Pré-processamento de dados para Modelos Hidrológicos com o algoritmo k-Medoids: O caso do Rio Pomba

Heitor G. B. Magacho[1], **Wagner R. Telles**[1], **Marcos Bedo**[2]

[1]Instituto do Noroeste Fluminense – Universidade Federal Fluminense (INFES/UFF)
S. A. Pádua/RJ – Brasil

[2]Faculdade de Medicina de Ribeirão Preto – Universidade de São Paulo (FMRP/USP)
Ribeirão Preto/SP – Brasil

{heitormagacho, wtelles, marcosbedo}@id.uff.br

Resumo. *Esse estudo propõe a aplicação do algoritmo k-Medoids como etapa de pré-processamento para reduzir a cardinalidade de conjuntos de dados topológicos em simulações hidrológicas do mundo real. Assim, é importante que os medóides identificados sejam representativos para o problema simulado e que os dados reduzidos não prejudiquem a qualidade dos Modelos Hidrológicos executados na sequência de processamento. Em particular, esse trabalho investiga o uso do pré-processamento junto a dois tipos de Modelos Digitais de Terreno para bacias hidrográficas no software MOHID Studio. Como estudo de caso, foi avaliada a simulação com os dados originais vs. pré-processados de um Modelo Hidrológico para o Rio Pomba, localizado em uma região de enchentes urbanas recorrentes. Os resultados indicam que o pré-processamento reduziu a entrada da simulação em até 90%, mantendo a qualidade do Modelo Hidrológico com relação a (i) construção da rede de drenagem e (ii) ao seu perfil de nível d'água. Os resultados também mostram que os medóides escolhidos formam grupos compactos e espaçados de acordo com os índices de Calinski-Harabasz e Davies-Bouldin, o que permitiu reduzir o custo das simulações.*

Abstract. *This study discusses the k-Medoids algorithm as a pre-processing step to reduce the cardinality of topological data in real-world hydrologic simulations. The hypothesis is that reduced data (medoids) are representative enough to not compromise the quality of the underlying Hydrologic Model. In particular, we coupled the data pre-processing strategy with two Digital Terrain Models for hydrologic basins in the MOHID Studio software to investigate the performance of a Hydrologic Model designed for the Pomba River (in an urban region with recurrent floodings). The results show the selected medoids represent compacted and separated clusters according to the Calinski-Harabasz and Davies-Bouldin indexes. Moreover, the results indicate data pre-processing has reduced the simulation input by up to 90% while maintaining the Model quality in terms of (i) water level profiles and (ii) drainage network levels.*

1. Introdução

Modelos Hidrológicos são capazes de simular o comportamento de precipitações e inundações para diversos tipos de terreno, permitindo que gestores públicos e privados atuem de forma preventiva para mitigar potenciais efeitos desses fenômenos natu-

rais [Sales et al. 2021, Oliveira et al. 2022]. A execução desses modelos é computacionalmente intensiva em função dos dados geográficos e topológicos ⟨latitude, longitude, altitude⟩ de entrada que são processados em múltiplos passos de tempo (interações). Além disso, migrar dados topológicos entre servidores e sistemas de arquivos diferentes para reproduzir a simulação também é um desafio, pois os arquivos de entrada são tipicamente da ordem de bilhões de pontos. Portanto, pré-processar esses dados topológicos de entrada sem que haja perda expressiva da qualidade do Modelo Hidrológico tem grande potencial para melhorar o desempenho dessas simulações e permitir a sua reprodutibilidade.

O método de agrupamento k-Medoids é capaz de identificar os k elementos mais centrais de um grupo (*cluster*) de pontos topológicos (denominados *medóides*), gerando um conjunto k-reduzido [Kaufman and Rousseeuw 1990]. A representatividade de cada elemento é explicitamente quantificada por meio das fronteiras de Voronoi induzidas entre os *medóides* selecionados nesses espaço tridimensional, sendo que os elementos descartados encontram-se agrupados na região delimitada pelo *medóide* e suas respectivas fronteiras de Voronoi. Esse trabalho investiga a *hipótese* de que os dados reduzidos (*medóides*) sejam também representativos o suficiente para não comprometer a qualidade do Modelo Hidrológico acoplado na sequência de processamento. Como estudo de caso, é examinado um Modelo Hidrológico real construído para um trecho do Rio Pomba no noroeste fluminense. Os resultados experimentais mostraram que os *medóides* escolhidos permitiram reduzir eficientemente os dados topológicos em até 90%.

O restante desse estudo é organizado da seguinte forma. A Seção 2 introduz os conceitos preliminares, enquanto a Seção 3 descreve os materiais e métodos. A Seção 4 apresenta a avaliação realizada e a Seção 5 discute os resultados encontrados.

2. Preliminares

Agrupamento particional. Um agrupamento particional *crisp* é uma função θ que divide os elementos do conjunto de dados $\mathcal{O} \subseteq \mathbb{O}$ em um conjunto de k *clusters* $\mathcal{C} = \{C_1, \ldots, C_k\}$ disjuntos, a função $\theta : \mathcal{O} \rightarrow \mathcal{C}$ atribui cada elemento $o \in \mathcal{O}$ para um único *cluster* $C \in \mathcal{C}$ de forma que $\cup_{C_i \in \mathcal{C}} C_i = \mathcal{O}$ e $\cap_{C_i \in \mathcal{C}} C_i = \emptyset$.

Um agrupamento θ é induzido através de comparações por distância. Uma função de distância métrica $\delta, \delta : \mathbb{O} \times \mathbb{O} \rightarrow \mathbb{R}_+$ opera sobre um domínio $\mathbb{O}$ e compara quaisquer dois elementos $o_i, o_j \in \mathcal{O} \subseteq \mathbb{O}$ tal que quanto maior o valor da distância, mais dissimilares são os dois elementos. Um agrupamento particional *crisp* que se beneficia dessa premissa é a abordagem k-Medoids resolvido pela heurística *Partition Around Medoids*, simplificada no Algoritmo 1. O objetivo desse agrupamento é encontrar um conjunto k de elementos "centrais" (*medóides*) cuja distância média aos demais elementos do *cluster* seja mínima, um problema de solução NP-difícil [Schubert and Rousseeuw 2019]. A heurística inicia-se com um conjunto aleatório de *medóides* e a cada iteração escolhe um novo conjunto de elementos que minimiza a média das distâncias, encerrando-se quando a solução converge ou quando um número máximo de iterações (Ω) é ultrapassado.

Coesão, Separação, Calinski-Harabasz e Davies-Bouldin. Além das medidas objetivas derivadas do uso do método k-Medoids como pré-processamento, a qualidade da partição gerada também é quantificada por indicadores internos. A Coesão mede a soma da compacidade dos *clusters*, enquanto a Separação mede a soma das distâncias entre os *medóides* e o "centro" da partição. Dada uma partição $\mathcal{C}$ so-

Entrada: Conjunto de dados $\mathcal{O}$, #*clusters* k, iterações máximas Ω
Saída: Conjunto escolhido de *medoides* $\mathcal{C}$

$\mathcal{C} = \mathcal{C}' = \{c_1, \ldots, c_k\} \leftarrow$ EscolhaAleatória($\mathcal{O}$); `/* Sem reposição */`
for $x \leftarrow [1, \ldots, \Omega] \wedge (\mathcal{C}' \neq \mathcal{C} \vee x = 1)$ **do**
  $\mathcal{C} \leftarrow \mathcal{C}'$; $C' \leftarrow \emptyset$;
  for $c_k \in C$ **do**
    $T \leftarrow \{o_i \mid o_i \in \mathcal{O}, \delta(o_i, c_k) \leq \delta(o_i, c_m)\ \forall c_m \in \{\mathcal{C} \setminus \{c_k\}\}$;
    $c'_k \leftarrow o_i \mid o_i \in T, \min\left(\sum_{o_j \in T} \delta(o_i, o_j)/|T|\right)$; $\mathcal{C}' \leftarrow \mathcal{C}' \cup \{c'_k\}$;
  end
end
Return $\mathcal{C}, \mathcal{C} \leftarrow \mathcal{C}'$;

Algoritmo 1: Solução heurística para minimizar a média de distâncias aos *medóides*.

bre $\mathcal{O}$, esses dois indicadores são unificados em medidas individuais, tais como os índices de Calinski-Harabasz (CH) e Davies–Bouldin (DB). O primeiro é dado por $\text{CH}(\mathcal{C}) = \left((\sum_{C_i \in \mathcal{C}} |C_i| \delta(c_i, \bar{\mu})/|\mathcal{C}| - 1)/\sum_{C_i \in \mathcal{C}} \sum_{o_x \in C_i} \delta(o_x, c_i)/(|\mathcal{O}| - |\mathcal{C}|)\right)$, onde c_i é o *medóide* de C_i e $\bar{\mu}$ é o *medóide* de $\mathcal{O}$. Já o segundo é calculado como $\text{DB}(\mathcal{C}) = 1/k \cdot \sum_{C_i, C_j \in \mathcal{C}} \max_{C_i \neq C_j} (\phi_i + \phi_j)/\delta(c_i, c_j)$, onde $\phi_i = 1/2 \cdot |C_i| \sum_{o_x, o_y \in C_i} \delta(o_x, o_y)$.

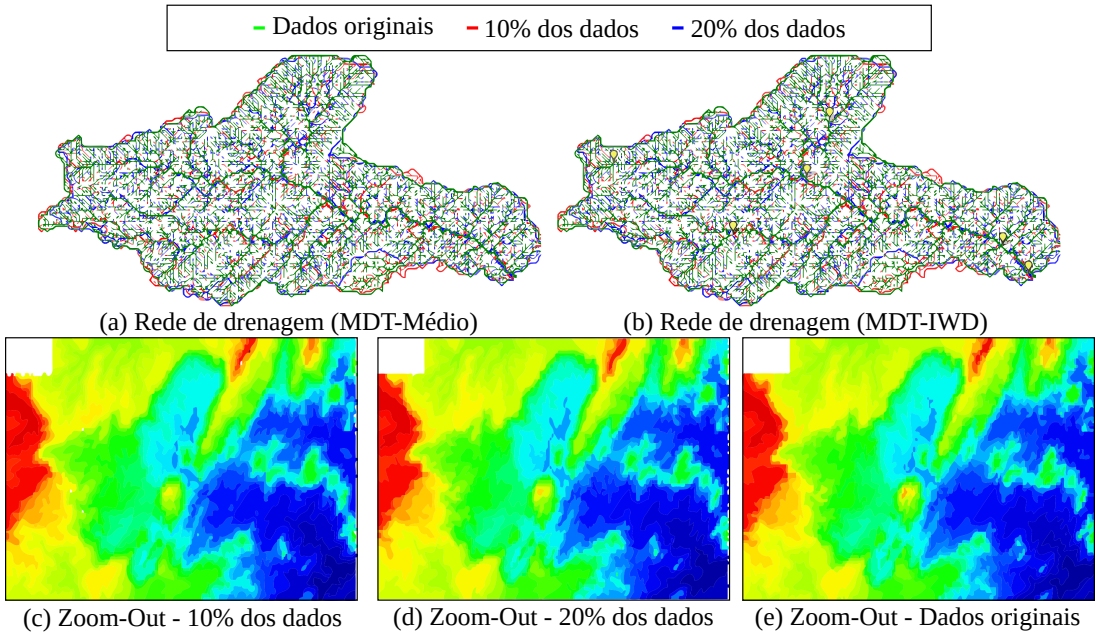
3. Materiais e Métodos

A proposta deste estudo consiste em aplicar o Algoritmo 1 para selecionar k elementos representativos dentre um conjunto de dados topológicos que será alimentado em um Modelo Hidrológico de previsão de inundações para o trecho urbano do Rio Pomba, que nasce em Barbacena/MG e deságua no Rio Paraíba do Sul em Itaocara/RJ – Brasil.

Modelo Hidrológico: Processos e Implementação. Os processos hidrológicos são implementados como um *workflow* intensivo em dados que calcula o comportamento da rede de drenagem de três modos: *(i)* unidimensional na direção do canal pelas equações de Saint-Venant de forma completa, *(ii)* bidimensional para o escoamento horizontal de acordo com o modelo de Saint-Venant na forma de Equação de Onda de Difusão e *(iii)* tridimensional para o escoamento de solo, seguindo a equação de Richards [Ross 1990]. As simulações do *workflow* são executadas na plataforma MOHID Studio com a ferramenta numérica MOHID Land [Braunschweig et al. 2012]. Além dos dados topológicos, são usados como entrada dados oficiais de séries de precipitação para o Rio Pomba[1].

Coleta dos dados topológicos. Foram usados na avaliação experimental os dados medidos pela missão espacial SRTM com a nave *Endeavour*, gerenciada pela agência NASA/EUA e com dados revisados e publicados pela agência INPE/Brasil no projeto TOPODATA[2]. Os dados da região estudada (56^oS e 60^oN) foram medidos em fevereiro de 2000, com quadrantes espaçados em 90 metros. Em particular, foram coletados dados dos quadrantes 20_435, 21_435 e 21_45, correspondentes ao trecho urbano do Rio Pomba, cada um com a cardinalidade de 2.343.600 instâncias e no mesmo domínio $\mathbb{O} = \mathbb{R}^3$.

[1]Disponível em: `http://www.snirh.gov.br/hidroweb/`
[2]Disponível em: `http://www.dsr.inpe.br/topodata/`



(a) Rede de drenagem (MDT-Médio) (b) Rede de drenagem (MDT-IWD)

(c) Zoom-Out - 10% dos dados (d) Zoom-Out - 20% dos dados (e) Zoom-Out - Dados originais

Figura 1. (a–b) Malhas MDT-Médio e IWD. (c–e) Mapas *zoom-out* para MDT-Médio.

Modelos Digitais de Terreno. Para as simulações do Modelo Hidrológico foram considerados duas construções de Modelos Digitais de Terreno (MDT) construídos sobre os dados topológicos: *(i)* uma utilizando uma malha composta pelo valor médio do quadrante (denominado MDT-Médio) e *(ii)* outra utilizando a distância inversa (*Inverse Weighted Distance* – IWD) entre os pontos do quadrante (denominado MDT-IWD).

Infraestrutura de Teste. Além da plataforma MOHID Studio, utilizou-se o *framework* R v3.4.4 com o pacote "cluster" para a implementação da solução k-Medoids, além do programa SciLab para automatizar as medidas de erro e comportamento derivados do Modelo Hidrológico. Todos os testes foram realizados em uma máquina com processador Intel® i5-6200U dual-core 2.30GHz, 8GB de RAM, rodando Windows 10 64-bits.

4. Avaliação Experimental

Foram comparados os dados topológicos originais de cada quadrante contra dois conjuntos reduzidos: *(i)* o primeiro com 10% do total de pontos (234.360) e *(ii)* o segundo com 20% (468.720). As Figuras 1(a–b) mostram as redes de drenagem para os dois tipos de MDT examinados, enquanto as Figuras 1(c–d) ilustram a representação do terreno para os conjuntos reduzidos na comparação com os dados topológicos originais do ponto de vista de seis estações de controle igualmente espaçadas. Poucas diferenças são observadas nessa comparação, especialmente fora dos limites (bordas) dos MDTs.

Qualidade da redução (avaliação interna). A Figura 2 detalha a qualidade dos agrupamentos k-Medoids com $\delta = L_2$ (Euclidiana). Os resultados indicam que os *medóides* geram grupos uniformemente coesos e separados, corroborando a hipótese de que é possível encontrar elementos representativos mesmo para baixos valores de k ($<1\%$). Além disso, os resultados mostram que mais *medóides* representam melhor o conjunto original.

Simulação dos níveis d'água (avaliação externa). A Figura 3 apresenta o resultado da simulação para os níveis d'água em cada uma das seis estações de monitoramento nos

Índice\Redução	0,1%	0,2%	0,3%	0,4%	0,5%	0,6%	0,7%	0,8%	0,9%	1,0%
DB	0,53	0,45	0,40	0,40	0,41	0,41	0,43	0,44	0,47	0,48
CH (.10^7)	2,25	4,77	7,00	10,94	15,02	19,18	23,81	28,24	32,06	35,49

Figura 2. Medidas internas Calinski-Harabasz (CH) e Davies–Bouldin (DB).

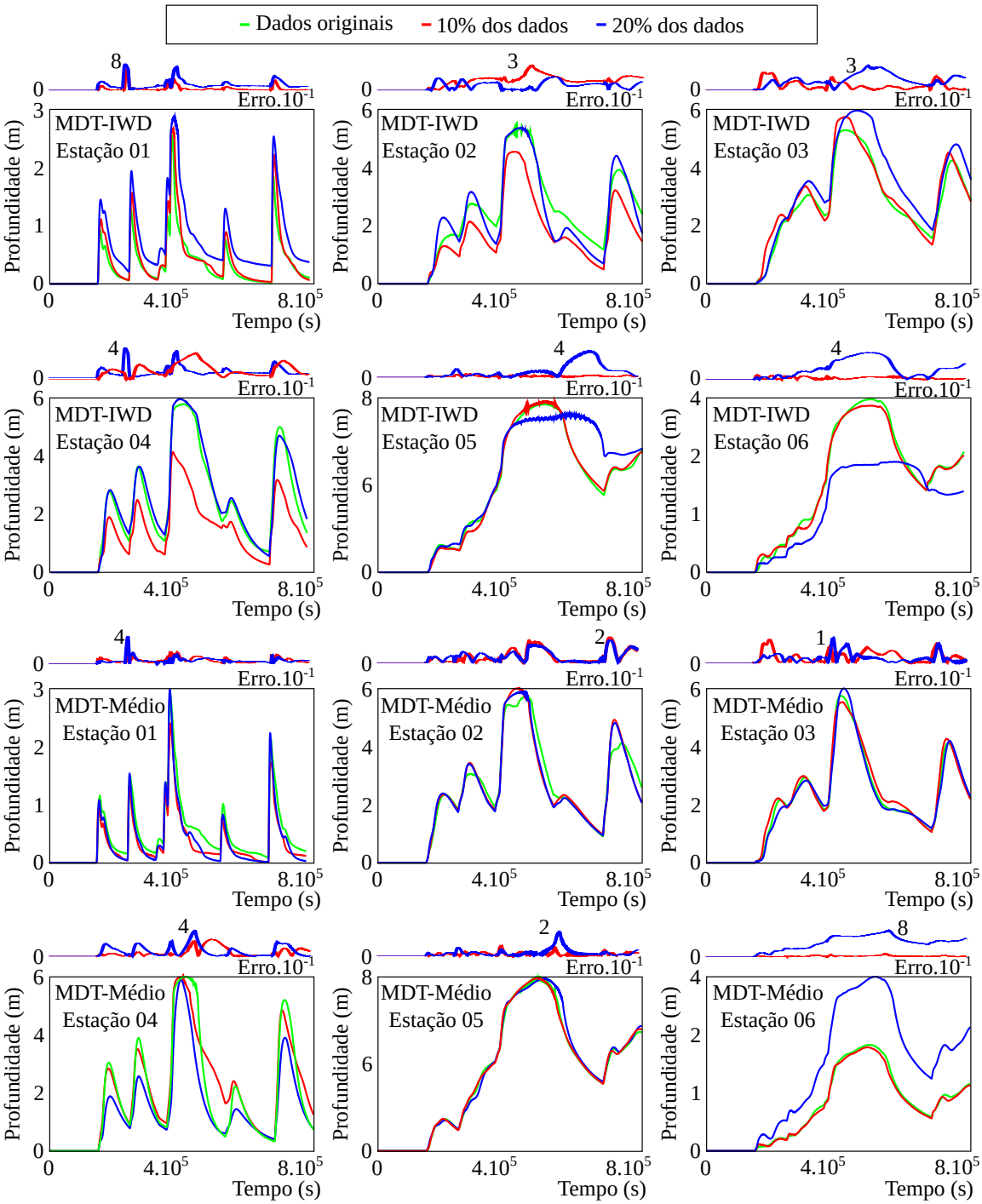


Figura 3. Erro relativo e profundidade: Dados originais *vs.* pré-processados.

dois MDTs avaliados. Nas simulações, um período inicial de erro nulo foi considerado para o "aquecimento" do Modelo Hidrológico (um número fixo de iterações na Figura 3). Os resultados mostram que a profundidade obtida pelas simulações com os dados pré-processados ao longo do tempo foi próxima ao valor obtido com os dados originais. Foram avaliados os perfis do nível d'água na rede de drenagem com base na profundidade do rio considerando o erro relativo $\in [0, 1]$ entre os dados reduzidos e o conjunto topológico original. A redução para 10% obteve um erro relativo de $6{,}85 \cdot 10^{-2} \pm 5{,}34 \cdot 10^{-2}$ ($1{,}01 \cdot 10^{-1} \pm 7{,}22 \cdot 10^{-2}$ para o MDT-IWD e $3{,}35 \cdot 10^{-2} \pm 3{,}46 \cdot 10^{-2}$ para o MDT-Médio) na comparação com a totalidade dos dados. Já a redução a 20% alcançou um erro relativo de $8 \cdot 10^{-2} \pm 2{,}99 \cdot 10^{-2}$ ($9{,}77 \cdot 10^{-2} \pm 8{,}73 \cdot 10^{-2}$ para o MDT-IWD e $7{,}28 \cdot 10^{-2} \pm 5{,}98 \cdot 10^{-2}$ para o MDT-Médio) comparado às simulações com os dados topológicos originais. Esses resultados indicam que não há, necessariamente, um *trade-off* entre a quantidade de pontos e a qualidade do modelo, *i.e.*, poucos dados de entrada definem a maior parte da saída da simulação e é necessário adicionar muitos pontos de entrada para reduzir apenas uma fração do erro relativo. Nesse sentido, os resultados sugerem que é possível manter a qualidade das simulações com apenas 10% de elementos *medóides* do conjunto original.

5. Conclusão e Trabalhos Futuros

Esse estudo discutiu o algoritmo k-Medoids como pré-processamento para Modelos Hidrológicos, analisando como estudo de caso um trecho do Rio Pomba. Os resultados mostraram que os *medóides* escolhidos representam grupos coesos e bem separados e que o pré-processamento permitiu reduzir os dados de entrada em até 90% mantendo o desempenho das simulações considerando a *(i)* construção da rede de drenagem e *(ii)* os seus perfis de niveis d'água. Como trabalho futuro, pretende-se investigar o desempenho de outros métodos de agrupamento e redes de drenagem com quilômetros de extensão.

Agradecimentos. Marcos Bedo está em afastamento do INFES/UFF na FMRP/USP (G. #21/06564-0 – Fundação de Amparo à Pesquisa do Estado de São Paulo). Esse estudo foi apoiado financeiramente pela Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (G. E47/2021-SEI260003/016517/2021-R210.107/2022 – Wagner R. Telles).

Referências

Braunschweig, F., Fernandes, L., and Lourenço, F. (2012). *MOHID Studio User Guide*.

Kaufman, L. and Rousseeuw, P. (1990). Partitioning around medoids (Program Pam). *Finding groups in data: An introduction to cluster analysis*, 344:68–125.

Oliveira, T., Simionesei, L., Gonçalves, M., and Neves, R. (2022). Modeling Streamflow at the Iberian Peninsula Scale Using MOHID-Land: Challenges from a Coarse Scale Approach. *Water*, 14(7):1013.

Ross, P. J. (1990). Efficient numerical methods for infiltration using Richards' equation. *Water Resources Res.*, 26(2):279–290.

Sales, D., Lugon Jr., J., Oliveira, V., and Silva-Neto, A. (2021). Rainfall input from WRF-ARW atmospheric model coupled with MOHID Land hydrological model for flow simulation in the Paraíba do Sul river. *J. Urban & Env. Eng.*, 15(2).

Schubert, E. and Rousseeuw, P. J. (2019). Faster k-Medoids clustering: improving the PAM, CLARA, and CLARANS algorithms. In *SISAP*, pages 171–187. Springer.