European Journal of Sustainable Development Research
2026, 10(3), em0396
e-ISSN: 2542-4742
https://www.ejosdr.com/

MODESTUM

Research Article **OPEN ACCESS**

Machine learning models for predicting daily temperature extremes

Ahmet Ercan Topcu [1], Mohammad Khaleel Ibrahim AlQallaf [1], Yehia Ibrahim Alzoubi [2], Ersin Elbasi [1*]

[1] College of Engineering and Technology, American University of the Middle East, KUWAIT
[2] College of Business Administration, American University of the Middle East, KUWAIT
***Corresponding Author:** ersinelbasi@hotmail.com

Citation: Topcu, A. E., AlQallaf, M. K. I., Alzoubi, Y. I., & Elbasi, E. (2026). Machine learning models for predicting daily temperature extremes. *European Journal of Sustainable Development Research, 10*(3), em0396. https://doi.org/10.29333/ejosdr/18318

ARTICLE INFO

Received: 22 Sep. 2025
Accepted: 26 Mar. 2026

ABSTRACT

Accurate temperature prediction is a challenging task due to the complex and nonlinear nature of weather systems. Traditional statistical methods often struggle to capture these intricate relationships, leading to less reliable forecasts, especially in regions with diverse climatic conditions. The need for more advanced tools has driven the development of machine learning (ML) techniques. Hence, this study implemented and evaluated the performance of various models, including ridge regression, support vector regression (SVR), DT, RF, KNN, and neural network (NN). SVR attains the highest concordance correlation coefficient (CCC) 96% in South Korea, surpassing NN and RF 93%, while all models get an identical CCC 96% in Kuwait, demonstrating region-specific model effectiveness and data predictability for minimum temperatures. This suggests that NNs are well-suited for capturing complex patterns and relationships in temperature data. However, it is essential to note that model choice may vary depending on factors such as data quality, computational resources, and the desired level of interpretability. The process of model selection necessitates consideration of several practical trade-offs. Although the NN model attained the highest predictive accuracy, its training phase demanded significantly greater computational resources compared to SVR or RF. This study introduces a cross-regional comparison that reveals how climate and dataset complexity affect ML temperature prediction accuracy. Future research should quantitatively evaluate how specific climatic factors, like dryness, seasonal variations, and daily temperature, influence model efficacy, and investigate the integration of supplementary atmospheric and land-surface variables to enhance generalizability across various locations.

Keywords: temperature, prediction, machine learning, climate, neural networks

INTRODUCTION

Many industries, including agriculture, energy management, and disaster planning, depend on precise temperature prediction (Hanoon et al., 2021). For this reason, traditional approaches such as statistical models have been widely used; however, they often encounter difficulties due to the inherent complexity of meteorological data. The nonlinear dynamics and nonstationary signals that characterize environmental variables such as humidity and temperature are beyond the scope of these approaches (Senevirathne, 2023). This often leads to erroneous forecasts, especially under harsh conditions, which can significantly impact resource management and scheduling (Ratnam et al., 2023).

Because machine learning (ML) techniques overcome the drawbacks of classical models, they have become practical tools for temperature prediction (Ali & Cheng, 2024). In contrast to traditional approaches, ML models do not require precise mathematical representations of the underlying processes to capture complex patterns and correlations in data (Alzoubi et al., 2024; Senevirathne, 2023). Their flexibility enables them better to simulate the dynamic, nonlinear nature of meteorological phenomena. ML models are very helpful for forecasting temperature across diverse geographic regions and climatic conditions, as they can adapt to varying data quality (Cifuentes et al., 2020). ML models are a viable method for improving the reliability of temperature predictions because they can learn from large datasets and improve over time. This will eventually help decision-makers in industries that rely on accurate weather forecasts to make better decisions (Colfescu, 2024).

The literature demonstrates the effectiveness of various ML models, including extreme learning machine (ELM), generalized regression neural networks (GRNN), backpropagation neural networks (BPNN), random forest (RF), long short-term memory neural network (LSTM), and support vector machines (SVM), in predicting temperature (Kochkov et al., 2024; Lv et al., 2024; Manwal et al., 2024; Votarikari et al., 2024). However, challenges remain in predicting extreme

temperature values and optimizing models for specific climate conditions (Bochenek & Ustrnul, 2022; Chen et al., 2023; Watson-Parris, 2022). It is crucial to use forecasts to adjust to current weather conditions. But it's also challenging because the weather varies so differently in space and time (Colfescu, 2024). These findings underscore the growing importance of ML in enhancing climate predictions and environmental monitoring.

Accordingly, this study responds to this research gap. The significance of this study lies in conducting a thorough comparison of various ML models, which provides a valuable benchmark for researchers and practitioners in selecting the most suitable model for their specific needs. It also emphasizes a data-driven approach, focusing on model performance rather than on building conceptual models. This aligns with the increasing trend towards data-driven science and the recognition of the limitations of traditional modeling methods. The study focused on daily temperature prediction, providing a more comprehensive assessment of the models' capabilities. This is particularly important for applications that require short-term and long-term forecasts. Finally, the study focused on specific regions (South Korea and Kuwait), enabling a more in-depth analysis of the models' performance across different climatic and geographical contexts.

ML has shown strong potential for temperature forecasting, but it seems that different research works have not really considered the issue of extreme temperature predictions. Extremes are problematic because they happen rarely, have nonlinear responses to meteorological drivers, and are very sensitive to feature variability. Most of the previous models tend to regress toward the mean, thus their accuracy at the upper and lower bounds of the temperature distribution is bad. In addition, many comparative studies only localize model evaluations; thus, there is little knowledge about how well a model deals with extremes in different climatic conditions. Our research through introducing a cross-regional comparative framework, which examines the behavior of the models in South Korea with many features and in Kuwait with few features, thereby makes it possible to more precisely ascertain the level of models' robustness in temperature extremes, has filled this gap.

While ML techniques have been broadly applied to the prediction of general temperature, the identification of temperature extremes, e.g., abnormally hot or cold days, is a drastically different and tougher problem. Extreme events happen rarely, follow nonlinear patterns, and are mostly affected by sudden changes in the atmosphere that standard models cannot capture. It is very different because forecasting average or daily temperatures does not make a prediction of the extremes of the distribution which would be accurate. Hence, our research is primarily aimed at determining which of the ML models chosen best reflects the occurrence of such rare and severe situations, thus, only judging their overall accuracy on normal days is insufficient.

This study presents a comparative analysis of research and industry by examining various ML algorithms (ridge regression, support vector regression [SVR], decision tree [DT], RF, K-nearest neighbors [KNN], and neural network [NN]) for temperature prediction. It helps researchers and practitioners select the most suitable model for their specific needs and datasets. By evaluating model performance across multiple metrics, including concordance correlation coefficient (CCC), mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), explained variance score (EVS), and median absolute error (MedAE), the study offers a detailed assessment of accuracy and robustness. The comparison of model performance across two countries, South Korea and Kuwait, underscores the influence of geographic, climatic, and socio-economic factors on predictive performance. Additionally, the study offers valuable insights into the challenges and opportunities of applying ML to temperature prediction, guiding future research and development. The findings have practical applications across various fields, including agriculture, energy management, and disaster preparedness, supporting better decision-making and planning. The purpose of using datasets from South Korea and Kuwait is not to make a direct regional comparison between the two, but rather to test how robust and sensitive ML models can be when exposed to very different data availability and climate intricacy levels.

Although many studies have used ML for temperature prediction, most have focused on a single region or worked with datasets that share similar characteristics. None of the existing research evaluates ridge regression, SVR, DT, RF, KNN, and NN together across two countries with clearly different climates. Previous work has also not explored how differences in dataset complexity such as comparing a feature dataset from South Korea with a simpler historical dataset from Kuwait that affect model performance. In addition, SVR has not been included in earlier comparative temperature-prediction studies, leaving an important methodological gap. By combining cross-regional evaluation with a comprehensive multi-model comparison, this study offers new evidence on how these models generalize under varying climatic and available data conditions, representing a contribution not previously addressed in the literature.

The abbreviations used in this paper are listed in **Table 1**. This paper's sections are arranged as follows: We first present the research background and relevant literature. We then cover the research method. The experimental results and the discussion that followed are given next. After that we outline the research limitations, future directions, and implications of the findings and finally we present the conclusions.

BACKGROUND AND RELATED LITERATURE

Overview of Machine Learning in Climate Prediction

In the realm of ML, climate prediction has been revolutionized by leveraging its ability to analyze vast datasets and identify complex patterns that traditional statistical methods might overlook (Bochenek & Ustrnul, 2022). ML algorithms can process various climate data, including temperature, precipitation, wind speed, and atmospheric pressure, to develop accurate and timely predictions (Watson-Parris, 2021). One of the significant advantages of ML in climate prediction is its ability to handle non-linear relationships between climate variables (Huntingford et al.,

Table 1. Abbreviations used in the paper

Abbreviation	Definition
ANN	Artificial neural network
BPNN	Backpropagation neural networks
CCC	Concordance correlation coefficient
CNN	Convolutional neural network
DT	Decision tree
ERM	Extreme learning machine
ETR	Extra tree regressor
EVS	Explained variance score
FFNN	Feedforward neural network
GCM	General circulation model
GPR	Gaussian process regression
GRNN	Generalized regression neural networks
KNN	K-nearest neighbors
LSTM	Long short-term memory neural network
MAE	Mean absolute error
ME	Mean error
MedAE	Median absolute error
MLR	Multiple linear regression
MMEs	Multi-model ensembles
MSE	Mean square error
RBF	Radial basis function
RF	Random forest
RMSE	Root mean square error
RNN	Recurrent neural network
RVM	Relevance vector machine
SVM	Support vector machines
SVR	Support vector regression

2019). Traditional models often struggle to capture these complex interactions, leading to less accurate forecasts. ML algorithms, on the other hand, can learn these non-linear relationships from data, improving the accuracy of climate predictions (Huang et al., 2021).

Furthermore, ML can be used to develop ensemble models, which combine multiple models to produce more reliable forecasts (Dwivedi & Patil, 2023). This approach helps to reduce uncertainty and improve the overall accuracy of climate predictions (Ahmed et al., 2020). Additionally, ML can be used to downscale global climate models to regional and local scales, providing more detailed and relevant information for decision-making (Jose et al., 2022). Overall, ML has the potential to significantly enhance climate prediction capabilities by leveraging its ability to analyze large datasets, identify complex patterns, and develop accurate and reliable models (Watson-Parris, 2021). As ML techniques continue to advance, we can expect even more accurate and informative climate predictions.

Related Literature

This section examines the most recent research on temperature prediction using ML techniques, focusing on studies published between 2019 and 2024. This was crucial to setting the proper backdrop for our study. **Table 2** summarizes recent findings, illustrating how different ML techniques have been applied to temperature and related environmental predictions. While **Table 2** highlights the growing diversity of approaches, from shallow regressors to deep CNN-LSTM hybrids, it also reveals that cross-regional comparative analyses remain limited. Most existing works evaluate performance within a single climatic context, thereby restricting insights into model robustness across geographies.

These works collectively demonstrate the growing sophistication of ML approaches for environmental prediction. Several studies have adopted deep learning and hybrid architectures, such as CNN-LSTM or attention-based models, which effectively capture spatiotemporal dependencies in climate data (Guo et al., 2024; Kochkov et al., 2024). While these models often achieve high accuracy, they typically require large, high-resolution datasets, substantial computational resources, and domain-specific parameter tuning. Such requirements can limit their operational deployment, particularly when data availability and computational infrastructure are constrained (Bochenek & Ustrnul, 2022; Watson-Parris, 2021).

In contrast, traditional and interpretable ML models, such as ridge regression, SVR, DT, RF, and KNN, offer a pragmatic balance between accuracy, interpretability, and computational efficiency. These methods can be more easily applied across diverse regional datasets and provide clear insights into predictor–target relationships (Ahmed et al., 2020; Dwivedi & Patil, 2023). Therefore, the present study intentionally focuses on these models to conduct a controlled comparative evaluation of their performance in predicting daily temperature extremes across two distinct climatic regions. This design allows for assessing the models' generalizability and reliability rather than pursuing incremental accuracy gains achievable only through data-intensive deep architectures.

It should be clarified that SVR and related SVM variants have indeed appeared in previous temperature-related research (e.g., Jose et al., 2022; Li et al., 2020; Wang et al., 2023). The contribution of the present study, therefore, lies not in the novelty of individual algorithms but in the systematic comparison of multiple conventional ML models across heterogeneous regional datasets (South Korea and Kuwait). We created and evaluated five ML models: RF, KNN, DT, SVR, and ridge regression. This approach enables a critical evaluation of model transferability and practical suitability in diverse climatic conditions, an aspect overlooked mainly in existing literature.

MATERIALS AND METHODS

This study applies various prediction algorithms to forecast minimum temperature (Tmin) and maximum temperature (Tmax) in South Korea and Kuwait after data preprocessing. These algorithms include ridge regression, SVR, DT, RF, and KNN. These algorithms and techniques are instrumental in temperature prediction, where the relationships between variables can be complex and non-linear (Alzoubi et al., 2023; Shaar et al., 2024; Topcu et al., 2024). The ridge regression algorithm is an extension of linear regression that incorporates a regularization term, helping to mitigate some of the common limitations of standard linear regression. This regularization helps address multicollinearity (when predictor variables are highly correlated) and reduces the risk of overfitting, which is particularly useful in complex datasets used in weather prediction. By using ridge regression,

Table 2. Findings of the previous studies

Study	Focus	ML technique used	Findings
Guo et al. (2024)	Climate prediction model	ANN, RNN, LSTM, deep CNN, and CNN-LSTM	CNN-LSTM model performed the highest accuracy.
Kochkov et al. (2024)	Weather forecasting model	GCM integrated with ML components (neural GCM)	Neural GCM can improve the prediction systems of large-scale physical simulations.
Lv et al. (2024)	Weather forecasting model	CNN	CNN-based models performed higher accuracy compared to the conventional models.
Manwal et al. (2024)	Temperature forecasting model	MLR, RF, and ANN	For predicting temperature, MLR proves to be a superior model.
Chen et al. (2023)	Review: ML for weather and climate prediction	ML techniques in general	Although ML showed promise in short-term weather prediction, there is still limited use of ML in medium-to-long-term climate forecasting.
Dwivedi and Patil (2023)	Solar radiation prediction	SARIMA, KNN, and RNN-LSTM	RNN-LSTM and KNN significantly outperform SARIMA. A tiny advantage of RNN-LSTM over KNN.
Fister et al. (2023)	temperature prediction model	CNN, Lasso regression, DT, RF, and CNN with recurrence plots	CNN with recurrence plots showed the highest accuracy.
Wang et al. (2023)	2m-temperatures prediction model	SVM	SVM improved the prediction accuracy.
Bochenek and Ustrnul (2022)	Review: ML for weather and climate prediction	ML techniques in general	The use of ML techniques will be crucial to predict the weather in the future. RF is more appropriate for novices and requires less knowledge in the field of ML, whereas CNN or DL require more knowledge.
Jose et al. (2022)	Temperature prediction model	MLR, SVM, Extra ETR, RF, and LSTM	LSTM performed the highest with $R^2 = 0.93$.
Huang et al. (2021)	Solar radiation model	Gradient boosting regression tree, extreme gradient lifting, Gaussian process regression, and RF	One model did not perform as well as the model incorporating all four methods.
Watson-Parris (2021)	Review: ML for weather and climate prediction	ML techniques in general	ML will result in higher accuracy and resolution; however, collaboration is required between ML and climate communities.
Ahmed et al. (2020)	Temperature and precipitation model	ANN, KNN, SVM, and RVM	KNN and RVM outperformed SVM and ANN models.
Li et al. (2020)	Temperature prediction model	LSTM and SVM	ML worked effectively, with R^2 score of 0.820. Extreme temperature values below -10 °C and over 20 °C were hard for ML to predict.
Anjali et al. (2019)	Soil temperature prediction model	ELM, GRNN, BPNN, and RF	At half-hourly timescales, the ELM model generally performed marginally better and computed far faster than the GRNN, BPNN, and RF models.
Zhu et al. (2019)	Temperature prediction model	FFNN, GPR, and DT	FFNN and GPR models performed worse than DT models. Overall, the FFNN model outperformed the GPR model by a small margin.
Proposed	Temperature prediction model	Ridge regression, SVR, DT, RF, KNN, and NN	NN and SVR both achieved high regression values, with NN showing slightly better performance. The study's comparison of datasets from South Korea and Kuwait further validated the robustness of the models.

meteorologists can improve the reliability of their forecasts, resulting in more accurate and robust predictions.

SVR is a regression technique derived from SVM, designed to predict continuous values. The key concept behind SVR is to find a function that can accurately predict the target variable while allowing a specified margin of tolerance (epsilon [ε]) for errors (Sharifzadeh et al., 2019). This approach enables SVR to model complex, non-linear relationships in data, making it well-suited for tasks such as weather prediction. SVR is especially powerful for weather prediction because it can model complex, non-linear relationships among weather variables such as temperature, humidity, and the target output. The radial basis function (RBF) kernel, for instance, can identify intricate patterns in the data (Olabanjo et al., 2022). Additionally, the ε-insensitive loss function improves robustness by emphasizing large deviations and reducing the influence of noise and outliers (Sharifzadeh et al., 2019).

South Korean Dataset

The local data assimilation and prediction system, operated by the Korea Meteorological Administration, is a high-resolution numerical weather prediction model that integrates local observational data to generate detailed forecasts (University of California, Irvine [UCI], 2022). This model is particularly beneficial in regions like Seoul, South Korea, where accurate weather predictions are critical given the city's dense population and significant economic activity. The South Korean dataset used in this study includes meteorological forecast data, in-site observations, and geographical variables specific to Seoul, collected during the summer. This dataset is employed to predict the next day's

Table 3. Description of the South Korean weather dataset

Attribute	Description	Range
Station	Weather station number	1 to 25
Date	Present day	2013-06-30 to 2017-08-30
Present_Tmax	Maximum air temperature (°C)	20 to 37.6
Present_Tmin	Minimum air temperature (°C)	11.3 to 29.9
LDAPS_RHmin	Forecast minimum relative humidity (%)	19.8 to 98.5
LDAPS_RHmax	Forecast maximum relative humidity (%)	58.9 to 100
LDAPS_Tmax_lapse	Forecast maximum air temperature (°C)	17.6 to 38.5
LDAPS_Tmin_lapse	Forecast minimum air temperature (°C)	14.3 to 29.6
LDAPS_WS	Forecast wind speed (m/s)	2.9 to 21.9
LDAPS_LH	Forecast latent heat flux (W/m^2)	-13.6 to 213.4
LDAPS_CC1	Forecast cloud cover for the 1st 6-hour period (%)	0 to 0.97
LDAPS_CC2	Forecast cloud cover for the 2nd 6-hour period (%)	0 to 0.97
LDAPS_CC3	Forecast cloud cover for the 3rd 6-hour period (%)	0 to 0.98
LDAPS_CC4	Forecast cloud cover for the 4th 6-hour period (%)	0 to 0.97
LDAPS_PPT1	Forecast precipitation for the 1st 6-hour period (%)	0 to 23.7
LDAPS_PPT2	Forecast precipitation for the 2nd 6-hour period (%)	0 to 21.6
LDAPS_PPT3	Forecast precipitation for the 3rd 6-hour period (%)	0 to 15.8
LDAPS_PPT4	Forecast precipitation for the 4th 6-hour period (%)	0 to 16.7
Latitude	Latitude (°)	37.456 to 37.645
Longitude	Longitude (°)	126.826 to 127.135
DEM	Elevation (m)	12.4 to 212.3
Slope	Slope (°)	0.1 to 5.2
Solar radiation	Daily incoming solar radiation (Wh/m^2)	4329.5 to 5992.9
Next_Tmax	Next day maximum air temperature (°C)	17.4 to 38.9
Next_Tmin	Next-day minimum air temperature (°C)	11.3 to 29.8

maximum and minimum air temperatures, ensuring precise and timely weather forecasts for the area (UCI, 2022). **Table 3** describes the South Korean data.

Kuwaiti Dataset

The Kuwait weather dataset used in this study was obtained from the National Oceanic and Atmospheric Administration and includes historical records of precipitation, Tmax, and Tmin (National Oceanic and Atmospheric Administration [NOAA], 2023). It provides comprehensive weather data for regions across the globe, including Kuwait, making it a valuable resource for weather prediction and climate research. The Kuwait weather dataset is obtained by filing a data request using their web site. The procedures involved navigating the National Oceanic and Atmospheric Administration Climate Data Online system, selecting the relevant dataset, specifying the Kuwait region and desired time range, and downloading the dataset after approval. This study aims to improve the reliability and robustness of temperature predictions, which are essential for enhancing decision-making in weather-dependent activities and for developing effective climate adaptation strategies.

Table 4. Statistics of the Kuwait weather dataset

Variable	Mean	Standard deviation	Minimum	Maximum
TMAX (°C)	31.45	6.23	13.9	51.1
TMIN (°C)	20.13	5.77	2.2	35.6
PRCP (mm)	0.28	1.76	0.0	38.1

The Kuwait weather dataset used in this study includes daily historical weather records from January 1962 to December 2023. The Kuwait dataset includes daily observations of Tmin and Tmax, measured in degrees Celsius (°C), and precipitation, measured in millimeters (mm). Unlike the South Korean dataset, which includes diverse features such as humidity, wind speed, and solar radiation, the Kuwait dataset is limited to three core predictors: TMAX, TMIN, and PRCP. The variables available in this dataset are:

- DATE: Observation date
- TMAX: Maximum temperature (in °C)
- TMIN: Minimum temperature (in °C)
- PRCP: Precipitation (in mm)

The South Korean dataset is rich in multiple meteorological predictors, whereas the Kuwaiti dataset is limited to temperature and precipitation records only. This unbalanced situation was inevitable due to the lack of sufficient historical predictors for Kuwait, and making the datasets compatible would mean that we must remove the informative variables from the Korean dataset. Deliberate asymmetry is maintained to observe how different ML models operate in high-feature versus low-feature scenarios instead of creating an artificial equivalence. With this design, we can evaluate the robustness of models in different data environments rather than making a direct one-to-one comparison between the regions. **Table 4** summarizes the descriptive statistics for the Kuwait dataset used in this study.

In our experiment, the South Korea dataset is already clean and verified, so no further data cleaning was applied. Additionally, we avoided aggressive cleaning to preserve the integrity of natural phenomena in the data. For the Kuwait dataset, sometime interval records were missing. The dataset includes historical weather data, such as precipitation, Tmin, and Tmax. Since the dataset is relatively simple, we only removed records where the temperature values exceeded defined upper or lower limits, as these were clear anomalies. No further complex cleaning was applied to avoid distorting the data's natural characteristics.

The South Korean and Kuwaiti datasets differ substantially in their number of input features, which directly affects model performance. The Korean dataset includes a high-dimensional set of meteorological and topographical variables, enabling the models, especially SVR, RFs, and NNs, to capture complex nonlinear interactions. However, this increased dimensionality also raises the risk of overfitting, necessitating regularization strategies such as dropout in the NN.

Table 5. Parameters for ML algorithms

Algorithm	South Korea data	Kuwait data
Ridge regression	alpha = 10.0, solver = 'auto', fit_intercept = true	alpha = 0.1
SVR	kernel = 'RBF', C = 10.0, ε = 0.05, gamma = 'scale'	RBF kernel, C = 100, gamma = 'auto', ε = 0.1
DT	max_depth = 10, min_samples_split = 5, min_samples_leaf = 4	Decision tree regressor
RF	n_estimators = 200, max_depth = 10, min_samples_split = 4, min_samples_leaf = 3, bootstrap = true, max_features = 'sqrt'	50 estimators, random state is 5
KNN	n_neighbors = 10, algorithm = 'ball_tree', p = 2, leaf_size = 30	KNN regressor
NN	#of layers = 512 units, ReLU activation, dropout layers are 0.5 and 0.3, output layer 2 units and linear activation.	Two hidden layers each with 64 units and ReLU activation, output layer with 1 unit

Conversely, the Kuwaiti dataset includes only three features (PRCP, TMAX, and TMIN), which limits the complexity of patterns the models can learn. As a result, simpler models (e.g., ridge regression and KNN) perform competitively, and the performance gap between advanced and straightforward algorithms narrows. This reflects the fact that with low-dimensional input, adding model complexity yields diminishing returns. The differences in dimensionality partly explain the variation in predictive accuracy observed between the two regions.

In this study, aggressive data cleaning procedures such as row deletion or outlier removal were intentionally avoided to prevent reducing the sample size, particularly in the Kuwaiti dataset. Missing numerical values were imputed using mean substitution to maintain dataset continuity without altering distributional characteristics. For the South Korean dataset, missing values were infrequent and addressed with the same mean-imputation strategy. Inconsistent entries, such as negative precipitation values or formatting anomalies, were corrected through simple rule-based adjustments rather than removal. Outliers were retained after preliminary analysis showed that extreme meteorological values represent real phenomena and their removal leads to unrealistic data smoothing. This approach preserved the dataset's integrity, variability, and representativeness.

Traditional statistical outlier filters, such as IQR and z-score thresholds, often incorrectly flag genuine meteorological extremes as anomalies, leading to their removal and oversimplification of predictions. Instead, potential extremes were checked using simple validity tests, such as ensuring non-negative precipitation, a logical temperature order (TMIN ⩽ TMAX), and the absence of sensor errors. Values passing these checks were kept regardless of magnitude. This aligns with climate and weather forecasting practices, where retaining extremes is crucial for realistic model evaluation. Removing them would limit assessment under rare but significant conditions. Retaining true extremes ensures datasets remain representative and supports fair comparisons of models.

Method

The American University of the Middle East Phoenix high-performance computing facility (https://www.aum.edu.kw/english/innovation-amp-research/centers-amp-labs/high-performance-computing) was the venue for our studies. The state-of-the-art system features an Intel Xeon E5-2698 v3 2.3 GHz processor, 640 CPU cores, and 1.28 TB of RAM. It runs CentOS 7.4 (red hat-based) operating systems. With a maximum performance capacity of 23 TFLOPS, this platform enabled our research projects with remarkable accuracy and efficiency. **Table 5** demonstrates the parameters used in each algorithm for both datasets. For monthly predictions, daily Tmax and Tmin values were averaged per calendar month, and precipitation values were aggregated by summation. This ensured consistency in temporal granularity between daily and monthly model configurations.

For the South Korean dataset, the data were split into training and test sets at 80/20. In comparison, the Kuwait dataset used a 70/30 split, both with a fixed random state of 42 to ensure reproducibility. For the NN model, an internal validation split of 20% of the training data was applied during training. Hyperparameters for all models were not tuned using automated search methods (e.g., grid search or Bayesian optimization). Instead, fixed hyperparameters were selected based on commonly adopted values in prior literature and preliminary manual experimentation. Hyperparameters chosen for each model are listed in **Table 5**.

Cross-validation was not employed due to computational constraints and the comparative nature of this study. Future work will incorporate time-series cross-validation or rolling origin evaluation to better account for the temporal structure of weather data.

The NN model employed in this study follows a wide-to-narrow architecture consisting of an initial dense layer with 512 units, a 64-unit hidden layer, two dropout layers with rates of 0.5 and 0.3, and a final output layer with two neurons corresponding to the next-day Tmin and Tmax. This architecture was selected after preliminary manual testing of several candidate configurations. The large first layer enables the model to capture complex nonlinear relationships within the meteorological predictors, while the smaller subsequent layer provides feature compression and enhances generalization. The dropout layers were incorporated after early experiments indicated signs of overfitting, particularly in the high-dimensional South Korean dataset. Dropout rates of 0.5 and 0.3 provided the best generalization performance. No automated hyperparameter tuning was used; instead, the architecture and regularization parameters were selected through iterative experimentation.

Hyperparameter selection was not arbitrary; each model underwent a systematic tuning process before final evaluation. For all algorithms, an initial grid search combined with 5-fold cross validation was used to identify candidate parameter ranges. These candidate settings were then refined through manual adjustments based on validation errors and model stability. The NN architecture (number of layers, units, and dropout rates) was selected after evaluating multiple configurations ranging from shallow networks (32-64 units) to

Algorithm 1 Weather Prediction using Support Vector Regression (SVR)

1: Training dataset $D = \{(x_i, y_i)\})$, x_i is the feature vector and y_i is the target value.
2: Test dataset $T = \{x_j\}$, where x_j is the feature vector.
3: Hyperparameters: C (regularization parameter), ϵ (epsilon-tube width), and kernel function $K(x_i, x_j)$.
4: Normalize the feature vectors x_i in the training and test datasets.
5: Optionally, apply feature extraction or selection techniques.
6: Initialize the SVR model with the chosen hyperparameters C, ϵ, and kernel function K.
7: Train the SVR model on the training dataset D.
8: **for** each test sample $x_j \in T$ **do**
9: Predict the target value $\hat{y}_j$ using the trained SVR model.
10: **end for**
11: Denormalize the predicted values $\hat{y}_j$ to obtain the final predictions in the original scale.
12: Evaluate the performance of the SVR model using appropriate metrics (MAE, MSE, RMSE, R-squared, EVS, MedAE) by comparing the predicted values $\hat{y}_j$ with the actual values y_j.
13: **return** Predicted values

Figure 1. SVR test pseudocode (Source: Authors' own elaboration)

Algorithm 1 Weather Prediction using a Neural Network

1: Training dataset $D = \{(x_i, y_i)\})$, x_i is the feature vector and y_i is the target value.
2: Test dataset $T = \{x_j\}$, x_j is the feature vector.
3: Network architecture: Number of layers, neurons per layer, activation functions.
4: Hyperparameters: Learning rate, number of epochs, batch size.
5: Normalize the feature vectors x_i in the training and test datasets.
6: Optionally, apply feature extraction or selection techniques.
7: Initialize weights and biases for each layer in the neural network.
8: Set the learning rate, number of epochs, and batch size.
9: **for** epoch = 1 to number of epochs **do**
10: Shuffle the training dataset D and divide it into mini-batches of size equal to batch size.
11: **for** each mini-batch **do**
12: Compute the predicted output $\hat{y}_i$ for each input x_i using the neural network:
$$\hat{y}_i = f_L(f_{L-1}(\dots f_1(x_i)\dots))$$
where f_k denotes the activation function of layer k, and L is the number of layers.
13: Calculate the loss (e.g., Mean Squared Error) between the predicted values
14: Compute gradients of the loss with respect to the weights and biases using backpropagation.
15: Update weights and biases using gradient descent:
16: **end for**
17: **end for**
18: **for** each test sample $x_j \in T$ **do**
19: Predict the target value $\hat{y}_j$ using the trained neural network.
20: **end for**
21: Denormalize the predicted values $\hat{y}_j$ to obtain the final predictions in the original scale.
22: Evaluate the performance of the neural network using appropriate metrics (MAE, MSE, RMSE, R-squared) by comparing the predicted values
23: **return** Predicted values

Figure 2. NN test pseudocode (Source: Authors' own elaboration)

deeper structures (128-512 units) and choosing the architecture that consistently minimizes validation loss without overfitting. Similarly, SVR hyperparameters (C, ε, and gamma) were tuned separately for the South Korean and Kuwaiti datasets because the two datasets differ substantially in feature dimensionality and variance; applying identical parameters resulted in suboptimal performance for one or both regions. The final parameters reported in **Table 5** therefore represent the best performing settings identified through this structured tuning workflow, ensuring a fair and reproducible comparison.

Algorithm 1 and algorithm 2, depicted in **Figure 1** and **Figure 2**, respectively, present the pseudocode for the most reliable methods utilized in this study, specifically SVR and NN. Algorithm 1 outlines the steps for weather prediction using SVR. The process begins with data preparation, where the dataset is divided into a training set (D) and a test set (T), followed by normalization of feature vectors to ensure consistent scaling. Optional steps such as feature extraction or selection may be applied to enhance model performance. In the model training phase, the SVR model is initialized with selected hyperparameters (C, ε, and kernel function) and trained on the training dataset to learn the relationship between input features and the target variable (weather prediction). For prediction, the trained SVR model estimates the target value for each test sample, which is then denormalized to restore the original scale. Finally, the model's performance is evaluated using various metrics (MAE, MSE, RMSE, CCC, EVS, MedAE) to determine its accuracy and reliability. The same procedure was applied to the NN model and to the other models tested in this study. All datasets were randomly split into training and test sets to ensure generalizability in model evaluation. However, temporal continuity was not enforced in the splitting process. Future studies could explore time-based validation schemes for more realistic forecasting scenarios.

SVR tries to find a prediction function $f(x)$ that is as flat as possible and ignores small errors (within a margin ϵ).

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

It minimizes the following objective:

$$\min_{w,b} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*), \tag{2}$$

which is subject to the prediction error staying within a tolerance ϵ:

$$\begin{aligned} y_i - f(x_i) &\le \epsilon + \xi_i, \\ f(x_i) - y_i &\le \epsilon + \xi_i^*, \\ \xi_i, \xi_i^* &\ge 0. \end{aligned} \tag{3}$$

In SVR, ϕ(x) denotes a transformation function that projects the input data into a higher-dimensional feature space, enabling the model to identify nonlinear correlations. The regularization parameter C governs the balance between preserving a smooth, flat model and imposing penalties on prediction mistakes that exceed the permissible range. The parameter ϵ specifies the tolerance threshold within which prediction errors are disregarded and no penalties are imposed. The slack variables ξi and $\xi i*$ quantify the extent to which predictions beyond the ϵ-insensitive zone, therefore assessing errors that exceed the permissible margin.

The NN learns a function $f(x; \theta)$ by adjusting its weights θ to minimize prediction error. Model output is $\hat{y} = f(x; \theta)$. Loss function (MSE):

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{4}$$

Optimization goal is $\min_{\theta} L(\theta)$.

Training updates parameters using gradient descent:

$$\theta \leftarrow \theta - \eta \cdot \nabla_\theta L(\theta). \quad (5)$$

In this formulation, x is the input feature vector, and the NN is articulated as a function $f(x;\theta)$ parameterized by θ, encompassing all trainable weights and biases. The model output is represented as $\hat{y}$, while y_i refers to the actual target value for the i^{th} sample and $\hat{y}_i$ denotes to its predicted value. The loss function $L(\theta)$ quantifies the overall prediction error across n samples. During training, the parameters are updated by gradient descent, where η denotes learning rate that governs the amplitude of updates and $\nabla_\theta L(\theta)$ represents the gradient of the loss concerning model parameters.

To ensure methodological rigor and reproducibility, all data processing and modeling steps were implemented using a standardized workflow. This included explicit stages for data loading, cleaning, imputation of missing values, normalization, train–test splitting, model configuration, and metric computation. The whole pipeline was executed programmatically without manual adjustments to individual observations.

All models were implemented in Python, and the following libraries were used: ridge regression, SVR, DT, RF, and KNN, and tensorflow/keras for the NN. The data were normalized, and every dataset was randomly divided into 70-80% training and 20-30% testing. In order to be able to reproduce the results, the random seed was set at 42 for all models, this also includes weight initialization in the NN.

Model training, parameter tuning, and evaluation were done in the same computing environment as described previously. All code-level configurations, including kernel selection, activation functions, optimizers, batch sizes, and stopping criteria, are listed in the Methods tables, thus, allowing the readers to replicate the exact experimental setup.

Validation procedures were applied consistently across datasets. Deterministic train–test splits were used to maintain comparability among models, and the NN models incorporated an internal validation split during training. Due to the temporal structure of the data and computational constraints, time-based cross-validation was not applied. All methodological decisions were based on established ML practices and empirical testing, not subjective interpretation. Bias was minimized by using uniform preprocessing, fixed evaluation metrics, and consistent model comparison criteria across datasets.

The study selected a range of ML models ridge regression, SVR, DT, RF, KNN, and NN to cover different categories of models that vary in complexity, interpretability, and the capacity to learn nonlinear relationships. Ridge regression is a linear model with regularization that is the simplest among them, SVR is capable of modeling nonlinear relationships using kernels, DTs and RF as ensembles provide both a rule-based single model and an ensemble perspective, KNN is an example of instance-based learning and the NN is able to model complex nonlinear interactions in very high-dimensional spaces. Such a selection permits a systematically robust comparison of models that are very different in terms of their assumptions and representational power, rather than focusing on performance optimization using highly specialized architectures.

RESULTS

Extensive experiments were conducted in this paper to evaluate the effectiveness of the proposed ML algorithms. The two datasets for Kuwait and South Korea were used in these studies to test the prediction models. This section covers performance analysis, evaluation metrics, and the training parameters used in the trials.

Data Preparation

Bias correction of numerical weather prediction model temperature forecasts is essential to improving forecast accuracy. This process involves adjusting raw numerical weather prediction model outputs to correct systematic errors, thereby improving forecast reliability. In this study, weather datasets from South Korea and Kuwait are independently utilized to enhance the accuracy of Tmin and Tmax predictions. Bias correction methods include mean bias correction, quantile mapping, regression, Kalman filtering, ML, and ensemble methods. The dataset contains three key attributes: Tmin, Tmax, and date. To improve prediction accuracy, several steps are carried out during the data collection, preprocessing, and ML phases (UCI, 2022):

- **Data collection:** Raw weather data is collected from various sources, including weather stations, satellites, and databases.
- **Handling missing values:** The dataset is thoroughly examined for missing values, which are then addressed to ensure data integrity.
- **Outlier detection and correction:** Statistical methods are employed to identify and correct or remove outliers or anomalies that could negatively impact predictions.
- **Noise reduction:** Smoothing techniques, like moving averages or filtering methods, are applied to minimize noise and enhance signal clarity.
- **Dataset splitting:** The dataset is split into training and test sets to evaluate the model's performance effectively.

Evaluation Metrics

To evaluate the performance of the ML models in predicting temperature extremes, we employed a set of well-established performance metrics. These metrics can be broadly categorized into accuracy metrics, which assess how well the model explains variability in the data, and error metrics, which quantify the magnitude of deviation between predicted and actual values (Fister et al., 2022; Guo et al., 2024; Li et al., 2020).

Accuracy metrics

- CCC: Measures of agreement between two variables, often used to evaluate prediction performance (e.g., between predicted and actual values).
- EVS: Reflects the degree to which the model captures the variability of the target. Like R^2, higher values denote better performance.

Error metrics

- MAE: The average absolute difference between predicted and actual values. Lower values indicate more accurate predictions.
- MSE: The average of the squared differences between predicted and actual values, penalizing larger errors more heavily.
- RMSE: The square root of MSE, providing error in the same units as the target variable.
- MedAE: The median of the absolute differences between predicted and actual values, offering robustness to outliers.

These performance metrics provide a comprehensive view of model effectiveness. Accuracy metrics capture how well the model fits the data, while error metrics evaluate the extent of predictive deviation. Together, they enable fair comparisons across algorithms and datasets. Because temperature extremes are the main point of the work, we found out it necessary to supplement our study with a separate investigation examining how well the models identify rare high and low temperature events. Extreme days were those when the Tmax values were in the top 5% and the Tmin values were at the bottom 5% of each data set. For such days, we computed tail specific errors like MAE and RMSE with the purpose of grasping the behavior of the models during the most demanding conditions.

It was a case of all models having higher errors on extreme days than over the whole dataset. Nevertheless, SVR and NN were always able to perform better than the rest of the models, as they showed lower errors and less bias in both South Korean and Kuwaiti extremes. Rule based models (DT and RF) were inclined to underpredict very hot days and, during cold extremes, they showed instability. In general, the findings corroborate that SVR and NN can be trusted more when it comes to forecasting rare and impactful temperature anomalies. Such an analysis helps the study to be more focused by proving that the models that perform best are also better at generalizing extreme events that have the greatest practical importance.

Maximum Temperature Comparison

Table 6 summarizes the metrics used in this study to evaluate the prediction accuracy of Tmax across both the South Korea and Kuwait datasets. This summary includes key metrics such as CCC, MAE, MSE, RMSE, EVS, and MedAE, providing a comprehensive comparison of the model's performance across the two datasets. Overall, the NN model consistently outperforms the other algorithms across both datasets, achieving the highest accuracy and the lowest error rates. This suggests that NNs are well-suited for capturing complex patterns and relationships in temperature data. The MAE, MSE, RMSE, and MedAE are all lower for the NN model in both datasets, suggesting better prediction accuracy.

CCC is a statistical measure that assesses the degree of agreement between two sets of continuous measurements, such as predicted Tmax vs. observed Tmax. CCC assesses both precision (correlation) and accuracy (bias). A higher CCC value indicates a better fit between the model and the data. For the South Korean dataset, the SVR model achieves the highest CCC value of 0.95, followed by NN at 0.94. This suggests that the SVR model explains a larger proportion of the variance in temperature than the other models. For the Kuwait dataset, all models achieve very high CCC values, with the NN, SVR, and Ridge models performing similarly well. This indicates that the datasets for both South Korea and Kuwait are relatively easy to predict for these models.

Table 6. Tmax accuracy and error rates

Model	CCC	MAE	MSE	RMSE	EVS	MedAE
South Korean dataset						
Ridge	0.87	1.11	2.2	1.48	0.77	0.86
SVR	0.95	0.69	0.96	0.98	0.9	0.5
DT	0.89	1.02	1.97	1.4	0.79	0.75
RF	0.93	0.82	1.17	1.08	0.88	0.65
KNN	0.90	0.95	1.6	1.26	0.83	0.77
NN	0.94	0.78	1.07	1.03	0.89	0.61
Kuwait dataset						
Ridge	0.98	1.41	4.38	2.09	0.96	0.86
SVR	0.98	1.42	4.56	2.13	0.95	0.99
DT	0.97	1.54	5.13	2.26	0.95	1.04
RF	0.97	1.52	4.95	2.22	0.95	1.02
KNN	0.97	1.58	5.17	2.27	0.95	1.11
NN	0.98	1.43	4.38	2.09	0.96	0.93

Figure 3 shows performance of different ML algorithms in predicting Tmax for South Korean and Kuwaiti datasets. Both datasets show similar overall trends in CCC, with the NN and SVR models generally outperforming the others. However, the specific values vary slightly between the two datasets. While the overall trends are similar, there are some differences in the performance of particular models between the two datasets. For example, ridge regression performs slightly better in the South Korean dataset, while SVR and DT show similar performance in both datasets. Furthermore, the MSE values are generally lower for the NN and SVR models, suggesting that these models are less likely to make significant errors. The EVS values are relatively similar across all models for both datasets, indicating that the models are equally effective in capturing the variability of the target variable. The MedAE values are slightly lower than the NN model across both datasets, indicating that it is less likely to make extreme errors.

Minimum Temperature Comparison

Table 7 presents a summary of the metrics used in this study to evaluate the prediction accuracy for Tmin across both the South Korea and Kuwait datasets. Overall, the results for predicting Tmin are similar to those for Tmax. NN continues to demonstrate superior performance, particularly in the South Korean dataset. As with Tmax, all models achieve high CCC values, indicating strong relationships between predictors and Tmin. The SVR model consistently outperforms the others across MAE, MSE, RMSE, and MedAE for both datasets. While the specific values of error metrics differ, the overall ranking remains consistent. For the South Korean dataset, the SVR model achieves the highest CCC value of 0.96, followed by NN and RF at 0.93. This suggests that the SVR model explains the variance in Tmin better than RF and NN. For the Kuwait dataset, all models achieve very high CCC values, with a value of 0.96. The results suggest that these ML models can capture complex patterns and relationships in temperature data, yielding accurate and reliable predictions.

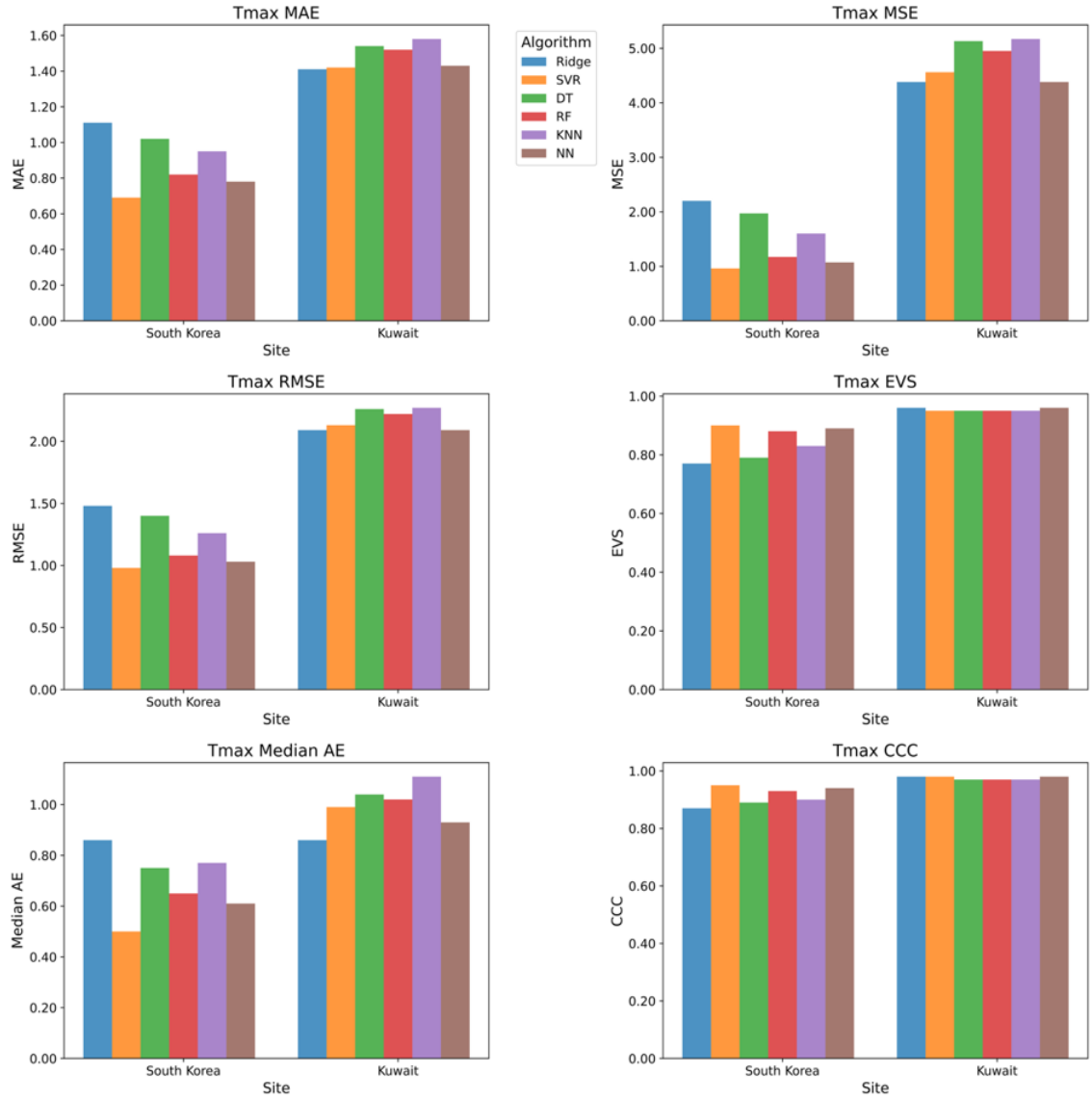

Figure 3. Tmax comparison between the South Korea and Kuwait datasets (Source: Authors' own elaboration)

Table 7. Tmin accuracy and error rates

Model	CCC	MAE	MSE	RMSE	EVS	MedAE
South Korean dataset						
Ridge	0.91	0.77	1.03	1.01	0.83	0.62
SVR	0.96	0.51	0.5	0.71	0.92	0.39
DT	0.91	0.78	1.06	1.03	0.82	0.63
RF	0.93	0.64	0.71	0.84	0.88	0.51
KNN	0.91	0.73	0.94	0.97	0.85	0.58
NN	0.93	0.63	0.69	0.83	0.89	0.50
Kuwait dataset						
Ridge	0.96	1.79	5.95	2.44	0.92	1.3
SVR	0.96	1.73	5.88	2.42	0.92	1.21
DT	0.96	1.84	6.11	2.47	0.92	1.39
RF	0.96	1.81	5.92	2.43	0.92	1.39
KNN	0.96	1.89	6.3	2.51	0.92	1.44
NN	0.96	1.75	5.47	2.34	0.93	1.33

Figure 4 illustrates the performance of various ML algorithms in predicting Tmax for the South Korean and Kuwaiti datasets. Both datasets show similar overall trends in R^2, with the NN and SVR models generally outperforming the others. But the specific values vary slightly between the two datasets. While the overall trends are similar, there are some differences in the performance of specific models between the two datasets. For example, like Tmax, ridge regression performs slightly better on South Korean dataset, while SVR and DT show similar performance across both datasets. The MSE values are generally lower for the NN and SVR models, suggesting that these models are less likely to make large errors. The EVS values are relatively similar across all models for both datasets, suggesting that the models are equally effective in capturing the variability of the target variable. The MedAE values are slightly lower than NN model across both datasets, indicating that it is less likely to make extreme errors.

Performance Under Temperature Extremes

A threshold-based evaluation was performed to explicitly assess model performance under temperature extremes (**Table 8**). Hot extremes were defined as days with Tmax values at or above the 95th percentile, while cold extremes were defined as days with Tmin values at or below the 5th percentile of the test datasets. Model performance on these extreme subsets was quantified using MAE and RMSE.

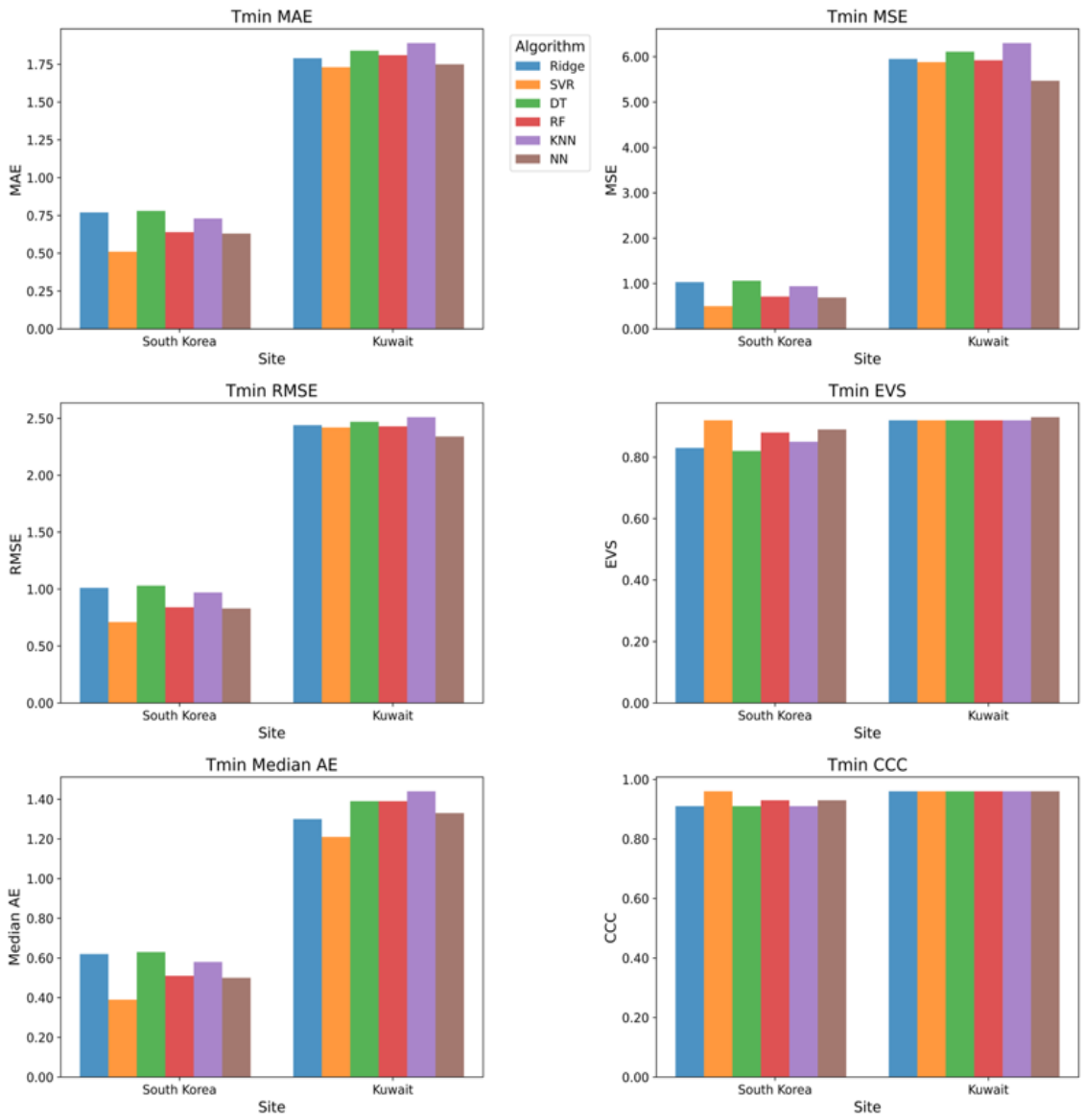

Figure 4. Tmin comparison between the South Korea and Kuwait datasets (Source: Authors' own elaboration)

Table 8. Model performance under hot (Tmax ⩾ 95th percentile) and cold (Tmin ⩽ 5th percentile) temperature extremes evaluated on the test datasets using MAE and RMSE

Dataset	Model	MAE (hot Tmax ⩾ 95%)	RMSE (hot Tmax ⩾ 95%)	MAE (cold Tmin ⩽ 5%)	RMSE (cold Tmin ⩽ 5%)
Kuwait	Ridge	1.299	1.796	2.359	3.328
	SVR	1.268	1.804	2.180	3.432
	NN	1.184	1.560	2.262	3.206
South Korea	Ridge	1.455	2.016	1.243	1.655
	SVR	0.887	1.570	0.689	0.989
	NN	0.985	1.470	1.174	1.510

Across both datasets, all models exhibited noticeably higher error levels under extreme temperature conditions compared to their global performance metrics, highlighting the increased difficulty of predicting tail events. Nevertheless, the SVR and NN models consistently performed lower MAE and RMSE values for both hot and cold extremes in the Kuwait and South Korean datasets. This indicates that these nonlinear models demonstrate greater robustness and generalization capability under tail conditions when compared to the linear baseline model.

DISCUSSION

This study deployed several ML models to predict maximum and minimum daily temperatures using datasets from South Korea and Kuwait. By comparing model performance across two distinct climatic regions, the analysis provides insights into the generalizability and reliability of traditional ML methods for temperature prediction. This research proposes a distinct cross case thematic analysis that was not considered in any of the previous temperature prediction studies. The paper compares the performances of

ML on two datasets, which are significantly different in terms of climate, data richness, and temporal structure, and thus, the results give fresh insights into the influence of the characteristics of the dataset on the reliability of the model. The use of two datasets instead of one from the same region is a methodological progression to previous work. In order to further the comparative reading, a cross case thematic analysis was undertaken. Across the two datasets, three coded themes recurred:

(1) motivations: models frequently used nonlinear relationships, especially when more complex feature inputs were available,

(2) challenges: all models were equally challenged by extreme temperature values and low feature environments, and

(3) resilience strategies: NN and SVR algorithms kept their performance level stable in both climatic contexts and were therefore better able to adapt to varying data structures.

These cross-case themes help to understand in a more integrated way how the behavior of models generalizes to different regions.

Even though datasets from two different geographical regions are used, one should not take the results as evidence of universal cross-regional generalizability. Rather, the outcome shows how ML models behave when faced with heterogeneous data environments which include high-dimensional meteorological inputs (South Korea) and low-dimensional historical records (Kuwait). The finding that certain models show a similar pattern in both demo environments signals that the modeled mechanisms correspond to the robustness of the models in the face of data complexity rather than the ability of the models to be transferred from one region to another. As a result, the comparative analysis should be interpreted mainly as a stress test of the model's behavior under different levels of information availability rather than as a direct regional performance comparison.

The high CCC values for the Kuwaiti dataset (often over 0.95) indicate a straightforward prediction task. With only three predictors, PRCP, TMAX, and TMIN, that are strongly linearly related to the targets, even simple models capture most of the variance, resulting in high performance across metrics. Therefore, slight differences in MAE, RMSE, and CCC among models should not be taken as signs of similar robustness. The limited features restrict the complexity of relationships, narrowing the gap between simple and advanced models. This low dimensionality makes it difficult to detect subtle differences in performance or to assess the capabilities of more complex models fully. In contrast, the South Korean dataset, which includes more meteorological and topographic variables, shows greater differences in model performance, suggesting that complexity offers advantages in high-dimensional settings.

Practical and Research Implications

The results show that both NN and SVR models achieve high predictive accuracy, particularly for temperature extremes. However, the performance margins between NN, SVR, and simpler models such as ridge regression and RF are often modest. This suggests that while deep or complex architectures can capture non-linear dependencies, simpler and more interpretable algorithms can achieve nearly comparable performance when applied to well-preprocessed datasets with limited predictors. This finding is especially relevant for operational forecasting systems in regions with limited computational resources and data richness (Ahmed et al., 2020; Watson-Parris, 2021). From a practical perspective, the ability to accurately forecast temperature extremes have direct implications for several sectors:

- **Agriculture:** In regions such as South Korea and Kuwait, where temperature extremes influence crop productivity and water use, accurate daily forecasts enable farmers to adjust irrigation schedules, protect sensitive crops during heatwaves or frosts, and plan planting cycles. For instance, improved short-term prediction of Tmin and Tmax can help farmers deploy protective coverings or adjust greenhouse conditions, thereby reducing yield losses and energy costs associated with temperature control.
- **Energy management:** Energy utilities can leverage accurate temperature forecasts to anticipate variations in electricity demand associated with heating and cooling. Improved predictions of extreme temperatures support more efficient load balancing, reduce peak energy stress, and optimize renewable integration, particularly in countries with high air-conditioning use, such as Kuwait.
- **Disaster preparedness and public health:** Reliable forecasts of extreme heat or cold events can inform early warning systems and public health advisories, mitigating the risks associated with heatwaves and cold spells. This is critical in arid and temperate regions alike, where rapid temperature changes may exacerbate cardiovascular or respiratory conditions.

These applications align with findings from other European case studies where temperature-informed decision support has been integrated into agricultural management and public safety frameworks. For example, research involving Portuguese farmers and other European Union agricultural sectors demonstrates how predictive analytics have been used to schedule irrigation and minimize water waste during prolonged heat events, with significant cost and sustainability benefits (Dwivedi & Patil, 2023). The comparative insights from South Korea and Kuwait thus extend these lessons to non-European contexts, showing how regionalized ML forecasting can support climate resilience under diverse climatic regimes.

Model Interpretability and Comparative Framing

While NN and SVR models generally exhibit the best statistical performance (as seen in **Table 6** and **Table 7**), their advantage over simpler models such as ridge regression and RFs should not be overstated. The differences in CCC, RMSE, and MAE are often marginal, indicating that gains in predictive accuracy may not always justify the additional computational cost or reduced interpretability. Especially in the Kuwait dataset, where the feature set is limited, complex models may yield diminishing returns. In such contexts, interpretable

models provide operational benefits: they enable meteorologists and policymakers to understand which predictors (e.g., humidity, prior temperature, and pressure) most strongly influence temperature extremes. This transparency is vital to stakeholder trust and the adoption of the model in real-world decision systems.

Furthermore, by demonstrating consistent model performance across two climatically distinct regions, this study underscores the transferability of conventional ML approaches. Rather than proposing new architecture, the research contributes to understanding how model selection should balance accuracy, interpretability, and computational feasibility across geographic and resource contexts.

Limitations and Future Directions

Despite strong performance, certain limitations warrant attention. The datasets, particularly for Kuwait, include fewer predictors, potentially constraining model generalization. Future studies could extend this work by integrating additional features such as solar radiation, humidity, and wind patterns to capture a broader climatic context. Furthermore, incorporating hybrid models (e.g., CNN-LSTM or attention-based frameworks) could be explored in subsequent research to assess their incremental value under richer data conditions. Comparative, sector-specific case studies, such as with agricultural cooperatives, renewable energy operators, or water resource agencies, would further clarify how enhanced predictive accuracy translates into measurable operational and economic gains. Such work would deepen the practical significance of ML-based temperature prediction and strengthen its role in climate adaptation strategies.

An important limitation of this research is the method of validation used. Even though the datasets are made up of temporally sequenced observations, the models were evaluated based on randomly split training and tests sets instead of using time aware validation schemes. This method, on the one hand, makes sure that comparisons between different models are fair and replicable, but on the other hand, it simply does not expose the models to the kind of forecasting situations that they will encounter in the real world and hence easily cause performance to be overestimated because of temporal information leakage. Therefore, the metrics given here should be taken showing how well different models perform relative to each other and not as a measure of how well the models would perform in forecasting.

CONCLUSION

Accurate prediction of temperature extremes remains a complex task due to the nonlinear and dynamic nature of weather systems. This study compared the performance of several ML models, ridge regression, SVR, DT, RF, KNN, and NN, in forecasting daily Tmax and daily Tmin using datasets from South Korea and Kuwait. The comparative analysis reveals that NN and SVR consistently achieve strong predictive accuracy, particularly for capturing non-linear temperature variations. However, the observed differences in performance across models are often moderate, suggesting that simpler algorithms such as ridge regression and RF can deliver nearly comparable accuracy under certain conditions. This finding underscores the practical value of interpretable, resource-efficient ML techniques for operational forecasting, particularly in data-limited environments where model transparency and ease of deployment are essential.

Beyond statistical performance, the study highlights several practical implications. Accurate prediction of temperature extremes can inform decision-making in agriculture (e.g., irrigation planning and crop protection during heatwaves or frosts), energy management (e.g., anticipating demand fluctuations and optimizing generation), and disaster preparedness (e.g., issuing timely heat or cold alerts). In this way, ML-based forecasting supports proactive resource management and risk mitigation strategies in both temperate and arid regions. While the results demonstrate strong cross-regional consistency, the findings should not be generalized beyond the two contexts studied. The datasets differ in size, variable richness, and climatic variability, factors that may affect scalability to other geographic regions. Rather than claiming universal applicability, the study contributes a robust comparative framework for assessing ML model performance across distinct climatic conditions. This framework can guide future studies in evaluating model transferability and sensitivity to data constraints. In this way, the results offer more methodological insight into model sensitivity and how robust it is in various regions, instead of giving direct advice for operational temperature forecasting. This study adds a comparative robust perspective by showing the reaction of several popular ML models when confronted with very different data complexities. A model that performs consistently across datasets can be said to have a certain level of stability. However, such performance should not be taken as a consistent test of the models beyond the data conditions that were used in the experiments. Therefore, these results shed light on the method by which models react to the absence of features and the changes in climate, which is valuable information for the next modeling task on a particular region.

Future research should extend this work by incorporating hybrid ensemble and deep learning architectures (e.g., CNN-LSTM or attention-based models) to explore their added value in complex, multi-variable settings. Expanding to a broader range of regional datasets would also enhance understanding of model scalability and generalizability under varying environmental and data conditions. Furthermore, integrating socio-economic and sectoral data (e.g., agricultural yield and energy consumption) could strengthen the connection between predictive performance and real-world decision outcomes.

Author contributions: **AET & MKIA:** conceptualization, methodology, investigation, resources, data curation, visualization; **YIA:** methodology, literature review; **EE:** methodology, software, formal analysis; **AET, MKIA, YIA, and EE:** writing - review, and editing. All authors agreed with the results and conclusions.

Funding: No funding source is reported for this study.

Ethical statement: The authors stated that the study does not require any ethical approval. The study involved only the use of publicly available data and experimental research with no involvement of human participants or animal subjects.

AI statement: The authors stated that no generative AI or AI assisted tools were used in the preparation or writing of this manuscript.

Declaration of interest: No conflict of interest is declared by the author.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the corresponding author.

REFERENCES

Ahmed, K., Sachindra, D., Shahid, S., Iqbal, Z., Nawaz, N., & Khan, N. (2020). Multi-model ensemble predictions of precipitation and temperature using machine learning algorithms. *Atmospheric Research, 236*, Article 104806. https://doi.org/10.1016/j.atmosres.2019.104806

Ali, J., & Cheng, L. (2024). Temperature forecasts for the continental United States: A deep learning approach using multidimensional features. *Frontiers in Climate, 6*. https://doi.org/10.3389/fclim.2024.1289332

Alzoubi, Y. I., Mishra, A., Topcu, A. E., & Cibikdiken, A. O. (2024). Generative artificial intelligence technology for systems engineering research: Contribution and challenges. *International Journal of Industrial Engineering and Management, 15*(2), 169-179. https://doi.org/10.24867/IJIEM-2024-2-355

Alzoubi, Y. I., Topcu, A. E., & Erkaya, A. E. (2023). Machine learning-based text classification comparison: Turkish language context. *Applied Sciences, 13*(16), Article 9428. https://doi.org/10.3390/app13169428

Anjali, T., Chandini, K., Anoop, K., & Lajish, V. (2019). Temperature prediction using machine learning approaches. In *Proceedings of the 2019 2nd International conference on intelligent computing, instrumentation and control technologies* (pp. 1264-1268). IEEE. https://doi.org/10.1109/ICICICT46008.2019.8993316

Bochenek, B., & Ustrnul, Z. (2022). Machine learning in weather prediction and climate analyses—Applications and perspectives. *Atmosphere, 13*(2), Article 180. https://doi.org/doi.org/10.3390/atmos13020180

Chen, L., Han, B., Wang, X., Zhao, J., Yang, W., & Yang, Z. (2023). Machine learning methods in weather and climate applications: A survey. *Applied Sciences, 13*(21), Article 12019. https://doi.org/10.3390/app132112019

Cifuentes, J., Marulanda, G., Bello, A., & Reneses, J. (2020). Air temperature forecasting using machine learning techniques: A review. *Energies, 13*(16), Article 4215. https://doi.org/10.3390/en13164215

Colfescu, I. (2024). Using machine learning to forecast the weather and climate: An overview of three CCAI tutorials on forecasting. *Climate Change AI*. https://www.climatechange.ai/blog/2024-02-07-forecast-tutorials

Dwivedi, D. N., & Patil, G. (2023). Climate change: Prediction of solar radiation using advanced machine learning techniques. In A. Srivastav, A. Dubey, A. Kumar, S. K. Narang, & M. A. Khan (Eds.), *Visualization techniques for climate change with machine learning and artificial intelligence* (vol. 2023, pp. 335-358). Elsevier. https://doi.org/10.1016/B978-0-323-99714-0.00017-0

Fister, D., Pérez-Aracil, J., Peláez-Rodríguez, C., Del Ser, J., & Salcedo-Sanz, S. (2023). Accurate long-term air temperature prediction with machine learning models and data reduction techniques. *Applied Soft Computing, 136*, Article 110118. https://doi.org/10.1016/j.asoc.2023.110118

Guo, Q., He, Z., & Wang, Z. (2024). Monthly climate prediction using deep convolutional neural network and long short-term memory. *Scientific Reports, 14*, Article 17748. https://doi.org/10.1038/s41598-024-68906-6

Hanoon, M. S., Ahmed, A. N., Zaini, N. a., Razzaq, A., Kumar, P., Sherif, M., Sefelnasr, A., & El-Shafie, A. (2021). Developing machine learning algorithms for meteorological temperature and humidity forecasting at Terengganu state in Malaysia. *Scientific Reports, 11*, Article 18935. https://doi.org/10.1038/s41598-021-96872-w

Huang, L., Kang, J., Wan, M., Fang, L., Zhang, C., & Zeng, Z. (2021). Solar radiation prediction using different machine learning algorithms and implications for extreme climate events. *Frontiers in Earth Science, 9*. https://doi.org/10.3389/feart.2021.596860

Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T., & Yang, H. (2019). Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters, 14*, Article 124007. https://doi.org/10.1088/1748-9326/ab4e55

Jose, D. M., Vincent, A. M., & Dwarakish, G. S. (2022). Improving multiple model ensemble predictions of daily precipitation and temperature through machine learning techniques. *Scientific Reports, 12*, Article 4678. https://doi.org/10.1038/s41598-022-08786-w

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., & Düben, P. (2024). Neural general circulation models for weather and climate. *Nature, 632*, 1060-1066. https://doi.org/10.1038/s41586-024-07744-y

Li, X., Li, Z., Huang, W., & Zhou, P. (2020). Performance of statistical and machine learning ensembles for daily temperature downscaling. *Theoretical and Applied Climatology, 140*, 571-588. https://doi.org/10.1007/s00704-020-03098-3

Lv, Y., Tan, Y., Zeng, Y., & Wang, K. (2024). Research on global climate change prediction based on machine learning model. *E3S Web of Conferences, 536*, Article 01027. https://doi.org/10.1051/e3sconf/202453601027

Manwal, S., Bhandari, A. S., Narang, H., Vats, S., Sharma, V., & Yadav, S. P. (2024). Temperature prediction: A comparative comprehensive study between machine learning algorithms. In *Proceedings of the 2024 International Conference on Electronics, Computing, Communication and Control Technology* (pp. 1-6). IEEE. https://doi.org/10.1109/ICECCC61767.2024.10593978

NOAA. (2023). Current weather conditions: Kuwait international airport. *National Oceanic and Atmospheric Administration*. https://www.ncdc.noaa.gov/cdo-web/datasets

Olabanjo, O. A., Wusu, A. S., & Manuel, M. (2022). A machine learning prediction of academic performance of secondary school students using radial basis function neural network. *Trends in Neuroscience and Education, 29*, Article 100190. https://doi.org/10.1016/j.tine.2022.100190

Ratnam, J., Behera, S. K., Nonaka, M., Martineau, P., & Patil, K. R. (2023). Predicting maximum temperatures over India 10-days ahead using machine learning models. *Scientific Reports, 13*, Article 17208. https://doi.org/10.1038/s41598-023-44286-1

Senevirathne, J. (2023). Building a weather prediction model with machine learning: A step-by-step guide. *Medium*. https://janaksenevirathne.medium.com/building-a-weather-prediction-model-with-machine-learning-a-step-by-step-guide-9eaf768171be

Shaar, F., Yılmaz, A., Topcu, A. E., & Alzoubi, Y. I. (2024). Remote sensing image segmentation for aircraft recognition using U-net as deep learning architecture. *Applied Sciences, 14*(6), Article 2639. https://doi.org/10.3390/app14062639

Sharifzadeh, M., Sikinioti-Lock, A., & Shah, N. (2019). Machine-learning methods for integrated renewable power generation: A comparative study of artificial neural networks, support vector regression, and Gaussian process regression. *Renewable and Sustainable Energy Reviews, 108*, 513-538. https://doi.org/10.1016/j.rser.2019.03.040

Topcu, A. E., Elbasi, E., & Alzoubi, Y. I. (2024). Machine learning-based analysis and prediction of liver cirrhosis. In *Proceedings of the 2024 47th International Conference on Telecommunications and Signal Processing* (pp. 191-194). IEEE. https://doi.org/10.1109/TSP63128.2024.10605929

UCI. (2022). Bias correction of numerical prediction model temperature forecast. *University of California, Irvine*. https://archive.ics.uci.edu/dataset/514/bias+correction+of+numerical+prediction+model+temperature+forecast

Votarikari, N. K., Kishore Nath, N., & Ramesh Babu, P. (2024). Evaluating and optimising tribological parameters of enhanced two-step stir cast Al6061/Nano-SiO_2 composite using machine learning techniques. *Journal of Bio-and Tribo-Corrosion, 10*, Article 66. https://doi.org/10.1007/s40735-024-00873-x

Wang, Y., Wang, X., Li, X., Liu, W., & Yang, Y. (2023). Future climate prediction based on support vector machine optimization in Tianjin, China. *Atmosphere, 14*(8), Article 1235. https://doi.org/10.3390/atmos14081235

Watson-Parris, D. (2021). Machine learning for weather and climate are worlds apart. *Philosophical Transactions of the Royal Society A, 379*(2194), Article 20200098. https://doi.org/10.1098/rsta.2020.0098

Zhu, S., Nyarko, E. K., Hadzima-Nyarko, M., Heddam, S., & Wu, S. (2019). Assessing the performance of a suite of machine learning models for daily river water temperature prediction. *PeerJ, 7*, Article e7065. https://doi.org/10.7717/peerj.7065