

Article

A Computational Framework to Assess Model Complexity Trade-Offs in Country-Level Temperature Anomaly Time Series

Rafael Rojas-Galván ¹, Luis E. Gallo-Gonzalez ¹, Juan S. Arteaga-Hernandez ¹, Omar Rodríguez-Abreo ^{1,2,*} and Juvenal Rodríguez-Reséndiz ^{1,*}

¹ Facultad de Ingeniería, Universidad Autónoma de Querétaro, Santiago de Querétaro 76010, Querétaro, Mexico; rafael.rojas@uaq.mx (R.R.-G.); luis.gallo@uaq.mx (L.E.G.-G.); juan.arteaga@uptc.edu.co (J.S.A.-H.)

² Programa de Ingeniería en Datos e Inteligencia Artificial, Universidad Politécnica de la Zona Metropolitana de Guadalajara, Av. Adolfo B. Horn 8941, Colonia Arvento, Cajititlán, Tlajomulco de Zúñiga 45670, Jalisco, Mexico

* Correspondence: omar.rodriguez.a@uaq.mx (O.R.-A.); juvenal@uaq.edu.mx (J.R.-R.)

Abstract

Accurate forecasting of country-level temperature anomalies is increasingly important for climate monitoring, policy planning, and environmental risk assessment. However, the trade-off between predictive performance, model complexity, and computational cost remains insufficiently explored, particularly across multiple countries using compact and interpretable feature representations. This study presents a comprehensive comparative evaluation of eight forecasting approaches for annual temperature anomaly prediction using country-level observations from the FAOSTAT Temperature Change dataset. The evaluated methods comprise a Persistence baseline, Ordinary Least Squares (OLS), Ridge regression, Support Vector Regression (SVR), Random Forest, a multilayer perceptron (MLP), and the classical time-series models ARIMA and ETS. Annual temperature anomalies were modeled using lagged observations, a temporal trend, and a trailing moving average under a temporally ordered 80/20 train–test split. Model performance was assessed using RMSE, MAE, R^2 , per-country win-rate, computational runtime, and pairwise statistical comparisons based on the Wilcoxon signed-rank test with Holm correction. Hyperparameters were optimized through expanding-window temporal cross-validation, and an ablation study was conducted to quantify feature contributions. Results indicate that the ETS model achieved the best overall predictive performance, obtaining the lowest median RMSE (0.3388 °C), the lowest MAE (0.2792 °C), and the highest per-country win-rate (40.07%). ARIMA provided competitive forecasting accuracy but incurred substantially higher computational cost, whereas OLS and Ridge offered an attractive compromise between predictive performance, robustness, interpretability, and computational efficiency. In contrast, the more flexible machine learning models (SVR, Random Forest, and MLP) did not consistently outperform the simpler approaches despite their higher complexity. Overall, the results demonstrate that classical statistical forecasting methods remain highly competitive for annual country-level temperature anomaly prediction and that increasing model complexity does not necessarily translate into improved predictive performance.

Keywords: temperature anomaly forecasting; climate time series; predictive modeling; machine learning regression



Academic Editor: Steven Prestwich

Received: 27 May 2026

Revised: 12 July 2026

Accepted: 16 July 2026

Published: 20 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

1. Introduction

Understanding and forecasting temperature anomalies at national and urban scales has become increasingly important in the context of global climate change. While global warming trends are well documented at planetary scales, regional and local temperature dynamics often exhibit heterogeneous patterns driven by geographical conditions, atmospheric circulation, land–surface interactions, and anthropogenic factors such as urbanization [1–3]. These processes lead to substantial variability across countries, where temperature series may display both linear and nonlinear dynamics influenced by climate variability and large-scale ocean–atmosphere interactions [4–6].

Urban environments introduce additional complexity due to land-use changes and localized climate effects. Studies on urban heat islands demonstrate that urbanization can significantly intensify warming patterns at the city level and influence temperature-related mortality [7,8]. These dynamics have direct societal implications, as temperature extremes are strongly associated with adverse health outcomes, including respiratory mortality and heat-related illnesses [9]. Vulnerable populations, particularly older individuals, are disproportionately affected by extreme temperature exposure, reinforcing the importance of accurate monitoring and forecasting [10]. Beyond health impacts, temperature variability also affects key socioeconomic systems, including agricultural productivity, water availability, and food security [11–14].

From a methodological perspective, time series analysis provides a natural framework for modeling climate variables. Classical statistical approaches—such as regression models and autoregressive structures—offer interpretable representations of temporal dynamics and have been widely used to study climate variability [15]. In parallel, advances in data availability and computational methods have enabled the use of machine learning models, including neural networks and hybrid architectures, which can capture nonlinear relationships and complex temporal dependencies [16–18]. These approaches have shown promising results in diverse applications such as crop yield forecasting, agroclimatic prediction, environmental monitoring, and energy planning [19–23].

Despite these advances, important challenges remain in climate forecasting. Model performance can be affected by data limitations, measurement errors, and non-stationary climate dynamics [24]. Furthermore, predictive accuracy often varies depending on model structure, spatial scale, preprocessing choices, and uncertainties associated with future climate scenarios [25–27]. Another key limitation concerns evaluation methodologies: many studies rely on in-sample metrics or conventional cross-validation schemes that do not fully reflect real-world forecasting conditions. In contrast, temporally consistent approaches such as rolling-origin or expanding-window validation provide more realistic assessments of predictive performance.

Although a substantial body of literature examines temperature dynamics at regional or national levels, systematic multi-country comparisons remain limited. Existing studies frequently focus on individual case studies or specific applications, which restricts the generalizability of their findings. Moreover, relatively few contributions explicitly analyze the trade-off between model complexity, predictive accuracy, and computational cost under consistent experimental conditions.

To address these gaps, this study proposes a unified computational framework for evaluating forecasting models of annual temperature anomalies across countries. Using FAOSTAT data derived from NASA–GISS temperature products, we perform a systematic multi-country comparison under temporally consistent validation schemes. The analysis includes classical statistical models, machine learning approaches, and baseline methods, evaluated using multiple criteria such as error distributions, win-rate analysis, cross-validation performance, and paired statistical testing.

The main contributions of this work are summarized as follows:

- A unified framework for evaluating forecasting models on country-level temperature anomaly series.
- A large-scale comparison of baseline, linear, machine learning, and classical time-series models across heterogeneous countries.
- An analysis of the trade-off between predictive accuracy, model complexity, and computational cost.
- An evaluation strategy based on temporal cross-validation and statistical comparisons.
- Practical benchmark results for model selection in climate forecasting applications.

The remainder of the paper is organized as follows. Section 2 presents the exploratory data analysis. Section 3 describes the modeling framework and validation strategy. Section 4 reports the empirical results. Finally, Section 6 concludes the study.

2. Exploratory Data Analysis

The FAOSTAT Temperature Change (ET) dataset was used, publicly available via Kaggle (<https://www.kaggle.com/datasets/sevgisarac/temperature-change> accessed on 23 march 2026). The ET values are derived from the NASA–GISS GISTEMP products and provide harmonized country-level annual temperature anomalies.

Temperature anomalies represent deviations from a long-term climatological baseline and are widely used to assess climate variability and long-term warming trends.

The original data were filtered to retain only records corresponding to the *Temperature change* element and the *Meteorological year*. After preprocessing, the resulting dataset consisted of 282 countries and territories covering the period from 1963 to 2019. Missing temperature anomaly values were removed without imputation, and additional observations were excluded during feature engineering when lagged predictors or moving-average features could not be computed or when country-level series contained fewer than 10 valid observations. Consequently, the resulting country-level time series are of unequal length. A summary of the dataset after preprocessing is provided in Table 1.

An exploratory data analysis (EDA) was subsequently conducted to characterize global warming trends, distributional properties, spatial heterogeneity, and temporal dependence. These characteristics provide the basis for selecting appropriate forecasting models and interpreting their predictive performance.

Table 1. Summary of the dataset after preprocessing.

Characteristic	Value
Countries/territories	282
Observations	14,393
Time span	1963–2019
Missing-value removal	1781 records
Removed during feature engineering	568 records
Removed (series with <10 observations)	14 records
Missing-value handling	Listwise deletion (no imputation)
Minimum series length	12 years
Maximum series length	57 years
Equal time-series length	No

2.1. Global Temperature Trend

Figure 1 shows the global average temperature anomaly over time.

A clear and persistent warming trend is observed, particularly from the late 1970s onward. Temperature anomalies increase steadily and exceed 1 °C in recent years. While short-term fluctuations remain present, they occur around a strong upward trend.

This suggests that trend components are a fundamental feature of the data and should be explicitly modeled.

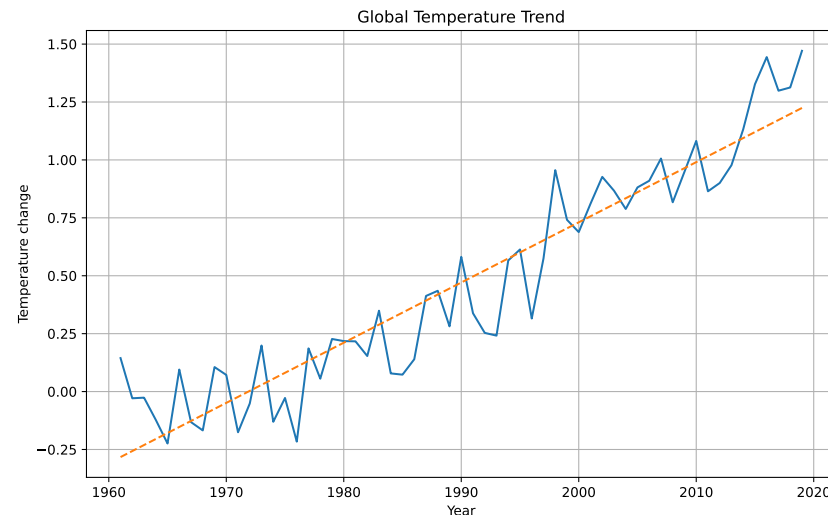


Figure 1. Global average annual temperature anomaly over time.

2.2. Distributional Shifts over Time

Figure 2 shows the distribution of temperature anomalies grouped by decade.

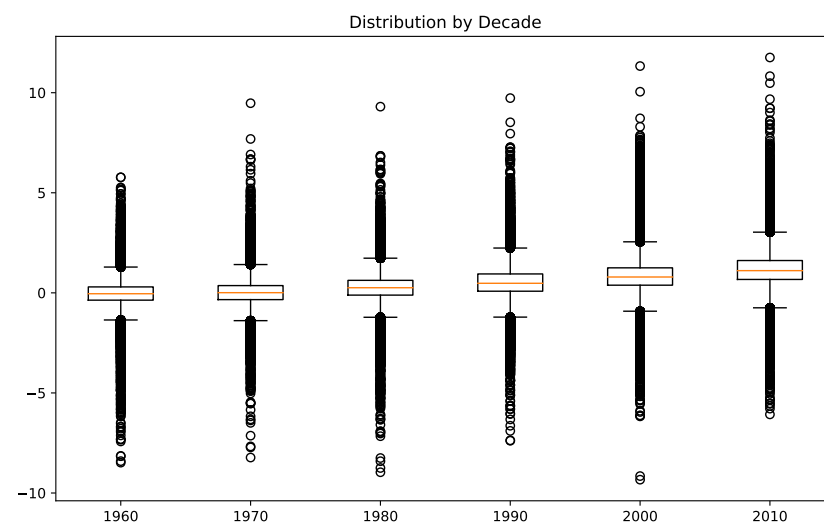


Figure 2. Distribution of temperature anomalies by decade.

The distribution shifts upward across decades, with increasing median values and a higher concentration of positive anomalies. This indicates a systematic warming process rather than isolated extreme events.

The interquartile range remains relatively stable, suggesting that the increase is driven by a shift in central tendency rather than increased dispersion.

2.3. Correlation Structure

Figure 3 presents the correlation matrix between the target variable and the engineered features.

The engineered features include lagged values (Lag 1 and Lag 2), a trend component, and a moving average. Lagged variables represent temperature anomalies observed in previous years (y_{t-1} and y_{t-2}), capturing temporal persistence. The trend variable accounts for long-term warming, while the moving average summarizes recent observations.

The correlation matrix reveals strong positive relationships between the target variable and its lagged values, confirming the presence of temporal dependence. In particular, the first lag exhibits the highest correlation, indicating that recent observations contain most of the predictive signal.

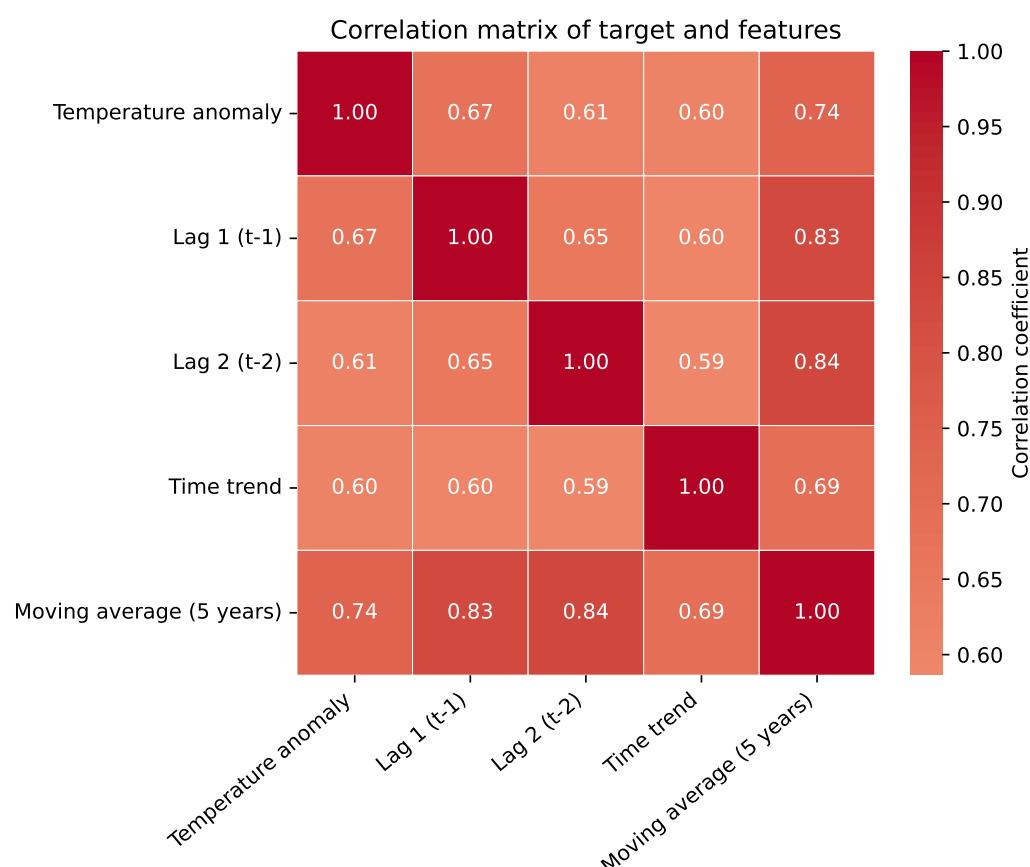


Figure 3. Correlation matrix between temperature anomalies and engineered features.

The moving average also shows a strong association with the target, reflecting its smoothing effect. In contrast, the trend variable presents a weaker correlation, suggesting that long-term drift is less informative than short-term dynamics.

Additionally, the high correlations among lagged variables indicate redundancy in the feature space, which helps explain the strong performance of regularized linear models such as Ridge regression.

Overall, the correlation structure supports the use of autoregressive features and provides a statistical explanation for the observed model performance.

2.4. Countries with Strongest Warming

Figure 4 compares two complementary indicators of long-term warming across countries. Figure 4a ranks countries according to the increase in mean annual temperature anomaly per decade, computed as the difference between the average anomaly during the last and first decades of the available record, normalized by the duration of the time series. Figure 4b presents the warming rate estimated from the temporal slope of the annual temperature anomaly series, expressed in °C per decade.

Both indicators consistently identify high-latitude regions, particularly Svalbard and Jan Mayen Islands, among the areas experiencing the most pronounced warming. Although the rankings differ slightly because each metric captures a different aspect of long-term temperature change, both reveal a common geographical pattern in which northern countries exhibit stronger warming than lower-latitude regions. This behavior is consistent with the well-documented phenomenon of polar amplification and highlights the spatial heterogeneity of climate change, motivating the need for region-specific forecasting models.

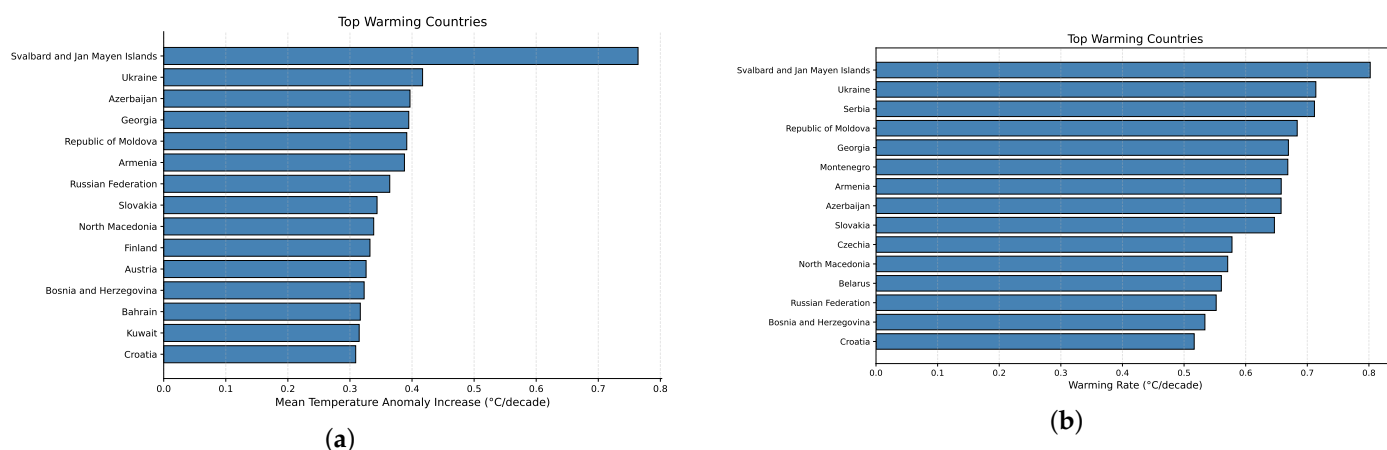


Figure 4. Comparison of countries exhibiting the strongest warming according to two descriptive metrics: (a) increase in mean annual temperature anomaly per decade and (b) warming rate estimated from the temporal trend of the annual temperature anomaly series.

2.5. Temporal Dependence

Temporal dependence was analyzed using the autocorrelation function (ACF), shown in Figure 5. The ACF was computed from the global average annual temperature anomaly series in order to quantify the correlation between observations separated by different temporal lags.

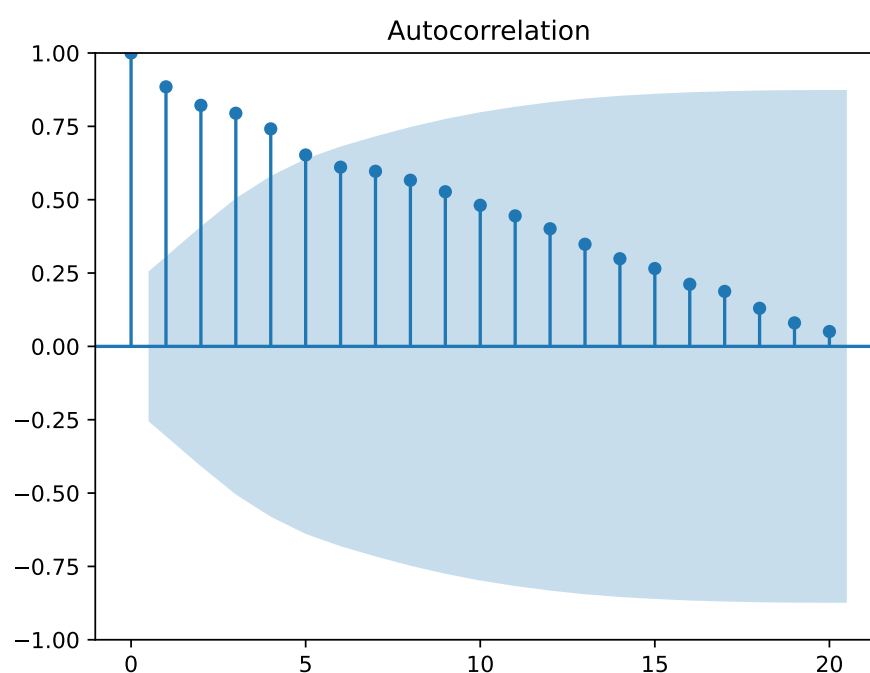


Figure 5. Autocorrelation function (ACF) of the global average annual temperature anomaly series. Shaded regions represent the approximate 95% confidence intervals under the null hypothesis of no autocorrelation.

The results reveal strong positive autocorrelation at low temporal lags, indicating substantial temporal persistence in the annual temperature anomaly series. The gradual decay of autocorrelation values across increasing lags suggests that temperature anomalies exhibit long-term memory and slow temporal dynamics rather than purely random fluctuations.

This behavior supports the inclusion of lagged observations and autoregressive structures within the forecasting framework, since previous years contain predictive information about future temperature anomalies. The observed persistence also explains why relatively simple autoregressive and trend-based models remain highly competitive under the evaluated forecasting setting.

2.6. Clustering of Climate Dynamics

Clustering analysis was applied to identify groups of countries exhibiting similar long-term temperature dynamics based on two aggregated descriptors: warming rate and temporal variability (Figure 6). The warming rate represents the estimated long-term temperature trend for each country, while variability measures the dispersion of annual temperature anomalies over time.

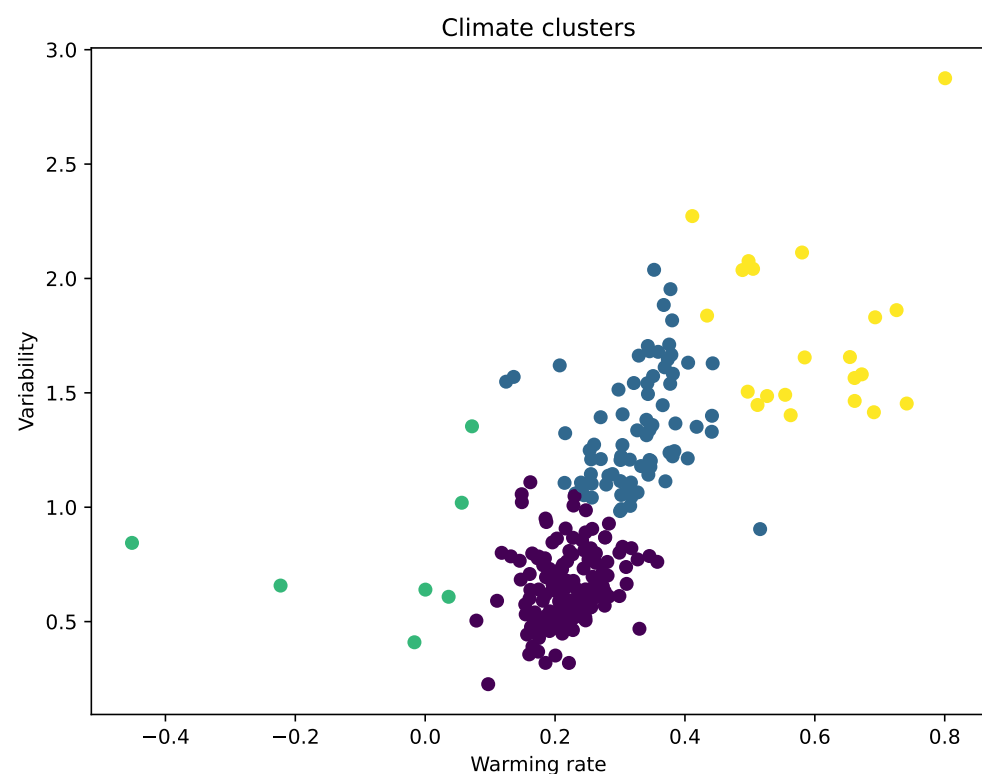


Figure 6. Climate clusters based on warming rate and temperature anomaly variability. Each point represents a country projected into a two-dimensional feature space defined by long-term warming trend and temporal variability. Colors indicate cluster membership identified through unsupervised clustering analysis.

The clustering results reveal the existence of several distinguishable climate regimes characterized by different combinations of warming intensity and temporal variability. Countries located toward the right side of the figure exhibit stronger long-term warming trends, while countries positioned higher on the vertical axis show greater interannual variability in temperature anomalies.

The central concentration of countries suggests that many regions share moderate warming rates and variability patterns, whereas smaller peripheral clusters correspond to countries with more extreme climatic behaviors. In particular, some groups exhibit both

strong warming trends and elevated variability, potentially reflecting regions more exposed to unstable climate dynamics.

Despite these differences in climate behavior, subsequent forecasting experiments showed that relative model performance remains relatively stable across clusters. This suggests that the comparative ranking between classical time-series models, linear methods, and machine learning approaches is robust under different climate regimes within the evaluated annual forecasting framework.

2.7. Summary

The EDA reveals:

- A strong global warming trend
- Upward shifts in temperature distributions
- Significant spatial heterogeneity
- Strong temporal dependence

These findings justify the use of time-series models with trend and autoregressive components and support the feature engineering choices adopted in this study.

3. Materials and Methods

3.1. Dataset and Preprocessing

The original dataset was transformed from wide format into a long panel structure, where each observation corresponds to a unique country–year pair. Missing temperature anomaly values were removed without imputation. Subsequently, lagged variables and moving-average features were generated using only historical observations, resulting in the exclusion of observations for which these predictors could not be computed. Finally, country-level series containing fewer than ten valid observations after feature engineering were discarded to ensure sufficient temporal support for model estimation.

After preprocessing, the final dataset comprised **282 countries and territories** and **14,393 country–year observations**, covering the period from **1963** to **2019**. A total of **1781** observations containing missing temperature anomaly values were removed, while an additional **568** observations were discarded during feature engineering because lagged predictors and moving-average features could not be computed. Finally, **14** observations belonging to short country-level series were excluded by the minimum-length criterion. Consequently, the resulting time series are not all of identical length, ranging from **12** to **57** annual observations.

Let $y_{c,t}$ denote the annual temperature anomaly (in °C) for country c observed in year t .

3.2. Feature Engineering

A supervised learning representation was constructed using lagged observations and trend-based predictors derived exclusively from historical temperature anomaly values. This feature engineering strategy was designed to provide a compact, interpretable, and temporally consistent representation while ensuring a fair comparison among the evaluated forecasting models. Unless otherwise specified (i.e., ARIMA, ETS, and the Persistence model), all supervised learning models (OLS, Ridge, SVR, Random Forest, and MLP) used the same predictor vector.

The feature vector for country c at year t is defined as

$$\mathbf{x}_{c,t} = [y_{c,t-1}, y_{c,t-2}, (t - t_0), \text{MA}_5(t)]^\top, \quad (1)$$

where $y_{c,t-1}$ and $y_{c,t-2}$ denote the temperature anomalies observed one and two years before year t , respectively. The term $(t - t_0)$ represents a temporal trend measured as the number of years elapsed since the first available observation t_0 for country c . Finally, $MA_5(t)$ is a trailing five-year moving average defined as

$$MA_5(t) = \frac{1}{k} \sum_{j=\max(t-4, t_0)}^{t-1} y_{c,j}, \quad k = t - \max(t - 4, t_0), \quad (2)$$

where $y_{c,j}$ denotes the observed temperature anomaly for country c at year j , and k corresponds to the effective number of historical observations available within the moving-average window. By relying exclusively on observations preceding year t , the moving average summarizes medium-term temperature evolution while avoiding the use of future information.

All predictors were computed exclusively from past observations, thereby preserving temporal causality and preventing information leakage during training and evaluation. Numerical predictors were standardized whenever required by the corresponding learning algorithm (Ridge, SVR, and MLP), whereas tree-based models (Random Forest) and Ordinary Least Squares were trained using the original feature scales. The same feature representation was employed throughout all supervised learning models to ensure that performance differences reflected the forecasting algorithms rather than differences in input information.

3.3. Predictive Models

Eight forecasting models were evaluated, spanning baseline, linear, machine learning, and classical time-series approaches. This selection allows the comparison of low-complexity interpretable models with more flexible nonlinear approaches under the same forecasting framework.

Unless otherwise specified, all supervised learning models were trained using the same feature representation defined in Section 3.2, ensuring that differences in predictive performance were caused by the learning algorithm rather than by differences in input information.

Persistence Naïve Model

The Persistence model was considered as a baseline forecasting approach. It assumes that the next temperature anomaly remains equal to the most recently observed value:

$$\hat{y}_{c,t} = y_{c,t-1}. \quad (3)$$

where $\hat{y}_{c,t}$ denotes the predicted temperature anomaly for country c at year t , and $y_{c,t-1}$ represents the temperature anomaly observed in the previous year.

This model provides a reference level of performance based exclusively on temporal persistence.

Ordinary Least Squares (OLS)

The Ordinary Least Squares (OLS) model was included as a linear benchmark to evaluate the relationship between historical temperature information and future temperature anomalies. The model is expressed as

$$\hat{y}_{c,t} = \beta_0 + \beta^\top \mathbf{x}_{c,t}, \quad (4)$$

where $\hat{y}_{c,t}$ is the predicted temperature anomaly for country c at year t , β_0 is the intercept term, and β represents the regression coefficients.

Unlike a generic regression formulation, the predictor vector $\mathbf{x}_{c,t}$ used in this study is explicitly defined as

$$\mathbf{x}_{c,t} = [y_{c,t-1}, y_{c,t-2}, (t - t_0), MA_5(t)]^\top. \quad (5)$$

Therefore, the independent variables used by the OLS model are: (i) the one-year lagged temperature anomaly $y_{c,t-1}$, (ii) the two-year lagged temperature anomaly $y_{c,t-2}$, (iii) the temporal trend $(t - t_0)$, and (iv) the five-year trailing moving average $MA_5(t)$.

The lagged variables represent short-term temperature persistence, the trend term captures the long-term warming evolution, and the moving average summarizes recent temperature behavior. The regression parameters were estimated by minimizing the residual sum of squares between observed and predicted temperature anomalies.

Ridge Regression

Ridge regression was evaluated as a regularized extension of the linear model to improve coefficient stability when predictors are correlated. The optimization problem is defined as

$$\min_{\beta} \sum (y_{c,t} - \beta_0 - \beta^\top \mathbf{x}_{c,t})^2 + \lambda \|\beta\|_2^2. \quad (6)$$

where $y_{c,t}$ is the observed temperature anomaly, β_0 is the intercept, β is the coefficient vector, and λ controls the regularization strength.

The Ridge model uses the same explicitly defined predictor variables as OLS:

$$\mathbf{x}_{c,t} = [y_{c,t-1}, y_{c,t-2}, (t - t_0), MA_5(t)]^\top. \quad (7)$$

Thus, Ridge regression receives four input variables: the one-year temperature lag, the two-year temperature lag, the temporal trend, and the five-year moving average.

The term $\|\beta\|_2^2$ corresponds to the squared Euclidean norm of the coefficient vector.

This penalty reduces excessively large coefficients and improves model robustness in the presence of correlated predictors, particularly among lagged temperature variables and moving-average features.

Support Vector Regression (SVR)

Support Vector Regression (SVR) was evaluated as a nonlinear kernel-based regression method capable of modeling nonlinear relationships between historical temperature features and future temperature anomalies.

SVR used the same input representation:

$$\mathbf{x}_{c,t} = [y_{c,t-1}, y_{c,t-2}, (t - t_0), MA_5(t)]^\top. \quad (8)$$

A radial basis function (RBF) kernel was employed:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|_2^2). \quad (9)$$

where $K(\mathbf{x}_i, \mathbf{x}_j)$ represents the similarity between two feature vectors and γ controls the kernel width.

The SVR model seeks a regression function that minimizes prediction error while maintaining a maximum-margin solution controlled by the regularization parameter C and the insensitive-loss parameter ϵ .

Random Forest

Random Forest was included as an ensemble-based nonlinear regression approach. The model was trained using the same predictor vector $\mathbf{x}_{c,t}$ consisting of lagged anomalies, temporal trend, and moving-average information.

Random Forest combines multiple decision trees trained using bootstrap aggregation and random feature selection.

The final prediction is obtained by averaging the predictions of all individual regression trees:

$$\hat{y}_{c,t} = \frac{1}{B} \sum_{b=1}^B T_b(\mathbf{x}_{c,t}), \quad (10)$$

where B represents the number of trees and $T_b(\mathbf{x}_{c,t})$ denotes the prediction generated by the b -th tree.

This ensemble strategy reduces prediction variance and improves robustness compared with individual decision trees.

Neural Network (MLP)

The multilayer perceptron (MLP) was evaluated as a nonlinear learning model capable of approximating complex relationships between historical temperature information and future anomalies.

The network receives the same input feature vector:

$$\mathbf{x}_{c,t} = [y_{c,t-1}, y_{c,t-2}, (t - t_0), MA_5(t)]^\top. \quad (11)$$

The MLP prediction can be represented as:

$$\hat{y}_{c,t} = \mathbf{W}_2 \sigma(\mathbf{W}_1 \mathbf{x}_{c,t} + \mathbf{b}_1) + b_2. \quad (12)$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ represent the network weights, $\mathbf{b}_1$ and b_2 are bias terms, and $\sigma(\cdot)$ denotes the nonlinear activation function.

This structure enables the model to capture potential nonlinear interactions among temporal persistence, warming trends, and recent temperature evolution.

ARIMA

The ARIMA model captures autoregressive and moving average dynamics:

$$y_t = \sum_{i=1}^p \phi_i y_{t-i} + \sum_{j=1}^q \theta_j \epsilon_{t-j} + \epsilon_t, \quad (13)$$

where y_t represents the observed temperature anomaly at time t , ϕ_i are autoregressive coefficients, θ_j are moving-average coefficients, ϵ_t denotes the residual error term, p is the autoregressive order, and q is the moving-average order. The differencing order is denoted by d . Model orders were selected via temporal cross-validation.

Stationarity Analysis

Since ARIMA assumes stationarity (or stationarity after differencing), the country-level temperature anomaly series were evaluated using both the Augmented Dickey–Fuller (ADF) and KPSS tests. The ADF test evaluates the null hypothesis of a unit root (non-stationarity), whereas the KPSS test assumes stationarity under the null hypothesis. Table 2 summarizes the percentage of country-level series classified as stationary before and after first-order differencing.

The results indicate that most country-level series are non-stationary in their original form but become stationary after first-order differencing. Therefore, the ARIMA hyperparameter search included both $d = 0$ and $d = 1$, allowing the differencing order to be selected objectively through temporal cross-validation.

Table 2. Summary of stationarity tests for the country-level temperature anomaly series.

Stationarity Test	Stationary Series (%)
ADF (original series)	13.0
KPSS (original series)	3.6
ADF (first difference)	94.2
KPSS (first difference)	90.3

Autocorrelation Analysis

To further characterize the temporal dependence structure of the temperature anomaly series, both the autocorrelation function (ACF) and the partial autocorrelation function (PACF) were analyzed using the average annual temperature anomaly across all countries (Figure 7).

The ACF exhibits a slow positive decay across multiple lags, indicating strong temporal persistence and supporting the presence of non-stationary behavior in the original series. This observation is consistent with the stationarity analysis based on the ADF and KPSS tests. In contrast, the PACF displays a dominant spike at lag 1, while the remaining lags remain comparatively small and mostly within the confidence bounds. This behavior suggests that short-order autoregressive structures are sufficient to capture most of the linear temporal dependence, providing additional justification for considering low-order autoregressive terms during ARIMA model selection.

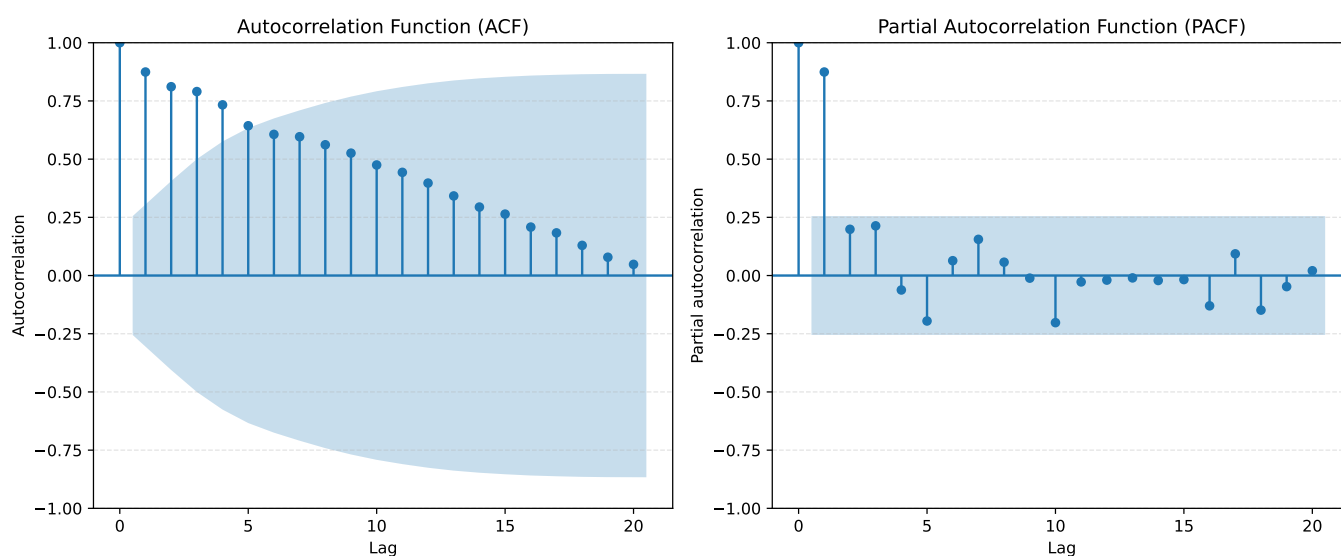


Figure 7. Autocorrelation function (ACF) and partial autocorrelation function (PACF) computed from the average annual temperature anomaly series. The ACF reveals strong temporal persistence, whereas the PACF indicates that most autoregressive dependence is concentrated at the first lag.

The differencing order was selected from candidate models with both $d = 0$ and $d = 1$, whereas the complete parameter selection procedure is described in the Hyperparameter tuning Section 3.9.

ETS

The ETS (Error–Trend–Seasonality) model decomposes the observed time series into level, trend, and error components according to the exponential smoothing framework. In its additive formulation, the model is expressed as

$$y_t = \ell_t + b_t + \epsilon_t, \quad (14)$$

where ℓ_t denotes the level component, b_t represents the trend component, and ϵ_t is the random error term. ETS models are particularly suitable for capturing smooth trend evolution and persistent temporal dynamics without requiring explicit autoregressive terms.

The model was implemented using the ExponentialSmoothing class from the statsmodels library. Since the dataset consists of annual temperature anomalies, no seasonal component was considered. Candidate configurations included additive and no-trend specifications, both initialized using the estimated initialization method. Hyperparameter selection was performed using the same expanding-window temporal cross-validation framework adopted for the remaining forecasting models. For each country, the ETS configuration yielding the lowest mean cross-validation RMSE was retained for the final evaluation. Forecasts were generated recursively using a one-step-ahead forecasting strategy.

To avoid restricting the ETS framework to simple exponential smoothing, the model selection procedure explicitly evaluated both level-only and additive-trend ETS formulations. The final ETS configuration for each country-level temperature anomaly series was selected using expanding-window temporal cross-validation according to the lowest RMSE. The resulting selected configurations are summarized in Table 3.

Table 3. Selection frequency of ETS configurations using expanding-window temporal cross-validation.

ETS Configuration	Description	Countries	Percentage
ETS(A,A,N)	Holt additive trend exponential smoothing	188	66.67%
ETS(A,N,N)	Level-only simple exponential smoothing	94	33.33%

The predominance of the ETS(A,A,N) specification indicates that the final ETS benchmark was not equivalent to simple exponential smoothing. Instead, the framework allowed the inclusion of an explicit trend component to represent long-term changes whenever supported by temporal cross-validation.

The benchmark intentionally includes forecasting methods with different levels of complexity and modeling assumptions. The objective of this study is not to compare only models specifically designed for trending climate series, but to evaluate whether increasing methodological complexity systematically translates into improved predictive performance under a common forecasting framework. Accordingly, the benchmark comprises a simple baseline approach (Persistence), feature-based linear regression models (OLS and Ridge), classical statistical time-series models (ETS and ARIMA), and nonlinear machine-learning methods (Random Forest, SVR, and MLP). Whenever applicable, model configurations were selected using expanding-window temporal cross-validation and implemented according to recommended practices for annual temperature anomaly forecasting. This benchmark design enables a balanced comparison between model complexity, predictive accuracy, and computational cost under a consistent evaluation protocol.

3.4. Hyperparameter Tuning

Hyperparameters for machine learning models (Ridge, SVR, Random Forest, MLP) were optimized using grid search combined with temporal cross-validation.

A subset of countries was used to estimate robust configurations, and the most frequently selected parameter combinations were retained for the final evaluation.

For ARIMA and ETS models, parameter configurations (orders and trend specification) were selected via rolling-origin cross-validation.

3.5. Training and Validation Protocol

Model evaluation followed two complementary strategies:

- A temporally ordered 80/20 train–test split for baseline comparison
- Temporal cross-validation using expanding window splits (TimeSeriesSplit)

The latter ensures robustness of results across multiple temporal partitions and reduces dependence on a single split.

3.6. Performance Metrics

Performance was evaluated using:

- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- Coefficient of Determination (R^2)

Country-level results were aggregated using median and interquartile range (IQR).

Additionally, a win-rate metric was computed, defined as the proportion of countries where a model achieved the lowest RMSE.

3.7. Statistical Comparison and Computational Cost

Pairwise statistical comparisons were performed using the Wilcoxon signed-rank test on country-level RMSE differences.

For each model pair, the following statistics were computed:

- Median Δ RMSE
- Fraction of countries with Δ RMSE > 0
- Associated p -value (Holm-adjusted)

The computational cost was assessed by documenting the running time for each model across all countries, facilitating the analysis of the trade-off between predictive precision and effectiveness.

3.8. Implementation Details

All experiments were implemented in Python 3.11 using the Pandas, NumPy, Scikit-learn, Statsmodels, and SciPy libraries. Machine learning models were implemented through Scikit-learn pipelines to ensure that feature standardization was performed exclusively on the training data and subsequently applied to the corresponding test data, thereby preventing data leakage. Random seeds were fixed whenever supported by the underlying implementation to improve reproducibility. All experiments followed the same preprocessing, feature engineering, and evaluation protocol to ensure a fair comparison among forecasting models.

The experiments were executed on a workstation equipped with a 13th Gen Intel® Core™ i7-13620H processor (Intel, Santa Clara, CA, USA), 16 GB RAM, and an NVIDIA GeForce RTX 4060 Laptop GPU with 8 GB of dedicated memory (NVIDIA, Santa Clara, CA, USA), running a 64-bit Windows operating system.

Linear and Kernel-Based Models

The Ordinary Least Squares (OLS) model was implemented using the default Linear-Regression estimator from Scikit-learn. Ridge regression employed feature standardization through StandardScaler followed by $\alpha = 10^{-2}$. Support Vector Regression (SVR) was implemented using a radial basis function (RBF) kernel with $C = 1.0$, $\gamma = 1/p$, and $\epsilon = 0.01$, where p denotes the number of predictor variables. Feature standardization was applied to both Ridge and SVR within Scikit-learn pipelines.

Ensemble Learning

The Random Forest model consisted of an ensemble of 200 regression trees with a minimum leaf size of five samples. A fixed random seed of 42 was adopted to ensure reproducibility, while parallel computation was enabled through Scikit-learn's multi-core implementation.

Neural Network

The multilayer perceptron (MLP) model was implemented as a feedforward neural network with a single hidden layer containing 10 neurons and ReLU activation. The network was trained using the Adam optimizer with the default mean squared error (MSE) loss function, an initial learning rate of 0.01, and a maximum of 250 training iterations. Early stopping was enabled to reduce overfitting. Input features were standardized using StandardScaler within a Scikit-learn pipeline, and a fixed random seed of 42 was used to improve reproducibility.

Time-Series Models

The ARIMA and ETS models were implemented using the statsmodels library. ARIMA models were fitted using recursive one-step-ahead forecasting, whereas ETS employed additive exponential smoothing without seasonal components and estimated initialization parameters. Both statistical forecasting models were evaluated under the same temporal validation framework as the machine learning models. The complete hyperparameter search space, temporal cross-validation protocol, forecasting horizon, and model selection procedure are described in the Hyperparameter tuning subsection.

3.9. Hyperparameter Tuning

Hyperparameters for Ridge, SVR, Random Forest, MLP, ARIMA, and ETS were optimized using temporal cross-validation based on three expanding-window splits (TimeSeriesSplit), thereby preserving the chronological order of the observations. For each country, the candidate configuration yielding the lowest mean cross-validation root mean squared error (RMSE) was selected. For machine learning models (Ridge, SVR, Random Forest, and MLP), hyperparameter optimization was performed using GridSearchCV. In contrast, ARIMA and ETS configurations were evaluated through a manual rolling-origin procedure using recursive one-step-ahead forecasting.

The search space consisted of $\alpha \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ for Ridge regression; $C \in \{0.1, 1, 10\}$, $\gamma \in \{1/p, 0.1, 1\}$, and $\epsilon \in \{0.01, 0.05, 0.1\}$ for SVR; 100 and 200 trees, maximum depths of {None, 5, 10}, and minimum leaf sizes of {3, 5, 10} for Random Forest; and one hidden layer with {5, 10, 20} neurons together with learning rates of {0.001, 0.01} for the MLP. For ARIMA, the evaluated candidate orders were (1, 0, 0), (1, 1, 0), (0, 1, 1), (1, 1, 1), and (2, 1, 1), thereby considering both $d = 0$ and $d = 1$. ETS models evaluated additive and no-trend specifications. The configuration yielding the lowest mean cross-validation RMSE was retained for each country.

Table 4 summarizes the most frequently selected hyperparameter configuration for each model across the evaluated countries.

The reported frequencies should be interpreted solely as descriptive statistics indicating how often a given configuration was selected during the tuning process. They do not imply universal optimality, statistical superiority, or causal relationships, since the optimal configuration remains dependent on the temporal behavior of each country-level temperature anomaly series.

Table 4. Most frequently selected hyperparameter configurations obtained through temporal cross-validation.

Model	Recommended Hyperparameters	Median CV RMSE	Mean CV RMSE	Std CV RMSE	Frequency
Ridge	$\alpha = 0.1$	0.2862	0.2833	0.0501	5
SVR	RBF kernel; $C = 0.1$; $\gamma = 1/p$; $\epsilon = 0.1$	0.3340	0.3710	0.1014	4
Random Forest	max depth = 5; minimum samples per leaf = 10; random seed = 42	0.3696	0.4092	0.0937	11
MLP	1 hidden layer with 10 neurons; ReLU; Adam; learning rate = 0.01; maxi- mum iterations = 250; random seed = 42	0.3481	0.3790	0.0820	17
ARIMA	Order (2, 1, 1)	0.2856	0.3150	0.1164	50
ETS	Additive trend	0.3143	0.3827	0.1617	227

4. Results

This part shows how the different forecasting methods that were tested compare to each other in the real world. Results are sorted by how accurate the predictions are, how well the model works, how much it costs to run, how much each feature adds, how statistically significant they are, and how well they work over time.

4.1. Predictive Performance Across Models

Figure 8 shows how RMSE values (measured in °C) differ between countries for each model, while Table 5 summarizes the corresponding performance statistics, including RMSE and MAE values reported in °C and runtime metrics reported in seconds (s).

The results show that classical time-series models are the most accurate overall. ETS achieves the lowest median RMSE (0.3388 °C) and the smallest interquartile range, indicating both high predictive accuracy and stability across heterogeneous country-level series. ARIMA follows with a median RMSE of 0.3514 °C, although with greater variability across countries.

Table 5. Summary of model performance across countries.

Model	RMSE Median	RMSE IQR	MAE Median	R2 Median	Runtime Total	Runtime Median
ETS	0.3388	0.2016	0.2792	−0.0121	27.4408	0.1030
ARIMA	0.3514	0.2285	0.2838	−0.1612	77.1010	0.2717
OLS	0.3695	0.2578	0.3018	−0.0758	0.5779	0.0020
Ridge	0.3695	0.2579	0.3011	−0.0756	0.8118	0.0030
Persistence	0.3997	0.2856	0.3275	−0.5897	0.0020	0.0000
Random Forest	0.4473	0.2584	0.3724	−0.7623	54.6724	0.1907
MLP	0.5573	0.3743	0.4718	−1.5411	4.1845	0.0110
SVR	0.5670	0.3194	0.4776	−1.7466	0.8225	0.0030

The linear models (OLS and Ridge) remain competitive, with median RMSE values around 0.3695 °C. Although slightly less accurate than ETS, these models exhibit strong robustness and low variability, suggesting that a substantial fraction of the predictive signal is captured through relatively simple linear relationships.

In contrast, machine learning approaches such as Random Forest, MLP, and SVR perform substantially worse, exhibiting larger median errors and greater variability. These findings suggest that increasing model flexibility does not necessarily improve forecasting performance under the evaluated setting.

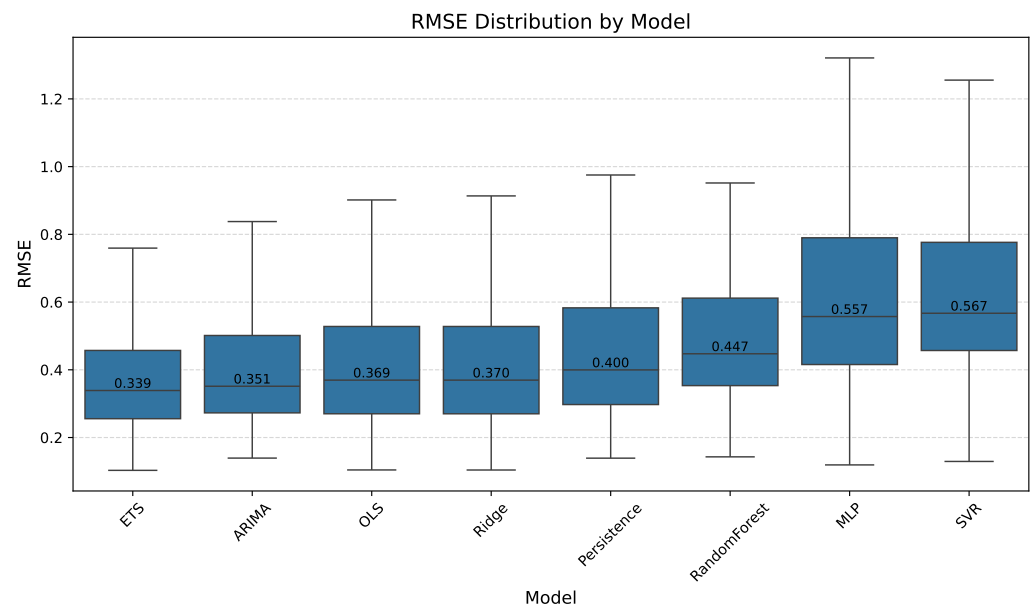


Figure 8. Distribution of per-country RMSE values under the temporal 80/20 train–test split.

4.2. Model Dominance and Robustness

We measure model dominance by a per country win-rate measure. Table 6 and Figure 9 show that ETS achieves the highest win-rate (40.07%), significantly outperforming all competing models in terms of frequency of best performance.

Although ETS achieved the best relative performance according to RMSE and win-rate, the median R^2 values remained negative across all evaluated models. This indicates that, under the selected feature representation and validation framework, none of the approaches consistently outperformed a simple mean-based predictor in terms of explained variance.

Interestingly, two linear models (OLS and Ridge) have lower win-rates but are still competitive, showing consistent instead of dominant achievement across countries.

Table 6. Per-country win-rate based on lowest RMSE.

Model	Countries Won	Win-Rate (%)
ETS	113	40.07
OLS	55	19.50
Ridge	35	12.41
Persistence	22	7.80
MLP	20	7.09
Random Forest	16	5.67
ARIMA	14	4.96
SVR	7	2.48

For robustness evaluation, temporal cross-validation was used. Table 4 shows ARIMA and Ridge have the lowest median RMSE under cross-validation, while ETS has higher variability. That is, ETS performs well for specific cases, but may be less stable on temporal splits.

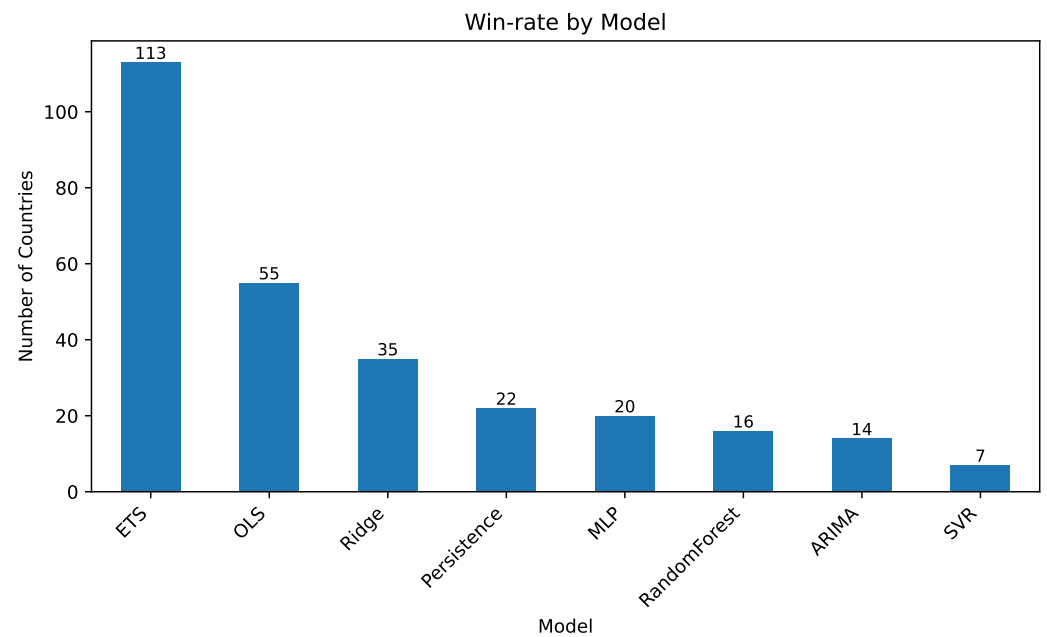


Figure 9. Per-country win-rate based on lowest RMSE.

4.3. Computational Cost and Complexity

The computational cost results (Figure 10) show that Random Forest and ARIMA are significantly more expensive than the other models. Linear models, on the other hand, can reach this level of precision at a small percentage of the computational expenses.

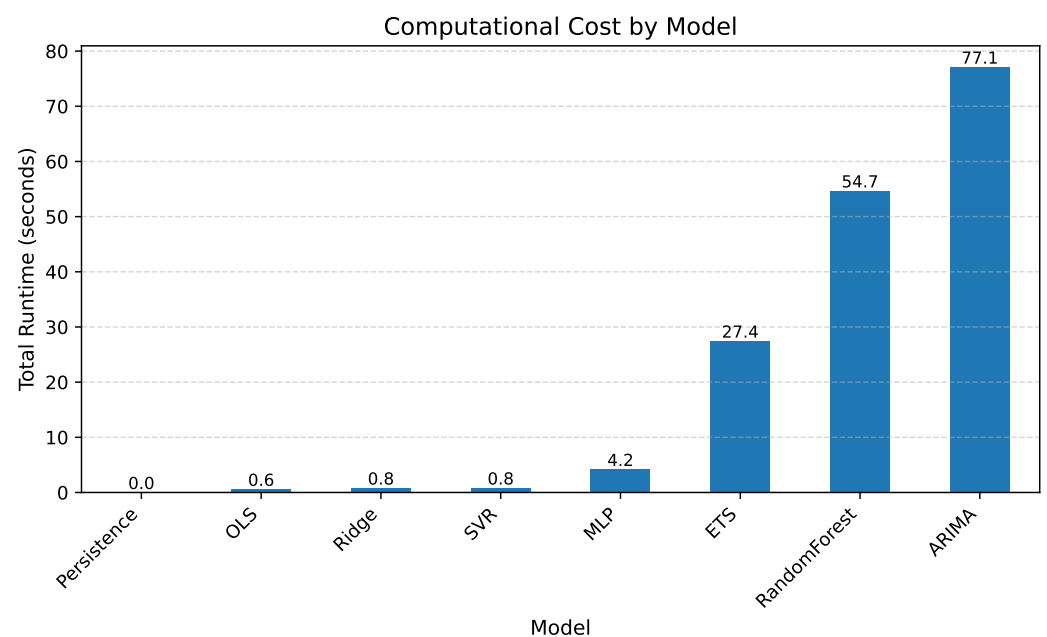


Figure 10. Computational cost across models.

4.4. Feature Contribution Analysis

The ablation study shows (Figure 11) that most of the predictive signal is contained in lagged variables. Adding trends and moving averages as extra features leads to only small improvements.

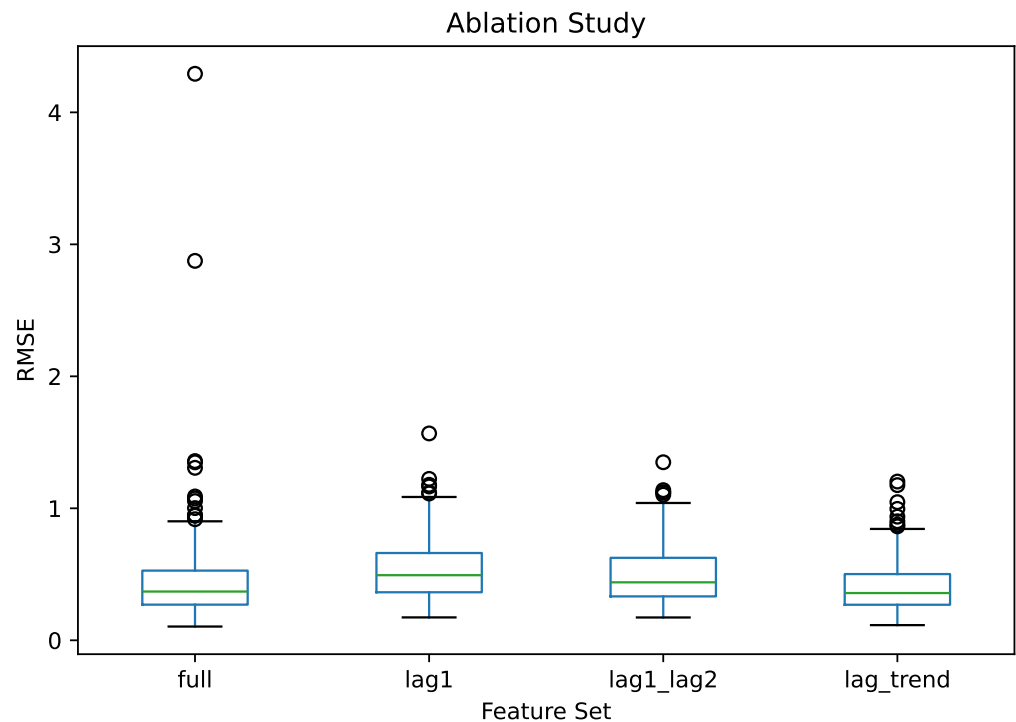


Figure 11. Ablation analysis of feature contributions. The boxplots represent the RMSE distribution for each feature subset, and the circles denote outlier observations.

4.5. Statistical Comparison of Models

Pairwise statistical comparisons were performed using the Wilcoxon signed-rank test with Holm correction for multiple comparisons. Figure 12 provides a global overview of the statistical evidence among all forecasting models through the adjusted p -value matrix. The color scale represents $-\log_{10}(\text{adjusted } p)$, where larger values indicate stronger statistical evidence against the null hypothesis of equal predictive performance.

The heatmap combines statistical significance information and effect-size interpretation in a single visualization. The color intensity represents the strength of the statistical evidence based on Holm-adjusted p -values, whereas the numerical annotations inside each cell correspond to the absolute median ΔRMSE between each pair of forecasting models.

To facilitate the interpretation of the pairwise comparisons, Table 7 summarizes the statistical results obtained by comparing the best-performing model (ETS) against the remaining forecasting approaches. The median ΔRMSE quantifies the typical signed performance difference between models, whereas the fraction of countries with $\Delta\text{RMSE} > 0$ indicates the proportion of country-level series for which the competing model achieved a lower RMSE than ETS. The adjusted p -values confirm that all observed differences are statistically significant after Holm correction.

Unlike the heatmap annotations, which report absolute median ΔRMSE values to facilitate visual comparison of effect magnitudes, Table 7 preserves the sign of ΔRMSE to indicate the direction of the performance difference between models.

The paired performance difference was defined as

$$\Delta\text{RMSE} = \text{RMSE}_A - \text{RMSE}_B,$$

where negative values indicate that Model A achieved lower prediction error than Model B.

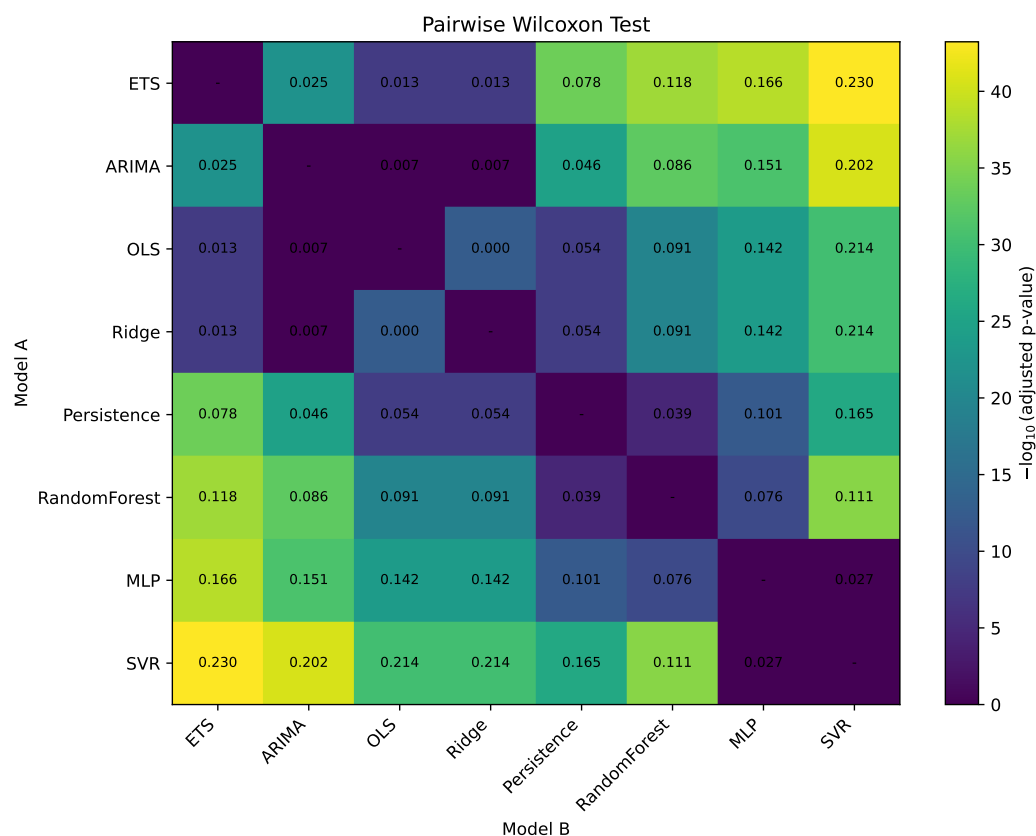


Figure 12. Heatmap of Holm-adjusted pairwise Wilcoxon tests. Colors represent $-\log_{10}$ of the adjusted p -values, where higher values indicate stronger statistical evidence against the null hypothesis of equal predictive performance. The annotated values correspond to the absolute median ΔRMSE between each pair of models.

Table 7. Summary of pairwise Wilcoxon signed-rank comparisons between ETS and the remaining forecasting models.

Model A	Model B	Median ΔRMSE	Fraction of Countries with $\Delta\text{RMSE} > 0$	Adjusted p -Value
ETS	ARIMA	−0.025	0.206	<0.001
	OLS	−0.013	0.401	
	Ridge	−0.013	0.401	
	Persistence	−0.078	0.131	
	Random Forest	−0.118	0.131	
	MLP	−0.166	0.128	
	SVR	−0.230	0.057	

Overall, the statistical analysis confirms that ETS consistently outperformed the remaining forecasting models. The largest performance improvements were observed with respect to SVR and MLP, whereas the smallest median differences were obtained against the linear models (OLS and Ridge) and ARIMA. Although these latter differences were numerically smaller, they remained statistically significant after multiple-comparison correction, indicating that the superior predictive performance of ETS was consistently observed across the evaluated country-level temperature anomaly series.

5. Discussion

This paper provides an empirical evaluation of the accuracy-complexity tradeoff in the prediction of country-level temperature anomalies. The results paint a nuanced picture where model performance is heavily influenced by the evaluation criterion.

A key observation is that the evaluated models produced negative median R^2 values under the temporally held-out test evaluation. Negative values occur when the squared prediction error exceeds the variance of the test observations around their mean.

Therefore, the negative median R^2 values indicate that, for a considerable fraction of country-level series, the improvement over a test-set mean predictor was limited according to this variance-based metric. This result highlights the difficulty of annual temperature anomaly forecasting using only compact autoregressive and trend-based predictors. However, it does not invalidate the relative comparison among forecasting approaches, since model ranking was primarily assessed using forecasting-oriented metrics such as RMSE and MAE, country-level win-rate analysis, and paired statistical comparisons.

Consequently, the results should be interpreted as a comparative assessment of model accuracy, robustness, and computational efficiency rather than as evidence of high explanatory power in terms of variance reconstruction.

Meanwhile, linear models with standard features such as OLS and Ridge continue to be highly competitive. However, they are not dominant in terms of win-rate, but they show solid results across countries with relatively small error dispersion. Ridge reaches similar performance than ARIMA under the temporal cross validation setting, but with a much lower computational cost. This implies that a large part of the foreseeable signal can be represented by simple autoregressive framework and trend components.

The Persistence baseline is also better than expected, which corroborates the strong temporal dependence present in annual temperature anomaly series. However, its inferior overall performance compared with ETS, ARIMA, and linear models demonstrates that inclusion of trend and smoothing information provides additional predictive value.

By comparison, the non-linear machine learning frameworks explored here do not produce systematic improvements. Random Forest, SVR, and MLP exhibit higher errors, greater variability, and substantially higher computational costs. This suggests that the additional flexibility of these models mainly increases variance without sufficiently reducing bias, which is consistent with the classical bias–variance tradeoff in time-series forecasting under limited feature representations.

An important implication of these findings is that forecasting performance depends not only on the theoretical assumptions of each model, but also on their practical suitability for the available data. Although the benchmark includes methods based on different modeling principles, all models were implemented using configurations appropriate for annual temperature anomaly forecasting and selected through temporally consistent validation procedures. The results therefore suggest that increasing methodological complexity does not necessarily yield better predictive performance under the considered feature representation. Instead, relatively simple statistical and linear approaches remain highly competitive, emphasizing that appropriate model selection and validation may be more influential than algorithmic complexity alone.

The ablation study further supports this conclusion, showing that most of the predictive signal is captured by lagged variables. Additional features such as trend and moving averages provide only modest improvements, indicating limited exploitable nonlinear structure under the current forecasting setup.

Statistical comparison using the Wilcoxon signed-rank test provides further evidence that ETS consistently outperforms several competing approaches, particularly Random Forest, SVR, MLP, and Persistence. Differences relative to linear models are smaller but still statistically significant, suggesting that linear methodologies remain highly competitive despite the stronger predictive accuracy achieved by ETS.

Temporal cross-validation results also demonstrate that model rankings may vary depending on the evaluation strategy. ETS achieves the best overall win-rate under the

80/20 temporal split, while ARIMA and Ridge exhibit stronger robustness across rolling temporal partitions. This highlights the importance of evaluating forecasting models under multiple temporally consistent validation schemes.

At a practical level, computational efficiency is an important consideration. Random Forest and ARIMA are considerably more expensive, whereas OLS and Ridge achieve competitive predictive performance with minimal runtime requirements. ETS provides an attractive balance between predictive precision and computational cost under the evaluated conditions.

There are several limitations that should be acknowledged. First, the analysis intentionally relies on a compact and interpretable feature set based on lagged temperature values, temporal trend, and moving averages. While this facilitates a controlled benchmarking framework, it may favor linear models and limit the representational advantages of more complex nonlinear approaches. Furthermore, annual-scale temperature anomaly prediction is influenced by external forcings, internal climate variability, and anthropogenic factors that are not explicitly incorporated into the present feature representation. Consequently, the predictive rankings and conclusions reported in this study are conditioned on the selected variables, temporal aggregation level, and evaluation framework. Different predictor configurations, temporal resolutions, or additional physical climate variables may therefore lead to different forecasting outcomes and model rankings. Future work should explore multivariate forecasting settings incorporating physical climate indicators, external forcings, and socio-environmental variables.

Second, the current framework does not explicitly address uncertainty quantification. Future studies may incorporate probabilistic forecasting strategies, confidence intervals, or Bayesian approaches to better characterize predictive uncertainty under climate variability. Third, although the clustering analysis identified meaningful groups of countries with different warming dynamics, model performance differences across clusters were not explored in detail and deserve further investigation.

In general, the results indicate that classical time-series models, particularly ETS, provide the strongest predictive performance under the evaluated framework, while linear models constitute highly efficient and robust alternatives with substantially lower computational cost. Increasing model complexity does not systematically improve predictive performance, highlighting the importance of model parsimony in annual temperature anomaly forecasting.

6. Conclusions

In this study, a comparative evaluation of multiple forecasting techniques for country-level annual temperature anomaly prediction was conducted using FAOSTAT data. The analysis compared classical time-series models, linear methods, and machine learning approaches across a large set of countries, providing evidence regarding the trade-off between predictive accuracy, robustness, interpretability, and computational cost.

The results indicate that classical time-series models provide the strongest predictive performance under the evaluated framework. In particular, ETS achieves the lowest median RMSE and the highest country-level win-rate, indicating consistent predictive reliability across heterogeneous countries. ETS emerged as the most competitive method within the evaluated benchmark, offering the lowest median RMSE and the highest win-rate. Nevertheless, the negative median R^2 values indicate that substantial room for improvement remains in achieving strong absolute forecasting accuracy for annual temperature anomalies. ARIMA also demonstrates competitive performance, especially under temporal cross-validation where it achieves the lowest median RMSE.

Linear models such as OLS and Ridge remain highly competitive despite their simplicity. These approaches exhibit low variability, strong robustness, and extremely low computational cost. In particular, Ridge performs comparably to ARIMA under temporal cross-validation, suggesting that a substantial portion of the predictive signal can be captured through relatively simple autoregressive and trend-based representations.

The study also highlights the importance of employing multiple evaluation criteria. Although ETS achieves the strongest overall accuracy and win-rate, cross-validation results suggest that ARIMA and Ridge may offer greater robustness across temporal partitions. This demonstrates that model rankings can be sensitive to the selected evaluation framework and reinforces the importance of temporally consistent validation strategies.

Overall, the results indicate that increasing model complexity does not systematically improve predictive performance for annual temperature anomaly forecasting under the evaluated feature representation. Instead, the findings emphasize the importance of model parsimony, where simpler linear and classical time-series models often provide the most reliable balance between predictive accuracy, robustness, interpretability, and computational efficiency.

Importantly, these findings should not be interpreted as evidence against the existence of long-term global warming trends. The exploratory analysis and temporal behavior of the FAOSTAT series clearly reveal persistent warming dynamics across multiple countries and regions. Rather, the results suggest that, within the evaluated forecasting framework and selected feature representation, much of the short-term predictive signal can be effectively captured through temporal persistence and autoregressive dependencies.

It is important to note that statistical relationships derived from time-series data should be interpreted carefully. In particular, correlation-based analyses may be affected by temporal dependence, autocorrelation, and common trends, which can lead to misleading associations if the temporal structure of the data is not properly considered. Future extensions of this work should further investigate these aspects using time-series-specific dependence measures and additional statistical validation approaches.

From a practical perspective, model selection should depend not only on predictive accuracy, but also on computational efficiency, interpretability, robustness, and deployment constraints. Linear models such as OLS and Ridge provide competitive performance with minimal computational requirements, while ETS offers the strongest predictive accuracy under the evaluated setting. Consequently, the proposed framework should be interpreted primarily as a comparative benchmarking methodology rather than a definitive operational climate forecasting system for real-world deployment.

Furthermore, the conclusions presented in this work are conditioned on the selected feature representation, temporal aggregation level, and evaluation framework. Since annual-scale temperature anomaly prediction is influenced by external forcings, internal climate variability, and anthropogenic factors, future research should explore richer multivariate forecasting settings incorporating physical climate indicators, external climate forcings, uncertainty quantification strategies, and broader validation protocols.

In addition, future research could evaluate uncertainty-aware forecasting approaches, including quantile regression and Bayesian neural networks. These methods would allow the estimation of predictive uncertainty intervals rather than only point forecasts, providing additional information regarding forecast reliability and the range of possible temperature anomaly outcomes.

These findings provide a strong foundation for future research on interpretable and computationally efficient climate forecasting methodologies.

Author Contributions: Conceptualization, R.R.-G. and O.R.-A.; methodology, R.R.-G. and L.E.G.-G.; software, R.R.-G.; validation, R.R.-G., J.S.A.-H. and J.R.-R.; formal analysis, R.R.-G. and L.E.G.-G.; investigation, R.R.-G.; resources, O.R.-A.; data curation, R.R.-G. and J.S.A.-H.; writing—original draft preparation, R.R.-G.; writing—review and editing, R.R.-G., L.E.G.-G., J.S.A.-H., J.R.-R. and O.R.-A.; visualization, R.R.-G.; supervision, O.R.-A. and J.R.-R.; project administration, O.R.-A.; funding acquisition, O.R.-A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

OLS	Ordinary Least Squares
Ridge	Ridge Regression
SVR	Support Vector Regression
MLP	Multilayer Perceptron
ARIMA	AutoRegressive Integrated Moving Average
ETS	Error, Trend, Seasonality
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
CV	Cross-Validation
FAOSTAT	Food and Agriculture Organization Statistical Database
R ²	Coefficient of determination

References

1. Kazemzadeh, M.; Noori, Z.; Jamali, S.; Abdi, A.M. Forty Years of Air Temperature Change over Iran Reveals Linear and Nonlinear Warming. *J. Meteorol. Res.* **2022**, *36*, 462–477. [\[CrossRef\]](#)
2. Wan, H.; Spassiani, A.C.; Vincent, L.A. Canada's Fourth Generation of Homogenized Surface Air Temperature and its Trends for 1948–2023. *Atmos.-Ocean* **2025**, *63*, 223–240. [\[CrossRef\]](#)
3. Nida, H.; Kashif, M.; Janjua, A.A.; Aslam, M. Impacts of Climate Change on Pakistan's Weather Patterns: A Comprehensive Study of Temperature and Precipitation Trends. *Environ. Monit. Assess.* **2025**, *197*, 509. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Sein, Z.M.; Zhi, X.; Ogou, F.K.; Nooni, I.K. Covariability of Decadal Surface Air Temperature Variability over Myanmar with Sea Surface Temperature. *Environ. Res. Lett.* **2024**, *19*, 044056. [\[CrossRef\]](#)
5. Derakhshandeh, A.; Khoorani, A.; Rezazadeh, M. Trend Analysis of Precipitation in Iran Based on MERRA2. *J. Earth Space Phys.* **2023**, *49*, 669–683. [\[CrossRef\]](#)
6. Alghamdi, A.S.; Harrington, J.A. The Symmetry and Structural Characteristics of the Core of the Arabian Subtropical Anticyclone and Associated Surface Climatology over Arabia. *Atmos. Res.* **2023**, *281*, 106482. [\[CrossRef\]](#)
7. Mohammadi, N.; Hejazizadeh, Z.; Zeaiean Firouzabadi, P. Detection of the Effect of Climate Change on the Mechanism of Heat Islands in Tehran Province. *Nivar* **2024**, *48*, 203–224. [\[CrossRef\]](#)
8. de Andres, J.M.; D'Elia, I.; Paz, D. City-Scale Revegetation Strategies Impact on Temperature-Related Mortality. *Forests* **2025**, *16*, 1089. [\[CrossRef\]](#)
9. Qi, J.; Zhang, J.; Wang, Y.; Huang, J. Temporal Variation in the Effects of Extreme Temperature on Respiratory Mortality. *Environ. Int.* **2024**, *189*, 108800. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Günsche, S.; Borg, M.A.; Anikeeva, O. Mortality, Morbidity and Healthcare Costs of Short-Term High Temperatures and Heatwaves Exposure in Older Populations. *Environ. Int.* **2026**, *208*, 110129. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Joseph, M.; Moonsammy, S.; Davis, H. Modelling Climate Variabilities and Global Rice Production. *Heliyon* **2023**, *9*, e15480. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Ahmed, M.; Saha, P. Economic Impact of Climate Change on Agriculture. *Nat. Environ. Pollut. Technol.* **2023**, *22*, 2173–2179. [\[CrossRef\]](#)

13. Acci, Y.; Ucar, E.; Ucar, M. Evaluating the Relationship between Climate Change, Food Prices, and Poverty. *Environ. Dev. Sustain.* **2025**, *27*, 28061–28085. [[CrossRef](#)]
14. Yañez-Arenas, A.; Lluch-Cota, D.B. Global Empirical Assessment of Marine Heatwaves Impacts on Fisheries Production and Economic Value. *Fishes* **2025**, *10*, 594. [[CrossRef](#)]
15. Chen, X.; Jiang, Z.; Cheng, H. A Novel Global Average Temperature Prediction Model Based on GM-ARIMA Combination Model. *Earth Sci. Inform.* **2024**, *17*, 853–866. [[CrossRef](#)]
16. Rasyid, F.R.; Adytia, D. Time Series Temperature Forecasting Using ConvLSTM. *J. RESTI* **2022**, *6*, 563–569. [[CrossRef](#)]
17. Haghramani, S.H. Enhancing Temperature Prediction in the UAE Using GRU-CNN Hybrid Models. *Model. Earth Syst. Environ.* **2025**, *11*, 100. [[CrossRef](#)]
18. Aydin, S. Predicting Climate Change with Artificial Neural Networks: The Case of the G20. *Int. J. Knowl. Process Stud.* **2025**, *5*, 18–44. [[CrossRef](#)]
19. Aboulhosn, S.; Akkawi, M.; Kadry, S.N. Comparative Evaluation of Machine Learning Models in Forecasting Crop Yields amid Climate Change. In *Lecture Notes in Networks and Systems*; Springer: Cham, Switzerland, 2026; Volume 1668, pp. 207–217. [[CrossRef](#)]
20. Tesfaye, K.; Takele, R.; Shelia, V. High Spatial Resolution Seasonal Crop Yield Forecasting for Heterogeneous Maize Environments in Ethiopia. *Clim. Serv.* **2023**, *32*, 100425. [[CrossRef](#)]
21. Corrales-Suastegui, A.; Gonzalez-Gonzalez, M.A. Statistical Forecast of GDDs and CHs for Northern Mexico Improved by Bias Correction. *Rev. Mex. Cienc. Agric.* **2025**, *16*, e4048. [[CrossRef](#)]
22. Aloyce, A.A. Climate Change and Plant Protection: Challenges and Innovations in Disease Forecasting Systems. *J. Plant Prot. Res.* **2025**, *65*, 174–186. [[CrossRef](#)]
23. Alif, H.A. Temperature-Based Solar Energy Forecasting: A Big Data Analysis for Sustainable Energy Planning. *Theor. Appl. Climatol.* **2025**, *156*, 526. [[CrossRef](#)]
24. Gler, M.; Turner, A.; Hirons, L. Systematic Biases over the Equatorial Indian Ocean and Their Influence on Seasonal Forecasts. *Clim. Dyn.* **2025**, *63*, 328. [[CrossRef](#)] [[PubMed](#)]
25. Sun, Z.; Liu, Y.; Chen, H. Evaluation of Future Climatology and Uncertainty Under SSP Scenarios. *Atmos. Res.* **2024**, *298*, 107134. [[CrossRef](#)]
26. Jiang, Z.; Choudhury, D.; Sharma, A. Spectrally Transformed CMIP6 Decadal Projections Improve Rainfall Forecasts. *J. Hydrol.* **2025**, *655*, 132870. [[CrossRef](#)]
27. Tigona, R.; Ongoma, V.; Weir, T. Towards Improved Seasonal Rainfall Prediction in the Tropical Pacific Islands. *Theor. Appl. Climatol.* **2023**, *154*, 349–363. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.