GAP FILLING BASED ON A QUANTILE PERTURBATION FACTOR TECHNIQUE

D. E. MORA (1,2), G. WYSEURE (3), P. WILLEMS (2,4)
(1): Universidad de Cuenca, Faculty of Engineering, PROMAS Research Unit, Av. 12 de abril, Cuenca – Ecuador
(2): Katholieke Universiteit Leuven, Hydraulics Divison, Kasteelpark Arenberg 40,3001 Leuven-Belgium
(3): Katholieke Universiteit Leuven, Soil and Water Management Divison, Celestijnenlaan 200E,3001 Leuven-Belgium
(4): Vrije Universiteit Brussel, Department of Hydrology & Hydraulic Engineering, Pleinlaan 2, 1050 Brussels- Belgium

The presence of gaps in hydro-meteorological series is a common problem at the moment of analyzing data series. That is the case of the Ecuadorian hydrological data series, presenting eventual gaps of short term duration. The Paute River Basin, located in the Southern Ecuadorian Andes, is one of the most monitored basins in Ecuador, with 25 rainfall observed sites during the period of 1963 till 1990. However, its data base suffers of about 20% of missing data.

For this research, two techniques were evaluated comparing their efficiency in the filling of missing gaps. The first one is based on multiple linear regressions, which applies a logarithmic transformation to the data and then converts the data to normalized standard variables. The second one is a new proposed technique based on quantile perturbation approach after a classical prior gap filling. It is used to shelter estimations for high and low intensities based on: i. Identification of the station with the highest monthly correlation ii. Selection and ranking of the stations for which the correlation is significant, tested by the t-test, iii. Gap filling based on the stations with the highest significant correlation, and iv. the application of a correction factor to the filled value.

For the evaluation, 3 un-interrupted daily rainfall data series were selected. Data series were deleted in a random way, simulating the 20% of missing data. The two filling techniques were applied separately. Finally, data series were evaluated by the different statistic criteria.

Results indicate that the proposed technique performs an efficient filling of missing gaps. It supports the definition of gaps corresponding to high or low events and avoids, in a certain range, the averaging of the series. However, it might lead to double counting of high/low extremes events.

INTRODUCTION

In the Southern Ecuadorian Andes, the rainfall variability is considerably high (Espinoza Villar et al., 2009) [5], therefore the importance to have a good and representative rainfall data records is high, to be used in different applications, i.e. by rainfall-runoff modeling. Within this region, the Paute River Basin, is one of the most monitored basins in Ecuador since 1963 till present, due to its important hydropower energy production. However, as in other river basins, records collected in long periods of time present data gaps. Gaps can be an obstacle at the moment of assessing and evaluate the region hydrology towards the management of water resources.

Gaps maybe present for different circumstances like the problems with the measuring device, loss of extreme events records for bad functioning of equipment, absence of observers and the lack of funds to continue the measurements. This last one is the one that affected the most in the region when unfortunately, monitoring was cancelled for most of the stations at the beginning of the 1990's due to the change of structure of governmental institutions (Galarraga-Sanchez, 2000) [6]. As this specific case, several basing located in developing countries deal with lack of data. Therefore the application of any hydrological model or hydrology analysis needs first of an analysis of data technique to complete the rainfall records.

Procedures to estimate missing data are well documented and generally use an interpolation method based on, for instance, normal ratio precipitations (Paulhus and Kohler, 1952) [10], regressions (Makhuva et al., 1997) [7] or a combination of both. In addition, advance interpolation techniques, as the Kriging technique, are available, but limited to complex and heavy computationally resources, showing small increase in accuracy (Teegavarapua and Chandramoulia, 2005) [12]. Another procedure to fill rainfall gaps is stochastic modeling of rainfall sequences (Zuccini et al., 1992) [14]. This procedure, which can be applied irrespective of gaps in the records, is used to generate artificial rainfall sequences. However, this procedure is limited to be used as input of rainfall-runoff models which uses a prior calibration.

In a previous study within the region, (Celleri et al., 2007) [3] estimated missing data using the method of linear regression, in which monthly estimates were calculated using linear regression with stations showing the highest correlation coefficient. This procedure was followed separately for each month (i.e all January's are analyzed separately from all February's, etc.). To finish, daily estimates were derived from monthly values assuming that each station had the same daily distribution as the nearest recording station. However, this method accounts a not-well identified daily distribution, resulting in a poor accurate rainfall depths with high residual errors, especially the ones regarding extreme events.

Within the regressions techniques, the multiple linear regressions bring an effective technique with lower errors than the linear regression technique (Villazón and Willems, 2010) [13]. However, there is the need to previously analyze and take into account criteria, i.e. rainfall vs. elevation, to group the stations into a specific order. This is relevant to preserve important correlations, although the criteria may be subjective depending of rainfall properties. Therefore, the present research focuses on the uses of a multiple linear regression and the quantile perturbation factors after a classical linear regression prior gap filling. The methods were applied in 3 un-interrupted daily rainfall data series for the period of 1964-1994, after simulating a missing data rate of 20%. The evaluation was performed considering the Mean Absolute Error (MAE) for all the series and peaks, Standard Deviation (STD), the Root Mean Squared Error (RMSE) and the Efficiency (EF), (Nash & Sutcliffe, 1970) [9].

MATERIALS AND METHODS

Study area

The Paute River basin is a mountain basin located in the Inter-Andean Depression which separates the "Western" and "Real Cordillera" (eastern) mountain ranges in the south of

Ecuador (Coltorti and Ollier, 2000) [4]. The study area considers the Paute River catchment area upstream to the hydropower dam "Amaluza", with 5066 km2, between Latitudes [2.3 -3.3] South and Longitude [79.4 – 78.4] West. Elevation ranges between 1824 to 4680 m a.s.l. The basin consists of an irregular pattern of mountain ranges intersected by deep valleys, draining its waters in southwest – northeast direction reaching the Atlantic Ocean trough the Amazon River. The western headwaters are part of the continental water divide of the Amazon River basin, at a distance smaller than 70 km. from the Pacific Ocean, see Figure 1.

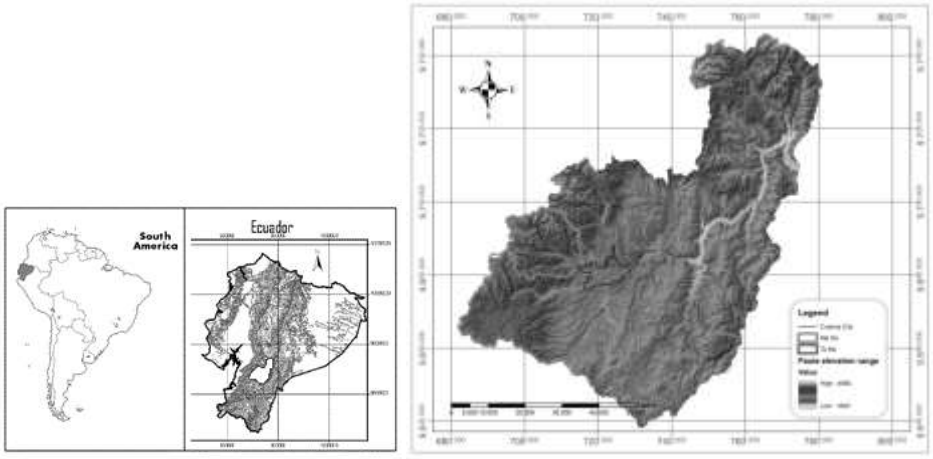

Figure 1 Location of the Paute River Basin in the Southern Ecuadorian Andes

Data availability

The monitoring data is dispersed in several databases ruled by different institutions: INAMHI (Instituto Nacional de Meteorología e Hidrología), INERHI (Instituto Ecuatoriano de Recursos Hídricos) and INECEL (Instituto Ecuatoriano de Electrificación). In 1987, the "Plan Nacional de Riego" project INERHI-ORSTOM (Office de la Recherche Scientifique et Technique d'Outre-Mer, France) developed a complete assessment of the existing data bases, including the revision on the quality of the data and merged the different databases in one database called BIDRIE (Ruf and Le Goulven, 1987) [11]. Unfortunately, monitoring was cancelled for most of the stations at the beginning of the 1990's due to the change of structure of governmental institutions (Galarraga-Sanchez, 2000) [6]. The monitoring continued only in few sites. However, local institutions have monitored additional data in an independent way. Recent data are available from ETAPA (the drinking water authority for the city of Cuenca) for the period 1997-2004. For this research, data were compiled from the BIDRIE, INAMHI and ETAPA databases, resulting in 25 rainfall stations, 11 started in 1962-1964 and 14 in 1972-75 till 1992-1993, from which 3 with recent data from 1997-2004. Figure 2 give an overview of the stations considered. Rainfall data series are in daily resolution. The sites that were considered for evaluation are sites that do not have gaps during 1975 and 1985. These sites are Cuenca Aeropuerto M067, Cumbe M418 and El Labrado M141.

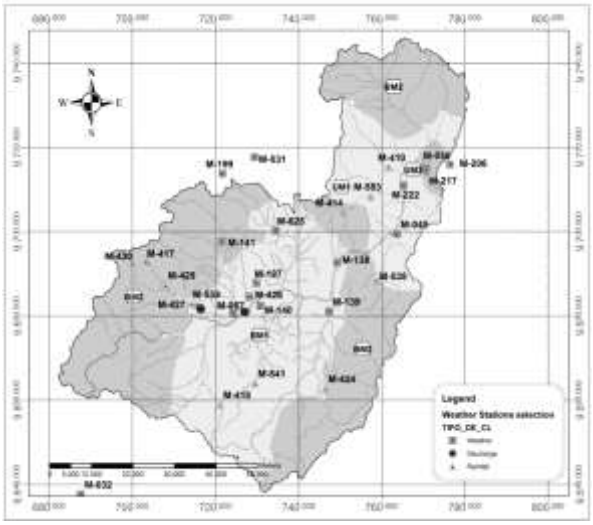

Figure 2 Location of weather, rainfall and discharge stations

Multiple linear regression

The monthly stream simulation tool (HEC4) developed by the U.S. Corps of Engineers (HEC, 1971) has been applied as a multiple linear regression approach. Its use in monthly rainfall accumulation is possible because of the correlated data used. The model applies a logarithmic transformation to the data, where after each value is converted to a normalized standard variate (Villazón and Willems**,** 2010) [13]. Correlation coefficients between all pairs of stations for each current and preceding calendar month are calculated and stored in a correlation matrix. For filling the missing months a multiple linear regression equation is applied, Eq.1 (Carter and Benson, 1970) [2]:

$$K_{i,j} = \beta_1 K_{i,1} + \beta_2 K_{i,2} + ... + \beta_{j-1} K_{i,j-1} + \beta_j K_{i-1,j} + \beta_{j+1} K_{i-1,j+1} + ... + \beta_n K_{i-1,n} + \sqrt{1 - R_{i,j}^2} \cdot z_{i,j} \quad (1)$$

K = Monthly logarithmically transformed flow, expressed as a normal standard deviate
β = Beta coefficient computed from correlation matrix
i = Month number
j = Station number
n = Number of interrelated stations
R = Multiple correlation coefficient
z = Random number from standard normal distribution

In order to maintain a reasonable number of stations for each computation, stations are first ranked by rainfall regime (Celleri et al, 2007, Mora et al, 2012) [3] [8] and grouped in sets of 6. The manner of grouping stations is extremely important, because it is important to include in each successive group as much information as possible that is pertinent to the computation of missing data for each station in the group (Beard et al., 1970) [1].

Quantile perturbation approach

For the case of daily rainfall series, a new methodology based on the application of a quantile correction factors after a classical prior gap filling is applied. This method is used to shelter estimations for high and low intensities. It is based on:

i. Identification of the station with the highest correlation with the station for which the gap filling will be performed (month per month).

ii. Selection and ranking of the stations for which the correlation is significant (statistical hypothesis is used based on t-test).

iii. Gap filling based on the stations with the highest significant correlation (unless none of them is significant, then the testing is dropped).

iv. Application of a correction factor to the filled value, based on the difference in monthly empirical frequency distribution between the two stations. The correction factor is quantile (or return period or cumulative probability) based. It is calculated as the ratio of the values (from the 2 stations) with the same quantile as the quantile of the value considered for filling.

The perturbations refer to relative changes between two series. The first one is the series of the site with the highest significant correlation x_N. The second one is the series of the site of interest x_L. The daily values that correspond to these return periods are denoted as the quantiles x_L, $x_L/2$,…,x_L/i for the second series and x_N, $x_N/2$,…x_N/i for the first one. The series are ranked and the return periods $T_{N(i)}$ and $T_{L(i)}$ are calculated, see Eq. 2 and Eq. 3.

$$T_{L(i)} = \frac{L}{i} \tag{2}$$

$$T_{N(i)} = \frac{N}{i} \tag{3}$$

A perturbation factor $PF_{(i)}$ is calculated in function of its return period, see Eq. 4. Then, this perturbation factor is applied to the missing value taking into account the date and the quantile to which belongs the value in the selected site.

$$PF_{(i)} = \frac{x_L\left(T_{L(i)}\right)}{x_N\left(T_{N(i)}\right)} \tag{4}$$

The missing values are filled considering the highest correlated station at a specific month, and with a different correction factor depending on the quantile of a daily intensity rainfall.

RESULTS AND DISCUSSION

These techniques were evaluated using statistical criteria as the Mean Squared Error (MSE), Standard Deviation (STD) and the Efficiency (EF), (Nash & Sutcliffe, 1970) [9] calculated in the filled series considering a monthly aggregation and in the daily series. To determine the statistical criteria, only filled values were considered.

Table 1 shows the different statistical criteria for series filled with both techniques when compared with the observed series at monthly and daily data. When monthly data are compared, it is shown that the multiple linear regression (M.L.) presents lower MSE than the QPA technique in two of the three cases. The same is shown in the other statistical criteria.

Table 1. Statistical criteria for the gap filling techniques for monthly and daily series at the sites M067, M418 and M141

		M067		M418		M141	
		M. L.	Q.P.A.	M. L.	Q.P.A.	M. L.	Q.P.A.
Monthly	MSE	99.33	165.54	46.45	62.12	111.68	108.78
	STD	47.23	47.07	33.79	33.67	53.41	53.19
	EF	0.89	0.68	0.95	0.91	0.74	0.75
Daily	MSE	43.51	31.82	24.61	16.86	42.58	33.90
	STD	5.56	5.44	3.57	3.50	5.89	5.79
	EF	0.85	0.88	0.84	0.87	0.83	0.86

The M.L. technique shows higher efficiencies that the QPA technique. This means that the average of gaps were filled considering all the distribution of monthly intensities and these values are closer to the observed ones than in the QPA technique.

The statistic criteria evaluated the daily series gap filling shows nearly the same performance as in the monthly series when STD and EF are compared. However, attention must be taken at the moment of evaluate the extreme values.

Figure 3 shows the filled gaps vs. the observed daily intensities. The figure shows that a better estimation is achieved for events with low intensities in both cases. However, for the M.L technique extreme values are underestimated or overestimated from the observed series in a higher range than in the QPA technique. Hence, a better performance of the gap filling is achieved for extreme events when using QPA. This improvement is also reflected in the MSE values in Table 1.

Results indicate that the proposed technique performs an efficient filling of missing gaps. It supports the definition of gaps corresponding to high or low events and avoids, in a certain range, the averaging of the series.

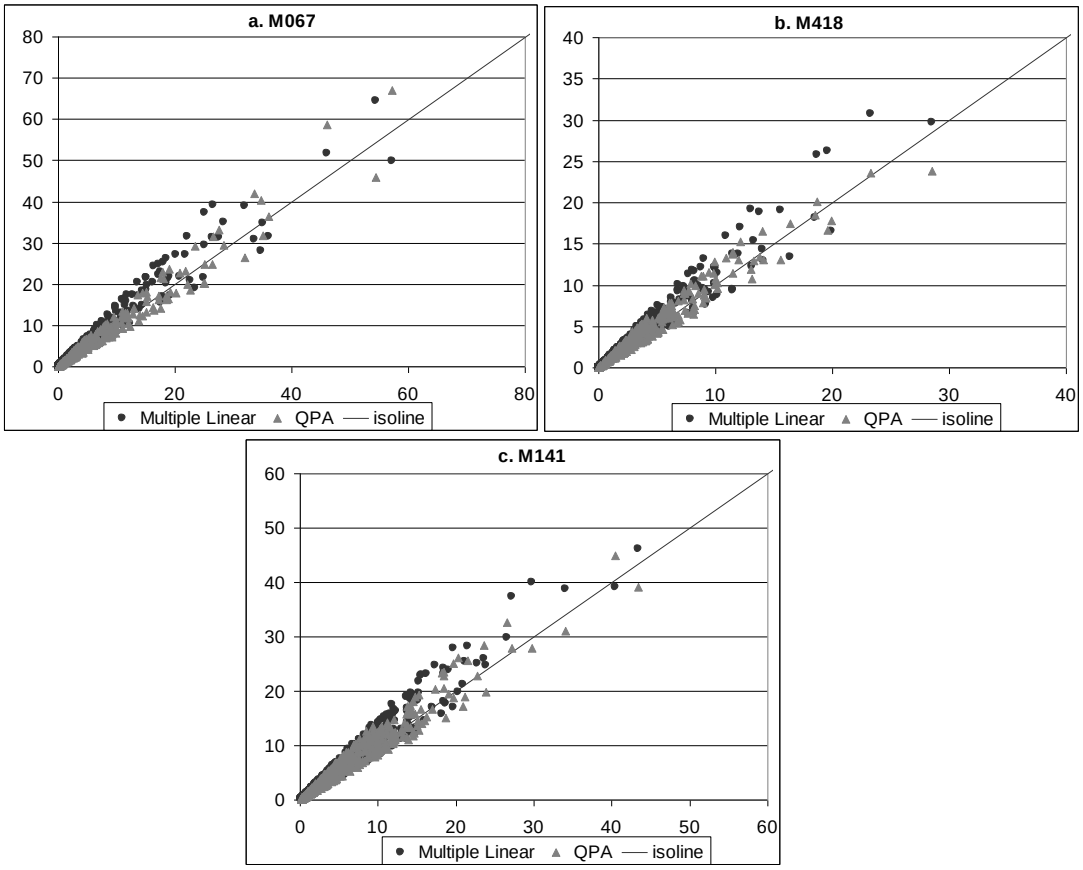

Figure 3 Observed vs. Filled daily series for sites a. M067, b. M418 and x. M141.

However, the QPA technique might present some uncertainties. One of them is that at low extreme values, low quantiles intensities may result into high perturbation factors, especially when intensities are below 1mm, which can overestimate rainfall depths. Therefore it is advised to consider a rainfall depth threshold in which perturbation factors are limited in their magnitude.

Another limitation of the QPA technique is that it might lead to double counting of high/low extremes events.

AKNOWLEDGEMENT

This research was feasible thanks to the Selective Bilateral Agreement (SBA) between KU Leuven and Universitites in Latin America. The research was also developed within the frame of the IWQM VLIR UOS project. The first author would like to thank to dr. eng. Felipe Cisneros (PROMAS – U.CUENCA) and to dr. eng. Mauricio Villazón (U. Mayor de San Simón) for their valuable cooperation.

REFERENCES

[1] Beard L. R., Fredrich A. J., Hawkins E. F., 1970: Estimating monthly streamflows within a region. Presented at ASCE National Meeting on Water Resources Engineering, held at Memphis, Tennesee. January 26-30.

[2] Carter, R., Benson, M., 1970. Concepts for the design of streamflow data programs: US Geol. Survey open-file report 33.
[3] Celleri, R., Willems, P., Buytaert, W., Feyen, J., 2007. Space–time rainfall variability in the Paute basin, Ecuadorian Andes. Hydrological Processes 21, 3316–3327.
[4] Coltorti, M., Ollier, C.., 2000. Geomorphic and tectonic evolution of the Ecuadorian Andes. Geomorphology 32, 1–19.
[5] Espinoza Villar, J.C., Ronchail, J., Guyot, J.L., Cochonneau, G., Naziano, F., Lavado, W., De Oliveira, E., Pombosa, R., Vauchel, P., 2009. Spatio-temporal rainfall variability in the Amazon basin countries (Brazil, Peru, Bolivia, Colombia, and Ecuador). International Journal of Climatology 29, 1574–1594.
[6] Galarraga-Sanchez, R., 2000. Informe Nacional sobre la gestion del agua en el Ecuador. Comite Asesor Tecnico de America del Sur (SAMTAC). Global Water Partnership (GWP), Quito-Ecuador, February 2000, pp. 80.
[7] Makhuvha T., Pegram G., Sparks R., Zucchini W. 1997a: Patching rainfall data using regression methods I. Best subset selection, EM and pseudo-EM methods: theory. Journal of Hydrology 198: pp 289-307
[8] Mora, D., Willems, P., 2012. Decadal oscillations in rainfall and air temperature in the Paute River Basin—Southern Andes of Ecuador. Theoretical and Applied Climatology 108, 267–282
[9] Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—A discussion of principles. Journal of Hydrology 10, 282–290
[10] Paulhus, J.L.H. and Kohler, M.A. 1952: Interpolation of missing precipitation records, Mon. Wea. Re6., 80, 129–133.
[11] Ruf, T., Le Goulven, P., 1987. L'exploitation des inventaires réalisés en Equateur pour une recherche sur les fonctionnements de l'irrigation. Bulletin de Liaison-ORSTOM. Département H 30–47.
[12] Teegavarapua R. and Chandramoulia, V. 2005: Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records, Journal of Hydrology, 312: 191-206.
[13] Villazón, M.F., Willems, P., 2010. Filling gaps and Daily Disaccumulation of Precipitation Data for Rainfall-runoff model. Presented at the Proc. 4th Int. Sci. Conf. BALWOI 2010 on Water Observation and Information Systems for Decision Support, Rep. Macedonia, pp. 25–29.
[14] Zucchini, W., P. Adamson and L. McNeill, 1992: Applications of stochastic daily rainfall model, S.Afr. J. Sci. 88: pp 103-109.