Artificial Intelligence and Discrete Wavelet Transform Applied to the Forecasting of Plastic Waste Leakage in the Ocean

Iwldson G Silva Santos [1, 3, *] - https://orcid.org/0000-0001-5663-9047
Isnaldo I Barbosa [2] - https://orcid.org/0000-0003-3147-1780
José F Oliveira Júnior [3] - https://orcid.org/0000-0001-5280-9600
Micejane S Costa [3] - https://orcid.org/0000-0002-5724-1441
Adalgisa Fornaro [4] - https://orcid.org/0000-0003-0086-9649
Daniel C Zacharias [4] - https://orcid.org/0000-0002-9355-8964
Rodrigo M S Brito [5] - https://orcid.org/0000-0002-4052-4541

[1] IAG, USP, São Paulo, SP, Brasil
Email: iwldson.santos@usp.br
* Corresponding author: iwldson@gmail.com

[2] IM, UFAL, Maceió, AL, Brasil
Email: isnaldo@pos.mat.ufal.br

[3] LAMMA, ICAT, UFAL, Maceió, AL, Brasil
Email: jose.junior@icat.ufal.br
Email: micejane@yahoo.com.br

[4] IAG, USP, São Paulo, SP, Brasil
Email: Adalgiza.fornaro@iag.usp.br
Email: danizach@gmail.com

[5] IGDEMA, UFAL, Maceió, AL, Brasil
Email: rodrigo.brito@igdema.ufal.br

RESEARCH ARTICLE

Annals of Data Analysis
https://periodicos.ufpe.br/revistas/ADA/

Submitted on: 12.12.2025.
Accepted on: 08.08.2026.
Published on: 18.08.2026.

Abstract

This study aimed to evaluate the forecasting of leak-prone plastic waste in the oceans (LPW) using artificial neural networks (ANN): Nonlinear Autoregressive (NAR); Nonlinear Autoregressive with Exogenous Inputs (NARX); Long Short-Term Memory (LSTM). The analysis focused on the 20 Brazilian municipalities with the highest LPW potential based on 2022 data. Methodologies included discrete wavelet transform, simple linear regression, Taylor diagrams, geospatial mapping, and others. The results indicated that São Paulo, Rio de Janeiro and the Southeast region exhibited the highest absolute LPW potential, followed by the Northeast and Central-West regions. In predictive modeling, the ANN-NARX achieved the highest accuracy, followed by the ANN-NAR and -LSTM. These findings can improve the monitoring and control of plastic waste leakage in the oceans.

Keywords

Marine pollution, waste management, artificial neural networks, wavelet analysis, Brazil.

1. Introduction

Global plastic production is estimated at approximately 400 million metric tons (Mt) annually, corresponding to an average of 57 kilograms per capita (Alencar et al., 2023). Brazil, the ninth largest economy and the sixth most population country, possesses an extensive coastline of about 8,000 km and a vast network of river basins, many of which discharge directly in the Atlantic Ocean. As the leading plastic producer in Latin America, generating approximately 7 Mt annually, Brazil plays a critical role in regional plastic flows. Consequently, it ranks 16th among global contributors to marine plastic leakage and the 20th in single-use plastic waste generation (Fraga Filho 2025; Guo et al., 2025).

According to the Monitor Mercantil (2023) approximately 46% of global plastic waste is disposed of in sanitary landfills, 17% is treated through controlled incineration, and 15% is recycled, indicating that 78% of plastic waste is properly managed. The remaining 22% is considered mismanaged plastic waste, totaling nearly 80 million metric tons annually (Fraga Filho, 2025). A substantial proportion of this mismanaged plastic waste ultimately reaches the marine environment. Plastics debris enters the oceans through multiple pathways, including improper disposal of solid waste on beaches, soils, and rivers; inefficiencies in waste collection systems; maritime activities such as fishing and shipping; and environmental disasters such as floods and storms (Barbosa et al., 2025; Velis et al., 2023; Alencar et al., 2023).

Brazil presents particularly concerning conditions in this regard (Barbosa et al., 2025; He et al., 2025). According to Fraga Filho (2025), the national solid waste recycling rate is only 4%, significantly lower than developed nations and also below to countries with similar economic development, such as Chile, Argentina, South Africa, and Turkey, which report an average recycling rate of 16%. The situation is even more critical regarding plastics: official data show that only about 1% of Brazil’s annual plastic production is recycled (Vergara et al., 2025).

Once in the oceans, plastic waste produced a broad range of detrimental effects, including threats to biodiversity, marine life, tourism, fisheries, aquaculture, and navigation. These environmental impacts are frequently accompanied by socioeconomic consequences, such as income loss, housing vulnerability, food insecurity, and public health risks (Pincelli et al., 2021; Meijer et al., 2021). Rivers serve as the primary pathways transporting plastic waste from land to sea, carrying debris through urban drainage systems, sewer networks, effluent discharges, and natural watercourses. These processes are often intensified by extreme weather events like heavy rainfall, flooding, and storms (2025Prakash & Zielinski, 2025; Islam, 2025).

In response to the urgent need for enhanced environmental monitoring and forecasting, artificial intelligence (AI) has become an important tool in addressing pollution (Prakash & Zielinski, 2025; Guo et al., 2025). A recent literature review by He et al. (2025) indicated that 57% of AI applications in the field of marine pollution are employed for monitoring, 24% for management, and 19% for prediction. Among the analytical methods applied in this context, wavelet techniques have gained increasing relevance, particularly for identifying coherence patterns between variables, preprocessing signals to reduce noise interference, and serving as components of hybrid models that incorporate artificial neural networks (Prakash & Zielinski, 2025; Guo et al., 2025; Vergara et al., 2025; Santos, 2024; Exame, 2023; Chamas et al., 2020).

According to the United Nations, plastic pollution is currently the second most significant threat to the global environment, surpassed only by climate change. In this context, the present study aims to statistically assess the predictive performance of artificial neural networks applied to municipal estimates of leak-prone plastic waste in the oceans, integrating discrete wavelet transform, Taylor diagrams, geospatial mapping, and complementary statistical metrics. By combining statistical, computational and spatial approaches, the study advances the understanding of patterns and drivers of marine plastic leakage in Brazil. The findings are expected to provide valuable insights into evidence-based policymaking and to support the design of targeted mitigation strategies at both local and regional scales.

2. Materials and methods

2.1. Study area

Figure 1 shows the map of Brazil, highlighting the 20 municipalities with the highest potential for plastic waste leakage in the oceans, organized by region: North (Manaus and Belém), Northeast (Teresina, Fortaleza, Natal, João Pessoa, Recife, Maceió, Aracaju and Salvador), Central-West (Brasília, Goiânia and Campo Grande), Southeast (Belo Horizonte, Contagem, Campinas, São Paulo and Rio de Janeiro), and South (Curitiba and Porto Alegre). The map was developed based on study from Alencar (2022).

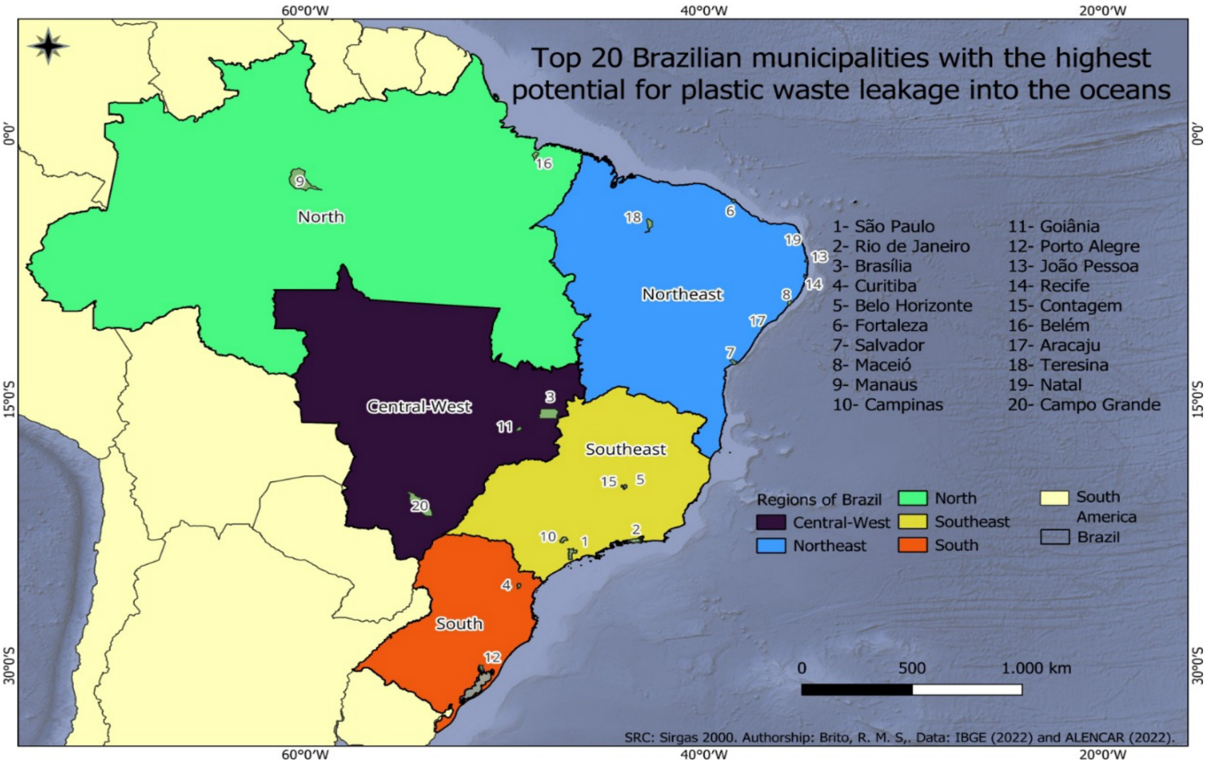

Fig. 1. Map of the study area showing Brazil subdivided by regions and highlighting the municipalities with the highest potential of leak-prone plastic waste in the oceans. Source: Alencar (2022).

2.2. Dataset

This study employs the dataset compiled by Alencar (2022), which estimated plastic waste leakage in the marine environment for all 5,570 Brazilian municipalities and emphasized the

20 with the highest leakage potential. Accordingly, the dataset analyzed in this research comprises these 20 Brazilian municipalities most susceptible to plastic waste leakage in the ocean in 2022. These data are referred to as leak-prone plastic waste in the oceans (LPW) with the unit of measurement being tons.year^{-1}. The estimates provided by Alencar et al. (2023) were generated by integrating demographic, environmental, and socioeconomic information with data on waste management and disposal practices, in order to quantify the amount of post-consumer plastic waste with inadequate final disposal potentially reaching the ocean at the municipal level.

The focus on a subset of municipalities was intentional and driven by the objective of analyzing plastic waste leakage dynamics in areas with the highest estimated contribution to marine pollution. Rather than aiming at statistical representativeness across all Brazilian municipalities, the study concentrates on locations classified as high risk according to Alencar (2022), where LPW processes are more pronounced and policy relevant. This targeted selection allows for a more detailed and meaningful evaluation of the modeling framework under critical conditions, while avoiding dilution effects that could arise from including municipalities with negligible leakage levels.

2.3. Discrete wavelet transform

The wavelet transform was employed as a preprocessing step to enhance the representation of the estimated LPW signal prior to its use as input in the artificial neural network. This choice is motivated by the inherently nonstationary and multiscale nature of LPW dynamics, which are influenced by socioeconomic, environmental, and infrastructural processes operating across different temporal scales. Traditional time or frequency domain approaches are limited in capturing such localized variations, whereas wavelet based methods allow simultaneous analysis in both time and frequency domains. The continuous wavelet transform (CWT) is defined by Eq. (1):

$$CWT(a,b) = \frac{1}{\sqrt{a}} \cdot \int x(t) \cdot \psi^{*}\left(\frac{t-b}{a}\right) dt \tag{1}$$

Where $\psi(t)$ denotes the mother wavelet function, characterized by finite duration, finite energy, compact support, and zero mean. The parameters a and b control the scale (dilation) and translation of the wavelet, respectively, enabling the detection of localized features and scale dependent patterns in the LPW signal (Santos, 2024). The term *x(t)* represents the LPW series, and (*) denotes the complex conjugate (Guo et al., 2025; He et al., 2025).

For computational efficiency and practical implementation within the neural network framework, the discrete wavelet transform (DWT, Eq. (2)) was adopted. The DWT is obtained by discretizing the scale and translation parameters of the CWT, such that $a = 2^j$ and $b = k2^j$ resulting in:

$$DWT(a,b) = \frac{1}{\sqrt{2^j}} \cdot \int x(t) \cdot \psi^{*}\left(\frac{t-k2^j}{2^j}\right) dt \tag{2}$$

Where *j* and *k* correspond to the scale and translation indices, respectively. The DWT enables a multiresolution decomposition of the LPW signal, separating low frequency components (long term trends) from high frequency components (short term fluctuations and noise). This decomposition improves the signal to noise ratio and facilitates the learning processes of natural networks by providing inputs that are more structured and physically interpretable.

In this study, the Symlets wavelet with four levels of decomposition was selected due to its near symmetry, good time and frequency localization properties, and robustness in representing complex environmental signals. This wavelet family has been widely applied in previous studies involving environmental and socioeconomic time series forecasting (Santos, 2024; Nascimento et al. 2023; Nuwairan et al., 2022). Recent methodological advances further support the use of DWT, including improvements to boundary treatment through polynomial fitting (Zhang, 2025) and comprehensive reviews highlighting the effectiveness of wavelet based approaches in analyzing nonstationary signals (Nguyen & Nguyen, 2024).

2.4. Artificial neural network

Given the nonlinear, dynamic, and nonstationary characteristics of the LPW series, deep learning models were adopted due to their ability to learn complex temporal dependences directly from data without requiring explicit assumptions about functional relationships. Recurrent neural network architectures are particularly suitable for this type of problem, as they incorporate memory mechanisms that allow past information to influence current predictions, which is essential for modeling environmental and socioeconomic processes characterized by delayed and cumulative effects.

In this study, the LPW forecasting models were developed using three recurrent neural networks architectures: (i) Nonlinear Autoregressive (NAR); (ii) Nonlinear Autoregressive with Exogenous Inputs (NARX); and (iii) Long Short-Term Memory (LSTM). Nar and NARX networks are well suited for nonlinear autoregressive modeling, with NARX explicitly incorporating exogeneous inputs to represent external influences. The LSTM architecture was included due to its enhanced capability to capture long term dependencies and mitigate the vanishing gradient problem through gated memory cells, making it particularly effective for complex and multiscale temporal patterns.

The complete dataset was divided in three mutually exclusive subsets: 70% for training, 15% for testing, and 15% for validation, following standard practices in time series forecasting to ensure robust model evaluation (Chen et al., 2025; Santos et al., 2022). The adoption of these networks was motivated by their proven effectiveness in modeling nonlinear and dynamic systems, particularly in environmental and socio-economic applications where data exhibit temporal dependencies and complex patterns.

According to Saraiva et al. (2021), the architecture of a neural network is defined by its topology, training algorithm, and nodes. Each nodes functions as a basic computational unit, inspired by biological neurons, that processes information locally. Artificial neurons calculate outputs by summing the weighted input vectors A with weights W_j, subtracting a bias term T_h, and applying an activation function $f(*)$ to the result (Guo et al., 2025; Izzaddin, 2024; Santos et al., 2020; Smyl, 2020). Accordingly, Eq. (3) is expressed as follows:

$$f\left(W_j \cdot A - T_{h_i}\right) = f\left(\sum_{i=1}^{n}\left(W_{ij} \cdot A_i - T_{h_i}\right)\right) \tag{3}$$

During the training phase, network parameters and weights were optimized using the Levenberg-Marquardt algorithm under a supervised learning framework. The training subset was used to iteratively adjust the model weights by minimizing the prediction error between observed and simulated LPW values. The testing subset was employed to monitor the generalization capability of the models and to prevent overfitting by evaluating performance on unseen data during training. Finally, the validation subset was used exclusively for an independent assessment of model performance after training was completed (He et al., 2025; Guo et al., 2025; Santos, 2024; Nascimento et al., 2023; Carmo, 2022).

Model performance was evaluated using statistical metrics commonly adopted in forecasting studies, including the mean square error (MSE) and root mean square error (RMSE). These metrics were computed separately for the training, testing, and validation phases to assess both the learning capacity and the generalization ability of each neural network.

The architectural configuration, including the number of hidden layers, neurons, delays, and activation functions, were defined based on previous studies and summarized in Table 1 (Guo et al., 2025; Santos, 2024; Nuwairan et al., 2022; Saraiva et al., 2021). Forecasts produced by the LSTM and NAR neural network are denoted as LSTM and NAR, both using LPW data as input and target. In contrast, the forecast from the NARX model is denoted as NARX, in which the LPW signal serves as the target variable, while the input consisted of LPW data preprocessed using the discrete wavelet transform, allowing the model to capture multi-scale temporal features.

Table 1. Configuration of the neural networks and parameters used during the training phase for LPW forecasting.

Parameters	NAR / NARX	LSTM
Input/target signals	LPW and LPW+DWT	LPW
Training algorithm	Levenberg-Marquardt	Levenberg-Marquardt
Performance metric	MSE	RMSE
Training set (%)	70	80
Test set (%)	15	20
Validation set (%)	15	-
Hidden layer(s)	1	5
Neurons in hidden layer	10	100
Activation function (hidden layer)	Sigmoid	Sigmoid / Hyperbolic Tangent
Output layer(s)	1	1
Activation function (output layer)	Linear	Linear
Neurons in output layer	1	1

2.5. Statistical analysis

Descriptive statistics were applied to characterize the LPW data and predicted data to provide a preliminary assessment of their distributional properties. The indicators considered were the mean (M, Eq. (4)), standard deviation (SD, Eq. (5)), coefficient of variation (CV, Eq. (6)), minimum (Min), first quartile (Q1), median (MD), third quartile (Q3), and maximum (Max). Together, these measures summarize central tendency, variability, and data dispersion, offering a comprehensive description of both datasets.

$$M = \bar{x}(t) = \mu = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i(t) \tag{4}$$

$$SD = \sigma = \sqrt{\frac{1}{N-1} \cdot \sum_{i=1}^{N} \left(x_i(t) - \bar{x}(t)\right)^2} \tag{5}$$

$$CV = \left(\frac{\sigma}{\mu}\right) \cdot 100\% \tag{6}$$

In these equations, $(\boldsymbol{N = 20})$ denotes the total number of observations, $(\boldsymbol{i})$ is the index ranging from 1 to *N*, $\boldsymbol{x_i(t)}$ corresponds to the values of the LPW datasets with their respective indices, and $\boldsymbol{\bar{x}(t)}$ is the mean value of the series.

To evaluate the robustness of the analyses given the limited sample size and the heterogeneous distribution of LPW values, complementary robustness checks were performed. These included the comparison between mean and median based statistics, the application of logarithmic transformation to LPW values, and a sensitivity analysis excluding the outlier municipalities.

From a statistical perspective, the adequacy of the dataset is ensured by the richness of the information used to estimate LPW and by the model evaluation strategy, which relies on out of sample validation and robustness analyses rather than on cross sectional inference.

To identify the relationship between the LPW data and your predicted, the scatter plot was combined with simple linear regression (SLR, Eq. (7)), the coefficient of determination (R^2, Eq. (8)), and Pearson's correlation coefficient (r, Eq. (9)). The statistical indicators applied to validate the neural network models (NAR, NARX, and LSTM) were the root mean square error (RMSE, Eq. (10)) and mean absolute percentage error (MAPE, Eq. (11)). The statistical metric RMSE shares the same unit of measurement as LPW, whereas MAPE is expressed as a percentage. The equations are shown below:

$$y(t) = a \cdot x(t) + b \tag{7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \left(x_i(t) - y_i(t)\right)^2}{\sum_{i=1}^{N} \left(x_i(t) - \bar{x}(t)\right)^2} \tag{8}$$

$$r = \frac{\sum_{i=1}^{N} \left(x_i(t) - \bar{x}(t)\right) \cdot \left(y_i(t) - \bar{y}(t)\right)}{\sqrt{\sum_{i=1}^{N} \left(x_i(t) - \bar{x}(t)\right)^2} \cdot \sqrt{\sum_{i=1}^{N} \left(y_i(t) - \bar{y}(t)\right)^2}} \tag{9}$$

$$RMSE = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} \left(y_i(t) - x_i(t)\right)^2} \tag{10}$$

$$MAPE = \frac{1}{N} \cdot \sum_{i=1}^{N} \left| \frac{y_i(t) - x_i(t)}{x_i(t)} \right| \cdot 100 \tag{11}$$

In this analysis, $y(t)$ and $x(t)$ represent the LPW and your predicted data, respectively, while (a) and (b) correspond to the slope and intercept coefficients. Here $y_i(t)$ represents the LPW values predicted by the neural networks, while $x_i(t)$ corresponds to the LPW values estimated by Alencar (2022). The coefficient of determination ranges from 0 to 1 and is employed to evaluate the goodness of the SLR fit between the LPW and your predicted data. Pearson's correlation coefficient complements this assessment by quantifying the strength and direction of the linear relationship between the variables. The classification of r values follow the criteria proposed by Hopkins (2009) (Table 2).

Table 2. Classification of Pearson's correlation coefficients (r) into positive and negative categories, with respective strength levels. Source: Hopkins (2009).

Correlation Negative	Correlation Positive	Classification
-0,1 a 0	0 a 0,1	Very Low
-0,3 a -0,1	0,1 a 0,3	Low
-0,5 a -0,3	0,3 a 0,5	Moderate
-0,7 a -0,5	0,5 a 0,7	High
-0,9 a -0,7	0,7 a 0,9	Very High
-1,0 a -0,9	0,9 a 1,0	Almost Perfect

The Taylor Diagram (Taylor, 2001) was employed to graphically assess and summarize the agreement between the predicted and estimated LPW values. This method integrates three statistical measures Pearson's correlation coefficient, root mean square error, and standard deviation in a single two-dimensional representation. The proximity of the predicted point to the estimated reference point in the diagram indicates the quality of model performance (Zhang, 2025; Tian and Chen, 2021).

To construct the Taylor Diagram, the estimated values $x(t)$ and the predicted values $y(t)$ were analyzed. The square root of the mean of the squared centered differences (E', Eq. (12)) was employed as a comparative metric. This formulation is conceptually similar to the law of cosines, which relates an internal angle of a triangle to the lengths of its sides $(a^2 = b^2 + c^2 - 2 \cdot b \cdot c \cdot cos\theta)$. Based on this analogy, Eq. (12) can be expressed as:

$$E'^2 = \sigma_{y(t)}^2 + \sigma_{x(t)}^2 - 2 \cdot \sigma_{y(t)}^2 \cdot \sigma_{x(t)}^2 \cdot r \tag{12}$$

Where $\sigma_{y(t)}$ and $\sigma_{x(t)}$ denote the standard deviations of the predicted and estimated data, respectively. This approach enables simultaneous correlation, variability and error, thereby providing a comprehensive assessment of model performance.

Taylor (2001) proposed a diagram that simultaneously represents four key statistics: E', $\sigma_{y(t)}$, $\sigma_{x(t)}$ and r, plotted over a half or quarter circle. The x- and y-axes display standard deviation values, with $\sigma_{x(t)}$ placed along the x-axis. The radial distance from the origin to the prediction point corresponds to $\sigma_{y(t)}$, while the azimuth angle from the origin to the prediction point reflects the correlation coefficient r. The distance between the estimated and predicted data points on the diagram represents the centered root mean square of E' (Izzaddin et al., 2024; Tian and Chen, 2021; Simão et al., 2020).

Finally, the map of the study area was developed using QGIS software, in its most recent version (v. 3.44.1 "Solothurn", released in July 2025). The Anaconda Navigator computational environment was used to perform descriptive statistical analyses, simple linear regression, and the visualization of results in geospatial maps subdivided by regions. The MATLAB® software (version R2021b) was employed for the implementation of predictive modeling techniques using artificial neural networks (NAR, NARX, and LSTM), discrete wavelet transform, as well as for the creation of various types of graphs, such as bar charts, boxplots and Taylor diagrams.

3. Results and discussion

Table 3 summarizes the descriptive statistics (DS) of the estimated LPW data reported by Alencar (2022) and the corresponding forecasts obtained using artificial neural network models (LSTM, NAR, and NARX). In addition to the original LPW dataset, descriptive statistics are also presented for the LPW sample excluding outliers associated with São Paulo and Rio de Janeiro, as well as for the natural logarithm transformed LPW data with and without outliers. These complementary analyses were performed to assess the robustness of the statistical properties and to examine the sensitivity of the results to data skewness and the influence of extreme observations.

The LPW presented a mean value of 57.033,65 tons.year^{-1}, considerably lower than the maximum value of 256.196,98 tons.year^{-1} recorded in São Paulo. The standard deviation of LPW (61.464,46) exceeded the mean, indicating substantial variability, high dispersion, and the possible presence of outliers in the LPW distribution. The coefficient of variation (%) was employed to compare datasets with different means, providing insights in the data homogeneity. The results indicated that LPW presented high heterogeneity. The statistical metrics of the first quartile (Q1), second quartile (median), and third quartile (Q3) for the LPW data revealed that the median values are closer to Q1 than to Q3, indicating a right-skewed distribution with positive skewness.

Among the predicted datasets, the NARX-ANN demonstrated central tendency and variability measure most closely aligned with the observed LPW values, followed by the NAR-ANN and LSTM-ANN. Both models underestimated the maximum LPW values, a result consistent with the findings of Santos (2024), who reported that ANN based predictions frequently underestimate observed extremes. The LSTM-ANN exhibited the highest degree of heterogeneity, which can be attributed to its low variability. While NARX appears more robust in capturing distributional properties of LPW, LSTM shows limitations in representing variability, highlighting the need for hybrid or ensemble approaches to improve predictive accuracy.

The robust analysis presented in Table 3 indicates that both the exclusion of extreme values and the application of the natural logarithm transformation substantially reduce data dispersion while preserving the central tendency of the LPW series. In particular, the close agreement between mean and median based measures across the alternative specifications suggests that the results are not driven by distributional asymmetry. The outliers excluded sample exhibits a marked decrease in variability, with the coefficient of variation dropping from over 100% in the original dataset to approximately 41%, highlighting the strong influence of extreme observations. The natural log-transformed datasets (NLT and OES-NLT) further enhance distributional stability, yielding low dispersion measures and closely aligned quartiles and medians. These results demonstrate that the statistical behavior of LPW remains consistent under different data treatments, supporting the robustness of the subsequent modeling and forecasting analyses.

Table 3. Descriptive statistics (DS) of LPW - leak-prone plastic waste in the oceans (tons.year^{-1}), LSTM - long short-term memory, NAR - nonlinear autoregressive, NARX - nonlinear autoregressive with exogenous inputs, OES - outliers excluded sample, and NLT - natural log-transformed data. Statistical metrics include mean (M), standard deviation (SD), coefficient of variation (CV (%)), minimum (Min), first quartile (Q1), median (MD), third quartile (Q3), and maximum (Max).

DS	LPW	LSTM	NAR	NARX	OES	NLT	OES-NLT
M	57.033,65	23.087,98	39.334,16	48.586,98	37.853,54	10,65	10,46
SD	61.464,46	786,20	27.556,10	31.941,90	15.629,88	0,70	0,42
CV (%)	107,77	3,41	70,06	65,74	41,29	6,57	4,02
Min	18.886,80	22.451,42	7.803,31	20.869,34	18.886,80	9,85	9,85
Q1	28.285,35	22.568,56	20.011,29	26.431,89	27.799,30	10,41	10,23
MD	33.052,19	22.882,69	28.210,72	39.365,26	31.540,91	10,95	10,36
Q3	57.059,91	23.231,49	56.333,86	59.872,19	56.237,96	10,24	10,94
Max	256.196,98	25.715,03	128.901,50	135.640,50	66.155,76	12,45	11,10

Fig. 2 shows the bar chart comparing the estimated data of plastic waste likely to leak into the ocean and the predictions made by artificial neural networks, highlighting the spatial variability of data for each municipality. Firstly, it is observed that none of the ANNs (NAR, NARX, and LSTM) were able to reproduce the extreme LPW values for the municipalities of São Paulo and Rio de Janeiro, especially the LSTM model, which showed the greatest underestimation. The NARX model came closest to these extreme values, followed by the NAR model. This result is consistent with the findings of Santos (2024), who reported that most ANN-based predictions tend to underestimate observed values.

The municipality of Manaus was also significantly underestimated by all ANNs, which may be due to its central position within the sample. The NARX model showed the highest overestimation values, especially for the municipalities of João Pessoa and Porto Alegre. Both the NAR and NARX models were able to extrapolate the estimated LPW values effectively, successfully reproducing the gradual decrease in plastic waste likely to leak into the ocean, particularly from the municipality of Brasília onwards. The exceptions were São Paulo and Rio de Janeiro, which exhibited extreme values.

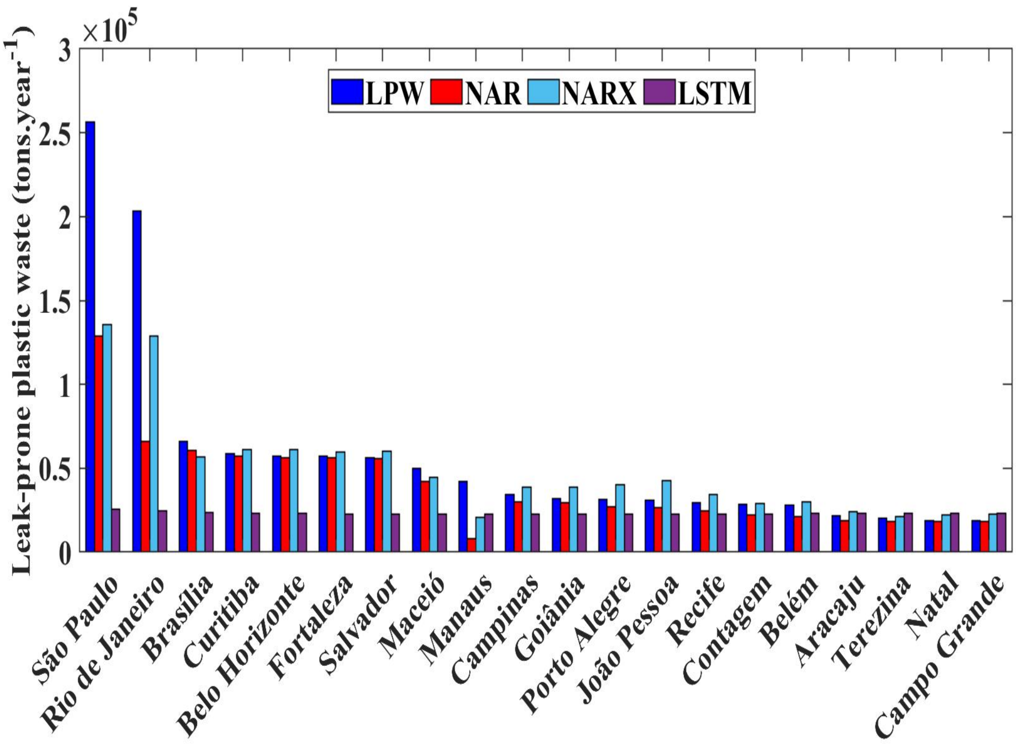

Fig. 2. Bar chart showing the intensity of the estimated data of plastic waste prone to leakage into the ocean and the corresponding forecasts using artificial neural networks.

Fig. 3 shows the boxplots of the estimated data of plastic waste prone to leakage into the ocean and the corresponding forecasts using artificial neural networks. This type of graph indicates the presence or absence of outliers, and it can be observed that all datasets exhibited outlier values. The estimated LPW data showed the highest outliers for the municipalities of São Paulo (256,196.98) and Rio de Janeiro (203,112.15). The NARX and LSTM neural networks also presented outliers for São Paulo (135,640.50 and 25,715.03, respectively) and Rio de Janeiro (128,794.23 and 24,436.70, respectively). The NAR neural network presented only one outlier for the municipality of São Paulo (128,901.50).

The boxplot of the NAR neural network showed the highest interquartile range (IQR) of 36,322.57, followed by the IQR of LPW at 29,007.65 and the IQR of the NARX neural network at 26,431.89. The LSTM neural network presented the smallest interquartile range, with only 662.94. This type of graph also allows for the visualization of positive skewness with a long tail to the right in all datasets. This is due to the median values being closer to the first quartile than to the third quartile, an observation previously discussed in the descriptive statistics for the estimated LPW and IBGE data and now confirmed for the neural network forecasts as well.

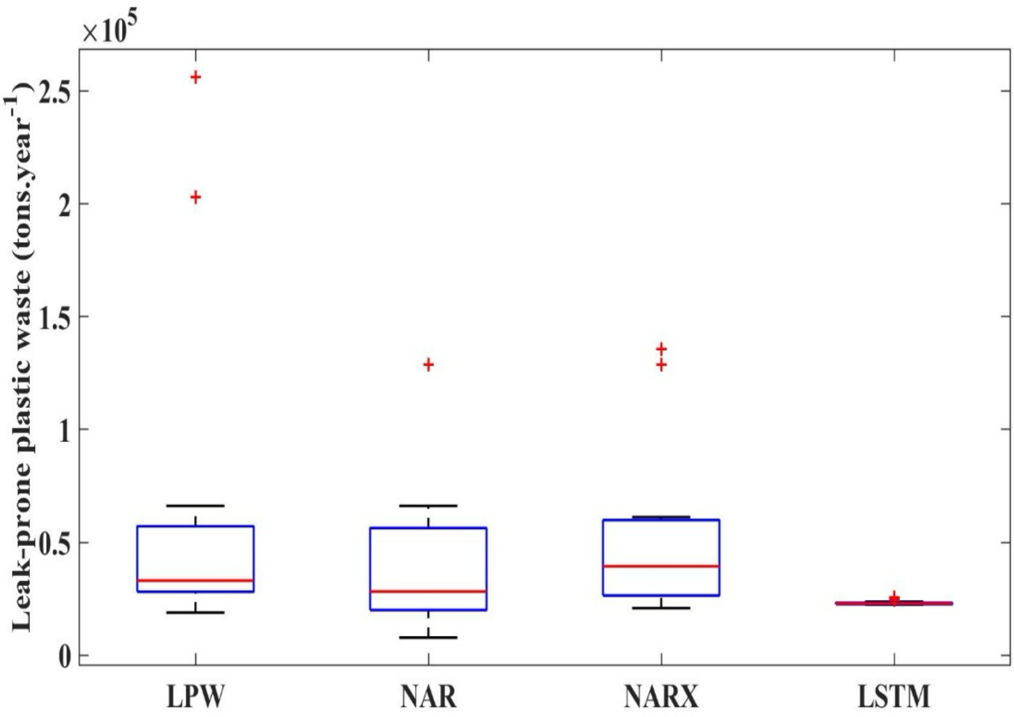

Fig. 3. Boxplot of the estimated plastic waste prone to ocean leakage and the forecasts of these data using artificial neural networks.

Figure 4 presents the simple linear regression between the estimated plastic waste prone to ocean leakage and the data predicted using artificial neural networks (LSTM, NAR, and NARX). This figure also includes additional information such as Pearson's correlation coefficient, coefficient of determination, root mean square error and mean absolute percentage error. All predictions showed an excellent linear fit with a positive slope, Pearson's correlation coefficients ranging from 0.86 to 0.96, classified as very high to nearly perfect according to Table 2 (Hopkins, 2009), and coefficients of determination explaining between 75% and 92% of the relationship.

Figure 4A displays the regression between LPW and LSTM, showing a more dispersed relationship. Consequently, it presented the weakest statistical metrics, such as RMSE = 30,203.21 and MAPE = 77.69%. The predictions performed by the NARX neural network (Figure 4C) and the NAR neural network (Figure 4B) showed data points closer to the linear fit and thus achieved better statistical performance, with RMSE between 16,897.42 and 29,441.94, and MAPE between 32.42% and 31.10%, respectively. Therefore, the initial statistical indicators suggest that the NARX neural network achieved the best forecasting performance for the plastic waste prone to ocean leakage.

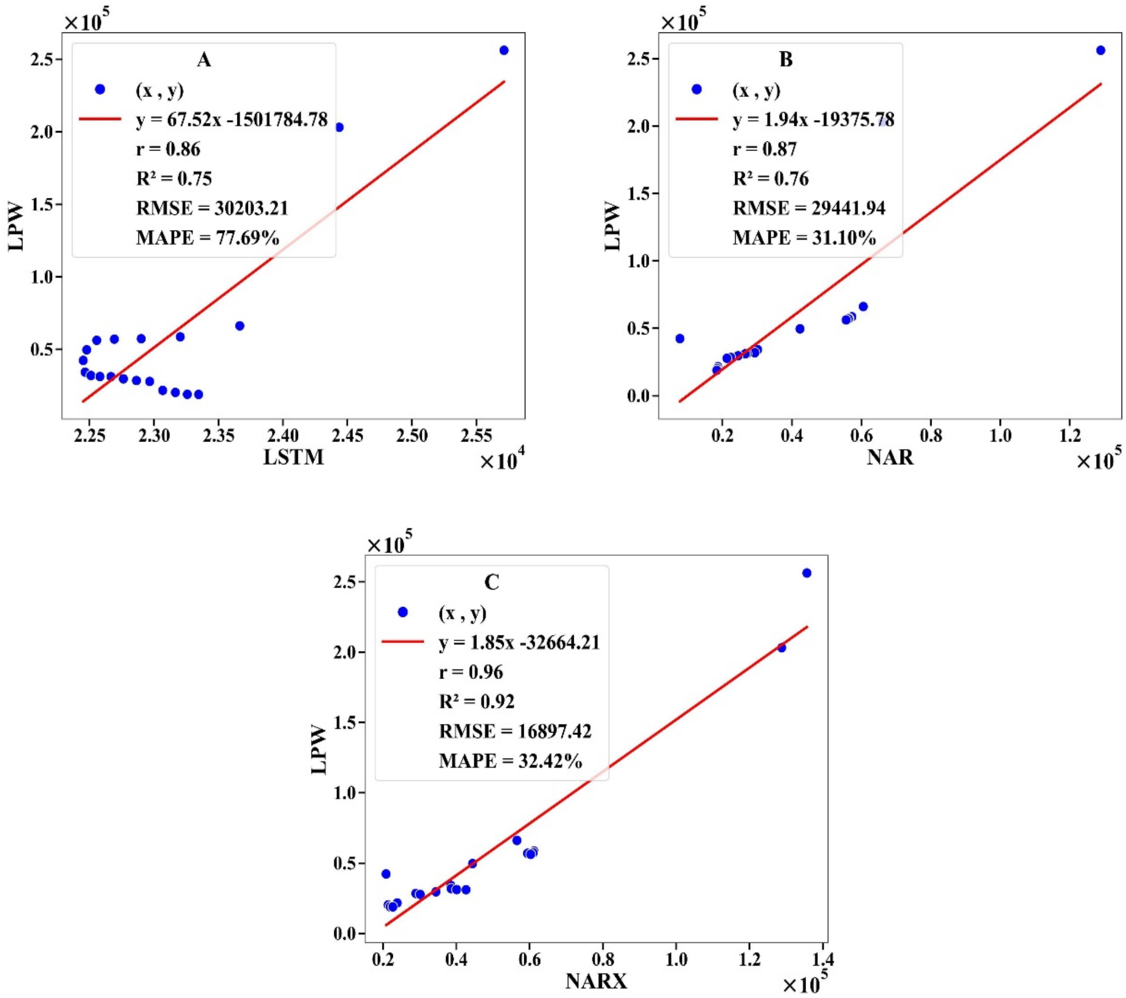

Fig. 4. Simple linear regression between the estimated LPW data and the predicted values generated by neural networks, including additional information on linear fit, Pearson correlation, coefficient of determination, root mean square error, and mean absolute percentage error. LPW vs LSTM (A), NAR (B), and NARX (C).

Another way to evaluate the best forecasting performance is through the graphical representation of the Taylor Diagram (Fig. 5). This diagram displays two dimensional points that represent the statistical metrics between the estimated LPW data, and the respective predictions made by the artificial neural networks NAR, NARX, and LSTM. The closer the predicted points (via ANN) are to the LPW reference point, the better the forecasting performance (Santos, 2024).

Overall, the prediction made by the NARX neural network was the closest to LPW and consequently showed the best forecasting performance compared to the other neural networks used. The LSTM neural network deviated the most from LPW, presenting the weakest forecasting performance. The NAR neural network showed the second-best performance, being closer to NARX than to LSTM.

Based on the Taylor Diagram, it is possible to observe that all neural networks NARX (31,941.90), NAR (27,556.10), and LSTM (786.20), presented variabilities significantly lower than the variability of the estimated LPW data (61,464.46) (Table 3). Therefore, the Taylor

Diagram also indicates that the NARX neural network exhibited the best predictive performance for estimating plastic waste prone to ocean leakage, followed by the NAR and LSTM networks, respectively.

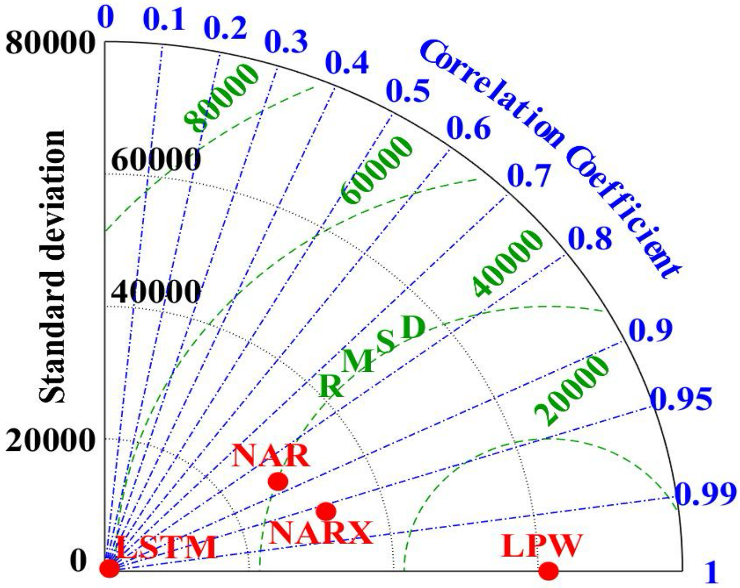

Fig. 5. Taylor diagram applied to the estimated data of plastic waste likely to leak into the ocean and the predicted data obtained through artificial neural networks.

Table 4 presents the statistics of the accumulated total and percentage of the accumulated total by region for plastic waste prone to ocean leakage and their respective forecasts using ANNs (LSTM, NAR, and NARX). The accumulated total of LPW was 1,140,672.93 tons·year^{-1}, indicating that Brazil contributes over one million tons of plastic waste prone to ocean leakage annually. The Southeast region recorded the highest accumulated LPW, with 579,239.47 tons·year^{-1}, representing more than 50% of the total. The Northeast had the second highest accumulated total, with 284,621.89 tons·year^{-1} (24.95%), followed by the Central-West (116,937.03 tons·year^{-1}), South (89,745.14 tons·year^{-1}), and North (70,129.40 tons·year^{-1}) regions, accounting for 10.25%, 7.87%, and 6.15% of the total, respectively.

Regarding the forecasts, the NARX neural network showed the closest regional totals, overall accumulated total, and percentage of the accumulated total compared to LPW, followed by the NAR and LSTM neural networks. The accumulated totals from all three neural networks underestimated LPW, with NARX showing the smallest underestimation and LSTM the largest. The NARX and NAR networks presented the highest accumulated values for the Southeast and Northeast regions, respectively. Only the LSTM model presented the highest accumulated value for the Northeast, which can be attributed to the fact that this region has the largest number of municipalities and that the LSTM model exhibited the lowest data variability.

Table 4. Statistics of TA - total accumulated ($tons.year^{-1}$) e PTA - percentage of the total accumulated (%) of LPW - leak-prone plastic waste ($tons.year^{-1}$), and respective forecasting via artificial neural network, LSTM - Long and Short Term Memory, NAR - Nonlinear Autorregressiva and NARX - Nonlinear Autorregressiva with eXogenous inputs, in the Brazilian regions (North, Northeast, Central-West, Southeast and South).

LPW	TA-LPW	PTA-LPW
North	70,129.40	6.15
Northeast	284,621.89	24.95
Central-West	116,937.03	10.25
Southeast	579,239.47	50.78
South	89,745.14	7.87
Total	1,140,672.93	100

LSTM	TA-LSTM	PTA-LSTM
North	45,418.34	9.84
Northeast	182,650.98	39.56
Central-West	69,522.54	15.06
Southeast	118,382.68	25.64
South	45,785.08	9.92
Total	461,759.62	100

NAR	TA-NAR	PTA-NAR
North	29,128.31	3.70
Northeast	260,902.69	33.16
Central-West	108,299.09	13.77
Southeast	304,010.53	38.64
South	84,342.49	10.72
Total	786,683.11	100

NARX	TA-NARX	PTA-NARX
North	51,043.85	5.25
Northeast	308,627.08	31.76
Central-West	117,874.73	12.13
Southeast	392,889.22	40.43
South	101,304.64	10.43
Total	971,739.52	100

Fig. 6 graphically represents Table 4, showing the geospatial map of Brazil subdivided into the North, Northeast, Central-West, Southeast, and South regions, illustrating the intensity of the total accumulated plastic waste prone to ocean leakage and its respective forecasts via artificial neural networks (LSTM, NAR, and NARX). The map in Fig. 6A highlights how the Southeast and Northeast regions stood out from the others, showing a high total accumulation. This is mainly due to São Paulo and Rio de Janeiro, which presented extreme LPW values in the Southeast region. In contrast, the Northeast region stood out because almost all of its state capitals exhibited high LPW values.

The map in Fig. 6B shows the total accumulated forecast produced by the LSTM neural network, which displayed more underestimated values, and the lowest data variability compared to the other forecasts. The forecast generated by the NAR neural network (Fig. 6C) also showed underestimated values, but with higher variability. The best performance was achieved by the NARX neural network, as shown in the map in Fig. 6D, as it was the only forecast that presented overestimated values for the Northeast, Central-West, and South regions, and it was the one that least underestimated the accumulated totals for the North and Southeast regions.

In summary, the geospatial representation of LPW and its forecasts highlights clear regional disparities across Brazil. The Southeast and Northeast regions consistently emerge as the most critical zones, driven by large urban centers and high population density. Among the forecasting models, the NARX neural network provided the most consistent and realistic spatial predictions, capturing both underestimations and occasional overestimations, which aligns more closely with the actual distribution of plastic waste leakage potential. These visual insights reinforce the importance of region-specific strategies and underscore the need for targeted public policies in areas with higher vulnerability to marine plastic pollution.

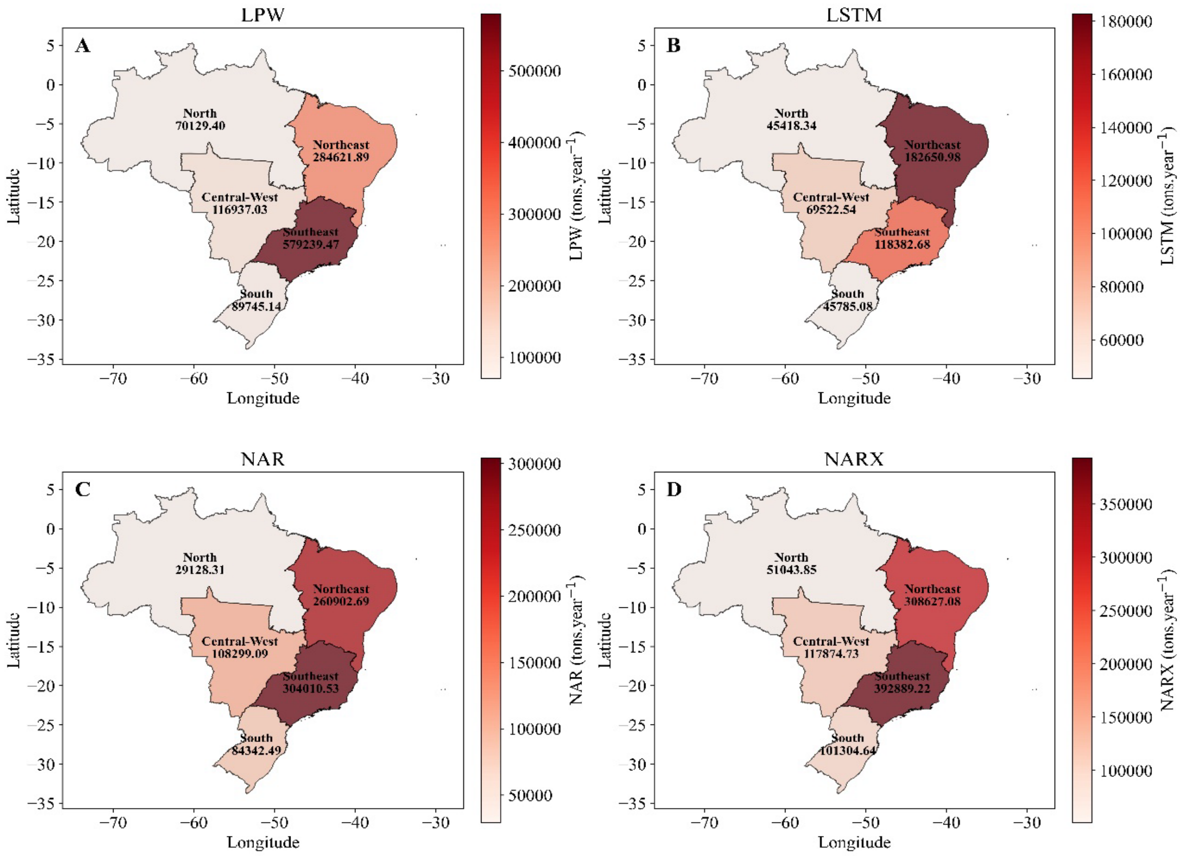

Fig. 6. Geospatial map of Brazil subdivided into the North, Northeast, Central-West, Southeast, and South regions, showing the intensity of the total accumulated LPW (A) and the forecasts generated by ANNs: LSTM (B), NAR (C), and NARX (D).

4. Conclusions

Millions of tons of mismanaged plastic waste are prone to leaking into the oceans each year, resulting in significant environmental, ecological, and socioeconomic impacts. In this study, the prediction of plastic waste leakage into the marine environment using artificial neural networks, as well as the analysis of its relationship with municipal socioeconomic variables, was successfully carried out. The main findings include.

Municipalities with the highest leakage potential: São Paulo and Rio de Janeiro showed the highest absolute potential for plastic waste leakage into the oceans, being the only municipalities with values exceeding 200,000 tons per year. As a result, the Southeast region presented the highest potential contribution to this issue, followed by Northeast and Central-West regions.

Performance of artificial neural networks: Error statistics metrics, Taylor diagrams, and bar plots indicated that the NARX neural network showed the highest accuracy in predicting LPW, followed by the NAR network, which exhibited the greatest variability, and the LSTM network, which demonstrated the lowest variability. However, all neural networks struggled to predict the extreme values for the municipalities of São Paulo and Rio de Janeiro.

Although the analysis is restricted to 20 municipalities, the conclusions are framed within the context of high leakage urban centers and do not seek to generalize results to all Brazilian municipalities. The findings are therefore statistically consistent with the study's scope and contribute methodological insights applicable to other regions facing similar waste management challenges.

Finally, the results indicated that the total accumulated LPW was 1,140,672.93 tons·year^{-1}, revealing that Brazil contributes more than one million tons of plastic waste prone to ocean leakage per year. The Southeast and Northeast regions consistently stood out as the most critical zones, reinforcing the importance of region-specific strategies. These findings may significantly contribute to the development of more effective public policies, supporting the management and monitoring of plastic waste with the potential to leak into marine ecosystems.

Acknowledgements

The authors gratefully acknowledge the support of the Brazilian funding agencies that made this research possible, including the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). This research was also supported by FAPESP (process 2023/01542-3), I.G.S.S. was supported by FAPESP (process 2024/06927-3), M.S.C. was supported by FAPEAL (process E:60030.0000000320/2023), and D.C.Z. is supported by CNPq (process 314837/2025-6).

We appreciate the assistance from Universidade de São Paulo (USP), Universidade Federal de Campina Grande (UFCG), Universidade Federal de Alagoas (UFAL), Instituto de Astronomia, Geofísica e Ciências Atmosféricas (IAG), Instituto de Ciências Atmosféricas (ICAT), Instituto de Matemática (IM), Instituto de Geografia, Desenvolvimento e Meio Ambiente (IGDEMA) Laboratório de Meteorologia Aplicada e Meio Ambiente (LAMMA), and the team EduWellTech.

Conflict of Interest Declaration

The authors have no conflicts of interest to declare. All co-authors have seen and agree with the contents of the manuscript and there is no financial interest to report.

References

1. Alencar, M.V., 2022. Improving the source-to-sea approach for marine litter in Brazil. PhD thesis in Oceanography, University of São Paulo, São Paulo. Available at: https://www.teses.usp.br/teses/disponiveis/21/21134/tde-19092022-144406/publico/Dissertacao_Alencar_Melanie_Corrigida.pdf. Accessed 30 January 2025.
2. Alencar, M.V., Gimenez, B.G., Sasahara, C., Elliff, C.I., Velis, C.A., Rodrigues, L.S., Conti, L.A., Gonçalves-Dias, S.L.F., Cetrulo, T.B., Scrich, V.M., Turra, A., 2023. Advancing plastic pollution hotspotting at the subnational level: Brazil as a case study in the Global South. *Mar. Pollut. Bull.* 194, 115382. https://doi.org/10.1016/j.marpolbul.2023.115382.
3. Barbosa, A., De Castro Gonçalves, T., Simões, P., 2025. Improving municipal solid waste services: Insights into efficiency, productivity, and recycling in Brazil. *Sustainability* 17, 2519. https://doi.org/10.3390/su17062519.
4. Carmo, N.R.M., 2022. Uso de redes neurais artificiais para previsão de GHI. *Rev. Virtual Quím.* 14, 56–60. https://doi.org/10.21577/1984-6835.20220013.
5. Chamas, A., Moon, H., Zheng, J., Qiu, Y., Tabassum, T., Jang, J.H., Abu-Omar, M., Scott, S.L., Suh, S., 2020. Degradation rates of plastics in the environment. *ACS Sustainable Chemistry & Engineering* 8, 3494–3511. https://doi.org/10.1021/acssuschemeng.9b06635.
6. Exame, 2023. Poluição plástica: 350 milhões de toneladas viram resíduos a cada ano, diz Credit Suisse. Available at: https://exame.com/esg/impacto-da-poluicao-plastica-no-mundo-e-medido-em-estudo-do-credit-suisse. Accessed 6 March 2025.
7. Fraga Filho, C.A.D., 2025. Analysis of Brazilian plastic waste management in the global context and case study of the city of Vitória, Espírito Santo. *Environmental Science and Pollution Research* 32, 5645–5684. https://doi.org/10.1007/s11356-024-35865-4.
8. Guo, Q., He, Z., Wang, Z., 2025. Assessing the effectiveness of long short – term memory and artificial neural network in predicting daily ozone concentrations in Liaocheng City. *Scientific reports*, 15, 6798. https://doi.org/10.1038/s41598-025-91329-w.
9. He, Z., Guo, Q., Wang, Z., Li, X., 2025.A Hybrid wavelet – based deep learning model for accurate prediction of daily surface PM2.5 concentrations in Guangzhou. *Toxics*, 13, 254. https://doi.org/10.3390/toxics13040254.
10. Hopkins, W.G., 2009. A new view of statistics: correlation coefficient. New York: Internet Society for Sport Science. Available at: http://www.sportsci.org/resource/stats/correl.html (accessed 18 January 2025).
11. Islam, F., 2025. The effects of plastic and microplastic waste on the marine environment and the ocean. *Eur. J. Environ. Earth Sci.* 6, 508. https://doi.org/10.24018/ejgeo.2025.6.3.508.
12. Izzaddin, A., Langousis, A., Totaro, V., Yaseen, M., & Iacobellis, V., 2024. A new diagram performance evaluation of complex models. *Stochastic Environmental Research and Risk Assessment*. https://doi.org/10.1007/s00477-024-02678-3.
13. Meijer, L.J.J., Emmerik, T.V., Ent, R.V.D., Schmidt, C., Lebreton, L., 2021. More than 1000 rivers account for 80% of global riverine plastic emissions into the ocean. *Science Advances* 7, eaaz5803. Available at: https://www.science.org/doi/pdf/10.1126/sciadv.aaz5803 (accessed 11 July 2025).

14. Monitor Mercantil, 2023. O Brasil é um dos que mais produz lixo no mundo; apenas 4% dos resíduos sólidos são reciclados. Available at: https://monitormercantil.com.br/brasil-e-um-dos-que-mais-produzem-lixo-no-mundo-so-4-dos-residuos-solidos-so-reciclados (accessed 06 March 2025).
15. Nascimento, E.G.S., Melo, T.A.C., Moreira, D.M., 2023. A transformer-based deep neural network with wavelet transforms for forecasting wind speed and wind energy. *Energy* 278, 127678. https://doi.org/10.1016/j.energy.2023.127678.
16. Nguyen, T.H., Nguyen, H.T., 2024. Advances in wavelet transform methods for non-stationary signal analysis: A review. *Signal Processing*, 215, 110145. https://doi.org/10.1016/j.sigpro.2024.110145.
17. Nuwairan, M.A.L., Sabir, Z., Raja, M.A.Z., Aldhafeeri, A., 2022. An advance artificial neural network scheme to examine the waste plastic management in the ocean. *AIP Advances* 12, 045211. https://doi.org/10.1063/5.0085737.
18. Pincelli, I., De Castilhos Júnior, A., Matias, M., Rutkowski, E., 2021. Post-consumer plastic packaging waste flow analysis for Brazil: The challenges moving towards a circular economy. Waste Manag. 126, 781–790. https://doi.org/10.1016/j.wasman.2021.04.005.
19. Prakash, N., Zielinski, O., 2025. AI-enhanced real-time monitoring of marine pollution: Part 1 – A state-of-the-art and scoping review. *Front. Mar. Sci.* 1486615. https://doi.org/10.3389/fmars.2025.1486615.
20. Prakash, N., Zielinski, O., 2025. AI-enhanced real-time monitoring of marine pollution: Part 2 — A spectral analysis approach. *J. Mar. Sci. Eng.* 13, 636. https://doi.org/10.3390/jmse13040636.
21. Santos, I.G.S., 2024. Estatísticas de cinco variáveis meteorológicas no município de Craíbas no Estado de Alagoas. Tese (Doutorado em Meteorologia), Universidade Federal de Campina Grande, Paraíba. Available at: http://dspace.sti.ufcg.edu.br:8080/jspui/handle/riufcg/38015 (accessed 09 March 2025).
22. Santos, I.G.S., Oliveira-Júnior, J.F., Sousa, F.A.S., Barbosa, I.I., 2022. Previsão de focos de calor na região metropolitana de Maceió utilizando rede neural artificial. *Revista Brasileira de Geografia Física* 15(5), 2313–2326. https://doi.org/10.26848/rbgf.v15.5.p2313-2326.
23. Santos, I.G.S., Lyra, R.F.F., Silva-Júnior, R.S., 2020. Comparativo de prognósticos da velocidade do vento utilizando o modelo WRF e rede neural artificial. *Revista Brasileira de Meteorologia* 35(Special issue), 1017–1027. https://doi.org/10.1590/0102-77863550103.
24. Saraiva, S.V., Carvalho, F.O., Santos, C.A.G., Barreto, L.C., Freire, P.K.M.M., 2021. Daily streamflow forecasting in Sobradinho Reservoir using machine learning models coupled with wavelet transform and bootstrapping. *Applied Soft Computing* 102, 107081. https://doi.org/10.1016/j.asoc.2021.107081.
25. Simão, M., Videiro, P., Silva, P., De Freitas Assad, L., Sagrilo, L., 2020. Application of Taylor diagram in the evaluation of joint environmental distributions' performances. *Marine Systems & Ocean Technology*, 15, 151–159. https://doi.org/10.1007/s40868-020-00081-5.

26. Smyl, S., 2020. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2019.03.017.

27. Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research* 126, 7183–7192. https://doi.org/10.1029/2000JD900719.

28. Tian, Z., Chen, H., 2021. Multi-step short-term wind speed prediction based on integrated multi-model fusion. *Applied Energy* 298, 117248. https://doi.org/10.1016/j.apenergy.2021.117248.

29. Velis, C.A., Wilson, D.C., Gavish, Y., Grimes, S.M., Whiteman, A., 2023. Socio-economic development drives solid waste management performance in cities: A global analysis using machine learning. *Science of the Total Environment* 872, 161913. https://doi.org/10.1016/j.scitotenv.2023.161913.

30. Vergara, D., De La Hoz-M, J., Ariza-Echeverri, E., Fernández-Arias, P., Antón-Sancho, Á., 2025. Evaluating solutions to marine plastic pollution. *Environments* 12, 86. https://doi.org/10.3390/environments12030086.

31. Zhang, B., 2025. Application of the Taylor Diagram in evaluating the performance of IOL formulas. *J. Refract. Surg.* 41 (1), e50–e55. https://doi.org/10.3928/1081597X-20241126-04.