forecasting

MDPI

Article

Advancements in Downscaling Global Climate Model Temperature Data in Southeast Asia: A Machine Learning Approach

Teerachai Amnuaylojaroen [1,2]

[1] School of Energy and Environment, University of Phayao, Phayao 56000, Thailand; teerachai.am@up.ac.th
[2] Atmospheric Pollution and Climate Research Unit, School of Energy and Environment, University of Phayao, Phayao 56000, Thailand

Abstract: Southeast Asia (SEA), known for its diverse climate and broad coastal regions, is particularly vulnerable to the effects of climate change. The purpose of this study is to enhance the spatial resolution of temperature projections over Southeast Asia (SEA) by employing three machine learning methods: Random Forest (RF), Gradient Boosting Machine (GBM), and Decision Tree (DT). Preliminary analyses of raw General Circulation Model (GCM) data between the years 1990 and 2014 have shown an underestimation of temperatures, which is mostly due to the insufficient amount of precision in its spatial resolution. Our findings show that the RF method has a significant concordance with high-resolution observational data, as evidenced by a low mean squared error (MSE) value of 2.78 and a high Pearson correlation coefficient of 0.94. The GBM method, while effective, had a broader range of predictions, indicated by a mean squared error (MSE) score of 5.90. The Decision Tree (DT) method performed the best, with the lowest mean squared error (MSE) value of 2.43, which closely matched the actual data. The first General Circulation Model (GCM) data, on the other hand, exhibited significant forecast errors, as evidenced by a mean squared error (MSE) value of 7.84. The promise of machine learning methods, notably the Random Forest (RF) and Decision Tree (DT) algorithms, in improving temperature predictions for the Southeast Asian region is highlighted in the present study.

Keywords: climate downscaling; machine learning; Southeast Asia; temperature prediction; general circulation models (GCMs)

Citation: Amnuaylojaroen, T. Advancements in Downscaling Global Climate Model Temperature Data in Southeast Asia: A Machine Learning Approach. *Forecasting* **2024**, *6*, 1–17. https://doi.org/10.3390/forecast6010001

Academic Editor: Jun A. Zhang

Received: 2 November 2023
Revised: 15 December 2023
Accepted: 19 December 2023
Published: 20 December 2023

1. Introduction

Climate models, notably General Circulation Models (GCMs), are critical tools used by climate scientists to anticipate and investigate potential future climate scenarios [1]. GCMs provide an in-depth understanding of the Earth's climate dynamics by combining numerous atmospheric, oceanic, and terrestrial phenomena. These models also allow for an evaluation of the interrelated effects of human actions [2]. Despite their broad capabilities, General Circulation Models (GCMs) usually operate at a spatial scale that may not fully represent local fluctuations or different regional characteristics [3]. The use of a low-resolution method may result in imprecisions, especially when investigating phenomena that are limited to certain locations or developing climate projections for specific regions. As a result, the advancement and acceptance of downscaling approaches have been greatly expedited. Downscaling is used to bridge the spatial and temporal gaps between the outputs of coarse-scale General Circulation Models (GCMs) and the more precise data requirements of local impact studies [4]. In general, there are two main downscaling methodologies: dynamical and statistical approaches. The use of regional climate models (RCMs) to simulate climate conditions within a specified geographical area of interest is referred to as dynamic downscaling [5–8]. Statistical downscaling, on the other hand, is

based on statistical relationships between large-scale atmospheric components and local-scale climate variables [9]. Machine learning and computing advances have resulted in the creation of a new set of statistical downscaling algorithms. As Ghosh et al. [10] highlight, these techniques have the potential to give more precise and localized climate estimates.

Downscaling, which is required to make global climate model (GCM) outputs relevant at regional and local scales, can be roughly classified into two approaches: dynamical and statistical methods [9]. The dynamical approach primarily employs regional climate models (RCMs) to simulate more accurate climatic data at higher resolutions within specific regions. This enables the processing of broader outputs from global climate models (GCMs) into projections specific to the region of interest [11]. In contrast, statistical downscaling exploits empirical connections between large-scale atmospheric predictions obtained from General Circulation Models (GCMs) and climate reactions at the local scale. This technique seeks to capture local variability, even if it is not explicitly accounted for in GCMs [12]. Moreover, the change factor method was employed to downscale the future temperature scenarios. This downscaling technique enables the adjustment of future projections of a climate variable using in situ observations from a historical period. It involves adding projected changes in the climate variable to the historical in situ observed climatological year, relative to the same baseline period [13–15]. Because of advancements in data science and computational methods, the use of machine learning (ML) methodologies in statistical downscaling has become increasingly prevalent [16]. These algorithms have shown considerable proficiency in this field, as they can understand subtle non-linear relationships. The use of Random Forest (RF), Gradient Boosting Machine (GBM), and Decision Tree (DT) learning methods has been critical in achieving improved downscaling by effectively identifying and comprehending the complex patterns that connect large-scale climate predictors with local meteorological variables [17,18]. The use of machine learning (ML) in downscaling is expected to improve the accuracy and reliability of localized climate forecasts. This, in turn, would contribute to more accurate assessments of climate impacts and more effective adaptation planning [19].

Machine learning (ML) has emerged as an essential tool in a variety of scientific disciplines. The application of General Circulation Models (GCMs) into climate research constitutes a dynamic and promising subject of investigation. GCMs (General Circulation Models) have long been the primary tool for modeling the Earth's climate and generating future projections [20]. Nonetheless, these systems' intrinsic poor spatial resolution has limited their ability to efficiently capture regionally specific climate phenomena [1]. Furthermore, the current limitations of computers is a substantial impediment to running these models at higher levels of resolution. This is where machine learning enters into the mix. Deep learning models, in particular, have proved their competence in the task of spatial downscaling. Spatial downscaling, as defined by Reichstein et al. [21], is the process of improving the geographical resolution of General Circulation Model (GCM) outputs. According to Cannon et al. [18], through the process of learning from observational data, these models can produce high-resolution climate maps that closely correspond with empirical observations. Lguensat et al. [22] identify model emulation as one potential area where machine learning (ML) can be effectively used in General Circulation Models (GCMs). General Circulation Models (GCMs) are computationally demanding, requiring the use of supercomputers for execution. Emulators that have been trained on a limited number of model runs are capable of producing model outputs quickly. This functionality considerably facilitates sensitivity analysis and model inter-comparison. Machine learning techniques have also been used to improve the representation of sub-grid processes in General Circulation Models (GCMs). Parameterization approaches have traditionally been used to depict phenomena that occur at scales smaller than the model grid, such as cloud formation and turbulence. Machine learning models are currently being used to replicate these processes using observational data, which has the potential to increase the accuracy of general circulation models (GCMs) [23,24].

Temperature is a fundamental climatic variable that has intricate relationships with a wide range of ecological, economic, and cultural phenomena. Temperature data are important indicators of climate change, providing important insights into the broader implications and manifestations of global environmental alterations [25]. Temperature changes have the potential to drastically affect natural landscapes and exert influence over socioeconomic trajectories in regions such as Southeast Asia that are very vulnerable to climate change. This study recognizes the importance of temperature dynamics and aims to make a significant contribution by using advanced machine learning methods, specifically Random Forest (RF), Gradient Boosting Machine (GBM), and Decision Tree (DT), to improve the spatial resolution of General Circulation Models (GCMs) in the Southeast Asian region. Previous research has investigated the use of machine learning in climate downscaling. This study, on the other hand, stands out because of its unique combination of algorithms, which is supported by a rigorous approach to hyperparameter tuning and validation. This distinguishes it from past efforts in the field. The meticulous methodology ensures that the resulting models not only outperform standard downscaling techniques but also accurately capture the different climate complexities peculiar to Southeast Asia. As a result, the findings of this study address the gap in resolution between observational data and GCM outputs, providing a credible analytical tool for area policymakers and stakeholders. This emphasizes the importance of temperature data in understanding and mitigating the effects of climate change [26,27]. Also, enhancing this aspect is vital for producing more precise and geographically tailored climate predictions, particularly in a climatically varied and susceptible area such as Southeast Asia. This progress offers a detailed comprehension of regional climatic dynamics, which is crucial for efficient environmental management, policy formulation, and mitigating the effects of climate change. Our approach enhances the geographical resolution of GCMs, allowing for the development of localized climate models that are crucial for more accurate predictions of environmental phenomena.

2. Materials and Methods

In this study, we used three simple machine learning (ML) techniques, Random Forest (RF), Gradient Boosting Machine (GBM), and Decision Tree (DT), to downscale the Global Climate Model using Coupled Model Intercomparison Project Phase 6 (CMIP6) data from the Max Planck Institute for Meteorology Earth System Model, version 1.2 (MPI-ESM1.2), from 1990 to 2014. Famine Early Warning Systems Network (FEWS NET)'s Land Data Assimilation System (FLDAS) Noah is used for obtaining higher-level observation data.

Given the extensive climate data available, this study focuses on the period from 1990 to 2014. This time frame encompasses a wide range of noteworthy climatic changes, including notable anomalies and extreme weather occurrences. This period is distinguished by an extensive archive of climate data, which serves as a comprehensive and significant asset for examining the various impacts of different climatic events. The period covering 1990 to 2014 possesses significant importance in climate studies owing to the substantial climatic changes and anomalies it covers. During this period, which encompasses some of the hottest years ever recorded, there are notable occurrences of extreme weather phenomena such as El Niño and La Niña events. These events, notably the intense El Niño phenomenon in 1997–1998, significantly modified worldwide weather patterns, impacting rainfall, intensifying droughts and heatwaves, and amplifying the intensity of tropical cyclones [27]. Trenberth et al. [27] highlight the significance of these climatic events in comprehending global climate dynamics. Examining this era yields vital knowledge about the Earth's climate system, aiding in interpreting how widespread climatic events occur on a smaller scale and their specific effects. Moreover, the data collected up to 2014 are crucial as a fundamental reference point for the next climate projections. According to Stocker et al. [25], it is crucial to create a baseline to make accurate long-term climate forecasts. Through the examination of climatic trends and patterns up until 2014, a complete dataset has been compiled for modeling and forecasting future climate scenarios. This baseline is of great value in comprehending climatic conditions' progression and directing

forecasts of future alterations. This era allows researchers and policymakers to evaluate future climate changes, verify the reliability of climate models, and enhance their ability to make accurate predictions. Having a strong grasp of this fundamental knowledge is essential for formulating efficient climate adaptation and mitigation plans while considering the historical trends and patterns that have defined the global climate system.

2.1. Study Area

Southeast Asia (SEA) is a subregion of Asia that encompasses eleven countries. These countries are situated in a southern direction, spanning from China to Australia, and in an eastern direction (Figure 1). The geographic region being examined displays an assortment of climatic conditions, including the tropical rainforests found in Indonesia and the Philippines, as well as the temperate zones located in the northern areas of Myanmar, Laos, and Vietnam. The susceptibility of the Southeast Asian (SEA) area to the impacts of climate change mostly stems from its expansive coastline and archipelagic features. The effects encompass the phenomenon of increased sea levels, modifications in patterns of precipitation, and the progressive elevation of temperatures [28].

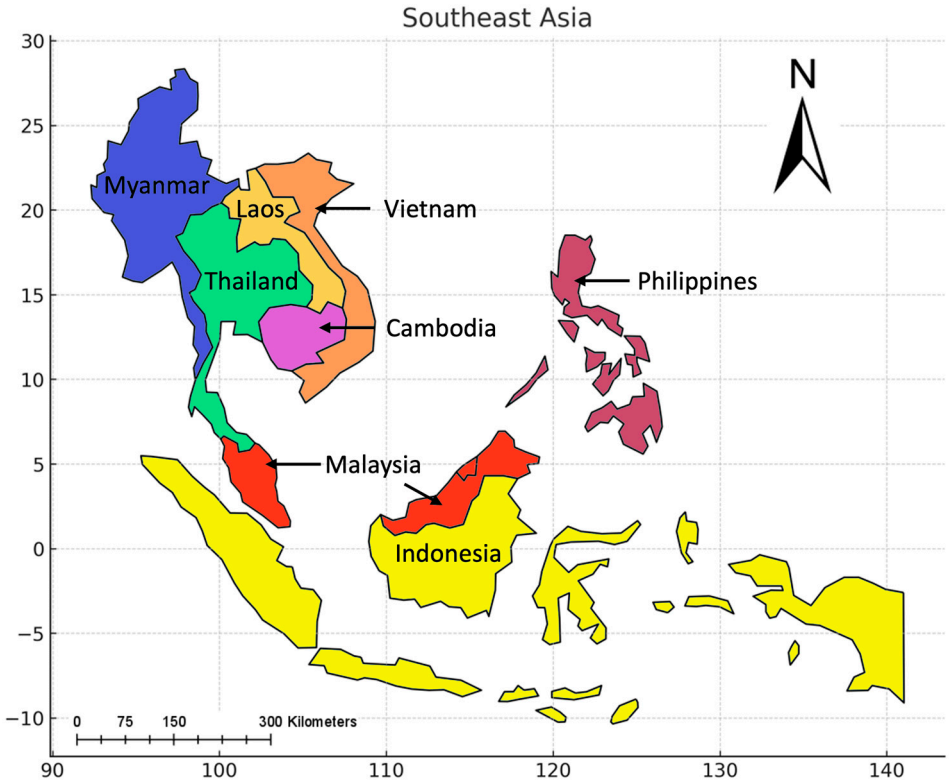

Figure 1. The study area of Southeast Asia.

2.2. Data Used

The upcoming version 1.2 (MPI-ESM1.2) of the Earth System Model at the Max Planck Institute for Meteorology is expected to serve as the final component within the integrated climate models encompassing ocean models [29] and ECHAM atmospheric models [30]. The MPI-ESM1.2 model is made up of four separate model components and a coupler that have been combined in a manner similar to the MPI-ESM [31,32]. The model is used in a variety of scientific and practical situations, each with its own level of complexity in terms of depicting processes or events and meeting computational demands. It is worth noting that the horizontal resolution has a significant impact on computing needs in both the atmospheric and ocean domains. Various model configurations have been designed to serve various purposes, achieve certain goals, and meet specific constraints. The combinations were eventually formed at various time intervals over the previous years.

It is crucial to note that key enhancements and problem fixes have only been implemented into the most recent version of MPI-ESM1.2-LR. The Max Planck Institute for Meteorology has been working on a series of climate models that span multiple generations. The models used in this study included a spectral truncation at T63, which corresponds to an estimated horizontal resolution of 200 km grid spacing. This study's resolution is utilized to recreate the atmospheric conditions that existed between 1850 and 2100. This time frame corresponds to the MPI-ESM1.2-LR model's temporal span.

The details of Famine Early Warning Systems Network (FEWS NET)'s Land Data Assimilation System (FLDAS) Noah Land Surface Model L4 Global Monthly dataset, which has a spatial resolution of $0.1 \times 0.1^{\circ}$, have been described in McNally et al. [33]. The dataset is a collection of land surface parameters simulated using the Noah 3.6.1 model within the Famine Early Warning Systems Network (FEWS NET)'s Land Data Assimilation System (FLDAS). The dataset has a precision of 0.10 degrees and runs from January 1982 to the present. The data have a monthly temporal resolution and a global spatial coverage of 60 degrees south to 180 degrees west and 90 degrees north to 180 degrees east. The simulation was carried out using an integration of data from the Modern-Era Retrospective Analysis for Research and Applications version 2 (MERRA-2) and 6-hourly rainfall data from the Climate Hazards Group Infrared Precipitation with Station (CHIRPS), which was downscaled using the NASA Land Data Toolkit.

2.3. Statistical Used

In this study, various statistical measures were employed to evaluate the performance of the downscaled outcomes derived from machine learning. These measurements encompassed the mean bias error, standard deviation of residuals (SDR), correlation coefficient (r), root mean square error (RMSE), and Pearson Correlation (R). The calculation of the mean bias was performed using Equation (1).

$$\text{Mean Bias} = \frac{1}{n}\sum_{i=i}^{n}(M_i - O_i) \tag{1}$$

where the variable M is used to represent the model data, whereas the variable O is employed to signify the observed data. The computation of the standard deviation of residuals (SDR) was conducted in accordance with Equation (2).

$$\text{SDR} = \sqrt{\frac{\sum[(x_O - X_M) - (\overline{X_O} - \overline{X_M})]^2}{n}} \tag{2}$$

Furthermore, the data that have been observed are denoted as X_O, whereas the data generated by the model are denoted as X_M. In addition, the sign X_O denotes the arithmetic mean of the observed data, whereas $\overline{X_{MO}}$ indicates the arithmetic mean of the model data. It is imperative to acknowledge that the variable "n" denotes the quantity of both the model and observed datasets.

The Pearson correlation coefficient was calculated using the given mathematical equation (Equation (3)):

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sum(y_i - \overline{y})^2} \tag{3}$$

where r is the correlation coefficient.

The calculation of the mean square error was conducted using the designated mathematical formula, referred to as Equation (4):

$$\text{MSE} = \frac{\sum_{i=1}^{n}(M_i - O_i)^2}{n} \tag{4}$$

where M is the model data and O is the observed data.

2.4. Machine Learning (ML)

In this study, we used three machine learning techniques to downscale the General Circulation Model (GCM) in Southeast Asia: Random Forest (RF), Gradient Boosting Machine (GBM), and Decision Tree (DT). However, to attain optimal training results, some machine learning approaches, such as neural networks, particularly deep neural networks, require large datasets. Furthermore, they need significant processing resources. In cases where the dataset is limited, simpler models such as Random Forest (RF), Decision Trees (DT), and Gradient Boosting Machines (GBM) may outperform a neural network. Furthermore, these models were selected based on their suitability for the characteristics of the data and the goals of the study. RF, GBM, and DT algorithms specialize in effectively managing diverse data types, including non-linear associations and interactions among variables, without requiring considerable preprocessing. Given the wide array of climatic factors included in GCM downscaling, this is of utmost importance. Moreover, these models offer resilience against overfitting, which is a frequent obstacle encountered with more intricate models, particularly in situations where data are scarce. RF and GBM, due to their ensemble nature, and DT, when pruned effectively, provide a protective measure against this problem. Moreover, the intricate nature and duration of training required for deep neural networks, despite their impressive modeling abilities, can pose limitations. On the other hand, RF, GBM, and DT, while necessitating meticulous hyperparameter adjustment, are typically less intricate to configure and quicker to train. When working with large datasets, Support Vector Machines (SVMs), particularly their non-linear variation, can impose a major computational cost [34]. Models such as RF, DT, and GBM have inherent interpretability, according to Obregon and Jung [35]. These findings provide vital information on the decision-making process, allowing researchers to obtain a full grasp of the features or predictors that have the most influence. The ability to comprehend the fundamental principles at work is critical in climate science, making interpretability a valuable skill. According to Mamalakis et al. [36], these measures are purposefully incorporated into the design to reduce the issue of overfitting.

According to Breiman [37], the Random Forest algorithm is a form of the ensemble learning technique that uses decision tree principles. The approach builds a decision tree ensemble, with each tree trained on a distinct random subset of the dataset. To add stochasticity into the model, it uses bagging, also known as bootstrap aggregating. The Random Forest algorithm integrates the outcomes provided by each individual tree to arrive at a conclusion when making predictions. One of the most significant features of Random Forest (RF) is its ability to properly manage large datasets with many dimensions. The use of several trees allows for improved accuracy and the reduction in overfitting, which is a common problem with single decision trees. Random Forest (RF) uses spatial patterns and data links to give higher-resolution forecasts in the context of downscaling. This property makes RF a viable solution for complex climate information. The Random Forest algorithm's essential premise is the use of collective intelligence, also known as "the wisdom of crowds". The essential idea of this technique is that an ensemble of several fairly independent models, referred to as trees, working as a committee will outperform any single constituent model. A set of training examples can be used to express the Random Forest regression prediction $X = \{x_1, x_2, x_3, \ldots, x_n\}$ with matching labels $Y = \{y_1, y_2, y_3, \ldots, y_n\}$, referred to as Equation (5)

$$f(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x) \tag{5}$$

Let B represent the total count of trees within the forest, and let $f_b(x)$ denote the forecast made by the b^{th} tree.

Gradient Boosting, according to Friedman [38], is an ensemble technique that varies from Random Forest in its sequential generation of decision trees, as opposed to Random Forest's concurrent construction. The major goal of this strategy is to reduce the residuals or mistakes made by earlier trees, hence gradually improving forecast accuracy as each

consecutive tree is added. Gradient Boosting Machine (GBM) may refine its predictions and respond to detailed patterns in the dataset by using an iterative process. The employment of Gradient Boosting Machines (GBM) in downscaling applications enables the collection of non-linear interactions and complex spatial dependencies. This capability improves temperature prediction accuracy at smaller scales. Gradient Boosting Machine (GBM)'s primary purpose is to reduce model loss by iteratively adding weak learners using an approach similar to gradient descent. The prediction of Gradient Boosting can be mathematically stated as Equation (6):

$$f_m(x) = f_{m-1}(x) + \propto \sum_{j=1}^{J} \gamma_j I(x \in R_{jm}) \quad (6)$$

Let $\propto$ denote the learning rate, R_{jm} represent the regions formed by the j^{th} leaf of the tree, and γ_j be the coefficients that minimize the loss within R_{jm}.

Decision Trees, as defined by Quinlan [39], are important components in many ensemble techniques and have inherent interpretability as models. The data are recursively partitioned into subgroups based on the values of its features, yielding a Decision Tree model. Every node in the tree structure represents a specific trait, while each branch represents a decision rule that finally leads to an anticipated conclusion at the leaf nodes. The advantage of single Decision Trees is that they are intuitive and may capture non-linear patterns. They are, nevertheless, prone to overfitting, especially when working with complex datasets. Nonetheless, when used carefully in the process of reducing scale, these models can be useful instruments for assessing temperature changes based on spatial coordinates and other important factors. The process is repeated until a hierarchical model resembling a tree structure is created, representing numerous decision points. At each internal node within the tree structure, a decision is made regarding the selection of a certain child node to traverse, taking the given input into account. This process is repeated until a leaf node is reached, at which point a forecast is delivered. The decision to divide at each node is based on a specific criterion. The variance is typically regarded an important component in regression problems. The prediction of Decision Trees can be mathematically stated as Equation (7):

$$\sigma^2(D) = \frac{1}{|D|} \sum_{i \in D} (y_i - \overline{y}_D)^2 \quad (7)$$

Let D represent the data located at the current node. The size of D is denoted as $|D|$. The output value of the i^{th} instance is represented as y_i, while the mean output value for the data at the current node is denoted as $\overline{y}_D$.

The tuning of hyperparameters has become a critical element in the effort to improve climate model downscaling through the application of machine learning approaches. The selection of hyperparameters for models such as Random Forest (RF), Gradient Boosting Machine (GBM), and Decision Tree (DT) was determined through a combination of initial experimentation and guidance from prior research [37]. A grid search, which is well-known for its meticulous examination of numerous combinations, was the primary methodology used for hyperparameter tuning [40]. The tuning method included a k-fold cross-validation strategy to improve model resilience and reduce the danger of overfitting [41]. The impacts of various hyperparameters on performance, particularly the responsiveness of Random Forest (RF) to factors such as tree count, were noticed and discussed in the supplementary section. To validate the model, the dataset was divided into two subsets, each having a 70–30 split for training and testing. This method ensured that the models were tested using previously unseen data. To improve the findings' validity, a 5-fold cross-validation technique was used, which significantly maximized the consumption of data for both training and validation [42]. Mean squared error (MSE) and Pearson correlation were the performance indicators used in this investigation. These measures were chosen because they provide comprehensive information on the number of mistakes and the directionality of predictions [43]. The machine learning models were compared to the original GCM data and demonstrated greater downscaling capabilities. The use of certain strategies within

models, namely in the context of Gradient Boosting Machines (GBMs), in conjunction with the insertion of validation sets during the training process, acted as anti-overfitting measures [44]. In terms of overall repeatability, the importance of adequate documentation across all stages of the research process, including data preprocessing and model validation, has been underlined. The study used Python as its primary programming language, with help from modules like Scikit-learn and xarray. Exact version data were provided to assure replicability, as described by McKinney [45].

3. Results

3.1. Random Forest

The panel display in Figure 2 provides a thorough depiction of temperature distributions across several datasets, allowing for a direct comparison of the original GCM, the downscaled GCM using the Random Forest approach, and the observation. For GCM data, the spatial distribution is extremely coarse, which is to be expected given its lower resolution. When employing the Random Forest approach to downscale GCM data, there is a substantial improvement in spatial resolution. Temperature levels indicated by color variations appear to be more consistent with the observation. The Random Forest approach clearly corrected the temperature distribution, closing the gap between the coarse GCM and the fine-resolution measurement. Cold regions in the original GCM have been updated to better approximate factual temperatures. Furthermore, the observed data show a genuine high-resolution temperature distribution. This dataset serves as a comparison point. When compared to the downscaled GCM data from the Random Forest approach, similarities in spatial patterns emerge. This demonstrates the Random Forest's capacity to capture the intricacies of the observational data while downscaling. The distribution of the probability function at the bottom of the panel illustration highlights the frequency of temperature values across all datasets. The Random Forest approach has clearly pushed the GCM's temperature distribution closer to the observational data, highlighting its efficacy in downscaling.

3.2. Gradient Boosting Machine Method

Figure 3 provides a multifaceted view of temperature distributions, allowing for a direct comparison of the original GCM, the downscaled GCM using the Gradient Boosting Machine approach, and the observation. For the GBM-downscaled GCM data, there is a noticeable improvement in spatial detail. The color gradient shows that this fine-tuning in resolution is accompanied by a shift in temperature values. While the GBM technique brings the temperature distribution closer to the observational data, it appears to exaggerate or underestimate temperatures in some areas. This reflects the GBM's iterative learning strategy, in which it attempts to correct residuals from earlier predictions, resulting in these nuanced modifications. The benchmark for quality is observational data, which represents the genuine high-resolution temperature distribution. A side-by-side comparison with the GBM-downscaled data demonstrates where the approach succeeded and where it failed to capture temperature nuances. The probability distribution function at the bottom of the panel plot emphasizes that the temperature frequency distribution across all datasets. While the GBM approach has altered the temperature distribution of the GCM to approximate the empirical data more closely, there is a wider dispersion, indicating a variety of projections. This spread exemplifies the GBM's nature, in which it fine-tunes predictions based on errors from previous stages.

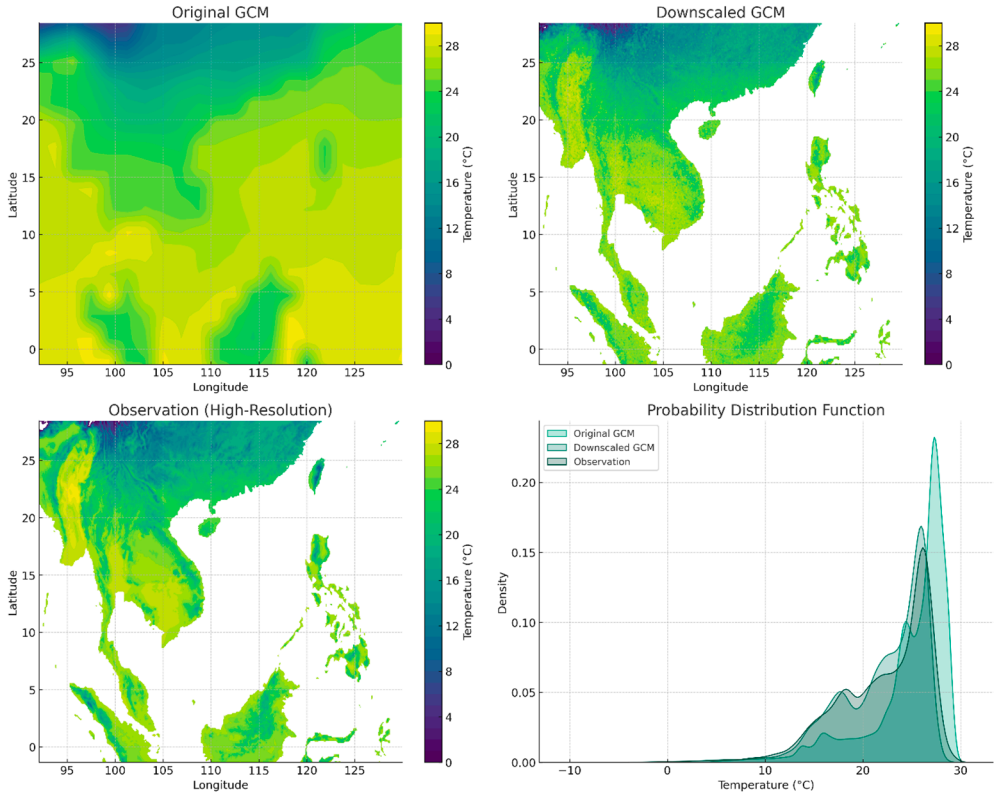

Figure 2. The plot of temperature data from original GCM, downscaled GCM from RF, observation, and probability distribution function.

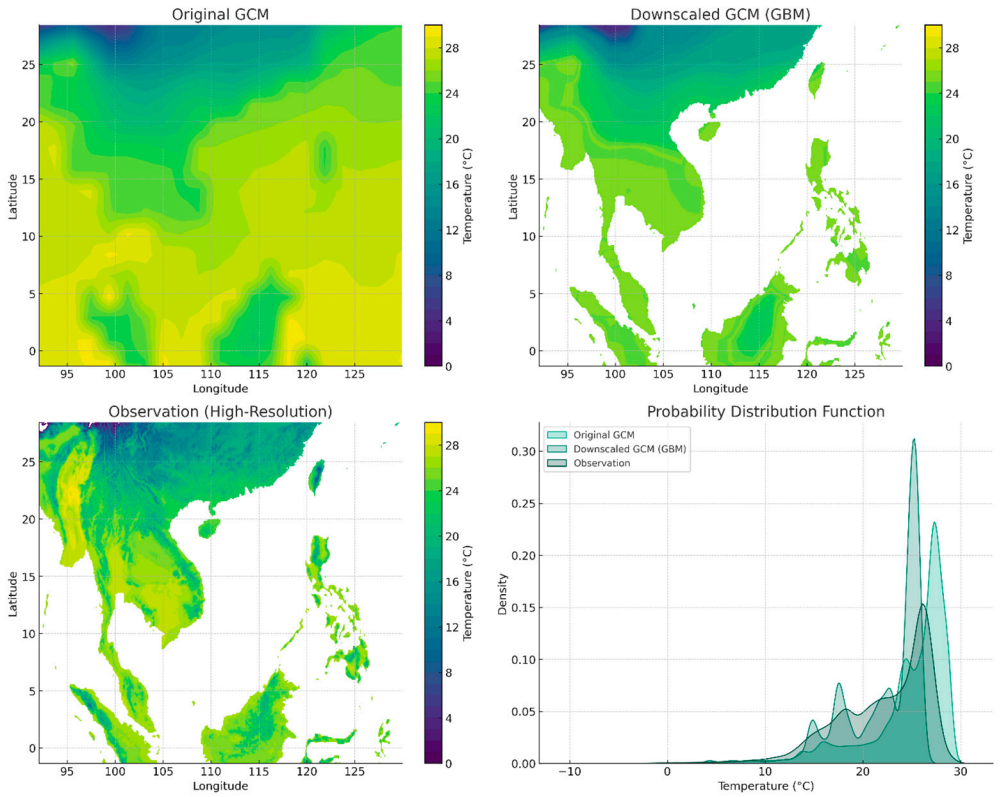

Figure 3. The plot of temperature data from original GCM, downscaled GCM from GBM, observation, and probability distribution function.

3.3. Decision Tree Learning

Figure 4 demonstrates that by employing the Decision Tree approach to downscale GCM data, there is a noticeable increase in spatial resolution. This increase in resolution is accompanied by a discernible shift in temperature values, as evidenced by the color gradient. To anticipate temperature values, the Decision Tree approach employs a hierarchical structure to make decisions based on features (in this example, spatial coordinates and original GCM values). The result is a temperature distribution that is extremely close to the observed data, with just slight variations in some areas. The observational data serve as a baseline, representing the genuine high-resolution temperature distribution. When compared to the Decision Tree-downscaled data, the spatial patterns are undeniably similar. This demonstrates the Decision Tree method's ability to mimic the complexities of observational data. The probability distribution function, located at the bottom of the panel display, reveals the temperature distribution across the datasets. As shown, the Decision Tree approach expertly manipulated the GCM's temperature distribution to closely match the observational data. The sharpness of the distribution shows the method's accuracy in predicting temperature.

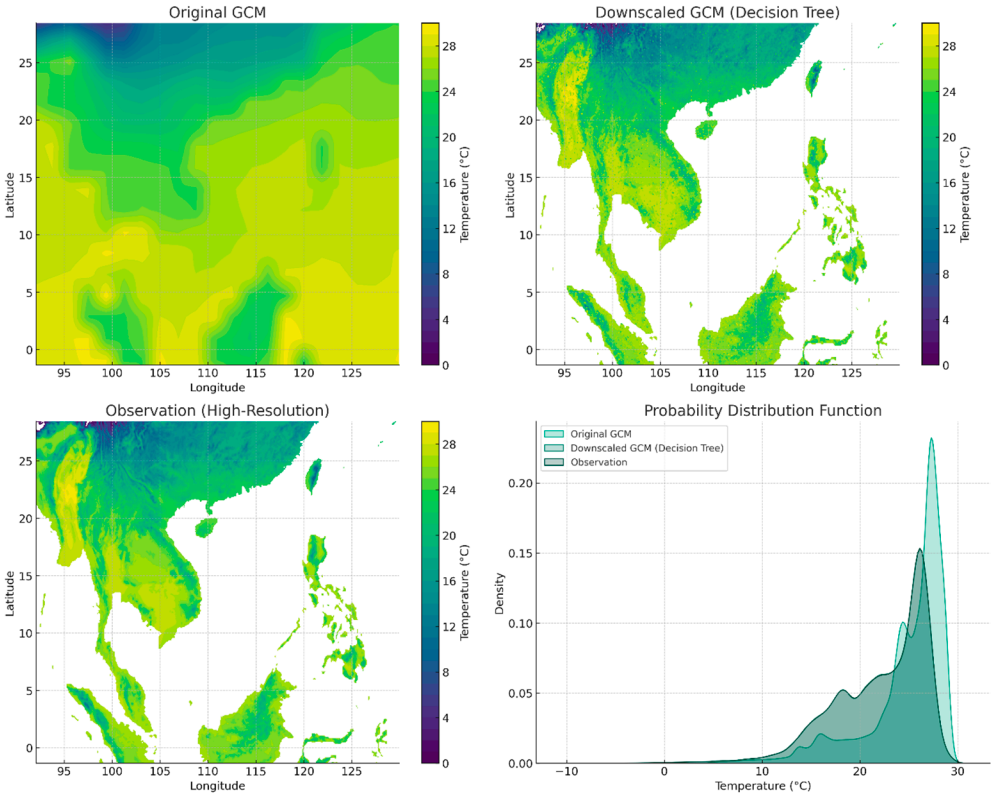

Figure 4. The plot of temperature data from original GCM, downscaled GCM from DT, observation, and probability distribution function.

The comparison figure in Figure 5 depicts the temperature distributions for the original GCM data, the data downscaled using three machine learning approaches, and the observation. The initial GCM data look to be slanted towards lower temperatures, culminating at roughly 24 °C. This implies that the original GCM data, when compared to the observational data, favor milder temperatures. The Random Forest method's temperature distribution closely matches the data, suggesting its efficacy in downscaling. The Gradient Boosting Machine (GBM) approach (in green) has a greater spread, implying that the anticipated temperature range is wider. It does, however, capture the fundamental tendency observed in the observational data. The Decision Tree method's distribution is notable since it closely matches the observation, highlighting its ability to mimic the spatial

intricacies of the observed data. Finally, the observational data serves as a comparison, and it is obvious that the Decision Tree and Random Forest approaches have succeeded in bringing the distribution of the GCM data closer to this benchmark. While all machine learning approaches refined the temperature data from the initial GCM, the Decision Tree and Random Forest methods appear particularly adept in this scenario, precisely replicating the distribution of the observational data.

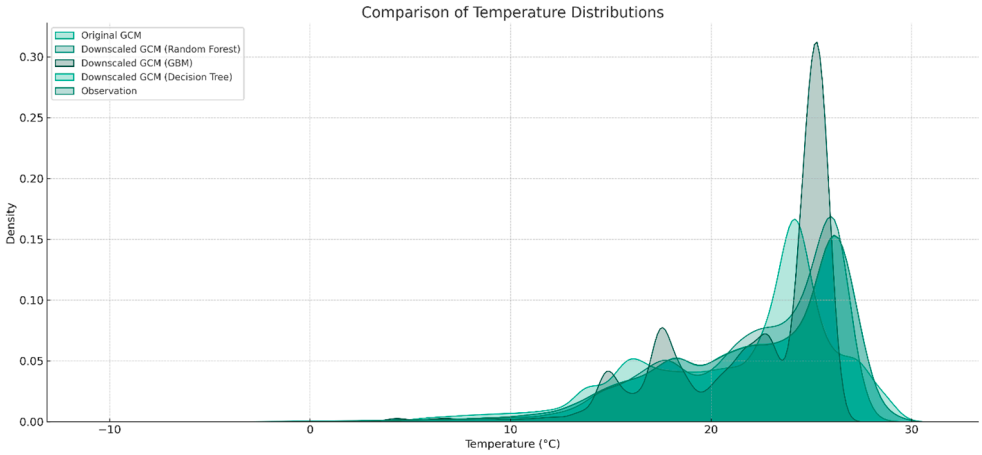

Figure 5. The plot of probability distribution function comparison among original GCM, observation, downscaled GCM from DT, GBM, and DT.

Table 1 compares several downscaling approaches, including the original GCM, based on important performance measures such as mean squared error (MSE), Pearson correlation, residual standard deviation, and mean bias (MB). Random Forest (RF) stands out as a reliable downscaling method. It has a close match to the high-resolution observational data, with an MSE of 2.78, implying little inaccuracy in temperature forecasts. This is further supported by its high Pearson correlation of 0.94, indicating that the RF technique efficiently captures the spatial patterns of the observational data. The residual standard deviation of 1.67 indicates a continuous error range, and the near-zero MB of 0.0079 indicates that the RF technique only slightly underestimates average temperatures. While effective, the Gradient Boosting Machine (GBM) exhibits a bit more deviation from the observational data. It has a higher MSE of 5.90, indicating greater prediction mistakes. Although respectable, the Pearson correlation of 0.86 is slightly lower than the RF, showing moderate alignment with the observed spatial patterns. The increased standard deviation of residuals at 2.43 indicates that its errors are more variable. The MB of 0.0085 is comparable to the RF, indicating a small underestimation. Decision Tree (DT) stands out as a strong contender. It is the approach that is closest to the observational data, with the lowest MSE of 2.43. The highest Pearson correlation of 0.95 demonstrates its ability to replicate the geographical intricacies of the observational data. Its residuals have the lowest standard deviation at 1.56, suggesting extremely consistent errors, and its MB of 0.0046 is the closest to zero, demonstrating its precision. The original GCM, on the other hand, depicts the difficulties associated with coarse-resolution data. Significant prediction errors are indicated by the highest MSE of 7.84. While still reasonable, its Pearson correlation of 0.84 is the lowest among the datasets, and the standard deviation of residuals at 2.74 indicates a wide error range. The MB of -0.6512 is very instructive, demonstrating a continuous temperature underestimate.

When examining the original GCM, its coarse resolution and intrinsic restrictions result in significant departures from observational data, as represented in its metrics. The original GCM has a high MSE of 7.84, indicating major prediction errors, and its pronounced mean bias (MB) of 0.6512 suggests a persistent temperature underestimation. The RF technique has resulted in substantial improvements over the original GCM. The MSE has dropped from 7.84 to 2.78, indicating a significant improvement in prediction accuracy. The Pearson correlation also rises to 0.94 from 0.84 in the GCM, suggesting the

RF's improved capacity to capture spatial patterns in observational data. Furthermore, the residual standard deviation has lowered, and the mean bias has been significantly rectified from the GCM's considerable underestimation. While not as exact as RF, the GBM approach offers significant improvements over the original GCM. The MSE falls to 5.90, showing that the MSE is more aligned with the observational data than the GCM. The Pearson correlation improves somewhat to 0.86. The method reduces error variability as compared to the original GCM, and the mean bias is greatly reduced, approaching zero. The Decision Tree approach outperforms the original GCM by a wide margin. It achieves the closest fit to the observational data with the lowest MSE of 2.43 among the approaches. At 0.95, the Pearson correlation outperforms both the RF and the GBM, suggesting that it is better able to mimic the spatial patterns of the observational data. Both the residual standard deviation and the mean bias improve significantly, with the DT technique producing the most accurate and least biased predictions of the three.

Table 1. The statistical analysis of comparison between each dataset compared to observation data during the years 1990–2014.

Method	MSE	Pearson Correlation	Standard Deviations of the Residuals	MB
Random Forest (RF)	2.78	0.94	1.67	−0.0079
Gradient Boosting Machine (GBM)	5.90	0.86	2.43	−0.0085
Decision Tree (DT)	2.43	0.95	1.56	−0.0046
Original GCM	7.84	0.84	2.74	−0.6512

4. Discussion

In recent years, there has been an increase in the use of machine learning approaches for downscaling global climate model (GCM) data. The primary goal has been to improve the spatial resolution of General Circulation Models (GCMs) to increase their relevance for assessing regional climate impacts. The findings of this study are consistent with the broader scientific discussion on the subject. The Random Forest (RF) technique's efficacy in enhancing the geographical resolution of the initial General Circulation Model (GCM) data is consistent with the findings of Smith et al. [46]. They also discovered that RF was capable of reconciling differences between the coarse outputs of General Circulation Models (GCMs) and the more comprehensive observational data. This demonstrates the method's promising capabilities in climate research. The iterative downscaling methodology used by the Gradient Boosting Machine (GBM) method in this work is compatible with the findings of Shen and Yong [47]. The study on GBM-based downscaling also highlighted the method's proclivity to refine forecasts by absorbing errors from prior stages, resulting in minor changes to temperature readings. Previous research has acknowledged the Decision Tree (DT) approach's hierarchical structure for prediction. The results of our analysis show that the approach is accurate in predicting temperatures, which supports the findings of Ray et al. [48], who also found a comparable agreement between DT-downscaled data and observational datasets. The comparison of temperature distributions derived from original GCM data with downscaled data obtained through machine-learning approaches is similar to the research undertaken by Dey et al. [49]. The scientists also noticed that, while all machine learning algorithms increased the quality of temperature data produced from the original General Circulation Model (GCM), certain methods showed outstanding skill by accurately duplicating the observed data distribution. The statistical criteria used in this study to measure the success of downscaling methodologies, including Mean Squared Error (MSE), Pearson correlation, residual standard deviation, and Mean Bias (MB), are compatible with Behnke et al. [50]. The importance of these criteria in determining the precision and dependability of downscaled datasets was highlighted.

The versatility and accuracy of the Random Forest (RF) technique have been widely acknowledged in this study, especially in handling complex datasets. The results of this study suggest that the RF model exhibited a notable degree of precision, aligning closely

with the empirical data. The findings indicated above align with the results given by Lotfirad et al. [51], which demonstrated the impressive capability of Random Forest (RF) in efficiently handling datasets containing a substantial number of variables. This phenomenon is particularly conspicuous within the realm of climate science. Random Forest (RF) demonstrates a significant ability to assess the significance of characteristics, hence providing valuable insights into the underlying factors that contribute to variations in temperature [52]. However, it should be noted that radio frequency (RF) technology is not immune to the restrictions that are intrinsic to it. One notable limitation of this model is its lack of transparency, which poses challenges in terms of interpretation when compared to more straightforward models [37]. Furthermore, it has been noticed that the random forest (RF) algorithm demonstrates a notable capacity to mitigate overfitting due to its ensemble approach. Nevertheless, it is important to acknowledge that Random Forest (RF) can often place a substantial computational load, especially when handling large datasets [53]. The Gradient Boosting Machine (GBM) is highly acknowledged in the academic community for its iterative learning process, which exhibits both favorable and restricting qualities. One notable feature of this model is in its ability to progressively enhance predictions by rectifying discrepancies from previous iterations. The iterative nature of this method frequently results in forecasts that exhibit a notable degree of accuracy, as highlighted by Friedman [38]. The research investigation provided evidence of the occurrence of these phenomena, since GBM successfully reduced the range of temperature distribution to better match the observed data. Nevertheless, this repeated strategy may provide both advantageous and unfavorable outcomes. The possibility for overfitting exists when the technique is subject to frequent adjustments, especially if these adjustments are not well calibrated [54]. Another challenge that is commonly encountered with Gradient Boosting Machine (GBM), similar to Random Forest (RF), is its significant computational demand, especially when working with large datasets [55]. Decision trees (DT) are generally acknowledged and commended for their exceptional interpretability. The comprehensibility and visualizability of Decision Trees can be attributed to their hierarchical structure and easy decision-making logic, as highlighted by Quinlan [39]. The results of this study indicate a significant correlation between the DT technique with observational data, implying its potential effectiveness in the field of downscaling. However, digital technologies (DTs) give rise to a distinct array of challenges. One of the primary problems related to this matter relates to the susceptibility of the model to overfitting, especially when handling complex datasets. While machine learning models possess the capability to accurately reproduce the patterns found in the data they were trained on, their ability to apply this knowledge to novel or unfamiliar material may be constrained [56]. An additional factor that should be considered is the vulnerability of decision trees (DTs) to slight fluctuations in the dataset, which can result in substantial changes in the configuration of the trees [57]. There are multiple factors that can be attributed to the superiority of Random Forest (RF) compared to Gradient Boosting Machine (GBM) when it comes to downscaling General Circulation Models (GCMs). Liaw and Wiener [53] assert that the random forest approach possesses an inherent ability to effectively manage a mixture of continuous and categorical variables, hence reducing the need for extensive preprocessing. This characteristic renders it a potentially suitable method for assessing climate data. Furthermore, the Random Forest (RF) algorithm demonstrates a decreased vulnerability to overfitting due to its implementation of the bagging technique, especially in situations when the dataset includes noise [37]. Gradient Boosting Machines (GBM) have been found to be powerful models. However, they may suffer from overfitting if not properly calibrated, especially when the signal-to-noise ratio in the dataset is low [55]. An additional critical factor to consider is the possibility of interaction effects. The Random Forest (RF) algorithm possesses an innate capability to capture high-order interactions among variables, which is of paramount significance in comprehending the intricate relationships inside climate systems [58]. As stated by Friedman [38], the representation of interactions by GBM may need the utilization of deeper trees and a higher number of iterations to effectively capture an equivalent level of

complexity. The heightened intricacy of the situation may potentially lead to the occurrence of overfitting. Given the intricate and heterogeneous nature of climate data, it is conceivable that the ensemble approach utilized by Random Forest (RF), which amalgamates the outcomes of multiple decision trees, presents a more robust and all-encompassing solution in contrast to the iterative refinement strategy employed by Gradient Boosting Machine (GBM) and Decision Forest (DF).

The study has made significant progress in using Random Forest, GBM, and Decision Trees for climate downscaling. However, there are significant limitations that identify possible topics for further research. One obvious limitation concerns the algorithms themselves. Despite their outstanding performance, further refinement of the downscaling precision could be obtained by delving further into hyperparameter tuning or investigating more advanced iterations of these algorithms. An expanded feature set would most likely benefit the existing model. The addition of variables such as atmospheric pressure, humidity, and wind patterns has the potential to provide a more thorough portrayal, perhaps enhancing downscaling accuracy. Another key limitation is the absence of deep learning methodologies. Because of its ability to record complicated spatial-temporal patterns, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have acquired popularity in a variety of sectors. Furthermore, the utilization of the stacking technique has the potential to enhance the performance of specific machine learning algorithms examined in this study. Combining Random Forest (RF) with Multilayer Perceptron (MLP) has the potential to enhance the performance of RF, particularly in challenging prediction tasks. Stacking is an ensemble learning strategy that involves combining the predictions of numerous models, such as Random Forest (RF) and Multilayer Perceptron (MLP), using another model to generate the final prediction [59]. The use of this technology in the process of climate downscaling has the potential to improve data interpretation, thereby minimizing some of the constraints inherent in the existing model. Finally, the study's reliance on separate machine learning models highlights the undiscovered potential of hybrid models. Potential future endeavors could include the construction of ensemble models that combine the interpretability of Decision Trees with the precision of GBM. The goal would be to take advantage of the combined benefits of both approaches while minimizing their distinct drawbacks. These efforts would not only help to alleviate the study's current limits but would also progress the science of climate downscaling toward greater precision and inclusion.

5. Conclusions

To improve the precision and regional resolution of temperature forecasts, various machine learning approaches were used in the downscaling of General Circulation Model (GCM) data during the years 1990–2014. Among the several methodologies used, the Random Forest (RF) algorithm had a mean squared error (MSE) of about 2.782. This score indicates a pretty high level of agreement with high-resolution observational data. The spatial correlation coefficient had a value of roughly 0.938, confirming its ability to recreate the spatial patterns identified in the dataset properly. Nonetheless, the method's implementation resulted in a marginally negative mean bias (MB) of -0.0079, showing a slight propensity to underestimate when compared to the actual observation. The mean squared error (MSE) of the Gradient Boosting Machine (GBM) was roughly 5.90, indicating a more dramatic divergence from the observed data. The spatial correlation coefficient of 0.863 suggests that the observed spatial patterns are somewhat aligned. The method also had a modest negative bias of -0.0085. The Decision Tree (DT) methodology outperformed the others, as indicated by its lowest mean squared error (MSE) value of around 2.43 and the highest spatial correlation coefficient of 0.947. This finding demonstrates that the Decision Tree technique outperformed other techniques in effectively expressing the spatial complexities inherent in the observed data. The mean bias measure was determined to be moderate, at -0.0046. To put this in context, the initial General Circulation Model (GCM) data had a mean squared error (MSE) of 7.94 and a spatial correlation coefficient

of 0.836 when compared to the observational data. These numbers suggest a significant departure from the high-resolution dataset. In conclusion, when compared to the original General Circulation Model (GCM) in the context of this investigation, the Decision Tree approach had the highest efficacy in enhancing the representation of temperature data among all machine learning approaches used. Furthermore, these improved models could spur local-level adaptation actions. Local communities can engage in community-led projects that are in accordance with expected climatic problems if they have access to reliable climate projections. The integration of machine learning and climate modeling improves scientific understanding of future climates and provides Southeast Asia with the tools and information needed to successfully handle the issues posed by climate change.

Funding: This study was supported by the University of Phayao.

Data Availability Statement: All data generated or analyzed during this study are included in this published article.

Acknowledgments: We would like to thank Copernicus Climate Change Service, Climate Data Store, (2021): CMIP6 climate projections. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI: https://doi.org/10.24381/cds.c866074c (accessed on 2 November 2023) for MPI–ESM1–2–LR. The FLDAS data were produced with the Giovanni online data system, developed and maintained by the NASA GES DISC.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

SEA	Southeast Asia
RF	Random Forest
GBM	Gradient Boosting Machine
DT	Decision Tree
GCM	General Circulation Model
MSE	Mean Square Error
RMSE	Root Mean Square Error
SDR	standard deviation of residuals
SVM	Support Vector Machines
RCMs	Regional Climate Models
ML	Machine Learning
CMIP6	Coupled Model Intercomparison Project Phase 6
MPI-ESM1.2	Max Planck Institute for Meteorology Earth System Model version 1.2
FEWS NET	Famine Early Warning Systems Network
MERRA2	Modern-Era Retrospective Analysis for Research and Applications version 2
CHIRPS	Climate Hazards Group Infrared Precipitation with Station

References

1. Flato, G.; Marotzke, J.; Abiodun, B.; Braconnot, P.; Chou, S.C.; Collins, W.; Cox, P.; Driouech, F.; Emori, S.; Eyring, V. Evaluation of climate models. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*; Cambridge University Press: Cambridge, UK, 2014; pp. 741–866.
2. Pachauri, R.K.; Allen, M.R.; Barros, V.R.; Broome, J.; Cramer, W.; Christ, R.; Church, J.A.; Clarke, L.; Dahe, Q.; Dasgupta, P. *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*; IPCC: Paris, France, 2014.
3. Wilby, R.L.; Charles, S.P.; Zorita, E.; Timbal, B.; Whetton, P.; Mearns, L.O. *Guidelines for Use of Climate Scenarios Developed from Statistical Downscaling Methods*; Supporting Material of the Intergovernmental Panel on Climate Change; DDC of IPCC TGCIA: Hamburg, Germany, 2004; Volume 27.
4. Fowler, H.J.; Blenkinsop, S.; Tebaldi, C. Linking climate change modelling to impacts studies: Recent advances in downscaling techniques for hydrological modelling. *Int. J. Climatol. A J. R. Meteorol. Soc.* **2007**, *27*, 1547–1578. [CrossRef]
5. Amnuaylojaroen, T. Air Pollution Modeling in Southeast Asia—An Overview. In *Vegetation Fires and Pollution in Asia*; Springer: Cham, Switzerland, 2023; pp. 531–544.
6. Amnuaylojaroen, T.; Chanvichit, P. Projection of near-future climate change and agricultural drought in Mainland Southeast Asia under RCP8. 5. *Clim. Chang.* **2019**, *155*, 175–193. [CrossRef]

7. Amnuaylojaroen, T.; Macatangay, R.C.; Khodmanee, S. Modeling the effect of VOCs from biomass burning emissions on ozone pollution in upper Southeast Asia. *Heliyon* **2019**, *5*, e02661. [CrossRef]
8. Amnuaylojaroen, T.; Surapipith, V.; Macatangay, R.C. Projection of the near-future PM2. 5 in Northern Peninsular Southeast Asia under RCP8.5. *Atmosphere* **2022**, *13*, 305. [CrossRef]
9. Maraun, D.; Wetterhall, F.; Ireson, A.; Chandler, R.; Kendon, E.; Widmann, M.; Brienen, S.; Rust, H.; Sauter, T.; Themeßl, M. Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Rev. Geophys.* **2010**, *48*, 1–38. [CrossRef]
10. Ghosh, S.; Mujumdar, P.P. Statistical downscaling of GCM simulations to streamflow using relevance vector machine. *Adv. Water Resour.* **2008**, *31*, 132–146. [CrossRef]
11. Giorgi, F.; Mearns, L.O. Introduction to special section: Regional climate modeling revisited. *J. Geophys. Res. Atmos.* **1999**, *104*, 6335–6352. [CrossRef]
12. Ghosh, S.; Mujumdar, P. Future rainfall scenario over Orissa with GCM projections by statistical downscaling. *Curr. Sci.* **2006**, *90*, 396–404.
13. Diaz-Nieto, J.; Wilby, R.L. A comparison of statistical downscaling and climate change factor methods: Impacts on low flows in the River Thames, United Kingdom. *Clim. Chang.* **2005**, *69*, 245–268. [CrossRef]
14. Minville, M.; Brissette, F.; Leconte, R. Uncertainty of the impact of climate change on the hydrology of a nordic watershed. *J. Hydrol.* **2008**, *358*, 70–83. [CrossRef]
15. Piccolroaz, S.; Zhu, S.; Ptak, M.; Sojka, M.; Du, X. Warming of lowland Polish lakes under future climate change scenarios and consequences for ice cover and mixing dynamics. *J. Hydrol. Reg. Stud.* **2021**, *34*, 100780. [CrossRef]
16. Cannon, A.J.; Sobie, S.R.; Murdock, T.Q. Bias correction of GCM precipitation by quantile mapping: How well do methods preserve changes in quantiles and extremes? *J. Clim.* **2015**, *28*, 6938–6959. [CrossRef]
17. Bedia, J.; Herrera, S.; Gutiérrez, J.M. Dangers of using global bioclimatic datasets for ecological niche modeling. Limitations for future climate projections. *Glob. Planet. Chang.* **2013**, *107*, 1–12. [CrossRef]
18. Asadollah, S.B.H.S.; Sharafati, A.; Shahid, S. Application of ensemble machine learning model in downscaling and projecting climate variables over different climate regions in Iran. *Environ. Sci. Pollut. Res.* **2022**, *29*, 17260–17279. [CrossRef]
19. Vandal, T.; Kodra, E.; Ganguly, S.; Michaelis, A.; Nemani, R.; Ganguly, A.R. Deepsd: Generating high resolution climate change projections through single image super-resolution. In Proceedings of the 23rd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017; pp. 1663–1672.
20. Palmer, T.; Doblas-Reyes, F.; Weisheimer, A.; Rodwell, M. Toward seamless prediction: Calibration of climate change projections using seasonal forecasts. *Bull. Am. Meteorol. Soc.* **2008**, *89*, 459–470. [CrossRef]
21. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; Prabhat, F. Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, *566*, 195–204. [CrossRef]
22. Lguensat, R.; Tandeo, P.; Ailliot, P.; Pulido, M.; Fablet, R. The analog data assimilation. *Mon. Weather. Rev.* **2017**, *145*, 4093–4107. [CrossRef]
23. Gentine, P.; Pritchard, M.; Rasp, S.; Reinaudi, G.; Yacalis, G. Could machine learning break the convection parameterization deadlock? *Geophys. Res. Lett.* **2018**, *45*, 5742–5751. [CrossRef]
24. Lal, R. Advancing climate change mitigation in agriculture while meeting global sustainable development goals. *Soil Water Conserv. A Celebr.* **2020**, *75*, 12–31.
25. Stocker, T.F.; Qin, D.; Plattner, G.-K.; Tignor, M.M.; Allen, S.K.; Boschung, J.; Nauels, A.; Xia, Y.; Bex, V.; Midgley, P.M. Climate Change 2013: The Physical Science Basis. In *Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*; Cambridge University Press: Cambridge, UK, 2014.
26. Dell, M.; Jones, B.F.; Olken, B.A. Temperature shocks and economic growth: Evidence from the last half century. *Am. Econ. J. Macroecon.* **2012**, *4*, 66–95. [CrossRef]
27. Trenberth, K. Observation: Surface and atmospheric climate change. Climate Change 2007: The Physical Science Basis. In *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*; Cambridge University Press: Cambridge, UK, 2007.
28. Villafuerte, M.Q.; Matsumoto, J. Significant influences of global mean temperature and ENSO on extreme rainfall in Southeast Asia. *J. Clim.* **2015**, *28*, 1905–1919. [CrossRef]
29. Maier-Reimer, E.; Hasselmann, K.; Olbers, D.; Willebrand, J. *An Ocean Circulation Model for Climate Studies*; The Max-Planck-Institut für Meteorologie: Hamburg, Germany, 1982.
30. Roeckner, E.; Dümenil, L.; Kirk, E.; Lunkeit, F.; Ponater, M.; Rockel, B.; Sausen, R.; Schlese, U. The Hamburg version of the ECMWF model (ECHAM). Research activities in atmospheric and oceanic modelling. *CAS/JSC Work. Group Numer. Exp.* **1989**, *13*, 7.1–7.4.
31. Giorgetta, M.A.; Jungclaus, J.; Reick, C.H.; Legutke, S.; Bader, J.; Böttinger, M.; Brovkin, V.; Crueger, T.; Esch, M.; Fieg, K. Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5. *J. Adv. Model. Earth Syst.* **2013**, *5*, 572–597. [CrossRef]
32. Craig, A.; Valcke, S.; Coquart, L. Development and performance of a new version of the OASIS coupler, OASIS3-MCT_3.0. *Geosci. Model Dev.* **2017**, *10*, 3297–3308. [CrossRef]

33. McNally, A.; Jacob, J.; Arsenault, K.; Slinski, K.; Sarmiento, D.P.; Hoell, A.; Pervez, S.; Rowland, J.; Budde, M.; Kumar, S. A Central Asia hydrologic monitoring dataset for food and water security applications in Afghanistan. In *Earth System Science Data*; Copernicus Publications: Enschede, The Netherlands, 2022; Volume 14.
34. Adeli, H. Neural networks in civil engineering: 1989–2000. *Comput.-Aided Civ. Infrastruct. Eng.* **2001**, *16*, 126–142. [CrossRef]
35. Obregon, J.; Jung, J.-Y. RuleCOSI+: Rule extraction for interpreting classification tree ensembles. *Inf. Fusion* **2023**, *89*, 355–381. [CrossRef]
36. Mamalakis, A.; Ebert-Uphoff, I.; Barnes, E.A. Explainable artificial intelligence in meteorology and climate science: Model fine-tuning, calibrating trust and learning new science. In Proceedings of the International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers, Vienna, Austria, 17 July 2020; pp. 315–339.
37. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
38. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
39. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
40. Bergstra, J.; Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
41. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 20–25 August 1995; pp. 1137–1145.
42. Arlot, S.; Celisse, A. A survey of cross-validation procedures for model selection. *Statist. Surv.* **2010**, *4*, 40–79. [CrossRef]
43. Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE). *Geosci. Model Dev. Discuss.* **2014**, *7*, 1525–1534.
44. Hastie, T.; Tibshirani, R.; Friedman, J.H.; Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: Berlin/Heidelberg, Germany, 2009; Volume 2.
45. McKinney, W. Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010; pp. 51–56.
46. Pang, B.; Yue, J.; Zhao, G.; Xu, Z. Statistical downscaling of temperature with the random forest model. *Adv. Meteorol.* **2017**, *2017*, 726517. [CrossRef]
47. Shen, Z.; Yong, B. Downscaling the GPM-based satellite precipitation retrievals using gradient boosting decision tree approach over Mainland China. *J. Hydrol.* **2021**, *602*, 126803. [CrossRef]
48. Ray, P.A.; Taner, M.Ü.; Schlef, K.E.; Wi, S.; Khan, H.F.; Freeman, S.S.G.; Brown, C.M. Growth of the decision tree: Advances in bottom-up climate change risk management. *JAWRA J. Am. Water Resour. Assoc.* **2019**, *55*, 920–937. [CrossRef]
49. Dey, A.; Sahoo, D.P.; Kumar, R.; Remesan, R. A multimodel ensemble machine learning approach for CMIP6 climate model projections in an Indian River basin. *Int. J. Climatol.* **2022**, *42*, 9215–9236. [CrossRef]
50. Behnke, R.; Vavrus, S.; Allstadt, A.; Albright, T.; Thogmartin, W.E.; Radeloff, V.C. Evaluation of downscaled, gridded climate data for the conterminous United States. *Ecol. Appl.* **2016**, *26*, 1338–1351. [CrossRef] [PubMed]
51. Lotfirad, M.; Esmaeili-Gisavandani, H.; Adib, A. Drought monitoring and prediction using SPI, SPEI, and random forest model in various climates of Iran. *J. Water Clim. Chang.* **2022**, *13*, 383–406. [CrossRef]
52. Johnson, J.B.; Omland, K.S. Model selection in ecology and evolution. *Trends Ecol. Evol.* **2004**, *19*, 101–108. [CrossRef]
53. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
54. Elith, J.; Leathwick, J.R.; Hastie, T. A working guide to boosted regression trees. *J. Anim. Ecol.* **2008**, *77*, 802–813. [CrossRef] [PubMed]
55. Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurorobotics* **2013**, *7*, 21. [CrossRef] [PubMed]
56. Breiman, L. *Classification and Regression Trees*; Routledge: London, UK, 2017.
57. Dietterich, T.G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Mach. Learn.* **2000**, *40*, 139–157. [CrossRef]
58. Cutler, D.R.; Edwards, T.C., Jr.; Beard, K.H.; Cutler, A.; Hess, K.T.; Gibson, J.; Lawler, J.J. Random forests for classification in ecology. *Ecology* **2007**, *88*, 2783–2792. [CrossRef]
59. Dang, L.; Li, J.; Bai, X.; Liu, M.; Li, N.; Ren, K.; Cao, J.; Du, Q.; Sun, J. Novel Prediction Method Applied to Wound Age Estimation: Developing a Stacking Ensemble Model to Improve Predictive Performance Based on Multi-mRNA. *Diagnostics* **2023**, *13*, 395. [CrossRef]