Abordagem Multivariada para Imputação de Temperatura Média na Região Amazônica baseada em Estações Meteorológicas e Dados de Reanálise ERA5-Land

Jean Arthur Costa Dias[1], André Vinicius N. Alves[1], Waldemiro J. A. G. Negreiros[1], Rodrigo Gonçalves Novais[4], Fernando Luiz Cyrino Oliveira[4], Leonardo de O. Tamasauskas[1], Pedro H. do V. Guimarães[1], Karla Figueiredo[3], Gabriel B. Costa[2], Marivan S. Gomes[5], Marcos César da Rocha Seruffo[1,2]

[1]Laboratório de Pesquisa Operacional (LPO)
Universidade Federal do Pará (UFPA), Belém-PA

[2]Programa de Pós-Graduação em Estudos Antrópicos na Amazônia
Universidade Federal do Pará (UFPA), Castanhal-PA

[3]Department of Informatics and Computer
Science, Institute of Mathematics and Statistics,
Rio de Janeiro State University (UERJ)

[4]Departamento de Engenharia Industrial
Pontíficia Universidade Católica do Rio de Janeiro (PUC-RIO)

[5]Núcleo de Robótica e Automação
Escola Superior de Tecnologia
Universidade do Estado do Amazonas (UEA), Manaus-AM

jean.dias@itec.ufpa.br,andre.neves.alves@itec.ufpa.br,
msgomes@uea.edu.br, cyrino@puc-rio.br, novaisrodrigo@outlook.com
leonardo.tamasauskas@icen.ufpa.br, pedro.guimaraes@itec.ufpa.br
karlafigueiredo@ime.uerj.br, Waldemiro.negreiros@ifpa.edu.br
gabrielbritocosta@gmail.com, seruffo@ufpa.br

Abstract. *This study investigates the imputation of missing values in mean air temperature time series in the state of Pará, Brazil, by integrating observational data from INMET and the ERA5-Land reanalysis. A multivariate non-linear approach based on the MissForest algorithm was developed and compared with models using only station data and with linear interpolation. The evaluation employed synthetic gaps together with the RMSE, MAE, and MAE in quantile 99 metrics. The results indicate that the inclusion of ERA5-Land reduced RMSE to 0.664–0.744 °C, MAE to 0.509–0.569 °C, and ΔP_{99} to 0.909–1.013 °C, while increasing model stability and outperforming the other strategies, thus supporting the methodological consistency of the approach for tropical climate time series.*

1. Introdução

A vastidão territorial da Amazônia e as barreiras logísticas de acesso a áreas remotas consolidam um obstáculo crítico à produção sistemática de conhecimento, resultando em um persistente ”vazio informacional”na região. Essa escassez de dados estruturados com-

promete severamente a representatividade das análises científicas, o que acaba por gerar paradigmas falhos e prescrições políticas inadequadas para a realidade dos trópicos [Metcalfe et al. 2025]. Sob essa ótica, a ausência de monitoramento sistemático e de dados orientadores culmina em políticas públicas ineficazes, incapazes de mitigar os impactos do desenvolvimento mal planejado e das profundas disparidades socioeconômicas locais [Hoinaski et al. 2024].

Nesse cenário, a precisão na recuperação de séries térmicas é fundamental para o cumprimento do ODS 13 (Ação Contra a Mudança Global do Clima) e do ODS 3 (Saúde e Bem-Estar), visto que a análise de tendências climáticas e picos de calor é a base para a formulação de políticas de saúde pública eficazes [United Nations General Assembly 2015]. Entretanto, a carência de evidências robustas dificulta a compreensão e o equacionamento dos trade-offs entre conservação e desenvolvimento, limitando a capacidade de gestores em avaliar arranjos de proteção frente a usos alternativos da terra [den Braber et al. 2024]. Tal cenário impede não apenas o monitoramento efetivo do desmatamento e da perda de biodiversidade, mas também a formulação de estratégias capazes de responder às complexidades locais e promover o desenvolvimento sustentável [Garrett et al. 2024]. Consequentemente, a persistência dessa lacuna informacional, que obstaculiza a integração de dados essenciais para a tomada de decisão [Han et al. 2024], constitui o elemento central que justifica a realização do presente estudo.

Reiterando o supracitado, a revisão de [Alejo-Sanchez et al. 2025] contextualiza o preenchimento de valores faltantes, ou imputação, e sua crescente relevância na literatura de séries temporais climáticas, dada a recorrência de valores ausentes nas principais fontes de dados meteorológicos, incluindo estações de observação e sensoriamento remoto. A revisão aponta um aumento na utilização de métodos estatísticos e de aprendizado de máquina para esse fim. No entanto, apesar de fornecer um panorama geral das metodologias empregadas atualmente, dedica atenção limitada à integração de informações externas, como dados de reanálise, com observações in situ.

Nesse contexto, recentemente diversas estratégias têm sido empregadas para aproveitar a relação consistente entre modelos de reanálise e a temperatura medida na torre. Em 2019, [Lompar et al. 2019] propôs um procedimento de imputação dos valores faltantes de temperatura em estações meteorológicas, com resolução temporal horária, baseado na correção de viés do ERA5, por meio de uma regressão linear, em uma janela temporal de 3 horas. No entanto, a utilização de modelos puramente lineares pode limitar a capacidade de capturar relações mais complexas, como não-lineares e condicionais, entre a reanálise e as observações, especialmente em casos de alta variabilidade.

Em 2024, [Lalic et al. 2024] realizaram uma análise comparativa entre diferentes modelos de aprendizado de máquina aplicados à imputação de dados meteorológicos, incluindo séries de temperatura do ar. O estudo avalia distintas configurações de conjuntos de features, incluindo informações temporais, dados provenientes de estações vizinhas, variáveis de reanálise do ERA5 e combinações entre observações de estações e dados de reanálise. Os resultados indicam que a integração de features oriundas de estações meteorológicas e do ERA5 apresenta, de forma consistente, os melhores desempenhos entre as configurações avaliadas. Contudo, em função da natureza predominantemente univariada da abordagem adotada e do foco na otimização de algoritmos de predição pon-

tual, o estudo deixa margem para a exploração de métodos de imputação que considerem explicitamente a estrutura multivariada dos dados.

Diante da persistência de lacunas observacionais em biomas tropicais e das limitações observadas quanto à integração de dados observacionais e de reanálise, este trabalho tem como objetivo desenvolver e avaliar uma abordagem de imputação de valores faltantes em séries temporais climáticas, obtidas em estações meteorológicas, que considere a estrutura multivariada dos dados, integrando informações provenientes de estações meteorológicas e de produtos de reanálise, comparando o método proposto a abordagens que utilizam apenas dados observacionais e a um método de interpolação linear, adotado como linha de base.

O presente estudo está organizado da seguinte forma: a Seção 2 descreve o contexto climático da área de estudo, as fontes de dados utilizadas e a metodologia adotada para imputação e avaliação; a Seção 3 apresenta e discute os resultados obtidos a partir da formulação metodológica descrita anteriormente; por fim, a Seção 4 delimita as limitações identificadas e as perspectivas futuras do estudo.

2. Metodologia

Esta seção descreve o fluxo metodológico adotado ao longo da pesquisa, estruturado em quatro etapas principais. A Figura 1 apresenta, de forma esquemática, a organização dessas etapas. Primeiramente, na etapa de **Área de Estudo**, é exposto o contexto climático da região analisada; em seguida, em **Fonte de Dados**, descrevem-se as principais bases utilizadas para a obtenção de dados observacionais e de reanálise; posteriormente, em **Modelos de Imputação**, são apresentados os métodos e técnicas empregados no processo de estimativa de valores ausentes; por fim, em **Avaliação dos Modelos**, descreve-se o procedimento adotado para a introdução controlada de valores ausentes artificiais e os critérios utilizados para avaliar o desempenho dos modelos de imputação.

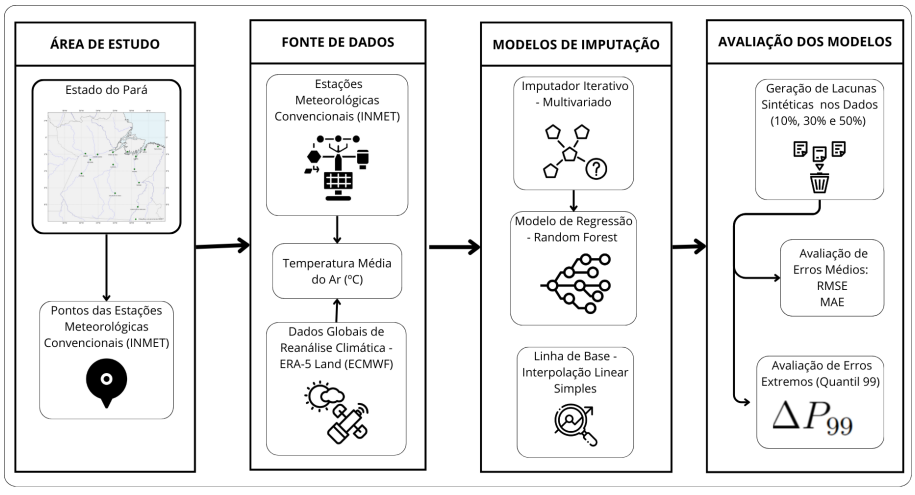

Figura 1. Figura metodológica do artigo

2.1. Área de Estudo

O sítio de estudo compreende o estado do Pará (Figura 2), localizado na região Norte do Brasil, situa-se predominantemente na zona intertropical, apresentando um regime

climático controlado pela dinâmica da Zona de Convergência Intertropical (ZCIT) e pela evapotranspiração da floresta amazônica, que atua na manutenção da umidade atmosférica através dos ”rios voadores”. Segundo a classificação de Köppen, o território paraense é dominado pelos climas tropical úmido e tropical de monção, caracterizados por elevadas temperaturas médias anuais, entre 24°C e 27°C, e por uma reduzida amplitude térmica diurna [Qin et al. 2025].

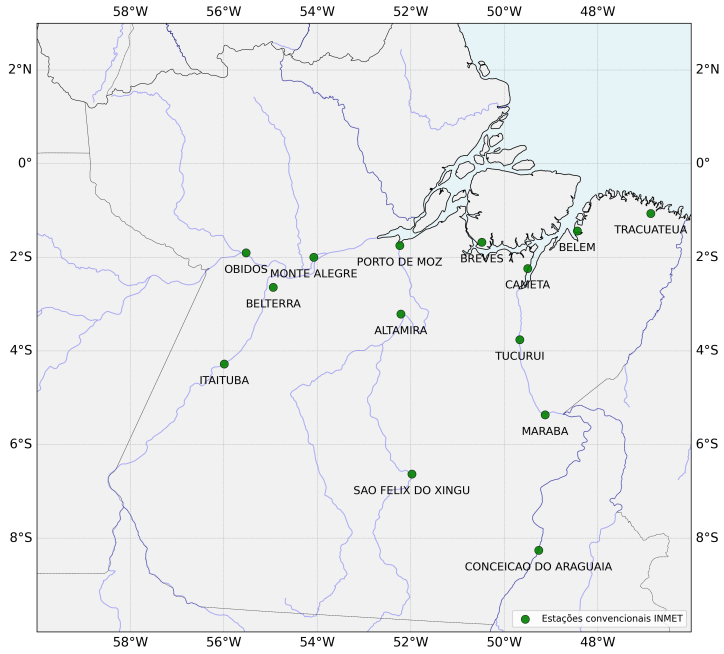

Figura 2. Área de Estudo

2.2. Fonte de dados

Para a coleta de dados de temperatura média do ar, a partir de observações terrestres, utilizaram-se todas as estações convencionais disponibilizadas pelo Instituto Nacional de Meteorologia (INMET)[1] no estado do Pará, dentro do período de 2000 a 2025 em escala temporal diária, totalizando 14 estações (Figura 2).

Além dos dados observacionais do INMET, foram utilizados dados de temperatura média do ar do conjunto de dados de reanálise ERA5-Land diário no mesmo período, coletados a partir da plataforma Google Earth Engine[2]. O ERA5-Land disponibiliza estimativas em resolução temporal horária e resolução espacial de aproximadamente 9 km, consistindo em um reprocessamento do componente terrestre do ERA5 com maior detalhamento espacial, forçado por variáveis atmosféricas do próprio ERA5 [Muñoz-Sabater et al. 2021]. Ao contrário dos dados observacionais do INMET, que apresentam descontinuidades temporais, o ERA5-Land apresenta cobertura temporal contínua de temperatura média no domínio e período analisados, sendo uma fonte de dados complementar para apoiar o procedimento de imputação.

A Figura 3 apresenta a cobertura temporal dos registros de temperatura das estações, em que a cor amarela indica a presença de dados para aquela data, a cor escura

[1]https://bdmep.inmet.gov.br/
[2]https://earthengine.google.com/

indica a ausência de dados e o indicador vermelho marca a data final de atuação daquela estação. Nesse contexto, observou-se que duas das estações, localizadas em Óbidos e São Félix do Xingu, foram desativadas em julho de 2021. Adicionalmente, uma estação apresentou disponibilidade inferior a 45% ao longo do período analisado, sendo excluída da análise por insuficiência amostral. Adicionalmente, a estação localizada no município de Soure foi removida da análise por não possuir dados correspondentes na base ERA5-Land.

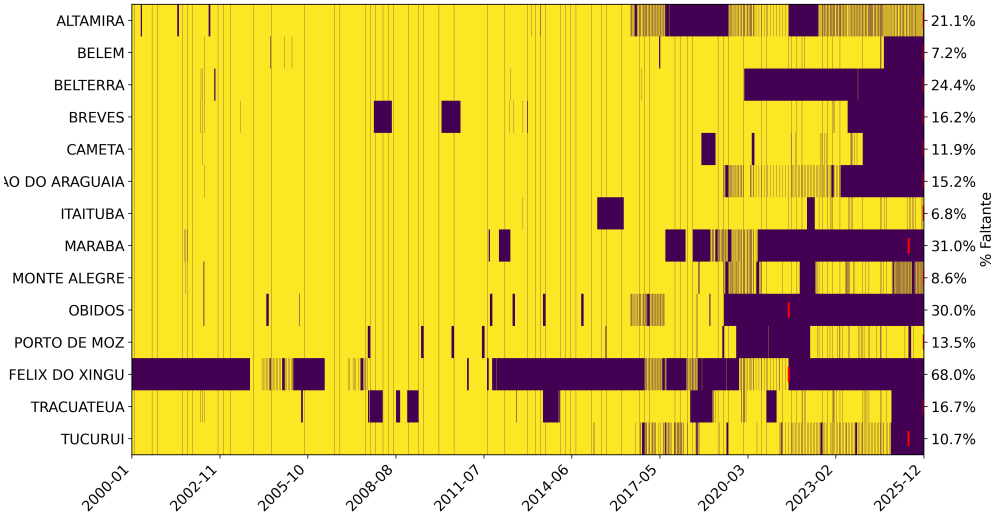

Figura 3. Disponibilidade temporal de dados de temperatura por estação

Por fim, com o objetivo de garantir compatibilidade temporal entre todas as estações, a análise foi restringida ao intervalo de operação simultânea das estações consideradas, compreendendo o período de 2000 a 2021-07-01. Dentro desse período, assume-se que o regime de dados faltantes é aproximadamente aleatório em relação ao processo observacional, pois está associado a falhas pontuais e/ou consecutivas de registro ou manutenção operacional.

2.3. Modelos de Imputação

O processo de imputação refere-se à substituição de valores ausentes por estimativas plausíveis, com o objetivo de preservar a estrutura estatística do conjunto de dados e viabilizar análises subsequentes [Abreu et al. 2024]. Essa substituição pode ser realizada por meio de abordagens determinísticas, como métodos de interpolação , ou por algoritmos de aprendizado de máquina, cuja adequação depende das características do conjunto de dados analisado.

Nesse contexto, adotou-se uma metodologia multivariada baseada em imputação iterativa sequencial, na qual, para cada instante de tempo t, o valor ausente da temperatura média $T_i(t)$ associada à estação i é estimado a partir de uma função de regressão $T_i(t) = f(T_j(t))$, para todo $j \neq i$. O conjunto de preditores é composto pelas n estações que apresentam observações válidas no instante t.

A implementação do procedimento descrito foi realizada por meio da função *Ite-*

rativeImputer, disponível na biblioteca scikit-learn[3], utilizando *Random Forest* como regressor e simulando o comportamento do modelo *MissForest* (MF), caracterizando uma função ($f(\cdot)$) não linear estimada iterativamente. A escolha desta forma de modelagem é justificada por considerar explicitamente a associação instantânea entre as observações de temperatura registradas nas diferentes estações meteorológicas, além de se tratar de uma abordagem já consolidada na literatura recente [Budiawan et al. 2025, Alejo-Sanchez et al. 2025].

De forma complementar, as estimativas provenientes da reanálise são incorporadas como covariáveis no modelo de imputação, de modo que valores faltantes são estimados a partir das temperaturas observadas nas demais estações T_j e das temperaturas do ERA5-Land no mesmo instante temporal. O ERA5-Land atua, assim, como fonte de informação exógena nos períodos sem observação, explorando a relação estatisticamente consistente entre dados observacionais e produtos de reanálise reportada na literatura [Tan et al. 2023].

Por fim, como linha de base adotou-se a interpolação linear simples, que consiste em um método de estimação de valores faltantes baseado na premissa de suavidade da função no tempo, assumindo continuidade na evolução da temperatura média e estimando o valor ausente a partir dos dois registros observados mais próximos temporalmente [Noor et al. 2015].

2.4. Avaliação dos Modelos

Para a avaliação dos modelos de imputação, adotou-se uma estratégia baseada na geração de lacunas sintéticas nas séries temporais de temperatura, inspirada em Zhang [Zhang 2023]. O procedimento consiste na remoção artificial de blocos contínuos de observações, de forma controlada, a partir das séries completas. Inicialmente, seleciona-se aleatoriamente uma estação meteorológica, assumindo-se uma distribuição uniforme entre todas as estações consideradas.

Em seguida, define-se aleatoriamente um instante inicial ao longo da série temporal, também segundo uma distribuição uniforme, garantindo que todos os dias possuam a mesma probabilidade de serem selecionados. O comprimento do bloco removido é então determinado a partir de uma distribuição uniforme, variando entre um dia e até duas vezes o comprimento médio dos blocos observados nos dados reais, de forma a simular um período de manutenção ou falha operacional da estação. O bloco então é removido, e o procedimento é repetido até atingir a porcentagem desejada.

Considerando que os dados originais já possuíam lacunas reais, foram introduzidos três níveis de contaminação artificial sobre as observações disponíveis, correspondendo à remoção de 10%, 30% e 50% das observações originalmente registradas em cada série temporal. Dada a natureza estocástica do experimento, cada configuração foi repetida cinco vezes, permitindo o cálculo da média e da variância das métricas de erro e proporcionando uma avaliação mais confiável da variabilidade do desempenho dos modelos em diferentes cenários de lacunas sintéticas.

Ademais, os modelos foram avaliados utilizando métricas de erro tanto no contexto pontual quanto em termos de distribuição. Para enfatizar erros de maior magni-

[3]https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html

tude, utilizou-se a Raiz Média do Erro Quadrático (RMSE), em virtude da ponderação quadrática aplicada aos desvios. Para caracterizar o erro médio típico, com menor sensibilidade a valores extremos, utilizou-se o Erro Médio Absoluto (MAE) [Tatachar 2021]. Por fim, para avaliar o comportamento do modelo na cauda superior da distribuição de erros, como forma empírica de análise de eventos extremos, tais como dias de calor intenso, analisou-se o MAE associado ao quantil 99 (ΔP_{99}) [Rodrigues et al. 2023].

3. Resultados e Discussão

A Tabela 1 condensa os resultados do experimento descrito na seção anterior, em que os melhores resultados estão em negrito. Para o treinamento dos modelos, utilizaram-se os parâmetros padrão disponibilizados na implementação presente no scikit-learn, correspondendo a 100 árvores sem profundidade previamente delimitada. O tamanho médio dos blocos foi de 4 dias, o que resultou na simulação de lacunas com comprimentos variando entre 1 e até 8 dias, de acordo com o procedimento descrito anteriormente.

Nesse contexto, é possível observar que a incorporação do ERA5-Land, como covariável auxiliar, resulta em melhorias consistentes na reconstrução da temperatura média, quando comparada à configuração que utiliza exclusivamente dados de estações meteorológicas.

Tabela 1. Resultados da imputação (média $\pm$ desvio padrão, 5 repetições).

% Faltante	Método	RMSE	MAE	ΔP_{99}
10%	MF	0.752 ± 0.017	0.567 ± 0.010	1.405 ± 0.140
	MF (+ERA5-Land)	$\mathbf{0.648 \pm 0.009}$	$\mathbf{0.500 \pm 0.005}$	$\mathbf{0.869 \pm 0.082}$
	Interp. Linear	0.940 ± 0.013	0.697 ± 0.007	0.996 ± 0.135
30%	MF	0.814 ± 0.002	0.613 ± 0.002	1.419 ± 0.065
	MF (+ERA5-Land)	$\mathbf{0.671 \pm 0.002}$	$\mathbf{0.517 \pm 0.002}$	$\mathbf{0.882 \pm 0.026}$
	Interp. Linear	0.966 ± 0.005	0.718 ± 0.003	1.083 ± 0.021
50%	MF	0.909 ± 0.005	0.684 ± 0.003	1.528 ± 0.108
	MF (+ERA5-Land)	$\mathbf{0.725 \pm 0.002}$	$\mathbf{0.558 \pm 0.002}$	$\mathbf{0.883 \pm 0.042}$
	Interp. Linear	1.281 ± 0.386	0.737 ± 0.007	1.154 ± 0.062

O MF com ERA5-Land apresentou, em média, reduções de 17.2% no RMSE, 15.3% no MAE e 39.4% no ΔP_{99} em relação ao MF usando apenas as estações, e de 35.0%, 26.9% e 18.3%, respectivamente, em relação à interpolação linear, com valores estimados a partir da média de cinco experimentos independentes com inicialização aleatória. Em termos absolutos, o modelo apresentou valores de MAE variando entre 0.500 °C e 0.558 °C, de RMSE entre 0.648 °C e 0.725 °C e ΔP_{99} entre 0.869°C e 0.883°C ao longo dos diferentes níveis de dados faltantes.

Destaca-se também que o desvio padrão das métricas permaneceu consistentemente baixo em todos os cenários ($\leq 0{,}009$ °C para o RMSE, $\leq 0{,}005$ °C para o MAE e $\leq 0{,}082$ °C para o ΔP_{99}) . Em contraste, a interpolação linear apresentou desvio padrão de até $0{,}386$ °C no RMSE no cenário de 50% de dados faltantes, indicando que a ausência de contexto climático instantâneo torna o método mais sensível à configuração das lacunas.

Dessa forma, constata-se que o MF, ao integrar ambas as fontes de dados, apresenta desempenho superior não apenas em todas as métricas avaliadas, mas também maior estabilidade frente à aleatoriedade do experimento, expressa por menores desvios padrão em todas as métricas e para todos os níveis de dados faltantes. Esse resultado evidencia, além da melhoria quantitativa na imputação, uma maior capacidade do método em reconstruir a temperatura média regional com fidelidade, preservando a variabilidade térmica mesmo em cenários associados a extremos empíricos e elevada perda de informação. A redução substancial do ΔP_{99} indica maior capacidade na reconstrução da intensidade e da persistência do calor extremo regional, evitando a suavização artificial desses episódios mesmo sob altos níveis de ausência de dados, aspecto particularmente relevante no estado do Pará, onde a atuação sazonal da ZCIT influencia a variabilidade térmica.

Nesse sentido, observou-se que há uma redução consistente em relação ao MF, apenas com as estações e a interpolação linear, nas métricas de erro médio; isto é, RMSE e MAE. Esse comportamento pode ser explicado pela incorporação do contexto climático adicional fornecido pelo ERA5-Land, que atua como uma representação física consistente da atmosfera. Tal abordagem é conceitualmente alinhada a metodologias de *correção de viés*, nas quais produtos de reanálise são utilizados como referência para a reconstrução de variáveis climáticas observadas e *in situ*, sendo uma técnica amplamente empregada na literatura devido a sua baixa taxa de erro [Lompar et al. 2019, Alejo-Sanchez et al. 2025].

Apesar das dificuldades enfrentadas por produtos de reanálise, como o ERA5-Land, em cenários tropicais, incluindo vieses sistemáticos de superestimação ou subestimação de eventos atípicos, estudos prévios têm destacado uma elevada correlação estatística linear entre os valores estimados e observados na região de estudo durante períodos extremos, indicando a necessidade da aplicação de técnicas de correção de viés como etapa prévia à análise desses períodos [Tan et al. 2023].

Em concordância com o supracitado, a utilização do ERA5-Land como variável de suporte em um modelo não linear permitiu uma correção de viés de forma implícita e complementar. Dessa forma, o maior ganho observado, em relação ao modelo baseado apenas nas estações, ocorreu na métrica ΔP_{99}, que avalia empiricamente a capacidade de reconstrução de episódios persistentes de calor extremo na região de estudo.

4. Conclusão

Este estudo avaliou o impacto da inclusão de dados de reanálise como fonte auxiliar na reconstrução de valores ausentes em séries de temperatura de estações meteorológicas do INMET, utilizando uma abordagem multivariada não linear baseada no algoritmo MF. A integração entre fontes permitiu que o modelo mantivesse precisão e estabilidade mesmo em cenários críticos de degradação da base de dados, como na remoção de 50% das observações originais, superando as limitações de modelos que dependem exclusivamente de estações vizinhas ou de métodos puramente lineares. O ganho mais expressivo dessa integração foi observado na reconstrução de extremos térmicos, onde o uso do ERA5-Land como fonte complementar reduziu o ΔP_{99} em quase 40% em relação ao uso isolado de outras estações.

Os resultados indicam que a integração de dados de reanálise ERA5-Land contribui diretamente para mitigar as incertezas associadas à escassez, irregularidade e descontinuidade dos registros observacionais na região amazônica. A inclusão dessas covariáveis

auxilia o modelo a capturar a variabilidade climática regional de forma mais consistente do que métodos convencionais, preservando adequadamente a estrutura térmica da série mesmo durante períodos de calor extremo ou interrupções operacionais nas estações.

Para estudos futuros, pretende-se investigar critérios de pré-seleção de estações auxiliares, avaliando quais séries carregam informação climática efetivamente complementar à série-alvo. Além disso, serão exploradas técnicas de clusterização como etapa prévia à imputação para agrupar estações com comportamentos semelhantes, bem como modelos temporais multivariados para analisar a evolução dinâmica conjunta das séries. Adicionalmente, propõe-se o desenvolvimento de um framework de imputação em tempo real para a rede do INMET no estado do Pará, visando o preenchimento contínuo de lacunas operacionais. Por fim, uma avaliação detalhada da qualidade por cidade permitirá o aprimoramento do monitoramento sistemático, fornecendo a base de dados contínua necessária para a formulação de políticas públicas e estratégias de desenvolvimento sustentável na região.

Referências

Abreu, J., Vidal, D., and Gonçalves, G. (2024). Serviço web para imputação de dados em séries temporais univariadas. In *Anais do XV Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais*, pages 131–140, Porto Alegre, RS, Brasil. SBC.

Alejo-Sanchez, L. E., Márquez-Grajales, A., Salas-Martínez, F., Franco-Arcega, A., López-Morales, V., Acevedo-Sandoval, O. A., González-Ramírez, C. A., and Villegas-Vega, R. (2025). Missing data imputation of climate time series: A review.

Budiawan, I., Wicaksana, H. S., Ekawati, E., and Kurniadi, D. (2025). Quality assurance of field temperature data from weather stations: A case study of missforest imputation. In *2025 9th International Conference on Instrumentation, Control, and Automation (ICA)*, pages 157–162. IEEE.

den Braber, B., Oldekop, J. A., Devenish, K., Godar, J., Nolte, C., Schmoeller, M., and Evans, K. L. (2024). Socio-economic and environmental trade-offs in amazonian protected areas and indigenous territories revealed by assessing competing land uses. *Nature Ecology & Evolution*, 8(8):1482–1492.

Garrett, R., Ferreira, J., Abramovay, R., Brandão, J., Brondizio, E., Euler, A., Pinedo, D., Porro, R., Cabrera Rocha, E., Sampaio, O., et al. (2024). Transformative changes are needed to support socio-bioeconomies for people and ecosystems in the amazon. *Nature Ecology & Evolution*, 8(10):1815–1825.

Han, H., Liu, Z., Li, J., and Zeng, Z. (2024). Challenges in remote sensing based climate and crop monitoring: navigating the complexities using ai. *Journal of cloud computing*, 13(1):1–14.

Hoinaski, L., Will, R., and Ribeiro, C. B. (2024). Brazilian atmospheric inventories–brain: a comprehensive database of air quality in brazil. *Earth System Science Data*, 16(5):2385–2405.

Lalic, B., Stapleton, A., Vergauwen, T., Caluwaerts, S., Eichelmann, E., and Roantree, M. (2024). A comparative analysis of machine learning approaches to gap filling meteorological datasets. *Environmental Earth Sciences*, 83.

Lompar, M., Lalić, B., Dekić, L., and Petrić, M. (2019). Filling gaps in hourly air temperature data using debiased era5 data. *Atmosphere 2019, Vol. 10,*, 10.

Metcalfe, D. B., Anders, E., Axén, H., Petter Axelsson, E., Bermudez, A. E., Bartholomew, D. C., Butt, N., Cadillo-Quiroz, H., Chaudhary, N., Callebaut, T., et al. (2025). Gaps in tropical science from unrepresentative distribution of sampling and citation across natural terrestrial environments. *Nature communications*.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., et al. (2021). Era5-land: A state-of-the-art global reanalysis dataset for land applications. *Earth system science data*, 13(9):4349–4383.

Noor, N. M., Al Bakri Abdullah, M. M., Yahaya, A. S., and Ramli, N. A. (2015). Comparison of linear interpolation method and mean method to replace the missing values in environmental data set. In *Materials science forum*, volume 803, pages 278–281. Trans Tech Publ.

Qin, Y., Wang, D., Ziegler, A. D., Fu, B., and Zeng, Z. (2025). Impact of amazonian deforestation on precipitation reverses between seasons. *Nature*, 639(8053):102–108.

Rodrigues, D. T., Gonçalves, W. A., Silva, C. M. S. E., Spyrides, M. H. C., and Lúcio, P. S. (2023). Imputation of precipitation data in northeast brazil. *Anais da Academia Brasileira de Ciencias*, 95.

Tan, M. L., Armanuos, A. M., Ahmadianfar, I., Demir, V., Heddam, S., Al-Areeq, A. M., Abba, S. I., Halder, B., Kilinc, H. C., and Yaseen, Z. M. (2023). Evaluation of nasa power and era5-land for estimating tropical precipitation and temperature extremes. *Journal of Hydrology*, 624:129940.

Tatachar, A. V. (2021). Comparative assessment of regression models based on model evaluation metrics. *International Research Journal of Engineering and Technology (IRJET)*, 8(09):2395–0056.

United Nations General Assembly (2015). Transforming our world: the 2030 agenda for sustainable development. A/RES/70/1. 21 October 2015. Available at: `https://www.refworld.org/legal/resolution/unga/2015/en/111816` [Accessed 15 February 2026].

Zhang, X. (2023). How to generate missing data for simulation studies. *The Quantitative Methods for Psychology*, 19:100–122.