Forest Ecosystems

RESEARCH **Open Access**

Alternative approaches for estimating missing climate data: application to monthly precipitation records in South-Central Chile

Alonso Barrios[1,2*], Guillermo Trincado[3] and René Garreaud[4]

Abstract

Background: Over the last decades interest has grown on how climate change impacts forest resources. However, one of the main constraints is that meteorological stations are riddled with missing climatic data. This study compared five approaches for estimating monthly precipitation records: inverse distance weighting (IDW), a modification of IDW that includes elevation differences between target and neighboring stations (IDW_m), correlation coefficient weighting (CCW), multiple linear regression (MLR) and artificial neural networks (ANN).

Methods: A complete series of monthly precipitation records (1995–2012) from twenty meteorological stations located in central Chile were used. Two target stations were selected and their neighboring stations, located within a radius of 25 km (3 stations) and 50 km (9 stations), were identified. Cross-validation was used for evaluating the accuracy of the estimation approaches. The performance and predictive capability of the approaches were evaluated using the ratio of the root mean square error to the standard deviation of measured data (RSR), the percent bias (PBIAS), and the Nash-Sutcliffe efficiency (NSE). For testing the main and interactive effects of the radius of influence and estimation approaches, a two-level factorial design considering the target station as the blocking factor was used.

Results: ANN and MLR showed the best statistics for all the stations and radius of influence. However, these approaches were not significantly different with IDW_m. Inclusion of elevation differences into IDW significantly improved IDW_m estimates. In terms of precision, similar estimates were obtained when applying ANN, MLR or IDW_m, and the radius of influence had a significant influence on their estimates, we conclude that estimates based on nine neighboring stations located within a radius of 50 km are needed for completing missing monthly precipitation data in regions with complex topography.

Conclusions: It is concluded that approaches based on ANN, MLR and IDW_m had the best performance in two sectors located in south-central Chile with a complex topography. A radius of influence of 50 km (9 neighboring stations) is recommended for completing monthly precipitation data.

Keywords: Climatological data, Cross-validation, Artificial neural networks, Multiple linear regression

* Correspondence: alonso.barrios@postgrado.uach.cl
[1]Escuela de Graduados, Facultad de Ciencias Forestales y Recursos Naturales, Universidad Austral de Chile, Box 567 Valdivia, Chile
[2]Departamento de Ciencias Forestales, Facultad de Ingeniería Forestal, Universidad del Tolima, Box 6299 Ibagué, Colombia
Full list of author information is available at the end of the article

Background

The effects of climate on natural resources have become highly relevant (Cannell et al. 1995). In forestry, there is an increasing interest to study the influence of climate on forest productivity (Álvarez et al. 2013), forest hydrology (Dai et al. 2011), soil water availability (Ge et al. 2013), and wood quality (Xu et al. 2013). Nowadays, climate data are also required for parameterizing process-based simulators of tree growth (Sands and Landsberg 2002) and for studying forest water balance (Huber and Trecaman 2002), phenology processes (Codesido et al. 2005) and to carry out pest and disease research (Ahumada et al. 2013). To perform these studies, complete and homogenous climate data that covers a sufficiently long period of time is required (Teegavarapu 2012; Khosravi et al. 2015).

Climate data often have missing information that limits their use (Alfaro and Pacheco 2000). Missing values in climate series affects parameter estimation when applying regression and multivariate analysis techniques (Ramos-Calzado et al. 2008). In most cases, some techniques must be applied to estimate missing data. In forestry, there are few studies that have compared the accuracy of different approaches. Furthermore, factors that might affect their precision have not been studied in detail.

The simplest approach for imputing missing values involves the data being filled-in. The main limitation is that these approaches are suitable for small gaps and can only be applied to climate variables with a high degree of autocorrelation (Khosravi et al. 2015), which is not the case for annual mean temperatures or precipitation values. A more common approach to complete missing data is to use information from neighboring meteorological stations (Vasiliev 1996), using techniques such as inverse distance weighting (IDW). Nonetheless, horizontal distance is not a measure of spatial autocorrelation (e.g., Ahrens 2006; Ramos-Calzado et al. 2008), especially when the region contains prominent topographic features or major water bodies. Indeed, two relatively close stations can feature substantial differences in their mean climate and climate variability if they are located at opposite sides of a mountain range. Spatial correlations could be quantified by calculating the correlation coefficient between time series obtained at different locations. Teegavarapu and Chandramouli (2005) found that replacing distances with correlation coefficients as weights improved estimation of missing precipitation data. The resulting method is known as a coefficient of correlation weighting (CCW), reported by Teegavarapu (2009).

Simple and multiple linear regressions have been successfully used to estimate precipitation (Pizarro et al. 2009), and temperature (Xia et al. 1999) in different topographical conditions. Alfaro and Pacheco (2000) compared different estimation approaches for missing precipitation data, including normal ratio and linear regression. They found that the best results were obtained when applying multiple linear regression; in agreement with the results reported by Xia et al. (1999) and Pizarro et al. (2009).

Recent studies used artificial neural networks for completing climate data (Kuligowski and Barros 1998; Khorsandi et al. 2011; Ghuge and Regulwar 2013). Kuligowski and Barros (1998) compared the performance of artificial neural networks for completing six-hour precipitation data at six test stations from nearby stations, to four other approaches, such as the simple nearest-neighbor estimate, the arithmetic average, the inverse distance weighting and linear regression. They found that artificial neural networks and linear regression approaches produced the lowest overall errors. Khorsandi et al. (2011) compared four approaches including the artificial neural network, normal ratio, inverse distance weighting, and a geographical coordinate approach for completing missing monthly precipitation data. They found that artificial neural networks produced the best results compared to other approaches. Different artificial neural network designs have been developed and tested for missing data estimation. Coulibaly and Evora (2007) compared six different types of artificial neural networks and found that the multilayer perceptron (MLP) appears to be the most effective for completing missing daily precipitation values and missing daily maximum and minimum temperature values.

Several studies evaluated the predictive capability of different approaches for completing missing climate data, but few have evaluated the effects of the radius of influence when selecting neighboring stations (e.g. Chen and Liu 2012) in regions with complex topography. We tested the predictive capability of five reported approaches at completing missing data of monthly precipitations from 1995 to 2012 from south-central Chile (around 37°S) along the west slope of the Andes mountain range. This region features a climate transition between semiarid conditions in the north and more humid conditions in the south (e.g. Viale and Garreaud 2015). More importantly, the region exhibits a complex topography including a central valley flanked by the Andes mountain range, reaching over 2.000 m asl (above sea level). Our specific objectives are (i) to compare different approaches for estimating missing monthly precipitation data based on measures of precision and bias, and (ii) to evaluate the effect of the number of available neighboring stations within a radius of influence (25 and 50 km) on estimation precision. We selected monthly precipitation as the target climate variable because it is a limiting factor for fast-growing radiata pine plantations in Chile (Gerding and Schlatter 1995; Álvarez et al. 2013).

Methods

Data

Twenty meteorological stations from the DGA (Dirección General de Aguas) located in central Chile (36°–38°S, 71°–72°W, Fig. 1) with complete monthly precipitation records from January 1995 to December 2012 were selected (Table 1). Annual mean rainfall in this region ranges from 1000 to 2000 mm.

We selected stations Diguillin (number 1) and Mulchen (number 11) as target stations because they were surrounded by an equal number of neighboring stations with a radius of influence of 25 and 50 km (Table 1). Meteorological stations 1 and 11 were located in the Andean foothills at an elevation of 670 m asl and in the Central valley at an elevation of 130 m asl, respectively (Fig. 1). In this part of the country there is marked seasonality, with winter (May to September) rainfall accounting for over 65% of annual accumulation and associated with widespread frontal systems crossing the region (e.g. Falvey and Garreaud 2007). Episodes of isolated convection are infrequent over this region and account for a very small fraction of the annual accumulation (Viale and Garreaud 2014). However, winter frontal rainfall is modified by the topography producing a marked enhancement over the western slope of the Andes relative to low-land values (Viale and Garreaud 2015). For instance, the horizontal distance between our target stations is less than 70 km but annual mean precipitation increases from about 1200 mm in the lower station (1) to 2100 mm in the higher station (11). On the other hand, annual precipitation across central Chile exhibits significant inter-annual variability where the standard deviation of annual accumulation is up to a third of the mean value due to the effects of the cold and warm phases of El Niño Southern Oscillation (ENOS; e.g., Montecinos and Aceituno 2003; Garreaud 2009).

The meteorological stations located in the Andes foothills show less variability in terms of mean annual precipitation than stations located in the Central valley (Fig. 2). This can be partially explained by an increased amount of stations at higher elevations located in the Central valley (CV = 62.8%) compared to the Andes foothills (CV = 54.5%).

The Euclidean distance between target and neighboring stations were computed using the formula $d_{mi} = \sqrt{(x_i - x_m)^2 + (y_i - y_m)^2}$, where x_m and y_m are the UTM coordinates of the target station and x_i and y_i are the UTM coordinates of the neighboring station. The radius of influence of 25 km included three neighboring stations and the radius of influence of 50 km included nine neighboring stations for each target station (Fig. 1). Although the neighboring station 20 was 52 km away from target station 11 it was maintained in the analysis in order to have the same number of neighboring stations for each target station (Table 1).

Minimum station density guidelines for different climatic and geographic zones have been established by the World Meteorological Organization (WMO

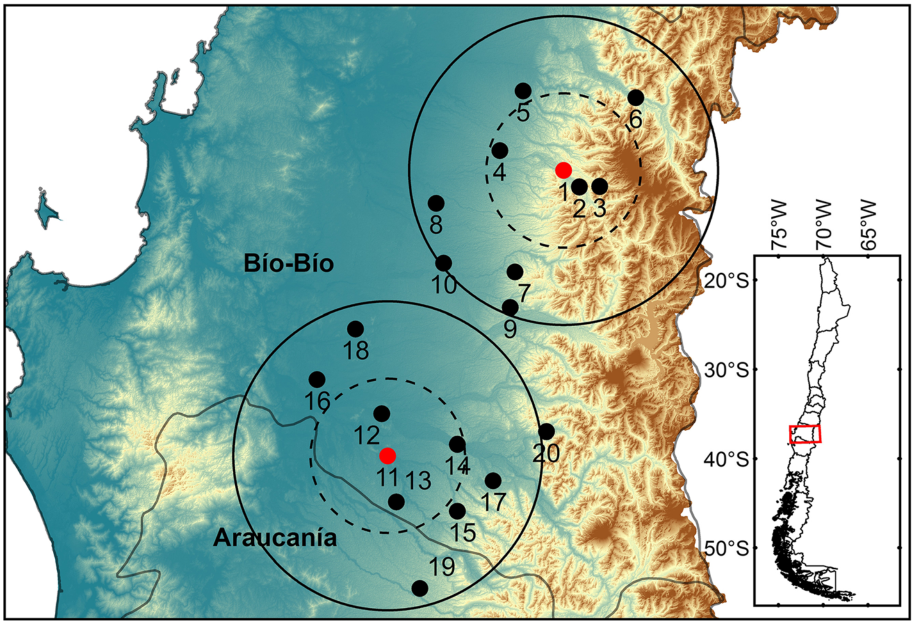

Fig. 1 Geographical location of the meteorological stations. Target stations are represented with a red dot (●) and neighboring stations with a black dot (●). Radius of influence of 25 km (dashed circle) and 50 km (continuous circle)

Table 1 Meteorological stations used for estimating monthly precipitation values

Zone	Station number	Station name	Latitude (S)	Longitude (W)	Elevation (m asl)	*d* (km)	Δ*h* (m)
Andean foothills	1 (Target)	Diguillín	36°52′07″	71°38′33″	670	0	0
	2	Fundo Atacalco	36°54′55″	71°34′58″	730	7	60
	3	Las Trancas	36°54′41″	71°30′34″	1200	13	530
	4	Mayulermo	36°49′02″	71°52′33″	385	22	285
	5	Coihueco Embalse	36°38′27″	71°47′57″	314	29	356
	6	Caracol	36°38′56″	71°23′25″	620	33	50
	7	Las Cruces	37°10′11″	71°48′22″	650	36	20
	8	Pemuco	36°58′35″	72°06′03″	200	43	470
	9	Trupán	37°16′25″	71° 49′ 09″	480	48	190
	10	Cholguán	37°09′02″	72°04′01″	225	49	445
Central valley	11 (Target)	Mulchén	37°43′02″	72°15′01″	130	0	0
	12	San Carlos de Purén	37°35′43″	72°16′37″	150	14	20
	13	Pilguén	37° 51′04″	72°12′49″	300	15	170
	14	Quilaco	37°40′38″	71°59′47″	225	23	95
	15	Poco a Poco	37°52′21″	71°59′17″	620	29	490
	16	Los Ángeles	37°30′02″	72°31′01″	90	34	40
	17	Cerro el Padre	37°46′49″	71°51′38″	400	35	270
	18	Las Achiras	37°20′59″	72°22′54″	125	42	5
	19	Encimar Malleco	38°06′02″	72°07′01″	520	44	390
	20	Quillaileo	37°37′53″	71°40′15″	500	52	370

2008). In the study area, the corresponding network density of meteorological stations is ~ 1.3 stations per 1000 km^2, which is less than the minimum recommended network density for mountainous areas (4 stations per 1000 km^2). Because existing network of climatological stations has a low density to explain the spatial variability of rainfall in mountainous regions at shorter time scales (e.g. hourly and daily) we used a monthly timescale for performing a comparison of approaches for estimating missing monthly precipitation data. Longer-timescales rainfall (e.g. monthly, seasonal and annual) tend to be more

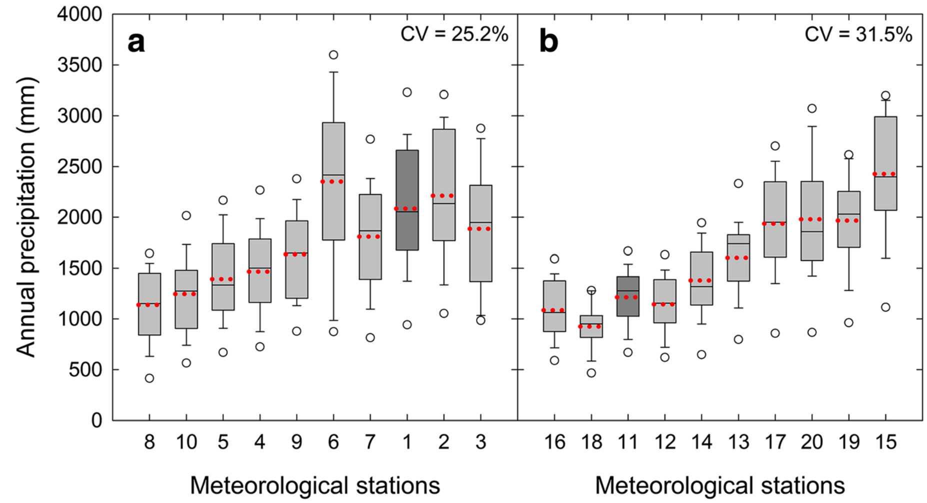

Fig. 2 Variation of the annual precipitation sum for meteorological stations at the Andean foothills (**a**) and Central valley (**b**). Meteorological stations are arranged by elevation, the target stations 1 and 11 are highlighted in dark gray. Red dotted lines indicate the inter-annual precipitation mean

spatially homogeneous than shorter-timescales rainfall (Cheng et al. 2008; Girons-Lopez et al. 2016). In addition, longer-timescale rainfall is of major importance for the evaluation of water availability for management of forest plantations (Álvarez et al. 2013).

Approaches for estimating missing data

We selected the following five reported approaches for estimating missing monthly precipitation data for the two target meteorological stations. All approaches were implemented and tested using the Statistical Analysis System-SAS (SAS Institute Inc. 2009).

Inverse distance weighting (IDW)

Missing data from target station m are determined from the values observed in neighboring stations weighted by the inverse distance between the target and the neighboring stations. The missing data $y_{j(m)}$ at station m, based on the values observed in neighboring stations is given by,

$$y_{j(m)} = \sum_{i=1}^{n} \left(d_{mi}^{-k} \cdot x_{j(i)} \right) / \sum_{i=1}^{n} d_{mi}^{-k} \tag{1}$$

where, n is the number of neighboring stations with information from the month to be estimated, d_{mi} is the Euclidian distance between station i and m, and $x_{j(i)}$ is the observed value at station i, and k is the distance of friction ranging from 1 to 6 (Vieux 2004). In this study, we used a value of $k = 2$ suggested by Teegavarapu (2009).

Modified inverse distance weighting (IDW$_m$)

Elevation has an important influence on precipitation (Golkhatmi et al. 2012; Viale and Garreaud 2015), therefore we used the elevation differences between the target and neighboring stations to adjust IDW estimates. A revised version of the approach proposed by Chang et al. (2005) ensuring that the sum of the weights equals 1 was used. This approach considers not only the effect of Euclidian distances but also differences in elevation. Elevation differences were added to the base IDW formula as;

$$y_{j(m)} = \sum_{i=1}^{n} \left(h_{mi}^{-a} \cdot d_{mi}^{-k} \cdot x_{j(i)} \right) / \sum_{i=1}^{n} \left(h_{mi}^{-a} \cdot d_{mi}^{-k} \right) \tag{2}$$

where h_{mi} is the absolute elevation difference between the target and neighboring stations, and exponent a is a power parameter. Thus, h_{mi} modifies the weights of IDW, prioritizing neighboring stations that are at the same or a close elevation of the target station giving them higher weights during the calculations. Values of the exponents a and k between 1 and 3 were tested, and a value of $a = 1$ and $k = 1$ were selected for computing the missing data.

Correlation coefficient weighting (CCW)

In this approach distance is replaced by Pearson's correlation coefficients. The missing value j in a given month at the target station m is completed as,

$$y_{j(m)} = \sum_{i=1}^{n} \left(r_{mi} \cdot x_{j(i)} \right) / \sum_{i=1}^{n} r_{mi} \tag{3}$$

where r_{mi} is the Pearson's correlation coefficient between the precipitation series of the neighboring station i and the incomplete series of the target station m, $x_{j(i)}$ is the monthly value observed at station i (Teegavarapu 2009).

Multiple linear regression (MLR)

The ordinary least squares method is used to fit a line between the observed data from the target station and several neighboring stations. We used a stepwise selection process to ensure that each station in the final linear model contributes to the accuracy of the estimate without compromising the goodness of fit. The linear model has the following form,

$$y_{j(m)} = \beta_0 \ + \sum_{i=1}^{n} \beta_i \cdot x_{j(i)} \tag{4}$$

where $y_{j(m)}$ is the observed monthly value from the target station m, $x_{j(i)}$ is the observed value in the neighboring station i and β_i are the parameters to be estimated (Freund et al. 2006).

Artificial neural networks (ANN)

An artificial neural network is a computational model inspired structurally and functionally in biological neural networks (Coulibaly and Evora 2007). The architecture of the designed artificial neural network corresponds to a feed forward multilayer perceptron with one hidden layer with ten neurons (see e.g. Dreyfus 2005; Teegavarapu and Chandramouli 2005). The observed values in the neighboring stations are used for the input layer and the estimated values for the target station are obtained for the output layer. To model the transformation of values through the layers a sigmoid function was used for the hidden layer and linear activation was used for the outer layer. Training of the artificial neural network was performed by using the standard error as criterion, applying the Levenberg-Marquardt training algorithm (Khorsandi et al. 2011; Ghuge and Regulwar 2013). The artificial neural network was built, trained and simulated using the SAS NEURAL procedure (SAS Institute Inc. 2009).

Cross-validation and statistical evaluation

Because complete monthly precipitation records were available for all meteorological stations, we simulated missing values using cross-validation for evaluating the accuracy of the estimation approaches. Cross-validation is a technique used for assessing how generalized the results of a statistical analysis are compared to an independent dataset (Chen and Liu 2012). For each target station, data were randomly partitioned into 10 nearly equally sized folds containing 21 or 22 monthly precipitation records (about 10% of total data). Subsequently, 10 estimation and validation iterations were performed, where 9 folds were used to estimate model parameters and the remaining fold was used to validate the method. Refaeilzadeh et al. (2009) reported that 10 folds are the most common because it allows estimations to be made with 90% of the data, producing representative data.

The performance and predictive capability of the approaches for completing missing monthly precipitation records were evaluated using the ratio of the root mean square error to the standard deviation of measured data (RSR).

$$\mathrm{RSR} = \left(\sum_{i=1}^{n} \left(y_{j(m)} - \hat{y}_{j(m)}\right)^2 \Big/ \sum_{i=1}^{n} \left(y_{j(m)} - \bar{y}_m\right)^2 \right)^{1/2} \quad (5)$$

the percent bias (PBIAS).

$$\mathrm{PBIAS} = 100 \cdot \left(\sum_{i=1}^{n} \left(y_{j(m)} - \hat{y}_{j(m)}\right) \Big/ \sum_{i=1}^{n} \left(y_{j(m)}\right) \right) \quad (6)$$

and the Nash-Sutcliffe efficiency (NSE),

$$\mathrm{NSE} = \left(1 - \sum_{i=1}^{n} \left(y_{j(m)} - \hat{y}_{j(m)}\right)^2 \Big/ \sum_{i=1}^{n} \left(y_{j(m)} - \bar{y}_m\right)^2 \right) \quad (7)$$

where $y_{j(m)}$ and $\hat{y}_{j(m)}$ are the observed and estimated expected monthly precipitations at station m during the month j, respectively, $\bar{y}_m$ is the observed mean and n is the number of missing values.

The RSR standardizes the root mean square error (RMSE) using the observed standard deviation. RSR varies from the optimal value of 0, which indicates zero RMSE or residual variation and therefore a perfect estimation, to a large positive value (Moriasi et al. 2007). Percent bias (PBIAS) measures the average tendency of the estimated data to be larger or smaller than their observed counterparts (Moriasi et al. 2007). On the contrary, Nash-Sutcliffe efficiency (NSE) is a normalized statistic that determines the relative magnitude of residual variance compared to measured data variance (Nash and Sutcliffe 1970). NSE indicates how well the plot of observed versus estimated data fits the 1:1 line (Moriasi et al. 2007).

For testing the main and interactive effects of the radius of influence (e.g. number of neighboring stations) and estimation approaches, we applied a two-level factorial design considering the target station as the blocking factor (Quinn and Keough 2002),

$$y_{ijkl} = u + S_i + R_j + A_k + (R \times A)_{jk} + e_{ijkl} \quad (8)$$

where y_{ijkl} is RSR calculated in the l^{th} cross-validation iteration within the k^{th} estimation approach within the j^{th} radius of influence within the i^{th} target station, S_i is the target station (block), R_j is the radius of influence, A_k is the estimation approach, $(R \times A)_{jk}$ is the interaction between radius of influence and estimation approach and e_{ijkl} is the error term. To confirm significant differences between factors (radius of influence or estimation approach) the Student–Newman–Keuls (SNK) test was used (Quinn and Keough 2002). A p-value of 0.05 was considered significant.

Results

Predictive capability of estimation approaches

The ANN and MLR approaches produced the best results for nearly all statistical criteria at both target stations 1 and 11, presenting a lower bias and higher precision compared to the other approaches (Table 2). On the contrary, the CCW approach showed the worst performance in terms of bias and precision for all target stations and radius of influence combinations. The variant IDW$_m$ produced better results than IDW for all target stations and radius of influence combinations, indicating that the inclusion of elevation differences improved the predictive capability. This result was somewhat expected given the existence of a vertical precipitation gradient in this mountainous region.

Estimation approaches showed a decrease in RSR and PBIAS, as well as an increase in NSE, when they were applied to the higher elevation target station 1 compared to the lower target station 11 (Table 2). In comparison to other approaches, IDW and CCW increase RSR and PBIAS and decrease NSE when the radius of influence increased from 25 to 50 km, that is, when the number of neighboring stations increased from 3 to 9.

Comparison of estimation approaches

The ANOVA showed significant differences ($p <$ 0.0001) between estimation approaches (Table 3). Even though ANN and MLR have the lower RSR values (Table 1), the SNK multiple comparison test showed no significant differences with IDW$_m$ (Fig. 3a). Additionally, IDW$_m$ had a more significant difference than IDW and CCW (Fig. 3a). This indicates that including elevation differences into the IDW significantly contributed to the improvement of its performance. The worst RSR values were obtained

Table 2 Predictive capability of the estimation approaches by target station and radius of influence (number of neighboring stations)

Estimation approach	25 km (3 neighboring stations)			50 km (9 neighboring stations)		
	RSR	PBIAS	NSE	RSR	PBIAS	NSE
Target station 1						
IDW	0.167	0.218	0.969	0.162	3.092	0.971
IDW$_m$	0.151	**−0.172**	0.975	0.147	−1.987	0.976
CCW	0.220	10.599	0.946	0.297	18.965	0.905
MLR	0.142	−0.183	0.978	0.138	**−0.190**	0.978
ANN	**0.131**	−0.357	**0.980**	**0.123**	0.356	**0.983**
Target station 11						
IDW	0.269	−11.956	0.911	0.350	−21.490	0.865
IDW$_m$	0.204	−2.811	0.937	0.162	1.993	0.968
CCW	0.270	−13.459	0.916	0.462	−31.408	0.766
MLR	0.191	**−0.900**	0.951	**0.134**	−0.512	**0.980**
ANN	**0.186**	−0.313	**0.953**	0.137	**0.066**	0.978

The best approach for each statistic is highlighted in bold

when applying the CCW approach (Fig. 3a) and its RSR values increased when the radius of influence was increased (Fig. 3b).

The ANOVA showed no significant effect of the radius of influence on RSR values, indicating that similar estimates of missing data can be obtained when considering 3 or 9 neighboring stations. However, as shown in Table 3 a significant interaction between the radius of influence (R) and the estimation approach (A) was detected (Table 3). This seemingly contradictory result is due to the opposite impact of the radius of influence on the method's performance: RSR increases when increasing the radius of influence (Fig. 3b) in IDW and CCW. In contrast, in the other approaches the RSR decreases when the radius of influence increases (Fig. 3b).

Discussion

In this study, we compared five alternative approaches for estimating missing monthly precipitation records in two sectors in south-central Chile with complex terrain. The ANN and MLR showed higher precision and in most cases a lower bias compared to the other approaches. However, the precision (as per RSR) of IDW$_m$ was not significantly different from ANN and MLR, according to the SNK test ($p <$ 0.05). The ANOVA

Table 3 Analysis of variance for estimation approaches

Source	DF	SS	MS	*F*-value	*p*-value
Target station	1	0.2363	0.2363	38.21	< 0.0001
Radius of influence (R)	1	0.0163	0.0163	2.64	0.1057
Estimation approach (A)	4	0.8214	0.2053	33.20	< 0.0001
R × A	4	0.2024	0.0506	8.18	< 0.0001
Error	189	1.1690	0.0062		

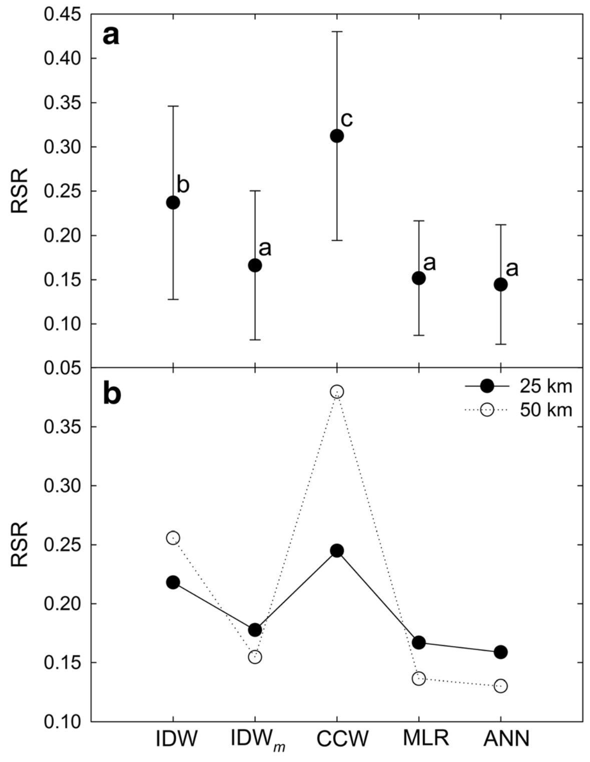

Fig. 3 SNK multiple comparisons for average RSR between estimation approaches (**a**) and interaction between radius of influence and estimation approach (**b**). In the upper panel, error bars represent the standard deviation and the letters (**a**, **b**, **c**) represent group methods that are not significantly different at $\alpha = 0.05$

indicated that the radius of influence in terms of RSR did not significantly affect their predictive capability. However, this result can be explained by the significant interaction between the radius of influence (R) and the estimation approach (A). Therefore, an additional ANOVA was performed to evaluate the effects of the radius of influence on the predictive capability considering only the best three approaches: ANN, MLR and IDW$_m$. For these approaches the radius of influence had a significant effect ($p = 0.036$). Therefore, we conclude that estimates based on nine neighboring stations located within a radius of 50 km are recommended for completing missing monthly precipitation data in these regions with complex topography.

Is there a “best” method?

Past studies have reported that the artificial neural network approach (ANN) was the best at estimating missing monthly precipitation records compared to other approaches (Teegavarapu and Chandramouli 2005; Khorsandi et al. 2011). Coulibaly and Evora (2007) tested different neural networks architectures for completing daily precipitation records and found that the best method was the multilayer perceptron used in our study. In contrast, Alfaro and Pacheco (2000) in Costa Rica and Pizarro et al. (2009) in central Chile found that multiple linear regression (MLR) was the best method for filling in gaps in annual and monthly precipitation series, respectively. Thus, past research showed that ANN and MLR have emerged as robust methods for completing missing data in different geographical and climate settings (Kuligowski and Barros 1998).

Impact of elevation

The inclusion of elevation differences between the target and neighboring stations as a weight modifier to the IDW$_m$ significantly improved its performance. This is in agreement with studies that showed that including elevation differences in IDW had a positive impact on its predictive capability (Chang et al. 2005; Golkhatmi et al. 2012). Recently, Khosravi et al. (2015) used an altitude ratio (elevation of the target station divided by elevation of the neighboring station) to enhance the efficiency of the geographical coordinate method for completing gaps in annual precipitation series.

The Pearson’s correlation coefficients between the target and surrounding stations are presented in Fig. 4. The values are moderately high (typically larger than 0.8) which is somewhat contradictory with the poor performance of the CCW method (e.g., Fig. 3a). We speculate that such a high correlation coefficient is due to the marked annual rainfall cycle, which is common among the seasons in this region and therefore this coefficient has little impact on the estimate of monthly precipitation values. Also there is a negative relationship between Pearson’s correlation coefficients and the elevation differences between the target station and its neighboring stations (Fig. 4). This allows us to conclude that neighboring stations located at similar altitudes to the target station have a close relationship.

Impact of the radius of influence

Even though the ANOVA showed that the radius of influence has a non-significant effect on precision (RSR), this factor interacted significantly with the evaluated approaches (Fig. 3b). An increase of the radius of influence around the target station improved the predictive capability of only three of the evaluated approaches: ANN,

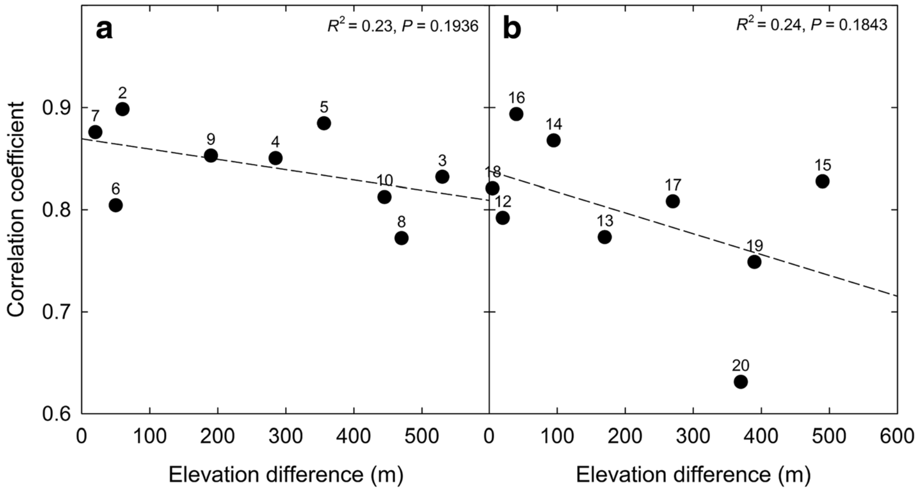

Fig. 4 Relationship between the Pearson’s correlation coefficients and absolute elevation differences between each target station and their neighboring stations: (**a**) target station 1 (Andean foothills) and (**b**) target station 2 (Central valley)

MLR and IDW_m. However, CCW and IDW showed a decreased performance when the radius of influence increased from 25 to 50 km, probably due to the association between decreased precipitations at the target and neighboring stations when distance from the target station increased (Johansson and Chen 2003; Mair and Fares 2011). Chen and Liu (2012) evaluated the IDW for interpolating rainfall data and found that the optimal radius of influence was in most cases up to 10–30 km. They also reported that the interpolation accuracy of this approach could become inferior when the number of considered rainfall stations exceeds the optimal value.

Conclusions

This study found that approaches based on artificial neural networks (ANN), multiple linear regression (MLR) and IDW_m had the best performance in two sectors located in central-south Chile with a complex topography. Inclusion of elevation differences and Euclidian distances between targets and neighboring stations as weight modifier in the IDW_m significantly improved overall estimates. Because the predictive capability of the three best approaches was significantly affected by the number of neighboring stations (radius of influence), we conclude that estimates based on nine neighboring stations located within a radius of 50 km are needed for completing missing monthly precipitation data.

Abbreviations

A × R: Interaction between estimation approach and radius of influence; A: Estimation approach; ANN: Artificial neural networks; ANOVA: Analysis of variance; CCW: Coefficient of correlation weighting; CV: Coefficient of variation; DGA: Dirección General de Aguas; ENOS: El Niño Southern Oscillation; IDW: Inverse distance weighting; IDW_m: Modified inverse distance weighting;; m asl: Meters above sea level; MLP: Multilayer perceptron; MLR: Multiple linear regression; NSE: Nash-Sutcliffe efficiency; PBIAS: Percent bias; R: Radius of influence; RMSE: Root mean square error; RSR: Root mean square error to the standard deviation; SAS: Statistical analysis system; SNK: Student–Newman–Keuls test; UTM: Universal Transverse Mercator

Funding

This work was supported by the National Fund for Scientific and Technological Development (FONDECYT) [Project 1151050] and the first author gratefully acknowledges funding from Chile's Education Ministry through the program MECESUP2 [Project UCO0702].

Availability of data and materials

The data used in this study are available in public repositories of the Dirección General de Aguas (DGA; available at http://snia.dga.cl/BNAConsultas/reportes).

Authors' contributions

AB collected the data and performed the statistical analysis. AB and GT drafted the manuscript. GT and RG revised it critically for important intellectual content. AB, GT and RG gave final approval of the version to be published.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Escuela de Graduados, Facultad de Ciencias Forestales y Recursos Naturales, Universidad Austral de Chile, Box 567 Valdivia, Chile. [2]Departamento de Ciencias Forestales, Facultad de Ingeniería Forestal, Universidad del Tolima, Box 6299 Ibagué, Colombia. [3]Instituto de Bosques y Sociedad, Facultad de Ciencias Forestales y Recursos Naturales, Universidad Austral de Chile, Box 567 Valdivia, Chile. [4]Departamento de Geofísica, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Santiago, Chile.

Received: 6 February 2018 Accepted: 3 July 2018
Published online: 30 July 2018

References

Ahrens B (2006) Distance in spatial interpolation of daily rain gauge data. Hydrol Earth Syst Sci 10:197–208

Ahumada R, Rotella A, Slippers B, Wingfield MJ (2013) Pathogenicity and sporulation of *Phytophthora pinifolia* on *Pinus radiata* in Chile. Australas Plant Pathol 42(4):413–420

Alfaro R, Pacheco R (2000) Aplicación de algunos métodos de relleno a series anuales de lluvia de diferentes regiones de Costa Rica. Tóp Meteor Oceanogr 7(1):1–20

Álvarez J, Allen HL, Albaugh TJ, Stape JL, Bullock BP, Song C (2013) Factors influencing the growth of radiata pine plantations in Chile. Forestry 86:13–26

Cannell MGR, Cruz RVO, Galinski W, Cramer WP (1995) Climate change impacts on forests. In: Watson RT, Zinyowera MC, Moss RH (eds) Climate change 1995: impacts, adaptations and mitigations of climate change, working group II. Cambridge University Press, Cambridge, pp 95–130

Chang CL, Lo SL, Yu SL (2005) Interpolating precipitation and its relation to runoff and non-point source pollution. J Environ Sci Health Part A 40:1963–1973

Chen FW, Liu CW (2012) Estimation of the spatial rainfall distribution using inverse distance weighting (IDW) in the middle of Taiwan. Paddy Water Environ 10:209–222

Cheng K, Lin Y, Liou J (2008) Rain-gauge network evaluation and augmentation using geostatistics. Hydrol Process 22:2554–2564

Codesido V, Merlo E, Fernández-lópez J (2005) Variation in reproductive phenology in a *Pinus radiata* D. Don seed orchard in northern Spain. Silvae Genet 54(4–5):246–256

Coulibaly P, Evora ND (2007) Comparison of neural network methods for infilling missing daily weather records. J Hydrol 341:27–41

Dai Z, Amatya DM, Sun G, Trettin CC, Li C, Li H (2011) Climate variability and its impact on forest hydrology on South Carolina coastal plain, USA. Atmosphere 2:330–357

Dreyfus G (2005) Neural networks: methodology and applications. Springer-Verlag, Heidelberg

Falvey M, Garreaud R (2007) Wintertime precipitation episodes in Central Chile: associated meteorological conditions and orographic influences. J Hydrometeorol 8:171–193

Freund RJ, Wilson WJ, Sa P (2006) Regression analysis: statistical modeling of a response variable, 2nd edn. Academic Press, San Diego

Garreaud R (2009) The Andes climate and weather. Adv Geosci 22:3–11

Ge ZM, Kellomäki S, Zhou X, Wang KY, Peltola H, Väisänen H, Strandman H (2013) Effects of climate change on evapotranspiration and soil water availability in Norway spruce forests in southern Finland: an ecosystem model based approach. Ecohydrol 6:51–63

Gerding V, Schlatter JE (1995) Variables y factores del sitio de importancia para la productividad de *Pinus radiata* D. Don en Chile. Bosque 16(2):39–56

Ghuge HK, Regulwar DG (2013) Artificial neural network method for estimation of missing data. Int J Adv Tech Civil Eng 2(1):1–4

Girons-lopez M, Wennerström H, Nordén L, Seibert J (2016) Location and density of rain gauges for the estimation of spatial varying precipitation. Geogr Ann A 97(1):167–179

Golkhatmi NS, Sanaeinejad SH, Ghahraman B, Pazhand HR (2012) Extended modified inverse distance method for interpolation rainfall. Int J Eng Invent 1(3):57–65

Huber A, Trecaman R (2002) The effect of the inter-annual variability of rainfall on the development of *Pinus radiata* (D. Don) plantations in the sandy soil zones of VIII region of Chile. Bosque 23(2):43–49

Johansson B, Chen D (2003) The influence of wind and topography on precipitation distribution in Sweden: statistical analysis and modelling. Int J Climatol 23:1523–1535

Khorsandi Z, Mahdavi M, Salajeghe A, Eslamian S (2011) Neural network application for monthly precipitation data reconstruction. J Environ Hydrol 19:1–12

Khosravi G, Nafarzadegan AR, Nohegar A, Fathizadeh H, Malekian A (2015) A modified distance-weighted approach for filling annual precipitation gaps: application to different climates of Iran. Theor Appl Climatol 119(1):33–42

Kuligowski RJ, Barros AP (1998) Using artificial neural networks to estimate missing rainfall data. J Am Water Resour As 34(6):1437–1447

Mair A, Fares A (2011) Comparison of rainfall interpolation methods in a mountainous region of a tropical island. J Hydrol Eng 16(4):371–383

Montecinos A, Aceituno P (2003) Seasonality of the ENSO-related rainfall variability in Central Chile and associated circulation anomalies. J Clim 16: 281–296

Moriasi DN, Arnold JG, van Liew MW, Bingner RL, Harmel RD, Veith TL (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans ASABE 50(3):885–900

Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models: part 1. A discussion of principles. J Hydrol 10(3):282–290

Pizarro R, Ausensi P, Aravena D, Sangüesa C, León L, Balocchi F (2009) Evaluación de métodos hidrológicos para la completación de datos faltantes de precipitación en estaciones de la región del Maule, Chile. Aqua-LAC 1(2):172–185

Quinn G, Keough M (2002) Experimental design and data analysis for biologists. Cambridge University Press, Cambridge

Ramos-Calzado P, Gómez-Camacho J, Pérez-Bernal F, Pita-López MF (2008) A novel approach to precipitation series completion in climatological datasets: application to Andalusia. Int J Climatol 28:1525–1534

Refaeilzadeh P, Tang L, Liu H (2009) Cross Validation. In: Ling L, Tamer ÖM (eds) Encyclopedia of database systems. Springer, New York, pp 532–538

Sands PJ, Landsberg JJ (2002) Parameterisation of 3-PG for plantation grown *Eucalyptus globulus*. For Ecol Manag 163(1–3):273–292

Statistical Analysis System Institute Inc (2009) User's Guide, 2nd edn Version 9.2 for Windows. Statistical Analysis System Institute Inc, Cary

Teegavarapu RSV (2009) Estimation of missing precipitation records integrating surface interpolation techniques and spatio-temporal association rules. J Hydroinf 11(2):133–146

Teegavarapu RSV (2012) Spatial interpolation using nonlinear mathematical programming models for estimation of missing precipitation records. Hydrol Sci J 57(3):383–406

Teegavarapu RSV, Chandramouli V (2005) Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. J Hydrol 312:191–206

Vasiliev IR (1996) Visualization of spatial dependence: an elementary view of spatial autocorrelation. In: Arlinghaus SL (ed) Practical handbook of spatial statistics. CRC Press, Boca Raton, pp 17–30

Viale M, Garreaud R (2014) Summer precipitation events over the western slope of the subtropical Andes. Mon Weather Rev 142:1074–1092

Viale M, Garreaud R (2015) Orographic effects of the subtropical and extratropical Andes on upwind precipitating clouds. J Geophys Res Atmos 120:4962–4974

Vieux BE (2004) Distributed hydrologic modeling using GIS, 2nd edn. Kluwer Academic Publishers, Dordrecht

WMO (2008) Guide to hydrological practices, volume I: hydrology – from measurement to hydrological information, 6th edn. World Meteorological Organization, Geneva

Xia Y, Fabian P, Stohl A, Winterhalter M (1999) Forest climatology: estimation of missing values for Bavaria. Germany Agric For Meteorol 96(1–3):131–144

Xu J, Lu J, Bao F, Evans R, Downes G (2013) Climate response of cell characteristics in tree rings of *Picea crassifolia*. Holzforschung 67(2):217–225