Comparison performance of machine learning and geostatistical methods for the interpolation of monthly air temperature over Costa Rica

M Méndez[1] and L.A Calvo-Valverde[2]

1 Associate Professor, Escuela de Ingeniería en Construcción, Instituto Tecnológico de Costa Rica, APDO 159-7050, Cartago, Costa Rica
2 Associate Professor, Escuela de Ingeniería en Computación, Instituto Tecnológico de Costa Rica, APDO 159-7050, Cartago, Costa Rica
E-mail: mamendez@itcr.ac.cr

Abstract. The performance of three machine learning (ML) methods; cubist regression (CR), random forest (RF) and generalized additive model using splines (GAM) in generating monthly air temperature grids over Costa Rica was evaluated against two heavily used geostatistical methods; ordinary kriging (OK) and kriging with external drift (KED). The skill of the interpolation methods was evaluated using a 10-fold cross-validation technique; selecting the root-mean square error (RMSE), the mean absolute error (MAE) and the Pearson correlation-coefficient (R) as agreement metrics. To this purpose, data from an irregularly-distributed observational-network comprised by 73 weather-stations were selected for the period 1950-1987. Several spatial fields derived from a high-resolution digital elevation model (DEM) were tested as covariants. Results from the 10-fold cross-validation test show that CR yielded the best individual performance followed by KED, whereas GAM performed worst. Elevation on the other hand, was the only covariant ultimately incorporated in the interpolation process, since the remaining spatial fields exhibited poor correlation with temperature or resulted in data redundancy. While the quantitative and qualitative evaluation of CR and KED can be said to be comparable, CR is considered the best approach since the method is unaffected by assumptions on data normality and homoscedasticity.

1. Introduction

Air surface temperature maps of high spatial resolution are crucial to a variety of applications including crop growth, ecology, hydrology, weather forecasting, environmental studies and climate change [1]. Observed meteorological data, collected from sparsely and unevenly distributed weather-stations are commonly used for the generation of gridded air temperature maps [2]. Therefore, spatial interpolation methods are frequently used to overcome the limitations of low observational-network densities and sparse distribution of weather-stations especially in mountainous areas, since they allow for the generation of continuous fields over locations that are not geographically covered by existing observation network [3]. Various spatial interpolation methods have been developed for predicting the spatial distribution of fields of interest. Interpolation methods however, can propagate important uncertainties into the generated climatologies, which underscores the importance of selecting suitable interpolation methods when characterizing climatic variables performance and biases [4]. Traditionally, geostatistical methods such as ordinary kriging (OK) have been used in the interpolation of air surface temperature [5]. Kriging estimates variable values at target locations in space based on the spatially dependent variance of known sampling values [6]. When using kriging, several products

besides the prediction field can be generated, including the estimation of the residual errors and the kriging variance. Kriging variance is also estimated at target locations and consequently provides a spatial view on the measure of performance [7]. Kriging interpolation techniques that incorporate covariants as auxiliary information, such co-kriging (CK) and kriging with external drift (KED) are applicable to areas with a clear relationship between the variables, where the expected variable is considered a linear function of one or more covariants [8]. A strong correlation between surface air temperature and spatial covariants such as elevation, slope and aspect been extensively analysed, which promotes their inclusion as covariant within various interpolation methods [9] [10] [11]. Nevertheless, kriging interpolation techniques rely on the assumptions of data normality and homoscedasticity, which cannot always satisfy [6]. More recently, machine learning (ML) interpolation methods such as cubist regression (CR), random forest (RF) and generalized additive model using splines (GAM), have been used in generating air temperature grids [4] [12] [13] [14]. Cubist regression (CR) and random forest (RF) are tree-based machine learning algorithms that predict a response from a set of predictors by creating multiple decision trees and aggregating their results [15]. The decision trees are constructed through recursive partitioning of a bootstrapped subset of the training data, which is split by defining an optimal threshold based on a randomly selected subset of predictor auxiliary variables. This provides two resulting data partitions, each with the least variation in the target variable. This process is then repeated successively on each data partitions until the terminal nodes are reached [16]. Both methods are unaffected by assumptions on data normality and homoscedasticity and are computationally less demanding than various kriging variants. Generalized additive model using splines (GAM) on the other hand, are a semi-parametric extension of generalized linear models (GLM), where no parametric coefficients are estimated for predictors. GAM allows a nonlinear relationship between the response and explanatory variables [17]. In this context, the objective of this study is to assess the performance of various interpolation methods in generating monthly air temperature grids over Costa Rica.

2. Methodology

2.1. Study area

Costa Rica is located across the Central American isthmus between Panama and Nicaragua and occupies an area of 51,060 km2 (Figure 1b). The country is bordered by the Caribbean Sea to the east and the Pacific Ocean to the west, which favours oceanic and climatological influences from both oceans. The country is meridionally divided by a northwest-southeast trending cordillera of topographic complexity, characterized by a mountainous topography with elevations ranging from sea-level to about 3400 m (Figure 1a). Consequently, the climate in Costa Rica is dependent on geographical altitude and has been divided into six separate climatic regions; North, Caribbean, North-Pacific, Central-Valley, Central-Pacific and South-Pacific, in which northeastern and southwestern domains are determined by the position and elevation of the aforementioned cordilleras (Figure 1a). Climate variability is driven by interactions between local elevation and a combination of the seasonal migration of the intertropical convergence zone (ITCZ), which includes sea breeze effects, monsoonal circulations, strong easterly trade winds, cold air masses from mid-latitudes in the winter and the perturbing influences of hurricanes and tropical cyclones in the Atlantic Ocean [42].

2.2. Data sources

Monthly temperature data were provided by Instituto Meteorológico of Costa Rica (IMN) for the period 1950-1987. The observational-network was comprised by 73 irregularly-distributed weather-stations (Figure 1). Only stations possessing at least 20 years of continuous records during the study period were included. Topographic information was derived from the Advanced Land Observing Satellite (ALOS) AW3D-30m Digital Elevation Model (DEM) and resampled to 1x1 km spatial resolution using bilinear resampling. Resampled DEM elevation was preferred over actual station elevations to improve spatial consistency. Spatial fields tested as covariants were mainly derived from the ALOS DEM (Table 1).

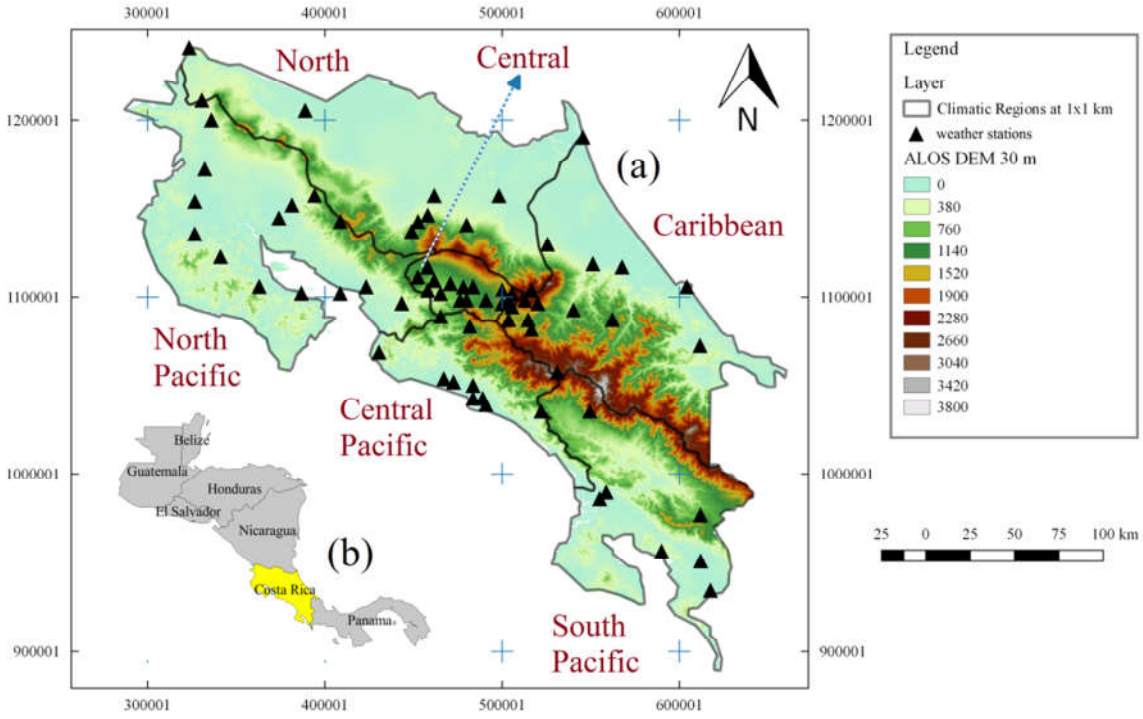

Figure 1. (a) Location of weather stations and Digital Elevation Model (DEM) for each climatic region in Costa Rica during the period 1950-1987. (b) Position of Costa Rica in Central America.

2.3. Interpolation methods

All spatial data processing was executed using the R programming language, along with specialized R packages. Geostatistical modelling, spatio-temporal data analysis and raster generation were implemented by combining functionalities of the **gstat**, **sp**, **raster**, **RSAGA** and **rgdal** packages. OK and KED automatic variogram fitting analysis was executed using the **automap** package. CR and RF were implemented by the **Cubist** and **randomForest** packages respectively. All spatial products followed the official Transverse Mercator projection system (CRTM05). Mechanical models included in the OK and KED automatic variogram fitting were: Spherical (*Sph*), Gaussian (*Gau*), Exponential (*Exp*) and M. Stein's parameterization (*Ste*). ORK and KED were run in parallel by means of the **doParallel** and **foreach** packages due to high computational demand.

2.4. Spatial covariants selection

The R package **VSURF** (*Variable Selection Using Random Forests*) was used to discretize among the spatial fields tested as covariants (Table 1). **VSURF** applies a three-step variable selection procedure based on random forests for supervised classification and regression problems. The first step is dedicated to eliminate irrelevant auxiliary variables from the dataset. The second step aims to select all variables related to the response for interpretation purpose, while the third step refines the selection by eliminating redundancy in the set of variables selected by the second step for prediction purposes.

2.5. Interpolation model performance

The skill of the interpolation methods was evaluated using a 10-fold cross-validation technique with a repeated cross-validation scheme with 5 repetitions using 80% of the data for training at each iteration. The root-mean square error (RMSE), the mean absolute error (MAE) and the Pearson correlation-coefficient (R) were selected as objective functions (OF). The median OF values obtained from the k-fold cross-validation were used as the primary metrics to compare the predictive power of methods with each other.

Table 1. Spatial fields tested as covariants for the interpolation process.

Variable	Description	VSURF-Status
X	CRTM05 X position (m)	Irrelevant
Y	CRTM05 Y position (m)	Interpretation
SLOPE	Terrain slope (radians)	Interpretation
ALOS_FILL	ALOS AW3D-30m DEM (m)	Prediction
ASPECT	Aspect (radians)	Irrelevant
TOPOGRAPHIC	Topographic Wetness Index (-)	Irrelevant
LS_FACTOR	USLE Slope length (LS) factor (-)	Interpretation
DIRECT_INSO	Potential incoming solar radiation (kW/m2)	Interpretation
CONVEXITY	Terrain Surface Convexity (-)	Interpretation
WIND_EXPOSI	Wind Exposition Index (-)	Interpretation
BUFFER	Euclidean distance to feature observations (m)	Irrelevant
COAST	Euclidean coastline proximity (m)	Interpretation
PREC	Mean monthly precipitation (mm/month)	Irrelevant

3. Results and discussion

3.1. Selection of spatial covariants

The application of the three-step variable selection procedure of the **VSURF** R-function identified seven potential covariants after the first screening step for interpretation purposes (Table 1). Nevertheless, after the following two remaining steps, elevation (*ALOS_FILL*) was the only spatial field ultimately incorporated as predictor in the interpolation process; since the remaining auxiliary variables exhibited poor correlation with temperature or resulted in data redundancy. This suggests that for Costa Rica, temperature changes over time independently of local factors other than elevation. It is worth noting that Y-position (latitude) scored highest in interpretation after ALOS_FILL. Nonetheless, the relatively narrow geographical extension of Costa Rica (Figure 1) does not allow sufficient temperature variation with latitude as to be taken into consideration by VSURF for prediction purposes. Correlation plots between mean monthly temperature and elevation (Figure 2), obtained after the application of robust linear regression (rlm) show that the slopes of temperature linear trend-lines are clearly related to elevation, and the amplitude of temperature variation is enlarged by high altitude. An overall linear trend for the vertical mean-temperature lapse-rate for the entire country, regardless of the months of year can be observed, with R^2 values oscillating between 0.87 and 0.96, suggesting that a considerable fraction of the air temperature variance can be explained by elevation. Moreover, temperature variations not only depend on elevation but also on seasonal scales, since the temperature lapse-rate varies within a range of 0.534°C/100 m to 0.602 °C/100 m for the months of August and March respectively. This suggests that during JFMA, which coincides with the driest months [18], the temperature lapse-rate is steeper and remains more stable to decrease towards the end of December. Contrary, as the rainy season begins, the lapse-rate becomes flatter, indicating a lower dependency of temperature as a function of elevation. For instance, ASO present the flattest lapse-rates as they represent the rainiest months of the year [18]. Seasonal variation of temperature at low elevations seems fairly stable throughout the year, probably because of more stable and constant surface heating, which is supported by a greater concentration of weather-stations at low elevations (Figure 1). It becomes clear that the seasonal variability of temperature is quite complex due to the combined action of local weather and climate conditions. Consequently alternative spatial covariants such as surface soil, vegetation cover, land use and relative humidity could be considered and eventually evaluated.

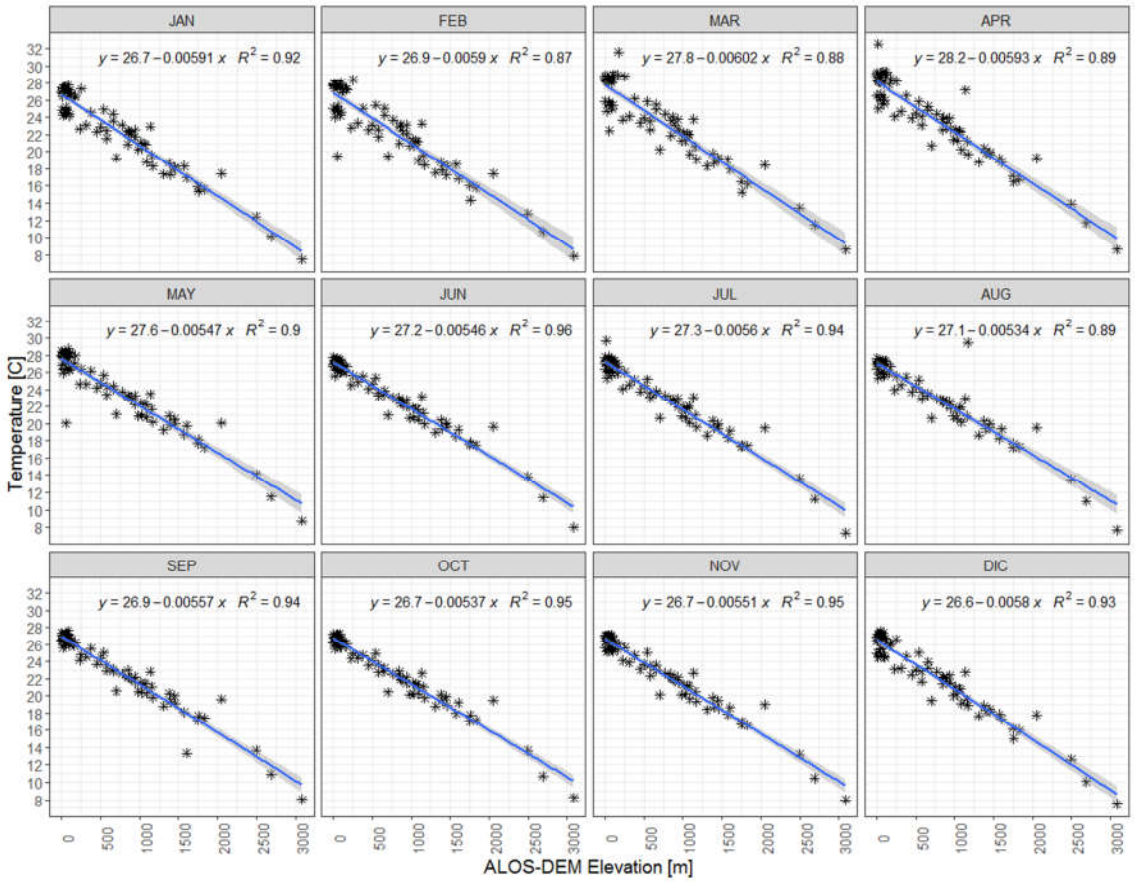

Figure 2. Correlation between mean monthly temperature and ALOS-DEM elevation during the period 1950-1987.

3.2. Performance of interpolation methods

Taylor diagrams (Figure 3) summarize the results of the 10-fold cross-validation process of the spatial agreement for temperature in terms of correlation coefficient (R), root mean square error (RMSE) and standard deviation (SD). As shown, CR yielded the best individual performance of all interpolation methods followed by KED and RF. In some instances, KED slightly outperformed CR, namely APR, JUN and OCT. For most cases however, CR was able to keep the RMSE close or below 1 [°C], with higher spatial R values and SD values closer to the observational-trend (around 3 [°C]). CR proved superior during the months of FEB and AUG, both of which exhibited the largest RMSE and the lowest R and SD values respectively. RF on the other hand, shows slightly higher RMSE values, but considerably lower SD values when compared to CR and KED. This could be an indication of model over-fitting which gradually starts to smooth-out the resulting temperature interpolated surface. In contrast, OK and GAM respectively performed worst, as they show the highest deviations and lowest correlations of all remaining methods; with RMSE over 2 [°C] for most of the year. At the same time, monthly boxplots (Figure 4) of the 10-fold cross-validation process, which used the same training and testing datasets for each of the interpolation methods; show the performance superiority of CR regardless of R, RMSE, MAE or months of the year. Concerning MAE, CR is able to keep the vast majority of points below 1 [°C], with median values in the order of 0.5 [°C]. MAE values for KED are marginally higher than CR in most cases but considerably lower than OK or GAM. While the quantitative evaluation of CR and KED can be said to be comparable, CR is considered the best approach since the method is unaffected by assumptions on data normality and homoscedasticity. The superior performance of CR probably stems from its ability to make better use of the elevation covariant field and from its insensitivity to non-normal distributions and outliers within the various

training datasets included in the 10-fold cross-validation process. Geostatistical methods such as KED and OK, assume second order stationarity and spatial autocorrelation; assumptions that are generally incorrect when dealing with temperature spatial fields. On the other hand, the mechanical models selected for variogram auto-fitting (*Sph, Exp, Gau, Mat and Sten*) may not be sufficient to capture specific spatial structures of particular months and therefore, could represent a disadvantage of the automation process executed by **automap**. Consequently, a wider family of mechanical variograms models could eventually be evaluated, although the present results indicate that such potential improvement should be considerably significant to outperform the skill of CR, which proved to be a powerful machine learning algorithm to integrate complex, non-linear relationship using auxiliary spatial variables. As expected, air temperature grid maps interpolated by CR (Figure 5) exhibit temperature pattern changes at spatial and seasonal scales. The North-Pacific, Central-Pacific and South-Pacific regions are by far the warmest areas of the country, with temperatures ranging 35 °C to 40 °C during the months of June, slowly decreasing towards December, when temperatures reaches values below 25 °C. Far more stable temperature ranges can be observed for the North and Caribbean regions, both of which face the Atlantic coast and therefore, exhibit milder temperature disparities. The Central-Valley region on the other side, experiences the lower average temperature values throughout the year, with values way below 25 °C.

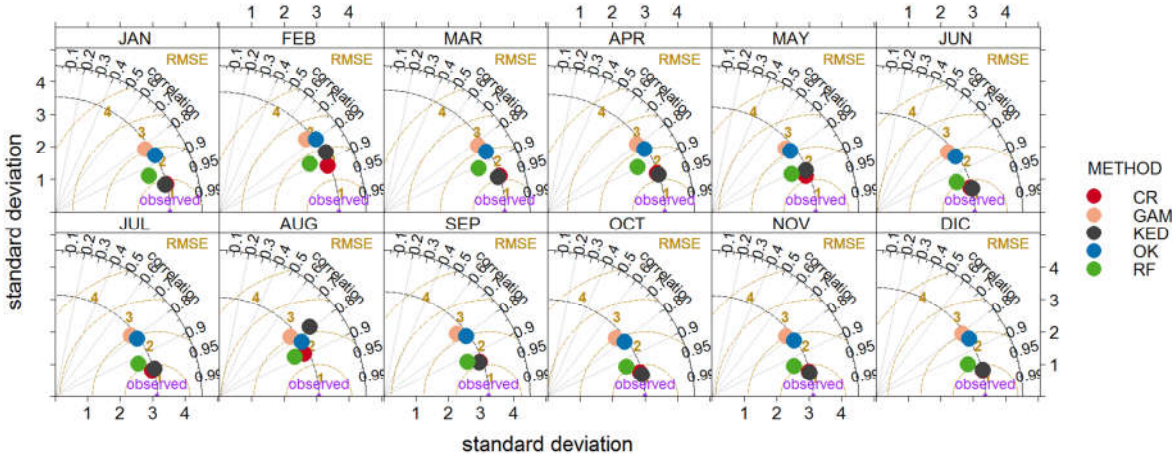

Figure 3. Taylor diagrams for the 10-fold cross-validation process in [°C].

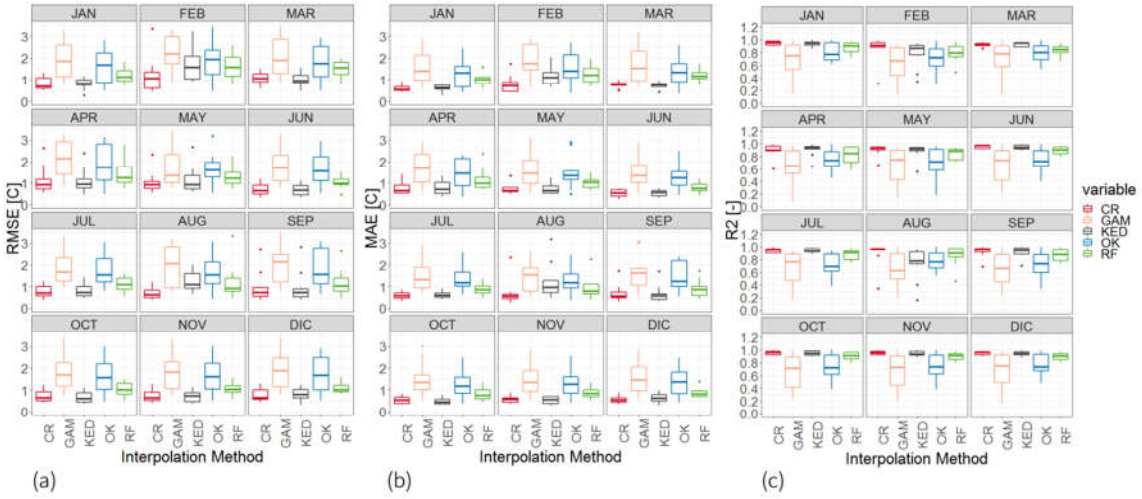

Figure 4. Boxplots of (a) RMSE, (b) MAE and (c) for the 10-fold cross-validation process.

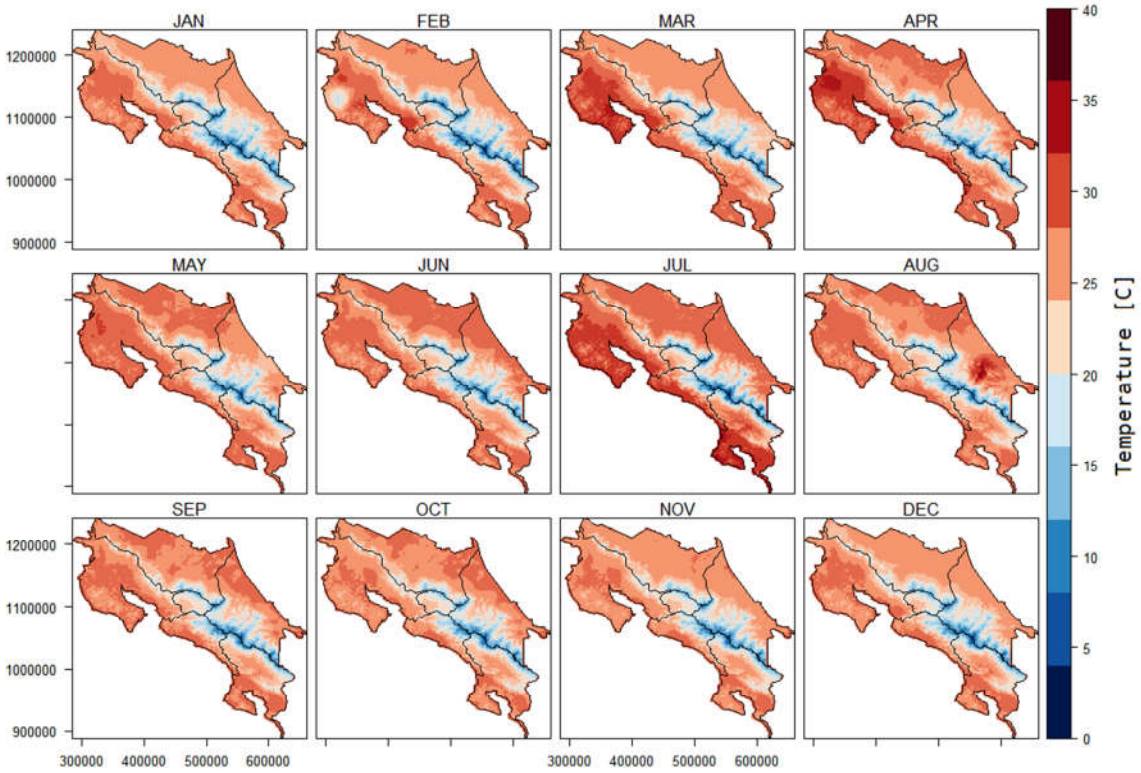

Figure 5. Mean monthly temperature maps for each climatic region of Costa Rica generated by cubist regression (CR) during the period 1950-1987.

4. Conclusions and recommendations

The performance of three machine learning methods and two geostatistical methods interpolation methods in generating monthly air temperature grids over Costa Rica was evaluated; the following conclusions can be drawn:

- After variable screening, elevation was the only covariant ultimately incorporated in the interpolation process, since the remaining spatial fields exhibited poor correlation with temperature or resulted in data redundancy, which at the same time suggests that temperature changes over time independently of local factors other than elevation.
- An overall linear trend for the vertical mean-temperature lapse-rate for the entire country can be observed regardless of temporal scales.
- CR yielded the best individual performance followed by KED and RF, whereas OK and GAM performed worst.
- While the quantitative evaluation of CR and KED can be said to be comparable, CR is considered the best approach since the method is unaffected by assumptions on data normality and homoscedasticity.
- Qualitative evaluation of mean monthly temperature maps generated by CR verify a decreasing temperature pattern with elevation, being July the hottest month and December the coolest month irrespectively of each climatic region.

The future evaluation of additional geostatistical and machine learning methods, along with different spatial covariants could be considered; although the present results indicate that such potential improvement should be considerably significant to outperform the skill of CR, which proved to be a powerful algorithm to integrate complex, non-linear relationship using auxiliary variables.

Acknowledgments
This research was supported by Vicerrectoría de Investigación & Extensión, Instituto Tecnológico de Costa Rica (TEC), specifically for the research project entitled "*Evaluación del impacto del Cambio Climático futuro sobre cuencas hidrológicas destinadas al abastecimiento de agua potable en Costa Rica*". The authors are grateful to Instituto Meteorológico of Costa Rica (IMN) for providing observed temperature data for this project.

References
[1] Carrera-Hernandez J J, and Gaskin S J 2007 Spatio temporal analysis of daily precipitation and temperature in the Basin of Mexico *J. Hydrol.* **336** 231-249
[2] Chen S and Guo J 2017 Spatial interpolation techniques: their applications in regionalizing climate-change series and associated accuracy evaluation in Northeast China *Geomat. Nat. Haz. Risk*. **8** 689-705
[3] Mendez M, Calvo-Valverde L A, Maathuis B and Alvarado-Gamboa L F 2019 Generation of Monthly Precipitation Climatologies for Costa Rica Using Irregular Rain-Gauge Observational Networks *Water*. **11** 70
[4] Aalto J, Pirinen P, Heikkinen J and Venäläinen A 2013 Spatial interpolation of monthly climate data for Finland: comparing the performance of kriging and generalized additive models *Theor. Appl. Climatol.* **112** 99-111
[5] Berndt C and Haberlandt U 2018 Spatial interpolation of climate variables in Northern Germany Influence of temporal resolution and network density *J. Hydrol. Reg. Stud*. **15** 184-202
[6] Cecinati F, Wani O and Rico-Ramirez M A Comparing Approaches to Deal With Non-Gaussianity of Rainfall Data in Kriging-Based Radar-Gauge Rainfall Merging *Water. Resour. Res*. **53** 8999-9018
[7] Yeh H C, Chen Y C, Chang C H, Ho C H and Wei C 2017 Rainfall Network Optimization Using Radar and Entropy *Entropy* **19** 553
[8] Daly C, Nielson R P and Phillips D L 1994 A Statistical-Topographic Model for Mapping Climatological Precipitation over Montainous Terrain *J. Appl. Meteor.* **33** 140-158
[9] Wang K, Sun J, Cheng G, Jiang H 2011 Effect of altitude and latitude on surface air temperature across the Qinghai-Tibet Plateau *J. Mt. Sci*. **8** 808-816
[10] Sanchez-Moreno J F, Mannaerts C and Jettena V 2014 Influence of topography on rainfall variability in Santiago Island, Cape Verde Int. *J. Climatol*. **34** 1081-1097
[11] Mendez M and Calvo-Valverde L A 2016 Assessing the Performance of Several Rainfall Interpolation Methods as Evaluated by a Conceptual Hydrological Model *Procedia. Eng*. **154** 1050-1057
[12] Thanh Noi P, Degener J and Kappas M 2017 Comparison of Multiple Linear Regression, Cubist Regression, and Random Forest Algorithms to Estimate Daily Air Surface Temperature from Dynamic Combinations of MODIS LST Data *Remote. Sens.* **9** 398
[13] Hengl T, Nussbaum M, Wright M N, Heuvelink G and Gräler B 2018 Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables *PeerJ* **6** e5518
[14] Appelhans T, Mwangomo E, Hardy D R, Hemp A and Nauss T 2015 Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania *Spat. Stat*. **14** 91-113
[15] Breiman L 2001 Random Forests *Mach. Learn.* **45** 5-32
[16] Bataille C P, von Holstein I, Laffoon J E, Willmes M, Ming Liu X and Davies G R 2018 A bioavailable strontium isoscape for Western Europe: A machine learning approach *Plos. One*. **13**(5): e0197386
[17] Hastie T and Tibshirani R 1990 Generalized additive models monographs on statistics and applied probability 43 Chapman and Hall, New York
[18] Rapp A D, Peterson A G, Frauenfeld O W, Quiring S M and Roark E B 2014 Climatology of Storm Characteristics in Costa Rica using the TRMM Precipitation Radar. *J. Hydrometeorol*. **15** 2615-2633