Comparison of LSTM and SARIMA Models for Air Temperature Forecasting in Belém, Amazônia, Pará

Leonardo de O. Tamasauskas[1] , Williane G. S. Pereira[1], Waldemiro J. A. G. Negreiros[1], Pedro H. do V. Guimarães[1], Jean A. C. Dias[1], Alan B. S. Corrêa[1], Gabriel B. Costa[2], Marcos C. da R. 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

{jean.dias,alan.correa}@itec.ufpa.br, Waldemiro.negreiros@ifpa.edu.br
{williane.pereira, leonardo.tamasauskas}@icen.ufpa.br, seruffo@ufpa.br
pedro.guimaraes@castanhal.ufpa.br, gabrielbritocosta@gmail.com

Abstract. *This study investigates air temperature forecasting in the city of Belém-PA, comparing the performance of the SARIMA and LSTM models. To this end, daily data from the ERA5-Land database was used and statistical metrics such as mean absolute error (MAE), mean squared error (MSE) and coefficient of determination (R^2) were evaluated, supported by ANOVA, Shapiro-Wilk, Levene and Tukey tests. The results indicate that the LSTM model was more accurate, capturing complex patterns better than SARIMA. The findings reinforce the potential of recurrent neural networks in climate modeling and suggest new approaches for improving weather forecasting.*

1. Introduction

Growing urbanization and changes in land use significantly affect the urban thermal environment, progressively replacing natural areas with impermeable surfaces, compromising the region's hydrological cycle [Karunaratne et al. 2022]. The increase in built-up areas results in serious climate changes, such as the intensification of the urban heat island, an increasingly common phenomenon in cities around the world [Yadav et al. 2023]. In addition, global warming exacerbates the urban microclimate, making heatwaves more intense and continuously contributing to rising air temperatures in cities [Du et al. 2019, Fan et al. 2024].

This increase in temperature rise events has consequences for various sectors of society, affecting human health in their daytime activities [Arsad et al. 2022], food production [Costa et al. 2022], the use of water from ecosystems [Zhang et al. 2022] and electricity for cooling [Stone Jr et al. 2021]. Therefore, it is essential for public authorities to be able to predict the increase in heat at local level in advance, in line with the guidelines of the United Nations Office for Disaster Risk Reduction (UNDRR), which promotes the management of Resilient Cities, capable of resisting, adapting and recovering from extreme weather events [UNDRR 2025].

In view of this, the study by [Gill et al. 2023] used statistical models, such as the *Seasonal AutoRegressive Integrated Moving Average* (SARIMA), to predict temperature.

The results showed that, although the model had some limitations, it was able to capture the dynamics of the time series and generate sensible forecasts for air temperature. Similarly, another study conducted by [Ayad 2022] used SARIMA for long-term monthly air temperature forecasting, showing that the model fitted the historical data well, demonstrating its effectiveness in predicting future temperatures.

As data becomes more complex, models based on neural networks stand out for capturing non-linear relationships and for their generalization capacity [Paspaltzis and Calheiros 2023]. In the study by [Khalil et al. 2022], a comparative analysis was carried out between two models for predicting the Earth's surface temperature: *Long Short-Term Memory* (LSTM) and *Artificial Neural Network* (ANN). The results indicated that the LSTM outperformed the ANN. Similarly, in [Park et al. 2019], with the aim of predicting air temperature, the authors compared a refined version of the LSTM with the *Deep Neural Network* (DNN), observing that the LSTM presented lower error metrics in its prediction.

In this context, this work carries out a comparative analysis between the LSTM neural network model and the SARIMA statistical model, with the aim of evaluating the performance of both in air temperature forecasting. The research was conducted using daily mean air temperature data for the city of Belém, in the state of Pará (PA), collected from the ERA5-Land dataset, an atmospheric reanalysis database.

2. Methodology

The methodology of this work has been structured into five main stages, which cover the entire process of developing and analyzing the research. The steps are presented in detail in the flowchart in Figure 1, which clearly illustrates the sequence of activities.

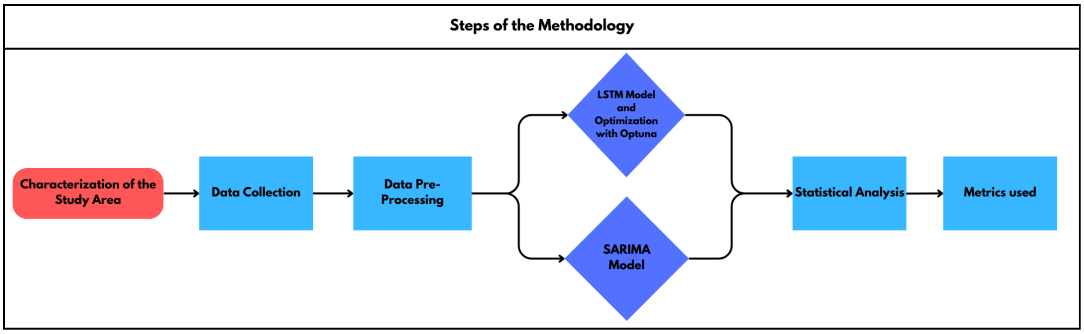


Figure 1. Article Methodology Flowchart

2.1. Characterization of the Study Area

To delimit the area of interest, we used the entire demarcation of the municipality of Belém, the capital of the state of Pará, approximately 160 km from the Equator, to the south, under the geographical coordinates 01°27' S and 48°28' W [Brasil 2023]. It has a total area of 1059 Km² and approximately 1.303 million inhabitants, according to the 2022 census by the Brazilian Institute of Geography and Statistics[1] (IBGE, in Portuguese). The climate is classified as hot and humid equatorial. The geographical outline of the city is described in Figure 2, using the geometry obtained from the IBGE portal[2].

[1]https://cidades.ibge.gov.br/
[2]https://portaldemapas.ibge.gov.br/portal.php

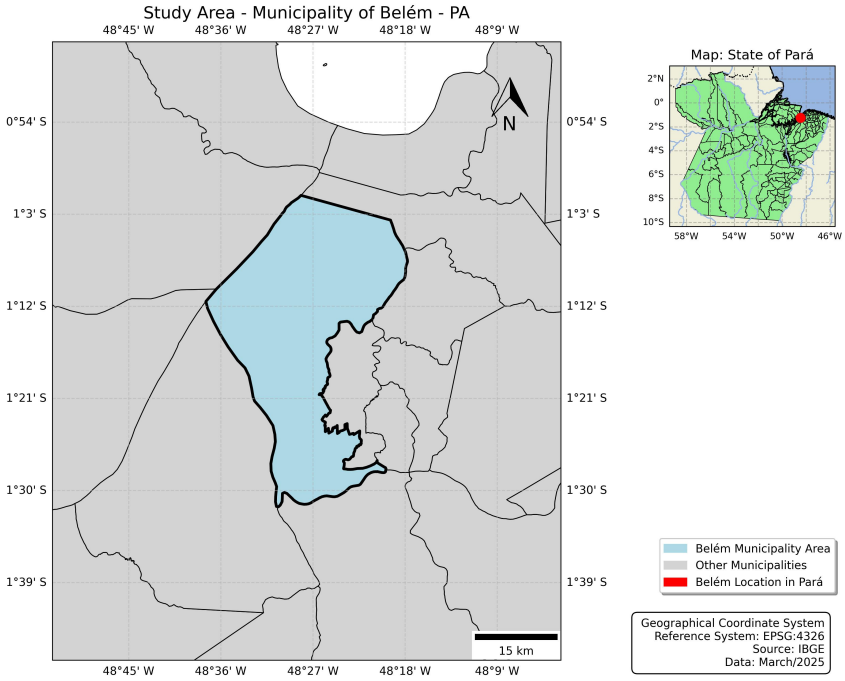


Figure 2. Map of the Study Area

2.2. Data Collection

ERA5-Land is an atmospheric reanalysis dataset that provides information on terrestrial variables, with an improved resolution compared to ERA5, and is made available by the *European Centre for Medium-Range Weather Forecasts* (ECMWF)[3]. This product provides air temperature data at a height of 2 meters above the earth's surface, with daily temporal resolution and spatial resolution of $0.1º \times 0.1º$ (approximately 9 km).

For this study, daily air temperature data was collected in the study area, covering the period from 1994 to 2024. The data provided by ERA5-Land was already aggregated for the daily period, with the daily average calculated at the time of retrieval. Considering that the values of this variable have a spatial resolution of 9 km, the median was calculated for the study area in order to obtain a single value representative of the region's daily air temperature, reducing the influence of extreme values.

2.3. Data Pre-Processing

2.3.1. LSTM

Data pre-processing consisted of three main stages: **cleaning**, **normalization** and **preparation**. Initially, missing values in the temperature column were removed, ensuring the integrity of the dataset. The date column was converted to the time index format, facilitating sequential analysis. To standardize the values, the *MinMaxScaler* [Pedregosa et al. 2011] technique was used, scaling the data to the $[0, 1]$ interval. In addition, the *sliding window* technique was applied with a 7-day interval, using each 7-day window as an input to predict the next day's temperature.

[3]https://cds.climate.copernicus.eu/datasets/derived-era5-land-daily-statistics?tab=overview

2.3.2. SARIMA

Data pre-processing involved three main steps: **data preparation**, **cleaning** and **stationarity test**. The date column was converted to the *datetime* format and defined as an index, transforming the data set into a time series. During cleaning, missing values in the temperature column were identified and removed, ensuring the integrity of the model. Descriptive statistics were calculated, including mean, standard deviation, minimum and maximum values and quartiles, to understand the distribution of the data. The stationarity of the time series was assessed using the Augmented Dickey-Fuller Test (ADF) [Silveira et al. 2022], which checks for the presence of a unit root, indicating whether the series is non-stationary. If the null hypothesis of the presence of a unit root is not rejected, differentiation of the series is applied to make it suitable for modeling with SARIMA.

2.4. LSTM Model and Optimization with Optuna

The model in this study is based on Recurrent Neural Networks (RNNs), specifically the LSTM variant, which is widely used for modeling time series due to its ability to capture long-term dependencies.

The architecture employs multiple stacked LSTM layers (*stacked LSTM*), followed by a dense layer for the final prediction, allowing hierarchical learning of temporal characteristics. To prevent overfitting, ***Dropout*** layers were inserted, which help to regularize the model, preventing overlearning from noise in the training data. The TensorFlow/Keras LSTM uses ***tanh*** activation for the state cell and ***sigmoid*** for the gates. [Abadi et al. 2015].

The hyperparameters were optimized using Optuna, a framework based on Bayesian searches [Akiba et al. 2019]. The hyperparameters adjusted were the number of LSTM layers (1 to 3), the number of units per layer (32 to 128), the batch size (16 to 64) and the learning rate (from 0.0001 to 0.01). Optuna minimized the Mean Absolute Error (MAE) loss function over 10 iterations (trials), using the Tree-structured Parzen Estimator (TPE) algorithm [Bergstra et al. 2011] to efficiently search for the best parameters. The best configuration found included n_{layers} LSTM layers, each with n_{units} units, a batch size of batch_size and a learning rate of learning_rate.

The LSTM model was then built with the best hyperparameters found by Optuna, consisting of up to 3 LSTM layers, with units ranging from 32 to 128 per layer. Training was carried out for 30 epochs, with a batch size of 32, using the validation set to monitor performance. After training, the model was evaluated on the test set, calculating the loss to verify its generalization.

2.5. SARIMA Model

The SARIMA model is an extension of the Autoregressive Integrated Moving Average (ARIMA), which incorporates seasonal components and is widely used for forecasting time series with seasonal patterns [Scheffer et al. 2014].

The model is characterized by two sets of parameters: the non-seasonal parameters (p, d, q) and the seasonal parameters (P, D, Q, s). The non-seasonal parameters are: p, which represents the order of the autoregressive (AR) term; d, which indicates the number of differentiations needed to make the series stationary; and q, which is the order of the

moving average (MA) term. The seasonal parameters are: P, which is the order of the seasonal autoregressive term; D, which denotes the number of seasonal differentiations; Q, which is the order of the seasonal moving average term; and s, which specifies the seasonal period (for example, $s = 7$ for weekly seasonality).

The parameters were selected using the function `auto_arima` from the pmdarima library, which tests various combinations of parameters and chooses the configuration that minimizes the Akaike Information Criterion (AIC) [Smith et al. 17]. Seasonality was set to weekly ($s = 7$), due to the periodicity observed in the data.

The model selected was SARIMA(1,1,2)(0,0,0)[7]. The model was implemented using the `SARIMAX` function from the `statsmodels` library, with the parameters identified [Seabold and Perktold 2010]. For evaluation, an iterative forecast was implemented for the next 8 days, continuously updating the training set with the actual values. The forecasts generated were stored for later comparison with the real data.

2.6. Statistical Analysis

For the systematic comparison between LSTM and SARIMA models, a sequential statistical approach was adopted following established specialized literature. Initially, a one-way analysis of variance (ANOVA) [Cuevas et al. 2004] was planned as the primary parametric method, aiming to test the null hypothesis (H_0) of equality between the performance metric means of the considered models. The significance level was set at $\alpha = 0.05$, using the F-statistic derived from the ratio between intergroup and intragroup variances as the decision parameter. Prior to conducting the ANOVA, verification of fundamental assumptions was planned, including assessment of residual distribution normality through the Shapiro-Wilk test [Hanusz et al. 2016] and analysis of variance homogeneity via Levene's procedure [Gastwirth et al. 2009]. Should the ANOVA indicate statistically significant differences, post hoc analyses were planned using Tukey's test [Abdi and Williams 2010], which would specifically identify differing models while properly controlling Type I error rate in multiple comparisons through calculation of the q statistic, simultaneously considering both the magnitude of mean differences and within-group variability.

2.7. Metrics used

To evaluate the performance of both models, we used metrics that are commonly used in regression tasks. The metrics chosen were Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Coefficient of Determination (R^2) [Nogueira and Moreira 2015].

Median Absolute Error (MAE): Average of the absolute differences between predicted and actual values, useful for its simplicity, but without penalizing large errors significantly.

Mean Squared Error (MSE): Average of the squares of the differences between forecasts and actual values, penalizing larger errors.

Root Mean Square Error (RMSE): Square root of the MSE, keeping the same unit as the data and penalizing large errors.

Coefficient of Determination (R^2): Measures the proportion of the variance explained by the model, ranging from 0 (no explanation) to 1 (perfect fit).

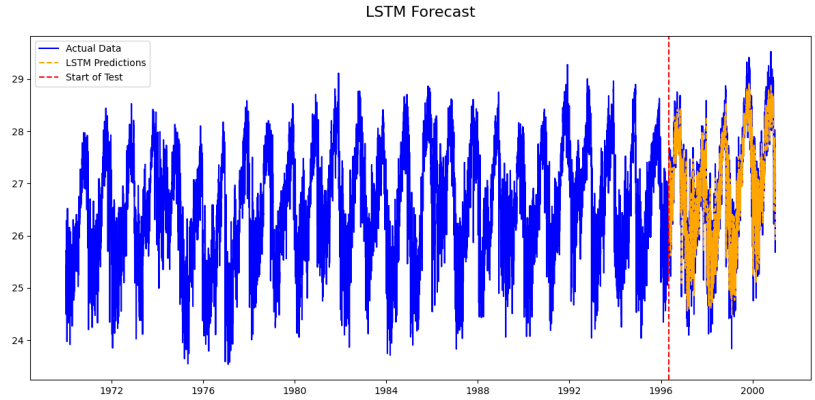
3. Results and Discussion

Table 1 shows the results of the metrics obtained after training the models, with the best ones highlighted in bold. It can be seen that the LSTM had the highest predictive capacity for air temperature, outperforming the other models in all metrics. The LSTM obtained an MAE of 0.35, which indicates more accurate predictions, as well as an MSE of 0.21 and RMSE of 0.46, which show extreme errors of low magnitude. The R^2 value, higher than 0.83, also suggests that the model can better explain temperature variability based on its predictions.

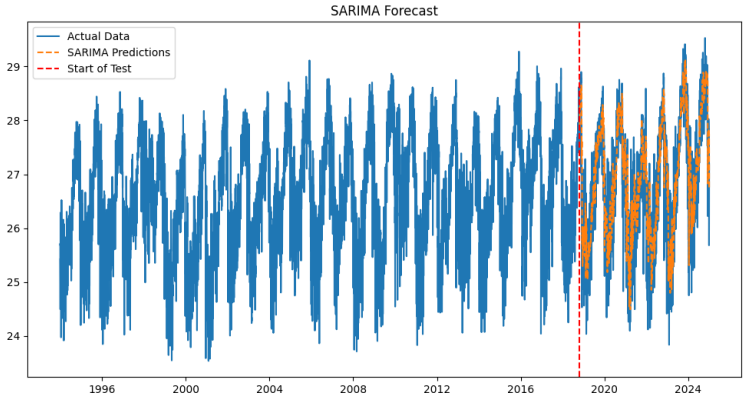
Model	MAE	MSE	RMSE	R^2
LSTM	**0.35**	**0.21**	**0.46**	**0.83**
SARIMA	0.46	0.39	0.62	0.68

Table 1. Comparison of LSTM and SARIMA values.

The graphs shown in Figure 3 illustrate the time series of real data (in blue), the values predicted by the models (in orange) and the start of the test set (in red). The greater overlap between the forecast curves and the actual values in the LSTM model (Figure 3(a)) indicates a more accurate modeling of air temperature variability over time. This behavior reinforces the previous results, highlighting the predictive superiority of the LSTM over the SARIMA model (Figure 3(b)).



(a) LSTM forecasts



(b) SARIMA forecasts

Figure 3. Comparison of LSTM and SARIMA forecasts.

The ANOVA revealed statistically significant differences between the LSTM and SARIMA models ($F = 8.265$, $p = 0.00406$), confirming LSTM's superiority across all evaluated metrics. The assumptions of normality and homogeneity of variances were satisfied, validating the results. Tukey's post hoc test demonstrated that LSTM performed significantly better than SARIMA ($p < 0.01$), with MAE values of 0.35 versus 0.46 and R^2 values of 0.83 versus 0.68, respectively.

The results obtained are in line with recent literature on modeling environmental time series [Balasooriya et al. 2022, Akbar et al. 2024]. The greater predictive capacity of LSTM can be attributed to the use of short- and long-term memory blocks, which allow the network to capture complex temporal dependencies, including local temperature peaks, and thus improve the accuracy of forecasts [Hochreiter and Schmidhuber 1997]. In contrast, SARIMA was developed to model less chaotic time series, with the assumption of linear relationships between the lagged components of the series, which makes it difficult to adapt to non-linear patterns or abrupt variations [Necula et al. 2025].

4. Conclusion

This study analyzed and compared the performance of the LSTM and SARIMA models in predicting air temperature, using daily data from the city of Belém, capital of the state of Pará, Amazonia, Brazil. The results indicate that the LSTM model showed greater accuracy, better ability to adapt to climate variations and lower errors - mean absolute error (MAE = 0.35), mean squared error (MSE = 0.21) and root mean squared error (RMSE = 0.46) - as well as a coefficient of determination (R^2 = 0.83), demonstrating its ability to capture complex patterns of temperature variation.

The statistical analysis reinforced LSTM's superiority. The one-way ANOVA test revealed significant differences between the models (F = 8.265; p = 0.00406), confirming that LSTM's lower error magnitude was not random. Additionally, the assumptions of normality (Shapiro-Wilk test) and homogeneity of variances (Levene's test) were met, validating the application of the Tukey post hoc test, which highlighted the statistically significant difference (p ¡ 0.01) between the models.

This research highlights the importance and efficiency of recurrent neural networks in modeling environmental time series, especially in forecasting meteorological variables. While the SARIMA model showed difficulties in capturing non-linear patterns, the LSTM showed greater capacity for adaptation and generalization, consolidating itself as a promising option for medium and long-term climate forecasts.

Despite the good results, this study has some limitations, such as the need to adjust hyperparameters for the LSTM and the dependence on high-quality historical data. In addition, extreme weather events can have an impact on forecast accuracy.

For future work, we suggest exploring hybrid models that combine the statistical robustness of SARIMA with the predictive capacity of neural networks. Furthermore, the inclusion of additional meteorological variables, such as humidity and wind speed, could improve the accuracy of the models. The evolution of machine learning techniques can make a significant contribution to climate change mitigation and adaptation strategies, aiding decision-making and sustainable urban planning.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Abdi, H. and Williams, L. J. (2010). Newman-keuls test and tukey test. *Encyclopedia of research design*, 2:897–902.

Akbar, A. A., Darmawan, Y., Wibowo, A., and Rahmat, H. K. (2024). Accuracy assessment of monthly rainfall predictions using seasonal arima and long short-term memory (lstm). *Journal of Computer Science and Engineering (JCSE)*, 5(2):100–115.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Arsad, F. S., Hod, R., Ahmad, N., Ismail, R., Mohamed, N., Baharom, M., Osman, Y., Radi, M. F. M., and Tangang, F. (2022). The impact of heatwaves on mortality and morbidity and the associated vulnerability factors: a systematic review. *International Journal of Environmental Research and Public Health*, 19(23):16356.

Ayad, S. (2022). Modeling and forecasting air temperature in tetouan (morocco) using sarima model. *J. Earth Sci. Geotech. Eng*, 12:1–13.

Balasooriya, S., Nguyen, C., Kavalchuk, I., and Yasakethu, L. (2022). Forecasting model comparison for soil moisture to obtain optimal plant growth. In *2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pages 1–7. IEEE.

Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24.

Brasil, I. (2023). Censo demográfico 2022. *Dados nacionais. Fundação Instituto Brasileiro de Geografia e Estatística. Brasil*.

Costa, H., Santana, T., Silva, R., Gomes, N., Gonçalves, G., Guiselini, C., Almeida, G., and Medeiros, V. (2022). Estudo da viabilidade técnica do emprego de detectores térmicos de baixa resolução na suinocultura 4.0. In *Anais do XIII Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais*, pages 21–30, Porto Alegre, RS, Brasil. SBC.

Cuevas, A., Febrero, M., and Fraiman, R. (2004). An anova test for functional data. *Computational statistics & data analysis*, 47(1):111–122.

Du, Q., Zhang, M., Wang, S., Che, C., Ma, R., and Ma, Z. (2019). Changes in air temperature over china in response to the recent global warming hiatus. *Journal of Geographical Sciences*, 29:496–516.

Fan, C., Zou, B., Li, J., Wang, M., Liao, Y., and Zhou, X. (2024). Exploring the relationship between air temperature and urban morphology factors using machine learning under local climate zones. *Case Studies in Thermal Engineering*, 55:104151.

Gastwirth, J. L., Gel, Y. R., and Miao, W. (2009). The impact of levene's test of equality of variances on statistical theory and practice.

Gill, K., Bhatt, K., Kaur, B., and Sandhu, S. S. (2023). Arima approach for temperature and rainfall time series prediction in punjab. *Journal of Agrometeorology*, 25(4):571–576.

Hanusz, Z., Tarasinska, J., and Zielinski, W. (2016). Shapiro–wilk test with known mean. *REVSTAT-statistical Journal*, 14(1):89–100.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Karunaratne, S., Athukorala, D., Murayama, Y., and Morimoto, T. (2022). Assessing surface urban heat island related to land use/land cover composition and pattern in the temperate mountain valley city of kathmandu, nepal. *Remote Sensing*, 14(16).

Khalil, U., Azam, U., Aslam, B., Ullah, I., Tariq, A., Li, Q., and Lu, L. (2022). Developing a spatiotemporal model to forecast land surface temperature: A way forward for better town planning. *Sustainability*, 14(19).

Necula, S.-C., Hauer, I., Fotache, D., and Hurbean, L. (2025). Advanced hybrid models for air pollution forecasting: Combining sarima and bilstm architectures. *Electronics (2079-9292)*, 14(3).

Nogueira, S. M. C. and Moreira, M. A. (2015). Avaliação das previsões de temperatura do modelo eta para o estado do paraná. In *Anais do XVII Simpósio Brasileiro de Sensoriamento Remoto (SBSR)*, pages 2071–2078, São José dos Campos. Instituto Nacional de Pesquisas Espaciais (INPE). Acesso em: 3 mar. 2025.

Park, I., Kim, H. S., Lee, J., Kim, J. H., Song, C. H., and Kim, H. K. (2019). Temperature prediction using the missing data refinement model based on a long short-term memory neural network. *Atmosphere*, 10(11):718.

Paspaltzis, V. and Calheiros, A. (2023). Uma abordagem usando redes neurais artificiais para a previsão de curto prazo de altura de ondas marítimas em região portuária. In *Anais do XIV Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais*, pages 141–150, Porto Alegre, RS, Brasil. SBC.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Scheffer, D., Souza, A. M., Zanini, R. R., et al. (2014). Utilização de modelos arima para previsão da arrecadação de icms do estado do rio grande do sul. *Simpósio de Pesquisa Operacional e Logística da Marinha*, 17.

Seabold, S. and Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Silveira, A., Mattos, V., Nakamura, L., Amaral, M., Konrath, A., and Bornia, A. (2022). Análise do valor-p determinado pela estatística τ na aplicação do teste de dickey-fuller aumentado. *Trends in Computational and Applied Mathematics*, 23(2):283–298.

Smith, T. G. et al. (2017–). pmdarima: Arima estimators for Python. [Online; accessed ¡today¿].

Stone Jr, B., Mallen, E., Rajput, M., Gronlund, C. J., Broadbent, A. M., Krayenhoff, E. S., Augenbroe, G., O'Neill, M. S., and Georgescu, M. (2021). Compound climate and infrastructure events: how electrical grid failure alters heat wave risk. *Environmental Science & Technology*, 55(10):6957–6964.

UNDRR (2025). Making Cities Resilient 2030. Avaliable in: `https://mcr2030.undrr.org`.

Yadav, N., Rajendra, K., Awasthi, A., Singh, C., and Bhushan, B. (2023). Systematic exploration of heat wave impact on mortality and urban heat island: A review from 2000 to 2022. *Urban Climate*, 51:101622.

Zhang, C., Yin, Y., Chen, G., Deng, H., Ma, D., and Wu, S. (2022). Water use efficiency-based assessment of risk to terrestrial ecosystems in china under global warming targets of 1.5° c and 2.0° c. *Ecological Indicators*, 143:109349.