Hydrol. Earth Syst. Sci., 16, 2253–2266, 2012
www.hydrol-earth-syst-sci.net/16/2253/2012/
doi:10.5194/hess-16-2253-2012

Hydrology and Earth System Sciences

Record extension for short-gauged water quality parameters using a newly proposed robust version of the Line of Organic Correlation technique

B. Khalil and J. Adamowski

Department of Bioresource Engineering, McGill University, Sainte-Anne-de-Bellevue, QC, Canada

Correspondence to: B. Khalil (bahaa.khalil@mail.mcgill.ca)

Received: 6 March 2012 – Published in Hydrol. Earth Syst. Sci. Discuss.: 13 April 2012
Revised: 16 June 2012 – Accepted: 25 June 2012 – Published: 23 July 2012

Abstract. In many situations the extension of hydrological or water quality time series at short-gauged stations is required. Ordinary least squares regression (OLS) of any hydrological or water quality variable is a traditional and commonly used record extension technique. However, OLS tends to underestimate the variance in the extended records, which leads to underestimation of high percentiles and overestimation of low percentiles, given that the data are normally distributed. The development of the line of organic correlation (LOC) technique is aimed at correcting this bias. On the other hand, the Kendall-Theil robust line (KTRL) method has been proposed as an analogue of OLS with the advantage of being robust in the presence of outliers. Given that water quality data are characterised by the presence of outliers, positive skewness and non-normal distribution of data, a robust record extension technique is more appropriate. In this paper, four record-extension techniques are described, and their properties are explored. These techniques are OLS, LOC, KTRL and a new technique proposed in this paper, the robust line of organic correlation technique (RLOC). RLOC includes the advantage of the LOC in reducing the bias in estimating the variance, but at the same time it is also robust in the presence of outliers. A Monte Carlo study and empirical experiment were conducted to examine the four techniques for the accuracy and precision of the estimate of statistical moments and over the full range of percentiles. Results of the Monte Carlo study showed that the OLS and KTRL techniques have serious deficiencies as record-extension techniques, while the LOC and RLOC techniques are nearly similar. However, RLOC outperforms OLS, KTRL and LOC when using real water quality records.

1 Introduction

In many cases, water resources management involves the use of different hydrologic or water quality data to simulate the outcomes of decisions (Hirsch, 1982). However, records available for many streams are either too short to contain a sufficient range of hydrologic and water quality conditions or have periods of missing data (Alley and Burns, 1983). One solution to this problem is to rely on the transfer of information from nearby stream gauges with available long-term records (Hirsch, 1982; Alley and Burns, 1983; Vogel and Stedinger, 1985). This can be done by extending historic hydrologic or water quality records of interest in time via extrapolation of the correlation between these records at the site of interest and concurrent records at a nearby site. This is commonly done using ordinary least squares regression (OLS). OLS is commonly applied to reconstitute information about short-gauged water quality variables (Harmancioglu and Yevjevich, 1986, 1987; Harmancioglu et al., 1999; Robinson et al., 2004; Khalil and Ouarda, 2009).

Water quality data by definition are non negative and have special characteristics, such as presence of outliers, positive skewness, non-normal distribution of data, censored records (e.g. concentrations below a detection limit), seasonal patterns and autocorrelation. Outliers and positive skewness are more common in water quality data. Due to the presence of outliers and positive skewness, water quality data often have a form resembling a log-normal distribution (Lettenmaier, 1988; Berryman et al., 1988).

There are two main deficiencies in using the OLS as a record extension technique for water quality data. First, it is not robust in the presence of outliers. Presence of outliers

significantly affects intercept and slope estimates in the OLS (Nevitt and Tam, 1989; Granato, 2006). Second, the variance in the extended records provides a negatively biased estimate of the true variance (Hirsch, 1982; Alley and Burns, 1983; Moog and Whiting, 1999; Helsel and Hirsch, 2002; Khalil et al., 2010; Koutsoyiannis and Langousis, 2011). If the record-extension technique that is used presents a bias in the estimation of extreme values, this leads to bias in the estimation of the probability of exceedance (Hirsch, 1982). For water quality, extreme values and percentiles are of interest for the assessment of compliance with standards or permissible limits (Khalil et al., 2010).

Several robust regression techniques have been proposed in the literature as analogues to OLS with the advantage of being robust in the presence of outliers (e.g. Huber M-estimation; Least Median of Squares; Least Absolute Deviations; Winsorized regression; and Trimmed Least Square estimation). In general, robust regression techniques have been developed mainly for the situation where symmetric error distributions have heavy tails because of the presence of outliers in the observed data (Dietz, 1987; Nevitt and Tam, 1989). In contrast, in nonparametric techniques (e.g. monotonic regression and Kendall-Theil robust line) methods for parameter estimation are regarded as distribution free (Nevitt and Tam, 1989).

Nevitt and Tam (1989) compared the behaviour of robust regression and nonparametric regression techniques with OLS with respect to the presence of outliers and deviation from normality. Their results showed that Kendall-Theil robust line (KTRL) is a very strong analogue to OLS regression with the advantage of being robust in the presence of outliers. It can provide accurate estimates of the population parameters under both the presence of outliers and deviation from normality. Although their results showed that Least Absolute Deviation (LAD) technique outperforms KTRL under heavily presence of outliers, KTRL was almost as strong as LAD regression. However, under deviation from normality, no estimator outperformed the KTRL technique. Nevitt and Tam (1989) concluded that the KTRL technique provides strong parameter estimation under presence of outliers and/or deviation from normality.

The Kendall-Theil robust line (KTRL) technique is an analogue to OLS with the advantage of being robust in the presence of outliers and/or deviation from normality. KTRL is almost as efficient as the OLS when normality assumptions are met, and more efficient under deviation from normality (Helsel and Hirsch, 2002). For the case where the data show a linear pattern, homoscedastic and normality of residuals, the KTRL and OLS will provide almost identical results (Hirsch et al., 1991). When outliers exist, the KTRL will produce a line with greater efficiency than OLS (Hirsch et al., 1991; Helsel and Hirsch, 2002). However, similar to OLS, KTRL also underestimates the variance in the extended records. KTRL has been widely applied not only for record extension but also for trend assessment (e.g. Albek, 2003; Granato, 2006; Olson et al., 2010; Déry et al., 2011).

The line of organic correlation (LOC) was first introduced in hydrology by Kritskiy and Menkel (1968). The LOC theoretical characteristics were presented by Kruskal (1953). The main advantage of the LOC is that it is able to maintain variability in the extended records (Helsel and Hirsch, 2002). However, it is not robust in the presence of outliers. Several studies applied the LOC for extending stream-flow records (e.g., Hirsch, 1982; Hirsch et al., 1991; Jia and Culver, 2006; Ryu et al., 2010), for estimation of missing precipitation values (e.g. Raziei et al., 2009, 2011), and also for extension of water quality records (e.g. Khalil et al., 2010, 2011).

The main objective of this paper is to evaluate the suitability of four record-extension techniques for the reconstruction of information about short-gauged water quality parameters. These techniques are OLS, KTRL, LOC and a modified version of the LOC that retains the LOC advantage of preserving the cumulative distribution function of predictions, but which is also robust in the presence of outliers. The modified version proposed in this study will be referred to hereafter as robust line of organic correlation (RLOC).

2 Theoretical background

Assume that the measured variables x and y have n_1+n_2 and n_1 years of data, respectively, of which n_1 are concurrent data as follows:

$$x_1, x_2, x_3, ..., x_{n_1}, x_{n_1+1}, x_{n_1+2}, ..., x_{n_1+n_2}$$
$$y_1, y_2, y_3, ..., y_{n_1}.$$

Assume that at the year n_1+n_2, it is desired to reconstitute information about the variable y by extending its records through the period from n_1+1 to n_1+n_2 years. In this case, record extension techniques can be used. In this study four record extension techniques were used, the OLS, KTRL, LOC and RLOC, as briefly described in the following subsections.

2.1 Ordinary Least Squares (OLS) regression

Ordinary Least Squares (OLS), commonly referred to as linear regression, is used to describe the covariation between a variable of interest (response variable) and one or more other variables (explanatory variable(s) or predictor(s)). OLS of y on x can be illustrated as follows (Hirsch, 1982):

$$\hat{y}_i = a + bx_i, \quad (1)$$

where $\hat{y}_i$ is the y estimates for $i = 1, ...n_1$, a is the intercept and b is the slope of the regression equation. In OLS, estimates of the intercept and slope are to minimise the squared error in the estimated $\hat{y}$ values. By solving normal equations, the intercept and slope optimal estimates can be defined as follows (Draper and Smith, 1966):

$$\hat{y}_i = \bar{y}_c + r(s_{y_c}/s_{x_c})(x_i - \bar{x}_c); \quad (2)$$

where $\bar{y}_c$ and $\bar{x}_c$ are the mean values of y_c and x_c, respectively, which represent the series of the concurrent records ($i = 1, ..., n_1$); s_{y_c} is the standard deviation of y_c; s_{x_c} is the standard deviations of x_c; and r is the sample correlation coefficient between y_c and x_c. The OLS has the properties of being unbiased with a small mean square error $\mathrm{MSE} = \sigma_y^2\ (1 - \rho_{xy})$ (Koutsoyiannis and Langousis, 2011), where σ_y^2 is the y population variance and ρ_{xy} is the population correlation coefficient between x and y. It should be emphasized that OLS has five assumptions (Helsel and Hirsch, 2002): y and x are linearly dependent; the data used to fit the model are representative; the variance of the residuals is constant; and the residuals are independent and normally distributed. If the assessment of uncertainty or confidence intervals is of concern, statistical hypothesis should be introduced (Serinaldi, et al., 2012). In this case, the last three assumptions must be fulfilled.

Water quality data sets are commonly characterized by the presence of outliers and skewed distributions, which are not the ideal characteristics for the application of parametric statistical techniques (Hirsch et al., 1991; Helsel and Hirsch, 2002; Granato, 2006). The slope and intercept in the OLS techniques rely on the means and sum of squares of the y_c and x_c, which are significantly affected by the presence of outliers (Helsel and Hirsch, 2002; Granato, 2006). In addition, underestimation of the extended records variability may result in underestimation of high percentiles and overestimation of low percentiles (Khalil et al., 2010), which consequently may affect compliance with standards assessment.

2.2 Kendall-Theil Robust Line (KTRL)

In contrast to OLS, the Kendall-Theil robust line (KTRL) is not strongly affected by outliers (Helsel and Hirsch, 2002). The KTRL robust slope estimator was first described by Theil (1950), its asymptotic properties were studied by Sen (1968), and it is also known as Sen's slope. The Kendall-Theil slope estimate is calculated as the median of all possible slopes computed from each data pair. An n-element data set of (x, y) pairs will result in $n\ (n-1)/2$ pair-wise comparisons. For each data pair, a slope $\Delta y/\Delta x$ is calculated and the nonparametric slope estimate (b_K) is the median of all possible pair-wise slopes (Theil, 1950):

$$b_K = \text{median}\ \frac{y_j - y_i}{x_j - x_i} \quad \forall i < j$$

$$i = 1, 2...n_1 - 1 \quad j = 2, 3...n_1. \quad (3)$$

As for the intercept, several estimates have been proposed in the literature for the KTRL. For instance, Theil (1950) proposed an intercept as the median of the term $(y_i - b_K x_i)$ computed using each data pair. Conover (1980) proposed an intercept computed using the b_K and the y_c and x_c median values. It was concluded by Dietz (1987) that the Conover (1980) intercept estimate was more robust than other estimates for the KTRL. The Conover intercept estimate was recommended by Helsel and Hirsch (2002) for its robustness, efficiency, and easy to calculate. Thus, the KTRL intercept (a_K) is defined as follows (Conover, 1980):

$$a_K = \text{median}\ (y_c) - b_K\ \text{median}\ (x_c). \quad (4)$$

This formula ensures that the KTRL line passes through the point (median (x), median (y)) (Helsel and Hirsch, 2002), which can be considered as an analogue to OLS, where the OLS line passes through the point (mean (x), mean (y)). As described in Helsel and Hirsch (2002), b from OLS and b_K from KTRL are both unbiased estimators of the slope of a linear relationship. However, on one hand, when the residuals follow the normal distribution, OLS is slightly more efficient than KTRL. On the other hand, when the residuals do not follow normal distribution, then b_K is much more efficient than b (Hirsch et al., 1991; Helsel and Hirsch, 2002).

2.3 Line of Organic Correlation (LOC)

The main advantage of the line of organic correlation (LOC) is that it maintains the variance and cumulative distribution function of the extended records. The goal guiding to the development of the LOC was to estimate the intercept and slope in the regression equation to fulfil the following criteria (Hirsch, 1982):

$$\sum_{i=1}^{n_1} \hat{y}_i = \sum_{i=1}^{n_1} y_i \quad (5)$$

$$\sum_{i=1}^{n_1} (\hat{y} - \bar{y}_c)^2 = \sum_{i=1}^{n_1} (y_i - \bar{y}_c)^2\ . \quad (6)$$

One such solution is (Hirsch, 1982):

$$\hat{y}_i = \bar{y}_c + \text{sign}(r)(s_{y_c}/s_{x_c})(x_i - \bar{x}_c), \quad (7)$$

where "sign" (r) stands for the algebraic sign (+ or −) of the correlation coefficient. The LOC has also been called the "maintenance of variance extension" or MOVE (Hirsch, 1982), and also the "geometric mean functional regression" (Halfon, 1985). Hirsch (1982) carried out a Monte Carlo experiment to evaluate the OLS and LOC for bias and standard error of extreme-order statistics. Results of the Monte Carlo experiment showed that LOC produces time series with properties almost similar to the properties of the observed records, while OLS provided records with underestimated variability. However, similar to the OLS regression, the slope and intercept of LOC rely on the sample (y_c and x_c) mean and standard deviation values, which are significantly affected by the presence of outliers.

Similar to the OLS, the LOC is unbiased but with relatively higher MSE ($\mathrm{MSE} = 2\sigma_y^2(1 - |\rho_{xy}|)$) (Koutsoyiannis

and Langousis, 2011). As described by Koutsoyiannis and Langousis (2011), when the $|\rho_{xy}|$ is less than 0.5, the LOC results in an MSE greater than the population variance, which can be considered as a threshold below which the LOC becomes pointless if used to substitute missing values.

2.4 Robust Line of Organic Correlation (RLOC)

The presence of outliers may affect the estimation of the intercept and slope of the OLS and LOC techniques. In addition, when using OLS or KTRL, the under estimation of the variance in the extended records may affect the estimation of extreme percentiles and consequently affect the assessment of compliance with standards or permissible limits. Consequently, a record extension technique that is robust for the presence of outliers and at the same time maintains the variance in the extended records is required. Thus, it is necessary either to modify the KTRL to be able to maintain variance in the extended records, or to modify the LOC to be robust in the presence of outliers. In this section the robust line of organic correlation (RLOC) is proposed as a modified version of the LOC with the advantage of being robust in the presence of outliers.

Presence of high or low outliers has a larger effect on computing the mean than on computing the median. The mean is very sensitive to the presence of outliers, while the median, or the 50th percentile, is slightly influenced by the presence of outliers (Helsel and Hirsch, 2002). Similarly, the sample variance is strongly influenced by outlying values. Since the variance is based on the squares of the deviations from the mean, the variance magnitude may be more influenced by the presence of outliers than the mean. A variance value computed in the presence of outliers may give an indication of greater spread than actually indicated by the majority of the data. The most frequently used outlier-resistant measure of spread is the inter-quartile range (IQR) (Helsel and Hirsch, 2002). The IQR is computed as the range of the central 50 percent of the data (75th percentile minus the 25th percentile) and is not affected by the 25 percent on both ends.

The proposed technique (RLOC) follows the LOC, with a modification of the intercept and slope estimators. The goal guiding to the development of the RLOC was to estimate the intercept and slope estimators in such a way that they become robust in handling the presence of outliers. Given a normal distribution ($N(\mu, \sigma)$) with mean (μ) and standard deviation (σ), the 25th and 75th percentiles ($y_{(25)}$ and $y_{(75)}$) are defined as follows:

$$y_{(25)} = z_1\sigma + \mu \tag{8}$$

$$y_{(75)} = z_3\sigma + \mu, \tag{9}$$

where z_1 and z_3 are the standard scores equal to -0.6745 and 0.6745, respectively. Thus, the IQR is:

$$\mathrm{IQR} = y_{(75)} - y_{(25)} = (z_3 - z_1)\sigma \approx 1.35\,\sigma. \tag{10}$$

In the RLOC, the slope (b_R) is equal to the IQR ratio of y_c to x_c, which is equivalent to the slope of the LOC with the advantage of being robust for the presence of outliers. As for the RLOC intercept estimator (a_R), the Conover (1980) estimator was followed using the RLOC slope estimator. Thus, b_R and a_R are defined as follows:

$$b_\mathrm{R} = \text{ sign } (r)\,\frac{y_{\mathrm{c}(75)} - y_{\mathrm{c}(25)}}{x_{\mathrm{c}(75)} - x_{\mathrm{c}(25)}} \tag{11}$$

$$a_\mathrm{R} = \text{ median } (y_\mathrm{c}) - b_\mathrm{R} \text{ median } (x_\mathrm{c}) \tag{12}$$

where $y_{\mathrm{c}(75)}$, $y_{\mathrm{c}(25)}$, $x_{\mathrm{c}(75)}$ and $x_{\mathrm{c}(25)}$ are the 75th and 25th percentiles of y_c and x_c estimated during the period of concurrent records. Thus, both the intercept and slope estimators of the RLOC are robust in the presence of outliers and also censored records. In addition, using such estimators takes advantage of the nonparametric technique in which the predictor does not take a predetermined form (e.g. normal distribution), but is constructed according to information derived from the data.

To illustrate the impact of the departure from normality, the LOC and RLOC slope estimators were examined under deviation from normality using a combination of two normal distributions. The primary or main distribution has a mean value equal to 10 and a standard deviation equal to 1. The secondary distribution has a mean value equal to 11 and a standard deviation equal to 3. Different mixture distributions (each of 100 samples) were generated containing between 100 and 80 percent of the main distribution and between 0 and 20 percent of the second distribution. Each mixture distribution was treated as the response variable in a regression, while the predictor was a generated random order variable. Thus, the true population slope is zero. The slope estimators of the LOC and RLOC techniques were calculated and their standard deviations around zero recorded as root mean squared error (RMSE). The ratio of the RMSE for the RLOC estimator to that of the LOC estimator was plotted for each distribution mixture (Fig. 1). A value larger than 1 indicates that the LOC estimate is superior, while a value smaller than 1 indicates that the RLOC estimate is superior.

As shown in Fig. 1, the LOC and RLOC estimators have almost the same error when the data are not mixed (normal distribution). However, the RLOC estimator showed better efficiency with small amounts of mixtures. For the distribution mixture, which consists of 80 % of the main distribution and 20 % of the secondary distribution, the RLOC estimator was about 30 % more efficient than the LOC slope estimator.

It should be emphasized that when the intercept is negative, sometimes these record extension techniques may produce negative values of y. As explained by Koutsoyiannis and Langousis (2011), if the OLS intercept is negative, it may sometimes produce negative values or ignore values less than the intercept if it is positive. For the LOC and RLOC, the intercept becomes negative if y is directly proportional to x and at the same time the coefficient of variation of y is larger than

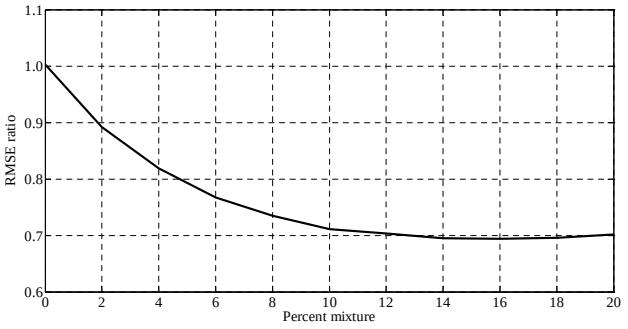

Fig. 1. Relative efficiency of the RLOC slope estimator as compared with the LOC slope estimator; the population is composed of a mixture of two normal distributions ($N(10,1)$ and $N(11,3)$); the x-axis tracks the percentage of the second distribution in the population; the y-axis represents the RMSE ratio = $\mathrm{RMSE_{RLOC}}/\mathrm{RMSE_{LOC}}$.

that of x. The occurrence of these two cases results in a negative intercept; however, the presence of a negative intercept does not necessarily lead to negative y-values.

3 Evaluation experiments

Monte Carlo and empirical experiments were conducted to evaluate the four record extension techniques. A Monte Carlo experiment allows for the comparison and evaluation of the four record extension techniques using records with predefined distributions and statistical properties. The empirical experiment permits evaluation of the four record extension techniques using real water quality data.

3.1 Monte Carlo experiment

In the Monte Carlo experiment, the x and y variable sequences of 120 cases (the same number of records available in the empirical study) were generated from a bivariate normal distribution with $\mu_x = \mu_y = 0$ and $\sigma_x^2 = \sigma_y^2 = 1$. Three cross-correlation coefficients ($\rho = 0.5; 0.7$ and 0.9) were considered. A correlation coefficient of 0.5 was selected to represent the threshold below which the MSE becomes larger than the variance when the LOC is used (Koutsoyiannis and Langousis, 2011). This allows assessing the performance of the modified version (RLOC) with respect to the LOC for the substitution of missing values. The correlation coefficients of 0.7 and 0.9 represent the range within which the correlation coefficients in the empirical experiment were observed (see next section). Different combinations of the number of records during the concurrent period (n_1) and the period to be estimated (n_2) were considered. The Monte Carlo experiments were carried out for (n_1, n_2) values of (96, 24), (72, 48), (48, 72) and (24, 96) and for the three correlation coefficient values. Monte Carlo experiments of 12 different combinations of ρ and (n_1, n_2) were conducted to assess the capability of the four record-extension techniques to extend records that reproduce the different statistical characteristics of the observed records. The estimation of the mean, standard deviation and the 5th to the 95th percentiles from the extended series was evaluated based on those estimated from the observed series.

3.2 Empirical experiment

In the empirical experiment, data from the Edko drainage system water quality monitoring network in the Nile Delta in Egypt were used. The Edko drainage system is one of the main drainage systems in the Nile delta. The Edko catchment area is about 96 000 ha (960 km^2) and its length is 48.8 km starting from Shubra-Kheit, flowing freely into Lake Edko then to the Mediterranean (El-Saadi, 2006). The Edko drainage system is covered by 11 water quality monitoring locations (Fig. 2) where monthly samples have been taken since August 1997. Ten years of monthly water quality records for Electric Conductivity (EC) and Chloride (Cl) measured at the 11 monitoring locations were used in the empirical experiment. The EC is used as an explanatory variable to extend the Cl records using the four record extension techniques. Preliminary analysis of the EC and Cl data at the 11 monitoring locations indicates the presence of outliers and that most of the variables are positively skewed (Fig. 2).

The experiment was designed to assess the usefulness of the four record-extension techniques for maintaining the statistical characteristics of the Cl data. Assessment of the usefulness of the four record-extension techniques was carried out using a split-sample cross validation method because it will provide a more general assessment of the techniques' performance than may be provided by the simple split-sample validation method. In the split sample cross validation, one year of monthly records was eliminated from the available ten years of data. The monthly values for the removed year were then estimated using the four record-extension techniques calibrated with the remaining nine years. At each of the Edko drain 11 monitoring locations, the four record-extension techniques were applied to estimate Cl using EC as a predictor. Thus, 110 (11 locations $\times$ 10 different samples combinations = 110) different realisations of extended Cl records were generated. For each trial, the extended series was evaluated based on the estimation of the mean, standard deviation and over the full range of percentiles (from the 5th to the 95th percentile). The correlation coefficient of y_c and x_c was computed for each of the 110 different realisations considered. Results showed that the correlation coefficient was always positive and ranges between 0.73 and 0.92.

3.3 Evaluation procedures

Record-extension techniques are commonly applied to extend streamflow records at short-gauged stations using the logarithm of the streamflow records. In general, transformed

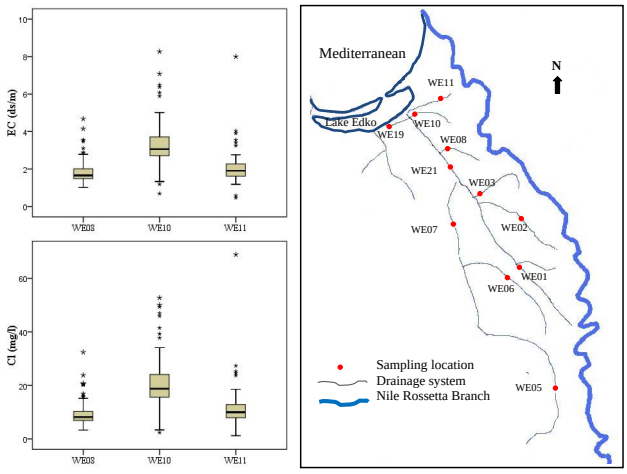

Fig. 2. Edko drain sampling locations and box-plots for Electric Conductivity (EC) and Chloride (Cl) records at three locations.

data were used instead of the raw records to improve the normality, when the data showed a strongly positive skew (Helsel and Hirsch, 2002; Granato, 2006).

In general, water quality data are usually positively skewed due to the presence of positive outliers (Lettenmaier, 1988; Berryman et al., 1988). Preliminary analysis for Edko drain data also confirms the presence of outliers and positive skewness. For the Edko drain data, preliminary analysis confirmed the linear dependency between EC and Cl, and that the data are serially independent and homoscedastic. In addition, preliminary analysis did not confirm any significant cycle or seasonal pattern in the data. As an example of the preliminary data analysis that was carried out, Fig. 3 shows the scatter plot for EC and Cl measured at WE11 as well as their probability density plots. Figure 3 shows a clear linear dependency between the EC and Cl (scatter plots), and also shows that both EC and Cl are positively skewed (probability density plots). For the log-transformed data, the probability density plots show a symmetric distribution. In addition, the Kolmogorov-Smirnov goodness-of-fit test was applied to test normality, where the null hypothesis is that the sample is drawn from the normal distribution. The test results show that for the raw data, the test null hypothesis cannot be accepted, while it is accepted for the log-transformed data (Table 1). Figure 4 shows the correlograms for EC and Cl, which indicate that the data are independent. In addition, although a set of positive autocorrelation values are followed by a set of negative autocorrelation values that may indicate seasonality (Fig. 4), these autocorrelation values are not significant. In the case of seasonality, the record-extension techniques can be applied to the data of each season or month as recommended by Alley and Burns (1983).

Khalil et al. (2010, 2011) applied record-extension techniques to the log-transformed data, while performance measures were computed based on the back-transformed estimated records. Similarly, in this study, the back-transformed extended series were compared to the observed series based on the estimation of different statistical parameters. It should be emphasized that although an appropriate transformation may be required to return normally distributed data for applying parametric techniques, the symmetry of the marginal distribution may be considered sufficient when applying the RLOC technique. However, for comparison purposes, in this study the four techniques under comparison were applied on the log-transformed data.

Two performance measures were used to evaluate the performances of the four record-extension techniques. These are the bias (BIAS) as a measure of accuracy and the root mean squared error (RMSE) as a measure of precision, which can be defined as follows:

$$\mathrm{BIAS} = \frac{1}{m}\sum_{i=1}^{m}\hat{S}_i - S_i \tag{13}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\hat{S}_i - S_i\right)^2}, \tag{14}$$

where $\hat{S}_i$ is the estimated statistics and S_i is the observed statistics of the response variable for $i = 1, \ldots m$, where m is the number of trials in the Monte Carlo or the empirical study.

It should be noted that mean square error is the second moment of the error and it incorporates both the variance of the estimate and its bias. Thus, by simultaneously examining the RMSE and the BIAS, one can assess if the error results more from the estimation variability or rather from the bias made on the estimate (Chokmani et al., 2008). Aside from computing the BIAS and RMSE for the estimated statistics, both measures were also applied to compare the extended records with the observed records. In this case the summation in equations 13 and 14 was for $(\hat{y}_i - y_i)$ from $i = n_1 + 1$ to $i = n_1 + n_2$, which is the size of extended records.

4 Results

4.1 Monte Carlo experiment results

In the Monte-Carlo experiment, 5000 trials were generated. This number of generated trials was selected based on a preanalysis carried out to examine the convergence of the error in estimating different statistics. The BIAS and RMSE values for the extended records are presented in Table 2. The values presented in Table 2 are the average values computed based on the 5000 trials. Results show that the hypothesis that the BIAS value is equal to zero could not be rejected at the 0.05 significance level for any of the extension techniques under any of the 12 designed combinations. For the RMSE, those

Table 1. Kolmogorov-Smirnov goodness of fit test for Electric Conductivity (EC) and Chloride (Cl) measured at WE11.

Data		Raw data		log-transformed	
Water quality variable		EC	Cl	EC	Cl
Number of samples		118	119	118	119
Mean		2.015	11.213	0.649	2.304
Standard deviation		0.772	6.863	0.315	0.477
Most Extreme Differences	Absolute	0.177	0.169	0.123	0.099
	Positive	0.177	0.166	0.105	0.082
	Negative	−0.174	−0.169	−0.123	−0.099
Kolmogorov-Smirnov Z-value		1.923	1.842	1.340	1.075
Probability of accepting the null hypothesis (p-value)		0.001	0.002	0.055	0.198

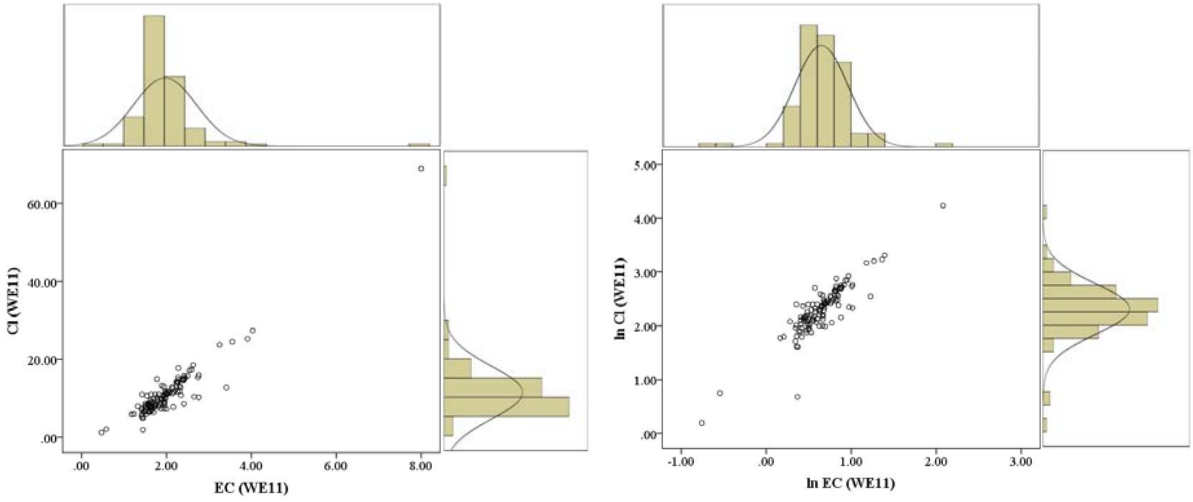

Fig. 3. Scatter plots and probability density plots for Electric Conductivity (EC) and Chloride (Cl) measured at WE11.

corresponding to the OLS are relatively lower than those corresponding to any of the other three techniques, while those corresponding to the KTRL are relatively lower than those corresponding to the LOC or RLOC. Similarly, RMSEs corresponding to the LOC are relatively lower than those corresponding to the RLOC. These results indicate that the four techniques are unbiased. However, the OLS is the most precise, followed by the KTRL. The margin of error exhibited by the KTRL as compared to the OLS is almost equal to that exhibited by the RLOC as compared to the LOC. These results indicate that when the objective is to substitute missing records, and the data show a linear pattern, constant variance and normality of residuals, the OLS is favorable.

The BIAS values for the estimation of the mean and standard deviation are shown in Table 3. For the estimation of the mean value, the hypothesis that the BIAS value is equal to zero could not be rejected at the 0.05 significance level for any of the extension techniques under any of the 12 designed combinations. In addition, the results show that there is no significant difference in the BIAS values between the four record-extension techniques and under any of the 12 different combinations considered. These results would be expected since the OLS and LOC lines pass through the point representing the mean values of the response and predictor. In the same manner, the KTRL and RLOC lines pass through the point representing the median values of the response and predictor, which is the same point representing the mean values, given that the data were generated from a bivariate normal distribution.

For the estimation of the standard deviation, when using the OLS or KTRL, the hypothesis that the BIAS value is equal to zero is rejected at the 0.05 significance level for all of the 12 designed combinations. When using the LOC or RLOC techniques, the hypothesis that the BIAS values are equal to zero cannot be rejected for most of the combinations. Using either the OLS or KTRL technique, the results show a significant underestimation of the standard deviation under any of the 12 different combinations of ρ and (n_1, n_2). Using the LOC technique, BIAS values ranged between −0.002

Table 2. BIAS and RMSE for the extended records (Monte-Carlo experiment).

n_1	n_2	ρ	BIAS				RMSE			
			OLS	LOC	KTRL	RLOC	OLS	LOC	KTRL	RLOC
96	24	0.5	0.000	0.000	0.001	0.003	0.837	0.956	0.869	1.004
72	48	0.5	0.000	0.000	0.004	0.007	0.854	0.977	0.879	1.016
48	72	0.5	0.000	0.000	0.001	0.001	0.863	0.986	0.890	1.029
24	96	0.5	0.000	0.000	0.001	0.000	0.872	0.990	0.919	1.061
96	24	0.7	0.000	0.000	0.001	0.002	0.684	0.736	0.718	0.779
72	48	0.7	0.000	0.000	−0.002	−0.001	0.704	0.759	0.730	0.792
48	72	0.7	0.000	0.000	0.004	0.004	0.708	0.763	0.736	0.799
24	96	0.7	0.000	0.000	0.002	0.001	0.716	0.766	0.764	0.826
96	24	0.9	0.000	0.000	0.006	0.006	0.416	0.426	0.444	0.456
72	48	0.9	0.000	0.000	−0.001	−0.001	0.425	0.434	0.448	0.458
48	72	0.9	0.000	0.000	0.000	0.000	0.429	0.439	0.455	0.468
24	96	0.9	0.000	0.000	−0.001	−0.002	0.431	0.440	0.472	0.484

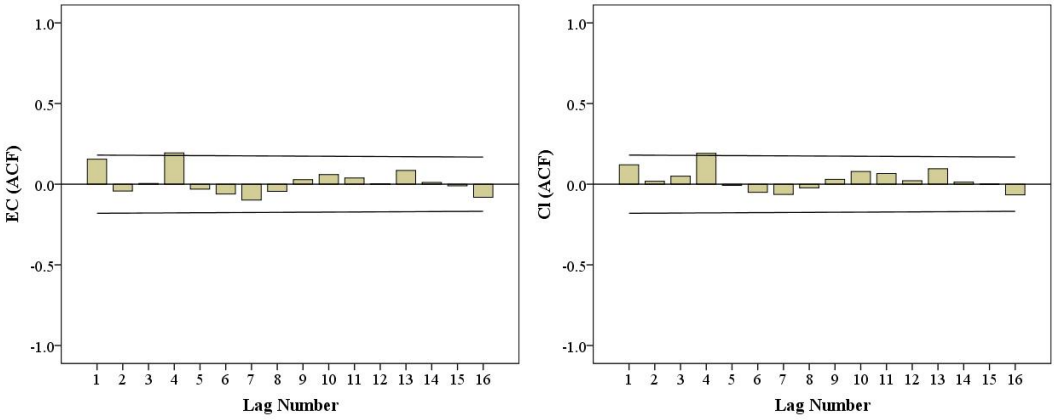

Fig. 4. Correlograms for Electric Conductivity (EC) and Chloride (Cl) measured at WE11.

and 0.008; they ranged between −0.001 and 0.010 when using the RLOC.

Table 3 shows that the BIAS values for the estimation of the standard deviation decrease with an increase in the correlation coefficient value and/or an increase of the size of the concurrent records (n_1). These results may indicate that when the size of the concurrent records is large enough with a high level of association, use of either the LOC or RLOC will estimate the standard deviation with high accuracy. In general, Table 3 shows that BIAS values corresponding to the LOC and RLOC are closer to zero than those corresponding to the OLS and KTRL under any of the 12 designed combinations.

Table 4 shows the RMSE values for the estimation of the statistical moments using the four record extension techniques for each of the 12 designed combinations of n_1, n_2 and ρ. From Table 4, it can be seen that the RMSE values decrease with an increase in the correlation coefficient value and/or the size of the concurrent records for the mean or standard deviation. For the standard deviation, using either the LOC or RLOC, the RMSE values are less than the values obtained when using the OLS or KTRL.

Figure 5 shows the BIAS values for the estimation of the non-exceedance percentiles using the four record extension techniques. In Fig. 5, six figures representing the two extreme cases (n_1 = 96 and n_1 = 24) under each of the three correlation coefficients considered are presented.

In general, Fig. 5 shows that when using OLS or KTRL, one may expect an overestimation of low percentiles and an underestimation of high percentiles. When using the LOC or RLOC techniques, these biases in the estimation of extreme percentiles were significantly reduced. Given that the data follow a normal distribution, underestimation of the variance leads to underestimation of high values and overestimation of low values, which leads to the shown bias in the estimation of extreme percentiles.

In the case where n_1 is equal to 96, the results obtained from using OLS show that the BIAS ranges between −0.14 and 0.14 when ρ is equal to 0.5, between −0.09 and 0.09 for ρ equal to 0.7, and between −0.04 and 0.04 for ρ equal to

Table 3. BIAS values for the estimation of the mean and the standard deviation (Monte-Carlo experiment).

n_1	n_2	ρ	BIAS				Standard deviation			
			OLS	LOC	KTRL	RLOC	OLS	LOC	KTRL	RLOC
96	24	0.5	−0.002	−0.001	−0.001	−0.001	−0.081*	−0.002	−0.080*	−0.001
78	48	0.5	−0.001	−0.001	−0.001	0.000	−0.167*	−0.001	−0.165*	0.001
48	78	0.5	−0.001	0.000	0.001	0.002	−0.261*	0.001	−0.259*	0.003*
24	96	0.5	−0.001	0.000	0.000	0.001	−0.369*	0.005*	−0.364*	0.009*
96	24	0.7	0.000	0.000	0.000	0.000	−0.055*	−0.001	−0.054*	0.000
78	48	0.7	0.000	0.001	0.000	0.001	−0.111*	−0.001	−0.109*	0.001
48	78	0.7	0.002	0.003	0.002	0.003	−0.170*	0.002	−0.168*	0.004*
24	96	0.7	0.004	0.006	0.003	0.004	−0.234*	0.008*	−0.231*	0.010*
96	24	0.9	−0.001	−0.001	−0.001	−0.001	−0.020*	−0.001	−0.020*	0.000
78	48	0.9	−0.001	−0.001	−0.001	−0.001	−0.041*	−0.001	−0.040*	0.000
48	72	0.9	0.000	−0.001	−0.001	−0.001	−0.061*	0.000	−0.059*	0.001
24	96	0.9	0.001	0.000	0.000	−0.001	−0.082*	0.002	−0.081*	0.003*

* The hypothesis that the BIAS is equal to zero is rejected at the 5 % level.

Table 4. RMSE values for the estimation of the mean and the standard deviation (Monte-Carlo experiment).

n_1	n_2	ρ	BIAS				Standard deviation			
			OLS	LOC	KTRL	RLOC	OLS	LOC	KTRL	RLOC
96	24	0.5	0.099	0.102	0.101	0.104	0.107	0.076	0.106	0.076
72	48	0.5	0.112	0.118	0.118	0.127	0.184	0.093	0.183	0.093
48	72	0.5	0.134	0.144	0.150	0.169	0.278	0.119	0.277	0.119
24	96	0.5	0.183	0.204	0.224	0.258	0.398	0.180	0.396	0.181
96	24	0.7	0.097	0.098	0.099	0.100	0.090	0.073	0.090	0.074
72	48	0.7	0.107	0.109	0.114	0.118	0.137	0.085	0.136	0.085
48	72	0.7	0.122	0.126	0.143	0.151	0.197	0.104	0.196	0.104
24	96	0.7	0.161	0.170	0.205	0.224	0.275	0.153	0.276	0.153
96	24	0.9	0.094	0.094	0.095	0.095	0.071	0.068	0.071	0.068
72	48	0.9	0.097	0.097	0.104	0.104	0.084	0.073	0.085	0.073
48	72	0.9	0.103	0.104	0.123	0.125	0.102	0.081	0.103	0.081
24	96	0.9	0.121	0.122	0.166	0.170	0.136	0.106	0.139	0.106

0.9. In the case where n_1 is equal to 24 records, the BIAS value when using OLS ranges between −0.6 and 0.6 for ρ equal to 0.5. For ρ equal to 0.7, the BIAS value ranges between −0.4 and 0.4, and between −0.2 and 0.2 for ρ equal to 0.9. For the other two cases (n_1 is equal to 24 and 48 records) not presented in this figure, the BIAS values were in-between the presented cases. The BIAS values corresponding to the LOC and RLOC are closer to zero than those corresponding to the OLS or KTRL under all of the combinations considered. Figure 5 shows that in general, under all combinations considered, use of LOC or RLOC will produce similar results, and both techniques reduce the bias exhibited by the OLS or KTRL when estimating extreme percentiles.

Similarly, Fig. 6 shows the RMSE values for the same six ρ, n_1 and n_2 combinations presented in Fig. 5. When using the RLOC, in the case where n_1 is equal to 96 and ρ is equal to 0.5, the RMSE value for the estimation of extreme percentiles (5th and 95th percentiles) reaches 0.14, whereas it is 0.33 when n_1 is equal to 24. When ρ is equal to 0.7, the RMSE value is 0.135 and only 0.13 when ρ is equal to 0.9, while when n_1 is equal to 24, the RMSE values are 0.23 and 0.21 respectively. These results indicate that the RMSE decreases with an increase in the size of concurrent records (n_1) and/or the correlation coefficient. In general, for the four record-extension techniques, plots in Fig. 6 show that for the estimation of percentiles, precision increases as the correlation coefficient (ρ) and/or the size of available records during the concurrent period (n_1) increases.

When n_1 is equal to 96 and ρ is equal to 0.9, Fig. 6 (the bottom right plot) shows that there is no difference between the four record-extension techniques for the estimation of any of the percentiles considered. When n_1 is equal to 24, Fig. 6 (the left column plots) shows that extended records produced when using the LOC provide more precise estimations of extreme percentiles than the OLS, KTRL and RLOC extended records. When n_1 is equal to 24 and ρ is equal to

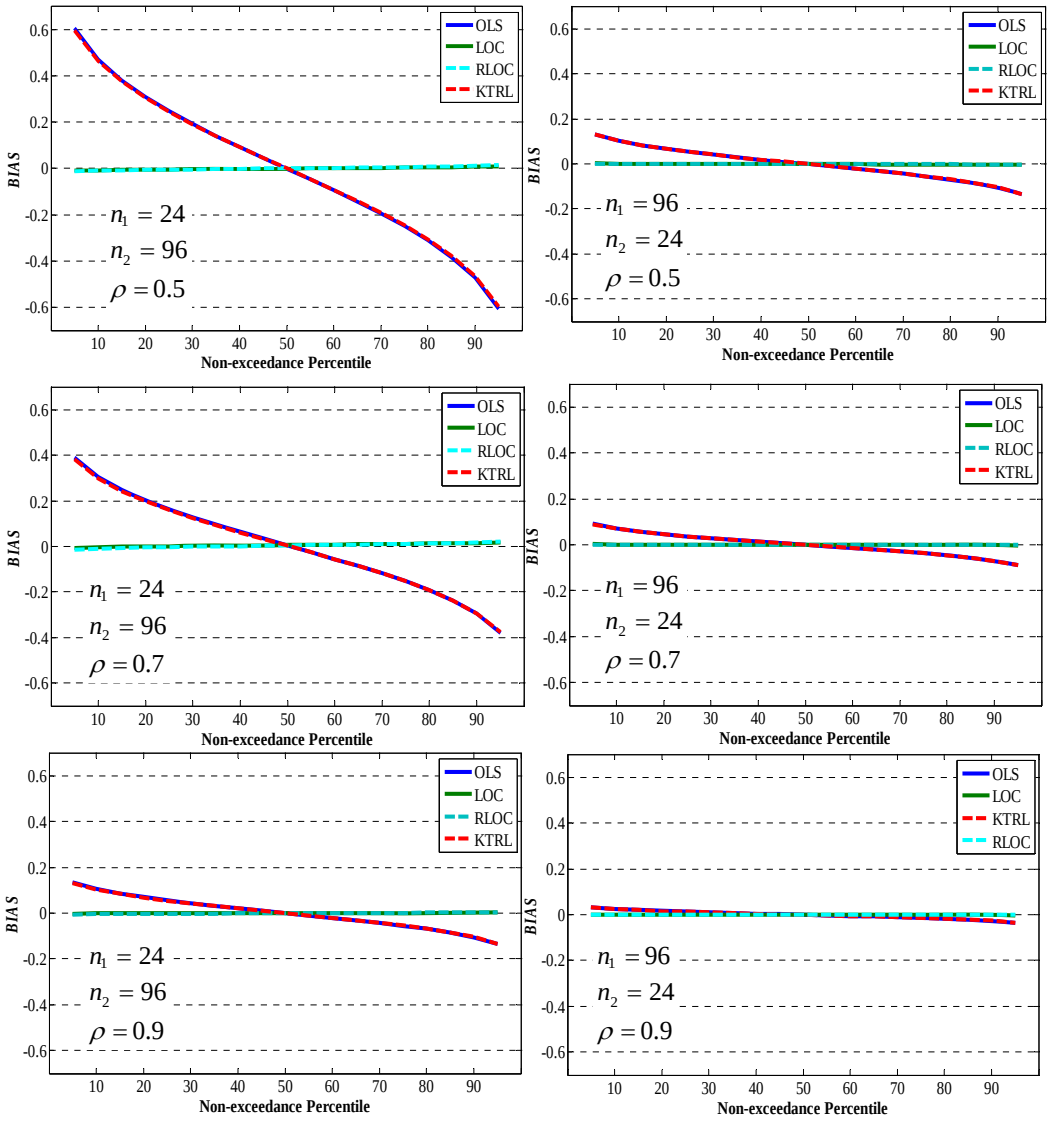

Fig. 5. BIAS values for the estimation of the non-exceedance percentiles (Monte-Carlo experiment).

0.9, the OLS outperforms the RLOC for the estimation of percentiles in the range from the 10th to the 90th percentiles. These results indicate that the RLOC requires larger sample sizes than regression techniques based on parametric models (OLS and LOC) because the limited size of concurrent records may not support the estimation of the RLOC slope estimator.

Thus, in summary, the Monte Carlo study results show that RLOC can be considered as an analogue of LOC. The main advantage of LOC and RLOC over OLS is that the cumulative distribution function of the forecasts estimates those of the observed records that they were estimated to represent. However, when the objective is to substitute individual missing records, the OLS and KTRL are preferable. Consequently, LOC and RLOC are preferable in cases where the probability distribution of the extended records is to be inferred and used. For the RLOC, the Monte Carlo experiment shows that it is as accurate as the LOC for the estimation of the standard deviation and extreme percentiles but not as precise as the LOC when a small number of records is available.

To confirm the impact of the size of the concurrent records on the performance of the RLOC, another Monte Carlo experiment was considered. In this experiment, the estimation of the IQR was evaluated based on different sizes of the concurrent records. A 5000 case time series was generated from a normal distribution with $\mu = 0$ and $\sigma^2 = 1$. This time series is considered as a population. A set of 1000 different subsamples were generated from the original sample (population) using sampling with the replacement technique for each of the following sizes: 12, 24, 36, 48, 60, 72, 84, 96, 108 and 120. For each of the 1000 subsamples representing each

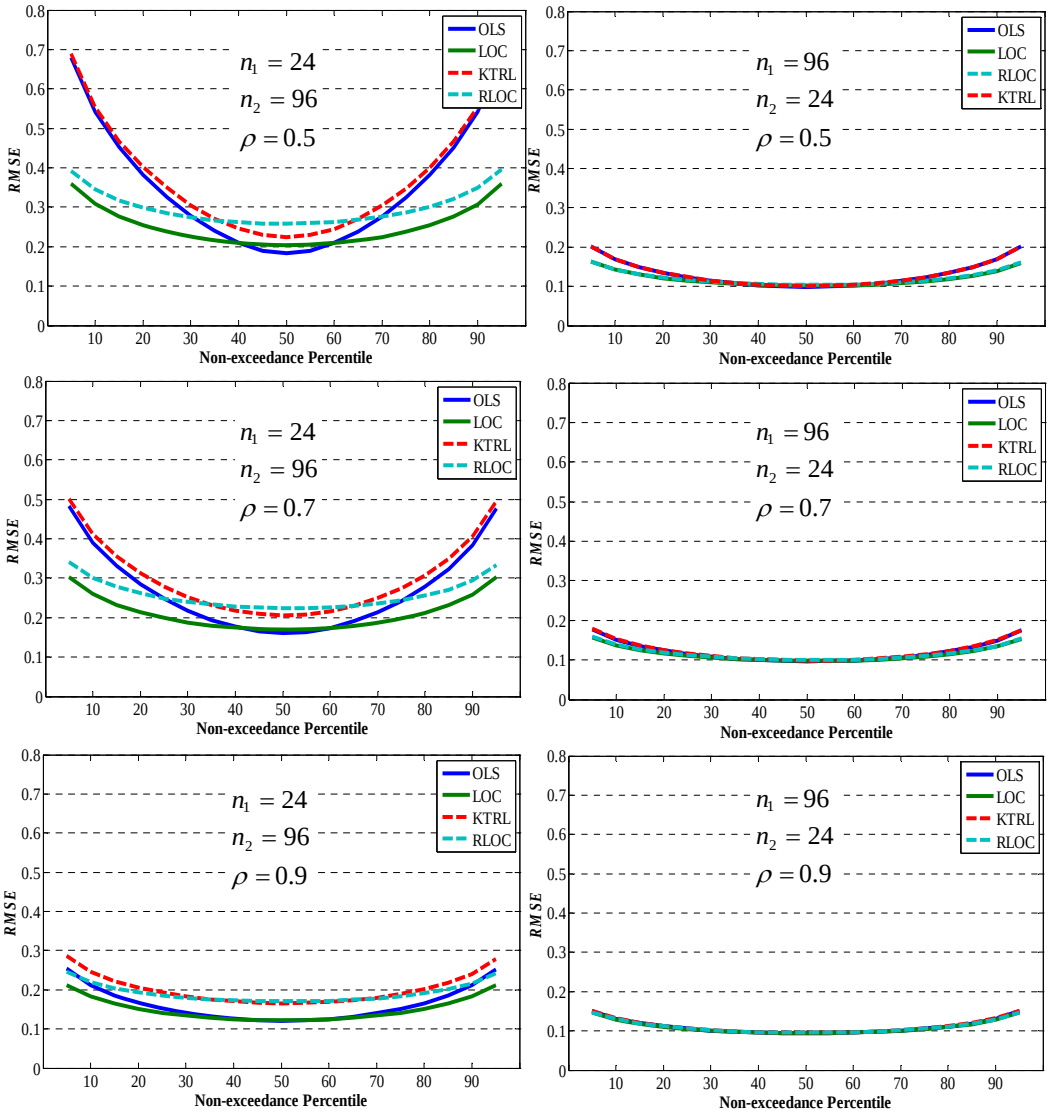

Fig. 6. RMSE values for the estimation of the non-exceedance percentiles (Monte-Carlo experiment).

of the sizes considered, an IQR ratio was computed, which is the ratio of the IQR value estimated from the subsample to that estimated from the original sample (population). An IQR ratio larger than 1 indicates overestimation and less than 1 indicates underestimation. Figure 7 shows box-plots for each of the data size considered. Each box-plot was drawn using the 1000 IQR ratios computed for each of the 1000 subsamples.

The level of accuracy of the IQR estimated from the subsamples is represented by how close the box-plot median value is to 1, while the level of precision is represented by the box-plot dispersion around the median value. From Fig. 7, the median values representing different subsample sizes are all close to 1, which indicates accuracy of the estimation even with a small sample size. However, for small sample sizes the box-plot dispersion around the median is larger than the dispersion exhibited by box-plots corresponding to relatively larger sample sizes. These results confirm results obtained from the Monte Carlo experiment (Figs. 5 and 6) that the RLOC produces records that allow for the estimation of extreme percentiles as accurate as the LOC, but not as precise as the LOC when only limited number of records are available.

4.2 Empirical experiment results

The BIAS and RMSE values for the estimated records, as well as those for the estimation of the mean and standard deviation are shown in Table 5. From Table 5, for the extended records, the BIAS values were almost comparable, while the lowest RMSE was observed when the KTRL was used, and the second lowest RMSE was observed when the RLOC was

Table 5. BIAS and RMSE values for the estimation of the Cl mean and the standard deviation.

Statistic	Metrics	OLS	LOC	KTRL	RLOC
Records	BIAS	0.034	0.033	−0.029	0.033
	RMSE	2.083	2.130	1.925	2.013
Mean	BIAS	0.001	0.001	−0.010	−0.005
	RMSE	0.081	0.085	0.075	0.079
Standard deviation	BIAS	−0.050	0.015	−0.04	0.003
	RMSE	0.086	0.075	0.072	0.065

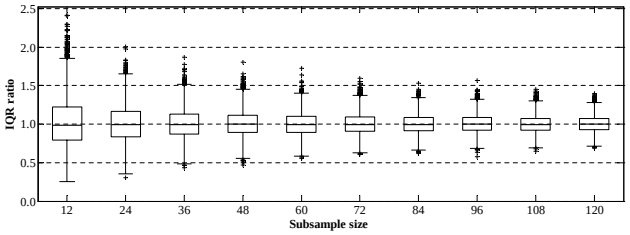

Fig. 7. Box plots of the Inter-quartile Range (IQR) ratio, RLOC technique.

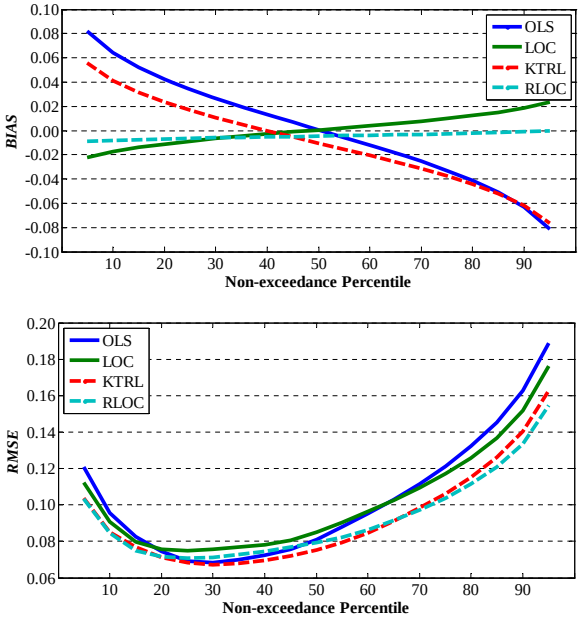

Fig. 8. BIAS and RMSE of the tested extension techniques in estimating Chloride (Cl) percentiles.

used. These results indicate that when the data exhibit outliers, a robust technique is preferable for the substitution of missing values.

For the estimation of the mean, the hypothesis that the BIAS value is equal to zero could not be rejected at the 0.05 significance level for any of the four record extension techniques. Results also show that there is no significant difference between the BIAS or RMSE values corresponding to the four record-extension techniques for the estimation of the mean.

For the estimation of the standard deviation, results showed that an underestimation was obtained when using the OLS or KTRL, while an overestimation was obtained when using the LOC or RLOC. However, the BIAS value corresponding to the RLOC is closer to zero and far below that corresponding to the LOC. This is due to the presence of outliers and/or deviation from normality, where the RLOC is robust while LOC is sensitive. Although RMSE values were almost comparable, when using the RLOC the lowest BIAS and RMSE values were obtained.

Figure 8 shows the BIAS values corresponding to the four record-extension techniques for the estimation of the Cl percentiles. From Fig. 8, when using OLS or KTRL, results showed an overestimation of low percentiles and underestimation of high percentiles. When using the LOC, Fig. 8 shows an underestimation of low percentiles and an overestimation of high percentiles, which is mainly because of the overestimation of the standard deviation. On the other hand, BIAS values for the estimation of the extreme percentiles using the RLOC are close to zero, while those corresponding to the OLS, KTRL and LOC are not. For the estimation of extreme percentiles, RLOC is better than LOC, but both of these are better than OLS and KTRL.

Figure 8 also shows the RMSE values corresponding to the four record-extension techniques for the estimation of the Cl percentiles. Results showed that the RMSE values corresponding to the RLOC are lower than those corresponding to the other three techniques and those corresponding to KTRL are lower than those corresponding to the LOC and OLS. Figure 8 shows that the estimation of Cl percentiles using records extended by KTRL was more precise than those using the LOC, which may be due to the presence of outliers. Figure 8 clearly illustrates that the OLS and KTRL overestimate low percentiles and underestimate of high percentiles, as expected from the tendency of OLS and KTRL to underestimate the variance in the extended records.

Thus, in summary, the results indicate that OLS and KTRL substantially reduce variability and that LOC and RLOC tend to preserve variability in the extended records. The OLS and KTRL techniques underestimate high percentiles and overestimate low percentiles, while the LOC and RLOC techniques reduce the bias in the estimation of both high and low percentiles. Also, both LOC and RLOC produce extended records that preserve extreme percentiles relatively well. However, when the objective is to substitute individual records and not to extend a time series, the OLS and KTRL are preferable. Using water quality data, RLOC outperforms the LOC in the estimation of extreme percentiles. This better performance shown by RLOC arises because real water quality data do not follow a normal distribution, and even after transformation, some deviation from normality may exist. This slight deviation from normality and/or presence of

outliers makes RLOC preferable. However, in the case of small sizes of concurrent records, the LOC outperforms the RLOC, as the RLOC slope estimator requires enough records to be estimated precisely.

5 Conclusions

The OLS, KTRL, LOC and the new RLOC technique were compared in this study using a Monte Carlo and empirical experiments using water quality data from the Edko drainage system, in the Nile Delta of Egypt. BIAS and RMSE were computed to evaluate each of the four record-extension techniques with respect to the errors in the extended records, as well as the extended record means, standard deviations and full range of percentiles. In the assessment of the errors of the extended records using BIAS and RMSE, the Monte Carlo experiment revealed that when OLS assumptions are fulfilled, it outperforms the other three techniques. However, the empirical experiment showed that the KTRL outperformed the OLS, which was mainly due to presence of outliers. Thus, it was concluded that the OLS and KTRL are recommended for substitution of missing records.

For the estimation of extended record statistics, both experiments showed that OLS and KTRL fail significantly to extend records that preserve main statistical characteristics. Mainly, both techniques cannot be expected to extend records with the appropriate variability or the appropriate distribution shape. The evaluation of biases of moments and non-exceedance percentiles showed that LOC and RLOC perform better than OLS and KTRL. The OLS and KTRL substantially underestimated the variance. Consequently, the frequency of extreme events such as exceedance of permissible limits would be underestimated when either the OLS or KTRL was used. On the other hand, the LOC estimates would substantially overestimate the variance. Thus, the frequency of extreme events would be overestimated. However, use of LOC reduces the bias exhibited by the OLS.

The RLOC slope estimator based on the interquartile ratio ensures that estimates of $\hat{y}_i$ from observed x_i have statistical parameters and distributional shape similar to those expected had y_i been measured. Using real water quality data, the empirical experiment showed RLOC to exhibit slightly more desirable properties than LOC. When records are to be extended and inference is to be made about probabilities of exceedance (such as probabilities of exceeding some water quality standard), LOC and RLOC should be used to extend the records rather than OLS or KTRL. RLOC is superior in cases of deviation from normality and/or the presence of outliers.

This study supports the idea that when the data or their transforms show a linear pattern and residuals are normality distributed, the LOC and RLOC techniques will give nearly identical results. However, when deviation from normality and/or presence of outliers is observed, the regression line fitted by the RLOC technique will be more efficient (lower variability and bias) as compared to LOC. The main advantages of the newly proposed RLOC technique are its robustness for handling the presence of outliers, that it maintains the variance of the extended records, and that it is simple to compute and implement.

It is recommended that the newly proposed RLOC technique be further investigated using simulated records with specific characteristics such as different degrees of data contamination, different sizes of concurrent records, deviation from normality, cyclic or seasonal pattern, heterosdasticity and different association levels. Further investigation using different hydrologic data sets from other geographical areas is also recommended. Additionally, a comparison with more advanced techniques such as Generalized Linear Models (GLM) and Generalized Additive Models (GAD) is recommended. Finally, modification of the RLOC to allow using multi predictors is also recommended for further study.

Acknowledgements. The authors are grateful to Shaden Abdel-Gawad, Chairperson of the National Water Research Center of Egypt, for providing the data used in this paper. Financial support provided by "Le Fonds de recherche du Québec – Nature et technologies", as well as an NSERC Discovery Grant is acknowledged. The authors wish to thank the editor, D. Koutsoyiannis, F. Serinaldi, and the anonymous reviewer whose comments and suggestions greatly improved the quality of the paper.

Edited by: D. Koutsoyiannis

References

Albek, E.: Estimation of point and diffuse contaminant loads to streams by non-parametric regression analysis of monitoring data, Water Air Soil Poll., 147, 229–243, 2003.

Alley, W. M. and Burns, A. W.: Mixed-station extension of monthly streamflow records, J. Hydraul. Eng., 109, 1272–1284, 1983.

Berryman, D., Bobée, B., Cluis, D., and Haemmerli, J.: Nonparametric tests for trend detection in water quality time series, Water Resour. Bull., 24, 545–556, 1988.

Chokmani, K., Ouarda, T. B. M. J., Hamilton, S., Ghedira, M. H., and Gingras, H.: Comparison of ice-affected streamflow estimates computed using artificial neural networks and multiple regression techniques, J. Hydrol., 349, 383–396, 2008.

Conover, W. L.: Practical nonparametric statistics, 2nd Edn., John Wiley and Sons, New York, 493 pp., 1980.

Déry, S. J., Mlynowski, T. J., Hernandez-Henriquez, M. A., and Straneo, F.: Interannual variability and interdecadal trends in Hudson Bay streamflow, J. Marine Syst., 88, 341–351, 2011.

Dietz, E. J.: A comparison of robust estimators in simple linear regression, Communication in Statistics-Simulation, 16, 1209–1227, 1987.

Draper, N. R. and Smith, H.: Applied regression analysis, John Wiley, New York, 736 pp., 1966.

El-Saadi, A.: Economics and uncertainty considerations in water quality monitoring networks design, Ph.D. dissertation, Faculty of Engineering, Ain-Shams University, Cairo, Egypt, 2006.

Granato, G. E.: Kendall-Theil Robust Line (KTRLine – version 1), A visual basic program for calculating and graphing robust nonparametric estimates of linear-regression coefficients between two continuous variables: Techniques and Methods of the US Geological Survey, Book 4, Chap. A7, 31 pp., 2006.

Halfon, E.: Regression method in ecotoxicology: A better formulation using the geometric mean functional regression, Environ. Sci. Technol., 19, 747–749, 1985.

Harmancioglu, N. B. and Yevjevich, V.: Transfer of Information among Water Quality Variables of the Potomac River, Phase III: Transferable and Transferred Information, Report to D.C. Water Resources Research Center of the University of the District of Columbia, Washington DC, 81 pp., 1986.

Harmancioglu, N. B. and Yevjevich, V.: Transfer of hydrologic information among river points, J. Hydrol., 91, 103–118, 1987.

Harmancioglu, N. B., Fistikoglu, O., Ozkul, S. D., Singh, V. P., and Alpaslan, M. N.: Water Quality Monitoring Network Design, Kluwer Academic Publishers, Dordrecht, The Netherlands, 290 pp., 1999.

Helsel, D. R. and Hirsch, R. M.: Statistical methods in water resources, Amsterdam, The Netherlands, Elsevier Science Publishers, 522 pp., 2002.

Hirsch, R. M.: A comparison of four streamflow record extension techniques, Water Resour. Res., 18, 1081–1088, 1982.

Hirsch, R. M., Alexander, R., and Smith, R. A.: Selection of methods for the detection and estimation of trends in water quality, Water Resour. Res., 27, 803–813, 1991.

Jia, Y. and Culver, T. B.: Bootstrapped artificial neural networks for synthetic flow generation with a small data sample, J. Hydrol., 331, 580–590, 2006.

Khalil, B. and Ouarda, T. B. M. J: Statistical approaches used to assess and redesign surface water quality monitoring networks, J. Environ. Monitor., 11, 1915–1929, 2009.

Khalil, B., Ouarda, T. B. M. J., St-Hilaire, A., and Chebana, F.: A statistical approach of the rationalization of water quality indicators in surface water quality monitoring networks, J. Hydrol., 386, 173–185, 2010.

Khalil, B., Ouarda, T. B. M. J., and St-Hilaire, A.: A statistical approach for the assessment and redesign of the Nile Delta drainage system water quality monitoring locations, J. Environ. Monitor., 13, 2190–2205, 2011.

Koutsoyiannis, D. and Langousis, A.: Precipitation, Treatise on Water Science, edited by: Wilderer, P. and Uhlenbrook, S., 2, 27–28, Academic Press, Oxford, 2011.

Kritskiy, S. N. and Menkel, J. F.: Some statistical methods in the analysis of hydrologic data, Soviet Hydrology Selected Papers 1, 80–98, 1968.

Kruskal, W. H.: On the uniqueness of the line of organic correlation, Biometrics, 9, 47–58, 1953.

Lettenmaier, D. P.: Multivariate nonparametric tests for trend in water quality, AWRA, Water Resour. Bull., 24, 505–512, 1988.

Moog, D. B. and Whiting, P. J.: Streamflow record extension using power transformations and application to sediment transport, Water Resour. Res., 35, 243–254, 1999.

Nevitt, J. and Tam, H. P.: A comparison of robust and nonparametric estimators under the simple linear regression model, Multiple Linear Regression Viewpoints, 25, 54–69, 1989.

Olson, O., Gassmann, M., Wegerich, K., and Bauer, M.: Identification of the effective water availability from streamflows in the Zerafshan river basin, Central Asia, J. Hydrol., 390, 190–197, 2010.

Raziei, T., Saghafian, B., Paulo, A. A., Pereira, L. S., and Bordi, I.: Spatial patterns and temporal variability of drought in western Iran, Water Resour. Manag., 23, 439–455, 2009.

Raziei, T., Bordi, I., and Pereira, L. S.: An application of GPCC and NCEP/NCAR datasets for draught variability analysis in Iran, Water Resour. Manag., 25, 1075–1086, 2011.

Robinson, R. B., Wood, M. S., Smoot, J. L., and Moore, S. E.: Parametric modelling of water quality and sampling strategy in a high-altitude Appalachian stream, J. Hydrol., 287, 62–73, 2004.

Ryu, J. H., Svoboda, M. D., Lenters, J. D., Tadesse, T., and Knutson, C. L.: Potential extents for ENSO-driven hydrologic drought forecasts in the United States, Climatic Change, 101, 575–597, 2010.

Sen, P. K.: Estimates of the regression coefficient based on Kendall's tau, J. Am. Stat. Assoc., 63, 1379–1389, 1968.

Serinaldi, F., Grimaldi, S., Abdolhosseini, M., Corona, P., and Cimini, D.: Testing copula regression against benchmark models for point and interval estimation of tree wood volume in beech stands, Eur. J. For. Res., online first: doi:10.1007/s10342-012-0600-2, 2012.

Theil, H.: A rank-invariant method of linear and polynomial regression analysis, 1, 2, and 3, Ned. Akad. Wentsch Proc., 53, 386–392, 521–525, and 1397–1412, 1950.

Vogel, R. M. and Stedinger, J. R.: Minimum variance streamflow record augmentation procedures, Water Resour. Res., 21, 715–723, 1985.