Environ Monit Assess (2021) 193:345
https://doi.org/10.1007/s10661-021-09097-9

# The influence of the correlation-covariance structure of measurement errors over uncertainties propagation in online monitoring: application to environmental indicators in SUDS

F. Peña-Heredia · S. Sandoval ·
J. A. Escobar-Vargas · A. Torres

Received: 31 August 2020 / Accepted: 19 April 2021


**Abstract** This paper presents a methodology to assess the influence of the correlation-covariance structure of measurement errors in online monitoring over the propagation of uncertainties, applied to wet-weather environmental indicators in sustainable urban drainage systems (SUDSs). The effect of auto-correlated and heteroskedastic errors in measured time-series over the estimated probability density function (PDF) of different environmental indicators is analyzed for a wide variety of possible error structures in the data. For this purpose, multiple correlation-covariance structures are randomly generated from exploring the parametric space of a linear exponent autoregressive (LEAR) model, employing a Bayesian-based Markov Chain Monte Carlo sampling technique. Significant differences tests are proposed to identify the most correlated parameters of the correlation-covariance error model with statistics of the environmental indicator PDFs. The method is applied to total suspended solids (TSS) and chemical oxygen demand (COD) time-series recorded during 13 rainfall events at the inlet and outlet of a SUDS train (stormwater settling tank—horizontal constructed wetland). In this case, results showed that the total error in the estimation of the analyzed environmental indicators is mostly explained by standard uncertainties (flattening of the PDFs) rather than bias contributions (displacement of the PDFs). The correlation-covariance model parameters related to the temporal delimitation of hydrographs/pollutographs and the intensity of the autocorrelation showed to have the strongest influence in the propagation of measurement errors (flattening/displacement of the PDFs).

**Keywords** Auto-correlated errors · Autocovariance matrix · Bayesian-MCMC methods · Noisy time-series · Positive definite matrix · Water quality metrology

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1007/s10661-021-09097-9.

F. Peña-Heredia (✉) · S. Sandoval · J. A. Escobar-Vargas · A. Torres
Department of Civil Engineering, School of Engineering, Pontificia Universidad Javeriana, Bogotá, Colombia
e-mail: felipe.pena@javeriana.edu.co

F. Peña-Heredia · S. Sandoval · J. A. Escobar-Vargas · A. Torres
Ciencia e Ingeniería del Agua y el Ambiente Research Group, Pontificia Universidad Javeriana, Bogotá, Colombia

## Introduction

Sustainable Urban Drainage Systems (SUDSs) (e.g., green roofs, rain gardens, infiltration trenches, swales, constructed wetlands, retention/infiltration basins) have emerged as an alternative to conventional sewer-based systems (Fryd et al., 2012). SUDSs are used to imitate the natural water balance, allowing the control of both flood and water quality impacts of stormwater runoff during and after rainfall events (Bockhorn et al., 2017; Hoang & Fenner, 2016). The operation and decision-making in SUDS is commonly nourished by estimating environmental indicators such as the runoff volume (RVol) (e.g., Andrés-Doménech et al., 2018), event mean concentration (EMC) (e.g., Métadier & Bertrand-Krajewski, 2012), hydraulic efficiency (HE) (e.g., Wang et al., 2017), or pollutant removal efficiency (RE) (e.g. Fardel et al., 2019). These indicators are evaluated principally by mathematical modeling (e.g., Andrés-Doménech et al., 2018) or water quality sampling (e.g., Grenni et al., 2019).

However, mathematical modeling approaches have limitations in representing the water quality dynamics properly, due to uncertainties related with the specific models used and the parameters they contain (Jia et al., 2018; Peng & Stovin, 2017). In addition, while sampling methods are of common practice (e.g., Grenni et al., 2019), the estimate of environmental indicators in SUDS based on this approach presents important limitations (Andrés-Valeri et al., 2014). These limitations principally arise from the low capacity of sampling strategies to represent the high temporal variability of stormwater quality during rainfall events (Sandoval et al., 2018). Online monitoring has therefore emerged during the past two decades as a well-recognized technique for water quality and quantity measurement (e.g., Métadier & Bertrand-Krajewski, 2012; Rizzo et al., 2020). However, the operation and decision-making in SUDS throughout monitoring have been addressed by a lesser number of authors (e.g., Hatt et al., 2009; Métadier & Bertrand-Krajewski, 2012; Leutnant et al., 2016).

On the other hand, the scientific community has recognized the requisite of uncertainty assessments related to this measurement principle and how these are propagated, according to the objective of the specific study (Bertrand-Krajewski & Bardin, 2002; Désenfant & Priel, 2017; Torres-Matallana et al., 2016). Specifically, the autocorrelation of measurement errors in time-series has been proved to have a significant impact over the uncertainties assessment in similar contexts (Carvajal & Sánchez, 2018; Horner et al., 2018; Lepot et al., 2017). These autocorrelated errors and their impact in the propagation of uncertainties in the case of high temporal resolution monitoring remain largely unstudied in the urban drainage and hydrological context (Horner et al., 2018). Bertrand-Krajewski and Bardin (2002) evaluated uncertainties of common environmental indicators for a stormwater tank given by three different autocorrelation structures of the measurement errors (zero, partially, and fully correlated). This study evaluated uncertainties employing the law of propagation of uncertainties (LPU), limiting its applicability to cases with symmetric probability distributions (Sega et al., 2016). More recently, Horner et al. (2018) presented a methodology aimed to describe and propagate uncertainties in hydrological time-series with two different types of autocorrelated measurement errors, based on Bayesian inferences and the Monte Carlo method. Because both approaches were aimed to consider specific error models, further correlation structures were not analyzed.

Nevertheless, time-series with similar characteristics have reported a diversity of complex structures of autocorrelated errors, ranging from compound symmetry (CS) type (slow decay) to first order autoregressive—AR(1)—(fast decay) (Ammann et al., 2019; Simpson et al., 2010). Therefore, exploring a wider variety of possible correlation structures for measurement errors arises as a promising alternative to better understand their influence in the propagation of uncertainties. For this purpose, correlation models appear as a flexible approach to explore, not only the entire spectrum between CS and AR(1), but also further structures with faster exponential decay rates than the AR(1) model such as first-order moving average—MA(1)—(Simpson et al., 2010). On the other hand, random explorations of complex parametric spaces as for the case of correlation-covariance models can be enforced by recently developed Bayesian-based Markov Chain Monte Carlo (MCMC) sampling techniques (e.g., He, 2018; Luo et al., 2019).

In sum, this paper presents a methodology to assess the influence of randomly generated correlation-covariance structures of measurement errors in online monitoring over the propagation of uncertainties in SUDS. The proposed method is illustrated for the evaluation of different wet-weather environmental indicators (RVols, EMCs, HEs, REs) for total suspended solids (TSSs) and chemical oxygen demand (COD) in a monitored SUDS train, composed of a connected system of stormwater settling tank-horizontal constructed wetland.

## Materials and methods

### Experimental setup and data description

The present study was developed by using a database from online monitoring of a connected settling tank-horizontal constructed wetland system, aimed to collect stormwater from a parking lot (3776 $m^2$) in the Pontificia Universidad Javeriana campus. This system was conceived, constructed, and operated within the framework of Galarza-Molina et al. (2015). A scheme of the studied system is presented (Fig. 1).

The dataset collected is composed of time-series with a temporal resolution of 1 min ($\Delta t = 1$ min) of water levels and multivariate UV–Vis spectrometry absorbances (in 214 wavelengths) at the inlet and outlet of the system (Fig. 1), for 13 rainfall events (see Tables 1 and 2). In addition, laboratory analyses for 63 stormwater samples were carried out to relate TSS and COD concentrations to absorbances measured at different wavelengths. The data processing consisted of a transformation of the water levels and UV–Vis spectrometry absorbance values to flow rates and TSS and COD concentrations time-series at the inlet and outlet of the system, respectively. Regarding the water levels, the transformation into flow rates was undertaken employing a rating curve, using the inlet and outlet weirs of the system. Standard uncertainties of this transformed flow rates were reported as 0.10 L/s and 0.20 L/s for the inlet and outlet, respectively (adapted from Galarza-Molina, 2017). TSS and COD concentration time-series were obtained by means of a local calibration from laboratory data with support vector machines (SVMs), using the function "tune.svm" from the package "e1071" (Meyer et al., 2019) in the R software (R Core Team, 2019). Indeed, local calibration of UV–Vis sensors from traditional laboratory analyses (EN 872, 2005) has been widely recommended to improve the representativeness of online water quality measurements. Calibration of UV–Vis sensors by using SVM has shown appropriate performances for predicting TSS and COD concentrations in similar settings (Lepot et al., 2016). The local calibrations for the TSS and COD concentrations reported the following validation results: coefficient of determination $r^2$ of 0.88 and 0.80, root mean square error (RMSE) of 2.62 and 9.24 mg/L, and relative root mean square error (RRMSE) of 0.05 and 0.14, respectively. Imputation of missing values for the resulting time-series was performed through linear interpolation after outlier removal using Tukey's fences (Tukey, 1977). The background concentrations of the constructed wetland for both TSS and COD time-series were subtracted and aimed to consider exclusively wet-weather–related pollutants. These were taken as the minimum values of the time-series at the beginning and end of rainfall periods (Merriman et al., 2017), reported as 2 mg/L for TSS and 10 mg/L for COD.

A summary of the rainfall events information is presented in Table 1, including the mean values for the flow rates and the TSS and COD concentrations at the inlet and outlet of the system, presented hereafter with the acronyms Q, TSS, and COD and subscripts "*_in*" and "*_out*," respectively. Additionally, the hydrological information of the rainfall events is presented in Table 2.

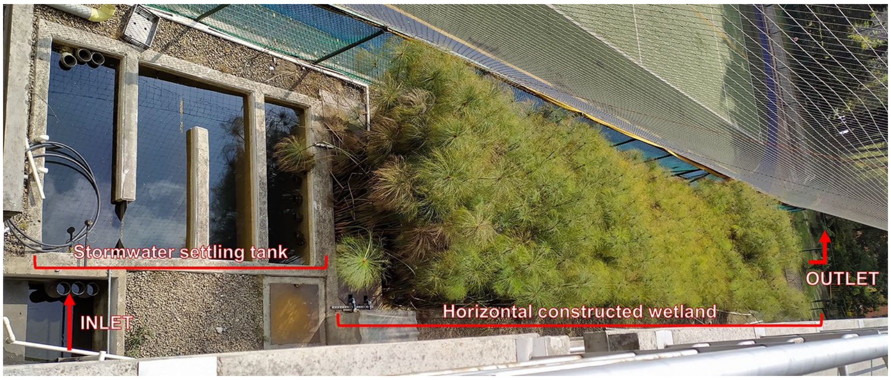


**Fig. 1** Scheme of the stormwater settling tank-horizontal constructed wetland in the Pontificia Universidad Javeriana campus

**Table 1** Mean values for Q, TSS, and COD concentrations at the inlet and outlet of the system, regarding the rainfall events

| Rainfall event | Event start (date-time) | Event end (date-time) | Mean values at the system inlet | | | Mean values at the system outlet | | |
|---|---|---|---|---|---|---|---|---|
| | | | Mean *Q_in* (L/s) | Mean *TSS_in* (mg/L) | Mean *COD_in* (mg/L) | Mean *Q_out* (L/s) | Mean *TSS_out* (mg/L) | Mean *COD_out* (mg/L) |
| 1 | 06 Feb 2015 13:24:00 | 07 Feb 2015 09:43:00 | 2.23 | 8.02 | 13.39 | 1.16 | 0.73 | 1.14 |
| 2 | 10 Feb 2015 16:23:00 | 10 Feb 2015 18:16:00 | 0.61 | 16.67 | 30.30 | 0.19 | 2.46 | 0.87 |
| 3 | 19 Feb 2015 16:56:00 | 19 Feb 2015 18:47:00 | 0.49 | 24.44 | 63.48 | 0.19 | 1.35 | 0.03 |
| 4 | 24 Feb 2015 15:40:00 | 24 Feb 2015 19:06:00 | 1.72 | 21.39 | 43.63 | 1.00 | 0.94 | 0.00 |
| 5 | 03 Mar 2015 15:11:00 | 03 Mar 2015 17:13:00 | 1.55 | 21.00 | 38.95 | 0.86 | 0.15 | 0.00 |
| 6 | 19 Mar 2015 11:05:00 | 20 Mar 2015 17:41:00 | 1.69 | 4.93 | 6.14 | 0.25 | 2.10 | 5.95 |
| 7 | 21 Mar 2015 16:29:00 | 22 Mar 2015 06:32:00 | 0.24 | 3.80 | 6.01 | 0.15 | 2.09 | 6.25 |
| 8 | 29 Mar 2015 05:42:00 | 29 Mar 2015 07:21:00 | 0.58 | 15.88 | 33.66 | 0.23 | 5.20 | 14.08 |
| 9 | 29 Mar 2015 11:51:00 | 29 Mar 2015 17:40:00 | 0.29 | 6.70 | 11.79 | 0.14 | 3.87 | 11.17 |
| 10 | 30 Mar 2015 13:42:00 | 31 Mar 2015 06:21:00 | 1.09 | 4.02 | 5.23 | 0.61 | 6.33 | 17.62 |
| 11 | 31 Mar 2015 10:51:00 | 31 Mar 2015 18:00:00 | 0.92 | 10.84 | 22.03 | 0.25 | 8.64 | 24.00 |
| 12 | 03 Apr 2015 09:00:00 | 03 Apr 2015 13:59:00 | 1.02 | 7.28 | 11.20 | 0.47 | 4.20 | 11.16 |
| 13 | 05 Apr 2015 16:24:00 | 05 Apr 2015 20:00:00 | 0.47 | 8.48 | 15.38 | 0.31 | 4.63 | 11.65 |

## Modeling the correlation-covariance structure of measurement errors

A Gaussian description of the random variable $\mathbf{X}$, aimed to represent the time-ordered measurements obtained from online monitoring (flow rate or water quality), from the beginning to the end of the measurement period (denoted hereafter as $beg : end$), can be proposed by Eq. (1) (e.g., Horner et al., 2018):

$$\mathbf{X}_{beg:end\times 1} = \mathbf{x}_{beg:end\times 1} + \boldsymbol{\varepsilon}_{beg:end\times 1} \quad (1)$$

where $\mathbf{x}_{beg:end\times 1}$ is the vector of the unknown "true" time-series, $\boldsymbol{\varepsilon}_{beg:end\times 1}$ is the vector of random error terms, assumed to be jointly normally distributed, i.e., $\boldsymbol{\varepsilon}_{beg:end\times 1} \sim N\left[0, \Sigma_{beg:end\times beg:end}\right]$, and $\Sigma_{beg:end\times beg:end}$ is the covariance matrix, denoting the covariance in the measurement error between any time-steps. The estimation of $\Sigma_{beg:end\times beg:end}$ is dependent on the correlation matrix $\mathfrak{C}_{beg:end\times beg:end}$ and the vector of standard deviations $\sigma_{beg:end\times 1}$, representing the set of standard deviations of the measurement error $\sigma_i$ at each time-step $i$. The covariance matrix $\boldsymbol{\Sigma}_{beg:end\times beg:end}$ can be directly obtained from $\mathfrak{C}_{beg:end\times beg:end}$, the identity matrix $\mathrm{I}_{beg:end\times beg:end}$ and $\sigma^{D}{}_{beg:end\times beg:end}$, which is a diagonal matrix composed of $\sigma_{beg:end\times 1}$, as follows in Eqs. (2) and (3):

$$\boldsymbol{\sigma}^{\mathbf{D}}{}_{beg:end\times beg:end} = \boldsymbol{\sigma}_{beg:end\times 1}\mathbf{e}_{1\times beg:end} \odot \mathbf{I}_{beg:end\times beg:end} \quad (2)$$

$$\boldsymbol{\Sigma}_{beg:end\times beg:end} = \boldsymbol{\sigma}^{\mathbf{D}}{}_{beg:end\times beg:end}\mathfrak{C}_{beg:end\times beg:end}\boldsymbol{\sigma}^{\mathbf{D}}{}_{beg:end\times beg:end} \quad (3)$$

**Table 2** Hydrological information of the rainfall events, collected from weather stations from Universidad Nacional de Colombia (UNAL station) and Bogotá District Institute for Risk Management and Climate Change (IDIGER station)

| Rainfall event | Event start (date-time) | Duration (min) | Total height (mm) | Max. intensity (mm/h) | Source |
|---|---|---|---|---|---|
| 1 | 06 Feb 2015 13:20:00 | 240 | 10.08 | 10.20 | UNAL station |
| 2 | 10 Feb 2015 04:40:00 | 30 | 0.40 | 1.20 | IDIGER station |
| 3 | 19 Feb 2015 16:40:00 | 60 | 3.80 | 16.80 | UNAL station |
| 4 | 24 Feb 2015 15:30:00 | 120 | 15.20 | 36.60 | UNAL station |
| 5 | 03 Mar 2015 14:30:00 | 70 | 3.40 | 5.40 | UNAL station |
| 6 | 19 Mar 2015 11:10:00 | 360 | 56.00 | 103.20 | UNAL station |
| 7 | 21 Mar 2015 15:30:00 | 50 | 3.40 | 15.60 | IDIGER station |
| 8 | 29 Mar 2015 04:20:00 | 130 | 2.30 | 7.20 | IDIGER station |
| 9 | 29 Mar 2015 11:50:00 | 45 | 1.60 | 4.80 | IDIGER station |
| 10 | 30 Mar 2015 12:25:00 | 125 | 15.30 | 39.60 | IDIGER station |
| 11 | 31 Mar 2015 13:10:00 | 60 | 8.00 | 26.40 | UNAL station |
| 12 | 03 Apr 2015 07:40:00 | 90 | 3.00 | 7.20 | IDIGER station |
| 13 | 05 Apr 2015 16:00:00 | 30 | 1.00 | 3.60 | UNAL station |

where $e_{1\times beg:end}$ is a row vector containing only values of one, and $\odot$ denotes the Hadamard product.

This study proposes an implementation based on the LEAR (linear exponent autoregressive) model of correlation structures (Simpson et al., 2010), aimed to obtain a parameterized version of $\mathfrak{C}_{beg:end\times beg:end}$ as presented in Eq. (4):

$$\mathfrak{C}_{(i,j)} = \begin{cases} \rho^{d_{min}+\delta\left[\frac{d_{i,j}-d_{min}}{d_{max}-d_{min}}\right]}, & i \neq j \\ 1, & i = j \end{cases} \tag{4}$$

$d_{i,j}$ corresponds to the number of time-steps between two observations, and $d_{min}$ and $d_{max}$ are numerical constants linked to the minimum and maximum number of time-steps between two correlated observations (Simpson et al., 2010). The $\rho$ and $\delta$ terms represent the correlation and decay speed parameters, respectively. Simpson et al. (2010) explored $\delta \geq 0$ and $0 \leq \rho \leq 1$, although he specified that $\delta < 0$ and $\rho < 0$ can generate valid autocorrelation functions. However, such patterns are hardly interpretable in biostatistical and environmental applications (adapted from Simpson et al., 2010). This correlation model can represent the entire spectrum between CS and AR(1) structures as a function of the model parameters (Simpson et al., 2010), which gives a flexible exploration approach that has not been studied before (e.g., Bertrand-Krajewski & Bardin, 2002). The LEAR model can also lead to correlation matrices that include faster exponential decays such as MA(1) (Simpson et al., 2010). For example: (i) when $\delta = 0$, LEAR structure reduces to a full-correlation model (e.g., Bertrand-Krajewski & Bardin, 2002); (ii) when $0 < \delta < d_{max} - d_{min}$, LEAR structures decrease in time at a slower rate than the imposed by the AR(1) model; (iii) when $\delta = d_{max} - d_{min}$, LEAR structure reduces to the AR(1) model; (iv) when $\delta > d_{max} - d_{min}$, LEAR structures decrease in time at a faster rate than the imposed by the AR(1) model; and (v) when $\delta \to \infty$, LEAR structures approach to the MA(1) model (Simpson et al., 2010).

On the other hand, setting $d_{min} = 1$ appears as natural selection, besides showing an appropriate convergence in the computations (Simpson et al., 2010). However, establishing $d_{max}$ as the total number of data points implies that the measurement errors between two far-apart observations can still be correlated, which remains unrealistic for urban drainage applications (Bertrand-Krajewski & Bardin, 2002). Replacing $d_{max}$ by a lag time $\tau$ and establishing any two time-steps distanced more than $\tau$ as zero-correlated might represent a more realistic model of the correlation structure in measurement errors for the case of online monitoring (Bertrand-Krajewski & Bardin, 2002). Hence, Eq. (5) is proposed as the family of correlation structures to be explored for the case study, where $\tau$ was established as an additional parameter for the proposed application.

$$\mathfrak{C}_{(i,j)} = \begin{cases} \rho^{1+\delta\left[\frac{d_{i,j}-1}{\tau-1}\right]}, & 0 < d_{i,j} \leq \tau \\ 1, & i = j \\ 0, & d_{i,j} > \tau \end{cases} \tag{5}$$