UDM 2018
11th International Conference on Urban Drainage Modelling
23-26 Sep | Palermo - Italy

Parametric inference in large water quality river systems

Antonio Moreno-Rodenas[1], Jeroen Langeveld[1] and Francois Clemens[1, 2]
[1]Delft University of Technology, Sanitary Engineering, Water Management, Delft, The Netherlands
[2] Deltares, Delft, The Netherlands

Abstract: Environmental models often contain parameters, which are not measurable, yet conceptual descriptions of some physical process. The value of such parameters is often derived by measuring internal state model variables in the system and indirectly tuning/calibrating the value of the parameters so some degree of match is achieved. Bayesian inference is a widely used tool in which the modeller can transfer some prior beliefs about the parameter space, which is updated when additional knowledge on the system is acquired (e.g. more measurements are available). However, the amount of simulations required to perform a formal inference becomes prohibitive when using computationally expensive models. In this work the inference of the hydraulic and dissolved oxygen processes is presented for a large scale integrated catchment model. Two emulator structures were used to accelerate the sampling of the river flow and dissolved oxygen dynamics. Posterior parameter probability distributions were computed using one year of measured data in the river.

Keywords: Integrated Catchment Modelling, Emulation and Water Quality

1. INTRODUCTION

The description of processes in environmental modelling is seldom purely physically based. This is due to an incomplete understanding of the real underlying dynamics, to the lack of field measurements or due to a need of simplification. This leads to the use of non-physical parameters, which cannot be directly measured or that lump several processes. The value of such parameters is calibrated such that the model and reality have a quantifiable degree of resemblance. The transferability of parameters from one system to others is mostly limited, yet the modeller often has some prior knowledge acquired by simulating similar cases, which could be used in the calibration process. This process is often approached from a Bayesian perspective, in which the modeller encodes its knowledge as a joint probability distribution of the parameters, which are updated in view of new data.

Integrated urban water modelling focuses on the joint simulation of processes affecting water dynamics through the urban-river system. These models jointly evaluate wastewater treatment processes, urban drainage and river dynamics. This often generates a rapid escalation of complexity. The representation of all subsystems involved produces highly parameterized conceptualizations, which requires a large amount of data in the identification-calibration process. Additionally, the dynamics of interest often occur at very different time-space scales, for instance, urban CSO discharges have a characteristic time of minutes-hours whereas river dissolved oxygen dynamics is at hourly-monthly scale. This sometimes leads to the inference been performed on long time-series and often in several measured state-variables. Although, when possible, calibration is done sequentially by decomposing the model in upstream to downstream independent regions, it is common to have an inference set-up, which depends in a computationally expensive model to be evaluated at long time-series samples.

Inference schemes often require a large number of model evaluations (~10,000s) to reach convergence. This renders the inference in the original model impractical. This problem has been approached from two main directions: a) reducing the number of required samples by creating optimal sampling schemes (Laloy and Vrugt 2012), b) accelerating model sampling through model emulation (Carbajal et al. 2017).

Data-driven model emulation focuses on reproducing the link between a set of given inputs/parameters to one or more outputs of interest. This is achieved by creating known samples at a series of given inputs-outputs and creating an interpolator, which approximates the model output at new given inputs. This process becomes a challenge when the dimensionality of the problem grows. In this work we present a case in which a flow and dissolved oxygen modelled time-series (1 year) is emulated to 4 and 8 parameters respectively in a large-scale integrated catchment system. This emulator is then used to infer the posterior probability distribution of the model parameters by using a measured dataset.

2. MATERIALS AND METHODS

2.1 Integrated Catchment modelling

The river Dommel is located in the south of The Netherlands. This river system presents severe oxygen depletion processes under heavy rainfall conditions. This is mainly originated by Combined Sewer Overflow (CSO) discharges in the river result of the overloading of the urban drainage system. The river receives the discharge of several municipalities through ~200 CSOs. A full-scale integrated catchment model has been used in the system in order to assess the origin of pollution and direct measures for its reduction. The model is developed in WEST (DHI) and includes urban drainage, wastewater treatment and river processes with the objective of simulating dissolved oxygen processes in the receiving water body. Further information on the case study can be found in (Langeveld et al. 2013, Moreno-Rodenas et al. 2017). One year of measured data are available in the system (Jan – Dec 2012).

2.4 Emulator

A polynomial chaos expansion was used to link the vector of hydraulic and water quality parameters to the targeted outputs of the model (Bellos et al. 2017, Xiu and Karniadakis 2002). A model response database was built assuming uniformly distributed parameters under a Latin hypercube sampling scheme. This was created by drawing 1000 samples of flow and 2000 of dissolved oxygen dynamics to combinations of the model parameters. From those simulations 200 were used to validate the emulator performance and the rest were used during the training process. Lagrange polynomials were used of order 4 and 5.

2.3 Inference

The prior knowledge was encoded as uniform distributions. An independently and identical Gaussian distribution was initially proposed as a likelihood description. However, a first inference attempt rendered a non-stationary variance in the residual structure. Thus the model error was updated using a linearly related variance with the variable's value:

$$\sigma^2(Q_t) = (std1 + std2 \cdot Q_t)^2 \quad (1)$$

this generated two extra error model hyper-parameters std1 and std2 which are also to be inferred from the available data, Q_t attended to the value at time t of the simulated variable. Thus the log likelihood took the form:

$$L(Y|\theta) = -\frac{1}{2}\log(\det(\boldsymbol{\Sigma})) - \frac{1}{2}\left(Y - M(\theta)\right)^T \cdot \boldsymbol{\Sigma}^{-1} \cdot (Y - M(\theta)) \quad (2)$$

where $\boldsymbol{\Sigma} = \mathbf{I} \cdot \bar{\sigma}^2(Q)$ with $\mathbf{I} \in \mathbb{R}^{mxm}$ the identity matrix and $\bar{\sigma}^2 \in \mathbb{R}^{mx1}$ the vector of variances (from eq(1)). $M(\theta) \in \mathbb{R}^{mx1}$ is the model's output and $Y \in \mathbb{R}^{mx1}$ the observations. A MCMC algorithm was used to sample from the posterior distribution of the parameters. 50,000 samples were drawn in each case. The process adopted for the inference was the following: 1) Create a database of model parameter-outputs. 2) Propose an emulator structure through a polynomial expansion. 3) Validate the emulator performance by using a testing dataset. 4) Propose a model/measurement error structure. 5) Use measured data in the system to infer the posterior parameter distribution. 6) Validate the assumptions made in the error model.

3. RESULTS AND DISCUSSION

Figure 1 shows the Nash-Sutcliffe efficiency performance indicator between emulator structures and the simulator output for one-year of flow and DO dynamics at 200 parameter combinations from the validation dataset. The values for both series are consistently close to 1, which indicates a good agreement with the simulator behaviour, thus allowing substituting the model by its emulator during the inference process.

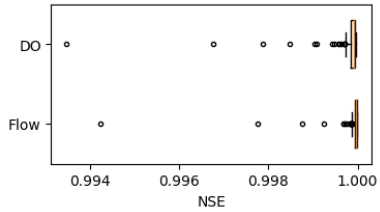

Figure 1. Box plot of NSE performance emulator vs simulator for the flow and DO dynamics

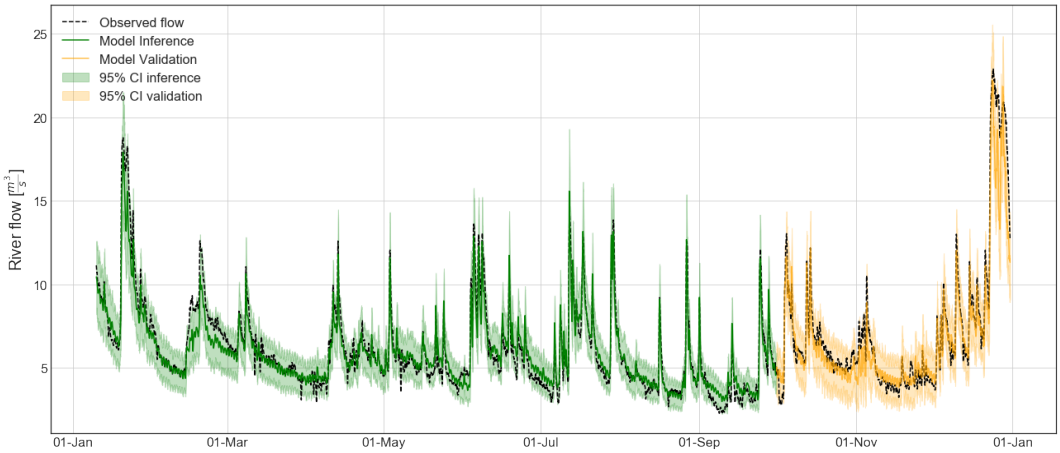

Figure 2. Measurements vs inferred model mean and 95% confidence interval for the inferred and validation river flow time-series.

Figure 2 and Figure 3 present the model expected value and its 95% confidence interval for the inferred dynamics of flow and DO respectively. 9 months were used in the inference process and 3 months where used for validation purposes. The residual structure presented a certain autocorrelation, which would require updating the error model. However, the generalization of the likelihood to accommodate time-dependency becomes highly

computationally expensive in this case (since the number of measured points is very large ~8000). The inference at the DO dynamics generated a large model error variance. This can be due to the fact that the model could not capture some oxygen depletion events (meaning that some process might be missing in the model conceptualization) or that the error description was insufficient. Which will be further explored.

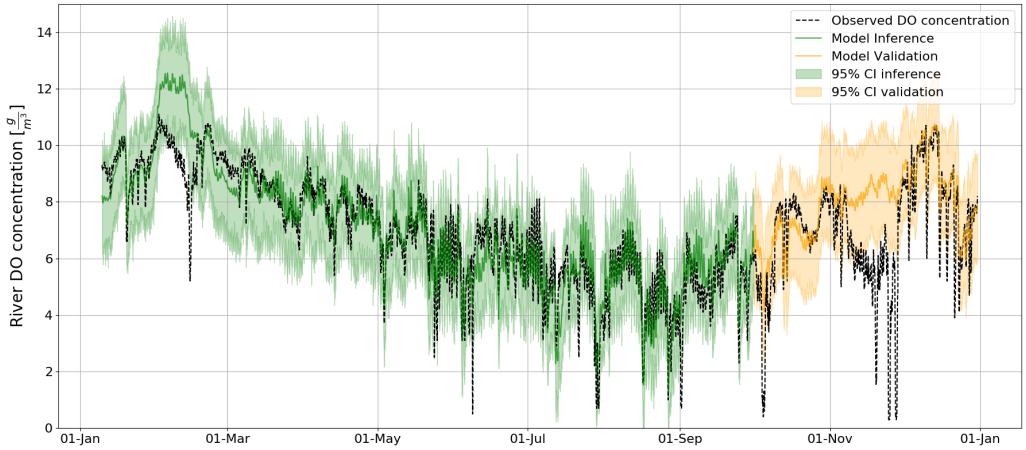

Figure 3. Measurements vs inferred model mean and 95% confidence interval for the inferred and validation river DO time-series.

4. CONCLUSIONS

Performing parameter inference becomes prohibitive when using computationally expensive models. This hampers severely the use of Bayesian inference in large scale Integrated Catchment Modelling (ICM). In this work an example is presented in which an ICM is used to simulate flow and dissolved oxygen depletion processes during one year. The use of an emulator made the inference of the model parameters feasible. Also, this can be easily extended to generate fast uncertainty quantification schemes in large integrated catchment simulators.

References

Bellos, V., Kourtis, I., Moreno-Rodenas, A. and Tsihrintzis, V. (2017) Quantifying Roughness Coefficient Uncertainty in Urban Flooding Simulations through a Simplified Methodology. Water 9(12), 944.

Carbajal, J.P., Leitão, J.P., Albert, C. and Rieckermann, J. (2017) Appraisal of data-driven and mechanistic emulators of nonlinear simulators: The case of hydrodynamic urban drainage models. Environmental Modelling & Software 92, 17-27.

Laloy, E. and Vrugt, J.A. (2012) High-dimensional posterior exploration of hydrologic models using multiple-try DREAM(ZS)and high-performance computing. Water Resources Research 48(1).

Langeveld, J.G., Benedetti, L., de Klein, J.J.M., Nopens, I., Amerlinck, Y., van Nieuwenhuijzen, A., Flameling, T., van Zanten, O. and Weijers, S. (2013) Impact-based integrated real-time control for improvement of the Dommel River water quality. Urban Water Journal 10(5), 312-329.

Moreno-Rodenas, A., Cecinati, F., Langeveld, J. and Clemens, F. (2017) Impact of Spatiotemporal Characteristics of Rainfall Inputs on Integrated Catchment Dissolved Oxygen Simulations. Water 9(12), 926.

Xiu, D. and Karniadakis, G.E. (2002) The Wiener--Askey polynomial chaos for stochastic differential equations. SIAM journal on scientific computing 24(2), 619-644.