Supplement of Hydrol. Earth Syst. Sci., 25, 193–216, 2021
https://doi.org/10.5194/hess-25-193-2021-supplement

Supplement of

Multivariate autoregressive modelling and conditional simulation for temporal uncertainty analysis of an urban water system in Luxembourg

Jairo Arturo Torres-Matallana et al.

Correspondence to: Jairo Arturo Torres-Matallana (arturo.torres@list.lu)

S1 Model inputs

Table S1 presents the most important general CSO input and output variables of EmiStatR. Table S2 presents the base values of input variables for each individual CSO.

Table S1: Most important general, CSO input and output variables of EmiStatR, with base values for the general input variables.

General input	Units	Base value	CSO input	Units
1. Wastewater			*1. Catchment data*	
Water consumption, qs	$L(PE^{a}\ d)^{-1}$	152	Total area, A_{total}	ha
Pollution COD[b], $C_{COD,S}$	$g\ PE^{-1}\ d^{-1}$	104.2	Impervious area, A_{imp}	ha
Pollution NH_4[c], $C_{NH4,S}$	$g\ PE^{-1}\ d^{-1}$	4.7	Run-off coeff.[d] for impervious area, C_{imp}	−
2. Infiltration water			Run-off coeff. for pervious area, C_{per}	−
Inflow, q_f	$L\ s^{-1}\ ha^{-1}$	0.116	Flow time structure, t_{fS}	[time step]
Pollution COD, COD_f	$g\ PE^{-1}\ d^{-1}$	0	Population equivalents, pe	PE
Pollution NH_4, $NH4_f$	$g\ PE^{-1}\ d^{-1}$	0	*2. CSO structure data*	
3. Rainwater			Volume, V	m^3
Rain time series, P	mm		Initial water level, Lev_{ini}	m
Pollution COD, COD_r	$mg\ L^{-1}$	71.0	Maximum throttled outflow, $Q_{d,\ max}$	$L\ s^{-1}$
Pollution NH_4, $NH4_r$	$mg\ L^{-1}$	2.0	Orifice diameter, D_d	m
			Orifice coefficient of discharge, C_d	−
Output variables				
1. Quantity				
Volume in the CSO chamber, $V_{Chamber}$	m^3			
Overflow spill volume, V_{Sv}	m^3			
Overflow spill flow, Q_{Sv}	$L\ s^{-1}$			
2. Quality				
Spill COD load, $B_{COD,Sv}$	g			
Average spill COD conc.[e], $C_{COD,Sv,av}$	$mg\ L^{-1}$			
99.9th perc.[f] spill COD conc., $C_{COD,Sv,99.9}$	$mg\ L^{-1}$			
Maximum overflow COD conc., $C_{COD,Sv,max}$	$mg\ L^{-1}$			
Spill NH_4 load, $B_{NH4,Sv}$	g			
Average spill NH_4 conc., $C_{NH4,Sv,av}$	$mg\ L^{-1}$			
99.9th perc. spill NH_4 conc., $C_{NH4,Sv,99.9}$	$mg\ L^{-1}$			
Maximum spill NH_4 conc., $C_{NH4,Sv,max}$	$mg\ L^{-1}$			

[a]PE = population equivalents; [b]COD = chemical oxygen demand; [c]NH_4 = ammonium;
[d]coef. = coefficient; [d]conc. = concentration; [f]perc. = percentile.

Table S2: The CSO structure input data for the EmiStatR model, after calibration. Structures 2 and 3, only C_d was calibrated.

CSO input			
1. Identification			
ID of the structure	1	2	3
Name of the structure	FBH Goesdorf	FBN Kaundorf	FBH Nocher-Route
2. Catchment data			
Name of the municipality	Goesdorf	Kaundorf	Nocher-Route
Name of the catchment	Haute-Sûre	Haute-Sûre	Haute-Sûre
Number of the catchment	1	1	1
Land Use[a]	R/I	R/I	R/I
Total area, A_{total} (ha)	30.0	22	18.6
Impervious area, A_{imp} (ha)	5.0	11.0	4.3
Run-off coefficient for impervious area, C_{imp} $(-)$	0.28	0.3	0.3
Run-off coefficient for pervious area, C_{per} $(-)$	0.07	0.10	0.10
Flow time structure, t_{fS} (time step)	1	2	2
Population equivalents, pe (PE)	611	358	326
3. CSO structure data			
Volume, V (m^3)	190	180	157
Initial water level, $\mathrm{Lev}_{\mathrm{ini}}$ (m)	0.57	1.8	1.8
Maximum throttled outflow, $Q_{\mathrm{d,max}}$ (L s^{-1})	5.0	9	4
Orifice diameter, D_{d} (m)	0.15	0.20	–
Orifice coefficient of discharge, C_{d} $(-)$	0.67	0.67	0.67

[a] R = residential, I = industrial.

S2 Selection of model inputs for uncertainty quantification

Regarding the selection of model inputs for uncertainty quantification, to better support our decisions we include Figure S1, as in Tscheikner-Gratl et al. (2017). Table S3 shows comparisons of the means and variances for $C_{\mathrm{COD,S}}$ and $C_{\mathrm{NH4,S}}$ based on 91 measurements in the Haute-Sure catchment and simulations at Goesdorf (note that for $\mathrm{COD_r}$ a comparison could not be made because we had no measurements of $\mathrm{COD_r}$ and a model for $\mathrm{COD_r}$ was based on expert judgement). The agreement between observed and simulated statistics is again quite close. We could not evaluate the autocorrelation functions of the observed $C_{\mathrm{COD,S}}$ and $C_{\mathrm{NH4,S}}$ because there were too few observations to be able to compute these (note that the 91 observations were from multiple locations within the catchment).

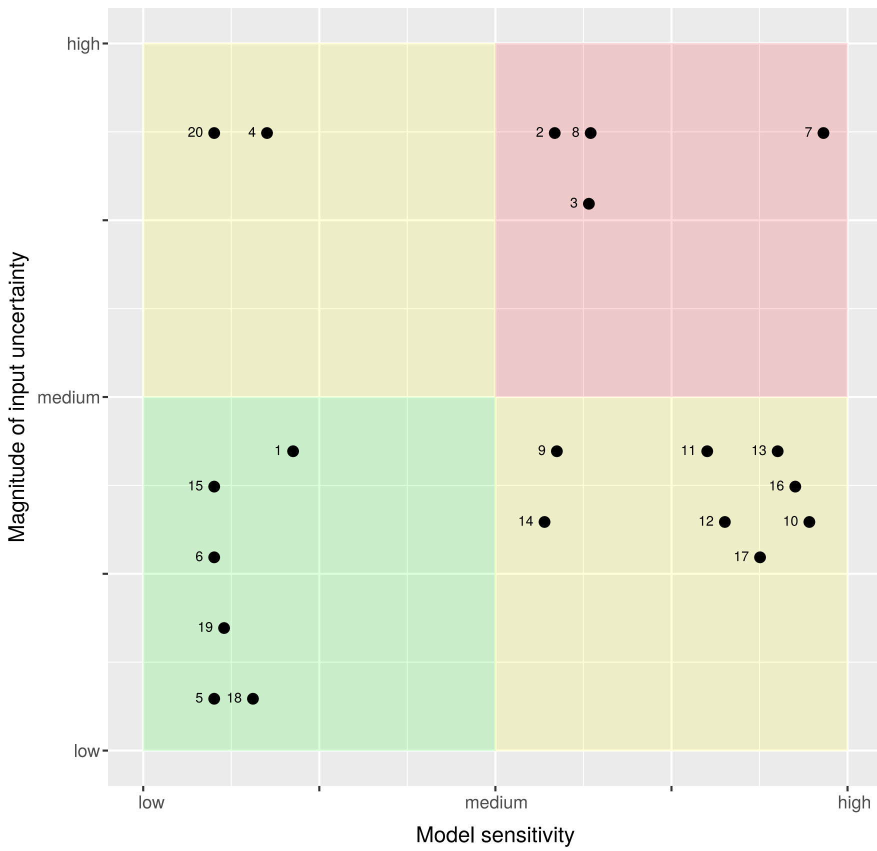

Figure S1: Graphical assessment of the contribution of input uncertainty to model output uncertainty. Numbers near each dot refer to the input variable number as defined in Table 2 of the manuscript. Panel layout after Tscheikner-Gratl et al. (2017).

Table S3: Mean and variance of log-transformed observed $C_{\mathrm{COD,S}}$ and $C_{\mathrm{NH4,S}}$ in the Haute-Sure catchment and of log-transformed simulated $C_{\mathrm{COD,S}}$ and $C_{\mathrm{NH4,S}}$ at Goesdorf (random selection of simulation numbers 1, 750, and 1500).

	Observations	Sim 1	Sim 750	Sim 1500	Sims (All)
Mean ($\log(C_{\mathrm{COD,S}})$)	4.3783	4.3752	4.3737	4.4106	4.3780
Variance ($\log(C_{\mathrm{COD,S}})$)	0.5637	0.5261	0.5257	0.5394	0.5640
Mean ($\log(C_{\mathrm{NH4,S}})$)	1.4733	1.4656	1.4639	1.4865	1.4730
Variance ($\log(C_{\mathrm{NH4,S}})$)	0.1679	0.1704	0.1684	0.1615	0.1681

S3 Model input characterisation and observations

The simulated precipitation time series captured the main statistics of the observed time series well. Evidence for this is presented in Table S4 and Figure S2.

Table S4: Mean and variance of the log-transformed observed precipitation time series at Esch-sur-Sure and Dahl rain gauges and the simulated precipitation time series at Goesdorf (random selection of simulation numbers 1, 750, 1500 and all).

	Esch-sur-Sure	Dahl	Sim 1	Sim 750	Sim 1500	Sims (All)
Mean	-6.6152	-6.5817	-6.3888	-6.3886	-6.3878	-6.3874
Variance	1.4188	1.5731	1.5636	1.5579	1.5594	1.5582

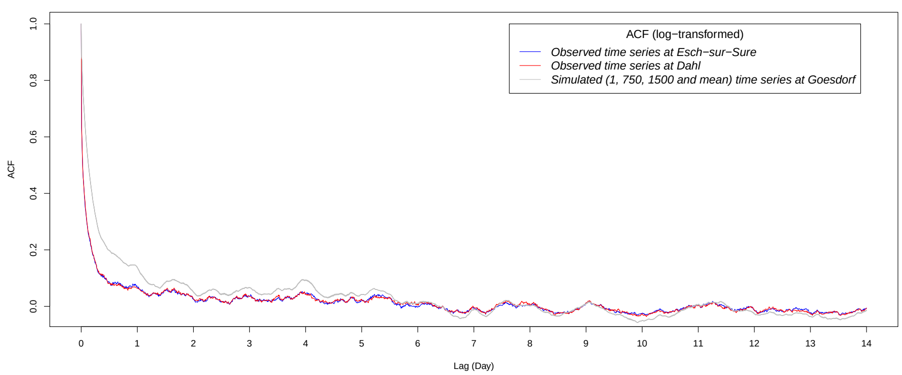

Figure S2: Autocorrelation function of the log-transformed observed precipitation time series at Esch-sur-Sûre and Dahl rain gauges and simulated precipitation at Goesdorf catchment.

S4 Input precipitation model: calibration and conditional simulation

S4.1 *Calibration*

We begin by relating the two precipitation time series as:

$$P_1(t) = P_2(t) \cdot \delta(t) \quad (1)$$

where $\delta(t)$ is a positive multiplicative factor that varies over time. We assume that $P_1(t)$, $P_2(t)$ and $\delta(t)$ are stationary and log-normally distributed stochastic processes. After log-transformation we get

$$\log(P_1(t)) = \log(P_2(t)) + \log(\delta(t)) \tag{2}$$

We apply a Kernel (daniell) smoothing to the precipitation time series to avoid rapid fluctuation of the time series for precipitation depth values smaller than 0.1 mm. This also solves problems associated with taking logarithms of near-zero values. Next, in order to estimate the parameters of $\delta(t)$, we filter the time series allowing the computation of a ratio between the two measured time series. This ratio represents the difference in precipitation as registered in two nearby rain gauge stations. It is computed only for those cases where the precipitation depth of the two time series is greater than 0.01 mm.

To simplify notation we write $LP_1(t) = \log(P_1(t))$, $LP_2(t) = \log(P_2(t))$ and $L\delta(t) = \log(\delta(t))$. Since two out of three determine the third, we need only define two processes. We model the joint distribution of $LP_1(t)$ and $L\delta(t)$ by a bivariate AR(1) process, as introduced before:

$$\begin{bmatrix} LP_1(t+1) \\ L\delta(t+1) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_\delta \end{bmatrix} + \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \left(\begin{bmatrix} LP_1(t) \\ L\delta(t) \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_\delta \end{bmatrix} \right) + \begin{bmatrix} \varepsilon_1(t+1) \\ \varepsilon_\delta(t+1) \end{bmatrix} \tag{3}$$

where ε_1 and ε_δ are zero-mean, cross-correlated and normally distributed white noise processes.

To calibrate this model, i.e. estimate its parameters μ_1, μ_δ, B_{11}, B_{12}, B_{21}, B_{22}, σ_1^2, σ_δ^2 and $\rho_{1\delta}$, where $\sigma_1^2 = \mathrm{var}(\varepsilon_1)$, $\sigma_\delta^2 = \mathrm{var}(\varepsilon_\delta)$ and $\rho_{1\delta}$ is the correlation between ε_1 and ε_δ, we used the R package mAr (Barbosa, 2015). Calibration is based on two time series of LP_1 and $L\delta$ derived from observed time series P_1 and P_2.

S4.2 *Conditional simulation*

To simulate a time series P for the target catchment from an observed time series P_o at a nearby location, we make use of the fact that the calibrated AR(1) model quantifies how precipitation at one location relates to that at a nearby location. We make use of Eq. 1:

$$P(t) = P_o(t) \cdot \delta(t) \tag{4}$$

This requires simulations of $\delta(t)$. These are obtained using the calibrated model Eq. 3, but now applied to the vector $[LP_o \quad L\delta]^T$, which characterises the joint pdf of LP_o and $L\delta$. We use this model to simulate $L\delta$ conditional to the observed time series LP_o. Since the two processes are jointly normally distributed we can make use of a well known property of the multivariate normal distribution (Searle, 1997, page 47). Let U and V be two jointly normally distributed random vectors. The conditional distribution of U given $V = v$ is then also normal and given by:

$$\{U|V = v\} \sim N\left(E[U] + cov(U,V) \cdot var(V)^{-1} \cdot (v - E[V]),\ var(U) - cov(U,V) \cdot var(V)^{-1} \cdot cov(V,U)\right) \tag{5}$$

We make use of this equation to simulate δ by substituting:

$$U = L\delta(t+1) \qquad V = \begin{bmatrix} L\delta(t) \\ LP_o(t+1) \\ LP_o(t) \end{bmatrix} \tag{6}$$

for all $t = 1, ..., T$, while substituting the observed time series LP_o for v.

For more details we refer to Torres-Matallana et al. (2017).

S5 Uncertainty quantification of selected model input

Figure S3 presents the histogram of observations, empirical density and theoretical normal density for $\log(C_{\mathrm{COD,S}})$, $\log(C_{\mathrm{NH4,S}})$ and $\log(\mathrm{COD_r})$.

S6 Monte Carlo simulation size and timing

In order to perform the MC propagation analysis, we first did a convergence test to estimate the number of simulations required. Besides this test, we also computed the standard error of all MC outputs. These two methods have the same aim and are closely related. In the convergence test, the standard deviation of two different MC simulations with different random seeds were computed and compared for the seven output variables of EmiStatR, three representing water quantity variables (V_{Chamber}, V_{Sv} and Q_{Sv}) and four for water quality ($B_{\mathrm{COD,Sv}}$, $B_{\mathrm{NH4,Sv}}$, $C_{\mathrm{COD,Sv,av}}$, and $C_{\mathrm{NH4,Sv,av}}$). The results of the test indicated that in most cases between 250 and 1,000 MC simulations are enough to reach stable results in terms of the Nash–Sutcliffe model efficiency coefficient (NSE), where a NSE of 1 means a perfect match between observations and model output. In this case we got a NSE $\approx$ 0.998 for overflow volume. Regarding the water quality variable $B_{\mathrm{COD,Sv}}$, the test showed that a larger number of MC simulations is required. Between 1,000 and 2,000 simulations are required to reach stable results (NSE $\approx$ 0.880 for overflow COD load and 0.998 for overflow NH_4 load). Therefore, a number of 1,500 MC simulations was used to perform the uncertainty analysis of the water quantity and water quality outputs. Figure S4 illustrates results of the convergence test for the cases

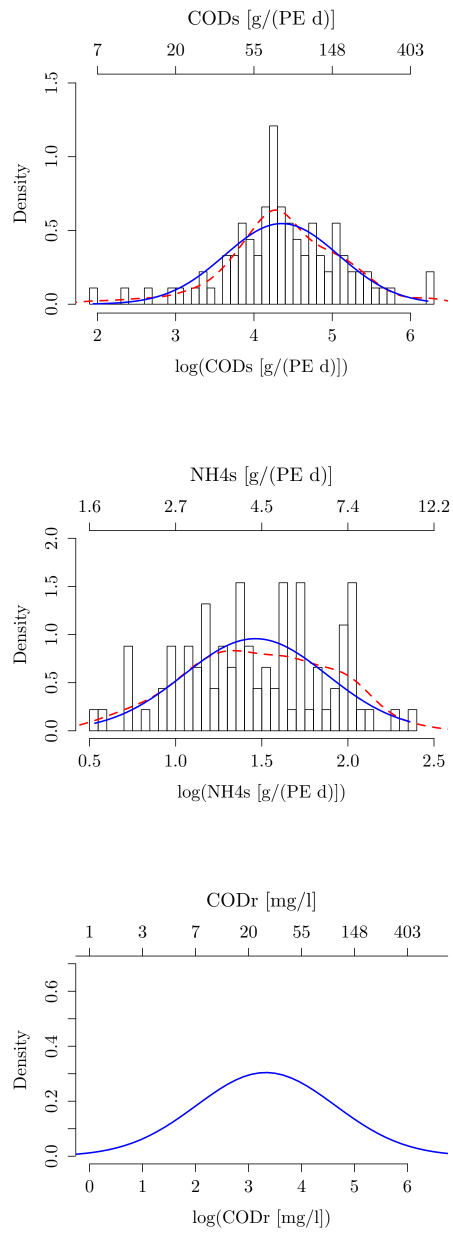

Figure S3: Histogram of observations, empirical density (red dashed line) and theoretical normal density (blue line) for (a) $\log(C_{\mathrm{COD,S}})$; (b) $\log(C_{\mathrm{NH4,S}})$; (c) $\log(\mathrm{COD_r})$. Note that the blue densities were used in the uncertainty propagation.

Table S5: Average running time in minutes for Monte Carlo (MC) replications and specific cores used with two different seeds for the pseudo-random number generator in R. The rainfall input used was a one-month length time series with 10 minutes time steps from 1 to 31 August 2010 (4,464 time steps).

Replications	250	500	1000	1500	2000
cores	3	3	50	50	50
MC1	7.12	14.23	4.84	7.33	9.40
MC2	7.09	14.63	4.96	7.26	9.53
Average	7.10	14.43	4.90	7.29	9.46

where the number of MC replications is 250, 1,000 and 1,500. In this figure the MC1 output is plotted on the x-axis and MC2 output on the y-axis. Although the model output corresponds to yearly time series at 10 minutes resolution, we only plotted those points where the overflow magnitude, and therefore COD and NH_4 load, is different from zero. As an indication, for a MC replication size of 1,500, the NSE values for overflow COD and NH_4 concentrations are 0.816 and 0.998, respectively.

The computing times per MC replication are presented in Table S5. The computations were performed with two different Linux machines, a laptop with four cores for simulations between 50 and 500 replications, and a server with 80 cores for performing the simulations above 500 replications. Similar execution times were reached for MC1 and MC2 for one-month time series at 10-min time steps (August 2010, 4,464 time steps), while substantial differences were obtained when the 80 cores server was used. We obtained similar timing for 1,500 replications with 50 cores as for 250 replications using three cores in the laptop. The timing reached demonstrates the feasibility to perform a solid MC uncertainty propagation analysis with EmiStatR.

References

Barbosa, S. M.: Package "mAr": Multivariate AutoRegressive analysis, The Comprehensive R Archive Network, CRAN, 1.1-2 edn., 2015.

Searle, S.: Linear Models, Wiley, 1997.

Torres-Matallana, J. A., Leopold, U., and Heuvelink, G. B. M.: Multivariate autoregressive modelling and conditional simulation of precipitation time series for urban water models, European Water, 57, 299–306, 2017.

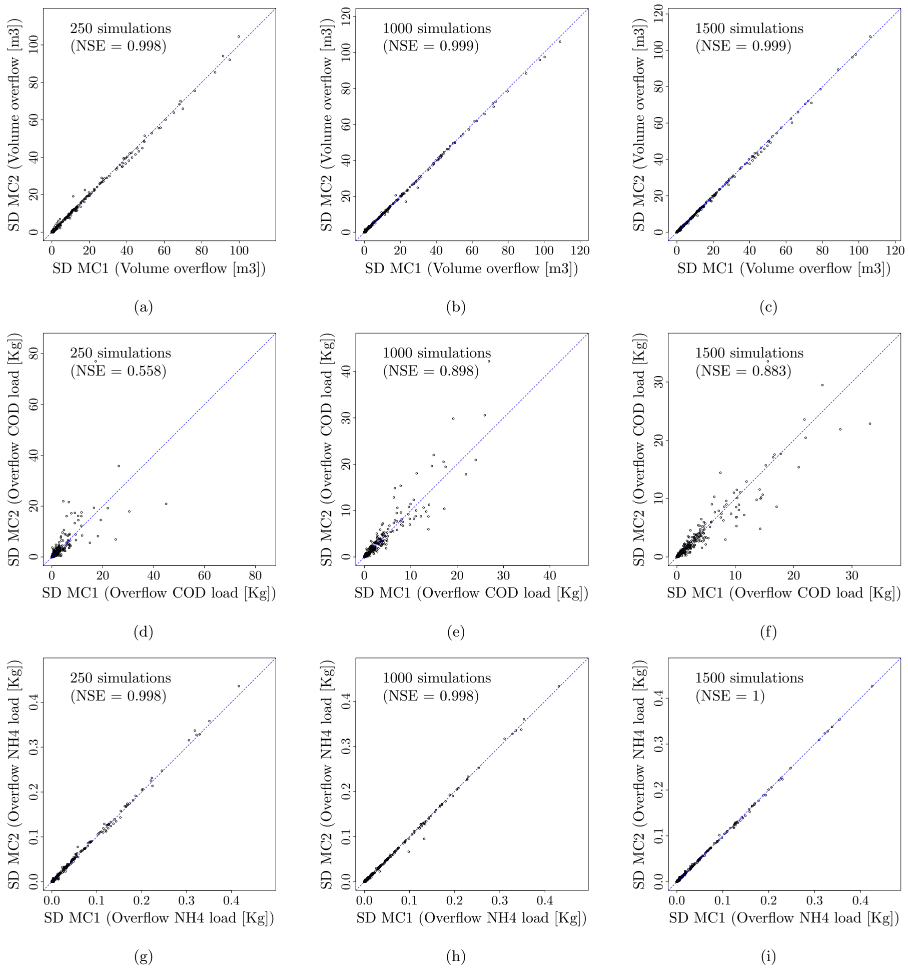

Figure S4: Results of the MC convergence test for (a, b, c) volume in overflow; (d, e, f) overflow COD load; (g, h, i) overflow NH_4 load. Each open circle refers to a ten minute time instant in 2010 were overflow happened. As an indication, for a MC replication size of 1,500, the NSE values for overflow COD and NH_4 concentrations are 0.816 and 0.998, respectively. Dotted line is the 1:1 line. SD = Standard Deviation.