Statistical emulation as a tool for analysing complex multiscale stochastic biological model outputs

O.K., Oyebamiji[*], D.J., Wilkinson[**]
* School of Mathematics & Statistics, Newcastle University, NE1 7RU, Oluwole.Oyebamiji@newcastle.ac.uk
** School of Mathematics & Statistics, Newcastle University, NE1 7RU, UK, darren.wilkinson@newcastle.ac.uk

The performance of credible simulations in open engineered biological frameworks is an important step for practical application of scientific knowledge to solve real-world problems and enhance our ability to make novel discoveries. Therefore, maximising our potential to explore the range of solutions at frontier level could reduce the potential risk of failure on a large scale. One primary application of this type of knowledge is in the management of wastewater treatment systems. Efficient optimisation of wastewater treatment plant focuses on aggregate outcomes of individual particle-level processes. One of the crucial aspects of engineering biology approach in wastewater treatment study is to run a high complex simulation of biological particles. This type of model can scale from one level to another and can also be used to study how to manage real systems effectively with minimal physical experimentation.

To identify crucial features and model water treatment plants on a large scale, there is a need to understand the interactions of microbes at fine resolution using models that could provide the best available representation of micro scale responses. The challenge then becomes how we can transfer this small-scale information to the macroscale process in a computationally efficient and sufficiently accurate way. It has been established that the macro scale characteristics of wastewater treatment plants are the consequences of microscale features of a vast number of individual particles that produce the community of such bacterial (Ofiteru et al. 2014).

Nevertheless, simulation of open biological systems is challenging because they often involve a large number of bacteria that ranges from order 10^{12} to 10^{18} individual particles and are physically complex. The models are computationally expensive and due to computing constraints, limited sets of scenarios are often possible. A simplified approach to this problem is to use a statistical approximation of the simulation ensembles derived from the complex models which will help in reducing the computational burden. Our aim is to build a cheaper surrogate of the Individual-based (IB) model simulation of biological particle. The main issue we address is to highlight the strategy for emulating high-level summaries from the IB model simulation data.

Our approach is to condense the massive, long time series outputs of particles of various species by spatially aggregating to produce the most relevant outputs in the form of floc and biofilms aggregates. The data compression has the benefit of suppressing or reducing some of the nonlinear response features, simplifying the construction of the emulator. Some of the most interesting properties at the mesoscale level like the size, shape, and structure of biofilms and flocs are characterised, see Figure 1. For instance, we characterize the floc size using an equivalent diameter. This strategy enables us to treat the flocs as a ball of a sphere

and or fractal depending on the shape, and we approximate the diameter of a sphere that circumscribes its boundary or outline.

We use Gaussian process emulation in the form of kriging metamodels where output data can be decomposed into a mixture of deterministic (non-random trend) and a residual random variation. In particular, we develop dynamic emulators for the multi-outputs simulation data using a multivariate kriging. The kriging model is formulated appropriately to filter the noise derived from replicate simulations. Due to the nature of output data from the simulation model, we use a dynamic emulation technique. Dynamic emulation models the evolution or trajectory of random variables over some time-steps (Conti et al. 2009). Finally, we perform the sensitivity analysis of the kriging model by calculating the total effects of each explanatory variable which helps to identify the relative importance of variables in the model.

Results

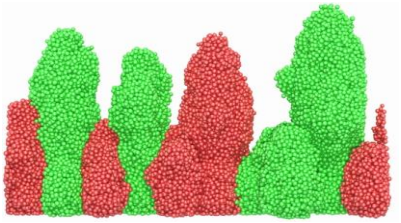

Figure 1b: rough surface flocs

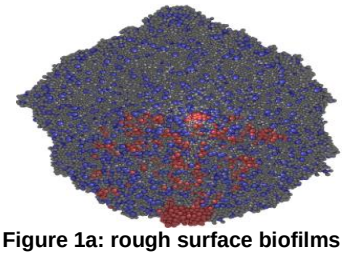

Figure 1a: rough surface biofilms

Figure (1) is the simulation data showing the transformation of microscale particles to biofilms and floc aggregates at the mesoscale for a particular time.

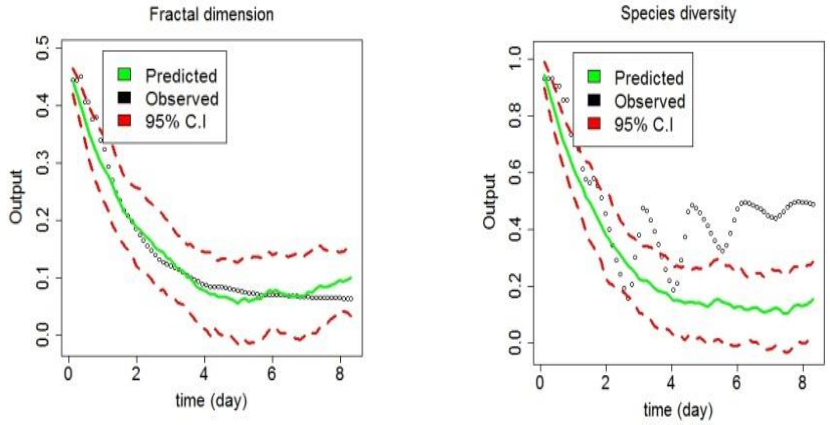

Figure 2: Comparison of the emulator performance with simulation data for two characterized outputs from IB model of floc simulation (black) and their emulator predictions (green) with 95% C.I (red).

In Figure (2), the emulator for the fractal dimension predicts the temporal behaviour relatively well, almost all the points lie within the 95% C.I. The predicted bands remain very small. The species diversity emulator produces similar pattern to the simulation data although after day ”3” the emulator deviates from the usual trend but not significantly. We note that the shape, size and structure of biofilms and floc are essential operation parameters in the management of wastewater.

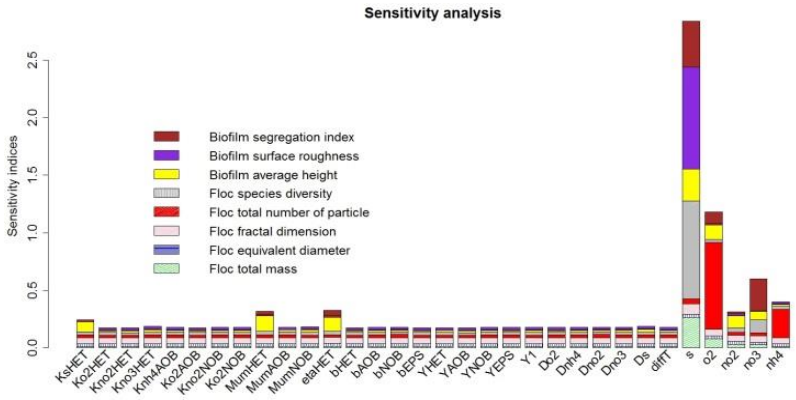

Figure 3: Barplots showing the kriging based sensitivity indices for the eight characterized outputs from IB model.

The sensitivity analysis in Figure (3) indicates that nutrient boundary conditions are the most critical parameters for predictions of most of the outputs. These parameters regulate the distribution and transports of nutrients across the computational domain thus determine the particle growth and division.

References:

Conti, S., Gosling, J. P., Oakley, J. E., &O'hagan, A. (2009), Gaussian process emulation of dynamic computer codes. *Biometrika*, asp028.

• Ofiteru, I. D., Bellucci, M., Picioreanu, C., Lavric, V., & Curtis, T. P. (2014), Multi-scale modelling of bioreactor–separator system for wastewater treatment with two-dimensional activated sludge floc dynamics. *Water Research*, 50, 382-395.