engineering proceedings

MDPI

*Proceedings Paper*

# Comparing Regression Techniques for Temperature downscaling in Different Climate Classifications†

**Ali Ilghami khosroshahi[1], Mohammad Bejani[1], Hadi Pourali[1] and Arman hosseinpour salehi[1]**

[1] Faculty of Civil Engineering, University of Tabriz, Tabriz, Iran aliikhosroshahi@gmail.com , m.bejani99@ms.tabrizu.ac.ir , h.pourali1400@ms.tabrizu.ac.ir and armanhsalehi@gmail.com

† Presented at the 4th International Electronic Conference on Applied Sciences, online conference 27 Oct–10 Nov 2023.

**Abstract:** This study aims to identify the optimum regression techniques for downscaling among ten commonly used methods in climatology, including SVR, LinearSVR, LASSO, LASSOCV, Elastic Net, Bayesian Ridge, RandomForestRegressor, AdaBoost Regressor, KNeighbors Regressor, and XGBRegressor. for Köppen climate classification system, including A (tropical), B (dry), C (temperate), and D (continental) a synoptic station data had been collected furthermore for downscaling purpose General Circulation Model (GCM) had been utilized. Additionally, to enhance the performance of downscaling accuracy, Mutual Information (MI) as feature selection was employed. The downscaling performance was evaluated using the Coefficient of Determination (DC) and Root Mean Squared Error (RMSE). results indicate that SVR had superior performance in tropical and dry climates. and, LassoCV with RandomForestRegressor had better results in temperate and continental climates.

**Keywords:** Downscaling; Regression techniques; Köppen climate classification; Mutual Information (MI)

## 1. Introduction

In recent years, downscaling techniques have emerged as practical methods in numerous fields, including climatology trend simulation [1, 2]. Therefore, identifying the optimal regression technique is critical for assessing, simulating, and predicting climate patterns. General Circulation Models (GCMs) play a crucial role as indispensable tools in the investigation of climate change and its associated consequences. Downscaling methods are crucial in improving the effectiveness of GCM impact models as the temporal limitations. Moreover, the outcomes of GCMs generated at lower spatial resolutions are not directly applicable to regional climate investigations. Therefore, it is necessary to employ appropriate downscaling approaches to convert GCM outputs into more refined local climatic data [3]. Furthermore, for several studies to identify the most relevant features or variables for a predictive model Mutual information (MI) has been utilized which increases the model's accuracy [4, 5].

The Köppen climate classification system, developed by Wladimir Köppen and later improved by Rudolf Geiger, is a widely utilized method for categorizing global climates based on temperature and precipitation patterns. Choosing the optimum regression model for Each Climate is challenging due to it non-linear nature of climate patterns. This study aims to utilize ten machine learning methods including SVR, LinearSVR, LASSO, LASSOCV, Elastic Net, Bayesian Ridge, Random-ForestRegressor, AdaBoost Regressor, KNeighbors Regressor, and XGBRegressor. for each koppen climate including; A (tropical), B (dry), C (temperate), and D (continental) to downscale and simulate the best machine learning method for temperature among utilized methods. To increase the accuracy of this study MI feature selection was utilized as predictor screening.

**Citation:** To be added by editorial staff during production.

Academic Editor: Firstname Lastname

Published: date

## 2. Methods and Materials

### *2.1. Study area and data set*

The cities investigated represent four main climate types: A (tropical), B (dry), C (temperate), and D (continental). Figure 1 shows the study area The initial city under scrutiny is Tabriz, situated in the northwestern region of Iran. Tabriz experiences a cold semi-arid climate (classified as B). Its geographical coordinates are 38° 5' N latitude and 46° 16' E longitude. Moving on, Miami, a coastal city located in southern Florida, United States represents a tropical monsoon climate (classified as A). Miami's geographical coordinates are 25° 45' N latitude, and 80° 11' W longitude. Edmonton, exemplifies a humid continental climate (classified as D). Edmonton's geographical coordinates are 53° 32' N latitude, and 113° 29' W longitude. Lastly, Madrid, situated in the heart of Spain, represents a Mediterranean climate (classified as C). Madrid's geographical coordinates are 40° 25' N latitude, 3° 42' W longitude.

### *2.2. data*

Monthly temperature data from Tabriz Airport, Miami Airport, Madrid Cuatrorientos, and Edmonton Saskatchewan stations were collected for the period spanning 1980 to 2014. These temperature records were utilized as part of the study. Additionally, GCMs by Can-ESM5 were collected from www.canada.ca. It is important to mention that predictors were. The grid employed in the analysis had a consistent longitudinal resolution of 2.8125° and a nearly uniform latitudinal resolution of 2.8125°.

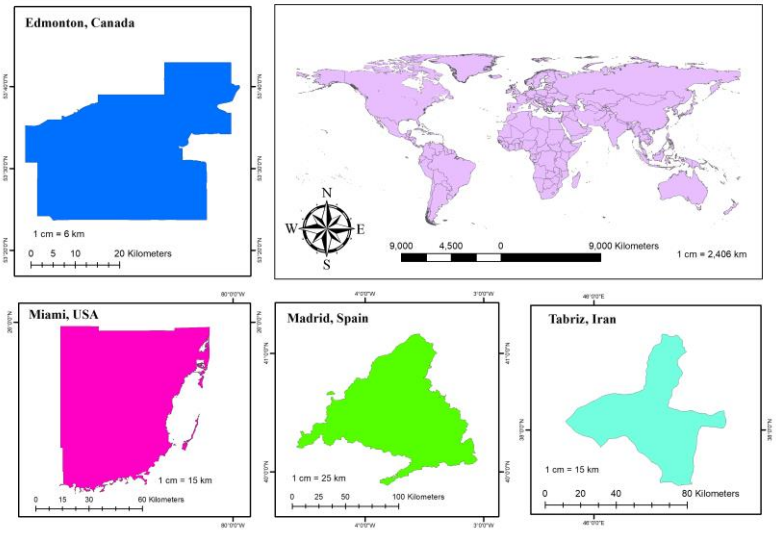


**Fig.** 1. The location of the study area classified by climate types.

### *2.3. Support vector regression (SVR)*

The Support Vector regression is a machine learning technique based on support vector machine(SVM), The primary objective of this algorithm is to construct an optimized function, denoted as f(x) that effectively captures the nonlinear relationship between a subset of training data points. The aim is to mitigate all errors that are smaller than a specified threshold [6].

### *2.4. Lasso Algorithm*

The Least Absolute Shrinkage and Selection Operator (LASSO) is a regularization method that serves as an alternative to ordinary least squares. It proves to be beneficial for feature selection and mitigating overfitting concerns. LASSO addresses overfitting by

shrinking the coefficient estimates toward zero and effectively reducing the number of variables through the utilization of a penalty parameter. [7, 8].

### *2.5. Random forest regression(RF)*

Random Forest Regression (RF) is a widely recognized and highly effective ensemble machine-learning algorithm that has gained significant popularity in the field. The main idea behind RF is to create a diverse ensemble of regression trees by randomly selecting subsets of samples and features through a process known as bootstrap sampling. RF excels in capturing complex non-linear relationships between input features and the target variable. Moreover, it demonstrates robustness against overfitting, a common issue in machine learning models [9].

### *2.6. Extreme gradient boosting*

XGBoost is an innovative and scalable machine learning framework that constructs a sequential ensemble of shallow regression trees using the gradient boosting technique [10]. During the training process, a regression tree undergoes division of the input dataset into progressively more homogeneous subsets at each decision node. The selection of splits is optimized to maximize the dissimilarity between distinct terminal nodes, ensuring effective discrimination [11].

### *2.7. k-Nearest Neighbor (kNN)*

K-nearest neighbor (kNN) is a fundamental technique in pattern recognition, falling under the umbrella of unsupervised machine learning methods. It operates by assigning class labels to objects based on their proximity to the nearest observed instances within the training dataset in the original feature space. [12].

### *2.8. AdaBoost*

AdaBoost algorithm can enhance the accuracy of a weak learning algorithm, which performs only slightly better than random guessing, and transform it into a strong learning algorithm with essentially unlimited accuracy. Its theoretical soundness has been a major factor in driving its success, both in academia and industry [13].

### *2.9. linear Support Vector Regression (LSVR)*

Support Vector Regression SVR is a widely employed linear regression model in the field of machine learning and data mining. It is an extension of least-squares regression that incorporates an $\epsilon$insensitive loss function. Additionally, to prevent overfitting of the training data, regularization is typically applied. In essence, SVR is formulated as an optimization problem that involves two key parameters: the regularization parameter and the error sensitivity parameter [6].

### *2.10. Lassocv*

LassoCV in sci-kit-learn is a valuable tool for radionics feature selection [8]. It combines cross-validation and Lasso regression, removing the requirement for manual regularization coefficient specification. By automatically exploring a range of $\lambda$ values through CV iterations, LassoCV identifies the optimal regularization parameter. This automated approach simplifies feature selection, improving the accuracy and effectiveness of radiomics analysis.

### *2.11. Elastic net*

The elastic net (ENET) is a method that builds upon the lasso technique and provides robustness against strong correlations among predictor variables**.** It addresses the instability issue faced by the lasso approach when dealing with highly correlated predictors,

such as SNPs exhibiting high linkage disequilibrium. The ENET was specifically developed for analyzing high-dimensional data [14].

*2.12. Bayesian ridge regression*

Bayesian ridge regression, similar to ridge regression, is a linear model that applies an L2 penalty to the coefficients. However, unlike ridge regression where the strength of the penalty needs to be manually set as a regularization hyperparameter, Bayesian ridge regression estimates the optimal regularization strength directly from the available data [15].

*2.13. evaluation criteria*

To evaluate the effectiveness of the methodologies employed in this investigation, two evaluation metrics, namely root mean square error (RMSE) and the coefficient of determination (also known as DC or Nash-Sutcliffe), were used.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (R_i - Z_i)^2}{N}} \quad (1)$$

$$\mathrm{DC} = 1 - \frac{\sum_{i=1}^{N} (Z_i - R_i)^2}{\sum_{i=1}^{N} (Z_i - \bar{Z})^2} \quad (2)$$

$R_i$ is used to symbolize the estimated value, $Z_i$ represents the target value, $\bar{Z}$ denotes the average value of the target observations, and N signifies the sample size. The RMSE retains the dimensionality of the observations, while the DC is dimensionless and lies within the interval of (-∞, 1]. A higher DC value converging towards 1 demonstrates an increased level of accuracy in the regression analysis.

## 3. Results

In this study, the objective was to accurately determine the optimal regression technique for downscaling climatology data based on GCMs. Among ten commonly used methods, the Can-Esm5 model was selected due to its higher nonlinearity and suitability for downscaling future parameters. The study considered four grid points and employed the MI method for feature selection, where the predictors with the highest MI values were identified as dominant predictors for each selected GCM (Table 1). Based on the MI feature selection method, variables related to temperature were consistently identified as the dominant predictors across all four grid points for the selected climates.

***Table 1.*** Appropriate predictors according to MI.

| Climate Type | Dominant predictors |
|---|---|
| A (tropical) Miami | Temp(2), Temp(4)<br>Temp(1), Shum(3)<br>Temp(3), Shum(2) |
| B (dry) Tabriz | Temp(1), Temp(3)<br>Temp(2), Temp(4)<br>Mslp(1), S850(4) |
| C (temperature) Madrid | Temp(3), Temp(1)<br>Temp(4) ,Temp(2)<br>S850(4), S850(3) |
| D (continental) Edmonton | Temp(3), Temp(4)<br>Shum(4), Temp(2)<br>Temp(1), S850(3) |

Before downscaling, the dominant predictors were standardized, and the data was split into calibration (75%) and validation (25%) sets to calibrate and validate the models. Ten regression methods were utilized to downscale the mean temperature for the four different climates (A: tropical, B: dry, C: temperature, D: continental). To evaluate the efficiency of these regression methods, RMSE and DC methods were employed. The results of mean temperature downscaling, as evaluated by RMSE and DC criteria, are presented in Fig. 2.

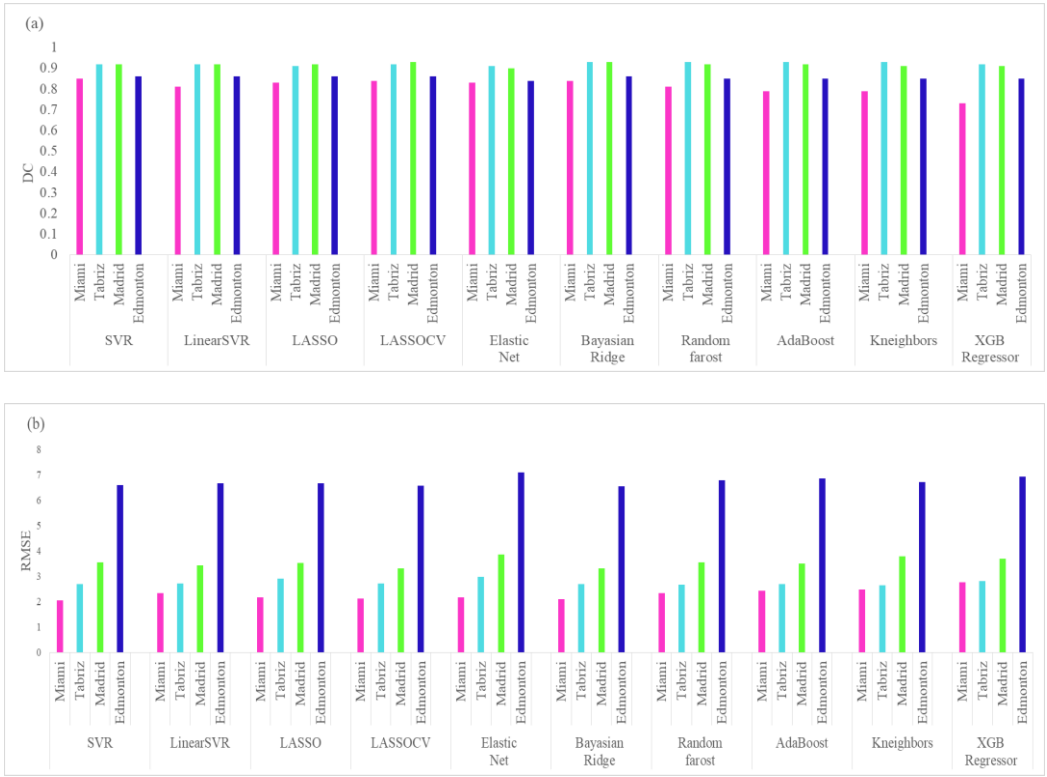


***Figure 2.*** Performance of the downscaling models for temperature a) DC and b) RMSE.

## 4. Conclusion

This study identifies the optimal regression technique among a set of ten methods for downscaling mean temperature in four distinct climatic regions. For increasing the accuracy of the models MI feature selection had been utilized. Study results indicated that for the city of Miami with a tropical climate (A), the Bayesian Ridge method outperformed the other methods. In Tabriz, a city with a dry climate (B), the LASSOCV method was identified as the most efficient. For Madrid, a city with a temperature climate (C), the KNeighbors regressor was the dominant method. Lastly, for Edmonton, a city with a continental climate (D), the SVR method exhibited superiority over other methods. For future studies, it is recommended to use other regression analyses for each climate sub-set of Köppen climate zones.

**Funding:** This research received no external funding

**Data Availability Statement:** The data is available upon request.

## References

1. Shahi, N.K., et al., *Assessment of the spatio-temporal variability of the added value on precipitation of convection-permitting simulation over the Iberian Peninsula using the RegIPSL regional earth system model.* Climate Dynamics, 2022. **59**(1): p. 471-498.

2. Shahi, N.K., *Fidelity of the latest high-resolution CORDEX-CORE regional climate model simulations in the representation of the Indian summer monsoon precipitation characteristics.* Climate Dynamics, 2022.
3. Mora, D.E., et al., *Climate changes of hydrometeorological and hydrological extremes in the Paute basin, Ecuadorean Andes.* Hydrol. Earth Syst. Sci., 2014. **18**(2): p. 631-648.
4. Okkan, U., *Assessing the effects of climate change on monthly precipitation: Proposing of a downscaling strategy through a case study in Turkey.* KSCE Journal of Civil Engineering, 2015. **19**(4): p. 1150-1156.
5. Jeong, D.I., et al., *Comparison of transfer functions in statistical downscaling models for daily temperature and precipitation over Canada.* Stochastic Environmental Research and Risk Assessment, 2012. **26**(5): p. 633-653.
6. Vapnik, V., S. Golowich, and A. Smola, *Support vector method for function approximation, regression estimation and signal processing.* Advances in neural information processing systems, 1996. **9**.
7. Friedman, J., et al., *Pathwise coordinate optimization.* The Annals of Applied Statistics, 2007. **1**(2): p. 302-332, 31.
8. Tibshirani, R., *Regression Shrinkage and Selection Via the Lasso.* Journal of the Royal Statistical Society: Series B (Methodological), 2018. **58**(1): p. 267-288.
9. Breiman, L., *Random Forests.* Machine Learning, 2001. **45**(1): p. 5-32.
10. Chang, C.-C. and C.-J. Lin, *LIBSVM: A library for support vector machines.* ACM Trans. Intell. Syst. Technol., 2011. **2**(3): p. Article 27.
11. Zhao, X., et al., *Estimation of Poverty Using Random Forest Regression with Multi-Source Data: A Case Study in Bangladesh.* Remote Sensing, 2019. **11**(4): p. 375.
12. Nasseri, M., H. Tavakol-Davani, and B. Zahraie, *Performance assessment of different data mining methods in statistical downscaling of daily precipitation.* Journal of Hydrology, 2013. **492**: p. 1-14.
13. Freund, Y. and R.E. Schapire. *Experiments with a new boosting algorithm*. in *icml*. 1996. Citeseer.
14. Zou, H. and T. Hastie, *Regularization and Variable Selection Via the Elastic Net.* Journal of the Royal Statistical Society Series B: Statistical Methodology, 2005. **67**(2): p. 301-320.
15. Faber, F.A., et al., *Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error.* Journal of Chemical Theory and Computation, 2017. **13**(11): p. 5255-5264.