Proceeding Paper

Comparing Regression Techniques for Temperature Downscaling in Different Climate Classifications [†]

Ali Ilghami Kkhosroshahi [1,*], Mohammad Bejani, Hadi Pourali 🅾 **and Arman Hosseinpour Salehi** 🅾

Faculty of Civil Engineering, University of Tabriz, Tabriz 51368, Iran; m.bejani99@ms.tabrizu.ac.ir (M.B.); h.pourali1400@ms.tabrizu.ac.ir (H.P.); armanhsalehi@gmail.com (A.H.S.)
* Correspondence: aliikhosroshahi@gmail.com
† Presented at the 4th International Electronic Conference on Applied Sciences, 27 October–10 November 2023; Available online: https://asec2023.sciforum.net/.

Abstract: This study aims to identify the optimal 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 the Köppen climate classification system, including A (tropical), B (dry), C (temperate), and D (continental), synoptic station data were collected. Furthermore, for the purpose of downscaling, a general circulation model (GCM) had been utilized. Additionally, to enhance the performance of downscaling accuracy, mutual information (MI) was employed for feature selection. The downscaling performance was evaluated using the coefficient of determination (DC) and root mean square 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 due to 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, in several studies that 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 optimal regression model for each climate is challenging due to the 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 Köppen 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, the MI feature selection was utilized as a mode of predictor screening.

2. Methods and Materials

2.1. Study Area and Data Set

The cities investigated represent the four main climate types of 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. The second city investigated is Miami, a coastal city located in southern Florida, United States, that represents a tropical monsoon climate (classified as A). Miami's geographical coordinates are 25°45′ N latitude and 80°11′ W longitude. Next, 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 and 3°42′ W longitude.
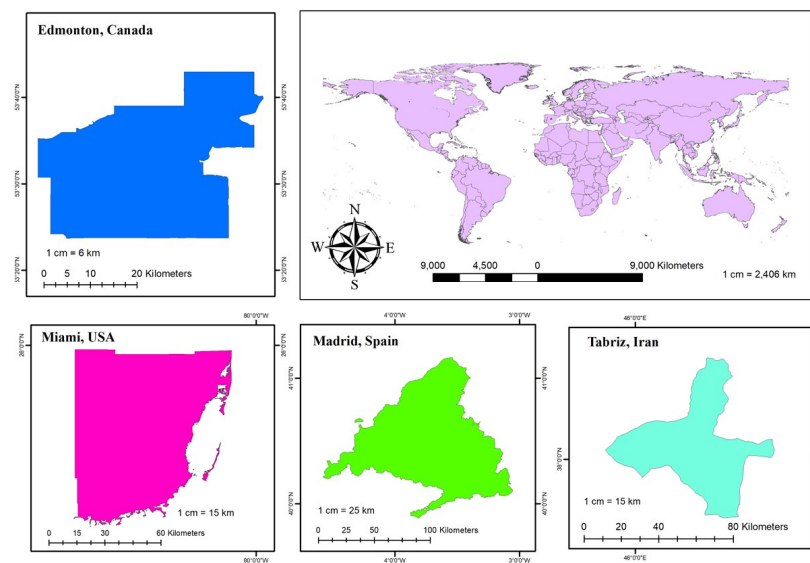
Figure 1. The location of the study area classified according to climate type.

2.2. Data

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

2.3. Support Vector Regression (SVR)

Support vector regression is a machine learning technique based a 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 non-linear 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 undertakes the division of an 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

The 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 square regression that incorporates an ϵinsensitive loss function. Additionally, to prevent the overfitting of training data, regularization is typically applied. In essence, SVR is formulated as an optimization problem that involves two key parameters of the regularization parameter and the error sensitivity parameter [6].

2.10. Lassocv

LassoCV in scikit-learn is a valuable tool for radionic feature selection [8]. It combines cross-validation and Lasso regression, removing the requirement for manual regularization coefficient specification. By automatically exploring a range of λ values through CV iterations, LassoCV identifies the optimal regularization parameter. This automated approach simplifies feature selection, improving the accuracy and effectiveness of radiomic 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 that exhibit 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.

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

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

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 $(-\infty, 1]$. A higher DC value converging towards one 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) Temp(1), Shum(3) Temp(3), Shum(2)
B (dry) Tabriz	Temp(1), Temp(3) Temp(2), Temp(4) Mslp(1), S850(4)
C (temperature) Madrid	Temp(3), Temp(1) Temp(4),Temp(2) S850(4), S850(3)
D (continental) Edmonton	Temp(3), Temp(4) Shum(4), Temp(2) 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 Figure 2.
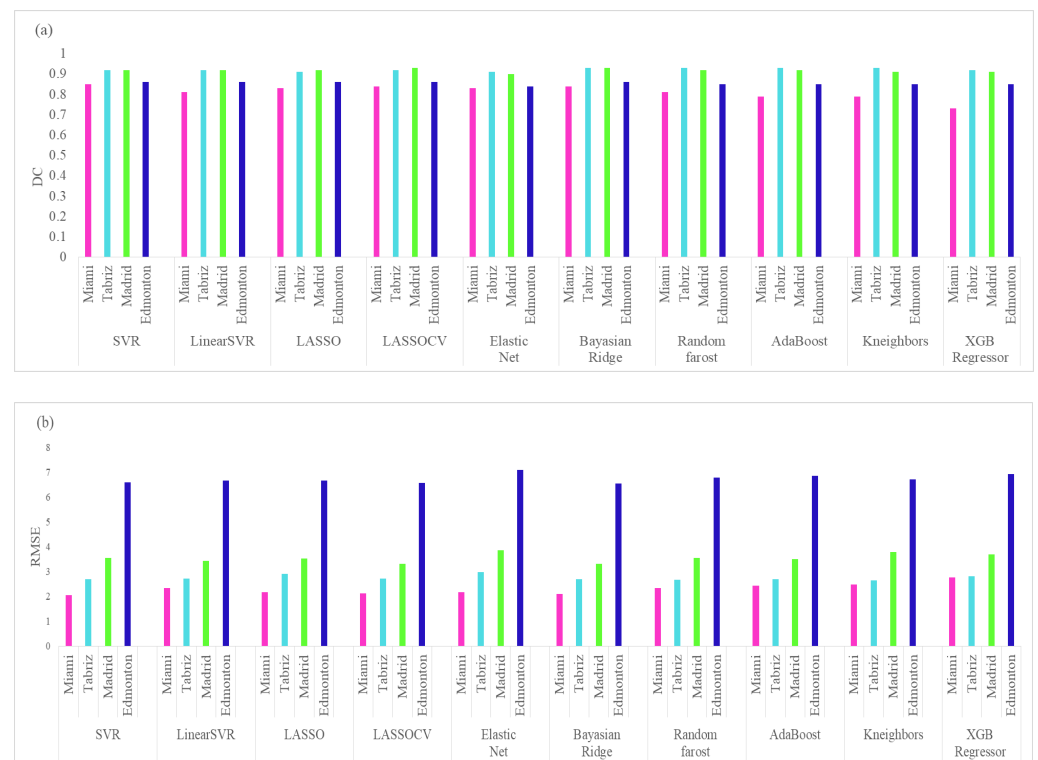
Figure 2. Performance of the downscaling models for temperature employing (**a**) DC and (**b**) RMSE.

4. Conclusions

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 was utilized. The study results indicated that the Bayesian Ridge method outperformed the other methods for the city of Miami with a tropical climate (A). 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 subset of Köppen climate zones.

References

1. Shahi, N.K.; Polcher, J.; Bastin, S.; Pennel, R.; Fita, L. 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. *Clim. Dyn.* **2022**, *59*, 471–498. [CrossRef]
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. *Clim. Dyn.* **2022**. [CrossRef]
3. Mora, D.E.; Campozano, L.; Cisneros, F.; Wyseure, G.; Willems, P. Climate changes of hydrometeorological and hydrological extremes in the Paute basin, Ecuadorean Andes. *Hydrol. Earth Syst. Sci.* **2014**, *18*, 631–648. [CrossRef]
4. Okkan, U. Assessing the effects of climate change on monthly precipitation: Proposing of a downscaling strategy through a case study in Turkey. *KSCE J. Civ. Eng.* **2015**, *19*, 1150–1156. [CrossRef]
5. Jeong, D.I.; St-Hilaire, A.; Ouarda, T.B.M.J.; Gachon, P. Comparison of transfer functions in statistical downscaling models for daily temperature and precipitation over Canada. *Stoch. Environ. Res. Risk Assess.* **2012**, *26*, 633–653. [CrossRef]
6. Vapnik, V.; Golowich, S.; Smola, A. Support vector method for function approximation, regression estimation and signal processing. *Adv. Neural Inf. Process. Syst.* **1997**, 281–287.
7. Friedman, J.; Hastie, T.; Höfling, H.; Tibshirani, R. Pathwise coordinate optimization. *Ann. Appl. Stat.* **2007**, *1*, 302–332. [CrossRef]
8. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B Methodol.* **2018**, *58*, 267–288. [CrossRef]
9. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
10. Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 27. [CrossRef]
11. Zhao, X.; Yu, B.; Liu, Y.; Chen, Z.; Li, Q.; Wang, C.; Wu, J. Estimation of Poverty Using Random Forest Regression with Multi-Source Data: A Case Study in Bangladesh. *Remote Sens.* **2019**, *11*, 375. [CrossRef]
12. Nasseri, M.; Tavakol-Davani, H.; Zahraie, B. Performance assessment of different data mining methods in statistical downscaling of daily precipitation. *J. Hydrol.* **2013**, *492*, 1–14. [CrossRef]
13. Freund, Y.; Schapire, R.E. Experiments with a new boosting algorithm. In Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, San Francisco, CA, USA, 3–6 July 1996; pp. 148–156.
14. Zou, H.; Hastie, T. Regularization and Variable Selection Via the Elastic Net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2005**, *67*, 301–320. [CrossRef]
15. Faber, F.A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S.S.; Dahl, G.E.; Vinyals, O.; Kearnes, S.; Riley, P.F.; von Lilienfeld, O.A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, *13*, 5255–5264. [CrossRef] [PubMed]