A Unified Benchmark for Multi-Horizon Outdoor Temperature Forecasting: Comparing Baseline, Statistical, and Machine Learning Models

Rudy Espinoza-Nima*, Rolyn Chuquihuanca-Carhuapoma, Darwin Vilcherrez-Vilela, José Hidalgo-Sullón and Víctor Ramírez-Ordinola

Universidad Nacional de Piura, Piura, Peru

Abstract

Currently, short-term outdoor temperature prediction faces difficulties associated with the quality of meteorological records, data processing, and differences in assessment procedures, which limit objective comparison between models and make it difficult to determine whether a predictive improvement responds to the algorithm used. Faced with this problem, the study aimed to compare reference, statistical, and machine learning models to predict temperature at horizons of 1, 3, 6, 12, and 24 hours. A total of 51 874 records from an automatic weather station at the National University of Piura, Peru, were analyzed. The data were normalized, subjected to physical range controls, aggregated by hour and processed against short periods of missing information. Subsequently, they were divided chronologically into 70% for training, 15% for validation and 15% for testing, avoiding incorporating future information into the predictors. Persistence, seasonal naïve, moving average, linear regression, Ridge, SARIMAX, XGBoost and a multilayer neural network model were compared, along with statistical tests and robustness analysis. Meanwhile, XGBoost obtained the lowest errors in the five horizons evaluated. Variant 3A reached RMSE values of 0.5991 °C and 1.1752 °C at 1 and 6 hours, while variant 3B recorded 0.9521 °C, 1.1972 °C and 1.2323 °C at 3, 12 and 24 hours. Therefore, it is concluded that rigorous temporal processing favors reliable predictions and allows establishing a reproducible procedure to compare different families of models under equivalent conditions.

Keywords

Temperature Forecasting, Machine Learning, XGBoost, Time Series,, Meteorological Data

1. Introduction

The Piura region, located on the northern coast of Peru, presents a marked meteorological variability throughout the year and, periodically, experiences alterations of greater intensity associated with the interaction between the Pacific Ocean and the tropical atmospheric circulation. The warm episodes of the eastern Pacific can modify the temperature, humidity and rainfall regime in the coast and mountains of Piura, with effects on agriculture, urban infrastructure, transport and risk management. The events recorded in 1925 and 2017 show the magnitude that these alterations can reach in northern Peru [1, 2]. This scenario makes it necessary to have reliable meteorological information that allows us to understand the behavior of atmospheric variables and anticipate their changes. Among them, temperature acquires special relevance due to its temporal variation and the possibility of analysing its evolution from historical records.

Therefore, the need to anticipate these variations has driven the development of different prediction methods ranging from statistical models to machine and deep learning techniques. Recent research has compared SARIMA, XGBoost, and deep networks [3], evaluated LSTM networks for hourly forecasts [4] and combined LSTM and GRU architectures optimized for daily forecasting [5]. In addition to these

IWSAI 2026: 1st International Workshop on Systems, Applications, Artificial Intelligence and Innovation, September 16–18, 2026, Piura, Peru

*Corresponding author.

respinozan@unp.edu.pe (R. Espinoza-Nima); rolynsmit97@gmail.com (R. Chuquihuanca-Carhuapoma); dvilcherrezv@unp.edu.pe (D. Vilcherrez-Vilela); 1302023041@alumnos.unp.edu.pe (J. Hidalgo-Sullón); vramirezo@unp.edu.pe (V. Ramírez-Ordinola)

0000-0002-5602-8921 (R. Espinoza-Nima); 0009-0008-9026-2149 (R. Chuquihuanca-Carhuapoma); 0000-0001-7632-6136 (D. Vilcherrez-Vilela); 0009-0005-5903-2351 (J. Hidalgo-Sullón); 0000-0002-7749-9247 (V. Ramírez-Ordinola)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-09-14

approaches, there are models based on Transformers, used to represent temporal dependencies on data [6], while XGBoost has shown favorable results in correcting short-term weather forecast errors [7]. In this context, the aforementioned advances show that there are multiple alternatives to address temperature prediction. However, the performance achieved may vary depending on the characteristics of the records, the time scale, the prediction horizon, and the way in which the data are prepared and evaluated. For this reason, the results obtained in other contexts or scales do not necessarily represent the behavior of a specific weather station.

This limitation is particularly important in Piura, where research has also been carried out aimed at the computational analysis of climate variables. Jiménez-Carrión et al. [8] and Peña-Cáceres et al. [9] they used artificial neural networks to model the "El Niño" phenomenon using climatic variables and used information related to the same station considered in this research. In addition, the regional need to strengthen the generation and use of information that contributes to climate risk planning and management has been pointed out [10]. These antecedents show the value of the meteorological data available in the region, but leave open the need to specifically study the hourly prediction of the outside temperature through models of different complexity evaluated on the same time series. On this basis, the present study aims to compare reference, statistical and machine learning models to predict the outdoor temperature at horizons of 1, 3, 6, 12 and 24 hours using records from an automatic weather station of the National University of Piura. The main contribution lies in establishing a temporal, progressive and reproducible comparison framework that evaluates the models under common criteria and allows determining whether the increase in complexity translates into consistent predictive improvements.

2. Research Questions

Based on the problem identified and to guide the comparison of the models under equivalent conditions, the following research questions were raised:

- How does the performance of reference, statistical, and machine learning models vary when predicting the outdoor temperature of the National University of Piura's weather station for horizons of 1, 3, 6, 12, and 24 hours?
- To what extent does the integration, filtering and structuring of meteorological records allow consistent data to be available for the construction and evaluation of predictive models?
- What contribution do current meteorological variables, their legs, and cyclical temporal components have to the predictive performance of the models?
- Are there statistically significant differences between the performance of the models when evaluated using chronological division and common metrics?
- How robust and stable is the performance of the models under different evaluation time periods and under different initializations of the multilayer neural network?
- How do collinearity and the elimination of predictor groups influence the predictive capacity of the models evaluated?

3. Methodology

The study had a quantitative approach applied to time series, dividing the work into six stages, which began with determining the data sources, controlling their quality, matching the hours, creating the predictor variables, defining the prediction horizons and, finally, comparing the models. Each stage was contrasted with the scripts provided, so that the description presented below corresponds to the operations carried out from the same meteorological station.

In the first stage, it consisted of integrating records from a historical TXT file and an Excel file with complementary records from the same weather station. Where headers, date formats, units and numerical types were reviewed, the record of the Excel file was preserved, as it contains more recent data. Data were ordered chronologically before any lag construction. In the second stage, range physical tests were applied. Values incompatible with the established limits were marked as missing, without automatically replacing them with extreme or average values. This criterion follows the control and traceability recommendations for automatic stations such as the World Meteorological Organization (WMO) or by national bodies such as the National Meteorology and Hydrology Service of Peru (SENAMHI) [11]. Because of this, a piece of data was not eliminated just because it was statistically extreme without exceeding the limits established considering the physical tests of rank.

The third stage consisted of time aggregation. Continuous variables were summarized by averages, maximums and minimums retained their corresponding operator, and precipitation accumulated within each hour. Indoor shortages of up to three hours were interpolated in temperature, humidity, pressure, wind speed, solar radiation and UV index; The longer gaps were not taken into account to extrapolate and estimate a value of a variable between two hours close to each other. The missing precipitation was assigned to zero only when other variables confirmed that the station had recorded that time.

In the fourth stage, the input variables were constructed: seven current measurements, eight lags per variable and four cyclical components. In the fifth stage, an independent objective variable was generated by training the models in each horizon. Rows that remained incomplete after creating lags and objectives were discarded entirely, avoiding artificially completing data in each section.

The sixth stage compared the models under a 70/15/15 time split. The validation set was used to select hyperparameters; The chosen configurations were then retrained with training plus validation and evaluated only once on test. As an additional check, three expansive temporary partitions were executed prior to the final test set [12, 13]. For the review, Diebold-Mariano tests were performed on quadratic losses, confidence intervals using a mobile bootstrap, an ablation analysis for XGBoost 3B, explicit adjustment of the Ridge parameter, repetition of the MLP with three seeds and a SARIMAX analysis that uses the meteorological variables available at the time of forecast. These extensions ran on the same partitions, and their output tables were preserved, so that each new result can be checked against the code and prediction files.

Persistence, 24-hour seasonal naïve, 6-hour moving average, linear regression, Ridge, SARIMAX, XGBoost and MLP were compared. The base models set minimum references; linear models evaluated the representable signal by additive combinations; SARIMAX functioned as a statistical reference; XGBoost examined nonlinear interactions; and MLP represented a tabular neural approximation.

3.1. Data sources and temporal scope

A wired Vantage Pro2 Plus station was used, which measures temperature, humidity, pressure, wind, rain, radiation and UV rays. Its thermal sensors have protective shields, are updated every 10 seconds, and have a minimum margin of error of ± 0.3 °C. A mathematical conversion was made with atmospheric pressure, going from millimeters of mercury (Excel) to hectopascals using equation 1:

$$P_{hPa} = 1.33322 P_{mmHg} \tag{1}$$

Where P_hPa represents the pressure in hectopascals and P_mmHg the pressure recorded in millimeters of mercury. The factor 1.33322 is derived from the ratio of pascals, hectopascals, and millimeters of mercury [14]. In total, 41,903 TXT records and 9,971 Excel records were integrated. After integrating the data and obtaining a database of 66 339 hours between November 23, 2018 and June 18, 2026, with gaps that reflect real interruptions in the record due to different reasons beyond the authors' control. The full array for the 1-hour horizon contains 30,891 rows, from April 29, 2019 to June 18, 2026, and 67 predictors and 70 for XGBoost A. To this end, Table 1 outlines the summary of the characteristics established for the procedure.

3.2. Quality control and filtering of meteorological data

This section presents the quality control applied to meteorological variables in order to detect observations that were not physically plausible for the conditions of the study area. To do this, each value recorded at a given time was checked against previously established limits, so that those outside

Table 1
Summary of the sources and the final matrix for the 1-hour horizon.

Description	Value
Source Records	51 874 (TXT: 41 903; Excel: 9 971)
Time grid with gaps	66,339 hours
Grid Period	23/11/2018 10:00 - 18/06/2026 12:00
Full rows, horizon 1 h	30 891
Modeling period, horizon 1 h	29/04/2019 09:00 - 18/06/2026 11:00
Effective predictors	67
Predictor structure	7 current + 56 lags + 4 cyclical
Training/validation/testing ($h = 1$)	21 623 / 4 634 / 4 634
Missing temp_c and target after filtering	0 / 0

the permissible range were identified during the debugging. This criterion is formalized by the rule presented in Equation 2.

$$x_t^* = \begin{cases} x_t, & si\ L_x \leq x_t \leq U_x \\ NaN, & si\ x_t < L_x\ o x_t > U_x \end{cases} \tag{2}$$

Where and are the minimum and maximum limits respectively, is the observed value and is the purified value, if the data was within the natural limits, we left it; if it registered an impossible value (such as temperatures out of all logic), we marked it as empty (NaN) without altering the rest of the information. This filtering was carried out before the hourly aggregation and was recorded by source and variable. For this reason, the verification was based on the WMO-No. 8, the SENAMHI technical manual and the recommendations for the validation of automatic stations $L_x,\ U_x,\ x_t,\ x_t^*$ [11]. In this study, tests that could be audited with the available files were implemented, a spatial test was not applied because a network of meteorological stations is needed that are not available for this research.

Similarly, the process was meticulous that we only had to cancel eight cells of the original 51,874, which is equivalent to a tiny 0.0077% of the data, whose characteristics are shown in Table 2. From this, in the text file we only find two temperatures with their maximum and minimum out of range, and in Excel we detect two unrealistic rainfall rates. A very important detail is that, if a piece of data was extreme, but physically possible, we keep it. Therefore, this decision was taken so as not to confuse an unusual weather event (such as a peak in strong heat) with a sensor error, thus allowing us to preserve the reality of these intense episodes.

3.3. Temporal homogenization, shortages and time aggregation

This section describes the procedure used to unify the temporal structure of the data and ensure that observations from different sources can be used consistently in the later stages of the analysis. The records from the two sources presented different temporal frequencies. The TXT file mainly contained observations made at intervals of less than an hour, while the Excel file had hourly records. To establish a common time scale and ensure the comparability of lags, both sources were integrated into a database with a regular frequency of one hour. For continuous variables, to calculate the value of each hour in measurements such as temperature, pressure or wind, UV radiation we take the average of all the valid records within that hour. For this process, equation 3 was used.

$$\bar{x}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} x_{i,h} \tag{3}$$

Where is the hourly average of the variable, is each valid observation and the number of observations

Table 2
Physical ranges used during quality control.

Variable	Symbol	Allowable range	Methodological action
Outside temperature	T	5 to 45 °C	Out of range: NaN
Maximum temperature	T_max	5 to 50 °C	Out of range: NaN
Minimum temperature	T_min	5 to 45 °C	Out of range: NaN
Relative humidity	HR	0 to 100%	Out of range: NaN
Dew Point	T_d	0 to 35 °C	Out of range: NaN
Atmospheric pressure	P	850 to 1100 hPa	Out of range: NaN
Wind Speed	V	0 to 60 m/s	Out of range: NaN
Maximum Rafache	V_max	0 to 80 m/s	Out of range: NaN
Precipitation	R	0 to 500 mm	Out of range: NaN
Precipitation rate	R_rate	0 to 500 mm/h	Out of range: NaN
Solar radiation	S	0 to 1400 W/m²	Out of range: NaN
UV Index	UV	0 to 20	Out of range: NaN
Evapotranspiration	ET	0 to 50 mm	Out of range: NaN

available. After obtaining the averages, we use a mathematical formula (interpolation) to fill in small gaps of up to three hours in the continuous variables. In the 1-hour matrix, this rule completed between 194 and 196 values per variable. Unlimited forward or backward spreads were not used in time and extensive gaps remained missing. Incomplete rows were completely removed. Therefore, the difference between the 66,339 hours of the initial files and the 30,891 rows of the 1-hour matrix should not be interpreted as a loss caused solely by physical control: it is also due to prolonged station interruptions caused by different reasons, as well as the requirement of the 168-hour lag and the availability of future temperature $\bar{x}_h,\ x_{i,h},\ n_h$.

4. Predictive Problem

This section formalises the prediction of the outside temperature for the different time horizons considered. Let Tt be the observed temperature at the hour t and h the prediction horizon, defined at 1, 3, 6, 12 and 24 hours. For each horizon, the target variable $y_t^{(h)}$ corresponds to the observed temperature h hour later, as set out in Equation 4:

$$y_t^{(h)} = T_{t+h} \tag{4}$$

This formulation gives rise to five independent regression tasks and preserves the temporal sequence of the data, because all the predictive characteristics are constructed only with information available up to the time t, while T_{t+h} is maintained exclusively as a response variable. Under this criterion, the prediction model for each horizon is expressed by Equation 5:

$$\hat{y}_t^{(h)} = f_h(X_t;\ \theta) \tag{5}$$

In Equation 5, $\hat{y}_t^{(h)}$ represents the predicted temperature for the horizon h, X_t contains the set of characteristics available up to the hour t, f_h corresponds to the function learned by the model for each horizon, and θ represents the parameters estimated during training. This definition allows maintaining a common structure for all the models evaluated and ensuring that predictions are made without incorporating future information.

4.1. Meteorological lags

This subsection describes the construction of lags used to incorporate historical information of meteorological variables into tabular models, which do not have implicit temporal memory. For each variable, the observed value k hours before the prediction instant were defined using Equation 6:

$$x_{t-k}; \quad K = \{1, 2, 3, 6, 12, 24, 48, 168\} \tag{6}$$

In Equation 6, x_{t-k} represents the lagging value of the variable x and k indicates the number of hours prior to instant t. Lags of 1, 2, 3, 6, 12, 24, 48 and 168 hours were considered. The first four allow to represent the short-term meteorological continuity, while the lags of 12, 24 and 48 hours incorporate information on the intraday evolution and the patterns between days. For its part, the lag of 168 hours provides a reference with respect to the behavior observed a week earlier. This structure was uniformly applied to temperature, humidity, pressure, wind speed, precipitation, solar radiation, and UV index. From the current variables and their lags, the input matrix used by the models was formed. Its general structure is expressed by Equation 7:

$$X_t = [M_t, M_{t-1}, M_{t-2}, M_{t-3}, M_{t-6}, M_{t-12}, M_{t-24}, M_{t-48}, M_{t-168}, C_T] \tag{7}$$

In Equation 7, x_t represents the set of predictors available at time t, x_t groups the current meteorological measurements, the terms x_{t-k} corresponds to their respective lags, and ct contains the cyclical temporal components. Thus, the input vector was constituted by seven current measurements, 56 lagging variables, and four temporal components, for a total of 67 predictors. In the matrix corresponding to the one-hour horizon, ten pairs of variables with absolute correlation greater than 0.95 and none greater than 0.99 were identified. The maximum correlation was 0.9685 between the current temperature and its one-hour lag. Likewise, the condition number of the standardized matrix reached 144.35, which shows dependence among some lags without indicating an exact singularity.

4.2. Cyclical temporal variables

In this section, the temporal dimension of the predictive problem was incorporated through variables capable of representing the periodic nature of the time of day and the day of the year. This representation is necessary because a conventional numerical coding does not preserve continuity between the beginning and the end of each cycle. For example, 23:00 and 00:00 correspond to consecutive hours, although their numerical values can be interpreted as distant. To preserve this temporal relationship, the time of day was transformed by sine and cosine functions, as established in Equations 8 and 9:

$$hora_{sen_t} = sin\left(\frac{2\pi.hora_t}{24}\right), \quad hora_{cos_t} = cos\left(\frac{2\pi.hora_t}{24}\right) \tag{8}$$

For the day of the year the same principle applied:

$$dia_{sen_t} = sin\left(\frac{2\pi.dia_t}{365.25}\right), \quad dia_{cos_t} = cos\left(\frac{2\pi.dia_t}{365.25}\right) \tag{9}$$

4.3. Temporal division and information leakage prevention

Once the predictor matrix was constructed, the separation of the data was carried out strictly preserving its chronological sequence. The observations were not mixed and, for each prediction horizon, the initial 70% was allocated to training, 15% after validation and the remaining 15% to testing. This strategy made it possible to maintain the past-future relationship of the predictive problem and prevent subsequent information from being available during training. In addition, standardization and other procedures whose parameters depended on the data were adjusted without using information from the test set [12, 13]. Thus, the temporal partition was 70% training, 15% validation and 15% test.

For the one-hour horizon, the training set was composed of 21,623 observations between 29/04/2019 at 09:00 and 08/01/2025 at 11:00. The validation set included 4,634 observations from 08/01/2025 at 12:00 to 05/12/2025 at 01:00, while the test set was made up of 4,634 observations from 05/12/2025 at 02:00 to 18/06/2026 at 11:00. To verify that the performance observed in the final test did not depend exclusively on the characteristics of the last period, the evaluation was complemented with three expansive temporal validation cuts whose test blocks covered different periods of 2024 and 2025. This analysis allowed us to examine the temporal stability of the results and determine if the predictive advantage of XGBoost was maintained by modifying the period used for the evaluation.

5. Model Comparison

This section presents the models used to predict the outdoor temperature and the criteria used for their configuration. The comparison was organized in a progressive manner, starting with reference methods and linear models, continuing with a statistical approach for time series, and moving towards machine learning models and a multilayer neural network. This organization made it possible to contrast methods with different levels of complexity under the same temporal structure and a common set of evaluation metrics.

5.1. Reference models and linear models

The reference models were incorporated to establish basic levels of performance against which to assess the improvements obtained by more complex methods. Persistence uses the last available temperature at time t as a forecast, while the seasonal naïve model uses the observed temperature at the same position in the previous daily cycle [15]. The six-hour moving average uses the average of the six most recent observations, which makes it possible to smooth out short-term fluctuations, although it can reduce the ability to represent rapid changes.

As a linear approximation, a regression was used based on the 67 previously defined predictors. The variables were standardized using StandardScaler, whose parameters were estimated exclusively with training data. Ridge was also evaluated to control the existing dependence between predictors by L2 regularization. The α parameter was explored between 0.001 and 1000, and the value $\alpha = 100$ presented the best validation performance in the five horizons. The main characteristics of these models, together with their respective formulations and interpretations, are summarized in Table 3.

Table 3
Reference models and linear models used in the comparison.

Method	**Equation**	**Interpretation**
Persistence	$\hat{y}_t^{(h)} = T_t$	Forecast with the observed temperature in t. Demanding reference at 1 h. T_{t+h}.
24-hour seasonal naïve	$\hat{y}_t^{(h)} = T_{t+h-24}$	It uses the temperature of the same relative position as the previous daily cycle [15].
Moving average 6 h	$\hat{y}_t^{(h)} = \frac{1}{6}\sum_{j=0}^{5} T_{j-t}$	Average the six most recent temperatures; Softens noise but attenuates quick transitions.
Linear regression	$\hat{y}_t^{(h)} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{j,t}$	Linear combination of the 67 predictors, with Standard-Scaler adjusted only in training.
Ridge	$min_\beta[\sum_{i=1}^{n}(y_i - \hat{y}_t)^2 + \alpha\sum_{j=1}^{p}\beta_j^2]$	L2 regression. Alpha was evaluated between 0.001 and 1000; = 100 was selected by validation in the five horizons α.

5.2. SARIMAX as a statistical reference

In this subsection, SARIMAX is incorporated as a statistical reference to explicitly represent the time dependence of the series and contrast its performance with the tabular models. Its general formulation combines autoregressive components, moving average terms, and exogenous variables, as expressed in Equation 10:

$$\phi(B)\Phi(B^s)(y_t - X_t\beta) = \theta(B)\Theta(B^s)\epsilon_t \tag{10}$$

In Equation 10, B represents the lag operator, the autoregressive and moving average polynomials describe the residual temporal dynamics, X_t contains the exogenous variables, and ϵ_t corresponds to the innovation term. Initially, AR orders of 1, 2 and 3 were evaluated, keeping d, q, P, D and Q equal to zero and incorporating cyclic calendar components. A window of 8,760 hours was used for its calibration and the selection of the order was made using the Akaike information criterion without using the test set. From this formulation, two variants were preserved to examine the effect of the variables incorporated into the model. SARIMAX-calendar reproduces the initial approach using temperature and four time components. The AR orders of 1, 2, and 3 were compared using the Akaike information criterion. However, the selected adjustments did not fully meet the convergence criteria, so their results were considered a preliminary baseline.

The second variant, called SARIMAX-Xt, was formulated directly for each horizon using an AR order of 1, seven meteorological variables available in t and four cyclic components. The exogenous variables were standardized using the recent window of 8 760 observations and the adjustment was made using the Powell method. During the evaluation, the predictions were generated in blocks without updating the model with target-future temperatures. This variant converged on the five horizons and allowed us to examine the effect of incorporating the meteorological conditions available at the source. However, its purpose was to perform a sensitivity analysis and not to reproduce the 67 predictors used by XGBoost or to develop an exhaustive search for SARIMAX configurations.

5.3. XGBoost: 3A and 3B variants

This section evaluates XGBoost as a machine learning method based on the sequential combination of decision trees. The prediction is obtained by adding the contributions generated by the trees progressively incorporated into the model, as represented in Equation 11:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in F \tag{11}$$

In Equation 11, K represents the total number of trees and each function belongs to the tree space considered by the algorithm. During training, new trees are incorporated in order to reduce residual errors produced by previous stages. This process is done by minimizing an objective function that combines predictive error with a regularization term, according to Equation 12:

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad \Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2 \tag{12}$$

Equation 12 allows you to simultaneously control the fit to the data and the complexity of the model. Regularization penalizes both the structure of trees and the magnitude of weights assigned to their leaves, while sampling observations and features reduces dependence on specific subsets. These properties allow us to represent nonlinear relationships and interactions between meteorological variables, their lags and temporal components [16, 17]. On this basis, two variants were evaluated. XGBoost 3A used the original 67 predictors and additionally incorporated the predictions generated by persistence, seasonal naïve and moving average, reaching a total of 70 entries. Therefore, this variant integrates information derived from the reference models as additional predictors and can be considered a combination of predictions strategy.

In contrast, XGBoost 3B exclusively used the 67 forecasters constructed from current weather variables, their lags, and cyclical time components. This configuration allowed XGBoost's performance to be evaluated without relying on output produced by other models and provided a direct reference to its ability to learn from the original feature matrix.

For both variants, three hyperparameter configurations were evaluated: 250 trees with depth 2 and learning rate of 0.05, 300 trees with depth 3 and rate of 0.03, and 300 trees with depth 4 and rate of 0.03. The third configuration achieved the lowest validation RMSE across all five horizons and was selected for final tuning. The model used the reg:squarederror target function, the hist construction method, an observational sampling ratio of 0.85, a trait-per-tree ratio of 0.85, and a random seed of 42. Once the configuration was selected, the model was retrained using the training and validation sets together.

5.4. Multilayer Perceptron

This section presents the multilayer neural network used as a neural approximation on the same tabular matrix used by XGBoost 3B and Ridge. For each hidden layer l, the transformation of the inputs combines a linear operation with an activation function, as expressed in Equation 13:

$$a^{(l)} = g(W^{(l)}a^{(l-1)}b^{(l)}) \tag{13}$$

In Equation 13, W^l represents the matrix of weights of the l-layer, $b^{(l)}$ corresponds to the bias vector, $a^{(l-1)}$ contains the activations from the previous layer, and g represents the activation function. Once the information has been processed through the hidden layers, the prediction for each horizon is obtained using the output layer, defined in Equation 14:

$$\hat{y}_t^{(h)} = W^{(L)}a^{(l-1)} + b^{(L)} \tag{14}$$

In Equation 14, $\hat{y}_t^{(h)}$ represents the predicted temperature for the horizon h, $a^{(l-1)}$ corresponds to the activations of the last hidden layer, while $W^{(L)}$ and $b^{(L)}$ represent the weights and bias of the output layer, respectively. To introduce nonlinear relationships during learning, the ReLU activation function was used, defined by Equation 115:

$$ReLU(z) = max(0, z) \tag{15}$$

According to Equation 15, positive values are conserved, while negative values are transformed into zero. On this structure, architectures with 32 and 16 neurons and with 64 and 32 neurons in the two hidden layers were compared, the first being selected according to the performance obtained in the validation set. Training employed StandardScaler, ReLU activation, Adam optimizer, $\alpha = 10^{-4}$, initial learning rate of 10^{-3}, batch size of 64, a maximum of 300 iterations, and a criterion of 20 iterations with no improvement. To evaluate stability against different initializations, the MLP was executed with seeds 42, 123 and 2026. The standard deviation of the RMSE ranged from 0.0081 °C to 0.0491 °C depending on the horizon. Some runs reached the limit of 300 iterations, so the variability between seeds was considered a stability analysis and not evidence of complete convergence.

5.5. Implementation

The implementation describes the computational organization used to reproduce the data processing, model training, and complementary analyses. The main flow was structured in seven scripts executed sequentially: integration and debugging of the registers, reference and linear models, XGBoost 3A, XGBoost 3B, MLP, SARIMAX and generation of tables and figures. To this flow were added the analyses of Ridge optimization, MLP stability, expansive temporal validation, and statistical comparison tests. Table 4 summarizes the final configurations of the models and the computational environment used. The cleanup integrated TXT and Excel files, physical range controls, and time aggregation over 51,874 records, with eight cells identified as invalid values. The predictive matrix consisted of seven meteorological variables, eight lags per variable and four cyclical components, for a total of 67 predictors. The division was performed chronologically into 70% for training, 15% for validation and 15% for testing, complemented by three expansive time windows.

Table 4
Final configurations and reference computing environment.

Component	Summary Implementation	Reproducible detail
Cleaning	TXT + Excel; physical ranks; Time Aggregation	51,874 registrations; 8 cells marked as NaN.
Features	7 variables; 8 lags; 4 cycles	67 predictors per horizon
Division	70/15/15 Chronological	Reserved test; 3 additional expansive windows.
Ridge	StandardScaler + $\alpha = 100$	Adjustment: 0.014–0.018 s; Inference: 0.0007–0.0013 s.
SARIMAX-X_t	Direct AR(1) + 11 exogenous available in t 300 trees; depth 4; $\eta = 0.03$	Final adjustment: 3.36–9.32 s; inference: 0.046–0.081 s; 5/5 Convergence Setting: 1.20–1.36 s; inference: 0.012–0.013 s.
MLP	(32,16); ReLU; Adam; 300 iterations max	Observed setting: 12.1–24.0 s; some runs reached 300 iterations.
Testing	DM-HLN + bootstrap móvil	3,000 replicas; 24-hour block; Seed 42.
Environment	Python 3.13.5; 5 vCPU AMD EPYC	5.9 GiB RAM; Linux; Review baseline measurements.
Libraries	scikit-learn 1.8.0; XGBoost 3.1.3	statsmodels 0.14.6; SciPy 1.17.0; Matplotlib 3.10.8.

The runtime used Python 3.13.5 on Linux, with five AMD epyc vCPUs and 5.9 GiB of RAM. The main libraries were scikit-learn 1.8.0, XGBoost 3.1.3, statsmodels 0.14.6, SciPy 1.17.0, and Matplotlib 3.10.8. Statistical tests were performed using Diebold-Mariano with HLN correction and mobile bootstrap with 3 000 replicates, 24-hour blocks and seed 42. These elements allow documenting the computational and methodological conditions under which the results of the comparison were obtained.

6. Evaluation Metrics

The comparison of predictive performance was carried out using four complementary metrics: MAE, RMSE, R^2 and MASE, whose formulations are presented in Table 5. The selection of these metrics made it possible to examine error from different perspectives and maintain a common criterion for the five prediction horizons. The RMSE was established as the main metric due to its greater sensitivity to large errors, while the MAE allowed the average error to be expressed directly in degrees Celsius. In turn, R^2 allowed quantifying the temperature variability reproduced by the predictions and MASE provided a scaled measure regarding the performance of a 24-hour seasonal naïve reference [15, 18, 19].

Based on these metrics, the comparison was complemented with an inferential analysis aimed at determining whether the differences observed between pairs of models could be attributed to statistically significant variations in their performance. To do this, the Diebold and Mariano test was applied [20] on the differences in square losses, incorporating the correction of Harvey et al. [21] associated with each horizon. In addition, 95% confidence intervals were estimated for RMSE differences using mobile bootstrap with 3,000 replicas and 24-hour blocks. Evidence of statistically significant difference was considered when $p < 0.05$ and the corresponding confidence interval did not include zero.

Table 5
Metrics used for the evaluation of the models.

Metrics	Equation	Interpretation
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i-\hat{y}_i\rvert$	Mean absolute error in °C; Less is better [18].
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$	Root of the mean square error; It gives more weight to large deviations.
R^2	$1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_t)^2}{\sum_{i=1}^{n}(y_i-\hat{y}_t)^2}$	Proportion of variability reproduced against the average; it is not a percentage of success [19].
MASE	$\frac{\frac{1}{n}\sum_{t=1}^{n}\lvert y_t-\hat{y}_t\rvert}{\frac{1}{n-s}\sum_{t=s+1}^{n}\lvert y_t-\hat{y}_{t-s}\rvert}$	MAE is scaled by the 24-h seasonal naïve error; values less than 1 indicate improvement [15].

7. Results

7.1. Predictive performance according to the horizon

The performance of the models varied as the prediction horizon increased. The lowest errors were recorded at one hour, while the intermediate horizons presented a greater predictive difficulty. At 24 hours, the increase in error was moderated and values close to those recorded at 12 hours were obtained. In this general behavior, the five lowest RMSE corresponded to the XGBoost variants, as summarized in Table 6.

Table 6
Best observed result by horizon.

Horizon (h)	Lower RMSE	IT (°C)	RMSE (°C)	R^2	THE BUTT
1	XGBoost 3A*	0.4729	0.5991	0.9719	0.4404
3	XGBoost 3B†	0.7232	0.9521	0.9290	0.6763
6	XGBoost 3A†	0.8526	1.1752	0.8921	0.7944
12	XGBoost 3B*	0.8697	1.1972	0.8882	0.8109
24	XGBoost 3B*	0.8922	1.2323	0.8797	0.8228

XGBoost 3A obtained the lowest RMSE at 1 and 6 hours, with 0.5991 °C and 1.1752 °C, while XGBoost 3B presented the minimum values at 3, 12 and 24 hours, with 0.9521 °C, 1.1972 °C and 1.2323 °C. The same pattern is observed in the MAE and in the values of R^2 and MASE. At 24 hours, XGBoost 3B achieved an MAE of 0.8922 °C, an R^2 of 0.8797 and a MASE of 0.8228, the latter being lower than one and therefore compared to the seasonal reference considered. The evolution of the minimum RMSE values is shown in Figure 1. Although the error increases from 0.5991 °C at one hour to 1.2323 °C at 24 hours, its growth is not uniform between horizons.

To place these minimums within the complete set of models, Table 7 presents the RMSE obtained by each method. XGBoost maintained the lowest values across all horizons, followed by varying performances of MLP, Ridge and reference models.

In this way, the optimized Ridge recorded an RMSE of 0.6428 °C at one hour and 1.2463 °C at 24 hours, the latter numerically close to the 1.2323 °C of XGBoost 3B. On the other hand, SARIMAX-Xt improved compared to SARIMAX-calendar at 6, 12 and 24 hours, although it presented errors greater than 1 and 3 hours. The reference models showed differentiated behaviors, with a more pronounced loss of precision of persistence in the intermediate horizons. The comparison by MAE, presented in Table 8, allowed us to verify whether the previous behavior was maintained when considering the average magnitude of the errors without penalizing to a greater extent the high deviations.

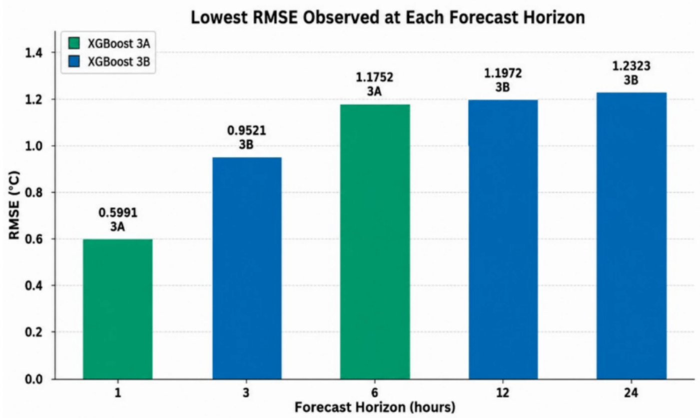

Figure 1: Lowest RMSE observed by prediction horizon.

Table 7
RMSE by model and horizon. Lower values indicate better performance.

Model	1 hour	3 hrs	6 hrs	12 p.m.	24 hrs
XGBoost 3B	0.6048	0.9521	1.2088	1.1972	1.2323
XGBoost 3A	0.5991	0.9578	1.1752	1.2084	1.2370
MLP	0.6495	1.0237	1.3179	1.3111	1.5062
Optimized Ridge	0.6428	1.2114	1.5815	1.2555	1.2463
Linear regression	0.6907	1.2374	1.5998	1.2845	1.2526
SARIMAX-X_t Direct	1.0094	3.4179	4.7997	3.8618	1.6730
SARIMAX-calendar	0.9242	2.4476	5.0501	7.4109	3.7745
Persistence	1.1306	2.8677	4.8205	6.0816	1.3876
Naïve 24 h	1.5667	1.5979	1.6419	1.6393	1.6370
Moving average 6 h	2.9645	4.2036	5.3650	5.5683	2.5286

The results preserved the pattern observed by RMSE. XGBoost variants had the lowest MAE across all five horizons, while seasonal naïve maintained relatively stable errors and the six-hour moving average recorded the largest deviations between reference methods. The agreement between both metrics reinforces the general pattern of observed performance.

7.2. Comparison between XGBoost 3A and XGBoost 3B

The numerical differences between the two variants of XGBoost were reduced in some horizons, so they were contrasted using the RMSE difference, 95% confidence intervals and the Diebold-Mariano test with Harvey-Leybourne-Newbold correction. The results are presented in Table 9.

At one hour, the -0.0057 °C difference significantly favored XGBoost 3A, with $p = 0.0025$. Over the 3 - and 6-hour horizons, the confidence intervals included zero and the differences were not statistically significant. The behavior changed at 12 and 24 hours, where XGBoost 3B presented a significant advantage, with p-values of 0.0139 and 0.0197, respectively. Therefore, the configuration with the lowest error depends on the horizon and the observed numerical differences do not always represent a statistically conclusive advantage. This behavior can be seen graphically in Figure 2, where the intervals corresponding to 3 and 6 hours cross the zero value, while the remaining horizons maintain defined intervals in favor of one of the variants.

Table 8
MAE by model and horizon. Lower values indicate lower mean deviation.

Model	1 hour	3 hrs	6 hrs	12 p.m.	24 hrs
XGBoost 3B	0.4762	0.7232	0.9058	0.8697	0.8922
XGBoost 3A	0.4729	0.7258	0.8526	0.8745	0.8958
MLP	0.4950	0.7696	0.9425	0.9485	1.1314
Optimized Ridge	0.5017	0.9428	1.2477	0.9154	0.8994
Linear regression	0.5459	0.9657	1.2510	0.9319	0.9140
SARIMAX-X_t Direct	0.8097	2.7048	3.7984	3.0449	1.2747
SARIMAX-calendar	0.7155	1.8942	3.8147	5.6350	3.2281
Persistence	0.8146	2.2795	4.0272	5.3249	0.9301
Naïve 24 h	1.0185	1.0287	1.0446	1.0545	1.0617
Moving average 6 h	2.4657	3.5759	4.6709	4.8336	2.0445

Table 9
Statistical RMSE Comparison Between XGBoost 3A and XGBoost 3B.

h	ΔRMSE 3A-3B	IC 95 %	p DM-HLN	Reading
1 hour	-0.0057	[-0.0094; -0.0020]	0.0025	Favors 3A
3 hrs	+0.0057	[-0.0187; +0.0358]	0.6159	No difference
6 hrs	-0.0336	[-0.0781; +0.0151]	0.0990	No difference
12 p.m.	+0.0112	[+0.0023; +0.0207]	0.0139	Favors 3B
24 hrs	+0.0047	[+0.0010; +0.0089]	0.0197	Favors 3B

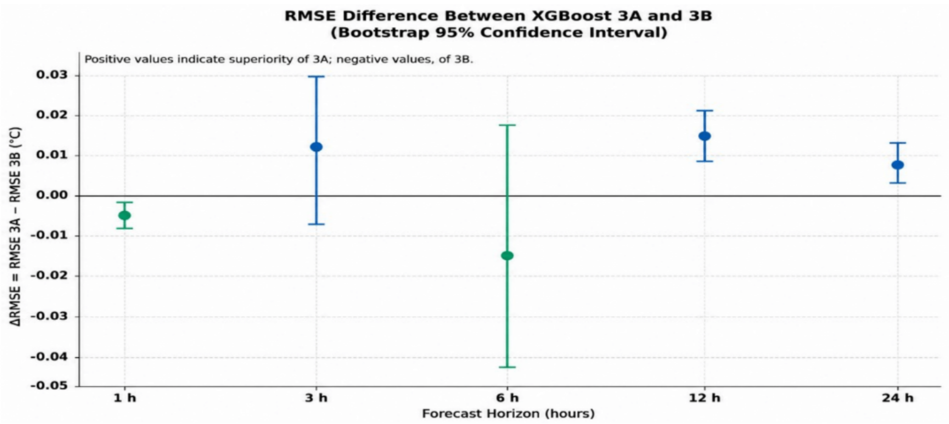

Figure 2: RMSE difference between XGBoost 3A and XGBoost 3B with confidence intervals.

7.3. Robustness and temporal stability

The stability of performance was examined through three expansive time windows, in order to verify whether the results obtained depended exclusively on the final probationary period. Table 10 presents the mean RMSE and its standard deviation for XGBoost 3B, optimized Ridge, and persistence.

XGBoost 3B maintained the lowest mean RMSE across all five horizons. The differences with respect to Ridge were wider at 3 and 6 hours and were reduced at 24 hours, where values of 1.1864 °C and 1.2065 °C were obtained, respectively. Persistence presented higher errors, particularly in the intermediate horizons. Consequently, the pattern observed in the final test was maintained when the evaluation

Table 10
Mean RMSE ± standard deviation in three expansive time windows.

h	XGBoost 3B	Optimized Ridge	Persistence
1 hour	0.5235 ± 0.0518	0.5665 ± 0.0303	1.1919 ± 0.0418
3 h	0.9021 ± 0.0431	1.1379 ± 0.0414	3.1727 ± 0.0888
6 h	1.2020 ± 0.0273	1.5873 ± 0.0908	5.3073 ± 0.1455
12 h	1.1186 ± 0.0900	1.2270 ± 0.1315	6.5988 ± 0.1464
24 h	1.1864 ± 0.1539	1.2065 ± 0.1859	1.3067 ± 0.1475

periods were modified, although the magnitude of the differences between models varied temporally.

7.4. Error and prediction behavior

In relation to the behavior of the errors and the predictions in this case, the distribution of the errors was examined to identify differences in their dispersion at 1, 6 and 24 hours. As Figure 3 shows, XGBoost exhibited more concentrated errors around zero, while differences between models increased mostly over the 6-hour horizon.

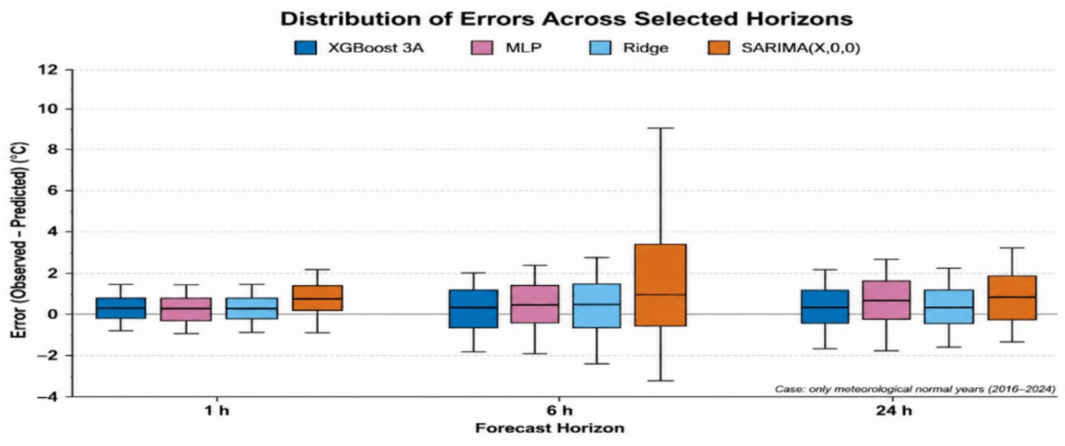

Figure 3: Distribution of errors to 1, 6 and 24 hours.

The time evolution of the predictions was subsequently examined over the 1, 6 and 24-hour horizons. At one hour, the estimates followed the observed temperature more closely and the main deviations appeared during certain peaks and transitions, as shown in Figure 4. The spaces without lines correspond to hours absent from the records and were preserved without artificial connection.

By extending the horizon to 6 hours, a greater separation between the predicted and observed trajectories was observed. Figure 5 shows that XGBoost maintained a greater proximity to the temperature evolution, while MLP and Ridge presented smoother trajectories and SARIMAX-Xt recorded deviations of greater magnitude in several segments.

Over the 24-hour horizon, shown in Figure 6, XGBoost 3B maintained the overall trajectory of the series, although it had differences in certain highs and lows. Ridge showed a visually close behavior in some periods, consistent with the smallest numerical difference recorded between both models in this horizon. This proximity is considered descriptive, because a specific paired test was not performed between both models.

Figure 7 presents the internal importance of the variables in the two XGBoost configurations. Current temperature and its lags were among the most important predictors, while in XGBoost 3A the predictions from the reference models also became relevant.

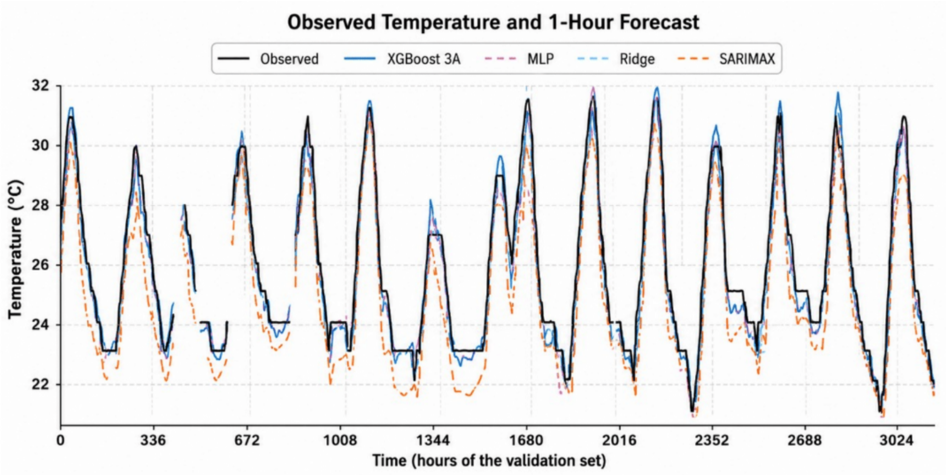

Figure 4: Observed and predicted temperature for the 1-hour horizon.

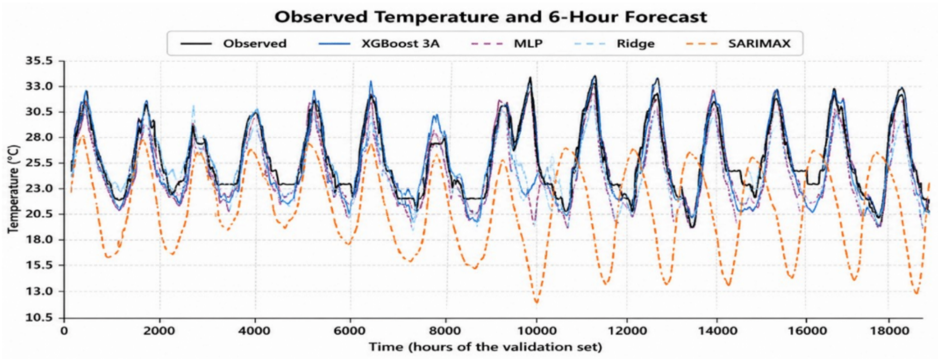

Figure 5: Observed and predicted temperature for the 6-hour horizon.

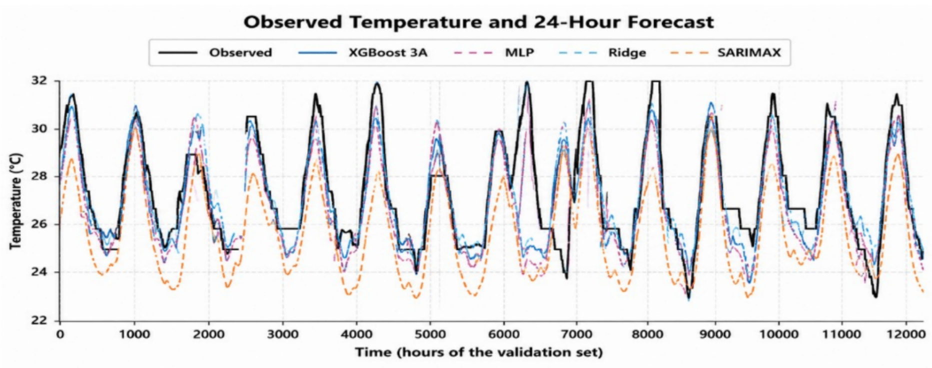

Figure 6: Observed and predicted temperature for the 24-hour horizon.

The presence of the base predictions among the relevant variables of XGBoost 3A confirms that this configuration takes advantage of additional information generated by other models, unlike XGBoost 3B, which operates exclusively with meteorological variables, lags, and time components. These importances express the relative utilization of predictors within trees and do not imply causal relationships.

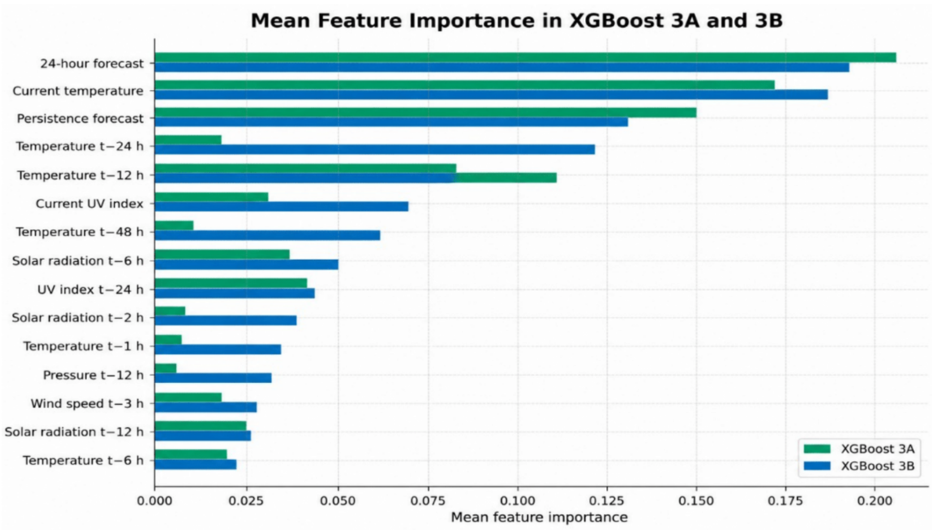

Figure 7: Average Importance of Variables in XGBoost 3A and XGBoost 3B.

8. Discussion

The results show that predictive performance cannot be interpreted solely from the algorithm used, but in relation to the way in which the temporal problem is structured. The filtering of the records, the hourly aggregation, the incorporation of lags and the chronological separation allowed to build a common evaluation scenario, while the ablation analysis evidenced a greater contribution of recent meteorological information compared to the cyclical components. This behavior is consistent with the temporal persistence of temperature and its relationship with variables such as solar radiation, humidity, pressure and wind. Consequently, the observed predictive capacity responds both to the use of time dependence and to the existing relationships between the variables available at each moment.

Within this framework, XGBoost achieved the lowest errors in the five horizons evaluated, a result consistent with studies that highlight the ability of the enhanced trees to represent nonlinear relationships in tabular meteorological data [3, 7]. However, the scope of the present work differs from research carried out with regional information, multiple stations or larger-scale architectures, since here the prediction is built from the records of a single meteorological station. This feature allows us to examine the extent to which a local source can sustain competitive forecasts when the data are organized under a controlled time structure. Therefore, the contribution does not intend to replace regional meteorological systems, but to provide reproducible evidence on the performance of different model families in a local scenario and with moderate computational resources.

This capability is therefore particularly visible over the 24-hour horizon. XGBoost 3B achieved an MAE of 0.8922 °C, an RMSE of 1.2323 °C, an R^2 of 0.8797 and a MASE of 0.822. The MASE value of less than one indicates that the model outperformed the 24-hour seasonal naïve reference, while the R^2 shows that it retained a high proportion of the observed variability. At the same time, the proximity of Ridge's performance on this horizon suggests that a part of the thermal dynamics can be approximated by relatively stable relationships between predictors. In contrast, XGBoost's advantage over the 3-, 6-, and 12-hour horizons highlights the value of representing nonlinear relationships and interactions that a linear model can capture in a more limited way.

However, the SARIMAX evaluation allows us to qualify this advantage and examine the effect of incorporating available meteorological information at the source of the forecast. The SARIMAX-Xt variant converged on the five horizons and improved compared to SARIMAX-calendar at 6, 12 and 24 hours, although it presented a lower performance at 1 and 3 hours. This behavior indicates that the initial difference cannot be attributed exclusively to the absence of exogenous meteorological variables. Aspects such as linear specification, autoregressive order, updating mechanism and the amount of information supplied to each model also play a role, given that SARIMAX-Xt used 11 predictors compared to the 67 used by XGBoost 3B. Consequently, this variant should be understood as a sensitivity analysis that allows assessing the effect of incorporating meteorological conditions available at the time of origin, but not as a completely symmetrical comparison between both families of models.

In this way, the results obtained do not imply that XGBoost is necessarily the best alternative to any deep learning architecture. Models such as LSTM, GRU, TCN and Transformer have shown the ability to represent temporal dependencies of different extensions in meteorological problems [22, 23]. These architectures were not incorporated into the main comparison because the study design prioritized a progression from reference models to machine learning methods on a common tabular matrix, while maintaining a moderate computational cost. In this sense, MLP should be interpreted only as a tabular neural reference and not as a representative of the set of models specialized in temporal sequences. Future evaluation of these architectures on the same records, time partitions, and horizons would determine whether explicit sequence modeling provides additional improvements over the performance achieved by XGBoost.

9. Limitations and future work

The scope of the results is conditioned by the use of records from a single meteorological station located in a university environment in Piura. Although the expansive temporal validation allowed us to examine the stability of performance in different periods, the results cannot be directly generalized to other areas of Piura or Peru without considering their own meteorological conditions, instrumental characteristics, and data availability. Also, some MLP runs reached the maximum limit of iterations and SARIMAX-Xt was configured with an AR order of 1 without exhaustively exploring other autoregressive structures, aspects that delimit the scope of the comparison. Based on these limitations, future research should incorporate records from other stations, analyze the behavior of errors according to seasons and extreme weather events, and document possible instrumental changes that affect the continuity of the series. It is also proposed to expand the search for SARIMAX configurations and evaluate architecture specialized in temporal sequences, such as LSTM, GRU, TCN and Transformers, using the same chronological partitions and evaluation criteria to determine if explicit modeling of temporal dependencies provides consistent improvements against the tabular methods analyzed.

10. Conclusions

Corresponding to the objective of comparing reference, statistical and machine learning models to predict the outdoor temperature at horizons of 1, 3, 6, 12 and 24 hours, the results show that XGBoost presented the most consistent overall performance under the same evaluation conditions. The integration and purification of the records of the meteorological station of the National University of Piura made it possible to obtain an adequate time series for modeling, whose matrix for the one-hour horizon was made up of 30,891 observations and 67 predictors. XGBoost achieved the lowest RMSE values in the five horizons, although the differences between its variants depended on the time of anticipation. XGBoost 3A was significantly higher than 3B at one hour, while at 3 and 6 hours the differences were inconclusive. In contrast, XGBoost 3B showed a significant advantage at 12 and 24 hours and, as it does not rely on auxiliary predictions, is the most suitable alternative for autonomous deployment. At 24 hours, it achieved an MAE of 0.8922 °C, an RMSE of 1.2323 °C, an R^2 of 0.8797 and a MASE of 0.8228.

These results should be interpreted considering that the superiority of XGBoost was not uniform against all models and horizons, as evidenced by temporal validation and statistical tests. Ridge maintained competitive performance at 24 hours, MLP exhibited variability between initializations, and SARIMAX-Xt improved over the calendar-based model at 6, 12, and 24 hours. In this way, the findings demonstrate that a properly structured local weather series can support short-term predictions with reduced errors and that the increase in model complexity must be justified by comparisons made under equivalent conditions. In this sense, horizons of 1 to 6 hours can contribute to the thermal monitoring of environments, crops or equipment, while predictions of 12 and 24 hours can provide information for the planning of temperature-sensitive activities.

Declaration on Generative AI

During the preparation of this work, the authors used generative artificial intelligence tools as support for organising ideas, reviewing the writing and improving the clarity of the text. The authors reviewed, edited and verified the resulting content, and assume full responsibility for the final content of the publication.

References

[1] K. Takahashi, A. G. Martínez, The very strong coastal el niño in 1925 in the far-eastern pacific, Climate Dynamics 52 (2019) 7389–7415.
[2] C. Rodriguez-Morata, H. Díaz, J. A. Ballesteros-Canovas, M. Rohrer, M. Stoffel, The anomalous 2017 coastal el niño event in peru, Climate Dynamics 52 (2019) 5605–5622.
[3] N. T. T. Hoa, Q. C. Thi, N. T. C. Ngoan, Time series prediction based on machine learning: A case study, temperature forecasting in vietnam, Journal of Military Science and Technology 85 (2023) 152–162.
[4] A. Ozbek, Deep learning approach for one-hour ahead forecasting of weather data, Energy Sources, Part A: Recovery, Utilization, and Environmental Effects 45 (2023) 7606–7628.
[5] M. Saleem, M. M. Saleem, F. Waseem, M. A. Bashir, An ensemble forecasting method based on optimized lstm and gru for temperature and humidity forecasting, Engineering, Technology & Applied Science Research 14 (2024) 18447–18452.
[6] E. Alerskans, J. Nyborg, M. Birk, E. Kaas, A transformer neural network for predicting near-surface temperature, Meteorological Applications 29 (2022) e2098.
[7] L. Hieta, M. Partio, Operational machine learning post-processing of short-range temperature, humidity, wind speed and gust forecasts, Meteorological Applications 32 (2025) e70074.
[8] M. Jiménez-Carrión, F. Gutiérrez-Segura, J. Celi-Pinzón, Modelado y predicción del fenómeno el niño en piura, perú mediante redes neuronales artificiales usando matlab, Información tecnológica 29 (2018) 303–316.
[9] O. Peña, M. More, R. Nima, H. Marchan, Artificial neural network model for the prediction of the" el niño" phenomenon in the region of piura (peru), TECHNO Review. International Technology, Science and Society Review/Revista Internacional de Tecnología, Ciencia y Sociedad 13 (2023) 10–37467.
[10] O. J. P. Cáceres, M. A. M. More, R. E. Nima, H. S. Marchan, Modelo de serie temporal para pronosticar la temperatura superficial del mar en la zona costera de paita (perú), TECHNO REVIEW. International Technology, Science and Society Review/Revista Internacional de Tecnología, Ciencia y Sociedad 11 (2022) 1–11.
[11] J. Estévez, P. Gavilán, J. V. Giráldez, Guidelines on validation procedures for meteorological data from automatic weather stations, Journal of Hydrology 402 (2011) 144–154.
[12] C. Bergmeir, J. M. Benítez, On the use of cross-validation for time series predictor evaluation, Information Sciences 191 (2012) 192–213.
[13] S. Kaufman, S. Rosset, C. Perlich, O. Stitelman, Leakage in data mining: Formulation, detection, and avoidance, ACM Transactions on Knowledge Discovery from Data (TKDD) 6 (2012) 1–21.
[14] B. N. Taylor, Guide for the Use of the International System of Units (SI), volume 811, US Department of Commerce, Technology Administration, National Institute of …, 1995.
[15] R. J. Hyndman, A. B. Koehler, Another look at measures of forecast accuracy, International journal of forecasting 22 (2006) 679–688.
[16] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[17] J. H. Friedman, Stochastic gradient boosting, Computational statistics & data analysis 38 (2002) 367–378.
[18] C. J. Willmott, K. Matsuura, Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance, Climate research 30 (2005) 79–82.
[19] D. Chicco, M. J. Warrens, G. Jurman, The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation, Peerj computer science 7 (2021) e623.
[20] F. X. Diebold, R. S. Mariano, Comparing predictive accuracy, Journal of Business & economic statistics 20 (2002) 134–144.
[21] D. Harvey, S. Leybourne, P. Newbold, Testing the equality of prediction mean squared errors, International Journal of forecasting 13 (1997) 281–291.
[22] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.
[23] S. Bai, J. Z. Kolter, V. Koltun, An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, arXiv preprint arXiv:1803.01271 (2018).