Response of authors to Dr. F. Serinaldi's comments

The authors are grateful to Dr. Serinaldi's for his valuable comments, which contributed to improve the quality of the paper.

Specific Comments

Dr. F. Serinaldi comment:

The Monte Carlo experiments reveal that the proposed RLOC provides estimates of the extreme quantiles less accurate than LOC for small datasets even though the data follow a bivariate Gaussian distribution. In my opinion, this shortcoming is due to the non-optimal use of the information contained in the data. In more detail, while the LOC slope is the ratio of standard deviations computed by the standard estimator based on n1 and n2 values of X and Y , respectively, the RLOC relies on only four values, that is, the first and third quartiles of X and Y . Therefore, even though RLOC is robust against outliers, it is also unavoidably more imprecise under uncertain estimates of the quartiles in small datasets. These remarks and a closer look at the KTRL estimator can suggest a straightforward modification of the RLOC that can be explored and can possibly improve the precision preserving the robust nature of the RLOC. From Eq. 10, it is rather clear that a similar relationship exists not only between the interquartile range and variance but also between every interquantile range and the variance for symmetric distributions. Therefore, instead of computing the RLOC slope by using only the interquartile range (four values), the median of the slopes resulting from a set of suitable interquantile ranges (say, $(y_{c(90)} - y_{c(10)}) / (x_{c(90)} - x_{c(10)})$, $(y_{c(80)} - y_{c(20)}) / (x_{c(80)} - x_{c(20)})$, $(y_{c(70)} - y_{c(30)}) / (x_{c(70)} - x_{c(30)})$ *) can be used. In this way, the information stored in the data can be better used by applying the same rationale of the KTRL method, and simultaneously the estimation procedure is kept robust by discarding the outliers from the computations. Of course, the effectiveness of this approach should be checked, but in principle, it seems to be a viable trade-off between unbiasedness and accuracy.*

Authors' response: The authors strongly agree with this comment. The idea is that the OLS technique has two main deficiencies: under estimation of the extended records variance, and that

it is not robust to the presence of outliers. On the one hand, the LOC (MOVE1) technique overcomes the first deficiency but is still not robust for the presence of outliers. On the other hand, the KTRL technique overcomes the second deficiency but underestimates the variance of the extended records. Thus, one may either modify the LOC technique to be robust to the presence of outliers (which is the subject of this paper), or modify the KTRL to be able to maintain the variance in the extended records. A modified KTRL that maintains the variance in the extended records was proposed by the first author and others (under revision in the journal of Water Resources Management "WARM2840"), where the modified KTRL slope (b_q) is based on the data percentiles rather than the observed values as in the case of KTRL, as follows:

$$b_q = median\ \frac{q(y)_j - q(y)_i}{q(x)_j - q(x)_i}$$

$$\forall\ i < j \qquad i = 5^{th}, 10^{th}, ...90^{th} \qquad j = 10^{th}, 15^{th}.....,95^{th}$$

where $q(y)$ and $q(x)$ are the percentiles of y and x estimated during the period of concurrent records. Percentiles are obtained for the range of the 5^{th}, 10^{th},… to the 95^{th} percentile. Thus, a set of 19 (x, y) pairs of percentiles will result in 171 ($n\ (n\text{-}1)/2$ = 19 (19-1)/2) pair-wise comparisons. For each of these comparisons, a slope $\Delta y / \Delta x$ is computed and the median of the 171 possible pair-wise slopes is taken as the slope estimate. Besides being robust to the presence of outliers, the modified KTRL is able to maintain the variance of the extended records. However, similar to the RLOC, the modified KTRL is also not precise in the case of small datasets compared to LOC (or MOVE1).

In addition, in an unpublished work (to be submitted soon), a comparison was carried out between the modified KTRL and the RLOC, as well as the four MOVE techniques. In this unpublished work, two Monte Carlo studies were carried out. The first Monte Carlo study consisted of generating x and y records from a bivariate normal distribution $N(0,1)$ under three different levels of association (ρ = 0.5, 0.7 and 0.9) crossed with five different sizes of the concurrent records (n_1 = 48, 72, 96, 120, and 144), while the size of the extended records (n_2) was 24 records. For each of the 15 (3 ρ × 5 n_1) combinations of the level of association and

sample sizes, 5000 simulations were generated. In the second Monte Carlo study, the same study was conducted after introducing just one extreme value to the *y*-series to assess the impact of the presence of an outlier value on the performance of the MOVE techniques as well as the two new techniques (modified KTRL and RLOC). The results of the first Monte Carlo study (without outlier) showed that the modified KTRL and RLOC were outperformed by MOVE techniques, while the results of the second Monte Carlo study (with outlier) showed that the modified KTRL and RLOC outperform the MOVE techniques. For the comparison between the modified KTRL and RLOC, based on the results of both studies, the modified KTRL outperforms the RLOC (sample of the results for the estimation of the standard deviation of the extended records is presented in the following table).

RMSE values for the estimation of the standard deviation (Monte Carlo experiments)

ρ	n_1	Without outlier		With outlier	
		KTRL2	RLOC	KTRL2	RLOC
0.9	144	0.102	0.127	0.104	0.131
	120	0.103	0.135	0.106	0.137
	96	0.106	0.143	0.111	0.146
	72	0.112	0.157	0.118	0.160
	48	0.122	0.184	0.135	0.190
0.7	144	0.166	0.189	0.167	0.192
	120	0.169	0.196	0.171	0.200
	96	0.174	0.207	0.178	0.212
	72	0.182	0.227	0.187	0.230
	48	0.196	0.257	0.208	0.268
0.5	144	0.200	0.221	0.208	0.231
	120	0.203	0.228	0.211	0.238
	96	0.209	0.238	0.218	0.249
	72	0.219	0.255	0.230	0.270
	48	0.235	0.287	0.254	0.312

Thus, as Dr. Serinaldi expected, using the median of a set of slopes allows for better use of the information stored in the data. The advantage of the modified KTRL slope estimate (using the median of a set of slopes) is that it is more precise than the RLOC estimate (using the interquartile range ratio). The advantage of the RLOC slope estimate is that it is simple to

compute and implement. The authors see that the RLOC technique as presented in this paper is a modified version of the LOC (or MOVE1), with the following advantages: it is robust to the presence of outliers; it maintains the variance of the extended records; and it is easy to compute and implement. Although the authors appreciate Dr. Serinaldi's comment, we will not really address it in this paper as it is addressed in another paper that has been submitted for publication (as explained in our response).

Dr. F. Serinaldi comment:

Referring to the above comments, I think that the Authors should distinguish the hypotheses of Gaussianity and symmetry throughout the paper. While the parametric methods require that the data follows a biviariate Gaussian distribution, the symmetry of the marginal distribution is enough for the non-parametric techniques, such as RLOC (otherwise, these methods should be referred to as "parametric"). A clear distinction is also fundamental to define an appropriate transformation of the original marginal distribution of the data. Namely, while we need that a logarithmic transformation (or whatever else) returns bivariate Gaussian data in order to apply the parametric regressions, on the other hand, we only need that the transformations simply adjust the symmetry of the marginal distributions when nonparametric RLOC is applied. Perhaps, the Authors can be interested to the work by Serinaldi et al. (2012), for a discussion on the effects of different marginal transformations and the role of the structure of dependence on the regression outcomes for skewed data. That work also points out the good performance (in terms of point and interval estimates) of a very simple weighted regression, which easily accounts for the heteroskedasticity of the errors of skewed data without applying preliminary marginal transformations. In this context, it is also worth mentioning that more refined techniques based on e.g. Generalized Linear Models (GLMs), Generalized Additive Models (GAMs) and their extensions are readily available and widely used in the industry and research. These techniques are not as simple as the closed form formulas provided in the paper under review, but I think that they must be taken into account when one requires a more refined augmentation of data that exhibit complex temporal and cross correlation patterns.

Authors' response: The authors agree with this comment, as the RLOC may require only symmetric distribution, and not essential a Gaussian distribution. This point is now emphasized in the revised version of the manuscript as follows:

"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."

For the point and confidence interval estimates, the confidence interval is based on the z statistic (the standard normal quantile at a specific probability or confidence level), and the variance of the errors. Thus, the normality and homoscedasticity of the errors are essential assumptions to estimate the confidence interval. As for the use of the weighted least-squares regression (WLS), it minimizes the weighted residuals $w_i(y_i - \hat{y}_i)^2$ rather than the residuals $(y_i - \hat{y}_i)^2$ as in the case of the OLS (unweighted LS). The WLS has the advantage that it accounts for heteroscedasticity of the residuals; however, similar to the OLS, it has two deficiencies: it is not robust to the presence of outliers; and it underestimates the variance of the extended records. In the presence of outliers Iteratively Weighted Least-Squares (IWLS) should be used instead of the WLS, where the residual weights are based on the IQR range, and not the standard deviation as in the case of the WLS. Similar to the OLS, neither WLS nor IWLS maintains the variance of the extended records, which is the advantage of LOC or RLOC, and which is the objective of this paper.

The regression assumptions are presented in the revised version, where the assumptions of a residuals homoscedasticity, normality and independency are highlighted as follows:

"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 assessment of uncertainty or confidence intervals is of concern, statistical hypothesis should be introduced (Serinaldi, et al., 2102). In this case, the last three assumptions must be fulfilled."

In the revised version, the set of recommendations presented was expanded to include a comparison with the more advanced techniques such as the GLM and GAM as follows:

"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. 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.

Dr. F. Serinaldi comment:

A final remark concerns the leave-one year-out cross validation used in the real world data analysis. As mentioned at P4669L12-15, water quality data show special characteristics such as seasonal patterns and autocorrelation. However, every regression technique relies on the basic hypothesis that data are time independent. Thus, I wonder if the Author can better clarify if the monthly data used in the case study show an evident seasonality (as it could be expected) and how they accounted for that. In my understanding, simple regression techniques, such as MOVE, work well for data augmentation when the signal-to-noise ratio between the seasonal pattern and the local fluctuations is high (as for monthly streamflow data). In this context, small and high values of the variables often correspond to particular seasons or months (summer, winter, etc.). In other words, an approximate linear relationship in the scatter plots results from clusters of data corresponding to particular months or seasons. In more detail, it can be worth specifying how the paper fits e.g. into the framework of cyclic and noncyclic procedures suggested by Alley and Burns (1983) and mentioned by Hirsh (1982). A figure showing some time series as well as the corresponding scatter plots of X versus Y can help the visual understanding of the data on hand and their statistical properties along with the regression outcomes.

Authors' response: The authors understand Dr. Serinaldi's concern about the nature of the data used in the empirical experiment. We would like to clarify that there are five assumptions

associated with linear regression. The obligation of satisfying them is decided based on the purpose of the regression equation. Reference is made to Helsel and Hirsch (2002), Subsection 9.1.1 Assumptions of Linear Regression, pages 225 (Table 9.1).

To predict y given x: only two assumptions are required, (1) y should be linearly related to x, and (2) the data used to fit the model are representative of the data of interest.

To predict y and variance of prediction: (3) variance of the residuals is homoscedastic beside the previous two assumptions.

To obtain best unbiased estimator of y: (4) the residuals should be independent beside the previous three assumptions.

To test hypothesis, estimate confidence or prediction intervals (which was one of the objectives of Serinaldi et al. (2012)): (5) the residuals should be normally distributed beside the previous four assumptions.

However, we would like to confirm that the preliminary analysis for the Edko drain data confirms the following:

- EC and Cl are linearly dependent;
- EC and Cl data are positively skewed;
- Presence of outliers;Data is serially independent;
- Absence of significant cycle or seasonal pattern;
- Data is homoscedastic.

The Nile Delta may be considered as a semi-arid region, where the annual total precipitation is about 100 mm (El-Saadi, 2006). The area is intensively cultivated (two or three cultivation seasons per year). Irrigation systems rely on water from the Nile River, while the agricultural drainage system receives the excess irrigation water from the agricultural lands. In addition, it also receives industrial and municipal effluents, because about only 40% of the villages within the Delta have accesses to sanitary services. The variation of the drainage system water quality is due more to the variation of the cultivated crop types rather than the normal seasonal variation, with expected irregular values due to industrial/municipal effluents and most of the time untreated effluents.

In the revised version of the manuscript these points are emphasized and an example of the preliminary analysis carried out at one of the monitoring sites is now presented (below are two figures and results of the Kolomogrov-Smirnov test for normality).

"For the Edko drain data, preliminary analysis confirmed the linear dependency between EC and Cl, and that the data is 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, Figure 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 indicates 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 (Figure 4), these autocorrelation values are not significant."

In the cases of seasonality, recommendations will be given in the revised version based on the work of Alley and Burns (1983):

"In the case of seasonality, the record-extension techniques can be applied for the data of each season or month as recommended by Alley and Burns (1983)."

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

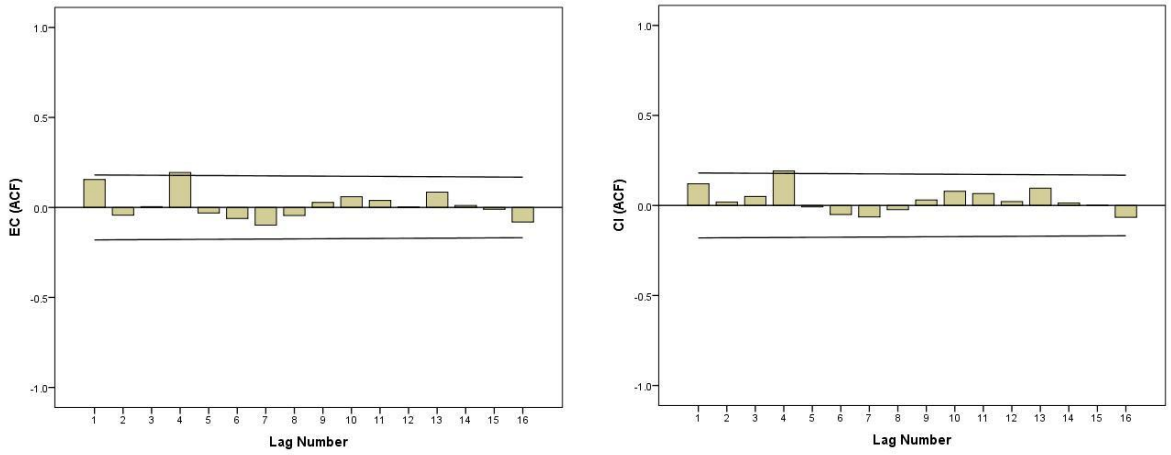

Figure 4. Correlograms for Electric Conductivity (EC) and Chloride (Cl) measured at WE11.

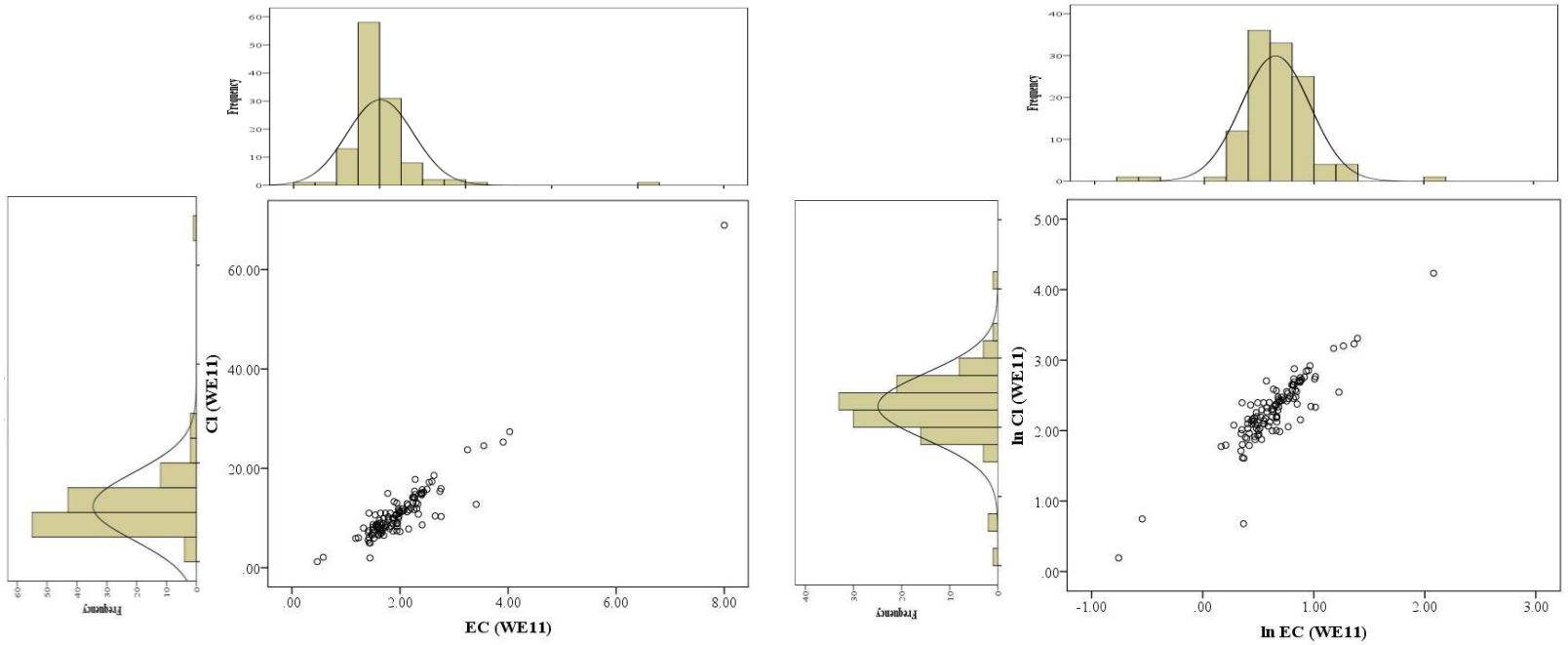

Figure 3. Scatter plots and probability density plots for Electric Conductivity (EC) and Chloride (Cl) measured at WE11.

Editing notes provided by Dr. F. Serinaldi:

In a previous work, the first Author mentions the Sen's slope. Actually, in my understanding the Sen's regression is a generalization of the Theil's method when there are ties in the covariate. Perhaps, it is fair to cite Sen (1968).

Authors' response: In the revised version of the manuscript, Sen (1968) is now cited as follows: "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." (Page 4673, lines 7-8).

I suggest avoiding the terms "independent" and "dependent" variables. Even though they are commonly used, perhaps, "response variable" and "explanatory variable" or "covariate" are more appropriate as X are rarely statistically independent.

Authors' response: In the revised version of the manuscript "independent" and "dependent" variables have been replaced by "explanatory" and "response variable".

Page 4672, lines 5-6: "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)."

Page 4677, lines 10-12: "Each mixture distribution was treated as the response variable in a regression, while the predictor was a generated random order variable."

In the rest of the manuscript "response" and "predictor" were used.

References:

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.

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