CLUSTER-BASED MACHINE LEARNING APPROACHES FOR PREDICTING DAILY MAXIMUM TEMPERATURES IN INDONESIA UNDER CLIMATE CHANGE

Uston Nawawi Christanto[1]; Brina Miftahurrohmah[1*]; Taufiqotul Bariyah[2]; Heri Kuswanto[3]; Niswatun Faria[4]

Departement of Information Systems[1], Departement of Informatics[2], Departement of Engineering Managemen[4]
Universitas Internasional Semen Indonesia, Gresik, Indonesia[1,2,4]
https://www.uisi.ac.id[1,2,4]
uston.christanto21@student.uisi.ac.id; brina.miftahurrohmah@uisi.ac.id*; taufiqotul.bariyah@uisi.ac.id; niswatun.faria@uisi.ac.id[4]

Department of Statistics[3]
Institut Teknologi Sepuluh Nopember, Indonesia[3]
https://www.its.ac.id/[3]
heri_k@statistika.its.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *Climate change is increasing the frequency of extreme temperatures in Indonesia, creating significant prediction challenges due to its geographical diversity. To address this, the study proposes a spatially adaptive framework using BNU-ESM and ERA5 data (1980–2005). The Indonesian region was classified into four climate clusters via K-Means, where Support Vector Regression (SVR), Random Forest (RF), and XGBoost models were evaluated. Results show SVR consistently outperformed other models across all clusters. In stable regions, SVR achieved the highest accuracy (RMSE 0.10; MAE 0.08) and remained superior even in the most volatile clusters. The study's novelty is the integration of clustering with comparative model evaluation, offering a robust methodology for precise, regionally adaptive climate early warning systems. Practically, this predictive model can support national mitigation strategies by enabling proactive resource allocation and targeted interventions in high-risk climate zones.*

Keywords: *climate change, climate clustering, early warning systems, spatially adaptive framework, support vector regression (SVR).*

Intisari— *Perubahan iklim meningkatkan frekuensi suhu ekstrem di Indonesia, menciptakan tantangan prediksi yang signifikan karena keragaman geografisnya. Untuk mengatasi hal ini, studi ini mengusulkan sebuah kerangka kerja yang adaptif secara spasial dengan menggunakan data BNU-ESM dan ERA5 (1980–2005). Wilayah Indonesia diklasifikasikan ke dalam empat klaster iklim melalui metode K-Means, di mana model Support Vector Regression (SVR), Random Forest (RF), dan XGBoost dievaluasi. Hasilnya menunjukkan bahwa SVR secara konsisten mengungguli model lainnya di seluruh klaster. Di wilayah yang stabil, SVR mencapai akurasi tertinggi (RMSE 0,10; MAE 0,08) dan tetap unggul bahkan di klaster paling fluktuatif. Kebaruan studi ini terletak pada integrasi klasterisasi dengan evaluasi komparatif model, menawarkan metodologi yang kuat untuk sistem peringatan dini iklim yang presisi dan adaptif secara regional. Secara praktis, model prediktif ini dapat memperkuat efektivitas kebijakan adaptasi dan mitigasi perubahan iklim di Indonesia dengan menyediakan prediksi suhu maksimum harian yang lebih akurat di tingkat lokal.*

Kata Kunci*: perubahan iklim, klasterisasi iklim, sistem peringatan dini, kerangka adaptif spasial, support vector regression (SVR).*

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6749

INTRODUCTION

Global climate change is becoming a concern, characterized by an increased frequency and magnitude of extreme climate events [1]. Experienced in Indonesia according to a recorded rising in the typical temperature by around 0.8°C each century besides the record-high maximum temperature in 2024 [2], [3], [4]. Indonesia as an archipelagic nation equipped with varied topography is most vulnerable to impacts from extreme heat, which are deadly threats for the stability of food security, public health, and the quality of energy infrastructure [5], [6].

In response to these growing challenges, there is an urgent need for predictive approaches capable of delivering localized and reliable estimates to support effective early warning systems and adaptation strategies. However, producing such predictions for a country as geographically and climatically diverse as Indonesia remains highly complex. As an archipelagic nation characterized by intricate topography and varied microclimates, Indonesia requires spatial modeling approaches that explicitly account for this heterogeneity. The absence of such considerations often results in substantial biases. Kurniadi et al. [7] demonstrated that low- and mid-resolution CMIP6 models tend to overestimate projections of extreme rainfall across various regions in Indonesia, particularly in areas with complex terrain such as North Sulawesi, northern Papua, and East Nusa Tenggara.

As a result, machine learning (ML) methodologies have been increasingly employed, based on their abilities for approximating complex and nonlinear patterns for large-scale spatiotemporal databases. Some of these ML-oriented methodologies, including Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and ensemble models like LightGBM and XGBoost, have been employed for temperature and rain forecast applications globally [8], [9], [10].

Although these models have proven effective, their direct application to regions with high spatial variability such as Indonesia tends to produce large bias values in predicted variables, especially when spatial variability is not explicitly accounted for. This aligns with the findings of Vivi Monita et al. [9], who used an LSTM model to predict heavy rainfall in Indonesia. They emphasized that Indonesia's unique geographical characteristics and rapidly changing weather dynamics make weather forecasting very challenging and difficult to predict accurately without carefully considering spatial and temporal variations [9], [11].

Although many studies have applied machine learning to climate prediction in Indonesia, most treat the region as spatially uniform and do not systematically integrate spatial classification techniques with tailored predictive models. Recent research, such as Najwa et al. [12], highlight the importance of accounting for local spatial variability.

However, these studies typically adopt a single-model approach or apply spatial clustering without integrating it directly into the predictive modeling workflow. To the best of our knowledge, no prior work has implemented a fully integrated framework that combines spatial clustering with comparative machine learning evaluation tailored for each spatial unit in the Indonesian context.

This study therefore introduces a novel direction by embedding K-Means-based spatial segmentation directly into a region-specific model selection pipeline, enhancing both the spatial adaptiveness and forecasting precision of climate prediction systems.

Thus, statistical downscaling and bias correction have also been considered as other ways of boosting the climatic prediction performance [13], as have machine learning techniques. Very recently, machine learning methodologies such as Support Vector Regression (SVR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) have been introduced as achieving very good precision for approximating intricate non-linear associations present within climatic data sets [14], [15], [16].

Nevertheless, accurate prediction also requires pre-modeling spatial classification. Among various clustering algorithms, K-Means, DBSCAN, Hierarchical Clustering, and Gaussian Mixture Models (GMM) are commonly used in climatological studies, as emphasized by Darmawan et al. [17]. While DBSCAN performs well for datasets with irregular shapes and noise, and Hierarchical Clustering is effective for capturing nested regional structures, these methods often suffer from high computational demands and parameter sensitivity when applied to massive datasets. This limitation is further highlighted by Harjupa et al. [18], who examined clustering applications on large-scale spatiotemporal climate datasets in Indonesia and noted that high spatial heterogeneity, coupled with computational and tuning constraints, presents significant challenges for timely and scalable implementation.

Therefore, the development of a scalable and spatially adaptive framework that integrates clustering with comparative machine learning evaluation represents a promising yet underinvestigated direction.

In this study, the selection of the K-Means clustering algorithm for Indonesia is motivated by three main considerations in the context of large-scale climate data. First, the computational efficiency and scalability of K-Means allow millions of spatial data points to be processed with minimal overhead [19]. Secondly, the climatologically interpretable output centroids show apparent regional signatures, enabling spatial delimitation of climate zones amenable to the latter stages of climate modeling [20]. Thirdly, empirically, K-Means has of late produced stable and robust results for climate zone classification, outperforming sophisticated clustering algorithms that are initiation- and parameter-setting-sensitive [21]. As such, K-Means is a justifiable and reasonable choice for managing spatial heterogeneity for Indonesian climate modeling.

In this study, we present a spatially adaptive forecasting framework for predicting the daily maximum temperature across Indonesia, utilizing climate data from the Beijing Normal University Earth System Model (BNU-ESM) and the European Centre for Medium-Range Weather Forecasts Reanalysis Fifth Generation (ERA5). The proposed framework integrates cluster-based classification with predictive modeling techniques. At the initial stage, the Indonesian region is divided into several clusters based on historical temperature patterns using K-Means Clustering. Within each cluster, three machine learning models Support Vector Regression (SVR), Random Forest (RF), and XGBoost are independently trained and evaluated. The best-performing model is then selected for each cluster, resulting in a hybrid forecasting system that is tailored and optimized according to the characteristics of each cluster.

The main contribution of this study lies not merely in the use of K-Means an approach that has been employed in previous research [22], [23] but rather in its integration into a spatially adaptive, multi-model prediction framework.

This methodological advancement bridges the gap between spatial segmentation and model-specific learning, allowing climate prediction systems to better reflect Indonesia’s regional climate complexities.

By strategically combining regional clustering with comparative model evaluation, this work proposes a new methodological direction for climate prediction in geographically diverse regions. The outcomes of this approach are expected to play a meaningful role in advancing the accuracy of early warning systems and supporting more informed climate adaptation efforts.

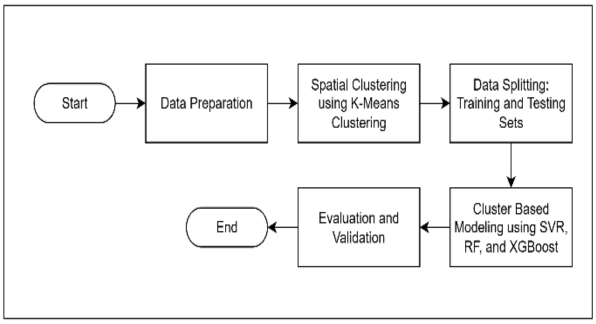
MATERIALS AND METHODS

This study takes a quantitative approach to develop mitigation strategies in response to the increasing frequency of extreme temperature events in Indonesia. The core focus is to build a predictive model for daily maximum temperatures using spatial data, based on two globally recognized climate datasets: the Beijing Normal University Earth System Model (BNU-ESM) and ERA5, developed by the European Centre for Medium-Range Weather Forecasts (ECMWF). The analysis is confined to the Indonesian region and covers the period from January 1, 1980, to December 31, 2005.

BNU-ESM is one of the climate models produced under the Coupled Model Intercomparison Project Phase 5 (CMIP5), which is designed to explore the dynamics of the global climate system and generate projections for future climate change scenarios. This model operates with a horizontal spectral resolution of T42, corresponding to a spatial grid of 2.81° longitude × 2.81° latitude, and includes 26 vertical layers based on a hybrid pressure-sigma coordinate system [24].

On the other hand, ERA5 is the fifth-generation global atmospheric reanalysis dataset developed by ECMWF. It provides hourly estimates of atmospheric, land, and oceanic variables from 1940 to the present. With a spatial resolution of about 31 km (using a 0.25° × 0.25° latitude-longitude grid), ERA5 supports detailed climate analysis at both regional and global levels.

This section presents the study’s methodology, covering data preprocessing, spatial clustering by temperature, model development, and evaluation. The stages are summarized in the flow diagram below.



Source : (Research Result, 2025)
Figure 1. Research Methodology

The methodology employed in this study involves a series of systematic steps designed to ensure data validity and enhance the accuracy of the predictive models. These steps are outlined as follows:

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6749

Data Preparation

1. Data Filtering and Cleaning
   The first step in data preparation is aimed at ensuring the quality and reliability of the datasets used in the study. Incomplete, inaccurate, or redundant data can significantly compromise the performance of predictive models. To address this, a thorough filtering and cleaning process is conducted to remove missing values, correct inconsistencies, and identify potential anomalies. This stage is essential for preserving the integrity of the data and establishing a solid foundation for subsequent analysis.

2. Resolution and Grid Alignment between BNU-ESM and ERA5
   The BNU-ESM and ERA5 datasets differ markedly in both spatial resolution and grid structure. BNU-ESM operates at a horizontal resolution of roughly 2.81°, while ERA5 provides a much finer resolution of 0.25°. This significant gap makes direct comparisons between the two datasets challenging and can potentially introduce biases into the predictive modeling results [25], [26]. To correct for this, spatial harmonization and alignment are conducted under the Climate Imprint technique—an interpolation-based approach that incorporates climatological context. This method enhances the accuracy of daily climate variables, such as temperature and precipitation, by integrating data from two different sources through the following steps:

   a) Calculating Climate Anomalies from GCM (BNU-ESM)
      Daily values from BNU-ESM are first used to compute climate anomalies, defined as the difference between each daily value and its long-term historical mean over a baseline period (1951–2005).
   b) Applying Anomalies to High-Resolution Grids
      These daily anomalies, initially derived from the coarse BNU-ESM grid, are then interpolated onto the finer-resolution grid used by observational data (e.g., ERA5). The result of this step is a high-resolution spatial field of daily anomalies, referred to as the Climate Imprint—a "fingerprint" of daily climate variability in a locally refined format.
   c) Combining with Observed Climatology
      In parallel, the ERA5 observational dataset is used to compute monthly mean climatology. This high-resolution monthly climatology is then combined with the Climate Imprint anomalies to reconstruct the final high-resolution daily fields.
   d) Saving the Final Output
      The result is a high-resolution dataset that preserves the temporal dynamics of BNU-ESM while inheriting the spatial detail and local representativeness of ERA5. This dataset is then saved for subsequent modeling steps.

   As described by Hunter and Meentemeyer [27], this method allows for enhanced spatial fidelity and reduced prediction error, particularly in regions with sparse observations or complex topography.

Spatial Clustering using K-Means Clustering

The next phase involves applying the K-Means Clustering algorithm, an unsupervised learning technique that groups data into distinct clusters based on similarities in patterns or characteristics [28]. In this study, clustering is performed on both the BNU-ESM and ERA5 datasets to create homogeneous groups of data. These clusters provide the contextual framework for developing the daily maximum temperature prediction models, allowing the models to be tailored to the unique characteristics of each cluster. To determine the optimal number of clusters (K), both the Elbow Method and Silhouette Score were employed. The Elbow Method was used to observe the point at which additional clusters result in diminishing returns in terms of within-cluster variance, while the Silhouette Score provided a quantitative measure of cluster cohesion and separation. The combination of these two methods ensures a more robust and data-driven justification for selecting the appropriate number of clusters, thus improving the spatial integrity of the forecasting framework.

Data Splitting: Training and Testing Sets

After the clustering process, the data is split into two sets: 80% for training and 20% for testing. This division is applied proportionally across all clusters to ensure balanced representation. By training the model on historical data and evaluating it on previously unseen data, this approach supports a more reliable assessment of model performance. Additionally, it helps mitigate the risk of overfitting—a condition where the model performs exceptionally well on the training set but struggles to generalize to new or unseen data [29].

Temperature Prediction Modeling

At this stage, temperature prediction models are developed using the selected algorithms and the training dataset. The following algorithms are employed:

1. Support Vector Regression (SVR)
   Support Vector Regression is a regression-based extension of the Support Vector Machine (SVM) algorithm [30]. It has demonstrated strong performance in handling large-scale datasets and is widely applied in temperature forecasting due to its high predictive accuracy [31]. SVR employs various kernel functions—such as linear, polynomial, sigmoid, and Radial Basis Function (RBF)—to transform input data into higher-dimensional spaces, enabling the model to capture complex, nonlinear relationships. Mathematically, the SVR model can be represented as:

$$f\ (x) = \ w^T \varphi(x) + b \tag{1}$$

   Where w is the weight vector, b is the bias, and $\varphi(x)$ is the mapping function from low-dimensional input space to high-dimensional feature space [32].
   The model also incorporates Structural Risk Minimization, formulated as:

$$\min \frac{1}{2}\ \|w\|^2 \ + \ C\ \sum \sum_{i=1}^{n} (\varepsilon k\ + \varepsilon k^*) \tag{2}$$

   Here, C controls the trade-off between model complexity and prediction tolerance, while $\varepsilon k\ +\ \varepsilon k^*$ are slack variables. The RBF kernel function used to evaluate proximity between input vectors is defined as:

$$(-\gamma \|x - x'\|^2)\ \gamma \tag{3}$$

2. Random Forest (RF)
   Random Forest (RF) is an ensemble learning algorithm widely applied in both regression and classification tasks [33]. In the context of temperature prediction, RF has shown strong performance and, in some studies, even outperformed other predictive models [34]. The final prediction of Random Forest is computed as the average of all individual decision trees:

$$\hat{y}(x) = \frac{1}{J} \sum_{j=1}^{J} h_j\ (x) \tag{4}$$

   Where $\hat{y}(x)$ is the final RF prediction, J is the number of trees, and $h_j\ (x)$ is the prediction of the j-th tree for input x. Minimizing the mean squared error loss function is expressed as:

$$E_{XY}[(Y - f(X))^2] \tag{5}$$

   to obtain accurate maximum temperature predictions.

3. Extreme Gradient Boosting (XGBoost)
   XGBoost is an enhancement of the Gradient Boosting algorithm, designed to increase both efficiency and predictive accuracy through advanced regularization and optimization techniques. The model constructs decision trees iteratively, minimizing loss functions at each step. Mathematically, XGBoost is formulated as:

$$\hat{y_i} = \ \emptyset(x_i) = \sum_{k=1}^{K} f_k(x_i)\ f_k\ \in F \tag{6}$$

   where y_i is the predicted value, and f_k is the regression tree function at iteration k, from the function space F. The model is optimized by sequentially adding trees to minimize the loss function and applying regularization to control model complexity.

4. Implementation Model for Each Cluster
   After clustering using K-Means, the predictive models (SVR, RF, and XGBoost) are applied independently to each cluster [35]. Since each cluster contains data with similar characteristics, this approach allows each model to be specifically trained and optimized, thereby improving the contextual accuracy of daily maximum temperature predictions [36].

These three models were specifically chosen because they have been consistently proven to deliver high accuracy in capturing complex, non-linear patterns in climatic datasets, as shown in recent studies [14], [15], [16]. Compared to other machine learning or deep learning methods, SVR, RF, and XGBoost offer a practical balance of predictive performance, computational efficiency, and model interpretability, making them highly suitable for this study's objectives and data scale.

Model Evaluation and Validation
To ensure the reliability and robustness of the developed models, evaluation is carried out using the testing data. Three primary performance metrics are applied:

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6749

1. Root Mean Square Error (RMSE)
   RMSE measures the average magnitude of prediction errors. A lower RMSE indicates better model performance [37].

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{x}_t - x_t)^2} \tag{7}$$

2. Mean Absolute Percentage Error (MAPE)
   MAPE expresses prediction error as a percentage of the actual value, making it easier to interpret:

$$\text{MAPE} = \frac{\sum_{i=1}^{n}\left|\frac{A_i - P_i}{A_i}\right| \times 100}{n} \tag{8}$$

   where A_i and P_i are actual and predicted values, respectively [38].

3. Mean Absolute Error (MAE)
   MAE calculates the average absolute difference between actual and predicted values, providing a direct assessment of prediction deviations [39]:

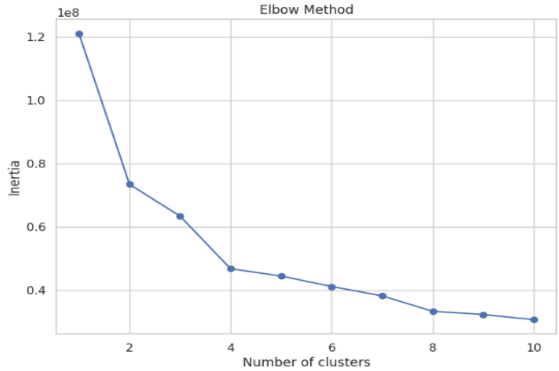
$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{9}$$

RESULTS AND DISCUSSION

In the data preparation stage, analysis showed no missing values or duplicate entries in the BNU-ESM or ERA5 datasets. Both datasets have been thoroughly cleaned and are ready for reliable use in this study.

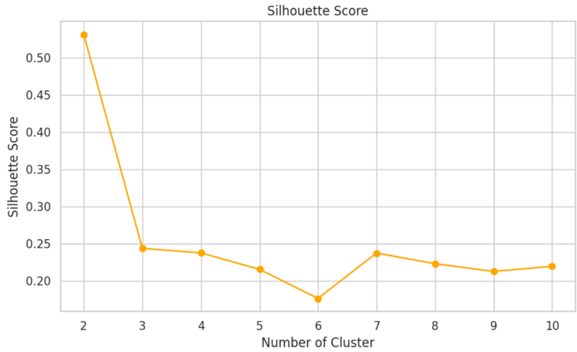
The next step involves aligning the spatial structure and harmonizing the resolution of the datasets using the Climate Imprint method. This process ensures that both datasets share consistent spatial characteristics, enabling more accurate and coherent integration for temperature prediction analysis.

Following spatial alignment, the BNU-ESM dataset undergoes further analysis using the K-Means Clustering algorithm to group regions based on similar temperature patterns. To determine the optimal number of clusters, both the Elbow Method and Silhouette Score were used. The Elbow Method identified the point of diminishing returns, while the Silhouette Score validated cluster cohesion and separation.



Source : (Research Result, 2025)
Figure 2. Elbow Method

Based on the Elbow Method illustrated in Figure 2; the inertia value decreases significantly from K = 1 to K = 4. Beyond K = 4, the reduction in inertia becomes more gradual, indicating a point of diminishing returns in reducing within-cluster variance. This inflection point suggests that K = 4 is a reasonable choice, as it balances clustering compactness with interpretability.



Source : (Research Result, 2025)
Figure 3. Silhoutte Score

As shown in Figure 3, the highest Silhouette Score occurs at K = 2, indicating the most distinct average separation between clusters at that point. However, the score gradually decreases for K > 2, suggesting that increasing the number of clusters leads to slightly less cohesive groups in terms of internal similarity and inter-cluster separation

Although the Silhouette Score peaks at K = 2, this study selected K = 4 as the optimal number of clusters based on the Elbow Method and the specific context of the data. In climate-related datasets, choosing K = 2 is often too coarse to capture the full complexity of daily temperature variation, which may involve transitional regimes, seasonal shifts, and localized patterns. Selecting K = 4 allows for a more detailed segmentation of temperature characteristics, which is particularly relevant for downstream applications such as daily maximum temperature prediction.

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v11i1.6749

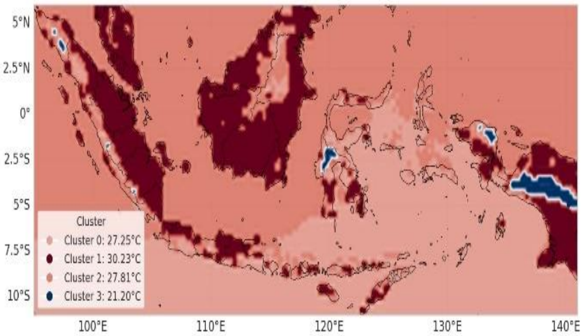
The clustering labels derived from the BNU-ESM data were subsequently used to classify the ERA5 dataset. This approach ensures structural consistency between the two datasets, enabling a more focused and coherent predictive analysis across both sources.

Table 1. BNU-ESM Clustering Characteristics

Cluster	Mean (°C)	Var	Min (°C)	Max (°C)
0	27.24	0.71	22.85	30.08
1	30.23	1.67	26.58	34.12
2	27.81	0.66	25.05	31.37
3	21.20	7.15	14.50	25.63

Source : (Research Result, 2025)

The visualization of the clustering results based on the BNU-ESM data is presented in Figure 3.



Source : (Research Result, 2025)
Figure 3. Cluster of BNU-ESM

Figure 3 presents the spatial distribution of clusters derived from daily maximum temperature data across Indonesia, covering the period from 1980 to 2025. The data were classified into four main clusters, each representing areas with distinct thermal characteristics. Average values for each cluster are shown in the figure legend and detailed in Table 1.

1. Cluster 0 is located in mountainous regions of western Sumatra, northern Java coast, and western Papua coastal areas, with a moderate mean maximum temperature of approximately 27.25°C and low variance (0.71). These regions are thermally stable with minimal daily fluctuations, likely influenced by elevation or persistent cloud cover.
2. Cluster 1 covers extensive regions including eastern Sumatra, central Kalimantan, southern Sulawesi, and parts of Papua. It exhibits the highest mean daily maximum temperature of 30.23°C and the greatest variability (variance of 1.67), suggesting hot tropical lowland zones or areas experiencing extreme dry seasons with significant daily temperature swings.
3. Cluster 2 is found in central Sulawesi, western coastal Kalimantan, and southern Papua, with a mean maximum temperature of 27.81°C and similarly low variability (variance of 0.66), reflecting warm yet thermally stable environments that may offer relatively comfortable conditions.
4. Cluster 3, primarily situated in highland areas such as mountainous zones of Papua and central Sulawesi, records the lowest mean maximum temperature of 21.20°C but the highest variance (7.15), indicative of cooler mountain climates characterized by wide diurnal temperature ranges.

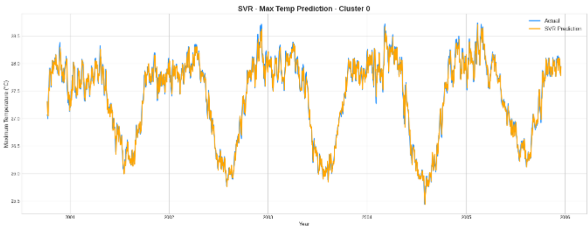
The minimum and maximum daily maximum temperatures observed from 1980 to 2025 also vary notably among the clusters. Cluster 0 ranges from about 22.86°C to 30.08°C, while Cluster 1 spans from 26.59°C to 34.12°C, indicating hotter extremes. Cluster 2 shows temperatures between 25.05°C and 31.37°C, and Cluster 3 has the widest range from 14.50°C up to 25.64°C, reflecting the pronounced diurnal and seasonal temperature swings typical of highland areas.

After determining the clusters and thoroughly understanding their distinct thermal characteristics, the next step is to develop machine learning models tailored for each cluster. This targeted modeling approach aims to correct biases inherent in the original temperature data by accounting for the unique climatic conditions within each cluster. By building specialized models per cluster, the prediction accuracy can be improved, leading to more reliable regional climate analyses and better-informed natural resource management decisions.

The visualization presented reflects the entirety of the testing data from 2000 to 2005. This period was intentionally selected to capture the dynamics of daily maximum temperatures over an extended timeframe, allowing for a more representative observation of temperature variation patterns.

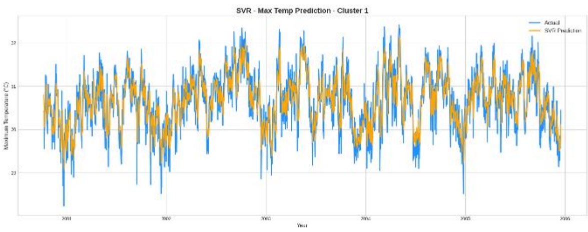
The first model employed in this process is Support Vector Regression (SVR). For Support Vector Regression (SVR), the parameter search included values for C [0.1, 1, 10, 100, 500], which controls the balance between model complexity and error tolerance; epsilon [0.01, 0.1, 0.5], which defines the margin where no penalty is applied during training; and gamma, tested with options ['scale', 0.001, 0.01, 0.1], which determines how far the influence of a single training point reaches in the feature space. The RBF kernel was chosen because it effectively captures non-linear relationships in the data.

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6749



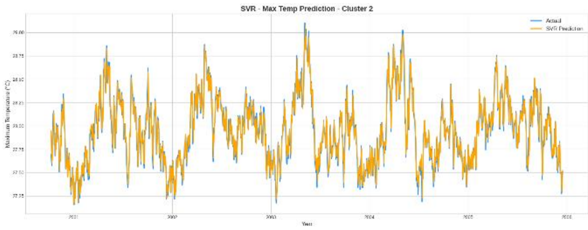
Source : (Research Result, 2025)
Figure 4. SVR Cluster 0

Based on the graph in Figure 4, the daily maximum temperature predictions for Cluster 0 using the Support Vector Regression (SVR) model on the testing dataset (2000–2005). The results demonstrate excellent performance, with the SVR model's prediction line (in orange) closely following the pattern of actual temperature fluctuations (in blue) with a high level of accuracy. The model effectively captures the consistent annual temperature cycle in this cluster, with maximum temperatures ranging from approximately 25.5°C at the lowest points to peaks around 28.5°C. This accurate performance was achieved after optimization, with the best-selected parameters being {'C': 100, 'epsilon': 0.1, 'gamma': 0.001, 'kernel': 'rbf'}. Consequently, the final SVR model SVR(C=100, gamma=0.001) has proven to be highly reliable for predicting temperatures in Cluster 0, which was previously identified as having moderate and stable temperature characteristics.



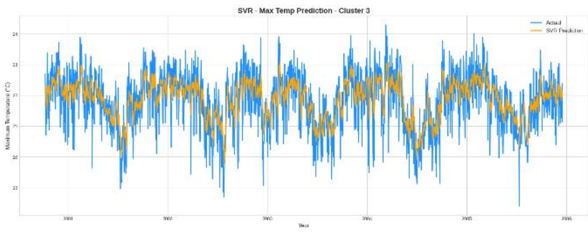
Source : (Research Result, 2025)
Figure 5. SVR Cluster 1

As illustrated in Figure 5, characterized by significantly higher and more fluctuating temperature variability, the SVR model demonstrates a fairly good performance in capturing the overall temperature trend. As shown in the graph, the SVR prediction line (orange) follows the movement pattern of the actual temperature data (blue), although the model tends to slightly smooth the data, making it challenging to precisely capture some extreme temperature peaks and the lowest valleys. Specifically, the SVR model produces temperature predictions mostly within the range of 29.5°C to 32.0°C. This optimal performance was achieved using the best parameters selected through tuning: {'C': 10, 'epsilon': 0.01, 'gamma': 0.001, 'kernel': 'rbf'}. Thus, the final SVR model established for this cluster is SVR(C=10, gamma=0.001), optimized to handle the more complex temperature dynamics in this hottest cluster.



Source : (Research Result, 2025)
Figure 6. SVR Cluster 2

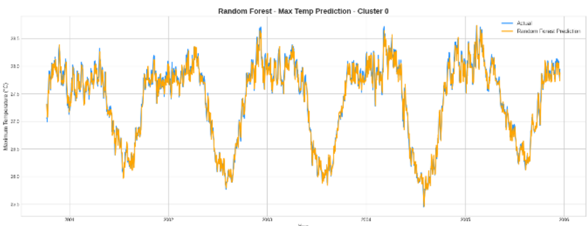
Based on the graph in Figure 6, the SVR model for Cluster 2 once again demonstrates outstanding performance in predicting daily maximum temperatures. The graph shows that the SVR prediction line (in orange) closely follows the fluctuations of the actual data (in blue) with remarkable precision, successfully capturing rapid and complex temperature variations. The model effectively maps temperature changes within a range of approximately 27.25°C to 29.0°C. This exceptional performance was achieved using the optimized best-fit parameters: {'C': 10, 'epsilon': 0.01, 'gamma': 0.001, 'kernel': 'rbf'}. Accordingly, the final SVR model—SVR(C=10, epsilon=0.01, gamma=0.001)—is established as a highly reliable and accurate model for regions exhibiting the temperature characteristics observed in Cluster 2.



Source : (Research Result, 2025)
Figure 7. SVR Cluster 3

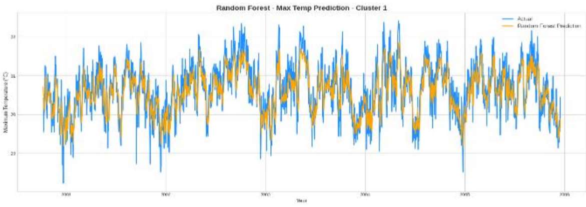
Based on the graph in Figure 7, for Cluster 3 which represents the most challenging region with the lowest temperature range and highest variability (characteristic of mountainous areas) the SVR model demonstrates a reasonably good ability to capture the general temperature pattern. The graph indicates that although the SVR predictions (in orange) follow the main trend of the actual data (in blue), the model tends to apply significant smoothing. As a result, it struggles to capture extreme temperature points, both the highest peaks and the lowest valleys, within this highly fluctuating dataset. The model is able to predict temperatures within a range of approximately 20.5°C to 23.0°C. This performance was achieved using the optimal parameter set: {'C': 1, 'epsilon': 0.1, 'gamma': 0.001, 'kernel': 'rbf'}. Therefore, the final model configuration SVR(C=1, gamma=0.001) is considered the most suitable for balancing trend detection and model complexity in regions with highly unpredictable climate conditions.

For the second algorithm, Random Forest (RF), hyperparameter tuning was conducted by adjusting several key parameters: n_estimators [100, 200], which defines the number of decision trees in the ensemble; max_depth [10, 20, None], which controls the maximum depth of each tree; min_samples_split [2, 5], specifying the minimum number of samples required to split an internal node; min_samples_leaf [1, 2], indicating the minimum number of samples needed at a leaf node; and max_features ['sqrt', 1.0], which determines the number of features considered when selecting the best split. These configurations were explored to optimize the trade-off between model complexity, accuracy, and generalization capability.



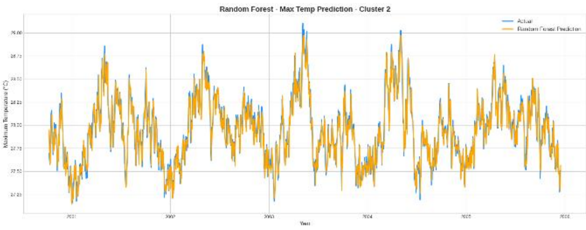
Source : (Research Result, 2025)
Figure 8. RF Cluster 0

Based on the graph in Figure 8, the Random Forest (RF) model for Cluster 0 demonstrates an exceptionally high level of accuracy. The graph clearly shows that the RF prediction line (in orange) almost perfectly overlaps the actual temperature line (in blue). This indicates the model's capability to capture not only the stable annual temperature cycles but also sharper daily fluctuations with remarkable precision. The model effectively predicts maximum temperatures across the full range, varying from approximately 25.5°C to 28.7. This superior performance was achieved using the optimized best parameters: {'max_depth': 10, 'max_features': 1.0, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 200}. Therefore, the final Random Forest model Random Forest Regressor (max_depth= 10, min_samples_leaf= 2, min_samples_split= 5, n_estimators= 200) proves to be a highly robust and accurate model for the Cluster 0 region.



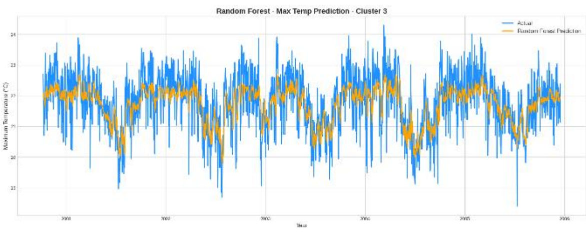
Source : (Research Result, 2025)
Figure 9. RF Cluster 1

For Cluster 1, characterized by high temperatures and volatility, the Random Forest (RF) model demonstrates highly accurate performance. The RF prediction line (orange) closely follows the actual temperature data (blue), successfully capturing most of the temperature peaks and valleys. The model predicts maximum temperatures ranging from 29.0°C to 32.5°C using the optimized parameters: {'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 200}. Consequently, this Random Forest Regressor proves to be reliable for Cluster 1.



Source : (Research Result, 2025)
Figure 10. RF Cluster 2

For Cluster 2, the Random Forest (RF) model once again demonstrates highly precise and reliable performance. The visualization shows the RF prediction line (in orange) accurately tracking the movements of the actual temperature data (in blue), including complex daily fluctuations. The model successfully predicts maximum temperatures within the range of 27.25°C to 29.0°C. This performance was achieved using the optimal parameters: {'max_depth': 10, 'max_features': 1.0, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 200} and proves to be highly effective for Cluster 2.



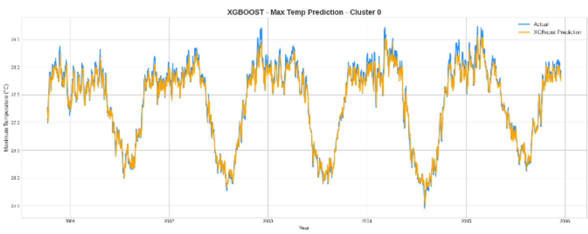
Source : (Research Result, 2025)
Figure 11. RF Cluster 3

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)
VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6749

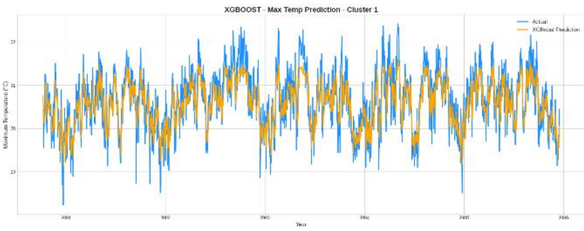
For Cluster 3, characterized by the highest temperature variability and the most challenging conditions, the Random Forest (RF) model demonstrates a significant improvement in performance. The RF prediction line (in orange) captures the main temperature patterns more effectively than the SVR model, responding more sensitively to temperature fluctuations despite reduced precision at extreme points. The model predicts maximum temperatures ranging from approximately 20.0°C to 23.0°C using the optimal parameters {'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 200}. These results confirm that RF offers a more adaptive and robust solution for mountainous regions with highly dynamic temperature patterns.

The final model applied in this study is Extreme Gradient Boosting (XGBoost). For the XGBoost Regressor, the hyperparameter tuning involved exploring a grid that included n_estimators [100, 200, 300] to determine the number of boosting rounds; max_depth [5, 7] to control the complexity of each tree; and learning_rate [0.05, 0.1] to regulate the contribution of each tree during training. Additionally, fixed values were set for subsample [0.8] and colsample_bytree [0.8], which specify the fractions of training data and features used in each boosting iteration, respectively. Model optimization was carried out using Grid Search, with negative mean squared error (neg_mean_squared_error) as the scoring metric, allowing the identification of the best parameter combination for accurate temperature prediction within each cluster.



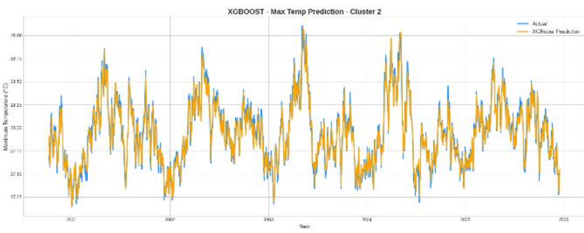
Source : (Research Result, 2025)
Figure 12. XGBoost Cluster 0

Figure 12 illustrates that the XGBoost model for Cluster 0 delivers highly accurate predictions, comparable to the outstanding performance of the Random Forest model. The graph shows an almost perfect alignment, where the XGBoost prediction line (in orange) precisely follows every detail of the actual temperature data (in blue), capturing both the annual cycles and daily fluctuations with remarkable accuracy. The model effectively predicts maximum temperatures within the range of 25.75°C to 28.75°C. This high level of precision was achieved using the optimized parameter set: {'colsample_bytree': 0.8, 'learning_rate': 0.05, 'max_depth': 5, 'n_estimators': 100, 'subsample': 0.8}. These results confirm that XGBoost is a highly robust and effective model for temperature prediction in Cluster 0.



Source : (Research Result, 2025)
Figure 13. XGBoost Cluster 1

In Cluster 1, characterized by high temperatures and extreme variability, the XGBoost model demonstrates strong and competitive performance. The graph shows that the XGBoost prediction line (in orange) closely tracks the actual temperature data (in blue), effectively capturing rapid dynamics and wide temperature swings—an indication of the model's robustness under challenging conditions. The predicted maximum temperatures accurately span the range from 29.0°C to 32.5°C. This level of accuracy was achieved using the optimal parameter configuration: {'colsample_bytree': 0.8, 'learning_rate': 0.05, 'max_depth': 5, 'n_estimators': 100, 'subsample': 0.8}. These results reaffirm that XGBoost, like Random Forest, is a highly suitable and reliable model for predicting temperatures in this highly dynamic cluster.

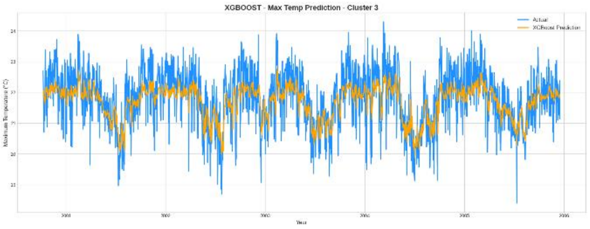


Source : (Research Result, 2025)
Figure 14. XGBoost Cluster 2

For Cluster 2, the XGBoost model continues its trend of exceptional performance, delivering highly accurate and precise results. The visualization reveals a very close alignment between the XGBoost prediction line (in orange) and the actual temperature data (in blue), with the model adeptly capturing the complex fluctuations that characterize this cluster. The predicted maximum temperatures consistently fall within the range of 27.25°C to 29.0°C. This strong performance is supported by the optimal parameter set:

P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v11i1.6749

{'colsample_bytree': 0.8, 'learning_rate': 0.05, 'max_depth': 5, 'n_estimators': 100, 'subsample': 0.8}. These findings confirm that XGBoost is a highly reliable and effective model for temperature prediction in Cluster 2.



Source : (Research Result, 2025)

Figure 15. XGBoost Cluster 3

Figure 15 presents the XGBoost model for Cluster 3—identified as the most challenging cluster due to its high temperature variability. Despite this complexity, the model delivers solid and competitive performance. The visualization illustrates how the XGBoost predictions (in orange) closely follow the main trends of the actual temperature data (in blue). Similar to Random Forest, this model outperforms SVR in handling high volatility, although a smoothing effect remains apparent, where the model struggles to fully capture the most extreme temperature points. It effectively predicts maximum temperatures within a range of approximately 20.0°C to 23.0°C. This reliable performance was achieved using the optimal hyperparameters: {'colsample_bytree': 0.8, 'learning_rate': 0.05, 'max_depth': 5, 'n_estimators': 100, 'subsample': 0.8}. These results complete the analysis by demonstrating that XGBoost remains a robust model, even under the most difficult climate prediction conditions.

Table 2. Evaluation and Validation Model

Cluster	Model	RMSE	MAPE	MAE
0	SVR	0.10	0.31	0.08
	RF	0.11	0.33	0.09
	XGBoost	0.12	0.34	0.09
1	SVR	0.43	1.13	0.34
	RF	0.44	1.17	0.35
	XGBoost	0.44	1.17	0.35
2	SVR	0.11	0.31	0.08
	RF	0.11	0.33	0.09
	XGBoost	0.11	0.32	0.09
3	SVR	0.72	2.63	0.56
	RF	0.73	2.68	0.58
	XGBoost	0.73	2.70	0.58

As shown in Table 2, the performance of SVR, Random Forest (RF), and XGBoost was assessed across four climatic clusters. The results revealed that SVR consistently achieved the best accuracy across all clusters, based on standard evaluation metrics—Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE).

In Cluster 0, all models performed well due to the relatively low variability of the temperature data. SVR recorded the lowest RMSE (0.10) and MAE (0.08), outperforming both RF and XGBoost, which produced slightly higher error rates (RMSE up to 0.12, MAE up to 0.09). This suggests that in stable climatic regions, SVR benefits from its ability to approximate continuous and smooth relationships without overfitting.

In Cluster 1, a region with slightly higher variability, SVR again outperformed both tree-based models with RMSE of 0.43, MAE of 0.34, and MAPE of 1.13%. RF and XGBoost both recorded RMSE of 0.44 and MAE of 0.35. The marginal advantage in error metrics demonstrates that SVR can capture subtle local fluctuations that are not easily modeled by discrete tree splits.

Cluster 2 also showed uniform model performance, with SVR maintaining a slight lead (RMSE = 0.11, MAE = 0.08) over RF and XGBoost. This reinforces SVR's robustness in modeling moderately variable data across geographical areas.

Cluster 3, characterized by the highest temperature volatility, posed the greatest challenge. Despite an increase in error values across all models, SVR still achieved the lowest RMSE (0.72) and MAE (0.56), compared to RF and XGBoost (RMSE = 0.73, MAE = 0.58). This indicates that SVR retains stability even under high-variance conditions, where overfitting and noise sensitivity typically degrade model performance.

In addition, SVR's uniform dominance across all clusters implies the presence of complex nonlinear interactions and localized behaviors in the underlying temperature data structure—patterns that ensemble learners like RF and XGBoost often struggle to capture. Unlike tree-based methods, which approximate functions via hierarchical partitioning, SVR—particularly when paired with the Radial Basis Function (RBF) kernel—projects input features into a higher-dimensional space, allowing the model to learn subtle, nonlinear spatial-temporal dependencies that are otherwise inaccessible in the original feature space.

This insight aligns with the nature of the data used: daily maximum temperature records over diverse Indonesian regions, characterized by topographic variability, region-specific climate drivers, and moderate noise. The ability of SVR to generalize across this heterogeneity stems from its structural risk minimization framework, which prioritizes optimal generalization rather than fitting all training samples tightly, as is often the case in ensemble tree learners.

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6749

Moreover, hyperparameter tuning revealed that SVR adapts well to local dynamics. For example, stable clusters performed best with higher penalty (C=100) and smaller gamma values, while volatile clusters required less aggressive regularization (C=1), reflecting its flexibility in controlling bias-variance tradeoffs depending on spatial complexity..

These findings are also consistent with prior research such as Syahreza et al. [31], which reported that XGBoost achieved a general MAE of 0.3744 and MSE of 0.2278 on unsegmented temperature datasets. In comparison, the cluster-based SVR approach in this study achieved an overall MAE of just 0.265, with values as low as 0.08 and RMSE down to 0.10 in certain clusters. This indicates that spatial segmentation, when combined with nonlinear modeling using SVR, effectively reduces prediction errors and enhances the model's capacity to capture local climatic patterns. While tree-based models like XGBoost and RF remain effective in generalized conditions, they are less capable in high-resolution, location-specific forecasting—highlighting the superior adaptability of cluster-based SVR in modeling spatial climate variability.

Overall, Support Vector Regression (SVR) emerged as the most consistently accurate and technically robust model across all climatic clusters. It achieved the lowest average RMSE and MAE scores, ranging from 0.10 to 0.72 and 0.08 to 0.56 respectively, outperforming both Random Forest and XGBoost in all evaluated segments. The model's effectiveness stems not only from superior numerical results but also from its underlying ability to capture spatial-temporal nonlinearities through kernel-based learning. This combination of statistical performance and technical soundness reinforces SVR as the most reliable model for high-resolution, cluster-based temperature forecasting in the context of Indonesia's diverse climatic zones.

CONCLUSION

This research successfully developed a daily maximum temperature prediction model for Indonesia using bias-corrected BNU-ESM and ERA5 data (1980–2005). Using climate-based clustering (Climate Imprint and K-Means), three models—SVR, Random Forest, and XGBoost—were trained across four clusters, with SVR consistently performing best across all segments.

The cluster-based approach enabled better capture of local temperature patterns, offering high-resolution insights that support early warning systems and local climate services.

However, this study has several limitations. First, the model only used daily temperature data, without other environmental variables like humidity, radiation, or land surface conditions. Second, performance in Cluster 1 and Cluster 3 showed that all models struggled to capture sharp fluctuations—likely due to not just high variance, but also coarse or noisy data in those areas. This suggests a need for more refined inputs and methods.

For future research, we recommend using richer predictors, higher-resolution data (e.g., sub-daily or satellite), and exploring deep learning models like LSTM that can better handle sequence and noise in volatile clusters.

REFERENCE

[1] B. Fallah and M. Rostami, "Exploring the impact of the recent global warming on extreme weather events in Central Asia using the counterfactual climate data," *Clim Change*, vol. 177, no. 5, May 2024, doi: 10.1007/s10584-024-03743-0.

[2] S. Prasetyo, N. Nuraini, P. Studi Meteorologi, and S. Tinggi Meteorologi Klimatologi dan Geofisika, "Analisis tren perubahan suhu udara dan curah hujan berdasarkan data observasi BMKG di Kota Denpasar," *Jl. Prof. Dr. Hamka, Kampus UNP Air Tawar*, 2021, doi: 10.24036/geografi/volx-issx/xxx.

[3] D. Bidang Klimatologi Badan Meteorologi and D. Geofisika Jakarta, "Catatan Iklim Kualitas Udara Indonesia 2024," 2024.

[4] World Meteorological Organization, "WMO report documents spiralling weather and climate impacts," World Meteorological Organization. Accessed: Jun. 06, 2025. [Online]. Available: https://wmo.int/news/media-centre/wmo-report-documents-spiralling-weather-and-climate-impacts

[5] T. M. Basuki *et al.*, "Improvement of Integrated Watershed Management in Indonesia for Mitigation and Adaptation to Climate Change: A Review," Aug. 01, 2022, *MDPI*. doi: 10.3390/su14169997.

[6] D. Colozza *et al.*, "The impact of climate change on child nutrition in Indonesia: a conceptual framework and scoping review of the available evidence," *BMJ Paediatr Open*, vol. 9, no. 1, p. e002980, Mar. 2025, doi: 10.1136/bmjpo-2024-002980.

[7] A. Kurniadi, E. Weller, Y. H. Kim, and S. K. Min, "Evaluation of Coupled Model Intercomparison Project Phase 6 model-

simulated extreme precipitation over Indonesia," *International Journal of Climatology*, vol. 43, no. 1, pp. 174–196, Jan. 2023, doi: 10.1002/joc.7744.

[8] I. Ebtehaj and H. Bonakdari, "CNN vs. LSTM: A Comparative Study of Hourly Precipitation Intensity Prediction as a Key Factor in Flood Forecasting Frameworks," *Atmosphere (Basel)*, vol. 15, no. 9, Sep. 2024, doi: 10.3390/atmos15091082.

[9] V. Monita, S. Raniprima, and N. Cahyadi, "Comparative Analysis of Daily and Weekly Heavy Rain Prediction Using LSTM and Cloud Data," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, vol. 10, no. 4, pp. 833–842, 2024, doi: 10.26555/jiteki.v10i4.30374.

[10] M. Putra, M. S. Rosid, and D. Handoko, "High-Resolution Rainfall Estimation Using Ensemble Learning Techniques and Multisensor Data Integration," *Sensors*, vol. 24, no. 15, Aug. 2024, doi: 10.3390/s24155030.

[11] K. Yadav, S. Malviya, and A. K. Tiwari, "Improving Weather Forecasting in Remote Regions Through Machine Learning," *Atmosphere (Basel)*, vol. 16, no. 5, May 2025, doi: 10.3390/atmos16050587.

[12] S. Najwa *et al.*, "Integrating Spatial Autoregressive Exogenous with Ordinary Kriging for Improved Rainfall Prediction in Java: Enhancing Accuracy with Climate Variables and Spatial Autocorrelation," *InPrime: Indonesian Journal of Pure and Applied Mathematics*, vol. 7, no. 1, pp. 25–43, Mar. 2025, doi: 10.15408/inprime.v7i1.42070.

[13] R. R. Wood, J. Janzing, A. van Hamel, J. Götte, D. L. Schumacher, and M. I. Brunner, "Comparison of high-resolution climate reanalysis datasets for hydro-climatic impact studies," Oct. 07, 2024. doi: 10.5194/egusphere-2024-2905.

[14] B. Miftahurrohmah, H. Kuswanto, D. S. Pambudi, F. Fauzi, and F. Atmaja, "Assessment of the Support Vector Regression and Random Forest Algorithms in the Bias Correction Process on Temperatures," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 637–644. doi: 10.1016/j.procs.2024.03.049.

[15] B. Miftahurrohmah, I. A. Cholilie, S. U. Wijaya, F. Atmaja, T. Bariyah, and C. Wulandari, "Future Growing Seasons: Bias Correction with SVR and QDM for Indonesian Temperature Projection under RCP 2.6 and RCP 8.5," *J Curr Sci Technol*, vol. 15, no. 2, Apr. 2025, doi: 10.59796/jcst.V15N2.2025.100.

[16] F. Atmaja, B. Miftahurrohmah, I. A. Cholilie, S. U. Wijaya, N. Haliza, and Y. A. Utami, "Improving Agricultural Productivity in Indonesia Using Xgboost Precipitation Prediction with Downscaling and Bias Correction," in *2024 7th International Conference on Informatics and Computational Sciences (ICICoS)*, 2024, pp. 485–490. doi: 10.1109/ICICoS62600.2024.10636822.

[17] Z. M. E. Darmawan, A. F. Dianta, K. Fathoni, O. C. R. Rachmawati, and K. I. Apriandy, "Comparison of Machine Learning Classification Methods for Weather Prediction: A Performance Analysis," *G-Tech: Jurnal Teknologi Terapan*, vol. 9, no. 2, pp. 715–727, Apr. 2025, doi: 10.70609/gtech.v9i2.6649.

[18] W. Harjupa and E. Nakakita, "Investigation and future projections of warm rain during the winter monsoon in the Western Java Sea, Indonesia," *Frontiers in Climate*, vol. 7, May 2025, doi: 10.3389/fclim.2025.1581382.

[19] D. Zhao, Z. Chen, Z. Li, X. Yuan, and I. Taniguchi, "Improving Building Temperature Forecasting: A Data-driven Approach with System Scenario Clustering," Feb. 2024, [Online]. Available: http://arxiv.org/abs/2402.13628

[20] Q. Van Doan, T. Amagasa, T. H. Pham, T. Sato, F. Chen, and H. Kusaka, "Structural k-means (S k-means) and clustering uncertainty evaluation framework (CUEF) for mining climate data," *Geosci Model Dev*, vol. 16, no. 8, pp. 2215–2233, Apr. 2023, doi: 10.5194/gmd-16-2215-2023.

[21] L. A. Pampuch, R. G. Negri, P. C. Loikith, and C. A. Bortolozo, "A Review on Clustering Methods for Climatology Analysis and Its Application over South America," *International Journal of Geosciences*, vol. 14, no. 09, pp. 877–894, 2023, doi: 10.4236/ijg.2023.149047.

[22] Q. Van Doan, T. Amagasa, T. H. Pham, T. Sato, F. Chen, and H. Kusaka, "Structural k-means (S k-means) and clustering uncertainty evaluation framework (CUEF) for mining climate data," *Geosci Model Dev*, vol. 16, no. 8, pp. 2215–2233, Apr. 2023, doi: 10.5194/gmd-16-2215-2023.

[23] O. Kisi, S. Heddam, K. S. Parmar, A. Petroselli, C. Külls, and M. Zounemat-Kermani, "Integration of Gaussian process regression and K means clustering for enhanced short term rainfall runoff modeling," *Sci Rep*, vol.

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6749

15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-91339-8.

[24] X. Yu, C. Liu, X. Wang, J. Cao, J. Dong, and Y. Liu, "Evaluation of Arctic Sea Ice Drift and its Relationship with Near-surface Wind and Ocean Current in Nine CMIP6 Models from China," *Adv Atmos Sci*, vol. 39, no. 6, pp. 903–926, 2022, doi: 10.1007/s00376-021-1153-4.

[25] M. D. Bailey, D. Nychka, M. Sengupta, A. Habte, Y. Xie, and S. Bandyopadhyay, "Regridding uncertainty for statistical downscaling of solar radiation," *Adv Stat Climatol Meteorol Oceanogr*, vol. 9, no. 2, pp. 103–120, Dec. 2023, doi: 10.5194/ascmo-9-103-2023.

[26] M. A. Stern *et al.*, "Selecting the Optimal Fine-Scale Historical Climate Data for Assessing Current and Future Hydrological Conditions," 2022, doi: 10.1175/JHM-D-21.

[27] R. D. Hunter and R. K. Meentemeyer, "Climatologically Aided Mapping of Daily Precipitation and Temperature," 2025.

[28] E. A. Saputra and Y. Nataliani, "Analisis Pengelompokan Data Nilai Siswa untuk Menentukan Siswa Berprestasi Menggunakan Metode Clustering K-Means," *Journal of Information Systems and Informatics*, vol. 3, no. 3, 2021, [Online]. Available: http://journal-isi.org/index.php/isi

[29] I. C. Azhari and T. Haryanto, "Modeling Of Hyperparameter Tuned RNN-LSTM and Deep Learning For Garlic Price Forecasting In Indonesia," *JOURNAL OF INFORMATICS AND TELECOMMUNICATION ENGINEERING*, vol. 7, no. 2, pp. 502–513, Jan. 2024, doi: 10.31289/jite.v7i2.10714.

[30] F. Setio Pribadi, F. Matematika dan Ilmu Pengetahuan Alam, U. Negeri Semarang, and B. Meteorologi Klimatologi dan Geofisika, "Penerapan Algoritma Linear Regression dan Support Vector Regression dalam Prediksi Temperatur Udara di Malang Application of Linear Regression and Support Vector Regression Algorithms in Air Temperature Prediction in Malang," 2025.

[31] A. Syahreza, N. K. Ningrum, and M. A. Syahrazy, "Perbandingan Kinerja Model Prediksi Cuaca: Random Forest, Support Vector Regression, dan XGBoost," *Edumatic: Jurnal Pendidikan Informatika*, vol. 8, no. 2, pp. 526–534, Dec. 2024, doi: 10.29408/edumatic.v8i2.27640.

[32] G. N. Lesnusa, Dwi Shinta Angreni, and R. Ardiansyah, "Perbandingan Akurasi Linear Regression dan Support Vector Regression dalam Prediksi Suhu Rata-Rata," *The Indonesian Journal of Computer Science*, vol. 13, no. 4, Jun. 2024, doi: 10.33022/ijcs.v13i4.3944.

[33] Z. A. Dwiyanti and C. Prianto, "Prediksi Cuaca Kota Jakarta Menggunakan Metode Random Forest," *Jurnal Tekno Insentif*, vol. 17, no. 2, pp. 127–137, Oct. 2023, doi: 10.36787/jti.v17i2.1136.

[34] F. Sevgin, "Machine Learning-Based Temperature Forecasting for Sustainable Climate Change Adaptation and Mitigation," *Sustainability (Switzerland)*, vol. 17, no. 5, Mar. 2025, doi: 10.3390/su17051812.

[35] H. Taneja, V. Jain, A. Jain, A. K. Dubey, M. C. Taplamacioglu, and M. Demirci, "Weather prediction using regression algorithm and neural network technique," *International Journal on "Technical and Physical Problems of Engineering*, 2024, [Online]. Available: www.iotpe.com

[36] A. Wickramasinghe, S. Muthukumarana, D. Loewen, and M. Schaubroeck, "Temperature clusters in commercial buildings using k-means and time series clustering," *Energy Informatics*, vol. 5, no. 1, Dec. 2022, doi: 10.1186/s42162-022-00186-8.

[37] Saylor Academy, "CS205: Metrics for Linear Regression Effectiveness: R-squared, MSE and RSE | Saylor Academy." Accessed: May 18, 2025. [Online]. Available: https://learn.saylor.org/mod/page/view.php?id=80811

[38] Arize Machine Learning Course, "Mean Absolute Percentage Error (MAPE): What You Need To Know - Arize AI." Accessed: May 18, 2025. [Online]. Available: https://arize.com/blog-course/mean-absolute-percentage-error-mape-what-you-need-to-know/

[39] Arize Machine Learning Course, "Mean Absolute Error In Machine Learning: What You Need To Know - Arize AI," 2023, Accessed: May 18, 2025. [Online]. Available: https://arize.com/blog-course/mean-absolute-error-in-machine-learning-what-you-need-to-know/