Int. J. Adv. Eng. Pure Sci. 2026, 38(3): <565-581>
DOI: 10.7240/jeps.1970358

RESEARCH ARTICLE / ARAŞTIRMA MAKALESİ

Enhancing Meta Fuzzy Functions via Dynamic Pruning: A Nationwide Air Temperature Forecasting Framework for Türkiye

Meta-Bulanık Fonksiyonların Dinamik Budama ile Geliştirilmesi: Türkiye İçin Ulusal Hava Sıcaklığı Tahmin Çerçevesi

Erdinç KARAKULLUKÇU[1]

[1] *Karadeniz Technical University, Faculty of Science, Department of Computer Science, 61080, Trabzon, Türkiye*

Abstract

Accurate air temperature forecasting across heterogeneous climatic regions remains a significant operational challenge. While existing studies predominantly focus on either single-model architectures or standard ensemble averaging, often evaluated on limited geographical scopes, these approaches face two critical limitations: first, they indiscriminately include all available predictors, allowing poorly performing or noise-inducing models to contaminate the final forecasts; and second, they lack the structural flexibility to adapt across diverse microclimates. Consequently, there is a pressing need for an adaptive forecasting architecture that systematically filters out unreliable predictors while dynamically adjusting to regional climatic patterns. To bridge this gap, this study proposes a novel Pruned Meta-Fuzzy Functions (PMFF) framework to forecast nationwide air temperatures across all 81 cities of Türkiye. The PMFF architecture integrates a heterogeneous pool of 10 base forecasting models encompassing statistical baselines, machine learning algorithms, and deep neural networks. The framework first executes a validation-driven Dynamic Pruning mechanism to eliminate divergent and noise-inducing predictors tailored to each city's microclimate. Subsequently, a Fuzzy C-Means (FCM) clustering strategy is applied to the meta-matrix of surviving predictions to adaptively determine continuous membership-based aggregation weights. Evaluated on an extensive multi-decadal weekly dataset (2010–2026), the proposed PMFF framework consistently outperforms individual base predictors and conventional fusion schemes, as corroborated by non-parametric Friedman and post-hoc Wilcoxon signed-rank tests ($p<0.0001$). The results demonstrate that coupling noise-resilient pruning with soft fuzzy meta-learning provides a highly scalable and robust solution for operational meteorological forecasting.

Keywords: air temperature forecasting, meta-fuzzy functions, dynamic pruning, ensemble learning, time series analysis

Öz

Heterojen iklim bölgelerinde doğru hava sıcaklığı tahmini yapmak, önemli bir operasyonel zorluk olmayı sürdürmektedir. Mevcut çalışmalar ağırlıklı olarak tekil model mimarilerine veya standart topluluk ortalamalarına odaklanırken ve sıklıkla sınırlı coğrafi alanlarda değerlendirilirken, bu yaklaşımlar iki kritik kısıtla karşılaşmaktadır: birincisi, mevcut tüm tahminleyicileri ayrım gözetmeksizin dahil ederek düşük performanslı veya gürültü üreten modellerin nihai tahminleri bozmasına izin vermeleri; ikincisi ise farklı mikro iklimlere uyum sağlayacak yapısal esneklikten yoksun olmalarıdır. Sonuç olarak, bölgesel iklim modellerine dinamik olarak uyum sağlarken güvenilmez tahminleyicileri sistematik olarak filtreleyen uyarlanabilir bir tahmin mimarisine acil bir ihtiyaç duyulmaktadır. Bu eksikliği gidermek amacıyla bu çalışma, Türkiye'nin 81 ilinin tamamında ülke çapında hava sıcaklıklarını tahmin etmek için özgün bir Budanmış Meta-Bulanık Fonksiyonlar (PMFF) çerçevesi önermektedir. PMFF mimarisi; istatistiksel temel modelleri, makine öğrenmesi algoritmalarını ve derin yapay sinir ağlarını kapsayan 10 temel tahmin modelinden oluşan heterojen bir havuzu bir araya getirir. Çerçeve ilk olarak, her şehrin mikro iklimine göre uyarlanmış, ıraksak ve gürültü üreten tahminleyicileri elemek için doğrulama tabanlı bir Dinamik Budama mekanizması çalıştırır. Ardından, sürekli üyelik tabanlı birleştirme ağırlıklarını uyarlanabilir şekilde belirlemek amacıyla, elenmeyip kalan tahminlerin meta-matrisine Bulanık C-Ortalamalar (FCM) kümeleme stratejisi uygulanır. On yılı aşkın kapsamlı bir haftalık veri seti (2010–2026) üzerinde değerlendirilen önerilen PMFF çerçevesi; parametrik olmayan Friedman ve post-hoc Wilcoxon işaretli sıra testlerinin ($p<0.0001$) de doğruladığı üzere, tekil temel tahminleyicilerden ve geleneksel birleştirme şemalarından tutarlı bir şekilde daha yüksek performans göstermektedir. Sonuçlar; gürültüye dayanıklı budama ile esnek bulanık meta-öğrenmenin birleştirilmesinin, operasyonel meteorolojik tahminleme için yüksek düzeyde ölçeklenebilir ve sağlam bir çözüm sunduğunu ortaya koymaktadır.

Anahtar kelimeler: hava sıcaklığı tahmini, meta-bulanık fonksiyonlar, dinamik budama, topluluk öğrenmesi, zaman serisi analizi

Corresponding Author: Erdinç Karakullukçu, **Tel:** 05445516725, **E-mail:** erdinc.karakullukcu@ktu.edu.tr

Submitted: 13.06.2026, **Revised:** 23.08.2026, **Accepted:** 14.09.2026

I.INTRODUCTION

The precise quantitative monitoring of air temperature is a cornerstone of environmental science, exerting a profound influence on diverse sectors ranging from energy resource management and agricultural productivity to the mitigation of climate-related risks [1, 2, 3]. As global climate dynamics become increasingly volatile, the ability to forecast temperature fluctuations with high spatial and temporal accuracy has moved from a scientific preference to a societal necessity [4]. Addressing this challenge requires a deep understanding of meteorological patterns, which often exhibit irregular behaviors that are difficult to capture through simplified observations alone.

For decades, researchers have relied on traditional statistical and machine learning techniques to translate historical observations into future insights. In the context of meteorological applications, Koçak [1] investigated the forecasting capabilities of seasonal ARIMA (SARIMA) against Long Short-Term Memory (LSTM) models for Ankara, providing a foundational benchmark for comparing statistical and neural architectures. Similarly, Aghelpour et al. [5] explored the use of SARIMA and Support Vector Regression (SVR) enhanced by the Firefly optimization algorithm (SVR-FA) to address the limitations of classical models in various climate types. Further developments in the field have been driven by Curceac et al. [6], who utilized non-parametric functional data analysis, and Elshewey et al. [7], who proposed a hybrid model combining Wavelet Decomposition with SARIMAX to improve stability. In parallel, Karabulut and Topçu [8] successfully applied LSTM to temperature series in Kars, demonstrating the efficacy of deep learning in handling chaotic meteorological data.

Complementary to statistical approaches, machine learning has fundamentally reshaped the forecasting landscape by enabling the mapping of complex, non-linear relationships. Alomar et al. [9] demonstrated that a data-driven approach could significantly improve short-term temperature forecasting performance. Apaydın et al. [10] evaluated a wide range of regression methods, including boosting-based algorithms, for air temperature forecasting in Seoul, illustrating the competitive edge of ensemble-based machine learning over static statistical approaches. The adaptability of these methods is further evidenced by Kumaş and Akyüz [11], who utilized Genetic Programming and Random Forest (RF) for Konya, and Kumaş et al. [12], who employed artificial neural networks for Antalya, each contributing unique insights into local climate variability.

As computational capabilities have evolved, the research focus has shifted toward deep learning architectures. Gong et al. [13] leveraged video prediction models to capture the diurnal cycle of temperature, showcasing the potential of deep neural networks to extract spatial-temporal features. Recent studies by Görücü et al. [14], Haque et al. [15], and Wu et al. [16] have further demonstrated the robustness of hybrid network configurations including Temporal Convolutional Networks (TCN) and Convolutional Neural Networks (CNNs). Furthermore, the systematic reviews and modeling efforts by Tran et al. [17], Nketiah et al. [18], and Xue et al. [19] confirm that hybridizing deep learning architectures with decomposition techniques consistently yields superior results compared to baseline methods. Zhang and Dong [4] specifically showcased that Convolutional Recurrent Neural Networks (CRNN) can effectively learn both temporal and spatial correlations, a finding that has influenced modern predictive modeling.

Despite these advancements, the literature largely focuses on the performance of single-model approaches or standard hybridizations. While effective, single models are often susceptible to overfitting and are unable to handle the intrinsic noise present in meteorological datasets [20, 21]. Furthermore, a model that performs exceptionally well in one city may fail in another, suggesting that relying on a monolithic forecasting algorithm is inherently risky [20, 22]. To overcome this, the Meta-Fuzzy Functions (MFF) approach, introduced by Tak [23], provides an elegant solution by utilizing the Fuzzy C-Means (FCM) clustering algorithm to dynamically aggregate the results of multiple methods based on their information content rather than simple averaging [23].

1.1. Multi-City and Regional Forecasting Studies

While a substantial portion of the existing meteorological forecasting literature has evaluated data-driven models on individual cities or single observation stations, an increasing number of recent investigations have recognized the necessity of multi-city, regional, or nationwide predictive frameworks. Cifuentes et al. [24] presented a comprehensive review of machine learning techniques for air temperature forecasting, highlighting that while isolated models like SVR and ANNs achieve satisfactory local accuracy, their performance fluctuates considerably when scaled across differing spatial and regional domains. Grover et al. [25] developed a deep hybrid model across 60 stations in the continental United States, utilizing Gaussian Process priors and deep belief networks to capture long-range spatiotemporal dependencies. Similarly, large-scale multi-city benchmarking studies have confirmed that deep learning models provide high representational power across diverse regimes. For instance, Gong et al. [26] applied CNN-LSTM across 321 global cities, Coutinho et al. [27] evaluated hybrid decomposition-based models across three climatic zones in Rio de Janeiro, Zaka et al. [28] compared LSTM and CNN models across four distinct climatic regions in Pakistan, and Uluocak and Bilgili [29] tested

LSTM-CNN and GRU-CNN models across Turkish provinces under Mediterranean and Continental climates. More recently, Kalimi et al. [30] employed a weighted ensemble of GRU, LSTM, and XGBoost for multi-horizon temperature forecasting at a tropical highland station in Indonesia, demonstrating that ensemble approaches can improve aggregate performance across multiple forecast horizons.

1.2. Limitations of Existing Multi-City Frameworks and Literature Synthesis

Despite these methodological advancements, existing multi-station and regional forecasting literature exhibits three critical limitations. First, most multi-regional frameworks apply a single global model or a fixed ensemble architecture across all stations, failing to accommodate sharp microclimatic contrasts, varying thermal inertia, and local topography, frequently leading to localized degradation of forecast skill [28, 29]. Second, while hybrid and stacking architectures generally outperform isolated base learners, standard ensembling schemes aggregate all candidate predictors indiscriminately without validation-based filtering, allowing divergent, noisy, or underperforming base models to contaminate the meta-fusion space [30, 31]. Third, existing ensemble approaches predominantly rely on static linear weighting, standard stacking regressors, or greedy selection, lacking the structural parsimony and soft adaptability required to dynamically balance multi-model consensus across complex geographical terrains without introducing excessive trainable parameters [32].

Synthesizing the current literature underscores an evident necessity for a scalable, nationwide forecasting framework that must couple station-adaptive noise filtering to discard erratic base predictors per region, with parsimonious soft fuzzy meta-aggregation to dynamically leverage complementary strengths across diverse climatic regimes, without enforcing rigid global assumptions.

Building upon these limitations, the efficacy of standard MFF architectures is often compromised by the indiscriminate inclusion of poorly performing or noise-inducing models. Including inaccurate predictors in the aggregation process introduces “forecast noise”, which disrupts the fuzzy clustering space and potentially biases the final ensemble output. To address this limitation, this study introduces a “Dynamic Pruning” (DP) mechanism integrated within the MFF framework. Before the fuzzy aggregation process, the proposed DP approach filters out underperforming base models based on their validation phase performance, ensuring that only the most reliable predictors contribute to the meta-learning process. By evaluating this Pruned MFF (PMFF) approach across a nationwide dataset covering all 81 cities of Türkiye, this study demonstrates that filtering out low-quality predictors significantly enhances predictive stability, offering a more resilient forecasting architecture for national meteorological applications.

The present study offers several key contributions to the field of temperature forecasting:

- Diverse Model Integration: The framework unites 10 heterogeneous architectures, including statistical (SARIMA, Exponential Smoothing (ETS), Seasonal Naïve(SN)), machine learning (XGBoost, RF, Ridge Regression, Support Vector Regression (SVR)), and deep learning (1D-CNN, LSTM, TCN) models, into a unified meta-learning pipeline.
- Dynamic Pruning (DP) Mechanism: Unlike conventional ensembles that indiscriminately include all base models, the DP mechanism identifies and excludes “noise-inducing” predictors by setting a validation-based performance threshold ($RMSE_{validation} < RMSE_{best} \times \theta$), thereby preserving the integrity of the fuzzy clustering space.
- Regional Adaptability and Multi-City Validation: By leveraging FCM-based weight learning, the framework dynamically adjusts the influence of different models based on local climatic patterns, effectively solving the “monolithic model” failure common in geologically and climatically heterogeneous countries like Türkiye. Unlike previous studies that often validate their models on a single city or a limited number of locations [28, 29], the proposed PMFF framework is rigorously evaluated across all 81 cities of Türkiye, providing a comprehensive assessment of its generalizability and robustness across diverse climatic zones ranging from coastal Mediterranean to continental and high-altitude regions.

The remainder of this study is organized as follows. Section 2 details the methodology, encompassing the base model pool, the DP mechanism, and the MFF aggregation process. Section 3 outlines the experimental design, including data acquisition, preprocessing, and the validation strategy. Section 3.4 presents the experimental results, providing a comparative analysis across Türkiye’s 81 cities supported by statistical validation. Finally, Section 4 summarizes the findings, discusses the implications of the PMFF architecture, and outlines potential avenues for future research.

II. METHODOLOGY

The forecasting framework proposed in this study integrates statistical, machine learning, and deep learning models into a unified meta-learning architecture. The overall pipeline consists of four main stages: data preprocessing, base model generation, dynamic pruning, and MFF aggregation.

2.1. Forecasting Model Pool

To ensure a robust meta-analysis, a diverse set of 10 base forecasting models, categorized into statistical,

machine learning, and deep learning domains, is employed:

- Statistical Models: SN [33], ETS [34], and SARIMA [35, 36].
- Machine Learning Models: XGBoost [37], RF [38], Ridge Regression [39], and SVR [40, 41].
- Deep Learning Models: 1D-CNN [42, 43], LSTM [44], and TCN [45].

These models were selected because, as noted by Tran et al. [17] and Nketiah et al. [18], different architectures offer complementary strengths in capturing non-linear and time-dependent features. For instance, while SARIMA excels in linear seasonality, models like LSTM and TCN are better suited for detecting long-term temporal dependencies [15, 20].

2.1.1. Architectural synergy and complementarity of the forecasting pool

The 10 base forecasting models included in the PMFF pool were deliberately selected across three fundamentally distinct computational paradigms to ensure high structural diversity and complementary feature representation:

- Statistical Baselines (SN, ETS, SARIMA): Provide strong, explicit inductive bias for rigid annual periodicities ($s = 52$) and linear mean drifts without overfitting high-frequency fluctuations. They anchor the ensemble during regular seasonal cycles where deep architectures may over-adapt to localized micro-variations.
- Non-Parametric and Regularized Machine Learning (Ridge, SVR, RF, XGBoost): Bridge linear dynamics and complex non-linear meteorological interactions. Ridge handles high multicollinearity among lagged features via L_2 shrinkage; SVR constructs robust non-linear boundaries in high-dimensional kernel spaces; and tree ensembles (RF, XGBoost) capture abrupt, non-monotonic threshold effects and reduce forecast variance via bagging and gradient boosting without assuming distributional normality.
- Deep Neural Networks (1D-CNN, LSTM, TCN): Capture hierarchical, multi-scale temporal dependencies that conventional models fail to extract. 1D-CNN extracts local temporal feature motifs via convolutional filters; LSTM tracks long-term recurring memory via gated recurrent units; and TCN models multi-scale receptive fields via dilated causal convolutions without suffering from vanishing gradients.

Complementary Fusion via PMFF: No single paradigm universally dominates all meteorological regimes. For instance, statistical models fail during sudden climatic anomalies, whereas deep architectures can exhibit high variance or phase lag in stationary periods. The PMFF framework harmonizes these trade-offs through a two-stage synergy (as illustrated in Figure 1):

1. Dynamic Pruning: Eliminates models whose structural assumptions fail in a given local climate (e.g., pruning rigid linear models in topographically complex zones) before they can introduce noise.
2. FCM-Based Meta-Aggregation: Dynamically discovers underlying consensus clusters among surviving models and computes adaptive fuzzy membership weights.

This allows the framework to dynamically shift reliance toward deep learning during complex non-linear transitions and toward regularized ML or statistical models during stable periodic cycles, achieving a synergistic ensemble superior to any isolated architecture. The mathematical formulations of these two stages are formalized in Section 2.2 and Section 2.3.

Seasonal Naïve. The SN method is one of the simplest benchmark forecasting approaches for seasonal time series. It assumes that future observations will repeat the values observed in the corresponding season of the previous cycle. Despite its simplicity, SN often provides a strong baseline and is widely used for evaluating the performance of more advanced forecasting models, particularly in seasonal datasets [33].

Exponential Smoothing (Holt–Winters). ETS methods generate forecasts by assigning exponentially decreasing weights to past observations, giving greater importance to more recent data. The Holt–Winters extension incorporates both trend and seasonal components, making it suitable for time series exhibiting recurring seasonal patterns. Due to its simplicity and effectiveness, Holt–Winters remains one of the most widely applied statistical forecasting techniques [34].

Seasonal Autoregressive Integrated Moving Average (SARIMA). SARIMA extends the classical ARIMA framework by incorporating seasonal autoregressive, differencing, and moving-average components. The model captures both short-term temporal dependencies and recurring seasonal structures within a time series. Owing to its flexibility and strong theoretical foundation, SARIMA has become a standard approach for forecasting seasonal data across numerous application domains [35, 36].

Extreme Gradient Boosting (XGBoost). XGBoost is an ensemble learning algorithm based on gradient boosting decision trees. The method sequentially constructs weak learners, where each new tree is trained to reduce the prediction errors of the existing ensemble through gradient-based optimization. XGBoost incorporates regularization mechanisms and efficient computational strategies, making it highly effective for modeling complex nonlinear relationships in structured data [37].

Random Forest (RF). RF is an ensemble machine learning method that combines predictions from multiple decision trees built on randomly selected subsets of observations and features. By aggregating the outputs of individual trees, the model reduces prediction variance and improves generalization performance. RF is particularly valued for its robustness against overfitting and its ability to capture nonlinear interactions among variables [38].

Ridge Regression. Ridge Regression is a regularized extension of ordinary least squares regression that introduces an L_2-norm penalty on model coefficients. This regularization reduces coefficient variability and mitigates multicollinearity problems, resulting in improved prediction stability when explanatory variables exhibit strong correlations [39].

Support Vector Regression (SVR). SVR extends the principles of Support Vector Machines to regression problems. By employing kernel functions, SVR maps input variables into higher-dimensional feature spaces and identifies a function that minimizes prediction error while maintaining model complexity. This framework enables effective modeling of nonlinear relationships and has been successfully applied to various forecasting tasks [40, 41].

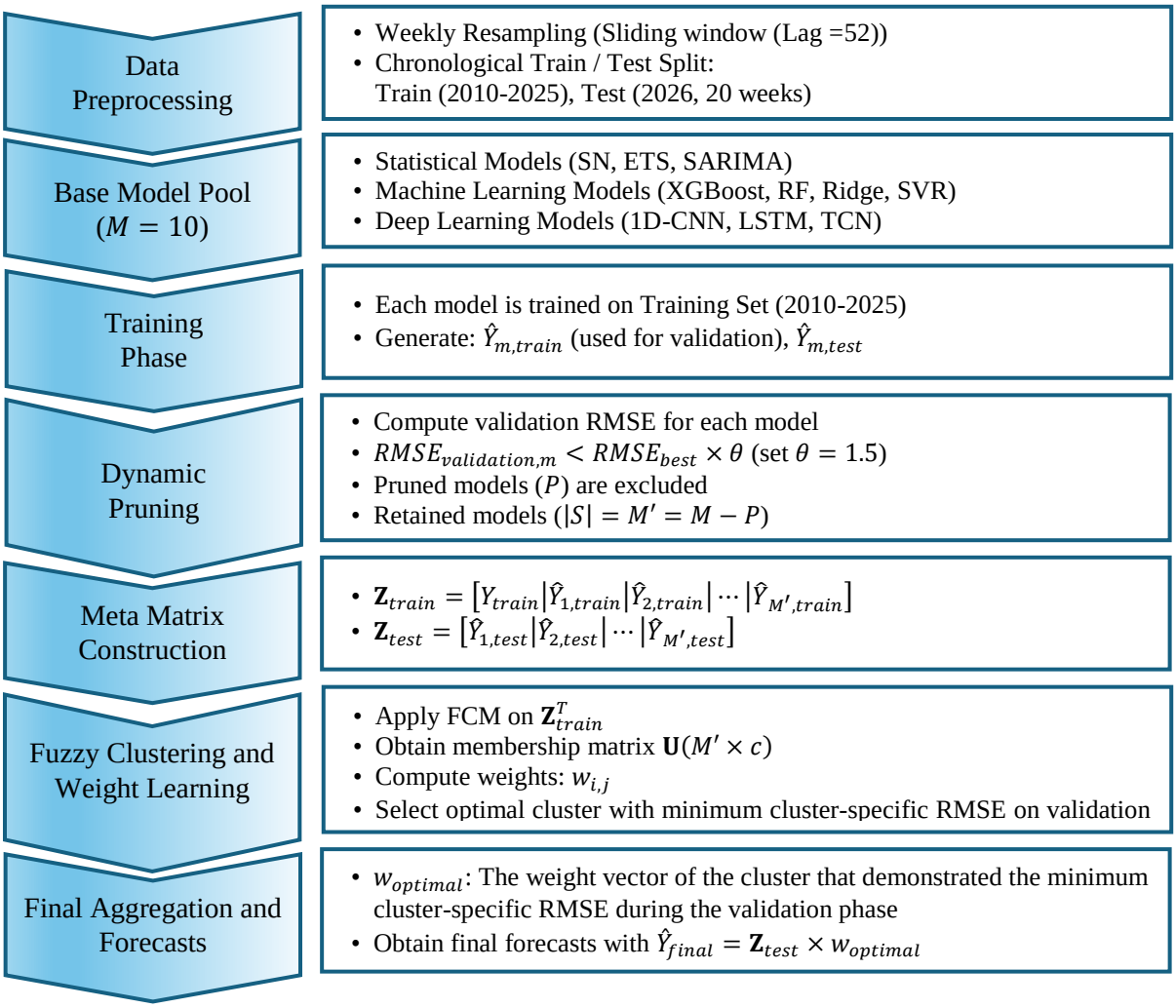


Figure 1. Systematic workflow and architectural pipeline of the proposed Pruned Meta-Fuzzy Functions (PMFF) forecasting framework. *The process begins with data preprocessing and partitioning into training and test sets. Subsequently, ten diverse base models are trained, and their forecasts are generated. A dynamic pruning mechanism then filters out underperforming models based on a validation-based RMSE threshold. The retained model outputs are used to construct meta-matrices, followed by FCM-based clustering and weight learning. Finally, the learned weights are applied to the test meta-matrix to produce the optimized final forecast.*

Algorithm 1: Pruned Meta Fuzzy Functions (PMFF)

Input: Historical temperature series T, Training length n_{train}, Test length n_{test}

Output: Final forecast vector $\hat{Y}_{final}$

Step 1: Base Forecasting
For each base model $m \in \{1, \dots, M\}$:
    Train model m on $T[0: n_{train}]$
    Generate forecast vectors $\hat{Y}_{m,train}$ and $\hat{Y}_{m,test}$
Step 2: Dynamic Pruning
    Compute validation RMSE for all base models on the training period
    Identify $RMSE_{best} = \min(RMSE_{validation})$
    Define the set of pruned models $S = \{m | RMSE_{validation,m} < RMSE_{best} \times \theta\}$
    Update $M' = |S|$ (number of retained models)
Step 3: Meta-Matrix Construction
    Construct the Training meta-matrix $\mathbf{Z}_{train} \in \mathbb{R}^{n_{train} \times (M'+1)}$:
        First column: Actual values Y_{train}
        Subsequent columns: Forecast vectors $\hat{Y}_{j,train}$ for all models $j \in S$
    Construct the Test meta-matrix $\mathbf{Z}_{test} \in \mathbb{R}^{n_{test} \times M'}$:
        Columns: Forecast vectors $\hat{Y}_{j,test}$ for all models $j \in S$
Step 4: Fuzzy Clustering and Weight Learning (Training Phase)
    Normalize $\mathbf{Z}_{train}$ using Z-score standardization to ensure feature uniformity
    Apply FCM on the transposed matrix $\mathbf{Z}_{train}^T$ to obtain the membership matrix $\mathbf{U} \in \mathbb{R}^{M' \times c}$, where c is the number of clusters
    Weight Calculation: Compute the weight vector $w \in \mathbb{R}^{M' \times c}$ for each cluster using the normalized membership values:
        $w_{ij} = \frac{\mu_{ij}}{\sum_{k=1}^{M'} \mu_{kj}}$ (where μ_{ij} represents the membership of model j to cluster i
Step 5: Final Aggregation
Apply the learned weight matrix w to the Test meta-matrix Z_{test}:
    $\hat{Y}_{final} = \mathbf{Z}_{test} \cdot w$
    (Note: The final prediction $\hat{Y}_{final}$ provides c different forecasts (one per cluster), from which the forecast corresponding to the minimum cluster-specific RMSE is selected as the optimal output)
Return final forecast vector $\hat{Y}_{final}$

One-Dimensional Convolutional Neural Network (1D-CNN). 1D-CNN apply convolutional filters directly to sequential data to automatically extract informative local patterns. Through hierarchical feature learning, 1D-CNNs can effectively identify short-term temporal structures and reduce the need for manual feature engineering. Their ability to learn representative features has led to widespread adoption in time-series forecasting applications [42, 43].

Long Short-Term Memory (LSTM). LSTM networks are a specialized form of recurrent neural networks designed to address the vanishing gradient problem encountered in standard Recurrent Neural Networks (RNNs). By employing memory cells and gating mechanisms, LSTMs selectively retain and update information over extended time horizons, enabling effective modeling of long-term temporal dependencies in sequential data [44].

Temporal Convolutional Network (TCN). TCNs are convolution-based architectures specifically developed for sequence modeling. They utilize causal convolutions and dilation mechanisms to capture long-range temporal dependencies while preserving temporal ordering. TCNs have demonstrated competitive or superior performance compared with recurrent architectures in many sequence prediction tasks due to their parallel computation capability and stable training behavior [45].

2.2. Dynamic Pruning (DP) Mechanism

A significant contribution of this study is the integration of a DP mechanism. In conventional ensemble methods, including poorly performing models can introduce "forecast noise", which degrades the meta-learner's performance [23]. To mitigate this, we define a performance threshold based on the validation phase. Let $RMSE_{best}$ be the minimum Root

Mean Square Error (RMSE) achieved by any base model in the validation set. A base model m is pruned if its validation error exceeds $RMSE_{best} \times \theta$, where θ is a sensitivity coefficient (set to 1.5 in this study). This ensures that only models with sufficient predictive merit contribute to the MFF aggregation.

2.3. Meta-Fuzzy Functions (MFF) Fusion

The pruned model outputs are aggregated using the MFF approach introduced by Tak [23]. Unlike standard ensemble averaging, MFF utilizes the FCM clustering algorithm to weight the contributions of base models based on their performance and similarity. The degree of membership μ_{ij} for model j in cluster i acts as a weight, allowing the MFF to dynamically select or combine models that are most relevant to the current temperature pattern [23].

The MFF framework is preferred over conventional stacking ensembles and simple weighted averaging for several fundamental reasons. Stacking regressors employ a fixed meta-learner that learns static weights from the entire training set, resulting in a single global combination rule that remains constant across all prediction instances [31]. Simple weighted averaging assigns predetermined or error-based weights that remain static once computed, offering no adaptability to changing climatic patterns [32]. In contrast, MFF leverages FCM clustering to partition the training data into c clusters based on the similarity of base model behaviors, allowing the system to dynamically emphasize the most accurate models for each specific meteorological regime. This soft clustering approach captures the inherent non-linearity and non-stationarity of meteorological time series, generalizing better across diverse climatic zones by adapting to local patterns rather than enforcing a single global rule [23].

2.4. Proposed Algorithm: Pruned Meta Fuzzy Functions (PMFF)

The pseudocode for the proposed framework is presented in Algorithm 1, while Figure 1 provides a complementary visual flowchart of the entire pipeline. The framework operates through a systematic integration of performance-based model selection and fuzzy-based aggregation, functioning through the following five primary steps:

Step 1. Base Forecasting

The process begins with the ingestion of the historical temperature series T. The dataset is partitioned into a training period (n_{train}) and a test period (n_{test}). A pool of M diverse forecasting models (comprising statistical, machine learning, and deep learning architectures) is trained on the interval $T[0{:}\, n_{train}]$. Upon completion of the training phase, each model m generates a training forecast vector $\hat{Y}_{m,train}$ (used for validation) and a prospective test forecast vector $\hat{Y}_{m,test}$.

Step 2: Dynamic Pruning (DP)

To mitigate the impact of "forecast noise", which occurs when poorly performing models are indiscriminately included in the meta-aggregation, a dynamic pruning threshold is implemented. Let $RMSE_{best}$ denote the minimum RMSE achieved by any base model in the validation set. A model m is retained if and only if it satisfies the following condition:

$$RMSE_{validation,m} < RMSE_{best} \times \theta \quad \textbf{(1)}$$

where θ represents the sensitivity coefficient controlling the tolerance level of the pruning mechanism. In this study, θ is set to 1.5, meaning that a base model is retained when its validation RMSE is no more than 50% higher than the minimum validation RMSE. The sensitivity coefficient θ governs the fundamental trade-off between ensemble diversity and forecast-noise filtering. Setting θ too close to 1.0 makes the threshold overly aggressive, prematurely filtering out viable models whose unique forecasting dynamics could otherwise enhance the meta-learning space. In contrast, a substantially larger value would render the pruning mechanism overly permissive, allowing severely underperforming predictors to enter the meta-matrix and disrupt the FCM clustering space with forecast noise.

Accordingly, $\theta = 1.5$ was adopted as an effective operational tolerance level that retains moderately competitive predictors while systematically purging models with severe validation errors. This setting is particularly well-suited for the heterogeneous model pool employed in this study, which integrates statistical, machine learning, and deep learning architectures with distinct behavioral characteristics. The primary objective of this threshold is not to find a single best value, but to find a good balance between model diversity and noise reduction.

Models failing the criterion in Equation 1 are excluded from the meta-learning phase. The set of retained models is denoted as S, and the number of surviving models is defined as $|S| = M'$.

In this step, RMSE was adopted as the pruning criterion due to its distinct advantages in meteorological modeling. First, by squaring the residuals, RMSE places a heavier penalty on large forecast errors, effectively preventing models with erratic variance or extreme peak deviations from contaminating the meta-learning stage. Second, unlike scale-independent percentage metrics (e.g., MAPE) which suffer from numerical instability when temperatures approach 0°C, RMSE provides continuous and robust evaluation across all thermal regimes. Third, RMSE preserves the original physical unit (˚C), offering direct interpretability while maintaining consistency with the MSE objective functions used during base model training. This choice ensures reliable exclusion of catastrophic failures and preserves the integrity of the FCM clustering space, while MAE, MAPE, and

SMAPE were reserved for final performance evaluation (Section 3.4).

Step 3: Meta-Matrix Construction
To ensure a robust meta-learning process and avoid data leakage, we construct two distinct matrices. The training meta-matrix, $\mathbf{Z}_{train} \in \mathbb{R}^{n_{train} \times (M'+1)}$, integrates the actual observations with the forecasts of the retained models:

$$\mathbf{Z}_{train} = \left[Y_{train} \middle| \hat{Y}_{1,train} \middle| \hat{Y}_{2,train} \middle| \cdots \middle| \hat{Y}_{M',train}\right] \quad \textbf{(2)}$$

where Y_{train} is the vector of actual observed values. Simultaneously, the test meta-matrix, $\mathbf{Z}_{test} \in \mathbb{R}^{n_{test} \times M'}$, is constructed using only the future forecasts generated by the retained models:

$$\mathbf{Z}_{test} = \left[\hat{Y}_{1,test} \middle| \hat{Y}_{2,test} \middle| \cdots \middle| \hat{Y}_{M',test}\right] \quad \textbf{(3)}$$

Step 4: Fuzzy Clustering and Weight Learning
FCM clustering algorithm is employed on the transposed training matrix $\mathbf{Z}_{train}^{T}$ to uncover latent relationships between model behaviors. This clustering process results in a membership matrix $\mathbf{U} \in \mathbb{R}^{M' \times c}$, where c represents the predefined number of clusters. The weight w_{ij} of model j within cluster i is derived by normalizing the membership values:

$$w_{ij} = \frac{\mu_{ij}}{\sum_{k=1}^{M'} \mu_{kj}} \quad \textbf{(4)}$$

where μ_{ij} denotes the degree of membership of model j to cluster i. This weighting mechanism allows the MFF framework to dynamically assign higher importance to models that perform optimally within specific temperature patterns.

Step 5: Final Aggregation
In the final inference phase, the learned weight matrix w is projected onto the prospective forecasts in the test meta-matrix. The final forecast $\hat{Y}_{final}$ is calculated as:

$$\hat{Y}_{final} = \mathbf{Z}_{test} \cdot w_{optimal} \quad \textbf{(5)}$$

where $w_{optimal}$ corresponds to the weight vector of the cluster that demonstrated the minimum cluster-specific RMSE during the validation phase. This final output represents the optimized, noise-reduced temperature forecast for the target period.

III. EVALUATION

3.1. Data Acquisition and Preprocessing

The dataset utilized in this study comprises daily average temperature records for 81 cities in Türkiye. Data were sourced from the Open-Meteo archive [46], which provides reliable reanalysis integration. The raw daily temperatures were resampled to weekly intervals to ensure seasonal consistency and reduce noise. The dataset is structured into a supervised learning format using a sliding window approach, where past observations are used to predict the subsequent temperature state. Table 1 provides a detailed statistical summary of minimum, maximum, and average temperatures for each city in Türkiye based on long-term weekly averaged data (2010–2026).

Table 1. Türkiye climate data: Temperature statistics by city (2010–2026)

City	Min. Temp. (°C)	Max. Temp. (°C)	Mean Temp. (°C)	City	Min. Temp. (°C)	Max. Temp. (°C)	Mean Temp. (°C)	City	Min. Temp. (°C)	Max. Temp. (°C)	Mean Temp. (°C)
Adana	3.84	32.3	19.2	Giresun	-0.37	26.37	14.6	Samsun	-1.2	27.53	14.51
Adıyaman	-2.26	36.27	16.99	Gümüşhane	-10.2	26.56	9.58	Siirt	-6.99	33.76	14.82
Afyonkarahisar	-9.43	29.17	11.7	Hakkari	-12.13	27.79	9.98	Sinop	-0.26	28.17	14.81
Ağrı	-15.09	26.93	8.11	Hatay	3.59	32.13	19.05	Sivas	-11.66	26.87	9.65
Amasya	-5.47	29.11	14.33	Isparta	-6.56	28.84	12.63	Tekirdağ	-1.26	30.13	15.51
Ankara	-6.89	30.96	12.94	Mersin	5.41	32.06	19.08	Tokat	-9.47	28.11	12.61
Antalya	4.56	34.36	18.79	İstanbul	1.09	29.31	15.28	Trabzon	0.21	27.14	14.82
Artvin	-8.09	27.37	11.68	İzmir	2.3	33.5	18.33	Tunceli	-7.57	32.34	13.42
Aydin	2.33	34.44	18.72	Kars	-13.6	23.46	6.56	Şanlıurfa	-1.96	36.93	18.69
Balıkesir	-0.54	30.33	15.29	Kastamonu	-7.21	27.77	10.95	Uşak	-7.09	29.19	13
Bilecik	-5.37	27.17	12.61	Kayseri	-8.7	30.67	12.48	Van	-10.01	27.17	9.83
Bingöl	-9.8	30.8	12.17	Kırklareli	-5.46	29.24	13.81	Yozgat	-12.69	27.03	9.58
Bitlis	-12.24	28.03	10.23	Kırşehir	-9.27	30.14	12.63	Zonguldak	-0.29	27.86	15.15
Bolu	-5.89	25.83	11.27	Kocaeli	-3.34	25.64	12.5	Aksaray	-9.04	31.79	13.35
Burdur	-4.81	30.61	13.88	Konya	-9.1	31.63	13.1	Bayburt	-14.16	24.7	6.62
Bursa	-1.3	30.73	14.67	Kütahya	-8.99	27.71	10.99	Karaman	-9.29	30.79	13.2
Çanakkale	0.49	31.36	16.48	Malatya	-6.13	33.64	14.37	Kırıkkale	-6.67	31.76	13.75
Çankırı	-7.47	30.16	12.87	Manisa	1.13	33.24	17.37	Batman	-4.57	35.44	17.15
Corum	-8.47	27.93	12.07	Kahramanmaraş	-1.77	32.04	16.42	Şirnak	-7.91	32.23	14.01
Denizli	-3.37	33.71	17.37	Mardin	-5.13	33.8	15.73	Bartın	-0.96	27.34	14.04
Diyarbakır	-5.2	36.21	16.55	Muğla	-0.96	31.33	15.9	Ardahan	-13.2	21.73	5.49
Edirne	-5.84	31.17	14.87	Muş	-12.8	29.5	10.73	Iğdır	-11.84	31.89	13.14
Elâzığ	-6.14	32.51	13.77	Nevşehir	-10.03	28.63	11.44	Yalova	0.9	28.94	15.67
Erzincan	-11.23	29.21	10.57	Niğde	-7.87	30.49	12.17	Karabük	-3.7	30.19	14.14
Erzurum	-14.03	25.71	6.59	Ordu	-1.93	26.46	13.95	Kilis	-0.04	33.39	17.14
Eskişehir	-6.57	28.69	11.87	Rize	-0.49	26.17	13.89	Osmaniye	3.89	31.9	18.43
Gaziantep	-2.87	34.17	16.24	Sakarya	-2.14	27.37	14.48	Düzce	-3.03	26.74	13.91

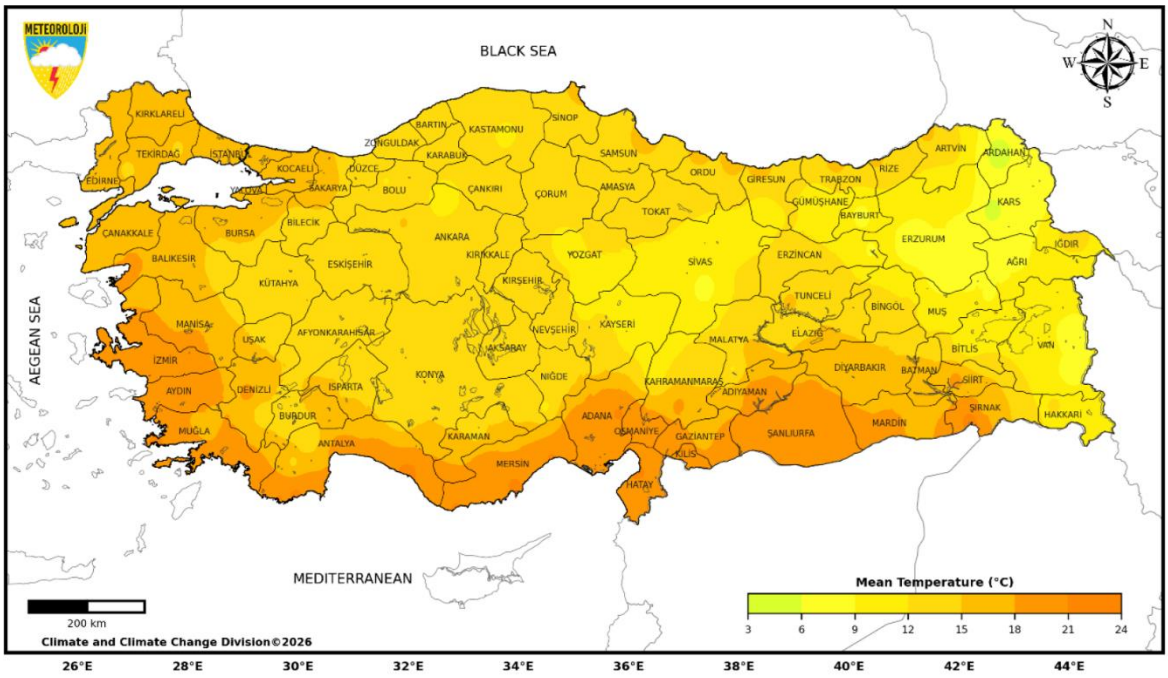


Figure 2. Türkiye mean temperature map (2025) [47]

Table 2. Hyperparameter configurations and optimization logic for the forecasting model pool

Category	Method	Configuration / Hyperparameters	Optimization Logic	Determination strategy	References / Basis
Statistical	Seasonal Naive	$s = 52$	Naive benchmark repeating the previous seasonal cycle	Annual seasonality of weekly data	Hyndman & Athanasopoulos [33]
	Exponential Smoothing	$Periodicity = 52$, $Trend = additive$, $Seasonality = additive$	Additive Holt-Winters formulation for fixed yearly seasonal swings	Weekly annual seasonality; model formulation based on [25]	Winters [34]
	SARIMA	Grid Search: $(p,d,q) \in \{(1,0,0),(0,0,1),(1,0,1)\}$, $(P,D,Q) \in \{(1,1,0),(0,1,1),(1,1,1)\}$	First-order seasonal differencing ($D = 1$) to remove yearly non-stationarity	Grid search over seasonal orders evaluated on AIC	Box et al. [36], Koçak [1]
ML	XGBoost	$n_estimators = 200$, $learning_rate = 0.1$	Minimizing residuals via gradient boosting	Fixed configuration based on benchmark practice	Chen & Guestrin [28], Apaydın et al. [10]
	Random Forest	$n_estimators = 100$, $max_depth = None$	Bagging ensemble averaging to reduce forecast variance without pruning individual trees	Fixed configuration	Breiman [29], Kumaş & Akyüz [11]
	Ridge	$\lambda = 1.0$ (Regularization strength)	Standard ridge penalty to shrink regression coefficients and mitigate high multicollinearity among lagged features	Standard regularization setting	Hoerl & Kennard [39]
	SVR	$Kernel = RBF$, $C = 1.0$, $\gamma = scale$	Non-linear mapping via Radial Basis Function; heuristic parameter scaling based on input variance	Standard configuration	Drucker et al. [41], Aghelpour et al. [5]
Deep Learning	1D-CNN	$Filters = 64$, $Kernel\ size = 3$	Local temporal feature extraction through 1D receptive fields	Empirically selected architecture	Kiranyaz et al. [43]
	LSTM	$Units = 50$, $Epochs = 10$, $Batch\ size = 16$	Hidden recurrent state capacity for long-term seasonal memory while preventing overfitting	Empirically selected	Hochreiter & Schmidhuber [44], Karabulut & Topçu [8]
	TCN	$DilatedConv(rate = 1,2,4)$, $Kernel = 3$	Exponentially expanding receptive field to cover full seasonal history without recurrent latency	Receptive-field design [45]	Bai et al. [45], Wu et al. [16]

Note: All deep learning models utilize the Adam optimizer with a learning rate of 0.001 and Mean Squared Error (MSE) loss function

Moreover, Figure 2 presents the mean temperature distribution map of Türkiye for 2025. Together, Figure 2 and Table 1 highlight the pronounced temperature heterogeneity in Türkiye, driven by topography, latitude, and proximity to seas. Such diversity confirms that a "one-size-fits-all" approach to forecasting is insufficient, as the factors driving temperature fluctuations in a coastal region significantly differ from those in high-altitude inland cities. The PMFF framework addresses this challenge by functioning as an adaptive ensemble. Instead of relying on a static model, the fuzzy clustering mechanism inherently accounts for these regional disparities, allowing the system to shift its reliance toward the most accurate predictors for any given micro-climate.

3.2. Training and Evaluation Pipeline

The evaluation of the models follows a rigorous time-series splitting strategy to prevent data leakage and ensure temporal consistency. The weekly averaged dataset comprises 854 observations per city, covering the period from 1 January 2010 to 24 May 2026. The training set is defined as the multi-decadal period from 1 January 2010 to 31 December 2025 (834 weeks), while the test set comprises the subsequent 20 weeks from 1 January 2026 to 24 May 2026.

This particular partitioning strategy is motivated by two primary operational considerations. First, the training cut-off at 31 December 2025 ensures that models learn long-term inter-annual trends from complete calendar years, while the 20-week test period covers the dynamic transition from winter to spring, a period characterized by high climatic variability. This design rigorously evaluates the framework under challenging, non-stationary conditions that are critical for real-world meteorological applications. Second, a 20-week forward-looking test horizon aligns directly with the medium-term forecasting windows utilized by national meteorological agencies. Concurrently, 24 May 2026 represents the most recent available observation boundary at the time of dataset acquisition, naturally forms the out-of-sample evaluation period's endpoint.

This strategy guarantees that all models are trained on historical observations up to the latest full annual cycle and validated on unseen forward-looking data, ensuring an unbiased assessment of predictive generalizability and robustness.

Hyperparameter selection utilizes specific configurations tailored to the nature of the temperature time series. For the SARIMA models, a grid search is performed across seasonal configurations to account for the annual 52-week cycle. The orders $(1,1,0,52)$, $(0,1,1,52)$, and $(1,1,1,52)$ are specifically tested. These parameters are selected because temperature data typically exhibit strong weekly seasonality (52 weeks/year), and differencing ($D = 1$) is necessary to stabilize the mean. The optimal configuration among these candidates is determined by minimizing the Akaike Information Criterion (AIC) and the validation set RMSE, ensuring that the SARIMA architecture effectively captures both short-term fluctuations and yearly periodicity [1, 36]. For the Exponential Smoothing model, the additive trend and seasonal components are adopted based on the observed linear trend and stable seasonal pattern in the weekly temperature series [33, 34].

Regarding machine learning models, hyperparameters are configured using commonly adopted default values from their original implementations, followed by light tuning on a validation subset to prevent overfitting and ensure generalization across diverse climatic regions [37–39]. Specifically, for XGBoost, the number of estimators ($n_estimators = 200$) and learning rate ($learning_rate = 0.1$) are selected after preliminary experiments showed that higher learning rates led to rapid overfitting while lower values increased convergence time without significant accuracy gains. For RF, the number of trees ($n_estimators = 100$) and unlimited maximum depth are chosen to balance model complexity with variance reduction through ensemble averaging [38]. For Ridge regression, the regularization strength is set to $\lambda = 1.0$ to mitigate multicollinearity among correlated meteorological features [39]. For SVR, the Radial Basis Function (RBF) kernel with $C = 1.0$ and $\gamma = scale$ is adopted to effectively model non-linear relationships in the temperature data [41].

For deep learning models, a sliding window (look-back) approach is implemented with a window size of 52 weeks to align with the seasonal nature of the data. Architecture complexity is empirically determined through iterative experimentation on the validation set: a hidden layer size of 50 neurons is utilized for LSTM and TCN-based networks to maintain a balance between model capacity and computational efficiency, preventing overfitting in complex terrains [8, 44, 45]. For the 1D-CNN, 64 filters with a kernel size of 3 are selected to extract local temporal features from the sequence [42, 43]. The TCN architecture employs dilated convolutions with dilation rates of 1, 2, and 4 to expand the receptive field and capture long-range temporal dependencies [45]. All deep learning models utilize the Adam optimizer with a fixed learning rate of 0.001 and MSE loss function, following standard practice in time-series forecasting literature, with early stopping applied based on validation loss to prevent overfitting. The number of epochs is set to 10 with a batch size of 16 to ensure stable gradient estimates without excessive computational cost. Importantly, all hyperparameter selection and tuning procedures are performed exclusively on the training and validation sets, and the test dataset remained strictly unseen throughout the entire model development process. The complete hyperparameter configurations and their determination strategies are summarized in Table 2.

3.3. Methodological Validity and Overfitting Mitigation

The validity and generalizability of the proposed PMFF are ensured through multiple architectural safeguards implemented across the model development pipeline:

- Strict Temporal Splitting (No Data Leakage): A chronological train-test partition is strictly enforced. All preprocessing scaling parameters, base model training, hyperparameter validation, dynamic pruning, and meta-aggregation are conducted strictly on historical data (2010–2025). The test period (1 January 2026 – 24 May 2026) remains completely unseen until final out-of-sample inference, eliminating look-ahead bias.
- Base Learner Regularization: Overfitting in individual predictors is controlled through specific structural constraints: Ridge regression incorporates L_2 regularization ($\lambda = 1.0$) against multicollinearity; SVR bounds margins via an RBF kernel; RF constrains variance through bootstrap aggregation; XGBoost applies shrinkage via learning_rate = 0.1; and deep learning models (1D-CNN, LSTM, TCN) utilize parsimonious architectures with early stopping.
- Noise Reduction via Dynamic Pruning: Rather than aggregating all models indiscriminately, PMFF evaluates validation RMSE and eliminates underperforming or unstable models satisfying: $RMSE_{validation,m} \geq RMSE_{best} \times \theta$. This prevents divergent base predictors from contaminating the fuzzy representation space with forecast noise.
- Parsimonious, Unsupervised Meta-Aggregation: In contrast to over-parameterized neural meta-learners that risk overfitting on limited validation windows, PMFF employs Fuzzy C-Means (FCM) clustering. FCM generates soft, bounded membership weights ($0 \leq w_{ij} \leq 1, \sum_j w_{ij} = 1$) based purely on geometric output proximity without introducing trainable gradient-descent parameters.

The empirical stability of PMFF is corroborated by consistent performance across 81 climatically and topographically diverse cities under multiple statistical tests. Nonetheless, because hyperparameter selection and pruning thresholds are derived from finite historical records, future work on longer horizons and higher resolutions will further evaluate the architecture's generalizability.

3.4. Experimental Results

The evaluation focuses on the forecasting accuracy of the proposed MFF framework compared to the ten benchmark methods across the 81 Turkish cities. The performance is assessed using a comprehensive set of error metrics (RMSE, Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (SMAPE)), providing a multi-dimensional perspective on model reliability.

To examine model behavior across geographically diverse settings, Table 3 presents the metric performance for Türkiye's largest cities, illustrating how the MFF framework adapts to local climatic patterns compared to individual base models.

Table 4 synthesizes the predictive rankings and dynamic framework parameters across Turkish cities. To provide a unified multi-metric evaluation, the model performance for each city is presented as a 4-tuple in parentheses: ($rank_{RMSE}$, $rank_{MAE}$, $rank_{MAPE}$, $rank_{SMAPE}$), representing the model's relative rank from best (1) to worst (11) for RMSE, MAE, MAPE, and SMAPE, respectively. For instance, a ranking tuple of (1, 1, 1, 7) indicates that PMFF achieved the 1st rank in RMSE, MAE, and MAPE, but ranked 7th in SMAPE for Diyarbakır.

Table 3. Comparative analysis of predictive performance metrics for the MFF framework and base forecasting models across the biggest Turkish cities

City	Metrics	PMFF	SN	ETS	SARIMA	XGBoost	SVR	RF	Ridge	CNN	LSTM	TCN
İstanbul	RMSE	1.894	16.130	3.401	2.053	2.430	2.305	1.901	1.999	2.150	2.066	**1.882**
	MAE	1.526	15.547	2.996	1.795	2.051	1.924	1.651	1.652	1.843	1.800	**1.498**
	MAPE	16.264	178.544	31.821	20.134	24.131	22.005	18.354	18.007	19.670	19.228	**15.566**
	SMAPE	16.359	86.169	26.229	18.516	21.576	20.072	17.615	17.830	20.049	18.447	**15.944**
Ankara	RMSE	2.687	19.977	3.956	3.264	**2.592**	4.021	2.794	2.884	2.843	2.644	2.985
	MAE	2.272	19.521	3.508	2.840	**2.104**	3.247	2.388	2.493	2.374	2.161	2.598
	MAPE	112.957	2281.005	154.000	138.550	**80.815**	186.886	129.290	170.217	109.270	107.360	84.757
	SMAPE	46.348	115.128	49.700	45.203	**36.132**	53.149	44.920	56.561	53.074	46.843	56.556
İzmir	RMSE	1.731	18.202	3.398	2.189	2.180	2.462	1.817	2.054	2.405	**1.713**	2.136
	MAE	1.414	17.692	3.103	1.803	1.817	2.031	1.553	1.675	2.038	**1.380**	1.857
	MAPE	11.912	147.213	23.463	14.948	14.155	16.841	12.753	13.817	16.489	**11.576**	14.308
	SMAPE	11.964	79.265	20.596	13.968	13.701	15.728	12.866	14.177	16.262	**11.759**	14.493

Furthermore, Table 4 details the adaptive behavior of the PMFF architecture through columns P (the number of underperforming base models excluded by the Dynamic Pruning threshold) and C (the optimal number of clusters determined by FCM meta-learning). For instance, in the case of Diyarbakır, $P = 1$ indicates that one underperforming base model was dynamically excluded from the pool, while $C = 4$ reflects that the framework identified four distinct fuzzy clusters to optimally aggregate the surviving predictors.

The findings presented in Table 4 reveal the superior convergence of the PMFF framework. A critical observation is the variability of model performance across different urban contexts. For instance, while LSTM secures the top rank in Izmir with a perfect (1,1,1,1) score, the PMFF framework remains highly competitive with a consistent (2,2,2,2) performance. Similarly, while TCN excels in Istanbul and XGBoost proves optimal for Ankara, the PMFF framework consistently adapts to these local dynamics. This heterogeneity in performance underscores the capability of individual models to capture specific patterns. When evaluating predictive power across all 81 cities, the MFF emerges as the most robust approach, achieving superior Mean Rank values of 2.39 (RMSE), 2.82 (MAE), 3.60 (MAPE), and 3.62 (SMAPE). The inclusion of P and C values quantifies the adaptability of the proposed framework.

Figure 3 highlights the comparative performance of the PMFF framework; while baseline methods exhibit significant deviation from the ground truth in seasonal peaks, the proposed approach remains robust across both the inland climate of Ankara and the coastal-influenced climate of Istanbul.

While the metrics provide a direct comparison of forecasting error, statistical validation is essential to rule out performance due to chance. Following the control-versus-all comparison strategy [48] commonly adopted in forecasting benchmark studies, the proposed method was designated as the control algorithm and evaluated exclusively against the 10 benchmark methods. To validate statistical significance, post-hoc analysis was executed using pairwise Wilcoxon signed-rank tests combined with the step-down Holm-Bonferroni correction.

To determine whether the observed performance disparities across the 81 Turkish cities were statistically significant rather than artifacts of random variation, a non-parametric Friedman test was initially conducted on the RMSE, MAE, MAPE, and SMAPE performance distributions across all competing models. The test strongly rejected the omnibus null hypothesis that all forecasting methods perform equivalently, yielding highly significant results across all four evaluation metrics: RMSE: $\chi^2(10) = 483.74$, MAE: $\chi^2(10) = 476.00$, MAPE: $\chi^2(10) = 351.88$, SMAPE: $\chi^2(10) = 296.16$ ($p < 0.001$ for all).

Table 4. Comparative performance ranking of forecasting models across Turkish cities

City	PMFF	SN	ETS	SARIMA	XGBoost	SVR	RF	Ridge	CNN	LSTM	TCN	PC
Ankara	(3,3,5,4)	(11,11,11,11)	(9,10,8,6)	(8,8,7,3)	**(1,1,1,1)**	(10,9,10,8)	(4,5,6,2)	(6,6,9,10)	(5,4,4,7)	(2,2,3,5)	(7,7,2,9)	1 2
Diyarbakır	**(1,1,1,7)**	(11,11,11,11)	(10,4,8,10)	(7,9,5,8)	(6,7,9,3)	**(2,2,3,1)**	(4,5,7,6)	(8,6,4,9)	(5,8,6,5)	(3,3,2,2)	(9,10,10,4)	1 4
Hatay	(2.5,2.5,2.5,2.5)	(11,11,11,11)	(5,6,6,6)	(9,10,7,9)	(10,8,10,10)	(7,5,5,4)	(4,7,8,7)	**(1,1,1,1)**	(6,4,4,5)	(2.5,2.5,2.5,2.5)	(8,9,9,8)	3 4
Mersin	**(1,1,1,2)**	(11,11,11,11)	(10,10,9,8)	(7,7,7,7)	(9,8,10,10)	(8,9,8,9)	(4,4,6,4)	(5,5,5,6)	(6,6,4,5)	(2,2,3,3)	(3,3,2,1)	1 5
İstanbul	(2,2,2,2)	(11,11,11,11)	(10,10,10,10)	(5,5,7,6)	(9,9,9,9)	(8,8,8,8)	(3,3,4,3)	(4,4,3,4)	(7,7,6,7)	(6,6,5,5)	**(1,1,1,1)**	1 5
İzmir	(2,2,2,2)	(11,11,11,11)	(10,10,10,10)	(7,5,7,5)	(6,6,5,4)	(9,8,9,8)	(3,3,3,3)	(4,4,4,6)	(8,9,8,9)	**(1,1,1,1)**	(5,7,6,7)	1 5
Kahramanmaraş	(2,5,5,6)	(11,11,11,11)	**(1,3,3,1)**	(9,10,2,3)	(4,2,4,9)	(8,9,6,10)	(3,4,7,7)	(5,1,1,4)	(7,7,10,8)	(10,8,8,5)	(6,6,9,2)	1 3
Kastamonu	**(1,2,2,3)**	(11,11,11,11)	(10,10,9,7)	(9,7,5,5)	(7,8,10,9)	(8,9,8,10)	(5,5,7,6)	(2,4,4,4)	**(3,1,1,1)**	(4,3,3,2)	(6,6,6,8)	2 5
Rize	(2,2,2,2)	(11,11,11,11)	(10,10,10,10)	(6,8,8,7)	**(1,1,1,1)**	(8,6,5,5)	(3,3,3,3)	(7,7,7,8)	(9,9,9,9)	(5,5,4,4)	(4,4,6,6)	2 4
Tekirdağ	(2,2,2,2)	(11,11,11,11)	(10,10,10,10)	(9,9,9,8)	(5,8,8,9)	(8,7,7,5)	**(1,1,1,1)**	(7,5,5,6)	(6,6,6,7)	(3,4,3,3)	(4,3,4,4)	2 3
RMSE-MR	***2.39***	*11*	*8.25*	*8.25*	*5.48*	*7.53*	*3.01*	*5.06*	*6.15*	*4.12*	*4.78*	
MAE-MR	***2.82***	*11*	*8.38*	*8.67*	*5.12*	*6.95*	*3.10*	*4.85*	*6.17*	*3.92*	*5.01*	
MAPE-MR	***3.60***	*11*	*7.77*	*7.77*	*6.34*	*6.27*	*4.83*	*4.57*	*5.58*	*3.97*	*4.31*	
SMAPE-MR	***3.62***	*11*	*7.42*	*6.58*	*5.48*	*5.73*	*3.81*	*6.20*	*6.09*	*4.76*	*5.32*	

Note: PMFF: Pruned Meta Fuzzy Functions (Proposed method); SN: Seasonal Naive; ETS: Exponential Smoothing; SARIMA: Seasonal Autoregressive Integrated Moving Average; RF: Random Forest; SVR: Support Vector Regression; CNN: One-Dimensional Convolutional Neural Network; LSTM: Long Short-Term Memory; TCN: Temporal Convolutional Network. Values in parentheses represent the relative rank of each model for the given city ordered across the four evaluation metrics: ($rank_{RMSE}$, $rank_{MAE}$, $rank_{MAPE}$, $rank_{SMAPE}$), where 1 denotes the best-performing model. P represents the number of pruned models; C represents the optimal cluster count selected by the PMFF framework. RMSE-MR, MAE-MR, MAPE-MR, and SMAPE-MR denote the Mean Rank of each model, calculated across all 81 cities for the corresponding metric; values in bold represent the best performing results for that city or metric.

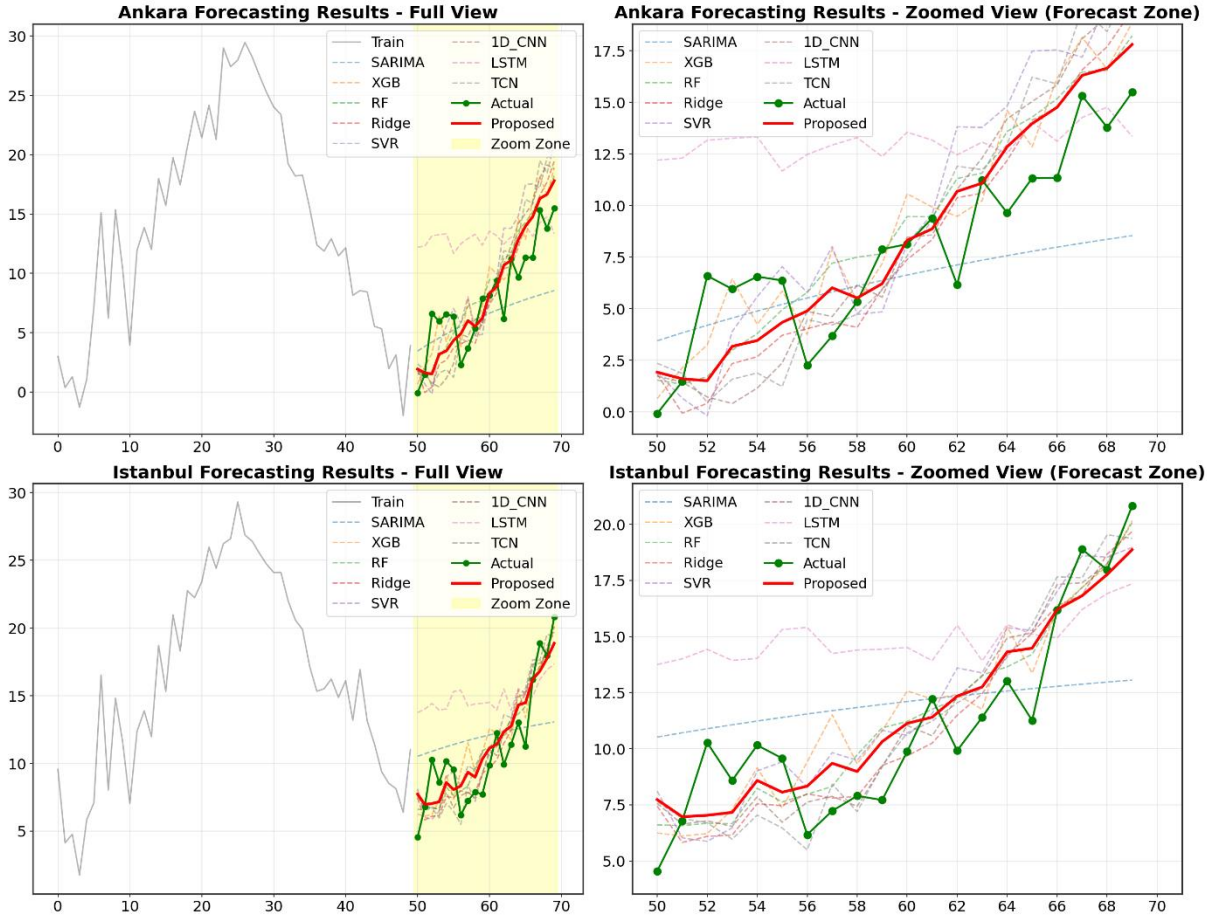


Figure 3. Comparative forecasting performance of the proposed PMFF framework against baseline models for Ankara and Istanbul

Following the control-versus-all benchmarking strategy [48], the proposed PMFF framework was designated as the control algorithm and evaluated against the 10 standalone benchmark models. To identify pairwise statistical differences while controlling the family-wise error rate, post-hoc analysis was performed using pairwise Wilcoxon signed-rank tests combined with the step-down Holm–Bonferroni correction procedure. Figure 4 illustrates the comparative average rankings across all 81 cities and highlights the post-hoc significance thresholds for each metric.

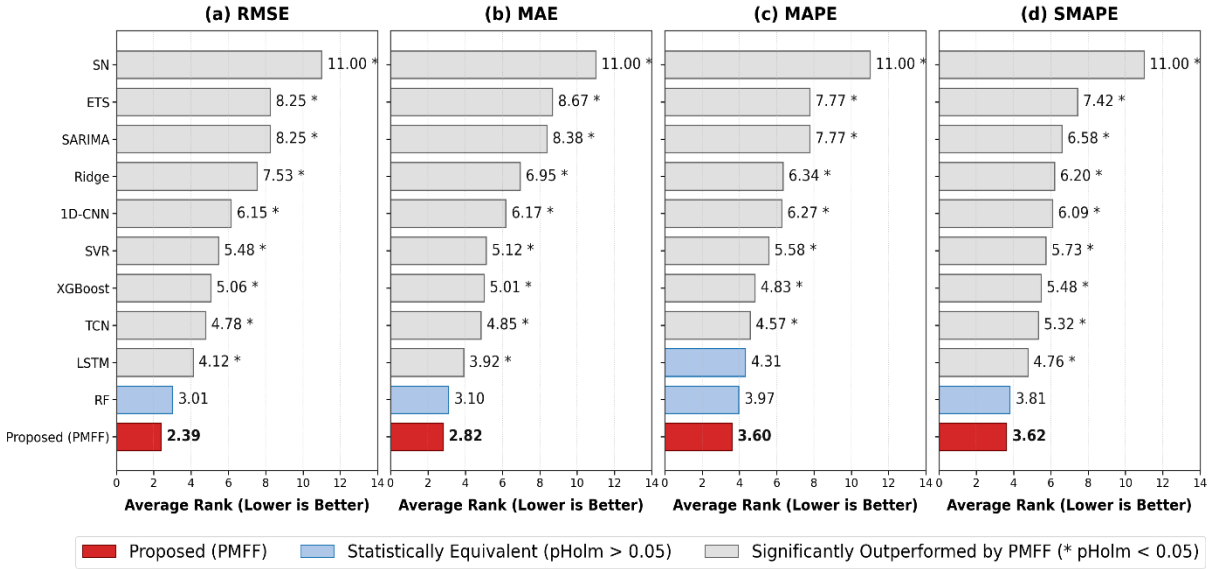


Figure 4. Average ranks and statistical significance of the forecasting methods across the 81 cities for RMSE, MAE, MAPE, and SMAPE

As shown in Figure 4, the proposed PMFF architecture consistently achieves the superior average rank across all four metrics: 2.39 for RMSE, 2.82 for MAE, 3.60 for MAPE, and 3.62 for SMAPE

Across the scale-dependent error metrics (RMSE and MAE), the post-hoc analysis confirms the statistical superiority of the proposed method ($p_{Holm} < 0.05$) against nearly all competing baselines. Specifically, PMFF establishes a profound statistical advantage over classical statistical baselines (SN, SARIMA, and ETS) as well as machine learning and deep learning architectures (XGBoost, SVR, 1D-CNN, TCN, and LSTM), with all comparisons demonstrating extreme statistical significance (all $p_{Holm} < 0.0001$). Notably, for RMSE, MAE, and SMAPE, the RF model did not show a statistically significant difference compared to our proposed method ($p_{Holm} > 0.05$ for all three metrics, specifically $p_{Holm} = 0.051$, 0.466, and 0.708 respectively), indicating competitive performance parity in absolute error minimization while PMFF still maintained the top overall average rank.

In the percentage error domains, PMFF maintained its statistical dominance against the majority of the models. In the MAPE metric, PMFF significantly outperformed 8 out of the 10 models (including RF with $p_{Holm} = 0.003$ and Ridge with $p_{Holm} = 0.031$), while both TCN ($p_{Holm} = 0.185$) and LSTM ($p_{Holm} = 0.390$) failed to reject the null hypothesis, demonstrating comparable performance in percentage scaling. In SMAPE, PMFF demonstrated statistically significant superiority against 9 out of the 10 benchmark models (including TCN with $p_{Holm} = 0.0003$ and XGBoost with $p_{Holm} = 0.0013$), with RF being the only model showing parity.

Crucially, across all 81 cities and all evaluation metrics, no competing baseline achieved a statistically superior performance compared to PMFF. These comprehensive statistical findings confirm that the predictive gains of the PMFF framework are robust, generalizable across heterogeneous geographical regions, and capable of effectively filtering out forecast noise.

3.5. Discussion on Regional Adaptability

A key challenge in nationwide temperature forecasting is the pronounced climatic heterogeneity across Türkiye, ranging from coastal Mediterranean and Aegean regions to continental interiors and high-altitude eastern provinces (as shown in Figure 2 and Table 1). A monolithic model that performs well in one region may fail in another due to fundamentally different temperature dynamics [20, 22]. To accommodate these regional disparities, the PMFF framework does not enforce a rigid, globally fixed ensemble structure. Instead, the Dynamic Pruning and FCM-based weight learning processes are executed station-wise based on local validation dynamics. This enables the base model pool and fuzzy fusion weights to dynamically adapt to the unique thermal inertia, microclimatic seasonality, and volatility of each specific province without requiring manual recalibration. For instance, the framework may assign higher weights to models that excel in capturing maritime influences for a coastal city like Antalya, while shifting its reliance toward models better suited for high-altitude, arid conditions for a continental city like Erzurum. This adaptability is validated through the nationwide experimental setup: the same PMFF pipeline is applied uniformly to all 81 cities without any region-specific manual tuning or model reselection. The consistent superior performance across diverse climatic zones, as evidenced by the low mean rank values in Table 4, confirms that the framework autonomously adapts to regional characteristics. This validation strategy ensures that the PMFF architecture is not overfitted to any particular region but rather possesses the flexibility to accommodate the full spectrum of climatic conditions observed across Türkiye.

IV. CONCLUSION

The empirical and statistical results of this study confirm that the PMFF framework serves as a robust and superior alternative to standalone forecasting models for national meteorological monitoring. By systematically integrating a diverse model pool with a performance driven pruning strategy, the framework successfully addresses the dual challenges of forecast noise and regional climate variability. When the data for each city is considered individually, none of the 10 baseline methods consistently emerges as the top performer. However, the proposed framework achieves the best average rank, as it utilizes an ensemble mechanism capable of self-adaptation to the provided datasets.

The statistical rigor applied through the Friedman and Wilcoxon post-hoc tests demonstrates that the proposed architecture achieves predictive gains that are not due to random chance, but rather reflect a consistent ability to capture non-linear, multi-scale temporal dependencies. Furthermore, the adaptability of the FCM based aggregation allows the system to autonomously navigate the complex climatic landscape of Türkiye, where coastal and high-altitude inland regions necessitate vastly different forecasting logic, without requiring manual feature engineering or model re-selection for each city.

Beyond its empirical performance across Türkiye's 81-city temperature network, an important strength of the proposed PMFF architecture is its potential generalizability across diverse geographical and climatic conditions. By evaluating the framework across cities characterized by substantially different topographical and climatic patterns, this study demonstrates its ability to adapt the relative contribution of heterogeneous forecasting models without requiring manual model re-selection for each location. Because the core mechanisms of PMFF are fundamentally model-agnostic and adaptable to different forecasting settings, the proposed framework has the potential to be transferred to other geographic territories confronting similar climatic heterogeneity. Furthermore, this modular architecture can potentially be extended beyond air temperature to other critical environmental and operational time-series applications, including renewable energy yield prediction, precipitation and flood forecasting, air quality index estimation, and grid-level electrical load management.

Despite these strengths, the current study has certain limitations. First, the framework relies solely on historical temperature series; integrating exogenous meteorological variables such as humidity, atmospheric pressure, and solar radiation could further refine its predictive power. Second, the computational overhead of training ten distinct base models and performing dynamic pruning may pose challenges for real-time applications requiring millisecond-level inference. Finally, in the present study, the sensitivity coefficient was set to 1.5 as a practical pruning tolerance. Future research should investigate automated, data-driven approaches for adapting this threshold to different climatic conditions. Ultimately, PMFF offers a scalable and high precision architecture that synthesizes the complementary strengths of classical and advanced machine learning techniques to support environmental management and climate risk mitigation.

REFERENCES

[1] Koçak, H. (2023). Time series prediction of temperature using seasonal ARIMA and LSTM models. Gazi Journal of Engineering Sciences, 9(3), 574–584.

[2] Utku, A. (2024). Hybrid CNN-LSTM model for accurate long-term and short-term temperature prediction: A case study for Bingöl and Tunceli. International Journal of Pure and Applied Sciences, 10(2), 550–562.

[3] Utku, A., & Kaya, S. K. (2025). LSTM-based deep learning model for air temperature prediction. International Journal of Natural and Applied Sciences, 8(1), 26–32.

[4] Zhang, Z., & Dong, Y. (2020). Temperature forecasting via convolutional recurrent neural networks based on time-series data. Complexity, 2020(1), Article 3536572.

[5] Aghelpour, P., Mohammadi, B., & Biazar, S. M. (2019). Long-term monthly average temperature forecasting in some climate types of Iran using the models SARIMA, SVR, and SVR-FA. Theoretical and Applied Climatology, 138, 1471–1480.

[6] Curceac, S., Ternynck, C., Ouarda, T. B. M. J., Chebana, F., & Niang, S. D. (2019). Short-term air temperature forecasting using nonparametric functional data analysis and SARMA models. Environmental Modelling & Software, 111, 394–408.

[7] Elshewey, A. M., Shams, M. Y., Elhady, A. M., Shohieb, S. M., Abdelhamid, A. A., Ibrahim, A., & Tarek, Z. (2023). A novel WD-SARIMAX model for temperature forecasting using daily Delhi climate dataset. Sustainability, 15(1), 757.

[8] Karabulut, M. A., & Topçu, E. (2022). Derin öğrenme tekniği kullanılarak Kars ilinin hava sıcaklık tahmini. Mühendislik Bilimleri ve Tasarım Dergisi, 10(4), 1174–1181.

[9] Alomar, M. K., Khaleel, F., Aljumaily, M. M., Masood, A., Razali, S. F. M., AlSaadi, M. A., Al-Ansari, N., & Hameed, M. M. (2022). Data-driven models for atmospheric air temperature forecasting at a continental climate region. PLoS ONE, 17(11), e0277079.

[10] Apaydın, M., Yumuş, M., Değirmenci, A., & Karal, Ö. (2022). Evaluation of air temperature with machine learning regression methods using Seoul City meteorological data. Pamukkale University Journal of Engineering Sciences, 28(5), 737–747.

[11] Kumaş, K., & Akyüz, A. Ö. (2024). Predicting Konya's air temperature: Genetic programming, gradient boosting and random forest approaches. International Journal of Sustainable Engineering and Technology, 8(2), 182–191.

[12] Kumaş, K., Ayan, M., Akyüz, A. Ö., & Güngör, A. (2020). Antalya ili meteorolojik verileri yardımıyla hava sıcaklığının yapay sinir ağları metodu ile tahmini. Gümüşhane Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 10(1), 146–154.

[13] Gong, B., Langguth, M., Ji, Y., Mozaffari, A., Stadtler, S., Mache, K., & Schultz, M. G. (2022). Temperature forecasting by deep learning methods. Geoscientific Model Development, 15(23), 8931–8956.

[14] Görücü, H. A., Özdemir, Y., Yıldırım, M., & Özdemir, G. (2026). Comparison of methods for

weather temperature forecast in Vancouver City. Journal of the Institute of Science and Technology, 16(2), 425–445.

[15] Haque, E., Tabassum, S., & Hossain, E. (2021). A comparative analysis of deep neural networks for hourly temperature forecasting. IEEE Access, 9, 160646–160660.

[16] Wu, S., Shao, J., Li, G., Pu, X., Dong, Y., Du, T., Kang, J., & Zhao, B. (2026). TS-TCN: A deep learning model for ultra-short-term outdoor air temperature forecasting in complex terrain with small sample size. Energy and Buildings, 351(2026), 116710.

[17] Tran, T. T. K., Bateni, S. M., Ki, S. J., & Vosoughifar, H. (2021). A review of neural networks for air temperature forecasting. Water, 13(9), 1294.

[18] Nketiah, E. A., Chenlong, L., Yingchuan, J., & Aram, S. A. (2023). Recurrent neural network modeling of multivariate time series and its application in temperature forecasting. PLoS ONE, 18(5), e0285713.

[19] Xue, X., He, X., Liu, B., Lyu, W., & Yang, M. (2025). Subseasonal forecasting and analysis of China temperature using an explainable deep learning model. International Journal of Climatology, 45(13), e70063.

[20] Mutinda, J. K., Langat, A. K., & Mwalili, S. M. (2025). Forecasting temperature time series data using combined statistical and deep learning methods: A case study of Nairobi County daily temperature. International Journal of Mathematics and Mathematical Sciences, 2025(1), Article ID 4795841.

[21] Estrada-Patiño, E., Castilla-Valdez, G., Frausto-Solis, J., González-Barbosa, J. J., & Sánchez-Hernández, J. P. (2024). A novel approach for temperature forecasting in climate change using ensemble decomposition of time series. International Journal of Computational Intelligence Systems, 17(1), Article ID 253.

[22] Toharudin, T., Pontoh, R. S., Caraka, R. E., Zahroh, S., Lee, Y., & Chen, R. C. (2023). Employing long short-term memory and Facebook Prophet model in air temperature forecasting. Communications in Statistics - Simulation and Computation, 52(2), 279–290.

[23] Tak, N. (2018). Meta fuzzy functions: Application of recurrent type-1 fuzzy functions. Applied Soft Computing, 73, 1–13.

[24] Cifuentes, J., Marulanda, G., Bello, A., & Reneses, J. (2020). Air temperature forecasting using machine learning techniques: A review. Energies, 13(16), 4215.

[25] Grover, A., Kapoor, A., & Horvitz, E. (2015). A deep hybrid model for weather forecasting. In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 379–386).

[26] Gong, Y., Zhang, Y., Wang, F., & Lee, C. H. (2024). Deep learning for weather forecasting: A CNN-LSTM hybrid model for predicting historical temperature data. arXiv preprint arXiv:2410.14963.

[27] Coutinho, E. R., Madeira, J. G. F., Borges, D. G. F., Springer, M. V., de Oliveira, E. M., & Coutinho, A. L. G. A. (2025). Multi-step forecasting of meteorological time series using CNN-LSTM with decomposition methods. Water Resources Management, 39(7), 3173–3198.

[28] Zaka, F., Nafisah, I.A., Lin, J., Almazah, M.M.A., Hussain, I., Almazroui, M., Chakrabortty, R., & Louati, H. (2025). Real-time temperature nowcasting using deep learning models across multiple locations. Modeling Earth Systems and Environment, 11(3), 165.

[29] Uluocak, I., & Bilgili, M. (2024). Daily air temperature forecasting using LSTM-CNN and GRU-CNN models. Acta Geophysica, 72(3), 2107–2126.

[30] Kalimi, K., Musyafa, A., Taryo, T., Sinambela, M., & Aji, T. W. (2026). Weighted ensemble of GRU, LSTM and XGBoost for multi-horizon temperature forecasting at a Tropical Highland Station. Journal of Applied Informatics and Computing, 10(4), 3576–3584.

[31] Mollick, T., Hashmi, G., & Sabuj, S. R. (2023). A perceptible stacking ensemble model for air temperature prediction in a tropical climate zone. Discover Environment, 1(1), 15.

[32] Oettli, P., Nonaka, M., Richter, I., Koshiba, H., Tokiya, Y., Hoshino, I., & Behera, S. K. (2022). Combining dynamical and statistical modeling to improve the prediction of surface air temperatures 2 months in advance: A hybrid approach. Frontiers in Climate, 4, 862707.

[33] Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and practice (3rd ed.). OTexts.

[34] Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. Management Science, 6(3), 324–342.

[35] Box, G. E. P., & Jenkins, G. M. (1970). Time series analysis: Forecasting and control. Holden-Day.

[36] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: Forecasting and control (5th ed.). Wiley.

[37] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In B. Krishnapuram, M. Shah, A. J. Smola, C. Aggarwal, D. Shen, & R. Rastogi (Eds.),

Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).

[38] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

[39] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67.

[40] Vapnik, V. N. (1995). The nature of statistical learning theory. Springer.

[41] Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J., & Vapnik, V. N. (1997). Support vector regression machines. In M. C. Mozer, M. I. Jordan, & T. Petsche (Eds.), Advances in Neural Information Processing Systems (Vol. 9, pp. 155–161). MIT Press.

[42] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324.

[43] Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., & Inman, D. J. (2021). 1D convolutional neural networks and applications: A survey. Mechanical Systems and Signal Processing, 151, 107398.

[44] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

[45] Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271.

[46] Open-Meteo. (2026). Open-Meteo Weather API. Retrieved from https://open-meteo.com (Accessed June 1, 2026).

[47] Turkish State Meteorological Service. (2026). The state of Türkiye's climate in 2025. Republic of Türkiye Ministry of Environment, Urbanization and Climate Change. Retrieved from http://www.emcc.mgm.gov.tr/files/State_of_the_Climate_in_Turkey_in_2025.pdf (Accessed June 1, 2026).

[48] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research. 7, 1–30.