Evaluation of infilling methods for time series of daily precipitation and temperature: The case of the Ecuadorian Andes

L. Campozano[1,4], E. Sanchez[1,2], A. Aviles[1,3], E. Samaniego[1,2]

[1] Departamento de Recursos Hídricos y Ciencias Ambientales, Universidad de Cuenca, Cuenca, Ecuador.

[2] Facultad de Ingeniería, Universidad de Cuenca, Av. 12 de Abril s/n, Cuenca, Ecuador.

[3] Facultad de Ciencias Químicas, Escuela de Ingeniería Ambiental, Universidad de Cuenca, Av. 12 de Abril s/n, Cuenca, Ecuador.

[4] Laboratory for Climatology and Remote Sensing (LCRS), Faculty of Geography, University of Marburg, Deutschhausstraße 10, D-35032 Marburg, Germany.

Corresponding author: lenin_camp@yahoo.com

Fecha de recepción: 4 de agosto 2013 - Fecha de aceptación: 28 de enero 2014.

ABSTRACT

Continuous time series of precipitation and temperature considerably facilitate and improve the calibration and validation of climate and hydrologic models, used inter alia for the planning and management of earth's water resources and for the prognosis of the possible effects of climate change on the rainfall-runoff regime of basins. The goodness-of-fit of models is among other factors dependent from the completeness of the time series data. Particular in developing countries gaps in time series data are very common. Since gaps can severely compromise data utility this research with application to the Andean Paute river basin examines the performance of 17 deterministic infill methods for completing time series of daily precipitation and mean temperature. Although sophisticated approaches for infilling gaps, such as stochastic or artificial intelligence methods exist, preference in this study was given to deterministic approaches for their robustness, easiness of implementation and computational efficiency. Results reveal that for the infilling of daily precipitation time series the weighted multiple linear regression method outperforms due to considering the ratio of the Pearson correlation coefficient to the distance, giving more weight to both, highly correlated and nearby stations. For mean temperature, the climatological mean of the day was clearly the best method, most likely due to the scarcity of weather stations measuring temperature, and because the few available stations are located at different elevations in the landscape, suggesting the need to address in future studies the impact of elevation on the interpolation.

Keywords: Infilling, deterministic infill methods, time series, daily rainfall, mean daily temperature, Andean Paute river basin.

RESUMEN

Series continuas de precipitación y temperatura facilitan y mejoran considerablemente la calibración y validación de modelos hidrológicos y climáticos, utilizados entre otras cosas, para la planificación y manejo de recursos hídricos y el pronóstico de los posibles efectos del cambio climático en el regimen lluvia-escorrentia de las cuencas hidrográficas. La bondad de ajuste de los modelos está entre los factores que dependen de la continuidad de las series temporales. En países en vías de desarrollo los vacíos en las series temporales de variables climáticas es común. Ya que los vacíos en las series temporales pueden comprometer severamente la utilidad de los datos, este estudio aplicado en la cuenca del río Paute en los Andes Ecuatorianos, examina el desempeño de 17 métodos determinísticos de relleno de datos diarios de las variables precipitación y temperatura media. A pesar de la existencia de métodos de relleno más sofisticados como métodos estocásticos o métodos de inteligencia artificial, en este estudio se dio preferencia a métodos determinísticos por su robustez, facilidad de

implementación, y eficiencia computacional. Los resultados revelan que para rellenar series temporales de precipitación diaria, el método de regresión lineal múltiple ponderada es el mejor, debido a la consideración de la razón entre el coeficiente de correlación de Pearson y la distancia con respecto a otras estaciones como factor de ponderación, dando mayor importancia a las estaciones más cercanas altamente correlacionadas. Para temperatura, la media climatológica del día fue claramente el mejor método, posiblemente debido a la escacez de datos de estaciones cercanas localizadas también en elevaciones diferentes, sugiriendo la necesidad de considerar en futuros estudios el impacto de la elevación en la interpolación de datos.

Palabras clave: Relleno de datos, métodos determinísticos de relleno, series temporales, precipitación diaria, temperatura media del día, cuenca del río Paute Andean.

1. INTRODUCTION

As stated by Harvey *et al.* (2010) time series of rainfall and river flow are vitally important assets, critical to the sustainable management of water resources and serving as indicators of past hydrological variability and fundamental contributors to hydrological models for future behavior prediction. The completeness of records is a crucial component of their utility. Even very short gaps preclude the calculation of important summary statistics, such as monthly runoff totals or *n*-day minimum flows, and inhibit the analysis and interpretation of flow variability. In fact hydrologic studies require complete time series data preferably collected over a long period of several stations, especially in large basin studies, spread out across the whole region of interest (Yozgatligil *et al.*, 2013). According to this and other authors (Beauchamp *et al.*, 1989; Hanson *et al.*, 2004; Gould *et al.*, 2008; Mwale *et al.*, 2012) having complete datasets one can form inferences based on equally spaced observations, preserving the statistical information of the system. As stated by Ng and Panu (2010) incomplete datasets raise the level of complexity and uncertainty in modeling. Encountering missing information in meteorological time series is inevitable, particular in developing and economic emerging countries. If data gaps are big, incomplete time series may hide the pattern of the data, and they may considerably distort the results of any statistical analysis. To avoid the effect of missing data on the results of climatological studies it is essential to handle meticulous the optimal infilling of missing data.

The National Meteorological and Hydrological Institute of Ecuador (INAMHI) collects since August 1961 systematically meteorological and surface hydrological data. Unfortunately political instability and economic constraints resulted at one hand that data monitoring is discontinuous, often of short duration, and that parallel other governmental and non-governmental institutions started collecting rainfall and temperature data. This situation is very prevalent for developing and economic emerging countries as stated by Gyau-Boake and Schultz (1994), Ilunga and Stephenson (2005), and Adeloye (2011). The main factors responsible for gaps and inconsistencies in available time series data include temporary absence of observers, use of different equipment, and the poor and infrequent calibration of sensors, the malfunctioning of equipment, among other factors, which can be very different from data collector to data collector. Dispersion in efforts, absence of coordination, poor follow-up and inconsistency in data storage are responsible that the time series of meteorological data in Ecuador are incomplete, often of short duration, and heterogeneous. As stated by Adeloye (1996, 2011) the low quality and incompleteness of time series data strongly hinders and effects the planning, operation and management of water resources systems, hindering the calibration and validation of modeling tools and their use for predicting the hydrological responses under changed climate conditions. A way to redress the issue in existing time series data is infilling the gaps using one of the many available techniques. Of course, it is evident that countries with a low monitoring profile in future should give more active policy attention to consistent monitoring, data storage and sharing.

The purpose of this paper is exploring the performance of 17 deterministic infilling methods with respect to completing the time series of daily precipitation and mean daily temperature, collected in different stations in the Andean Paute river basin. Performance is measured using 3 metrics. The paper is organized as follows: Section 2 gives an overview of infilling methods, with justification why for

this study only deterministic methods were selected. Section 3 is devoted to a detailed description of the study area, the infill methods and time series used. Results and discussions are presented in the Section 4, and Section 5 presents conclusions and prospects for future research.

2. INFILLING METHODS

Fifty three (53) precipitation stations are available in the Paute river basin, but after quality control only the time series data of 14 stations met the acceptance criteria. This clearly indicates that major effort is needed to better coordinate between the different stakeholders involved in data monitoring, so that as from today more and better information is available that is useable as to study for example the effect of climate change on the basin hydrology. According to the World Meteorological Organization (WMO) for model calibration and validation at least a time series of 30 years is needed, for which often the record period 1960 to 1990 is selected. The infill methods described in literature can be grouped in three major classes: (i) the deterministic, (ii) stochastic and (iii) artificial intelligence methods. A brief description of each group is given in the following.

2.1. Deterministic methods

Deterministic methods are mathematical models that always produce the same output from a given initial condition, and neither contemplates the existence of randomness nor attribute a probability of occurrence. A simple deterministic method is the imputation of the arithmetic mean of the corresponding day value of the stations near the station where infill is required. This method is suitable for areas where the variable under consideration possesses small spatial variability (Ramos-Calzado *et al.*, 2008). Dingman (1994) suggests selecting the stations based on sound meteorological judgment and expertise. The inverse distance weighting method (IDW) is probably the most commonly used to estimate missing data in hydrology and geographical sciences (Di Piazza *et al.*, 2011). The success of this method depends on the existence of a positive spatial autocorrelation (Vasiliev, 1996). Therefore one problem of the IDW method is the arbitrary selection of time series data from neighboring stations (Di Piazza *et al.*, 2011). Besides, Xia *et al.* (1999a) examined six methods for estimating missing climatological data: simple arithmetic averaging, inverse distance interpolation, normal ratio method, single best estimator, multiple regression analysis and universal kriging. The multiple regression analysis predominantly gave the best estimation for temperature, water vapor pressure and wind speed. On the contrary, Wagner *et al.* (2012) compared the performance of the spatial interpolation approach to infill precipitation time series data applying seven methods, including the deterministic Thiessen polygon method, the stochastic, and statistical and geostatistical approaches. From this evaluation the methods based on regression showed the best performance.

2.2. Stochastic methods

Stochastic methods provide probabilistic estimates of the outcome, in contrast to deterministic methods. There are many studies that have dealt with these methods for infilling of climate time series data (Ashraf *et al.*, 1997; Di Piazza *et al.*, 2011; Teegavarapu and Chandramouli, 2005; Yozgatligil *et al.*, 2013). Most of these studies focused on the comparison between the deterministic and stochastic infill methods. Findings revealed that depending from the datasets sometimes one method outperforms the other; however the high computational cost of stochastic methods has been highlighted as a major constraint. Also Ashraf *et al.* (1997) compared deterministic and stochastic methods (Inverse Distance Weighted, Kriging and Co-Kriging) to infill gaps in precipitation time series, and found that stochastic methods in general perform better. At the other hand, Teegavarapu and Chandramouli (2005) worked on filling gaps in precipitation datasets, by improving the IDW method and the development of databased models using concepts of artificial neural network (ANN) and stochastic interpolation by Kriging. Results showed the advantages of these methods when compared to traditional schemes. At the other hand, Di Piazza *et al.* (2011) compared different spatial interpolation techniques to create

complete monthly precipitation series. The algorithms used for interpolation were the IDW method, regression methods, ANNs and geostatistical models such as ordinary and residual kriging. Results revealed that the linear regression residuals between precipitation and elevation, incorporated into an ordinary kriging model gave the best results. Some authors used the elevation in stochastic methods to produce better results in the interpolation of climate variables. Because the lack of a significant correlation between the annual precipitation depth and the elevation, in this study elevation was not considered as additional parameter not withstanding the considerable elevation differences present in the Paute river basin. Further, Diodato and Ceccarelli (2005) compared three methods of interpolation, IDW, linear regression and co-kriging, concluding that the latter is the best method when considering the elevation. One of the main limitations of spatial interpolation methods used to infill climate time series is that they neglect the space-time structure of the time series (Di Piazza *et al.*, 2011).

2.3. Artificial intelligence methods

Modern methods of artificial intelligence are also widely used for infilling weather data time series, especially ANNs and support vector machines (SVM) (Mwale *et al.*, 2012; Khalil *et al.*, 2001; Mileva-Boshkoska and Stankovski, 2007). These methods have a complex mathematical formulation, consequently more difficult to be implemented, and require intensive calculations with high computational cost; however they are effective especially when dealing with non-linear relations (Dawson *et al.*, 2002). For example, Yozgatligil *et al.* (2013) used methods based on ANNs and multiple imputation strategies (MI) to fill monthly series of precipitation and temperature. Results showed that the creation of a Monte Carlo Markov Chain algorithm based on expectation-maximization (EM-MCMC) performed better and reduced the uncertainty of the results. Further, Khalil *et al.* (2001) presented an alternative approach to fill streamflow time series data using ANNs and a data grouping approach. The results indicated that ANN models showed good performance infilling streamflow time series data when performing a grouping of seasonal periods. Coulibaly and Evora (2007) conducted a comparison of six different types of ANN for filling daily time series of precipitation and temperature. The results clearly indicated that the multilayer perceptron (MLP) network was the most effective in filling missing daily precipitation and daily maximum and minimum temperature data. Also, a combination of ANN and stochastic methods can be found in Demyanov *et al.* (1998), where residues of the ANN are analyzed applying an ordinary Kriging method. Results showed good performance in representing the statistical characteristic of climate variables, besides representing adequately large-scale structure, periodicity and small-scale effects. Recently, Kim and Pachepsky (2010) presented a new technique to reconstruct missing daily precipitation data by combining ANNs with regression trees (RT). The RT+ANN method significantly improved accuracy and was more robust when compared with RT and ANN alone. This method was also more accurate and robust in streamflow predictions with reconstructed precipitation. Despite having multiple methods to infill data gaps in climate time series, this paper focuses on the evaluation and comparison of deterministic methods because they are computationally efficient, robust especially when dealing with extreme events or high spatial variability as the Andes (Ramos-Calzado *et al.*, 2008), and are easy to implement compared with more sophisticated techniques. These strengths make them very useful especially in countries with poor monitoring tradition where gaps of information in climatological and hydrologic time series are ubiquitous.

3. MATERIALS AND METHODS

3.1. Study area and data

The performance of the 17 deterministic infilling techniques was assessed using time series data of the Paute river basin. The catchment is of high economic and ecologic value covering an area of 6148 km^2 including the lower part eastwards towards the Amazon plateau. The basin is located in the inter-Andean depression between the western and the eastern sides of the cordillera in south Ecuador, draining into the Amazon basin. Its elevation ranges between 500 and 4680 m above mean sea level. Around 40% of the basin is covered with Páramo, a very fragile ecosystem (Célleri *et al.*, 2007). The

Paute river basin fosters several hydroelectric projects, such as the Amaluza (1075 MW), El Labrado y Chanlud (38,4 MW), Mazar (162,6 MW) and Sopladora (500 MW), generating roughly 50% of the national hydroelectricity production (Salazar and Rudnick, 2008). The southern region of Ecuador depends directly upon the hydro-ecological services of the basin.

The climatology of the basin is very diverse, presenting the typical inter-Andean bi-modal seasonality of rainfall in the central part, with peaks in March-April and October-November, and a profound dry season in July and August (Bendix and Lauer, 1992; Mejía *et al.*, 1996; Célleri *et al.*, 2007). The seasonality is mainly convective in nature due to the interplay of the intertropical convergence zone responsible for the rainy seasons and the Walker circulation accountable for the tendency to subsidence and high pressure inhibiting cloud formation in the dry season. Amazon influence to the east of the basin, accompanied by moist easterlies, shifts the rainy season to June-July-August (JJA) (Laraque *et al.*, 2007). The remainder of the year is continuously rainy. To the higher elevations, on the western side of the basin a weak rainfall peak in June-July is present, departing from the typical inter-Andean seasonality (Bendix and Lauer, 1992), advective in nature with lower rain rates in contrast to higher amounts of rainfall during the convective peaks in the March-April-May (MAM) and September-October-November (SON) seasons. During ENSO periods, precipitation in the costal plains is strongly influenced, but no link between ENSO and the annual rainfall in the inter-Andean valleys, as the Paute basin, has been detected (Rossel and Cadier, 2009; Célleri *et al.*, 2007).

Daily precipitation data of 53 stations and mean temperature of 13 weather stations, for the period January 1981 to December 2010 were used for this study. The stations are operated by the National Institute of Meteorology and Hydrology, INAMHI, and the datasets made available by El Grupo de las Ciencias de la Tierra y del Ambiente of Universidad de Cuenca.

3.2. Methods

Selection of stations

In this study the quality of time series was checked on the following three criteria: (i) gaps in the time series, (ii) homogeneity of the time series, and (iii) the importance of the station.

Gaps quantification

The first criterion applied to control the quality of the data was checking the gaps in the time series of precipitation and temperature. The percentage of gaps threshold for the whole study period was limited to 25% of the data. Recommendations found in literature advise lower thresholds, as 10% (Baddour and Kontongomde, 2007), but due to the scarcity of the data the presented value was considered. The next step in the assessment of the quality is the detection of outliers. For outlier detection a logarithmic transformation of the daily data was performed and data exceeding the range $Q_1 - 3(Q_3 - Q_1) \leq x \leq Q_3 + 3(Q_3 - Q_1)$ was considered outlier (Montgomery and Runger, 2011). In the proposed range, Q_1 is the first quartile or the 25^{th} percentile, and Q_3 is the third quartile or the 75^{th} percentile. The datasets were logarithmic transformed as to bring the distribution closer to a normal distribution, because the histogram of daily precipitation data presented a positive skewness.

Homogeneity check

The second applied quality criterion is assessment of the homogeneity of the data time series. For considering homogeneous a climate time series, it is necessary that variations in the data are caused only by variations in climate rather than other external factors (Aguilar *et al.*, 2003). External factors that might cause inhomogeneities in data are: monitoring stations relocations, changes in instrumentation, changes of the surroundings, instrumental inaccuracies, and changes of observational and calculation procedures (Costa and Soares, 2008). Especially for stations in remote locations with limited access to retrieve information, checking for homogeneity of the data is strongly recommended. The homogeneity was checked on monthly data, since daily time series can be very noisy. The package RHTestsV3 (Wang *et al.*, 2007; Wang, 2008a; Wang, 2008b) was used to detect and adjust for multiple change points that exist in time series data possessing first order autoregressive errors. The change point detection test is based on the penalized maximal F-test, which allows the time series

being tested to have a linear trend throughout the whole period of the data record. The trend estimates are robust to the first-order autocorrelation of the respective series, because the annual cycle, linear trend, and lag-1 autocorrelation of the time series were estimated in tandem through iterative procedures (Wan *et al.*, 2010).

Importance of station

Besides the amount of gaps and homogeneity, another used quality criteria is the spatial importance of the station, which considers the relative position of the station with respect to others. Stations located near the center of the basin, where normally the density of stations is greater than near the border, are less important than those located close to the border of the basin. Furthermore, stations located near the border are important in the case of gridded interpolation based data generation. A plus sign (+) was assigned to the most important stations by location.

Deterministic methods for gaps infilling

Deterministic methods for infilling gaps were applied to daily time series of precipitation and mean daily temperature. Table 1 presents a summary of the applied infill methods with their respective abbreviation. A brief outline of each method is given in the following.

Table 1. List of the 17 deterministic infilling methods with their respective code.

#	Infill method	Code
1	Average value nearer stations by distance	AVNSd
2	Average value nearer stations by R	AVNSR
3	Climatological mean of the day	CMD
4	Nearest neighbour value by distance	NNVd
5	Nearest neighbour value by R	NNVR
6	Inverse distance weight power 1	IDW+1
7	Inverse distance weight power 2	IDW+2
8	Inverse distance weight power ½	IDW+1/2
9	Linear regression with the nearest station by distance	LRNd
10	Linear regression with the nearest station by R	LRNR
11	Linear regression with the nearest station by R/d ratio	LRNR/d
12	Multiple linear regression weighted by distance power -1	MLRWd-1
13	Multiple linear regression weighted by distance power -2	MLRWd-2
14	Multiple linear regression weighted by R power 1	MLRWR+1
15	Multiple linear regression weighted by R power 2	MLRWR+2
16	Multiple linear regression weighted by R/d ratio power 1	MLRWR/d+1
17	Multiple linear regression weighted by R/d ratio power 2	MLRWR/d+2

Legend: R = Pearson coefficient; d = distance

Average value nearer stations by distance d (AVNSd). Missing data are obtained by arithmetically averaging data of the closest weather stations around the station of interest (Xia *et al.*, 1999b; Bennett *et al.*, 2007), applying Eq. 1.

$$V_{est} = \frac{\sum_{i=1}^{n} v_i}{n} \qquad \text{Eq. 1}$$

Where V_{est} is the estimated value of the missing data, v_i is the value of the same parameter at the i^{th} nearest weather station and n is the number of the nearest stations, from which information was used for the estimation of the missing value.

Average value nearer stations by the Pearson coefficient R (AVNSR). The Pearson correlation coefficient, R, is a measure of the linear relationship between two random variables. R is an index that can be used to measure the degree of relationship of two variables. The R values are converted into weights by using the weighting formula for the i^{th} R; where n is the total number of stations (Dumedah and Coulibaly, 2011; Bennett *et al.*, 2007). Estimation of the missing variable is computed as the weighted sum of the available variable-values (n) and their respective weights (w) (Eqs. 2 and 3).

$$w_i = \frac{R_i}{\sum_{j=1}^{n} R_j} \qquad \text{Eq. 2}$$

$$V_{est} = \sum_{j=1}^{n} V_j * w_i \qquad \text{Eq. 3}$$

Climatological mean of the day (CMD). This method uses the long-term average value of the same day of interest. V_{est} is the estimated value, V_i is the value of the variable for the i^{th} day of year j, and T is the number of years data are available (Narapusetty *et al.*, 2009). The output of Eq. 4 is simply a temporal average of the j^{th} day value.

$$V_{est,i} = \frac{\sum_{j=1}^{T} v_{ij}}{T} \qquad \text{Eq. 4}$$

Nearest neighbor value by d and R (NNV). Observations from neighboring stations are used in missing data reconstruction. It has been proposed to use geometrical distances to the stations and to apply the data from the closest station. However, the method gives poor results when the climate variable under analysis has a high spatial variability. For this reason, in this study a modified version of the criterion was used imposing conditions on the correlation coefficient (Ramos-Calzado *et al.*, 2008).

Inverse distance weight (IDW). The inverse distance weighting method estimates the missing value of an observation, V_{est}, using the observed values at other stations applying Eq. 5 (Teegavarapu *et al.*, 2009; Bennett *et al.*, 2007).

$$V_{est} = \frac{\sum_{i=1}^{n} V_i d_i^{-k}}{\sum_{i=1}^{n} d_i^{-k}} \qquad \text{Eq. 5}$$

Where V_{est} is the estimated value; n the number of stations; V_i the observation at station i, d_i the distance from the station to the i^{th} station, and k the power of distance, referred to as a friction distance ranging between 0,5 and 2.

Linear regression with the nearest station by d, R and by the ratio R/d (LRN). In this method the nearer station is selected by the closets distance, the higher Pearson coefficient, or the higher ratio of the Pearson coefficient and the distance (Dumedah and Coulibaly, 2011; Xia *et al.*, 1999b; Villazón and Willems, 2010). The value in the nearer station is defined V_i. Then a linear fit between the target station V_{est} and the selected station is calculated to obtain the parameters a and b.

$$V_{est} = b * V_i + a \qquad \text{Eq. 6}$$

Multiple linear regression weighted by d, R and by the ratio R/d (MLRW). In multiple linear regression, a linear combination of two or more predictor variables is used to explain the variation in a response (Abatzoglou *et al.*, 2009). In essence, the additional predictors are used to explain the variation in the response not explained by a simple linear regression. Several powers are used to identify the best performing one. The weight used is the distance, the Pearson coefficient or the ratio of the Pearson coefficient to the distance, with power k. The coefficients a_i and b_i are calculated as a linear fit between V_{est} and V_i (see Eq. 7).

$$V_{est} = \frac{\sum_{i=1}^{n} w_i^k * (b_i * V_i + a_i)}{\sum_{i=1}^{n} w_{mi}^k} \quad \text{Eq. 7}$$

For the methods using data of several stations four stations (n = 4) were used for the infilling of precipitation, assuming that one station is representative of one cardinal direction. This criterion might be important in case of directionality of precipitation. For temperature two stations (n = 2) were considered because of the scarcity of stations at diverse elevations, and taking into account that too distant stations could affect the prediction of closer stations.

Statistical metrics for the evaluation of methods

For the performance assessment of the 17 infilling methods 1,5% of the data in the time series was sampled at random and considered for further comparison with infilled values (Kotsiantis *et al.*, 2006). The comparison between infilled and observed values was quantified by the mean error (ME), the mean absolute error (MAE) and the root mean square error (RMSE). For precipitation, these metrics of error were quantified for respectively bi-modal and uni-modal stations. For temperature all stations were considered together because the seasonality of temperature is similar in the bi- and uni-modal regime. The three metrics of error (Eqs. 8, 9 and 10) were valuated seasonally and annually.

$$ME_k = \frac{\sum_{i=1}^{n} I_{ki} - Obs_i}{n} \quad \text{Eq. 8}$$

$$MAE_k = \frac{\sum_{i=1}^{n} |I_{ki} - Obs_i|}{n} \quad \text{Eq. 9}$$

$$RMSE_k = \sqrt{\frac{\sum_{i=1}^{n} (I_{ki} - Obs_i)^2}{n}} \quad \text{Eq. 10}$$

In these equations ME_k , MAE_k , $RMSE_k$ are the mean error, the mean absolute error and the root mean square error, respectively for each of the 17 infilling methods (k varies from 1 to 17). I_{ki} is the i^{th} infilled value of method k, n is the number on sampled values not considering gaps, and Obs_i is the observed value corresponding to the i^{th} value.

4. RESULTS AND DISCUSSIONS

4.1. Selection of stations

For precipitation the cumulative percentage of gaps in the period January 1981 until December 2010 for the 14 selected stations out of a total of 53 stations is presented in Fig. 1. This figure clearly reveals that for some stations an increase in percentage of gaps is noticeable in the 90's in comparison to other decades. The percentages include the outliers that finally were considered as gaps. Figure 2 depicts for the temperature stations the cumulative percentage of gaps in the period 1981-2010. From the 13 weather stations with temperature data only 5 stations had daily observations after the early 90s. For this reason the evaluation of gaps, homogeneity and spatial importance was only developed for these

stations. From Fig. 2 it can be derived that since 1982 all 5 stations became fully operative. The cumulative percentages encompass the gaps due to outliers' detection.

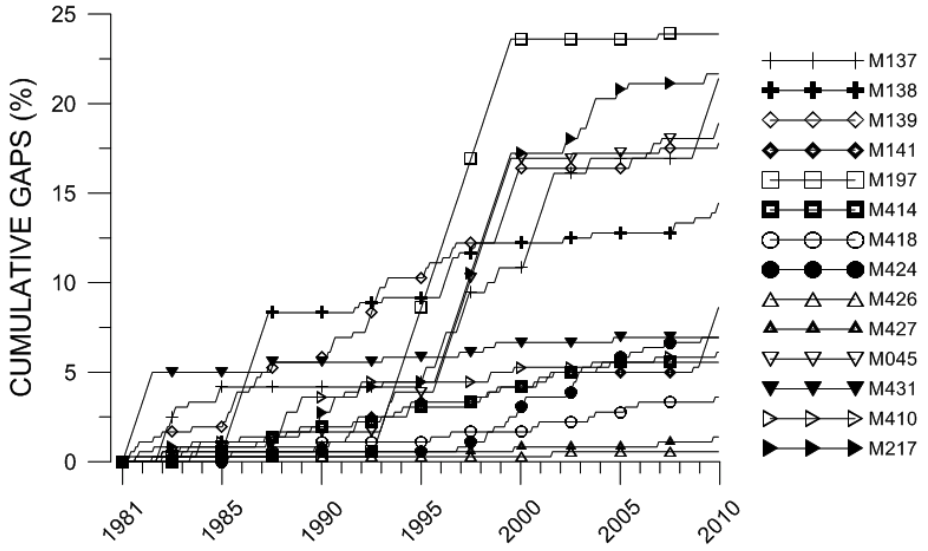

Figure 1. Cumulative percentage of gaps for the 14 selected precipitation stations.

The quality criteria applied to the available time series data of precipitation are summarized in Table 2. The percentage of gaps was derived for the entire study period. The identified inhomogeneities are presented in the sixth column and the last column presents the importance of the station by location. The Paute station present one minor inhomogeneity and does not justify rejection of the station. The stations Ricaurte, Gualaceo and Sigsig-INAMHI present weak inhomogeneities and are important by location. Since only 14 rainfall stations were retained it is evident that the density of stations is low; especially in the northern and western part of the basin.

Table 2. List of retained precipitation stations.

#	Station	Code	Elevation (m a.s.l.)	% gaps 1981-2010	Homogeneity Shift date*	Importance
1	Ricaurte	M426	2545	1	Oct-00	+
2	Sevilla de Oro	M431	2360	8		
3	Sayausí	M427	2780	1		+
4	Mazar rivera	M410	2450	7		+
5	Chanín	M414	3270	5		+
6	Paute	M138	2194	8	Oct-00	
7	El Labrado	M141	3335	9		+
8	Cumbe	M418	2720	4		+
9	Jacarín	M197	2700	24		
10	Palmas	M045	2400	17		+
11	Gualaceo	M139	2230	11	Jan-09	+
12	Sigsig INAMHI	M424	2600	6	Jan-09	+
13	Peñas Coloradas	M217	2321	19		+
14	Biblián	M137	2640	26		

* Shift date: date detected of mean-shift, although a change in the distribution without a shift in the mean could go undetected (Wang *et al.*, 2007).

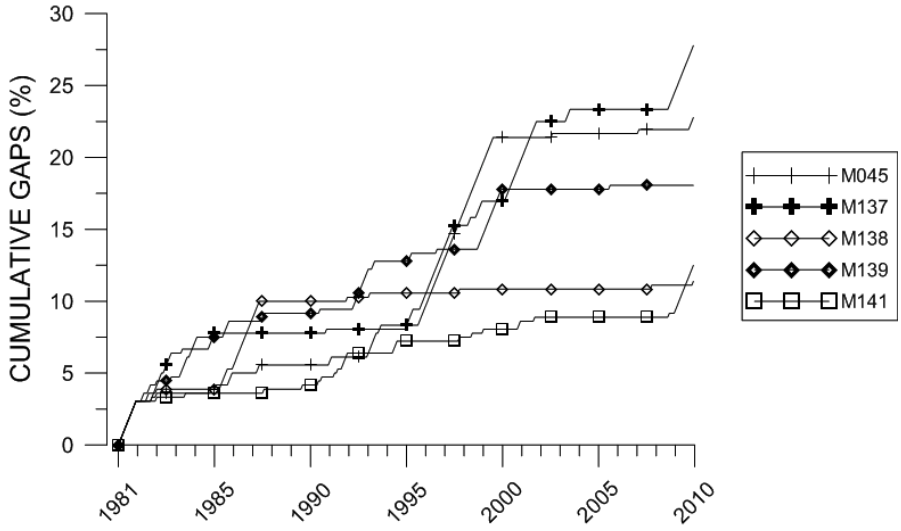

Figure 2. Cumulative percentage of gaps for the 5 selected temperature stations.

Table 3. List of retained temperature stations.

#	Station	Code	Elevation (m a.s.l.)	% gaps 1981-2010	Homogeneity Shift date*	Importance
1	Palmas	M045	2400	22		+
2	Biblián	M137	2640	26		+
3	Paute	M138	2194	10		+
4	Gualaceo	M139	2230	14		+
5	El Labrado	M141	3335	10		+

* Shift date: date detected of mean-shift, although a change in the distribution without a shift in the mean could go undetected (Wang *et al.*, 2007). In temperature time series all stations retained are homogeneous during the period of study.

Table 3 summarizes the quality criteria applied for the selection of the temperature stations. The percentage of gaps holds for the entire study period. The sixth column in the table presents the date of identified inhomogeneities and the last column the importance of the station by location. Biblián station, with 26% of gaps, surpassing by 1% the threshold of 25% is considered because of the lack of data. Temperature data did not present inhomogeneities during the study period, and most retained stations are located in the center of the basin.

4.2. Evaluation of the deterministic infilling methods

For infilling the gaps in the precipitation and temperature time series, 17 methods were applied. The evaluation of their performance was quantified by the ME, MAE and RMSE, between sampled observed values and infilled values. The sampled values roughly represent 1,5% of the data (Kotsiantis *et al.*, 2006). The list of the 17 evaluated deterministic infilling methods is given in Table 1.

Precipitation

For precipitation the performance of the infilling methods is presented separately for the bi-modal (BM) and the uni-modal (UM) stations because it was observed that metric errors differed between them especially during the rainy season JJA. Figures 3 and 4 present the ME, MAE and RMSE for respectively the BM and UM precipitation stations, and are calculated as the annual and seasonal average infilling errors.

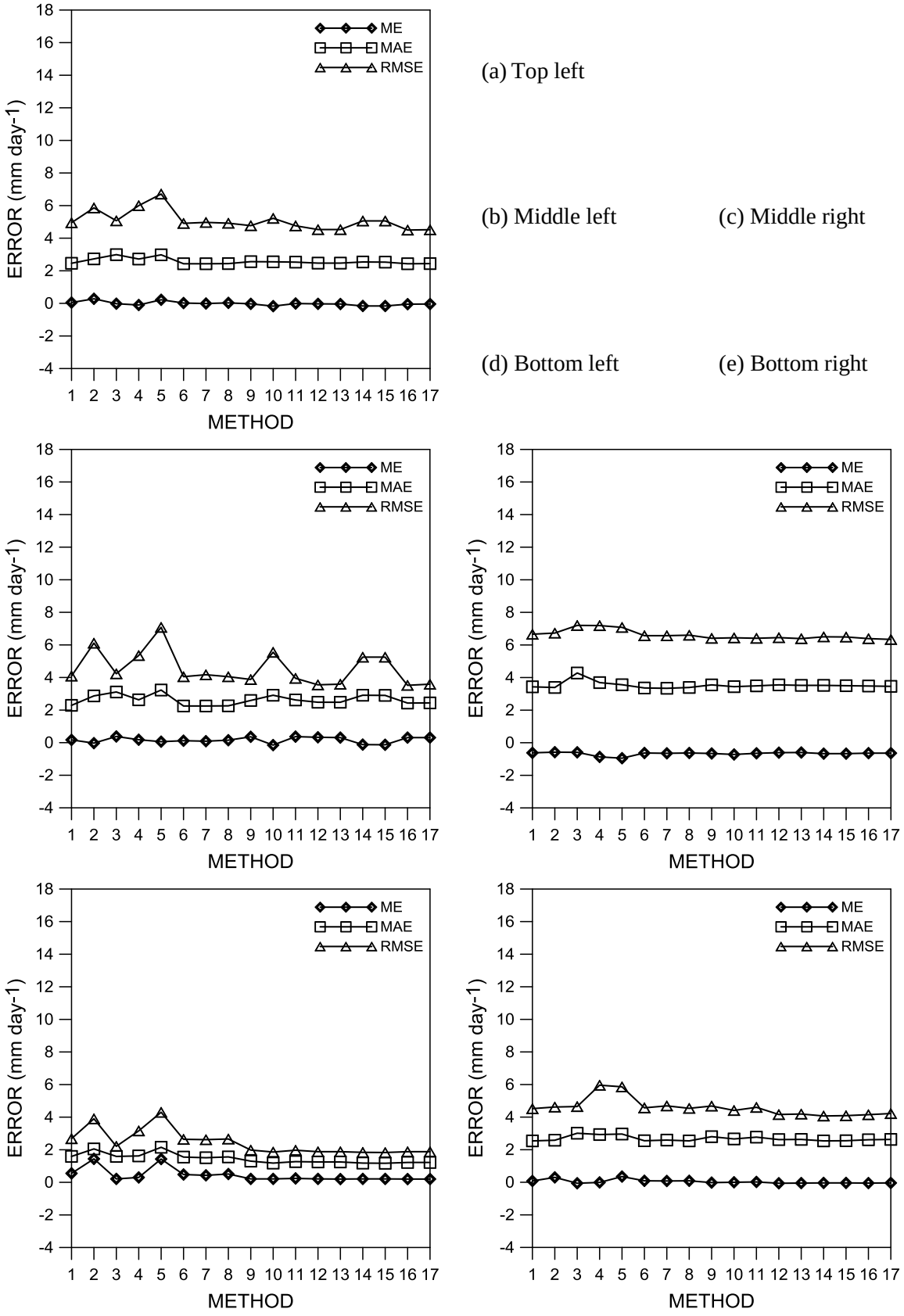

Figure 3. ME, MAE and RMSE of the annual (a) and seasonal (b = DJF; c = MAM; d = JJA; e = SON) average infilling errors of daily precipitation for bi-modal stations.

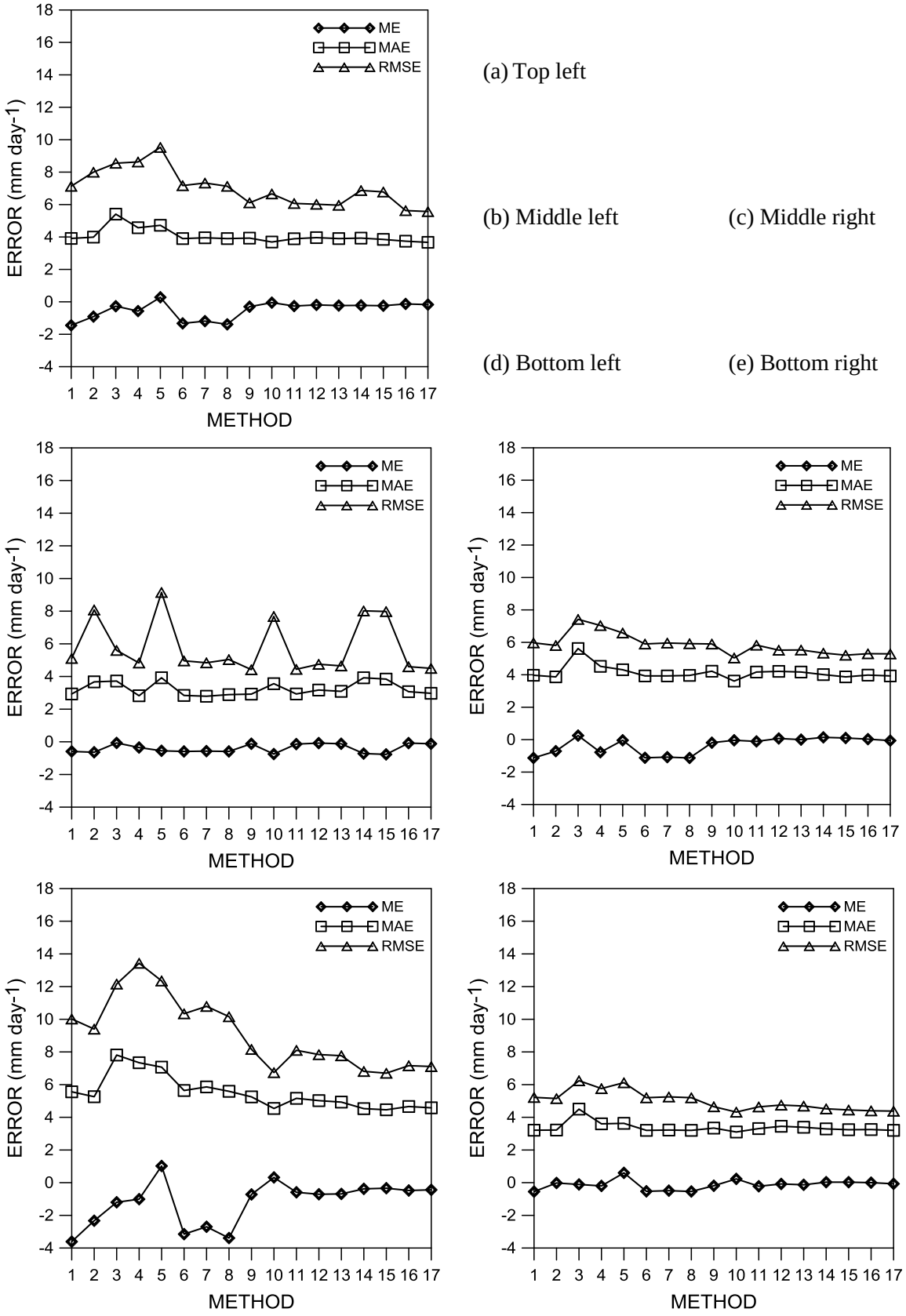

Figure 4. ME, MAE and RMSE of the annual (a) and seasonal (b = DJF; c = MAM; d = JJA; e = SON) average infilling errors of daily precipitation for uni-modal stations.

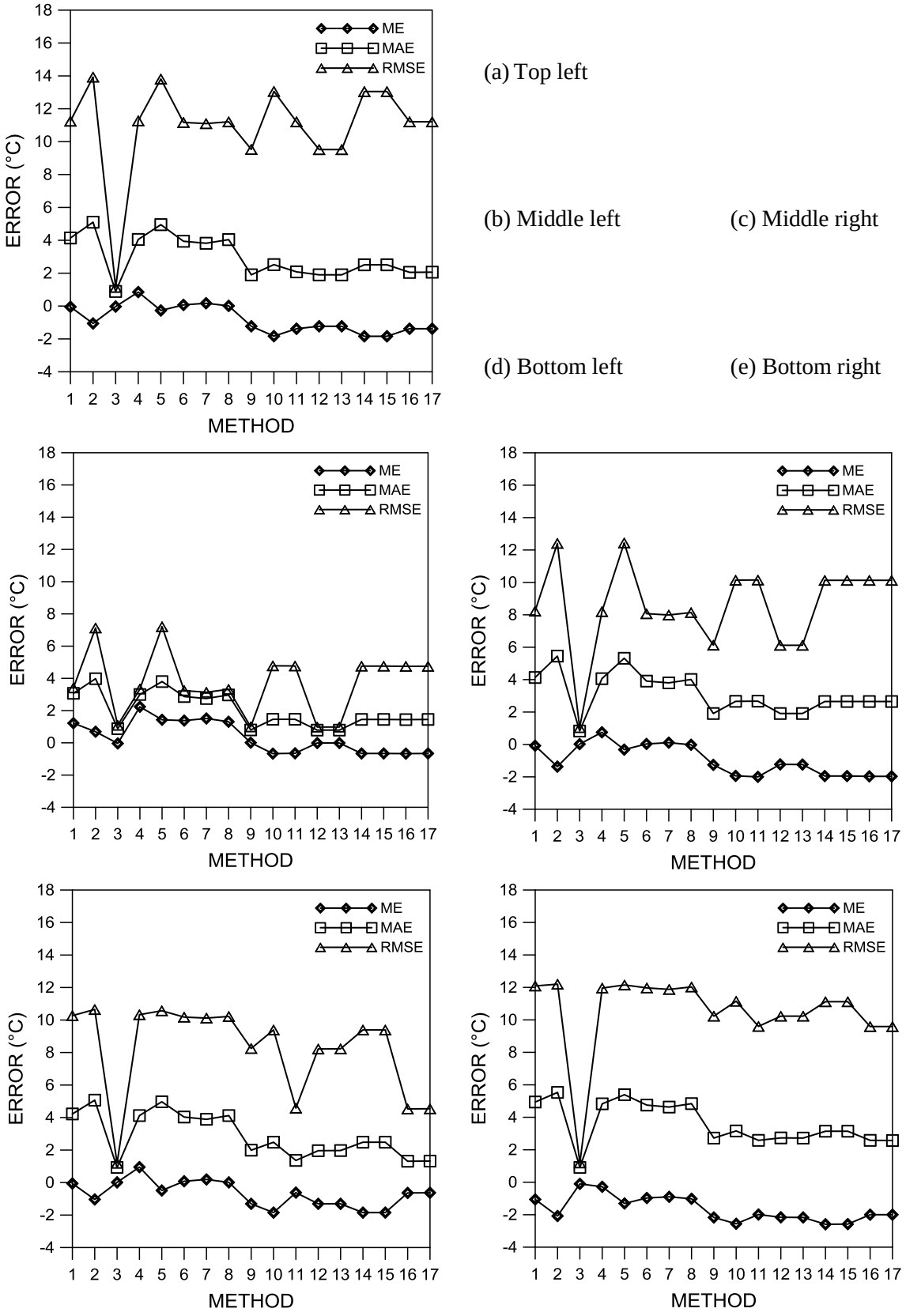

Figure 5. ME, MAE and RMSE of the annual (a) and seasonal (b = DJF; c = MAM; d = JJA; e = SON) average infilling errors of mean daily temperature.

For precipitation is the averaged annually value of RMSE lower for the infill methods [12, 13, 16, 17], and this for the BM and UM stations (Figs. 3a and 4a). All of these methods are multilinear regression methods, the first two weighted by R and the second two weighted by the R/d ratio, and the two weight powers (1 and 2). Annually ME is similar for all methods in BM and UM stations with exception of the [1, 2, 6, 7, 8] methods. Those methods present higher values for the UM stations. From the Figs. 3a and 4a it is clear that RMSE is more sensible than MAE and ME, that RMSE and MAE depict a similar pattern for respectively the BM and UM stations, and that for both station types the error value of RMSE and MAE is larger than for ME.

During DJF the RMSE of the different methods for both BM and UM stations varies considerably (Figs. 3b and 4b). The methods [12, 13, 16, 17] score best (lowest value for RMSE). All methods perform similarly with respect to the ME criteria. The linear and especially the multilinear methods perform well for the three metrics and this for all seasons (MAM see Figs. 3c and 4c; JJA see Figs. 3d and 4d; SON see Figs. 3e and 4e). In summary the analysis reveals that based on the ME, MAE and RMSE values, for both the BM and UM stations, that MLRWd-1 (#12), MLRWd-2 (#13), MLRWR/d+1 (#16) and MLRWR/d+2 (#17) are the best performing methods. In case one single method should be used for the annual and seasonal analysis, the MLRWR/d+2 (#17) would be most advisable, because the weighting considers R and d as factors of influence, and power 2 weights more close stations, which is most appropriate for study regions with high spatial variability, as is the case in the Paute basin.

Temperature

For temperature the BM and UM stations are considered as one group, because of the temperature similarity in both regions. Figure 5 depicts the ME, MAE and RMSE for the 17 infilling methods applied to the mean daily temperature time series, calculated respectively for the annual (Fig. 5a) and the seasonal (Figs. 5b, c, d, e) average infilling errors. The lowest RSME value for the annual average infilling error is obtained with method [3] (Fig. 5a). The calculated ME is similar for the methods [3, 6, 7, 8]. For MAE the lowest value is obtained by method [3], followed by the linear and multiple linear regression methods. For the seasonal mean daily temperature in the DJF (Fig. 5b) and MAM (Fig. 5c) seasons four methods [3, 9, 12, 13] score equally well with respect to RMSE and ME. During the JJA (Fig. 5d) and SON (Fig. 5e) season only the CMD method scores best for all three metrics. The lack of temperature stations might account for the poor performance of the regression methods, given stations are located too far apart. Therefore, for infilling gaps of temperature the lowest values of ME, MAE, and RMSE are realized by filling gaps with the long-term average of the same day of interest.

5. CONCLUSIONS

The research presented in this paper aimed assessing the performance of 17 deterministic infilling methods applied to the daily time series of precipitation and mean daily temperature, with application to the Paute basin, a highly variable region in the Andes mountain range in south Ecuador. Because of the robustness, the simplicity in implementation and computational efficiency, given also the complexity of the study area and the limitations in number of useable monitoring stations, preference was given to the deterministic methods over the more sophisticated techniques, such as the stochastic and artificial intelligence methods. The tested methods included the climatological mean, the nearer neighbor value, the average neighbor value, the inverse distance weighted, and the linear and multilinear regression methods. Useable stations were selected using the following tree criteria: (i) a maximum gaps threshold of 25% in the period January 1981 to December 2010; (ii) data homogeneity, and (iii) the relative importance of the station by location. Fourteen stations out of 53 available stations were selected for precipitation, and 5 out of 13 stations for temperature. The performance of the infilling methods was quantified by the mean error, the mean absolute error and the mean square error for annual and seasonal averages of infill errors in a random sample of 1,5% of the data.

The results for daily precipitation show that in general multiple linear regression methods outperform the others, annually and in all seasons. Specifically the multiple linear regression weighted by the ratio $(R/d)^2$ is the best performing method, consistently showing good performance in contrast to the linear regression with the nearest station by Pearson coefficient. The latter performed during some seasons similar as the multiple linear regression methods, but annually they performed inferior. For temperature the results show that the best performing infilling method for mean daily time series was the climatological mean of the day, annually and seasonally. The low performance of multiple linear regression methods might be the consequence of the low number of temperature stations and the scattering of the stations in a very heterogeneous region, affecting the difference between stations.

A shortcoming in this study, which will be the subject of further research, is that the elevation was not taken into account as a factor for interpolation. Also, in further work a comparison between the best performing deterministic methods with more sophisticated artificial intelligence methods as neural networks, support vector machines, and stochastic methods as Kriging and co-Kriging, will be pursued.

ACKNOWLEDGMENTS

This work has been funded by the Dirección de Investigación de la Universidad de Cuenca (DIUC) through the project "Análisis de los efectos del cambio climático en los caudales en las cuencas Andinas del Sur del Ecuador (Paute), debido a los cambios en los patrones de lluvia y temperatura", and by the Secretaria de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT) through a PhD grant for the first author. The authors would like to thank INAMHI for providing the meteorological data and the project SENESCYT PIC 728 which formed the platform for this research.

REFERENCES

Abatzoglou, J.T., K.T. Redmond, L.M. Edwards, 2009. Classification of regional climate variability in the State of California. *J. Appl. Meteorol. Clim*., 48(8), 1527-1541.

Adeloye, A.J., 1996. An opportunity loss model for estimating value of streamflow data for reservoir planning. *Water Resour. Manag*., 10(1), 45-79.

Adeloye, A.J., 2011. In: *Proceedings of the Symposium HS03 - Risk in Water Resources Management*, Melbourne, Australia, IAHS 347, pp. 121-126.

Aguilar, E., I. Auer, M. Brunet, T.C. Peterson, J. Wieringa, 2003. Guidelines on climate metadata and homogenization. World Meteorological Organization, WMO/TD No. 1186, 55 pp. Downloaded from *http://www.wmo.int/pages/prog/wcp/wcdmp/documents/ WCDMP-53.pdf* in December 2012.

Ashraf, M., J.C. Loftis, K.G. Hubbard, 1997. Application of geostatistics to evaluate partial weather station networks. *Agric. For. Meteorol*., 84(3-4), 255-271.

Baddour, O., H. Kontongomde (Eds.), 2007. The role of climatological normals in a changing climate. World Climate Data Monitoring Program, World Meteorological Organization, 46 pp. Downloaded from *http://www.wmo.int/datastat/documents/WCDMPNo61_1.pdf* in October 2012.

Bendix, J., W. Lauer, 1992.: Die Niederschlagsjahreszeiten in Ecuador und ihre klimadynamische Interpretation. *Erdkunde*, 46, 118-134.

Beauchamp, J.J., D.J. Dowing, S.F. Railsback, 1989. Comparison of regression and time-series methods for synthesizing missing streamflow records. *Water Resour. Bull*., 25, 961-975.

Bennett, N.D., L.T.H. Newham, B.F.W. Croke, A.J. Jakeman, 2007. Patching and disaccumulation of rainfall data for hydrological modelling. In: Oxley, L., D. Kulasiri (Eds.), Int. Congress on Modelling and Simulation (MODSIM 2007), *Modelling and Simulation Society of Australia and New Zealand Inc., New Zealand*, 2520-2526.

Celleri, R., P. Willems, W. Buytaert, J. Feyen, 2007. Space-time rainfall variability in the Paute basin, Ecuadorian Andes. *Hydrol. Process*., 21(24), 3316-3327.

Costa, A.C., A. Soares, 2008. Homogenization of climate data: Review and new perspectives using geostatistics. *Math. Geosci*., *41*(3), 291-305.

Coulibaly, P., N.D. Evora, 2007. Comparison of neural network methods for infilling missing daily weather records. *J. Hydrol*., 341, 27-41.

Dawson, C.W., C. Harpham, R.L. Wilby, Y. Chen, 2002. Evaluation of artificial neural network techniques for flow forecasting in the River Yangtze, China. *Hydrol. Earth Syst. Sci.,* 6(4), 619-626.

Demyanov, V., M. Kanevski, S. Chernov, E. Savelieva, V. Timonin, 1998. Neural network residual kriging application for climatic data. *J. Geogr. Inf. Decis. Anal*., 2, 215-232.

Di Piazza, A., F.L. Conti, L.V. Noto, F. Viola, G. La Loggia, 2011. Comparative analysis of different techniques for spatial interpolation of rainfall data to create a serially complete monthly time series of precipitation for Sicily, Italy. *Int. J. Appl. Earth Obs. Geoinf*., 13, 396-408.

Dingman, S.L., 1994. Physical hydrology. *Prentice Hall Englewood Cliffs, NJ*, USA, 575 pp.

Diodato, N., M. Ceccarelli, 2005. Interpolation processes using multivariate geostatistics for mapping of climatological precipitation mean in the Sannio Mountains (southern Italy). *Earth Surf. Process. Landf*., 30, 259-268.

Dumedah, G., P. Coulibaly, 2011. Evaluation of statistical methods for infilling missing values in high-resolution soil moisture data. *J. Hydrol*., 400(1-2), 95-102.

Gould, P.G., A.B. Koehler, J.K. Ord, R.D. Snyder, R.J. Hyndman, F. Vahid-Araghi, 2008. Forecasting time series with multiple seasonal patterns. *Eur. J. Oper. Res*., 191, 207-222.

Gyau-Boake, P., G.A. Schultz, 1994. Filling gaps in runoff time series in West Africa. *Hydrol. Sci. J*., 39(4), 621-636.

Hanson, R.T., M.W. Newhouse, M.D. Dettinger, 2004. A methodology to assess relations between climatic variability and variations in hydrologic time series in the southwestern United States. *J. Hydrol*., 287, 252-269.

Harvey, C.L., H. Dixon, J. Hannaford, 2010. Developing best practice for infilling daily river flow data. *British Hydrological Society, Third International Symposium, Managing Consequences of a Changing Global Environment, Newcastle*, UK, 8 pp.

Ilunga, M., D. Stephenson, 2005. Infilling streamflow data using feed-forward back-propagation (BP) artificial neural networks: application of standard BP and Pseudo Mac Laurin power series BP techniques. *Water SA*, 31(2), 171-176.

Khalil, M., U.S. Panu, W.C. Lennox, 2001. Groups and neural networks based streamflow data infilling procedures. *J. Hydrol*., 241, 153-176.

Kim, J.-W., Y.A. Pachepsky, 2010. Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. *J. Hydrol*., 394, 305-314.

Kotsiantis, S., A. Kostoulas, S. Lykoudis, A. Argiriou, 2006. Filling missing temperature values in weather data banks. *2nd IEE International Conference on Intelligent Environments, Athens*, Greece, 1, 327-334.

Laraque, A., J. Ronchail, G. Cochonneau, R. Pombosa, J.L. Guyot, 2007. Heterogeneous distribution of rainfall and discharge regimes in the Ecuadorian Amazon Basin. *J. Hydrometeorol., 8*(6), 1364-1381.

Mejia, R., D. Molinaro, G. Ontaneda, F. Rossel, 1996. Homogenización y regionalización de la pluviometria en la cuenca del río Paute. *Serie INSEQ Vol. 3. Republica del Ecuador, Ministerio de Energía y Minas, INAMHI, ORSTOM, Quito*, Ecuador.

Mileva-Boshkoska, B., M. Stankovski, 2007. Prediction of missing data for ozone concentrations using support vector machines and radial basis neural networks. *Informatica*, 50-52(31), 425-430.

Mwale, F.D., A.J. Adeloye, R. Rustum, 2012. Infilling of missing rainfall and streamflow data in the Shire River Basin, Malawi: A self organizing map approach. *Phys. Chem. Earth*, 50-52, 34-43.

Narapusetty, B., T. DelSole, M.K. Tippett, 2009. Optimal estimation of the climatological mean. *J. Climate*, 22, 4845-4859.

Ng, W.W., U.S. Panu, 2010. Infilling missing daily precipitation data at multiple sites using the multivariate truncated normal distribution model for weather generation. *Water*, 8 pp.

Ramos-Calzado, P., J. Gómez-Camacho, F. Pérez-Bernal, M.F. Pita-López, 2008. A novel approach to precipitation series completion in climatological datasets: Application to Andalusia. *Int. J. Clim*., 28, 1525-1534.

Rossel, F., E. Cadier, 2009. El Niño and prediction of anomalous monthly rainfalls in Ecuador. *Hydrol. Process.*, 23(22), 3253-3260.

Salazar, G., H. Rudnick, 2008. Hydro Power plants in Ecuador: A technical and economical analysis. *Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century, IEEE.* Pittsburg, PA, USA, 5 pp.

Teegavarapu, R.S., V. Chandramouli, 2005. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *J. Hydrol*. 312, 191-206.

Teegavarapu, R.S.V., M. Tufail, L. Ormsbee, 2009. Optimal functional forms for estimation of missing precipitation data. *J. Hydrol*., 374, 106-115.

Vasiliev, I.R., 1996. Visualization of spatial dependence: An elementary view of spatial autocorrelation. In: Arlinghaus, S.L. (Ed.), Practical Handbook of Spatial Statistics, *CRC Press, Boca Raton, Florida*, USA, 17-3.

Villazón, M., P. Willems, 2010. Filling gaps and daily disaccumulation of precipitation data for rainfall-runoff model. *Proc. 4th Int. Sci. Conf. BALWOI 2010 on Water Observation and Information Systems for Decision Support, Rep. Macedonia*, 9 pp.Wagner, P.D., P. Fiener, F. Wilken, S. Kumar, K. Schneider, 2012. Comparison and evaluation of spatial interpolation schemes for daily rainfall in data scarce regions. *J. Hydrol.*, 464-465, 388-400.

Wan, H., X.L. Wang, V.R. Swail, 2010. Homogenization and trend analysis of Canadian near-surface wind speeds. *J. Climate*, 23, 1209-1225.

Wang, X.L., Q. H. Wen, Y. Wu, 2007: Penalized maximal t test for detecting undocumented mean change in climate data series. *J. Appl. Meteor. Climatol*., 46 (6), 916-931.

Wang, X.L., 2008a. Accounting for autocorrelation in detecting mean-shifts in climate data series using the penalized maximal t or F test. *J. Appl. Meteor. Climatol*., 47, 2423-2444.

Wang, X.L., 2008b. Penalized maximal F-test for detecting undocumented mean-shifts without trend-change. *J. Atmos. Oceanic Technol*., 25(3), 368-384.

Xia, Y., P. Fabian, A. Stohl, M. Winterhalter, 1999. Forest climatology: Estimation of missing values for Bavaria, Germany. *Agric. For. Meteorol*., 96, 131-144.

Yozgatligil, C., S. Aslan, C. Iyigun, I. Batmaz, 2013. Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. *Theor. Appl. Climatol*., 112, 143-167.