World Academy of Science, Engineering and Technology
International Journal of Environmental and Ecological Engineering
Vol:3, No:9, 2009

Reconstitute Information about Discontinued Water Quality Variables in the Nile Delta Monitoring Network Using Two Record Extension Techniques

Bahaa Khalil, Taha B. M. J. Ouarda, and André St-Hilaire

***Abstract*—**The world economic crises and budget constraints have caused authorities, especially those in developing countries, to rationalize water quality monitoring activities. Rationalization consists of reducing the number of monitoring sites, the number of samples, and/or the number of water quality variables measured. The reduction in water quality variables is usually based on correlation. If two variables exhibit high correlation, it is an indication that some of the information produced may be redundant. Consequently, one variable can be discontinued, and the other continues to be measured. Later, the ordinary least squares (OLS) regression technique is employed to reconstitute information about discontinued variable by using the continuously measured one as an explanatory variable. In this paper, two record extension techniques are employed to reconstitute information about discontinued water quality variables, the OLS and the Line of Organic Correlation (LOC). An empirical experiment is conducted using water quality records from the Nile Delta water quality monitoring network in Egypt. The record extension techniques are compared for their ability to predict different statistical parameters of the discontinued variables. Results show that the OLS is better at estimating individual water quality records. However, results indicate an underestimation of the variance in the extended records. The LOC technique is superior in preserving characteristics of the entire distribution and avoids underestimation of the variance. It is concluded from this study that the OLS can be used for the substitution of missing values, while LOC is preferable for inferring statements about the probability distribution.

***Keywords*—**Record extension, record augmentation, monitoring networks, water quality indicators.

I. Introduction

THE quality of a water body is usually described by sets of physical, chemical, and biological variables, which are mutually interrelated. Water quality can be defined in terms of one variable to hundreds of compounds. Many researchers recognize that it is impossible to measure everything in the environment and that some logical means of selecting variables to measure must be part of every water quality information system [20]. Consideration should be given to reducing the number of variables sampled without substantial loss of information. Fewer variables make it easier to analyze and establish dependencies or correlations between various water quality variables, saving time and effort [19].

The literature reveals that correlation and regression analyses are commonly used to reduce the number of variables being measured. Correlation analysis is used to assess the level of association among the measured variables. If two variables show high correlation, it is an indication that some of the information produced may be redundant. Consequently, the measurement of one variable may be discontinued while maintaining the other. A regression technique is then used to reconstitute information about the discontinued variable using the continuously measured one as an explanatory variable.

Several previous studies have shown that the concentration of major ionic constituents can be related to specific conductance (e.g., [3, 15]). Specific conductance can serve as an indicator from which concentrations of major ionic solutes can be determined, as long as suitable regression functions can be found [18]. Yevjevich and Harmancioglu [23] investigated the transfer of information by bivariate correlations between daily water quality variables observed along the Upper Potomac River Estuary, USA. Their objective was to determine pairs of variables that are strongly correlated, in order to select variables that should be sampled and variables that can be estimated. Harmancioglu and Yevjevich [8, 9] studied the effects of removing deterministic components (trends, periodicity, and stochastic dependence) in order to understand the effects of these characteristics on the amount of information transfer. They concluded that basic similarities in deterministic components are the main contributors to the information transfer.

The use of regression analysis often results in an underestimation of the variance in extended records [1]. In addition, if the technique used for record extension introduces a bias into the value of more extreme order statistics, this will lead to bias in the estimates of the probability of exceedaance of selected extreme values or, conversely, bias in the

Bahaa Khalil is an Assistant Lecturer, Faculty of Engineering, Helwan University, Cairo, Egypt (corresponding author, phone: 1 450-667-0447; fax: 1 418-654-2600; e-mail: bahaa_kh2003@yahoo.com).

Taha B.M.J. Ouarda, Professor, Canada Research Chair on estimation of Hydrometerological variables, INRS-ETE, Quebec, Quebec, Canada (e-mail: taha.ouarda@ete.inrs.ca).

André St-Hilaire, Professor, INRS-ETE and Canadian River Institute, Quebec, Quebec, Canada (e-mail: andré.st-hilaire@ete.inrs.ca).

estimation of distribution percentiles [11]. In water quality, one may be interested not only in statistical moments but also in percentiles, which are used to assess compliance with standards or objectives. The line of organic correlation (LOC) was proposed as a linear fitting procedure in hydrology by Kritskiy and Menkel [12] and was applied to geomorphology by Doornkamp and King [4]. The line has also been called the "Maintenance of Variance Extension" or MOVE [11]. The LOC is widely applied to stream flow record extension at short-gauged stations (e.g., [11, 16, and 22]). The main advantage of the LOC is that the cumulative distribution function of the predictions, including the variance and probabilities of extreme events such as floods and droughts, estimates those of the actual records they are generated to represent [10].

The main goal of this study is to assess the usefulness of both the OLS and LOC techniques in reconstituting information about discontinued water quality variables. In the following section, methods and material are provided. The results obtained are presented and discussed in section 5. Finally, conclusions are presented in section 6.

II. Materials and Methods

This section consists of three subsections. In the first subsection, a theoretical background is provided. The second subsection gives a description of the Nile Delta water quality monitoring network. The third subsection details the empirical experiment.

A. Theoretical Background

Assume that the measured variable y has n_1 years of data, and the measured variable x has n_1+n_2 years, of which n_1 are concurrent with the data observed for y, illustrated as follows:

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

For water quality variables reduction, one can consider that the assessment and selection occurred in year n_1. After n_1 years, the measurement of variable y is discontinued, and the variable x continues to be measured. Assume that after n_2 years, it is desired that information be reconstituted about the variable y. To estimate records of the discontinued variable y for the period n_1+1 through n_2 years, simple linear regression of y on x can be used.

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

where $\hat{y}_i$ is the estimated value of y for $i = n_1+1,....n_2$, and a and b are the constant and slope of the regression equation, respectively. The parameters a and b are the values that minimize the squared error in the estimated y values. The solution of a and b is found by solving the normal equations [5]. The optimal solution to equation 1 is:

$$\hat{y}_i = \bar{y}_1 + r\,(s_{y1}/s_{x1})\,(x_i - \bar{x}_1) \quad (2)$$

Regression analysis often underestimates the variance in extended records [1]. Matalas and Jacobs [14] demonstrated that unbiased estimates of mean and variance are achieved if the following equation is used:

$$\hat{y}_i = \bar{y}_1 + r\,(s_{y1}/s_{x1})\,(x_i - \bar{x}_1) + \alpha\left(1-r^2\right)^{1/2} s_{y1}\, e_i \quad (3)$$

where α is a constant that depends on n_1 and n_2 (see Hirsch, [11]), r is the product-moment correlation coefficient between the n_1 concurrent measurements of x and y, and e_i is a normal independent random variable with a zero mean and unit variance. However, due to the presence of an independent noise component (e_i), the problem in using equation 3 is that studies of the same sequence of x and y by different investigators will almost surely lead to different values of $\hat{y}_i$ [1, 11].

An alternative to linear regression is to specify that the extension equation must be of the form given in equation 1, but that a and b are to be set not to minimize the squared error, but rather to maintain the sample mean and variance. The idea which leading to the development of LOC was to find the values of a and b in equation 1 that satisfy the following equations [11]:

$$\sum_{i=1}^{n_1} \hat{y}_i = \sum_{i=1}^{n_1} y_i \quad (4)$$

$$\sum_{i=1}^{n_1} (\hat{y}_i - \bar{y}_1)^2 = \sum_{i=1}^{n_1} (y_i - \bar{y}_1)^2 \quad (5)$$

One such solution is:

$$\hat{y}_i = \bar{y}_1 + (s_{y1}/s_{x1})\,(x_i - \bar{x}_1) \quad (6)$$

In spite of the obvious similarity between equations 2 and 6, it should be recognized that they have completely different origins [11]. If the OLS equation were used to estimate water quality variable, the variance of the resulting estimates would be smaller than it should be by a factor of r^2. OLS reduces the variance of estimates because the OLS slope is a function not only of the ratio of the standard deviations (s_y/s_x) but also of the magnitude of the correlation coefficient. r Only when absolute r =1 do OLS estimates possess the same variance as would be expected based on the ratio of variances for the original data.

When r = 0, there is no relationship between y and x. The slope then equals 0, and all OLS estimates would be identical and equal to $\bar{y}$. The variance of the estimates is also zero. As r^2 decreases from 1 to 0, the variance of OLS estimates is proportionately reduced [10]. This variance reduction is eliminated from the LOC by removing the correlation coefficient from the regression equation. The estimates

resulting from the LOC have a variance that is proportional to the ratio of the variances (s_y^2/s_x^2) from the original data.

Consider the situation in which a dependent variable *y* is to be estimated from values of an explanatory variable *x*. Slope and intercept estimates for the OLS equation are obtained by minimizing the sum of squares of residuals in units of *y*. Thus, its purpose is to minimize errors in the *y* direction only, without regard to errors in the *x* direction [10]. In contrast, situations occur in which it is just as likely that *x* should be estimated from *y*. It is easy to show that the two possible OLS lines (*y* on *x* and *x* on *y*) have different slopes. Assume that we are interested in estimating *x* from *y*; the regression model can be shown as follows:

$$\hat{x}_i = \bar{x}_1 + r\,(s_{x1}/s_{y1})\,(y_i - \bar{y}_1) \tag{7}$$

By solving this equation to estimate *y*:

$$\hat{y}_i = \bar{y}_1 + \frac{1}{r}\,(s_{y1}/s_{x1})\,(x_i - \bar{x}_1) \tag{8}$$

Thus, the slope is $(1/r * s_{y1}/s_{x1})$, which is not equal to the slope shown in equation 2. The two regression lines will therefore differ unless the correlation coefficient *r* equals 1. When only a single line describing the functional relationship between two variables is desired, the OLS line is not the appropriate approach. Neither OLS line uniquely or adequately describes that relationship. However, the LOC has a unique solution, which is one of its advantages. The LOC describes the functional relationship between two variables, and one can use it to estimate both sides.

B. Nile Delta Drainage WQM Network

The drainage system in the Nile Delta is composed of 22 catchment areas. Depending on their quality, effluents are either discharged into the Northern Lakes or pumped into irrigation canals at 21 sites along the main drains to augment freshwater supplies [6]. Numerous programs have been developed in the past to monitor the quality of the Nile water and agricultural drainage water in Egypt. In 1977, the National Water Research Center (NWRC) began to monitor a few volumetric and qualitative water parameters (predominantly concerning salinity) in several major Nile Delta drains. Since 1997, the NWRC has continuously expanded its monitoring activities to include an ever-increasing number of sampling sites and water quality variables. Thirty-three water quality variables are measured monthly at 94 sites in the Nile Delta drainage system (Fig. 1).

Fig. 1 The Nile Delta drainage system water quality monitoring network

Four water quality variables are selected for this study, the Specific Conductance (EC), Total Dissolved Solids (TDS), Sodium (Na), and Chloride (Cl). The selection of the four water quality variables is based on the evaluation and redesign of the National Water Quality Monitoring Network carried out in 2001 [17], where a high correlation was shown to exist between these variables. Data available for the selected variables at the 94 monitoring locations from August 1997 to July 2007 are used in this study.

C. Empirical Experiment

Three different models are considered to estimate TDS, Na, and Cl using EC as an explanatory variable. An empirical experiment is designed to examine the utility of the two extension techniques for preserving the statistical characteristics of the discontinued water quality variables. In order to evaluate the performance of the two record extension techniques, a cross-validation (jackknife) method is applied. In cross-validation, records of two years are removed from the available ten years of data. All possible successive or non-successive two years are considered. Thus, *C(10, 2)= 45* possible combinations are considered. The records for the removed two years are then estimated using the two record extension techniques calibrated with the remaining eight years.

The experimental design is as follows. For each of the three selected models at the 94 monitoring sites, the two record extension techniques are applied. Thus, 12,690 (94 locations x 3 models x 45 different sample combinations = 12,690) different realizations of extended water quality variable records are generated.

Water quality variables generally exhibit a positive skew [2, 7, 13], which is also confirmed by the preliminary analysis of the four selected variables. Consequently, the two record extension techniques are applied to the logarithms of the water quality variables.

Evaluation of the estimated records by the two extension techniques is comprised of three components. The first component involves testing the residual series for normality and serial correlation. An assumption associated with the record extension techniques is that the residual series is random and approximates a normal distribution. The normality of residuals is tested on the log-transformed

generated time series to determine whether the log-transformed parent series is normally distributed. The Shapiro-Wilk test for normality is applied to the residuals generated from each of the trials. Autocorrelation of residuals is an indicator of model inadequacy, which may arise when the model does not accounting for time-varying factors. In linear regression, autocorrelation of residuals increases the uncertainty associated with the estimated parameters; the mean square error (*MSE*) may underestimate the variance of the error terms and the confidence interval, so the tests on model parameters are no longer strictly applicable [21].

The second evaluation component involves the ability of the record extension techniques to estimate water quality concentration records with minimum errors. The Multiplicative Mean Error (*MME*) and the Normalized Mean Error (*NME*) are applied to individual estimated concentration records; the *MME* and *NME* are defined as:

$$MME = \exp\left[\frac{\sum_{i=n_1+1}^{n_2}\left|\ln \hat{y}_i - \ln y_i\right|}{n_2}\right] \tag{9}$$

$$NME = \frac{1}{n_2}\sum_{i=n_1+1}^{n_2}\frac{\hat{y}_i - y_i}{y_i} \tag{10}$$

where $\hat{y}_i$ and y_i are the estimated and measured values of the dependent variable for $i=n_1+1,\ldots n_2$, respectively. The *MME* of the logarithms is equal to the geometric mean of ($\hat{y}/y$), and thus, *MME* values equal to unity indicate an ideal performance of the record extension technique. The *MME* is applied to the logarithms of the extended records, while the *NME* is applied to the reverse transformed records.

The third evaluation component involves determining the ability of the extension techniques to reproduce different statistical parameters of the water quality concentrations during the extension period. The deviation of a statistical parameter (calculated from the estimated records) from the target value (calculated from the observed records), expressed as a fraction of the target value, is calculated for each of the 12,690 trials, and the fraction deviation is defined as follows:

$$f_{kj}(t) = \frac{v_{kj}(t) - tar_j(t)}{tar_j(t)} \tag{11}$$

where $f_{kj}(t)$ is the fractional deviation of the calculated statistic $v_{kj}(t)$ from the target value $tar_j(t)$ for trial (t), the statistical parameter is j, and the extension technique is k. For each record extension technique, the average fractional deviation is calculated over the 12,690 trials for each of the following statistical parameters:

- The cross-correlation between the estimated and observed concentrations;
  - The target value is 1, which would correspond to a perfect correlation.
- The lag-1 autocorrelation of the estimated concentrations;
  - The target value is the lag-1 autocorrelation of the observed concentrations.
- The mean value of the estimated concentrations;
  - The target value is the mean value of the observed concentrations.
- The variance of the estimated concentrations;
  - The target value is the variance of the observed concentrations.

The full range of non-exceedance percentiles from the 5th percentile to the 95th percentile is compared to target values during the extension period. A ratio *U* of each statistic for the extended records to that for the observed records is computed. The ratio *U* is used to assess the performance of the record extension techniques in preserving the historical characteristics. If the ratio *U* for a given statistical parameter is larger than 1, then the applied technique overestimates this statistic. If it is less than 1, the technique underestimates the target statistic. The ratio *U* is used to compare statistical parameters with the target equivalent. Additionally, the *MME* and *NME* are computed for different non-exceedance percentiles during the extended period.

III. Results

The results are divided according to the three evaluation components. The three subsections discuss the results for residual tests, error measures for individual water quality records, and the estimation of statistical parameters.

A. Residual tests

The first evaluation component examines the residual series for normality and autocorrelation. Table I summarizes the results obtained for the first evaluation component. Table I shows the average lag-1 autocorrelation (*SER*) obtained for all of the trials conducted for both extension techniques. The 95% confidence limits are shown in the columns entitled ±95%. The values given in Table I are average values obtained from the 12,690 trials. The standard deviations (*stdv*) of prediction are shown below the average values. The confidence limits depend on the level of significance of the test and the sample size. The level of significance is constant for all trials, but the sample size varies due to the existence of missing values in the data set. Table I also shows the percentage of trials in which the lag-1 autocorrelation is significant, as well as the percentage of trials in which the residuals fail the normality test.

Strictly speaking, the confidence limits shown have no specific meaning associated with the averaged test statistics due to the different sample sizes used in obtaining both test statistics and confidence limits. However, one may note that the standard deviation associated with the confidence limits is quite small, indicating that the confidence limits are not very dispersed. Thus, the average confidence limits, with the associated standard deviations, provide a general indication as to the range of confidence limits encountered in all trials.

Similarly, the average test statistic values, along with the associated standard deviations, provide a general indication as to the range of autocorrelation of residuals encountered over all trials.

TABLE I
RESULTS FOR THE RESIDUALS AUTOCORRELATION AND NORMALITY TESTS

Extension technique		-95%	*SER*	+95%	% *SER* Failed	Normality Failed %
OLS	Mean	-0.41	0.20	0.41	22	29
	stdv	0.01	0.27	0.01		
LOC	Mean	-0.41	0.17	0.41	19	26
	stdv	0.01	0.26	0.01		

Table I shows that the test statistics display a large variance, which indicates that the values of the test statistics obtained from the 12,690 trials are widely dispersed about the mean value. The standard deviation of the lag-1 autocorrelation is 135% of the mean value for OLS and 153% of the mean value for LOC. Although the mean value for the two techniques falls within the average 95% confidence limits, one would expect a large portion of the trials to "fail" the test due to the high standard error associated. Table I shows the fraction of trials for which each technique fails the autocorrelation test as well as the normality test. Autocorrelation of the residuals is found in 22% and 19% of the trials conducted for the OLS and LOC, respectively. For the normality test, 29% and 26% of the trials show that the residuals are not following the normal distribution for the OLS and LOC, respectively.

The assumption of a normal distribution is involved only when testing hypotheses, requiring the residuals from the regression equation to be normally distributed. The most important hypothesis test in regression is determining whether the slope coefficient is significantly different from zero. The slope in the regression model (b) is a linear function of the observations x_i and y_i, and a linear combination of normally distributed variables is itself normally distributed. Inferences regarding the variance of b are thus drawn from estimating the *MSE* from the sample, rather the true variance. The difficulty with variables that are not normally distributed is that the *MSE* may underestimate the variance of b. Consequently, the *MSE* is no longer an unbiased estimator of the variance of b and indeed may actually underestimate the uncertainty in b, resulting in more confidence being placed in the regression model parameters than is warranted.

Autocorrelation in residuals also increases the uncertainty in b. In the case of autocorrelated residuals, the constant and the slope in the regression model are still unbiased, but no longer have a minimum variance. If serial correlation of the residuals occurs, the estimates of the regression coefficients are no longer the most efficient estimates possible, although they remain unbiased. This means that the confidence intervals are too narrow. Thus, the OLS may indicate a greater precision in the regression coefficients than is actually the case. For the second and third evaluation components, evaluation is therefore carried out for the total number of trials (12,690) and for only the number of trials that pass the residual tests. Only 6,790 trials passed both residual tests for both record extension techniques.

B. Individual Values

The *MME* and *NME* are computed for individual concentration records, where each error represents the difference between the estimated and actual water quality record. The average *MME* values of the estimated water quality concentrations over all of the trials are 1.17 and 1.19 for the OLS and LOC techniques, respectively (Table II). The average *NME* values are 4.50 and 5.25 for OLS and LOC, respectively. The standard deviation is greater with LOC than with OLS, which indicates not only that the average error with LOC is greater than that with OLS, but also that the error obtained with LOC is more widely dispersed around the mean value. An application of the t-test average *MME* as well as the *NME* results shows that there is a significant difference between the two techniques in estimating discontinued water quality records.

TABLE II
AVERAGE ERROR MEASURES FOR RECORD EXTENSION TECHNIQUES

Error Measure (*number of trials*)		Extension Technique		t-test	
		OLS	**LOC**	**t**	**p-value**
MME (12,690)	mean	1.17	1.19	-14.9	0
	stdv	0.11	0.14		
MME (6,790)	mean	1.15	1.17	-10.2	0
	stdv	0.08	0.11		
NME (12,690)	mean	4.50	5.25	-2.6	0.01
	stdv	22.20	24.22		
NME (6,790)	mean	0.28	0.72	-2.5	0.01
	stdv	9.86	11.06		

When using only trials that meet the record extension technique assumptions, the *MME* and *NME* still show a significant difference, which indicates that OLS shows better performance than LOC in the estimation of individual water quality records. Results also show a significant reduction in the *NME* values when using only trials that meet the record extension technique assumptions. Overall, better results may be obtained if the record extension technique assumptions are met.

C. Statistical Parameters

Table III shows the average fractional deviation illustrated by the two record extension techniques for the correlation, lag-1 autocorrelation, mean, and variance values. Table III shows that, the two extension techniques have no difference in the fractional deviation from the target correlation coefficient. However, using only trials that meet the statistical technique assumptions (6,790 trials), both the average fractional

deviation and the standard deviation are reduced. This indicates that these trials produce better results. This is not the case for the lag-1 autocorrelation, where a significant increase of the average deviation is detected. However, a significant reduction in the standard deviation is associated with an increasing average deviation. The standard deviation is about 150% of the average deviation when using all of the trials (12,690), while it is 75% when using the selected trials (6,790). The mean and variance estimations of the observed records did not exhibit this significant reduction. The deviation from the variance results indicates that the OLS is negatively deviated, or in other words, the variance of the observed values is underestimated, while the LOC overestimates the variance. Both techniques are equivalent in estimating the mean values.

TABLE III
AVERAGE FRACTIONAL DEVIATION FOR DIFFERENT STATISTICS

Statistical parameter	technique	No. of trials	Average deviation	Standard deviation
Cross-Correlation	OLS	12,690	-0.23	0.22
		6,790	-0.20	0.20
	LOC	12,690	-0.23	0.22
		6,790	-0.20	0.18
Lag-1 autocorrelation	OLS	12,690	0.23	32.61
		6,790	0.41	28.25
	LOC	12,690	0.21	32.36
		6,790	0.39	28.01
Mean	OLS	12,690	-0.01	0.10
		6,790	-0.03	0.09
	LOC	12,690	0.01	0.12
		6,790	-0.00	0.10
Variance	OLS	12,690	-0.30	0.49
		6,790	-0.34	0.47
	LOC	12,690	0.40	1.82
		6,790	0.35	1.74

As a summary of Table III, there is no significant difference between the two record extension techniques in the cross-correlation, lag-1 autocorrelation, or the mean value estimation. The OLS underestimates the variance, while the LOC overestimates it. The LOC average deviation for the estimation of the variance is associated with a standard deviation about 4.5 times the average value, while the standard deviation for the OLS is about 1.5 times the average value.

Figs. 2 and 3 summarize the results obtained for the ratio *U*. Fig. 2 shows the *U* ratio distribution for the estimation of the statistical moments, and Fig. 3 shows the *U* ratio distribution for the estimation of the full range of non-exceedance percentiles. The box plots in Figs. 2a and 3a are plotted using the 6,790 selected trials, while Figs. 2b and 3b are plotted using the total 12,690 trials. In these figures, the box plots represent the distribution of *U* for a given statistical parameter and record extension technique. The accuracy of each approach can be judged by the degree of dispersion in the box plots, by the degree that the median approaches the value of 1, and by the symmetry of the box plot about the value of 1.

Fig. 2 shows box plots that represent the distribution of *U* for the estimation of the mean and standard deviation. For the mean estimation, the OLS and LOC techniques give values of *U* with median values of 0.98 and 0.99, respectively. This is true for both groups of trials. Box plots for both extension techniques are symmetric around 1 and have similar dispersions. For the estimation of the standard deviation, the box plots show that the OLS technique tends to underestimate the standard deviation. Figs. 2a and 2b show that more than 75% of the computed standard deviations are less than the historical values, with median values of 0.77 and 0.79 for 6,790 and 12,690 trials, respectively. For the LOC technique, the *U* median values are 1.03 and 1.02 for the selected and total trials, respectively. The box plots of LOC are more symmetric around 1 than those of the OLS technique. The box plots for the OLS and LOC techniques give similar dispersions.

For the OLS estimation of non-exceedance percentiles, Fig. 3 indicates that the median values of *U* for low percentiles are greater than 1 and are less than 1 for high percentiles. The LOC median values of *U* for the percentile range are between 0.97 and 1.00 using 6,790 trials (Fig. 3a) and between 0.99 and 1.01 using 12,690 trials (Fig. 3b).

The OLS box plots show that the median values range between 0.90 and 1.06 using the 6,790 trials and between 0.92 and 1.07 using the 12,690 trials. In general, the LOC box plots are symmetrical about 1 and show relatively less dispersion than that corresponding to the OLS technique. The OLS box plots show that, for low percentiles, almost 75% of the records are greater than 1, and almost 75% of the records are less than 1 for high percentiles. These results suggest that the OLS technique tends to overestimate low percentiles and underestimate high percentiles. The LOC technique reduces the bias toward overestimation of low concentrations and underestimation of high concentrations exhibited by OLS.

Fig. 4 illustrates the *MME* exhibited by the two extension techniques in estimating the non-exceedance percentiles. Figs. 4a and 4b show that the *MME* values for OLS are greater than 1 for the estimation of low percentiles and less than 1 for the estimation of high percentiles. The LOC *MME* values are closer to 1 than for the OLS, except at the median, where both values are equal.

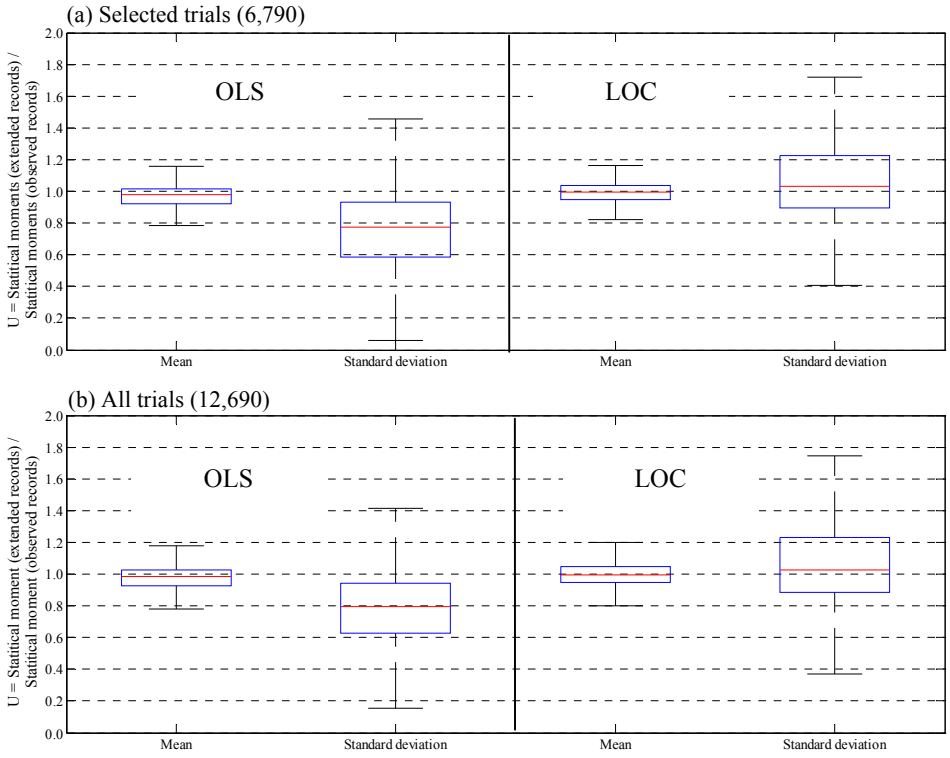

Fig. 2 Box plots of the *U* ratio for the statistical moments

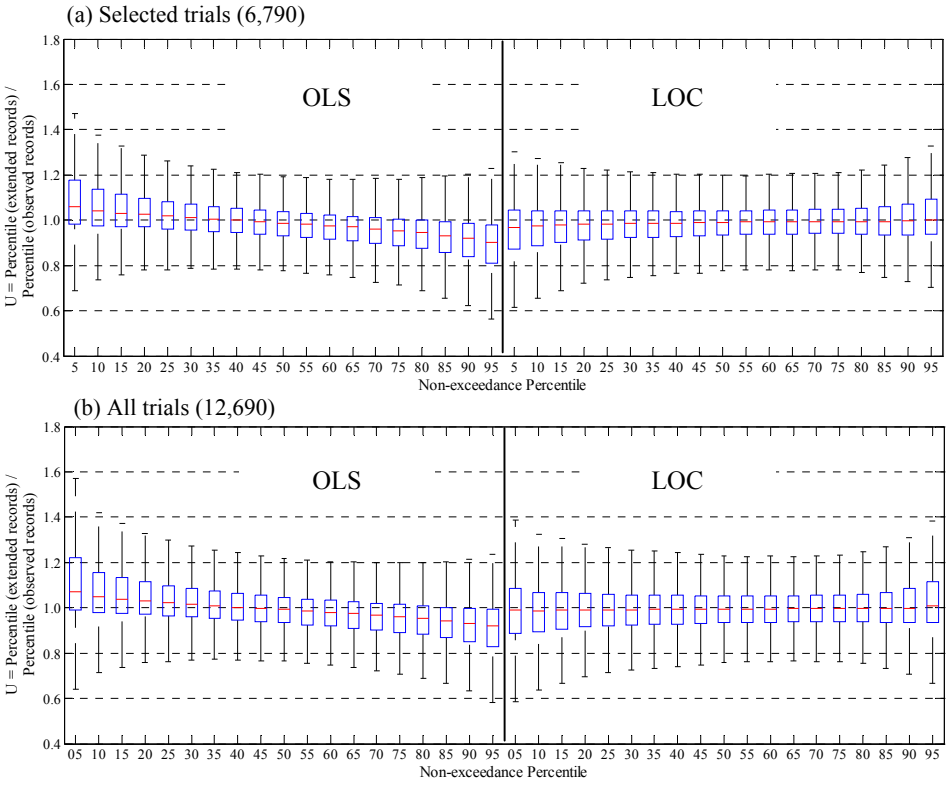

Fig. 3 Box plots of the ratio *U* for different percentiles

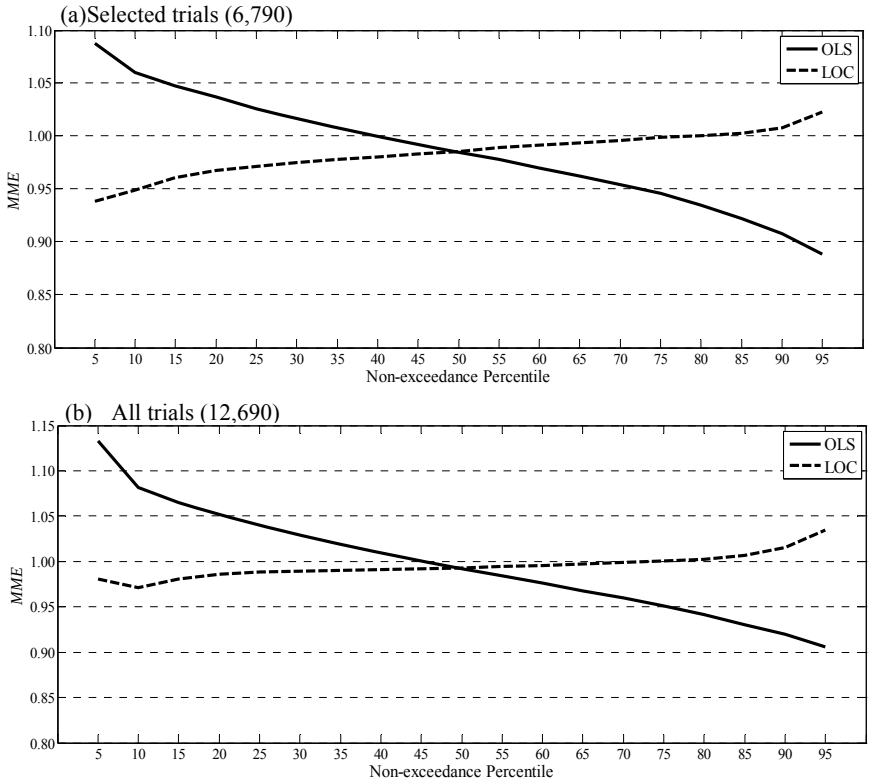

Fig. 4 *MME* of the tested extension techniques for the estimation of various percentiles

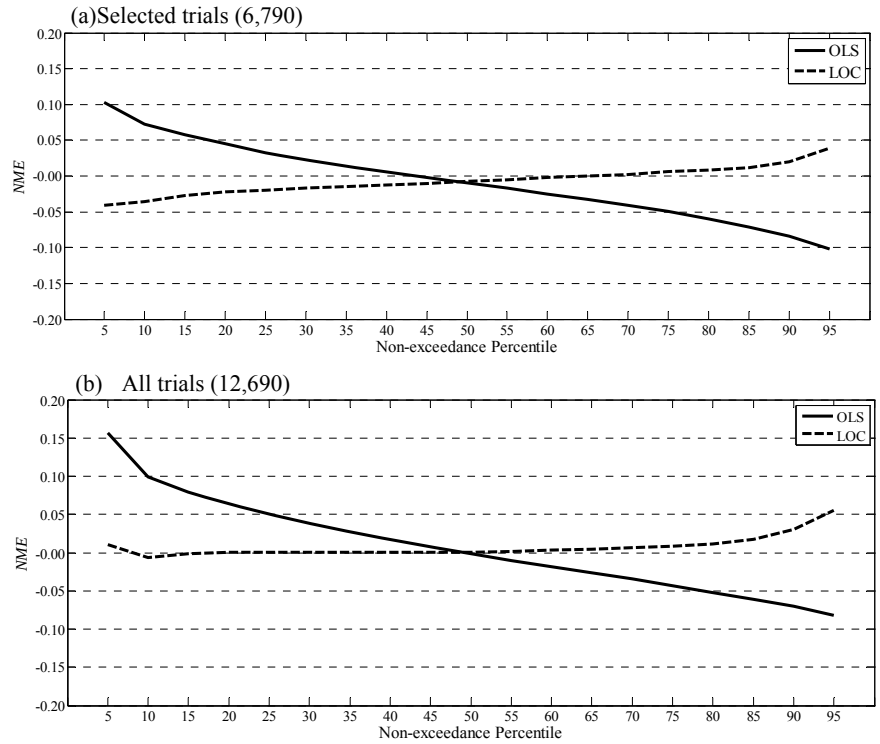

Fig. 5 *NME* of the tested extension techniques for the estimation of various percentiles

Fig. 5 illustrates the *NME* exhibited by the two extension techniques in estimating the non-exceedance percentiles. The *NME* shows behavior similar to that of the *MME*, since the LOC *NME* values are closer to zero than those for the OLS, except at the median, where both values are nearly identical. Thus, error measures performed on the logarithms of the extended records or on the reverse transformed values show equivalent results. Figs. 3, 4, and 5 clearly depict regression overestimation of low concentrations and underestimation of high concentrations, as would be expected from the tendency to produce an extended record with low variance.

IV. Conclusion

Ordinary least squares (OLS) regression and LOC techniques are applied to reconstitute information about discontinued variables using data from the Nile Delta water quality monitoring network. Different statistical performance measures are used to assess each of the extension techniques and their ability to maintain statistical characteristics of the water quality records. Verification of the model assumptions ensures better estimation of the individual records while also preserving the statistical characteristics.

The two techniques produce extended records that have unbiased mean or median values. The OLS substantially reduces variability, and the LOC preserves variability. The OLS technique underestimates high concentration values and overestimates low values. On the other hand, the LOC technique tends to reduce the bias in the estimation of both high and low concentration values. The LOC technique produces extended records that relatively preserve both high and low percentiles.

The LOC was better in preserving the statistical characteristics of the discontinued water quality variables. However, the OLS was superior in the estimation of individual records. If only individual water quality records are of interest, then the OLS technique is preferable, since it gave better results than the LOC. Therefore, it is recommended that the OLS be used for the substitution of missing values, while the LOC is preferable for gaining insight into the probability distribution.

Acknowledgment

The authors wish to thank Prof. Shaden Abdel-Gawad, president of the National Water Research Center of Egypt, for providing the data used in this study. The financial support provided by Helwan University, Cairo, Egypt, and the Natural Sciences and Engineering Research Council of Canada (NSERC) is acknowledged.

References

[1] Alley, W.M. and Burns, A.W., "Mixed-station extension of monthly streamflow records", Journal of Hydraulic Engineering, 109 (10), 1983, pp. 1272 - 1284.

[2] Berryman, D., Bobée, B., Cluis D. and Haemmerli, J., "Nonparametric Tests for Trend Detection in Water Quality Time Series", Water Resources Bulletin, 24(3), 1988, pp. 545 - 556.

[3] Briggs, J.C. and Ficke, J.F., "Quality of rivers of the United States, (1975) water year- based on the National Stream Quality Accounting Network", U.S. Geological Survey Open-File Report 78-200, 1978, p. 436.

[4] Doornkamp, J.C. and King C.A.M., "Numerical Analysis in Geomorphology, An Introduction", St. Martins Press, New York, NY, 1971, p. 372.

[5] Draper, N.R. and Smith, H., "Applied regression analysis", John Wiley, New York, 1966, p. 736.

[6] DRI (Drainage Research Institute) - MADWQ, "Monitoring and analysis of drainage water quality in Egypt", Interim Report, DRI, Cairo, 1988.

[7] 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, 1999, p. 290.

[8] 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, D.C., 1986, p. 81.

[9] Harmancioglu, N.B. and Yevjevich, V., "Transfer of hydrologic information among river points", Journal of Hydrology, 91, 1987, pp. 103 - 118.

[10] Helsel, D.R., and Hirsch, R.M., "Statistical methods in water resources", U.S. Geological Survey, Chapter A3 in Hydrologic Analysis and Interpretation, 2002, p. 522.

[11] Hirsch, R.M., "A comparison of four streamflow record extension techniques", Water Resources Research, 18(4), 1982, pp. 1081 - 1088.

[12] Kritskiy, S.N. and Menkel, J.F., "Some statistical methods in the analysis of hydrologic data", Soviet Hydrology Selected Papers 1, 1968, pp. 80-98.

[13] Lettenmaier, D.P., "Multivariate nonparametric tests for trend in water quality", AWRA, Water Resources Bulletin, (24)3, 1988, pp. 505 - 512.

[14] Matalas, N.C. and Jacobs, B., "A correlation procedure for augmenting hydrologic data", U.S. Geological Survey Prof. Pap., 434-E, 1964.

[15] McKenzie, S.W., "Long-term water quality trends in Delaware streams", U.S. Geological Survey open-files report 76-71, 1976, p. 85.

[16] Moog, D.B. and Whiting P.J., "Streamflow record extension using power transformations and application to sediment transport", Water Resources Research, 35 (1), 1999, pp. 243 - 254.

[17] NAWQAM, National Water Quality and Availability Management Project, "Evaluation and Design of Egypt National Water Quality Monitoring Network", Report No.: WQ-TE-0110-005-DR, NAWQAM, National Water Research Center, Cairo, Egypt, 2001.

[18] Sanders, T.G., Ward, R.C., Loftis, J.C., Steele, T.D., Adrian, D.D. and Yevjevich, V., "Design of Networks for Monitoring Water Quality", Water Resources Publications, Littleton, Colorado, 1983, p. 328.

[19] Strobl, R.O. and Robillard, P.D., "Network design for water quality monitoring of surface freshwaters: A review", Journal of Environmental Management, 87, 2008, pp. 639 - 648.

[20] Ward, R.C., Loftis, J.O. and McBride, G.B., "Design of Water Quality Monitoring systems", Van Nostrand Reinhold, New York, USA, 1990, p. 231.

[21] Wasserman, W., Kutner, M.H. and Neter, J., "Applied linear regression models" (2nd ed.). Richard D. Irwin Inc., Boston, 1989.

[22] Vogel, R.M. and Stedinger, J.R., "Minimum variance streamflow record augmentation procedures", Water Resources Research, 21(5), 1985, pp. 715 - 723.G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in Plastics, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

[23] Yevjevich, V. and Harmancioglu, N.B., "Modeling Water Quality Variables of Potomac River at the Entrance to its Estuary, Phase II (Correlation of Water Quality Variables within the Framework of Structural Analysis)" Report to D.C. Water Resources Research Center of the University of the District of Columbia, Washington, D.C., 1985, p. 59.