as seguintes opções na camada estática:

- Sem camada estática, fornecendo somente as 5 séries temporais;
- Camada estática com 27 atributos físicos da bacia;
- Camada estática com 16 assinaturas hidrológicas;

A lista completa de atributos físicos utilizados é apresentada na Tabela 13. Em relação à escolha das assinaturas hidrológicas, tomou-se como base as principais dentre as

mais bem preditas espacialmente de acordo com o Capítulo **Erro! Fonte de referência não encontrada.** Todas as assinaturas fornecidas como camada estática são apresentadas na Tabela 13.

Tabela 13 - Entradas da rede LSTM

Séries Temporais Diárias	Atributos Físicos	Assinaturas hidrológicas
Temperatura máxima do ar	Precipitação média	Vazão mediana
Temperatura mínima do ar	Evapotranspiração potencial média	Q10
Precipitação	Índice de aridez	Q25
Radiação	Sazonalidade da precipitação	Q90
Evapotranspiração potencial	Elevação média de captação.	Média mensal fevereiro
	Inclinação média da captação.	Média mensal junho
	Área de captação.	Média mensal outubro
	Fração da bacia coberta por floresta.	Coefficiente de escoamento superficial
	Fração da bacia coberta por plantações	Taxa de subida
	Fração da bacia coberta por grama	Elasticidade da precipitação mensal
	Fração da bacia coberta por musgo	Elasticidade da evapotranspiração diária
	Fração da bacia coberta por arbusto	Índice do escoamento de base
	Fração da bacia coberta por área urbana	Mínimo anual na média de 3 dias
	Fração da bacia coberta por água	Máximo anual na média de 90 dias
	Fração de carbono orgânico no solo.	Dia do ano que se atinge metade da vazão anual
	NDVI médio do outono	Q90/Q10
	NDVI médio do verão	
	NDVI médio da primavera	
	NDVI médio do inverno	
	Densidade aparente do solo	
	Profundidade do leito rochoso	
	Fração de areia no solo.	
	Fração de silte no solo.	
	Fração de argila no solo.	
Porosidade do subsolo.		
Permeabilidade do subsolo		
Condutividade hidráulica do subsolo		

Nesses experimentos, variou-se o número de passos anteriores fornecidos à rede entre 120 e 330 dias, em incrementos de 30 dias, fixando-se o número de neurônios em 128. Variou-se o número de neurônios entre 128 e 512, em incrementos de 128, fixando o número de entradas anteriores em 240. Treinou-se as redes nos primeiros 15 anos hidrológicos, enquanto os 15 anos seguintes foram utilizadas para validá-las, de maneira similar a Kratzert et al. (2018).

Os resultados de cada experimento, com sua arquitetura otimizada, foram comparados com o modelo conceitual calibrado do FUSE, obtido no Capítulo 5 em cada uma das 351 bacias. Nas bacias com mais de um modelo do FUSE possível utilizou-se aquele de maior NSE. Nas bacias em que não se obteve uma estrutura com as métricas mínimas, buscou-se o modelo com melhor NSE na fase de calibração.

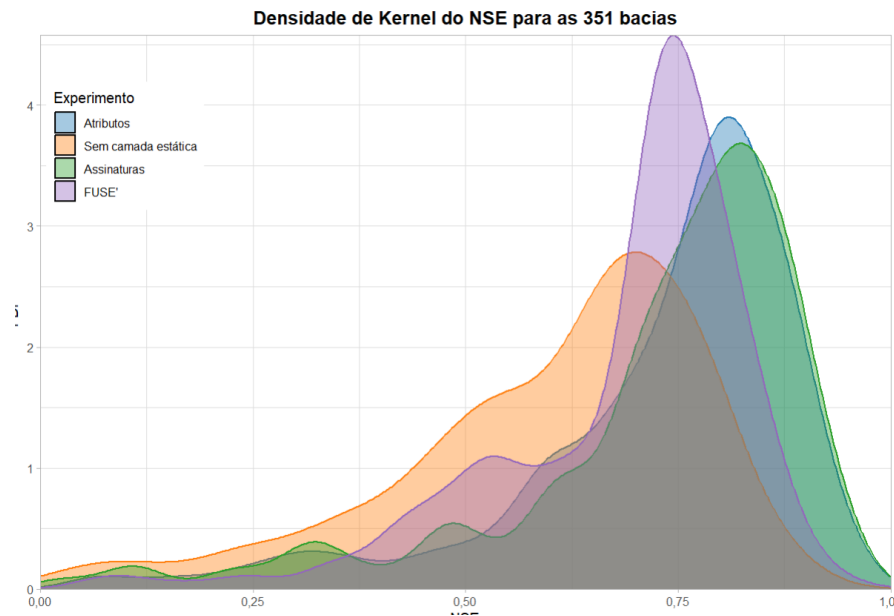
As redes supracitadas, tiveram seu treinamento nas 351 bacias. O que se conjectura neste estudo é que redes treinadas em grupos menores, mas com comportamento similar poderia levar a resultados superiores e pavimentar um caminho para a interpretação física desses modelos. Nesse sentido, avaliou-se em conjunto, o treinamento das redes individualmente para cada um dos 5 agrupamentos obtidos nos Capítulos 2 e 3, doravante designados como agrupamentos por assinaturas e mecanismos, respectivamente. Nesses casos, utilizou-se também a camada estática com as assinaturas hidrológicas. A hipótese nesse caso é que a seleção *a priori* do comportamento das bacias pode auxiliar no treinamento das redes e sustentar futuras interpretações físicas desse modelo orientado à dados. Para validar essa hipótese, além de comparar os resultados com uma única rede treinada com todas as bacias, criou-se três conjuntos de agrupamentos aleatórios com também 5 *clusters*.

Cada treinamento da rede neural foi uma média de um *ensemble* de 3 treinamentos em 10 épocas. Testes preliminares revelaram um sobreajuste na validação com um número maior de épocas. O tamanho do lote foi fixado em 1.000, a taxa de esquecimento em 0,4 e taxa de aprendizado em 0,001. Em testes preliminares não se identificou valores ótimos para esses parâmetros.

6.4 Resultados e discussão

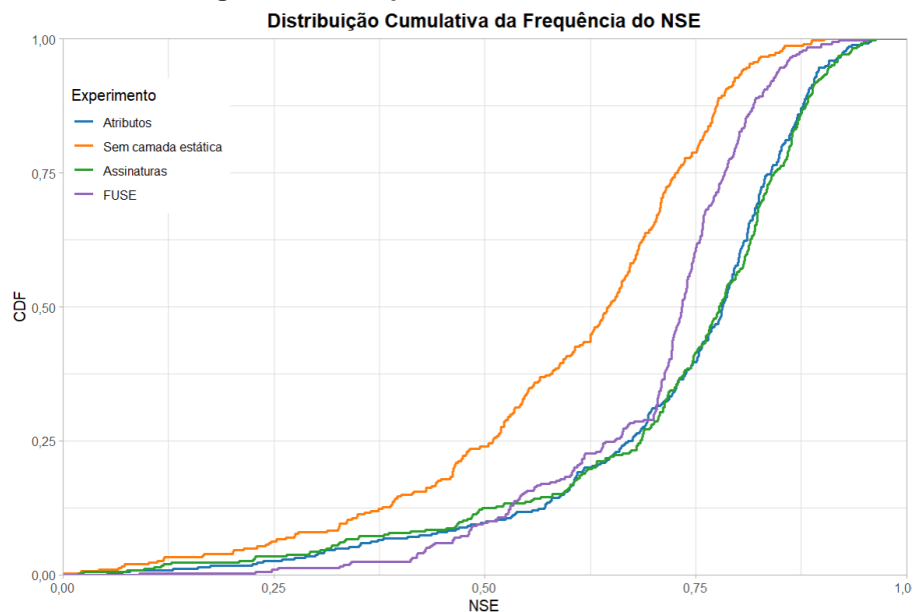
A Figura 57 apresenta as distribuições do NSE durante o período de validação para o FUSE, e as redes LSTM utilizando atributos físicos ou assinaturas hidrológicas como camada estática, e sem camada estática. Não utilizar uma camada estática nas redes LSTM degrada significativamente os resultados, com a média em torno do zero. Ao adicionar a cama estática os resultados melhoram significativamente. O resultado do FUSE é um pouco inferior ao das redes com camada estática (média de 0,59 do NSE). A Tabela 15 sumaria essas métricas.

Figura 57 – Função de densidade de probabilidade não paramétrica do NSE



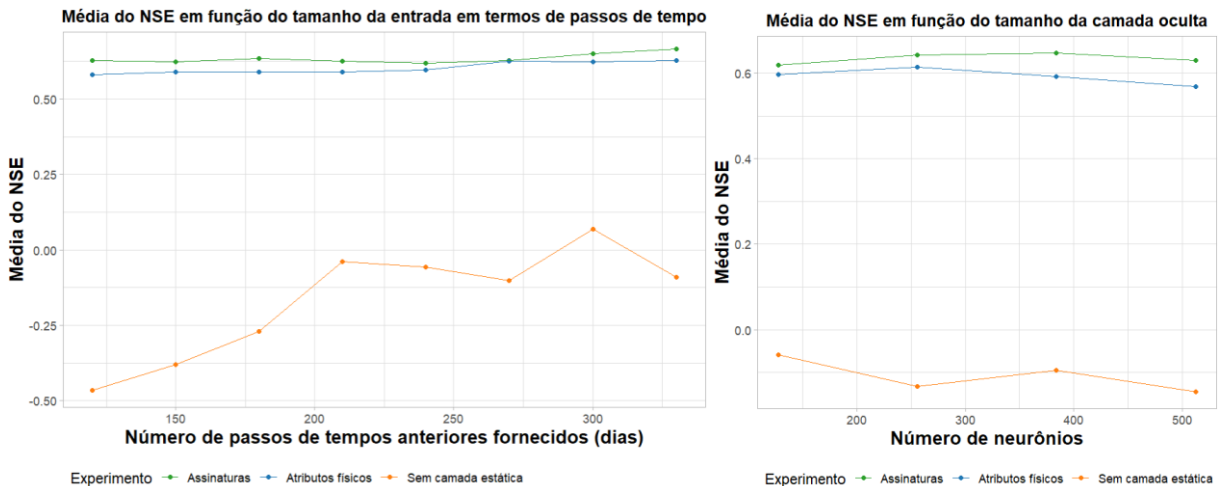
Por sua vez, a Figura 58 apresenta a frequência cumulativa do NSE. Utilizar assinaturas hidrológicas, em detrimento dos atributos físicos das bacias, melhora ligeiramente os resultados do NSE. Embora essa diferença seja na realidade pouco significativa, há de se salientar que a rede, no caso das assinaturas, recebe menos 11 entradas, o que já é um ganho significativo em termos de variáveis a serem estimadas para a utilização em bacias fora do CABra. Além disso, sua variação espacial mais suave se comparada aos atributos físicos permite um quadro otimizado para estimativa em bacias sem monitoramento.

Figura 58 – Frequência cumulativa do NSE



A Figura 59 apresenta o gráfico da média dos NSE para cada rede LSTM variando-se o número de passos de tempos anteriores fornecidos (tendo essas redes 240 neurônios), e o número de neurônios (sendo fixado a quantidade de passos de tempos anteriores em 240). Percebe-se que as redes com camada estática são pouco sensíveis a variação desses parâmetros, sendo a rede sem a camada mais suscetível ao número de passos de tempos anteriores. De uma maneira geral, fornecer mais passos de tempo pode ser benéfico, mas, no caso das camadas estáticas, usar o menor passo de tempo e o menor tamanho da camada oculta produz praticamente os mesmos resultados, e uma rede muito mais leve, e, portanto, de menor custo computacional.

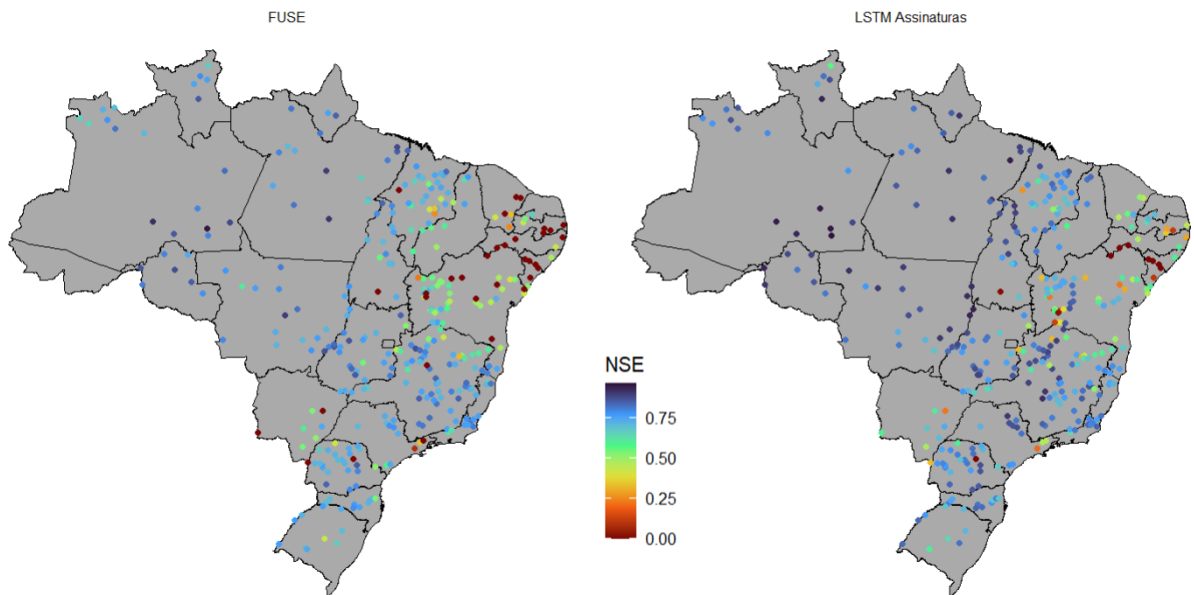
Figura 59 – Sensibilidade dos resultados das redes LSTM em função do número de passos de tempo anteriores fornecidos e o número de neurônios



A Figura 60 e a Figura 61 (excluindo as 4 bacias que tem NSE negativo) comparam o NSE obtidos pelo melhor modelo do FUSE ao da rede LSTM utilizando assinaturas hidrológicas como camada estática, em cada bacia. Em resumo, a rede LSTM consegue melhores ajustes especialmente quando resultados do FUSE são inferiores a 0,6. Geograficamente, essas bacias estão localizadas no semiárido brasileiro e na transição entre o Cerrado e Caatinga. Entretanto, ainda há bacias que não são bem representadas por nenhum dos modelos.

Figura 60 – NSE de cada bacia no FUSE e na rede LSTM com assinaturas como camada estática

Distribuição espacial do NSE



Por outro lado, a Figura 62 apresenta as diferenças relativas em relação a Q90 simulada e observada pela melhor estrutura do FUSE e a rede LSTM utilizando como camada estática as assinaturas hidrológicas. Naturalmente, era esperado um melhor desempenho do FUSE nesse quesito, uma vez que a calibração multiobjetivo garantia o atendimento dessa métrica. Mas havia uma expectativa que a inclusão desse percentil como camada estática na rede LSTM conduziria a resultados superiores, contudo isso não se concretizou.

Figura 61 – Comparação entre o NSE obtido no melhor modelo do FUSE e a rede LSTM com assinaturas hidrológicas como camada estática

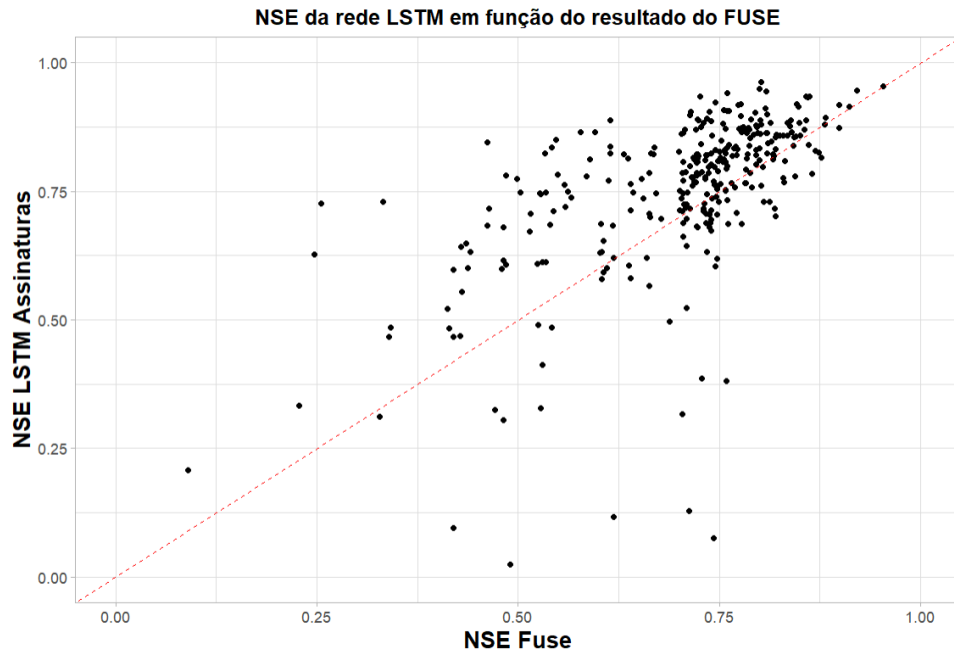
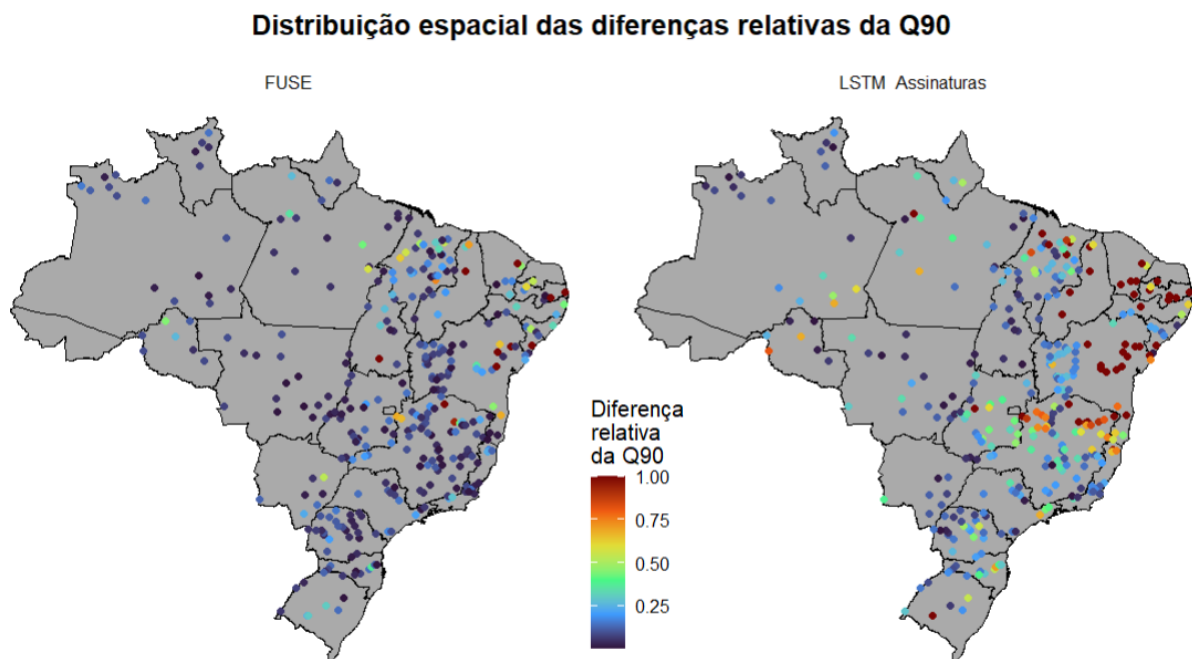


Figura 62 – Distribuição espacial da diferença relativa entre as Q90 simuladas e observadas para as redes LSTM e a melhor estrutura do FUSE



Na sequência, se avaliou o benefício em se treinar grupos de bacias similares em redes distintas, em detrimento a uma única rede com todas as bacias. Esses resultados estão sumariados na Tabela 14. Embora o desempenho das redes LSTM treinadas em agrupamentos hidrológicamente consistentes seja ligeiramente superior, agrupar as bacias aleatoriamente em 5 *clusters* resulta praticamente nos mesmos

resultados. Assim, no que tange ao desempenho do modelo, não se identificou qualquer benefício ao treinamento decorrente dessa segmentação. Possivelmente as informações contidas nos *clusters* aleatórios ainda é suficiente para que a rede aprenda a distinguir regimes hidrológicos distintos.

Tabela 14 – Desempenho das redes treinadas em agrupamentos

Tipo de agrupamento	NSE			
	Mínimo	Média	Mediana	Máximo
Mecanismos do FUSE	-2,73	0,66	0,78	0,96
Assinaturas hidrológicas	-9,60	0,65	0,78	0,97
Aleatório 1	-6,89	0,64	0,76	0,96
Aleatório 2	-8,72	0,63	0,76	0,96
Aleatório 3	-7,73	0,64	0,76	0,96

A Tabela 15 resume os desempenhos das arquiteturas de rede otimizadas para cada um dos cenários anteriormente descritos, assim como aqueles obtidos pelo melhor modelo conceitual do FUSE em cada bacia. Para as redes treinadas com todas as 351 bacias, estender o número de passos de tempos anteriores fornecidos melhora ligeiramente o desempenho, enquanto treinar redes em grupos de bacias, e, portanto, com menor volume de dados, tendem a conduzir a um tamanho maior da camada oculta.

Tabela 15 – Variação do NSE para as arquiteturas otimizadas das redes LSTM e do melhor modelo conceitual em cada bacia no FUSE

Experimento	Cluster	Parâmetros		NSE			
		Passos de tempo (dias)	Neurônios	Mínimo	Média	Mediana	Máximo
Sem camada estática	-	300	128	-34,87	-0,09	0,61	0,95
Atributos físicos	-	330	128	-14,99	0,63	0,78	0,96
Assinaturas hidrológicas	-	330	128	-10,84	0,67	0,78	0,96
Agrupamento de assinaturas hidrológicas	1	240	512				
	2	240	384				
	3	240	384	-8,55	0,67	0,78	0,97
	4	210	128				
	5	240	256				
Agrupamento de mecanismos	0	240	512				
	1	240	512				
	2	240	384	-1,35	0,69	0,78	0,96
	3	330	128				
FUSE	-	-	-	-4,22	0,59	0,72	0,95

6.5 Conclusão

Em termos da métrica NSE, o desempenho das redes LSTM foi similar à de outros estudos em outras regiões (ver FUENTE; ANDRÉS, 2021; KOCH; SCHNEIDER, 2022; KRATZERT et al., 2018; LEES et al., 2021), reforçando a dificuldade desse modelo nas regiões mais áridas. O desempenho do *framework* de estruturas flexíveis é bastante similar ao das redes LSTM quando se avalia o melhor modelo conceitual do FUSE para cada bacia. Por outro lado, as estruturas mais adequadas do FUSE para cada bacia conseguem atingir diferenças relativas das assinaturas hidrológicas melhores, em razão de terem sido calibrados em função delas. Esperava-se que se fornecer o indicador hidrológico na camada estática também levasse a uma consistência com as assinaturas.

Este estudo também reforça o aumento de desempenho ao se utilizar uma camada estática no treinamento das redes, e mostra que os atributos físicos podem ser substituídos pelas assinaturas hidrológicas, com uma ligeira melhora no desempenho, mas com um número menor de entradas.

Como já vem sendo sistematicamente demonstrado, os métodos de aprendizado profundo, sobretudo as redes LSTM, frequentemente superam os modelos hidrológicos conceituais na modelagem da chuva e vazão. Tentativas de investigar os aspectos internos de tais redes estão sendo feitas, mas a rastreabilidade dos estados e processos do modelo e suas inter-relações com a entrada e saída ainda não são compreendidas. A interpretabilidade direta dos processos é um recurso valioso dos modelos conceituais, no sentido de obter entendimento do sistema além da previsibilidade. Algum esforço para essa interpretação *a posteriori* tem sido feito (ver HÖGE et al., 2022; LEES et al., 2022). A similaridade de assinaturas hidrológicas ou de mecanismos indicaria uma consistência da interpretação física, no entanto, não se traduz em benefícios no treinamento das redes. A expectativa era que ao isolar processos dominantes a ativação da camada oculta revelasse interpretações físicas sobre esses processos, o que poderia levar a uma aprimoração dos modelos com base física.

7 CONCLUSÃO GERAL E RECOMENDAÇÕES

O Capítulo 3, explora a calibração multiobjetivo, utilizando assinaturas hidrológicas, dos parâmetros de modelos hidrológicos. Emboque seja um estudo incipiente e que carece de mais análises, inclusive na fase de validação do modelo, ele sugere um caminho interessante para a construção de uma métrica única, considerando o NSE e a pior diferença relativa das assinaturas hidrológicas, o que pode ser bem mais simples do que utilizar vários objetivos concomitantemente em uma frente de Pareto, e, aparentemente, mais efetivo. Além disso, reitera as conclusões de outros pesquisadores ao indicar que a calibração apoiada nas assinaturas hidrológicas é adequada e estende a interpretabilidade física dos modelos conceituais.

Na sequência o Capítulo 4, identifica as assinaturas mais facilmente estimadas utilizando as características das bacias. Além de uma extensa lista de assinaturas, o estudo avalia diversos algoritmos de regressão. Embora a regressão linear múltipla tenha um desempenho interessante, o estudo reforça a superioridade das técnicas baseadas no *Bagging* e, especialmente, em *Gradient Boosting*. Utilizando as assinaturas mais bem estimadas, as bacias brasileiras foram agrupadas por similaridade destas assinaturas. O estudo sugere cinco grupos de bacias, embora um deles possa ser futuramente desmembrado em um melhor monitoramento da região norte do país. Estes agrupamentos tiveram uma boa consistência espacial, sendo bem marcados pelo regime pluviométrico, zonas climáticas, biomas e os principais aquíferos.

O Capítulo 5, avalia 624 combinações de mecanismos físicos, calibrando os parâmetros destas estruturas para 351 bacias hidrográficas localizadas em diferentes porções do território brasileiro. Estabelecer relações explícitas entre as estruturas ótimas (ou adequadas) a representação dos fenômenos hidrológicos e as características das bacias (físicas e hidrológicas) é ainda desafiador. A heterogeneidade espacial e temporal das bacias além de diversas bacias poderem ser representadas por uma gama de estruturas são agravantes nesta análise. Independentemente, uma série de regras de associação foram descritas, sob o propósito de fundamentar a seleção de estruturas apropriadas a representação dos fenômenos.

Por fim, o Capítulo 6 demonstrou-se que o emprego de assinaturas hidrológicas como camada estática durante o treinamento das redes LSTM permite o alcance de resultados ligeiramente melhores aos alcançados quando se considera atributos físicos. O menor número de assinaturas fornecidas, quando comparada aos atributos, indica que esses índices condensam e transmitem a informação derivada do monitoramento hidrométrico de forma otimizada. O conjunto de assinaturas fornecidas, no entanto, tem potencial para alterar esses resultados. Recomenda-se que sua seleção se apoie nos resultados descritos no Capítulo 4. Por fim, a comparação das abordagens revelou que os modelos baseados em dados, apesar de apresentarem globalmente um NSE superior ao apresentado pelos modelos baseados em processos, a exemplo de outros estudos, representou as assinaturas de forma menos acurada. Esse fato, reforça a importância do aperfeiçoamento de abordagens híbridas e relativiza as estratégias metodológicas e conclusões de estudos recentes.

Segundo Seibert e McDonnell (2002) "a hidrologia de bacias encontra-se numa encruzilhada". Os autores se referiam complexidade crescente das teorias que se propunham a descrever as componentes do escoamento e a inobservância de avanços que permitissem, além de aperfeiçoar as teorias atuais à época, extrair toda informação contida nos dados empregados para calibração e desenvolvimento de modelos hidrológicos conceituais. Após 20 anos, o dilema culmina em publicações que exaltam efusivamente o desempenho de modelos baseados em dados frente aos baseados em processos físicos (Kratzert et al. (2018)) e discutem o papel do aprendizado de máquina na hidrologia, enunciando que "os hidrólogos não têm sequer uma compreensão básica da razão pela qual os seus modelos falham" (Nearing et al. (2021)).

Não obstante, a recomendação (óbvia) de que a melhor alternativa para modelar bacias não monitoradas seja iniciar o seu monitoramento, Blöschl (2006) pretere o fato de que, não raro, modelos hidrológicos conceituais são incapazes de representar adequadamente a resposta hidrológica, ainda que diversos anos de registros hidrométricos estejam disponíveis. Ademais, soma-se a isto observar na literatura experimentos em que modelos baseados em dados sejam aplicados a bacias não utilizadas na fase de treinamento e seu desempenho supere os de modelos conceituais calibrados exclusivamente nessas bacias (KRATZERT et al., 2019b). Por

isso, Nearing et al. (2021) afirmaram que os modelos conceituais já chegam mortos. Mas se todos os modelos são errados, como dito por George Box (“*All models are wrong; some are useful*”), cabe então aos hidrólogos identificar aqueles que são úteis, e nem sempre serão aqueles orientados a dados.

Certamente, há mais informações nos dados observados do que os hidrólogos conseguem usualmente explorar. Contudo, desconsiderar ou até menosprezar a base conceitual já estabelecida é um erro. Ademais, até o presente momento, salve melhor entendimento, as redes LSTM não podem ser treinadas em múltiplos objetivos. Como demonstrado, índices de qualidade de ajuste baseados exclusivamente na similaridade entre vazões observadas e modeladas não garante a consistência hidrológica do modelo. No presente trabalho, essa consistência foi perseguida para os modelos conceituais durante a calibração, por meio da busca de parâmetros que conduzissem a diferenças percentuais entre assinaturas observadas e modeladas limitadas explicitamente. Com efeito, as redes conduziram reiteradamente a assinaturas com diferenças percentuais superiores àquelas obtidas via modelos conceituais.

Além da comparação, há outros aspectos a serem considerados. O emprego de modelos baseados em processos em bacias fundamenta-se no conceito de similaridade hidrológica. Assume-se que as extrapolações temporais, representadas por simulações de eventos hidrológicos que excedam os limites da série empregada na calibração, e espacial, no caso de bacias não monitoradas, por meio da transferência de parâmetros, estruturas conceituais e/ou assinaturas hidrológicas, seja plausível pois a resposta hidrológica às forçantes climáticas ocorre de forma “suficientemente similar” nessas bacias. No caso dos modelos baseados em dados, as extrapolações espaciais requerem a interpretabilidade do comportamento do algoritmo frente às características físicas das bacias avaliadas. Nos estudos consultados essa avaliação ocorre *a posteriori*, i.e., busca-se identificar as “razões físicas” que justifiquem a dinâmica do algoritmo sem o estabelecimento de qualquer hipótese *a priori* (ex.: a construção de um modelo de percepção dos processos físicos ou a própria estrutura conceitual assumida). Apesar da fragilidade da hipótese, retratada pelos resultados alcançados neste trabalho, uma série de estudos recentes se dedicou a essa busca (HÖGE et al., 2022; LEES et al., 2022). No que tange às

extrapolações temporais, identificou-se um único estudo que se propôs a avaliá-la (FRAME et al., 2022). Nele, os autores indicam a superioridade da abordagem baseada em dados, mas limitam-se à leitura de métricas de qualidade de ajuste sem estabelecer bases conceituais sólidas que permitam fundamentar diretrizes para sua aplicação e, assim, estender o conhecimento atual.

Não parece crível a hipótese de que cientistas da computação substituirão hidrólogos. De fato, entende-se que aproximação desses profissionais das geociências, como a hidrologia, representa uma oportunidade única para avanços em ambas as áreas, seja pelo notório requisito de interpretabilidade dos algoritmos à luz de conceitos físicos, seja pela incômoda necessidade de revisar e validar hipóteses há muito assumidas como verdadeiras e que, frequentemente, contradizem observações e resultados da modelagem..

É importante ainda destacar que, ao contrário dos modelos baseados em processos físicos, os modelos de aprendizado profundo, como as redes LSTM, funcionam melhor quando treinados simultaneamente para diversas bacias. Esses modelos aprendem as relações a partir da amostra fornecida e são capazes de traduzir essas relações individualmente para cada bacia. No entanto, recomenda-se, que em estudos futuros, a combinação de características das bacias (atributos físicos e/ou assinaturas hidrológicas) seja explorada. Infere-se que a interpretabilidade do algoritmo seja afetada por informações redundantes e incertas. Para além, recomenda-se em estudos futuros:

- Avaliar outros algoritmos de otimização para calibração de modelos hidrológicos conceituais, não só multiobjetivos, mas os de objetivo único com uma função agregadora de todas as métricas desejadas;
- Implementar e testar uma versão do SCE-UA que evolua os complexos por evolução diferencial com mutação;
- Estender os testes de calibração para situações em que sejam fornecidas exclusivamente estimativas regionalizadas das assinaturas;
- Analisar o período de validação do melhor conjunto de parâmetros de cada estrutura em cada bacia, em uma tentativa de determinar relações entre as métricas de desempenho e atributos físicos e/ou assinaturas hidrológicas,

assim como foi realizado por Knoben et al. (2020), mas agora para as 624 estruturas do FUSE;

- Desenvolver uma estrutura de modelos conceituais em uma abordagem hierárquica progressiva. Ainda que não represente uma ideia inovadora, parece coerente buscar estruturas o mais simples possíveis e realizar um incremento da complexidade de maneira automática, ao invés de sugerir uma sequência de mecanismos como proposto pelo FUSE ou o SUPERFLEX;
- Avaliar a possibilidade de ativar mecanismos específicos ao longo do tempo de simulação;
- Avaliar o modelo de resíduos e a incerteza preditiva nas estruturas calibradas utilizando o FUSE (HUNTER et al., 2021);
- Explorar o treinamento de redes LSTM em que o objetivo seja a reprodução de um conjunto de assinaturas hidrológicas, ao invés das séries temporais de vazão. Buscar e/ou propor métodos que permitam uma calibração multiobjetivo também é premente;
- Explorar a combinação de redes LSTM com Rede Neurais Convolucionais (CNN);
- Investigar o emprego de novos produtos do sensoriamento remoto para seleção da estrutura conceitual do modelo, inclusive a partir da busca por padrões espaciais de similaridade no caso de modelos distribuídos;
- Avaliar a influência da extrapolação dos limites da série temporal utilizada durante o treinamento, na fase de validação de modelos baseados em dados;
- Explorar o número de incrementos de tempo fornecidos a rede LSTM; e
- Avaliar o SHAP sobre as redes LSTM.

REFERÊNCIAS

ADDOR, N. et al. The CAMELS data set: Catchment attributes and meteorology for large-sample studies. **Hydrology and Earth System Sciences**, v. 21, n. 10, p. 5293–5313, 20 out. 2017.

ADDOR, N. et al. A Ranking of Hydrological Signatures Based on Their Predictability in Space. **Water Resources Research**, v. 54, n. 11, p. 8792–8812, 1 nov. 2018.

ADDOR, N. et al. Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges. **Hydrological Sciences Journal**, v. 65, n. 5, p. 712–725, 3 abr. 2020.

ADDOR, N.; MELSEN, L. A. Legacy, Rather Than Adequacy, Drives the Selection of Hydrological Models. **Water Resources Research**, v. 55, n. 1, p. 378–390, 2019.

ADEYERI, O. E. et al. Conceptual hydrological model calibration using multi-objective optimization techniques over the transboundary Komadugu-Yobe basin, Lake Chad Area, West Africa. **Journal of Hydrology: Regional Studies**, v. 27, p. 100655, 1 fev. 2020.

AGÊNCIA NACIONAL DE ÁGUAS (ANA). **Mapeamento das Áreas Aflorantes dos Aquíferos e Sistemas Aquíferos do Brasil - 1:1.000.000.** , 2022. Disponível em: <https://dadosabertos.ana.gov.br/datasets/8355dc70dee94888bc7ac4b79030134a_1/about>

AGRAWAL, R.; IMIELIŃSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. **ACM SIGMOD Record**, v. 22, n. 2, p. 207–216, 1 jun. 1993.

AGRAWAL, R.; SRIKANT, R. **Fast Algorithms for Mining Association Rules in Large Databases**. Proceedings of the 20th International Conference on Very Large Data Bases. **Anais...: VLDB '94**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 12 set. 1994. . Acesso em: 9 nov. 2022

ALJAMAAN, H.; ALAZBA, A. **Software defect prediction using tree-based ensembles**. Proceedings of the 16th ACM International Conference on Predictive Models and Data Analytics in Software Engineering. **Anais...: PROMISE 2020**. New York, NY, USA: Association for Computing Machinery, 8 nov. 2020. Disponível em: <<https://dl.acm.org/doi/10.1145/3416508.3417114>>. Acesso em: 16 mar. 2023

ALMAGRO, A. et al. CABra: A novel large-sample dataset for Brazilian catchments. **Hydrology and Earth System Sciences**, v. 25, n. 6, p. 3105–3135, 9 jun. 2021.

ALTHOFF, D.; RODRIGUES, L. N. Goodness-of-fit criteria for hydrological models: Model calibration and performance assessment. **Journal of Hydrology**, v. 600, p. 126674, 1 set. 2021.

ALTHOFF, D.; RODRIGUES, L. N.; SILVA, D. D. DA. Addressing hydrological modeling in watersheds under land cover change with deep learning. **Advances in Water Resources**, v. 154, p. 103965, 1 ago. 2021.

ANAND, M.; VELU, A.; WHIG, P. Prediction of Loan Behaviour with Machine Learning Models for Secure Banking. **Journal of Computer Science and Engineering (JCSE)**, v. 3, n. 1, p. 1–13, 15 fev. 2022.

ANDRÉASSIAN, V. et al. Introduction and synthesis: Why should hydrologists work on a large number of basin data sets? p. 1, 2006.

ARAZA, A. et al. **Data-driven streamflow modelling in ungauged basins: regionalizing random forest (RF) models**. [s.l.] Ecology, 16 nov. 2020. Disponível em: <<http://biorxiv.org/lookup/doi/10.1101/2020.11.14.382598>>. Acesso em: 20 mar. 2023.

ARCHFIELD, S. A. et al. AN OBJECTIVE AND PARSIMONIOUS APPROACH FOR CLASSIFYING NATURAL FLOW REGIMES AT A CONTINENTAL SCALE: CLASSIFICATION OF NATURAL FLOW REGIMES. **River Research and Applications**, v. 30, n. 9, p. 1166–1183, nov. 2014.

ARNOLD, J. et al. SWAT: Model use, calibration, and validation. v. 55, 1 jul. 2012.

ARNOLD, J. G. et al. Automated Base Flow Separation and Recession Analysis Techniques. **Groundwater**, v. 33, n. 6, p. 1010–1018, 1995.

ARSENAULT, R.; BRISSETTE, F. P. Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches. **Water Resources Research**, v. 50, n. 7, p. 6135–6153, 2014.

BECK, H. E. et al. Global-scale regionalization of hydrologic model parameters. **Water Resources Research**, v. 52, n. 5, p. 3599–3622, 2016.

BENGIO, Y. **Practical recommendations for gradient-based training of deep architectures**. arXiv, , 16 set. 2012. Disponível em: <<http://arxiv.org/abs/1206.5533>>. Acesso em: 16 jul. 2023

BENGIO, Y.; SIMARD, P.; FRASCONI, P. Learning long-term dependencies with gradient descent is difficult. **IEEE Transactions on Neural Networks**, v. 5, n. 2, p. 157–166, mar. 1994.

BENTÉJAC, C.; CSÖRGŐ, A.; MARTÍNEZ-MUÑOZ, G. A comparative analysis of gradient boosting algorithms. **Artificial Intelligence Review**, v. 54, n. 3, p. 1937–1967, 1 mar. 2021.

BERGHUIJS, W. R. et al. Dominant flood generating mechanisms across the United States. **Geophysical Research Letters**, v. 43, n. 9, p. 4382–4390, 2016.

BERGHUIJS, W. R.; WOODS, R. A.; HRACHOWITZ, M. A precipitation shift from snow towards rain leads to a decrease in streamflow. **Nature Climate Change**, v. 4, n. 7, p. 583–586, jul. 2014.

BERGSTRA, J. et al. **Algorithms for Hyper-Parameter Optimization**. Advances in Neural Information Processing Systems. **Anais...**Curran Associates, Inc., 2011. Disponível em:

<https://papers.nips.cc/paper_files/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html>. Acesso em: 16 jul. 2023

BERHANU, B. et al. Flow Regime Classification and Hydrological Characterization: A Case Study of Ethiopian Rivers. **Water**, v. 7, n. 6, p. 3149–3165, jun. 2015.

BEVEN, K. Changing ideas in hydrology — The case of physically-based models. **Journal of Hydrology**, v. 105, n. 1, p. 157–172, 30 jan. 1989.

BEVEN, K. J. **Rainfall-Runoff Modelling 2e: The Primer**. 2nd ed. edição ed. Chichester, West Sussex ; Hoboken, NJ: Wiley-Blackwell, 2012.

BEVEN, K. J.; KIRKBY, M. J. A physically based, variable contributing area model of basin hydrology / Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. **Hydrological Sciences Bulletin**, v. 24, n. 1, p. 43–69, 1 mar. 1979.

BISHOP, C. M. **Pattern recognition and machine learning**. New York: Springer, 2006.

BLÖSCHL, G. Rainfall-Runoff Modeling of Ungauged Catchments. Em: **Encyclopedia of Hydrological Sciences**. [s.l.] John Wiley & Sons, Ltd, 2006.

BOMHOF, J.; TOLSON, B. A.; KOUWEN, N. Comparing single and multi-objective hydrologic model calibration considering reservoir inflow and streamflow observations. **Canadian Water Resources Journal / Revue canadienne des ressources hydriques**, v. 44, n. 4, p. 319–336, 2 out. 2019.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, n. 2, p. 123–140, 1 ago. 1996.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 1 out. 2001.

BROWN, G.; CRAIG, J. R. Structural calibration of an semi-distributed hydrological model of the Liard River basin. **Canadian Water Resources Journal / Revue canadienne des ressources hydriques**, v. 45, n. 4, p. 287–303, 1 out. 2020.

CALIŃSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics**, v. 3, n. 1, p. 1–27, 1 jan. 1974.

CHADALAWADA, J.; HERATH, H. M. V. V.; BABOVIC, V. Hydrologically Informed Machine Learning for Rainfall-Runoff Modeling: A Genetic Programming-Based Toolkit for Automatic Model Induction. **Water Resources Research**, v. 56, n. 4, p. e2019WR026933, 2020.

CHATURVEDI, A.; GREEN, P. E.; CAROLL, J. D. K-modes Clustering. **Journal of Classification**, v. 18, n. 1, p. 35–55, 1 jan. 2001.

CHEN, B. et al. Research on weighting method of geological hazard susceptibility evaluation index based on apriori Algorithm. **Frontiers in Earth Science**, v. 11, 2023.

CHEN, M.-S.; HAN, J.; YU, P. S. Data mining: an overview from a database perspective. **IEEE Transactions on Knowledge and Data Engineering**, v. 8, n. 6, p. 866–883, dez. 1996.

CHEN, T.; GUESTRIN, C. **XGBoost: A Scalable Tree Boosting System**. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **Anais...**13 ago. 2016. Disponível em: <<http://arxiv.org/abs/1603.02754>>. Acesso em: 5 mar. 2023

CHEUNG, D. W. et al. Efficient mining of association rules in distributed databases. **IEEE Transactions on Knowledge and Data Engineering**, v. 8, n. 6, p. 911–922, dez. 1996.

CHILKOTI, V.; BOLISETTI, T.; BALACHANDAR, R. Multi-objective autocalibration of SWAT model for improved low flow performance for a small snowfed catchment. **Hydrological Sciences Journal**, v. 63, n. 10, p. 1482–1501, 27 jul. 2018.

CHILKOTI, V.; BOLISETTI, T.; BALACHANDAR, R. Diagnostic Evaluation of Hydrologic Models Employing Flow Duration Curve. **Journal of Hydrologic Engineering**, v. 24, n. 6, p. 05019009, jun. 2019.

CHOUBIN, B. et al. Streamflow regionalization using a similarity approach in ungauged basins: Application of the geo-environmental signatures in the Karkheh River Basin, Iran. **CATENA**, v. 182, p. 104128, 1 nov. 2019.

CHOW, V.; MAIDMENT, D.; MAYS, L. **Applied Hydrology**. 1st edition ed. New York: McGraw-Hill Science/Engineering/Math, 1988.

CHOWDHURY, S. U. et al. Shapley-Additive-Explanations-Based Factor Analysis for Dengue Severity Prediction using Machine Learning. **Journal of Imaging**, v. 8, n. 9, p. 229, set. 2022.

CLARK, M. P. et al. Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. **Water Resources Research**, v. 44, n. 12, dez. 2008.

CLARK, M. P. et al. Hydrological field data from a modeller's perspective: Part 2: process-based evaluation of model hypotheses. **Hydrological Processes**, v. 25, n. 4, p. 523–543, 2011.

CLARK, M. P. et al. A unified approach for process-based hydrologic modeling: 1. Modeling concept. **Water Resources Research**, v. 51, n. 4, p. 2498–2514, 1 abr. 2015.

COXON, G. et al. Diagnostic evaluation of multiple hypotheses of hydrological behaviour in a limits-of-acceptability framework for 24 UK catchments. **Hydrological Processes**, v. 28, n. 25, p. 6135–6150, 2014.

CRAIG, J. R. et al. Flexible watershed simulation with the Raven hydrological modelling framework. **Environmental Modelling & Software**, v. 129, p. 104728, 1 jul. 2020.

DAGGUPATI, P. et al. Impact of model development, calibration and validation decisions on hydrological simulations in West Lake Erie Basin. **Hydrological Processes**, v. 29, n. 26, p. 5307–5320, 2015.

DAL MOLIN, M. et al. Exploring Signature-Based Model Calibration for Streamflow Prediction in Ungauged Basins. **Water Resources Research**, v. 59, n. 7, p. e2022WR031929, 2023.

DAL MOLIN, M.; KAVETSKI, D.; FENICIA, F. SuperflexPy 1.3.0: an open-source Python framework for building, testing, and improving conceptual hydrological models. **Geoscientific Model Development**, v. 14, n. 11, p. 7047–7072, 19 nov. 2021.

DANIELL, T. **Neural networks. Applications in hydrology and water resources engineering**. National Conference Publication- Institute of Engineers. Australia. **Anais...**1991.

DAS, I.; DENNIS, J. E. Normal-Boundary Intersection: A New Method for Generating the Pareto Surface in Nonlinear Multicriteria Optimization Problems. **SIAM Journal on Optimization**, v. 8, n. 3, p. 631–657, ago. 1998.

DAVID, P. C. et al. Correspondence Between Model Structures and Hydrological Signatures: A Large-Sample Case Study Using 508 Brazilian Catchments. **Water Resources Research**, v. 58, n. 3, p. e2021WR030619, mar. 2022.

DE BOER-EUSER, T. et al. Looking beyond general metrics for model comparison – Lessons from an international model intercomparison study. **Hydrology and Earth System Sciences**, v. 21, n. 1, p. 423–440, 25 jan. 2017.

DEB, K. et al. A fast and elitist multiobjective genetic algorithm: NSGA-II. **IEEE Transactions on Evolutionary Computation**, v. 6, n. 2, p. 182–197, abr. 2002.

DEB, K. Multi-objective Optimisation Using Evolutionary Algorithms: An Introduction. Em: WANG, L.; NG, A. H. C.; DEB, K. (Eds.). **Multi-objective Evolutionary Optimisation for Product Design and Manufacturing**. London: Springer, 2011. p. 3–34.

DEB, K.; JAIN, H. An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints. **IEEE Transactions on Evolutionary Computation**, v. 18, n. 4, p. 577–601, ago. 2014.

DESAI, S.; OUARDA, T. B. M. J. Regional hydrological frequency analysis at ungauged sites with random forest regression. **Journal of Hydrology**, v. 594, p. 125861, mar. 2021.

DOOGE, J. C. I. Looking for hydrologic laws. **Water Resources Research**, v. 22, n. 9S, p. 46S-58S, 1 ago. 1986.

DUAN, Q.; SOROOSHIAN, S.; GUPTA, V. Effective and Efficient Global Optimization for Conceptual Rainfall-Runoff Models. **Water Resources Research - WATER RESOUR RES**, v. 28, p. 1015–1031, 1 abr. 1992.

DUAN, Q.; SOROOSHIAN, S.; GUPTA, V. K. Optimal use of the SCE-UA global optimization method for calibrating watershed models. **Journal of Hydrology**, v. 158, n. 3, p. 265–284, 15 jun. 1994.

DUAN, S.; ULLRICH, P.; SHU, L. Using Convolutional Neural Networks for Streamflow Projection in California. **Frontiers in Water**, v. 2, 2020.

ECKHARDT, K. A comparison of baseflow indices, which were calculated with seven different baseflow separation methods. **Journal of Hydrology**, v. 352, n. 1, p. 168–173, 30 abr. 2008.

EFSTRATIADIS, A.; KOUTSOYIANNIS, D. One decade of multi-objective calibration approaches in hydrological modelling: a review. **Hydrological Sciences Journal**, v. 55, n. 1, p. 58–78, 10 mar. 2010.

EL BILALI, A. et al. An interpretable machine learning approach based on DNN, SVR, Extra Tree, and XGBoost models for predicting daily pan evaporation. **Journal of Environmental Management**, v. 327, p. 116890, fev. 2023.

ERCAN, M. B.; GOODALL, J. L. Design and implementation of a general software library for using NSGA-II with SWAT for multi-objective model calibration. **Environmental Modelling & Software**, v. 84, p. 112–120, 1 out. 2016.

FANG, K.; PAN, M.; SHEN, C. The Value of SMAP for Long-Term Soil Moisture Estimation With the Help of Deep Learning. **IEEE Transactions on Geoscience and Remote Sensing**, v. 57, n. 4, p. 2221–2233, abr. 2019.

FANG, Z. et al. Predicting flood susceptibility using LSTM neural networks. **Journal of Hydrology**, v. 594, p. 125734, 2021.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 37–37, 15 mar. 1996.

FENICIA, F. et al. From spatially variable streamflow to distributed hydrological models: Analysis of key modeling decisions. **Water Resources Research**, v. 52, n. 2, p. 954–989, fev. 2016.

FENICIA, F. et al. Signature-Domain Calibration of Hydrological Models Using Approximate Bayesian Computation: Empirical Analysis of Fundamental Properties. **Water Resources Research**, v. 54, n. 6, p. 3958–3987, 1 jun. 2018.

FRAME, J. M. et al. Deep learning rainfall–runoff predictions of extreme events. **Hydrology and Earth System Sciences**, v. 26, n. 13, p. 3377–3392, 5 jul. 2022.

FREUND, Y.; SCHAPIRE, R. E. A Short Introduction to Boosting. 1999.

FRIEDMAN, J. H. Greedy Function Approximation: A Gradient Boosting Machine. **The Annals of Statistics**, v. 29, n. 5, p. 1189–1232, 2001.

FRIEDMAN, J. H.; MEULMAN, J. J. Multiple additive regression trees with application in epidemiology. **Statistics in Medicine**, v. 22, n. 9, p. 1365–1381, 15 maio 2003.

FUENTE, D. LA; ANDRÉS, L. Using Big-Data to Develop Catchment-Scale Hydrological Models for Chile. 2021.

GAN, Y. et al. A systematic assessment and reduction of parametric uncertainties for a distributed hydrological model. **Journal of Hydrology**, v. 564, p. 697–711, 1 set. 2018.

GARCIA, F.; FOLTON, N.; OUDIN, L. Which objective function to calibrate rainfall–runoff models for low-flow index simulations? **Hydrological Sciences Journal**, v. 62, n. 7, p. 1149–1166, 19 maio 2017.

GOBEYN, S.; GOETHALS, P. L. M. Multi-objective optimisation of species distribution models for river management. **Water Research**, v. 163, p. 114863, 15 out. 2019.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [s.l.] MIT Press, 2016.

GRAYSON, R. B.; MOORE, I. D.; MCMAHON, T. A. Physically based hydrologic modeling: 2. Is the concept realistic? **Water Resources Research**, v. 28, n. 10, p. 2659–2666, 1992.

GUNTURI, S. K.; SARKAR, D. Ensemble machine learning models for the detection of energy theft. **Electric Power Systems Research**, v. 192, p. 106904, 1 mar. 2021.

GUO, J. et al. A Novel Multi-Objective Shuffled Complex Differential Evolution Algorithm with Application to Hydrological Model Parameter Optimization. **Water Resources Management**, v. 27, n. 8, p. 2923–2946, jun. 2013.

GUO, Y. et al. Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: A comprehensive review. **WIREs Water**, v. 8, n. 1, p. e1487, 2021a.

GUO, Y. et al. Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: A comprehensive review. **Wiley Interdisciplinary Reviews: Water**, v. 8, n. 1, p. e1487, 1 jan. 2021b.

GUPTA, H. V. et al. 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, 20 out. 2009.

GUPTA, H. V. et al. Large-sample hydrology: a need to balance depth with breadth. **Hydrology and Earth System Sciences**, v. 18, n. 2, p. 463–477, 6 fev. 2014.

GUPTA, H. V.; WAGENER, T.; LIU, Y. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. **Hydrological Processes**, v. 22, n. 18, p. 3802–3813, 30 ago. 2008.

HALLOUIN, T.; BRUEN, M.; O'LOUGHLIN, F. E. Calibration of hydrological models for ecologically relevant streamflow predictions: a trade-off between fitting well to data and estimating consistent parameter sets? **Hydrology and Earth System Sciences**, v. 24, n. 3, p. 1031–1054, 4 mar. 2020.

HAND, D. J.; MANNILA, H.; SMYTH, P. **Principles of data mining**. Cambridge, Mass: MIT Press, 2001.

HANUS, S. et al. Future changes in annual, seasonal and monthly runoff signatures in contrasting Alpine catchments in Austria. **Hydrology and Earth System Sciences**, v. 25, n. 6, p. 3429–3453, 17 jun. 2021.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition**. 2nd edition ed. New York, NY: Springer, 2016.

HE, Y.; BÁRDOSSY, A.; ZEHE, E. A review of regionalisation for continuous streamflow simulation. **Hydrology and Earth System Sciences**, v. 15, n. 11, p. 3539–3553, 23 nov. 2011.

HER, Y.; SEONG, C. Responses of hydrological model equifinality, uncertainty, and performance to multi-objective parameter calibration. **Journal of Hydroinformatics**, v. 20, n. 4, p. 864–885, 20 mar. 2018.

HERATH, H. M. V. V.; CHADALAWADA, J.; BABOVIC, V. Hydrologically informed machine learning for rainfall–runoff modelling: towards distributed modelling. **Hydrology and Earth System Sciences**, v. 25, n. 8, p. 4373–4401, 11 ago. 2021.

HERNANDEZ-SUAREZ, J. S. et al. Evaluation of the impacts of hydrologic model calibration methods on predictability of ecologically-relevant hydrologic indices. **Journal of Hydrology**, v. 564, p. 758–772, 1 set. 2018.

HERNANDEZ-SUAREZ, J. S.; NEJADHASHEMI, A. P.; DEB, K. A novel multi-objective model calibration method for ecohydrological applications. **Environmental Modelling & Software**, v. 144, p. 105161, out. 2021.

HOCHREITER, S.; SCHMIDHUBER, J. Long Short-Term Memory. **Neural Computation**, v. 9, n. 8, p. 1735–1780, 15 nov. 1997.

HÖGE, M. et al. Improving hydrologic models for predictions and process understanding using neural ODEs. **Hydrology and Earth System Sciences**, v. 26, n. 19, p. 5085–5102, 11 out. 2022.

HÖGE, M.; WÖHLING, T.; NOWAK, W. A Primer for Model Selection: The Decisive Role of Model Complexity. **Water Resources Research**, v. 54, n. 3, p. 1688–1715, mar. 2018.

HRACHOWITZ, M. et al. A decade of Predictions in Ungauged Basins (PUB)—a review. **Hydrological Sciences Journal**, v. 58, n. 6, p. 1198–1255, 1 ago. 2013.

HRACHOWITZ, M. et al. Process consistency in models: The importance of system signatures, expert knowledge, and process complexity. **Water Resources Research**, v. 50, n. 9, p. 7445–7469, 2014.

HRACHOWITZ, M.; CLARK, M. P. HESS Opinions: The complementary merits of competing modelling philosophies in hydrology. **Hydrology and Earth System Sciences**, v. 21, n. 8, p. 3953–3973, 1 ago. 2017.

HSU, K. LIN; GUPTA, H. V.; SOROOSHIAN, S. **Application of a recurrent neural network to rainfall-runoff modeling: Proceedings of the 1997 24th Annual Water Resources Planning and Management Conference**. 1997. Disponível em: <<http://www.scopus.com/inward/record.url?scp=0030645841&partnerID=8YFLogxK>> . Acesso em: 15 jul. 2023

HUANG, G. et al. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. **Journal of Hydrology**, v. 574, p. 1029–1041, jul. 2019.

HUANG, S. et al. Monthly streamflow prediction using modified EMD-based support vector machine. **Journal of Hydrology**, v. 511, p. 764–775, 16 abr. 2014.

HUNT, K. M. R. et al. **Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States**. [s.l.] Global hydrology/Modelling approaches, 8 fev. 2022. Disponível em: <<https://hess.copernicus.org/preprints/hess-2022-53/hess-2022-53.pdf>>. Acesso em: 12 mar. 2023.

HUNTER, J. et al. Achieving high-quality probabilistic predictions from hydrological models calibrated with a wide range of objective functions. **Journal of Hydrology**, v. 603, p. 126578, 1 dez. 2021.

HUO, J.; LIU, L. Application research of multi-objective Artificial Bee Colony optimization algorithm for parameters calibration of hydrological model. **Neural Computing and Applications**, v. 31, n. 9, p. 4715–4732, 1 set. 2019.

HUO, J.; LIU, L.; ZHANG, Y. An improved multi-cores parallel artificial Bee colony optimization algorithm for parameters calibration of hydrological model. **Future Generation Computer Systems**, v. 81, p. 492–504, abr. 2018.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTIC (IBGE). **Mapa de Climatologia do Brasil**, 2002.

INSTITUTO BRASILEIRO DE GEOGRAFIA (IBGE). **Biomass e Sistema Costeiro-Marinho do Brasil - 1:250 000**. Rio de Janeiro:IBGE, , 2019. Disponível em: <<https://www.ibge.gov.br/geociencias/cartas-e-mapas/informacoes-ambientais/15842-biomass.html?=&t=sobre>>

JAFFRÉS, J. B. D. et al. Hydrological characteristics of Australia: national catchment classification and regional relationships. **Journal of Hydrology**, v. 612, p. 127969, 1 set. 2022.

JAIN, A. K. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters**, v. 31, n. 8, p. 651–666, 1 jun. 2010.

JAIN, H.; DEB, K. An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point Based Nondominated Sorting Approach, Part II: Handling Constraints and Extending to an Adaptive Approach. **IEEE Transactions on Evolutionary Computation**, v. 18, n. 4, p. 602–622, ago. 2014.

JAKEMAN, A. J.; HORNBERGER, G. M. How much complexity is warranted in a rainfall-runoff model? **Water Resources Research**, v. 29, n. 8, p. 2637–2649, 1993.

JEHN, F. U. et al. Using hydrological and climatic catchment clusters to explore drivers of catchment behavior. **Hydrology and Earth System Sciences**, v. 24, n. 3, p. 1081–1100, 5 mar. 2020.

JENA, R. et al. Earthquake vulnerability assessment for the Indian subcontinent using the Long Short-Term Memory model (LSTM). **International Journal of Disaster Risk Reduction**, v. 66, p. 102642, 2021.

JI, H. et al. Adaptability of machine learning methods and hydrological models to discharge simulations in data-sparse glaciated watersheds. **Journal of Arid Land**, v. 13, n. 6, p. 549–567, 1 jun. 2021.

JOLLIFFE, I. T. **Principal Component Analysis**. [s.l.] Springer Science & Business Media, 2006.

KAN, G. et al. Fast hydrological model calibration based on the heterogeneous parallel computing accelerated shuffled complex evolution method. **Engineering Optimization**, v. 50, n. 1, p. 106–119, 2 jan. 2018a.

KAN, G. et al. Improving water quantity simulation & forecasting to solve the energy-water-food nexus issue by using heterogeneous computing accelerated global optimization method. **Applied Energy**, v. 210, p. 420–433, 15 jan. 2018b.

KAVETSKI, D. et al. Signature-Domain Calibration of Hydrological Models Using Approximate Bayesian Computation: Theory and Comparison to Existing Applications. **Water Resources Research**, v. 54, n. 6, p. 4059–4083, 2018.

KAVETSKI, D.; FENICIA, F. Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights. **Water Resources Research**, v. 47, n. 11, p. 11511, 1 nov. 2011.

KE, G. et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 2017.

KESKAR, N. S. et al. **On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima**. arXiv, , 9 fev. 2017. Disponível em: <<http://arxiv.org/abs/1609.04836>>. Acesso em: 16 jul. 2023

KHU, S. T.; MADSEN, H. Multiobjective calibration with Pareto preference ordering: An application to rainfall-runoff model calibration. **Water Resources Research**, v. 41, n. 3, 2005.

KIESEL, J. et al. When is a hydrological model sufficiently calibrated to depict flow preferences of riverine species? **Ecohydrology**, v. 13, n. 3, p. e2193, 2020.

KIRCHNER, J. W. Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. **Water Resources Research**, v. 42, n. 3, 1 mar. 2006.

KITTEL, C. M. M. et al. Informing hydrological models of poorly gauged river catchments – A parameter regionalization and calibration approach. **Journal of Hydrology**, v. 587, p. 124999, 1 ago. 2020.

KNOBEN, W. J. M. et al. A Brief Analysis of Conceptual Model Structure Uncertainty Using 36 Models and 559 Catchments. **Water Resources Research**, v. 56, n. 9, p. e2019WR025975, 2020.

KOCH, J.; SCHNEIDER, R. Long short-term memory networks enhance rainfall-runoff modelling at the national scale of Denmark. **GEUS Bulletin**, v. 49, 13 jan. 2022.

KOHAVI, R. **A study of cross-validation and bootstrap for accuracy estimation and model selection**. Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2. **Anais...: IJCAI'95**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 20 ago. 1995. . Acesso em: 8 ago. 2023

KONAPALA, G.; KAO, S.-C.; ADDOR, N. Exploring Hydrologic Model Process Connectivity at the Continental Scale Through an Information Theory Approach. **Water Resources Research**, v. 56, n. 10, p. e2020WR027340, 2020.

KRATZERT, F. et al. Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. **Hydrology and Earth System Sciences**, v. 22, n. 11, p. 6005–6022, 22 nov. 2018.

KRATZERT, F. et al. Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. **Water Resources Research**, v. 55, n. 12, p. 11344–11354, 1 dez. 2019a.

KRATZERT, F. et al. 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, 17 dez. 2019b.

KUENTZ, A. et al. Understanding hydrologic variability across Europe through catchment classification. **Hydrology and Earth System Sciences**, v. 21, n. 6, p. 2863–2879, 12 jun. 2017.

KUNNATH-POOVAKKA, A. et al. Remotely sensed ET for streamflow modelling in catchments with contrasting flow characteristics: an attempt to improve efficiency. **Stochastic Environmental Research and Risk Assessment**, v. 32, n. 7, p. 1973–1992, 1 jul. 2018.

LALOY, E. et al. Inversion using a new low-dimensional representation of complex binary geological media based on a deep neural network. **Advances in Water Resources**, v. 110, p. 387–405, 1 dez. 2017.

LANE, R. A. et al. Benchmarking the predictive capability of hydrological models for river flow and flood peak predictions across over 1000 catchments in Great Britain. **Hydrology and Earth System Sciences**, v. 23, n. 10, p. 4011–4032, 30 set. 2019.

LASHENG, C.; YUQIANG, L. **Improved initial clustering center selection algorithm for K-means**. 2017 Signal Processing: Algorithms, Architectures, Arrangements, and

Applications (SPA). **Anais...** Em: 2017 SIGNAL PROCESSING: ALGORITHMS, ARCHITECTURES, ARRANGEMENTS, AND APPLICATIONS (SPA). set. 2017.

LE, M.-H. et al. **Streamflow Estimation in Ungauged Regions using Machine Learning: Quantifying Uncertainties in Geographic Extrapolation**. [s.l.] Water Resources Management/Uncertainty analysis, 9 set. 2022. Disponível em: <<https://hess.copernicus.org/preprints/hess-2022-320/>>. Acesso em: 16 mar. 2023.

LEAVESLEY, G. H. et al. **Precipitation-runoff modeling system; user's manual**. Precipitation-runoff modeling system; user's manual: Water-Resources Investigations Report. [s.l.] U.S. Geological Survey, Water Resources Division, 1983. Disponível em: <<http://pubs.er.usgs.gov/publication/wri834238>>. Acesso em: 4 jul. 2023.

LEBECHEREL, L.; ANDRÉASSIAN, V.; PERRIN, C. On evaluating the robustness of spatial-proximity-based regionalization methods. **Journal of Hydrology**, v. 539, p. 196–203, 1 ago. 2016.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, maio 2015.

LEES, T. et al. Benchmarking data-driven rainfall-runoff models in Great Britain: A comparison of long short-term memory (LSTM)-based models with four lumped conceptual models. **Hydrology and Earth System Sciences**, v. 25, n. 10, p. 5517–5534, 21 out. 2021.

LEES, T. et al. Hydrological concept formation inside long short-term memory (LSTM) networks. **Hydrology and Earth System Sciences**, v. 26, n. 12, p. 3079–3101, 20 jun. 2022.

LEONTJEVA, A.; KUZOVKIN, I. **Combining Static and Dynamic Features for Multivariate Sequence Classification**. 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). **Anais...**IEEE, out. 2016. Disponível em: <<http://ieeexplore.ieee.org/document/7796887/>>. Acesso em: 27 maio. 2022

LEY, A.; BORMANN, H.; CASPER, M. Intercomparing LSTM and RNN to a Conceptual Hydrological Model for a Low-Land River with a Focus on the Flow Duration Curve. **Water**, v. 15, n. 3, p. 505, jan. 2023.

LEY, R. et al. Comparing classical performance measures with signature indices derived from flow duration curves to assess model structures as tools for catchment classification. **Hydrology Research**, v. 47, n. 1, p. 1–14, 12 maio 2015.

LI, F. et al. Equipment Quality Information Mining Method Based on Improved Apriori Algorithm. **Journal of Sensors**, v. 2023, p. e2155590, 3 maio 2023.

LI, L. et al. Optimized Apriori algorithm for deformation response analysis of landslide hazards. **Computers & Geosciences**, v. 170, p. 105261, 1 nov. 2022.

LI, N.; SHEPPERD, M.; GUO, Y. A systematic review of unsupervised learning techniques for software defect prediction. **Information and Software Technology**, v. 122, p. 106287, 1 jun. 2020.

LIANG, X. et al. A simple hydrologically based model of land surface water and energy fluxes for general circulation models. **Journal of Geophysical Research: Atmospheres**, v. 99, n. D7, p. 14415–14428, 1994.

LIANG, Z. et al. Simulate the forecast capacity of a complicated water quality model using the long short-term memory approach. **Journal of Hydrology**, v. 581, p. 124432, 2020.

LIONG, S.-Y.; KHU, S.-T.; CHAN, W.-T. Derivation of Pareto Front with Genetic Algorithm and Neural Network. **Journal of Hydrologic Engineering**, v. 6, n. 1, p. 52–61, 1 jan. 2001.

LIU, Y.; WU, X.; SHEN, Y. Automatic clustering using genetic algorithms. **Applied Mathematics and Computation**, v. 218, n. 4, p. 1267–1279, 15 out. 2011.

LIU, Z.; ZHOU, P.; ZHANG, Y. A Probabilistic Wavelet–Support Vector Regression Model for Streamflow Forecasting with Rainfall and Climate Information Input. **Journal of Hydrometeorology**, v. 16, n. 5, p. 2209–2229, 1 out. 2015.

LOGANATHAN, P.; MAHINDRAKAR, A. B. Intercomparing the robustness of machine learning models in simulation and forecasting of streamflow. **Journal of Water and Climate Change**, v. 12, n. 5, p. 1824–1837, 24 nov. 2020.

LU, Y. et al. Machine Learning Models Using SHapley Additive exPlanation for Fire Risk Assessment Mode and Effects Analysis of Stadiums. **Sensors**, v. 23, n. 4, p. 2151, jan. 2023.

LUNDBERG, S. M.; ERION, G. G.; LEE, S.-I. **Consistent Individualized Feature Attribution for Tree Ensembles**. arXiv, , 6 mar. 2019. Disponível em: <<http://arxiv.org/abs/1802.03888>>. Acesso em: 9 abr. 2023

LUNDBERG, S. M.; LEE, S.-I. **A Unified Approach to Interpreting Model Predictions**. Advances in Neural Information Processing Systems. **Anais...Curran Associates, Inc.**, 2017. Disponível em: <<https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>>. Acesso em: 9 abr. 2023

LYNE, V.; HOLLICK, M. **Stochastic Time-Variable Rainfall-Runoff Modeling**. . Em: AUSTRALIAN NATIONAL CONFERENCE PUBLICATION. 1 jan. 1979.

LYON, S. W.; TROCH, P. A. Development and application of a catchment similarity index for subsurface flow. **Water Resources Research**, v. 46, n. 3, 1 mar. 2010.

MA, K. et al. **Transferring hydrologic data across continents -- leveraging US data to improve hydrologic prediction in other countries**. [s.l.] Hydrology, 2 set. 2020. Disponível em: <<https://essopenarchive.org/doi/full/10.1002/essoar.10504132.1>>. Acesso em: 12 mar. 2023.

MADSEN, H. Automatic calibration of a conceptual rainfall–runoff model using multiple objectives. **Journal of Hydrology**, v. 235, n. 3, p. 276–288, 31 ago. 2000.

MADSEN, H. Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. **Advances in Water Resources**, v. 26, n. 2, p. 205–216, 1 fev. 2003.

MAI, J.; CRAIG, J. R.; TOLSON, B. A. Simultaneously determining global sensitivities of model parameters and model structure. **Hydrology and Earth System Sciences**, v. 24, n. 12, p. 5835–5858, 8 dez. 2020.

MARTINEZ, G. F.; GUPTA, H. V. Hydrologic consistency as a basis for assessing complexity of monthly water balance models for the continental United States. **Water Resources Research**, v. 47, n. 12, 1 dez. 2011.

MASSMANN, C. Identification of factors influencing hydrologic model performance using a top-down approach in a large number of U.S. catchments. **Hydrological Processes**, v. 34, n. 1, p. 4–20, jan. 2020.

MATOS, A. C. S. **Estimação Bayesiana dos parâmetros de um modelo hidrológico a partir do emprego de assinaturas hidrológicas**. Dissertação (Mestrado em Saneamento, Meio Ambiente e Recursos Hídricos)—[s.l.] Universidade Federal de Minas Gerais, 2021.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics 1943 5:4**, v. 5, n. 4, p. 115–133, dez. 1943.

MCINERNEY, D. et al. Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. **Water Resources Research**, v. 53, n. 3, p. 2199–2239, 2017.

MCMILLAN, H. K. A review of hydrologic signatures and their applications. **Wiley Interdisciplinary Reviews: Water**, v. 8, n. 1, p. e1499, 1 jan. 2021.

MENDOZA, P. A. et al. Are we unnecessarily constraining the agility of complex process-based models? **Water Resources Research**, v. 51, n. 1, p. 716–728, 2015.

MERZ, R.; TARASOVA, L.; BASSO, S. Parameter's Controls of Distributed Catchment Models—How Much Information is in Conventional Catchment Descriptors? **Water Resources Research**, v. 56, n. 2, p. e2019WR026008, 2020.

MIEBS, G. et al. Efficient Strategies of Static Features Incorporation into the Recurrent Neural Network. **Neural Processing Letters 2020 51:3**, v. 51, n. 3, p. 2301–2316, 31 jan. 2020.

MIZUKAMI, N. et al. On the choice of calibration metrics for “high-flow” estimation using hydrologic models. **Hydrology and Earth System Sciences**, v. 23, p. 2601–2614, 17 jun. 2019.

MONTAVON, G.; ORR, G. B.; MÜLLER, K.-R. (EDS.). **Neural Networks: Tricks of the Trade: Second Edition**. Berlin, Heidelberg: Springer, 2012. v. 7700

MONTEIL, C. et al. Multi-objective calibration by combination of stochastic and gradient-like parameter generation rules – the caRamel algorithm. **Hydrology and Earth System Sciences**, v. 24, n. 6, p. 3189–3209, 19 jun. 2020.

MOSTAFAIE, A. et al. Comparing multi-objective optimization techniques to calibrate a conceptual hydrological model using in situ runoff and daily GRACE data. **Computational Geosciences**, v. 22, n. 3, p. 789–814, 1 jun. 2018.

MULVANEY, T. J. On the use of self-registering rain and flood gauges in making observations of the relations of rainfall and flood discharges in a given catchment. **Proceedings of the Institution of Civil Engineers of Ireland**, v. 4, p. 19–31, 1851.

MUSHTAQ, H. et al. **Hydrologic Interpretation of Machine Learning Models for 10-daily streamflow simulation in Climate sensitive Upper Indus Catchments**. [s.l.] Water Resources Management/Modelling approaches, 2 set. 2022. Disponível em: <<https://hess.copernicus.org/preprints/hess-2022-213/>>. Acesso em: 9 abr. 2023.

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, abr. 1970.

NEARING, G. S. et al. What Role Does Hydrological Science Play in the Age of Machine Learning? **Water Resources Research**, v. 57, n. 3, p. e2020WR028091, 1 mar. 2021.

NELDER, J. A.; MEAD, R. A Simplex Method for Function Minimization. **The Computer Journal**, v. 7, n. 4, p. 308–313, 1 jan. 1965.

NERI, M.; PARAJKA, J.; TOTH, E. Importance of the informative content in the study area when regionalising rainfall-runoff model parameters: the role of nested catchments and gauging station density. **Hydrology and Earth System Sciences**, v. 24, n. 11, p. 5149–5171, 6 nov. 2020.

NEWLAND, C. P. et al. Multi-objective optimisation framework for calibration of Cellular Automata land-use models. **Environmental Modelling & Software**, v. 100, p. 175–200, 1 fev. 2018.

NIJZINK, R. et al. The evolution of root-zone moisture capacities after deforestation: a step towards hydrological predictions under change? **Hydrology and Earth System Sciences**, v. 20, n. 12, p. 4775–4799, 5 dez. 2016.

NOHARA, Y. et al. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. **Computer Methods and Programs in Biomedicine**, v. 214, p. 106584, 1 fev. 2022.

NOORI, N.; KALIN, L. Coupling SWAT and ANN models for enhanced daily streamflow prediction. **Journal of Hydrology**, v. 533, p. 141–151, 1 fev. 2016.

NOROUZI, M.; FLEET, D. J.; SALAKHUTDINOV, R. R. **Hamming Distance Metric Learning**. Advances in Neural Information Processing Systems. **Anais...Curran Associates, Inc., 2012.** Disponível em:

<<https://proceedings.neurips.cc/paper/2012/hash/59b90e1005a220e2ebc542eb9d950b1e-Abstract.html>>. Acesso em: 4 jul. 2023

OLDEN, J. D.; POFF, N. L. Redundancy and the choice of hydrologic indices for characterizing streamflow regimes. **River Research and Applications**, v. 19, n. 2, p. 101–121, 2003.

ORUCHE, R. et al. Transfer learning to improve streamflow forecasts in data sparse regions. 6 dez. 2021.

LOUDIN, L. et al. Spatial proximity, physical similarity, regression and ungauged catchments: A comparison of regionalization approaches based on 913 French catchments. **Water Resources Research**, v. 44, n. 3, 2008.

PENG, T. et al. Streamflow Forecasting Using Empirical Wavelet Transform and Artificial Neural Networks. **Water**, v. 9, n. 6, p. 406, jun. 2017.

PONCE, V. M. **Engineering Hydrology: Principles and Practices**. [s.l.] Pearson, 1994.

POOL, S.; VIS, M.; SEIBERT, J. Regionalization for Ungauged Catchments — Lessons Learned From a Comparative Large-Sample Study. **Water Resources Research**, v. 57, n. 10, p. e2021WR030437, 1 out. 2021.

POULIN, A. et al. Uncertainty of hydrological modelling in climate change impact studies in a Canadian, snow-dominated river basin. **Journal of Hydrology**, v. 409, n. 3, p. 626–636, 9 nov. 2011.

PRECHELT, L. Early Stopping - But When? Em: ORR, G. B.; MÜLLER, K.-R. (Eds.). **Neural Networks: Tricks of the Trade**. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 1998. p. 55–69.

PRIETO, C. et al. Flow Prediction in Ungauged Catchments Using Probabilistic Random Forests Regionalization and New Statistical Adequacy Tests. **Water Resources Research**, v. 55, n. 5, p. 4364–4392, 2019.

PRIETO, C. et al. Identification of Dominant Hydrological Mechanisms Using Bayesian Inference, Multiple Statistical Hypothesis Testing, and Flexible Models. **Water Resources Research**, v. 57, n. 8, p. e2020WR028338, 1 ago. 2021.

PROKHORENKOVA, L. et al. CatBoost: unbiased boosting with categorical features. 2018.

PURSHOUSE, R. C.; FLEMING, P. J. On the Evolutionary Optimization of Many Conflicting Objectives. **IEEE Transactions on Evolutionary Computation**, v. 11, n. 6, p. 770–784, dez. 2007.

PUSHPALATHA, R. et al. A review of efficiency criteria suitable for evaluating low-flow simulations. **Journal of Hydrology**, v. 420–421, p. 171–182, 14 fev. 2012.

QI, W. et al. **A framework to regionalize conceptual model parameters for global hydrological modeling**. [s.l.] Global hydrology/Modelling approaches, 15 jun. 2020.

Disponível em: <<https://hess.copernicus.org/preprints/hess-2020-127/>>. Acesso em: 19 mar. 2023.

QIN, H. et al. Multi-objective differential evolution with adaptive Cauchy mutation for short-term multi-objective optimal hydro-thermal scheduling. **Energy Conversion and Management**, v. 51, n. 4, p. 788–794, abr. 2010.

QIN, Z. et al. Predicting invasions of *Wedelia trilobata* (L.) Hitchc. with Maxent and GARP models. **Journal of Plant Research**, v. 128, n. 5, p. 763–775, 1 set. 2015.

QIU, R. et al. River water temperature forecasting using a deep learning method. **Journal of Hydrology**, v. 595, p. 126016, 2021.

RAHMAN, A.; DASH, S. Data Mining for Student's Trends Analysis Using Apriori Algorithm. **International Journal of Control Theory and Applications**, v. 10, p. 107–115, 1 abr. 2017.

RAZAVI, T.; COULIBALY, P. Streamflow Prediction in Ungauged Basins: Review of Regionalization Methods. **Journal of Hydrologic Engineering**, v. 18, n. 8, p. 958–975, 1 ago. 2013.

REDDY, M. J.; NAGESH KUMAR, D. Multi-objective particle swarm optimization for generating optimal trade-offs in reservoir operation. **Hydrological Processes**, v. 21, n. 21, p. 2897–2909, 2007.

REED, P.; DEVIREDDY, V. Groundwater monitoring design: a case study combining epsilon dominance archiving and automatic parameterization for the nsga-ii. Em: **Applications of Multi-Objective Evolutionary Algorithms**. Advances in Natural Computation. [s.l.] WORLD SCIENTIFIC, 2004. v. Volume 1p. 79–100.

REED, P.; MINSKER, B. S.; GOLDBERG, D. E. Simplifying multiobjective optimization: An automated design methodology for the nondominated sorted genetic algorithm-II. **Water Resources Research**, v. 39, n. 7, 2003.

RIGOTTO, J. P. L. **Previsão de Demanda para Sistemas de Abastecimento de Água em Tempo Real: Modelo Híbrido Baseado em Métodos de Agrupamento e de Redes Neurais**. Dissertação (Mestrado em Saneamento, Meio Ambiente e Recursos Hídricos)—[s.l.] Universidade Federal de Minas Gerais, 2022.

ROJAS-SERNA, C. et al. How should a rainfall-runoff model be parameterized in an almost ungauged catchment? A methodology tested on 609 catchments. **Water Resources Research**, v. 52, n. 6, p. 4765–4784, 2016.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, p. 53–65, 1 nov. 1987.

RUDER, S. **An overview of gradient descent optimization algorithms**. arXiv, , 15 jun. 2017. Disponível em: <<http://arxiv.org/abs/1609.04747>>. Acesso em: 16 jul. 2023

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **Nature** **1986 323:6088**, v. 323, n. 6088, p. 533–536, 1986.

SAAVEDRA, D. et al. A multi-objective approach to select hydrological models and constrain structural uncertainties for climate impact assessments. **Hydrological Processes**, v. 36, n. 1, p. e14446, 2022.

SADEGH, M. et al. The stationarity paradigm revisited: Hypothesis testing using diagnostics, summary metrics, and DREAM(ABC). **Water Resources Research**, v. 51, n. 11, p. 9207–9231, 2015.

SAHRAEI, S.; ASADZADEH, M.; SHAFII, M. Toward effective many-objective optimization: Rounded-archiving. **Environmental Modelling & Software**, v. 122, p. 104535, 1 dez. 2019.

SAHRAEI, S.; ASADZADEH, M.; UNDUCHE, F. Signature-based multi-modelling and multi-objective calibration of hydrologic models: Application in flood forecasting for Canadian Prairies. **Journal of Hydrology**, v. 588, p. 125095, 1 set. 2020.

SAMANIEGO, L.; KUMAR, R.; ATTINGER, S. Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. **Water Resources Research**, v. 46, n. 5, 2010.

SANKARASUBRAMANIAN, A.; VOGEL, R. M.; LIMBRUNNER, J. F. Climate elasticity of streamflow in the United States. **Water Resources Research**, v. 37, n. 6, p. 1771–1781, jun. 2001.

SAREEN, K. et al. An integrated decomposition algorithm based bidirectional LSTM neural network approach for predicting ocean wave height and ocean wave energy. **Ocean Engineering**, v. 281, p. 114852, 2023.

SAVENIJE, H. H. G.; HRACHOWITZ, M. HESS Opinions Catchments as meta-organisms – a new blueprint for hydrological modelling. **Hydrology and Earth System Sciences**, v. 21, n. 2, p. 1107–1116, 22 fev. 2017.

SCHMIDHUBER, J. Deep Learning in Neural Networks: An Overview. **Neural Networks**, v. 61, p. 85–117, jan. 2015.

SCHMIDT, L. et al. Challenges in Applying Machine Learning Models for Hydrological Inference: A Case Study for Flooding Events Across Germany. **Water Resources Research**, v. 56, n. 5, p. e2019WR025924, 2020.

SEIBERT, J.; BERGSTRÖM, S. A retrospective on hydrological catchment modelling based on half a century with the HBV model. **Hydrology and Earth System Sciences**, v. 26, n. 5, p. 1371–1388, 14 mar. 2022.

SEIBERT, J.; MCDONNELL, J. J. On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration. **Water Resources Research**, v. 38, n. 11, p. 23-1-23–14, 2002.

SEILLER, G.; ANCTIL, F.; ROY, R. Design and experimentation of an empirical multistructure framework for accurate, sharp and reliable hydrological ensembles. **Journal of Hydrology**, v. 552, p. 313–340, set. 2017.

SEMENOVA, O.; BEVEN, K. Barriers to progress in distributed hydrological modelling. **Hydrological Processes**, v. 29, n. 8, p. 2074–2078, 15 abr. 2015.

SEONG, C.; HER, Y.; BENHAM, B. L. Automatic Calibration Tool for Hydrologic Simulation Program-FORTRAN Using a Shuffled Complex Evolution Algorithm. **Water**, v. 7, n. 2, p. 503–527, fev. 2015.

SHAFII, M.; TOLSON, B. A. Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives. **Water Resources Research**, v. 51, n. 5, p. 3796–3814, 2015.

SHAPLEY, L. S.; ROTH, A. E. (EDS.). **The Shapley value: essays in honor of Lloyd S. Shapley**. Cambridge [Cambridgeshire]; New York: Cambridge University Press, 1988.

SHEN, C. A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. **Water Resources Research**, v. 54, n. 11, p. 8558–8593, 2018.

SHEN, H.; TOLSON, B. A.; MAI, J. Time to Update the Split-Sample Approach in Hydrological Model Calibration. **Water Resources Research**, v. 58, n. 3, p. e2021WR031523, 2022.

SINGH, N. M.; DEVI, T. T. Regionalization methods in ungauged catchments for flow prediction: review and its recent developments. **Arabian Journal of Geosciences**, v. 15, n. 11, p. 1019, 19 maio 2022.

SINGH, V. P. **Computer Models of Watershed Hydrology**. Reprinted Edition ed. Highlands Ranch, Colorado: Water Resources Pubns, 2012.

SIVAPALAN, M. Pattern, Process and Function: Elements of a Unified Theory of Hydrology at the Catchment Scale. Em: **Encyclopedia of Hydrological Sciences**. [s.l.] John Wiley & Sons, Ltd, 2006.

SMITH, A.; SAMPSON, C.; BATES, P. Regional flood frequency analysis at the global scale. **Water Resources Research**, v. 51, n. 1, p. 539–553, 2015.

SMITH, K. A. et al. A multi-objective ensemble approach to hydrological modelling in the UK: an application to historic drought reconstruction. **Hydrology and Earth System Sciences**, v. 23, n. 8, p. 3247–3268, 8 ago. 2019.

SMITH, M. B. et al. Results of the DMIP 2 Oklahoma experiments. **Journal of Hydrology**, The Distributed Model Intercomparison Project (DMIP) - Phase 2 Experiments in the Oklahoma Region, USA. v. 418–419, p. 17–48, 8 fev. 2012.

SPIELER, D. et al. Automatic Model Structure Identification for Conceptual Hydrologic Models. **Water Resources Research**, v. 56, n. 9, p. e2019WR027009, 2020.

STORN, R.; PRICE, K. Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. **Journal of Global Optimization**, v. 11, n. 4, p. 341–359, 1 dez. 1997.

ŠTRUMBELJ, E.; KONONENKO, I. Explaining prediction models and individual predictions with feature contributions. **Knowledge and Information Systems**, v. 41, n. 3, p. 647–665, 1 dez. 2014.

SWATHI, V. et al. Automatic calibration of SWMM using NSGA-III and the effects of delineation scale on an urban catchment. **Journal of Hydroinformatics**, v. 21, n. 5, p. 781–797, 18 jul. 2019.

SZCZEPANEK, R. Daily Streamflow Forecasting in Mountainous Catchment Using XGBoost, LightGBM and CatBoost. **Hydrology**, v. 9, n. 12, p. 226, dez. 2022.

TALUKDAR, S. et al. Trend analysis and forecasting of streamflow using random forest in the Punarbhaha River basin. **Environmental Monitoring and Assessment**, v. 195, n. 1, p. 1–27, 1 jan. 2023.

TAO, Y. et al. A Deep Neural Network Modeling Framework to Reduce Bias in Satellite Precipitation Products. **Journal of Hydrometeorology**, v. 17, n. 3, p. 931–945, 1 mar. 2016.

TAREK, M. et al. Daily streamflow prediction in ungauged basins: an analysis of common regionalization methods over the African continent. **Hydrological Sciences Journal**, v. 66, n. 11, p. 1695–1711, 18 ago. 2021.

TEGEGNE, G.; KIM, Y.-O. Modelling ungauged catchments using the catchment runoff response similarity. **Journal of Hydrology**, v. 564, p. 452–466, 1 set. 2018.

THORNDIKE, R. L. Who belongs in the family? **Psychometrika**, v. 18, n. 4, p. 267–276, 1 dez. 1953.

TUO, Y. et al. A multi-objective approach to improve SWAT model calibration in alpine catchments. **Journal of Hydrology**, v. 559, p. 347–360, 1 abr. 2018.

TYRALIS, H. et al. Explanation and Probabilistic Prediction of Hydrological Signatures with Statistical Boosting Algorithms. **Remote Sensing 2021, Vol. 13, Page 333**, v. 13, n. 3, p. 333, 20 jan. 2021.

VEGA GARCÍA, M.; AZNARTE, J. L. Shapley additive explanations for NO2 forecasting. **Ecological Informatics**, v. 56, p. 101039, 1 mar. 2020.

VRUGT, J. A. et al. Effective and efficient algorithm for multiobjective optimization of hydrologic models. **Water Resources Research**, v. 39, n. 8, 2003.

WAGENER, T. et al. Catchment Classification and Hydrologic Similarity. **Geography Compass**, v. 1, n. 4, p. 901–931, 1 jul. 2007.

WANG, H.; CAO, L.; FENG, R. Hydrological Similarity-Based Parameter Regionalization under Different Climate and Underlying Surfaces in Ungauged Basins. **Water**, v. 13, n. 18, p. 2508, jan. 2021.

WANG, Q. et al. Comparison of Multiobjective Optimization Methods Applied to Urban Drainage Adaptation Problems. **Journal of Water Resources Planning and Management**, v. 144, n. 11, p. 04018070, nov. 2018.

- WANG, S. et al. Analysis of runoff generation driving factors based on hydrological model and interpretable machine learning method. **Journal of Hydrology: Regional Studies**, v. 42, p. 101139, 1 ago. 2022.
- WESTERBERG, I. K. et al. Regional water balance modelling using flow-duration curves with observational uncertainties. **Hydrology and Earth System Sciences**, v. 18, n. 8, p. 2993–3013, 14 ago. 2014.
- WESTERBERG, I. K. et al. Uncertainty in hydrological signatures for gauged and ungauged catchments. **Water Resources Research**, v. 52, n. 3, p. 1847–1865, 2016.
- WESTERBERG, I. K.; MCMILLAN, H. K. Uncertainty in hydrological signatures. **Hydrology and Earth System Sciences**, v. 19, n. 9, p. 3951–3968, 24 set. 2015.
- WHEATER, H. S.; PEACH, D.; BINLEY, A. Characterising groundwater-dominated lowland catchments: the UK Lowland Catchment Research Programme (LOCAR). **Hydrology and Earth System Sciences**, v. 11, n. 1, p. 108–124, 17 jan. 2007.
- WHITE, B. W.; ROSENBLATT, F. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. **The American Journal of Psychology**, v. 76, n. 4, p. 705, dez. 1963.
- WILKS, D. S. **Statistical Methods in the Atmospheric Sciences**. 4th ed. edição ed. Cambridge: Elsevier, 2019.
- WILLEMIN, J. H. Hack's Law: Sinuosity, convexity, elongation. **Water Resources Research**, v. 36, n. 11, p. 3365–3374, 2000.
- WILSON, A. C. et al. **The Marginal Value of Adaptive Gradient Methods in Machine Learning**. arXiv, , 21 maio 2018. Disponível em: <<http://arxiv.org/abs/1705.08292>>. Acesso em: 16 jul. 2023
- WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining: Practical Machine Learning Tools and Techniques**. 3rd edition ed. Burlington, MA: Morgan Kaufmann, 2011.
- WORLAND, SCOTT. C. et al. Copula Theory as a Generalized Framework for Flow-Duration Curve Based Streamflow Estimates in Ungaged and Partially Gaged Catchments. **Water Resources Research**, v. 55, n. 11, p. 9378–9397, 2019.
- WU, H. et al. Runoff Modeling in Ungauged Catchments Using Machine Learning Algorithm-Based Model Parameters Regionalization Methodology. **Engineering**, p. S2095809922000613, fev. 2022.
- WU, Z. et al. On Scalability of Association-rule-based Recommendation: A Unified Distributed-computing Framework. **ACM Transactions on the Web**, v. 14, n. 3, p. 13:1-13:21, 21 jun. 2020.
- YAEGER, M. et al. Exploring the physical controls of regional patterns of flow duration curves – Part 4: A synthesis of empirical analysis, process modeling and catchment classification. **Hydrology and Earth System Sciences**, v. 16, n. 11, p. 4483–4498, 26 nov. 2012.

YAN, X. et al. Deep learning for daily potential evapotranspiration using a HS-LSTM approach. **Atmospheric Research**, v. 292, p. 106856, 2023.

YANG, J.; CASTELLI, F.; CHEN, Y. Multiobjective sensitivity analysis and optimization of distributed hydrologic model MOBIDIC. **Hydrology and Earth System Sciences**, v. 18, n. 10, p. 4101–4112, 15 out. 2014.

YANG, X. et al. Dependence of regionalization methods on the complexity of hydrological models in multiple climatic regions. **Journal of Hydrology**, v. 582, p. 124357, 1 mar. 2020.

YANG, X. et al. Comparison of the Calibrated Objective Functions for Low Flow Simulation in a Semi-Arid Catchment. **Water**, v. 14, n. 17, p. 2591, jan. 2022.

YANNIBELLI, V. et al. A Comparative Analysis of NSGA-II and NSGA-III for Autoscaling Parameter Sweep Experiments in the Cloud. **Scientific Programming**, v. 2020, p. e4653204, 28 ago. 2020.

YAPO, P. O.; GUPTA, H. V.; SOROOSHIAN, S. Multi-objective global optimization for hydrologic models. **Journal of Hydrology**, v. 204, n. 1–4, p. 83–97, jan. 1998.

YOUNG, P. Top-down and data-based mechanistic modeling of rainfall-flow dynamics at the catchment scale. **Hydrological Processes**, v. 17, p. 2195–2217, 15 ago. 2003.

YOUNG, P. C.; BEVEN, K. J. Data-based mechanistic modelling and the rainfall-flow non-linearity. **Environmetrics (United Kingdom)**, 1994.

ZELENY, M. Multiple Criteria Decision Making (MCDM): From Paradigm Lost to Paradigm Regained? **Journal of Multi-Criteria Decision Analysis**, v. 18, n. 1–2, p. 77–89, 2011.

ZHANG, H. et al. Dynamic Streamflow Simulation via Online Gradient-Boosted Regression Tree. **Journal of Hydrologic Engineering**, v. 24, n. 10, p. 04019041, 1 out. 2019.

ZHANG, J. et al. Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas. **Journal of Hydrology**, v. 561, p. 918–929, 2018a.

ZHANG, L. et al. Air quality predictions with a semi-supervised bidirectional LSTM neural network. **Atmospheric Pollution Research**, v. 12, n. 1, p. 328–339, 2021.

ZHANG, R. et al. Can multi-objective calibration of streamflow guarantee better hydrological model accuracy? **Journal of Hydroinformatics**, v. 20, n. 3, p. 687–698, 27 jan. 2018b.

ZHANG, Y. et al. Predicting hydrological signatures in ungauged catchments using spatial interpolation, index model, and rainfall–runoff modelling. **Journal of Hydrology**, v. 517, p. 936–948, 19 set. 2014.

ZHANG, Y. et al. Multi-metric calibration of hydrological model to capture overall flow regimes. **Journal of Hydrology**, v. 539, p. 525–538, 1 ago. 2016.

ZHANG, Y. et al. Predicting Runoff Signatures Using Regression and Hydrological Modeling Approaches. **Water Resources Research**, v. 54, n. 10, p. 7859–7878, 1 out. 2018c.

ZHANG, Y.; ZHAO, Z.; ZHENG, J. CatBoost: A new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China. **Journal of Hydrology**, v. 588, p. 125087, set. 2020.

ZHONG, M. et al. Using the Apriori Algorithm and Copula Function for the Bivariate Analysis of Flash Flood Risk. **Water**, v. 12, n. 8, p. 2223, ago. 2020.

ZITZLER, E.; THIELE, L. Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. **IEEE Transactions on Evolutionary Computation**, v. 3, n. 4, p. 257–271, nov. 1999.