Journal of Computational Science 22 (2017) 69–84

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

Gaussian process emulation of an individual-based model simulation of microbial communities

O.K. Oyebamiji[a,*], D.J. Wilkinson[a], P.G. Jayathilake[b], T.P. Curtis[c], S.P. Rushton[d], B. Li[e], P. Gupta[d]

[a] School of Mathematics & Statistics, Newcastle University, United Kingdom
[b] Department of Mechanical & Systems Engineering, Newcastle University, United Kingdom
[c] School of Civil Engineering and Geosciences, Newcastle University, United Kingdom
[d] School of Biology, Newcastle University, United Kingdom
[e] School of Computing Science, Newcastle University, United Kingdom

ARTICLE INFO

Article history:
Received 20 January 2017
Received in revised form 3 August 2017
Accepted 8 August 2017
Available online 14 August 2017

Keywords:
Individual-based models
Multivariate Gaussian process
Biofilms
Flocs
Emulator

ABSTRACT

The ability to make credible simulations of open engineered biological systems is an important step towards the application of scientific knowledge to solve real-world problems in this challenging, complex engineering domain. An important application of this type of knowledge is in the design and management of wastewater treatment systems. One of the crucial aspects of an engineering biology approach to wastewater treatment study is the ability to run a simulation of complex biological communities. However, the simulation of open biological systems is challenging because they often involve a large number of bacteria that ranges from order 10^{12} (a baby's microbiome) to 10^{18} (a wastewater treatment plant) individual particles, and are physically complex. Since the models are computationally expensive, and due to computing constraints, the consideration of only a limited set of scenarios is often possible. A simplified approach to this problem is to use a statistical approximation of the simulation ensembles derived from the complex models at a fine scale which will help in reducing the computational burden. Our aim in this paper is to build a cheaper surrogate of an individual-based (IB) model simulation of microbial communities. The paper focuses on how to use an emulator as an effective tool for studying and incorporating microscale processes in a computationally efficient way into macroscale models. The main issue we address is a strategy for emulating high-level summaries from the IB model simulation data. We use a Gaussian process regression model for the emulation. Under cross-validation, the percentage of variance explained for the univariate emulator ranges from 83–99% and 87–99% for the multivariate emulators, and for both biofilms and floc. Our emulators show an approximately 220-fold increase in computational efficiency. The sensitivity analyses indicated that substrate nutrient concentration for nitrate, carbon, nitrite and oxygen as well as the maximum growth rate for heterotrophic bacteria are the most important parameters for the predictions. We observe that the performance of the single step emulator depends hugely on the initial conditions and sample size taken for the normal approximation. We believe that the development of an emulator for an IB model is of strategic importance for using microscale understanding to enable macroscale problem solving.

1. Introduction

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 provide the best possible representation of micro-scale responses. The challenge then becomes how we can transfer this small-scale information to the engineered 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 populations [37]. In other words, the properties of cells or particles at a micro level dictate the behaviour of a wastewater treatment plant at a macro scale.

* Corresponding author.
E-mail address: Oluwole.Oyebamiji@newcastle.ac.uk (O.K. Oyebamiji).

http://dx.doi.org/10.1016/j.jocs.2017.08.006

We know that there is a wide separation in the spatial and temporal dimensions at which biological and physical processes occur which complicates the complete understanding of the emergent behaviour of the system. The scale transition for modelling biofilms and flocs in this study ranges from micro- to meso- to macro-scales (although, we only consider micro-meso-scales in this study) (see Fig. 1) for details. This multiscale approach was used in this study for passing aggregate information from one level to the other. The complex nature of the transitions from cellular level (microscale) to a group of bacteria (floc/biofilm) at mesoscale introduces a scaling problem in addition to model complexity, thus making the simulation from the micro model a computationally expensive task. A robust strategy is required to handle this issue efficiently.

One useful approach for addressing this problem is via the use of statistical emulators, sometimes called metamodels. Emulation is a statistical technique for simplifying models that leads to reduced-form representations of complex models which are computationally much faster to run. Emulators offer rapid and relatively quick alternatives for projection of model outputs [41,42]. A further benefit of emulation is the provision of a measure of uncertainty associated with the projections.

There have been a significant number of research applications dealing with the statistical emulation of expensive computer models. This ranges from a univariate Gaussian process emulation to multi-output predictions [8]. Similarly, [35] developed a Bayesian framework for the uncertainty analysis for the distribution of unknown input. In particular, [35] used a univariate Gaussian process for emulating computationally expensive simulator outputs with uncertain inputs. [19] extended the univariate GP approach in [35] to a multivariate GP and combined this with a principal component analysis (PCA) for calibrating high dimensional outputs from a computationally demanding computer model against the field data from an experiment. The experimental data was used to constrain uncertainty in the calibration parameters. The PCA reduces the dimension of the problem and computation time required for obtaining posterior distributions from Bayesian inference.

Another application of this sort of modelling is to separate stochastic from deterministic variations, the procedure for handling stochastic noise in emulation was described in [17] and [6]. However, there is a limited amount of literature that treats the emulation of stochastic simulators. Earlier work of [26] performed ordinary kriging emulation of detrended and standardised response $\mathbf{y}'$ from stochastic outputs where the scale response was derived by repeating the simulation several times at each design point. This approach was extended by [4] where an independent GP emulator is developed for both the mean response and stochastic (noise) variance. A related approach was documented in [24] and [15] where an additional GP model was built to estimate the noise variance of the noise-free dataset.

On a different note, [58] described the behaviour of large linear dynamic models that used statistical principles of dynamic emulation. Their approach identifies a low-order model that approximates the behaviour of the high-order dynamic simulator that is much cheaper. [36] described a Bayesian method for quantification of uncertainty in complex computer models while [23] presented some notable examples where GP modelling applications have been implemented.

The aim of this paper is to describe how to use an emulator as an effective tool for incorporating microscale processes in a computationally efficient way into macroscale models. The focus is to train the dynamic emulator with micro-level simulation data from an individual-based (IB) model for the predictions of an aggregate of particles, of varying species, called floc and biofilms. Biofilms are the aggregated microbial communities attached to surfaces. Flocs are aggregated microbial communities suspended in water. Their characteristic size is around 500 μm. The morphological features depend on the growth conditions. For example, high nutrient conditions may promote a smooth surface while a rough surface structure is more likely to emerge at low nutrient concentration. We have modelled their biological and chemical functions as listed in the supporting document.

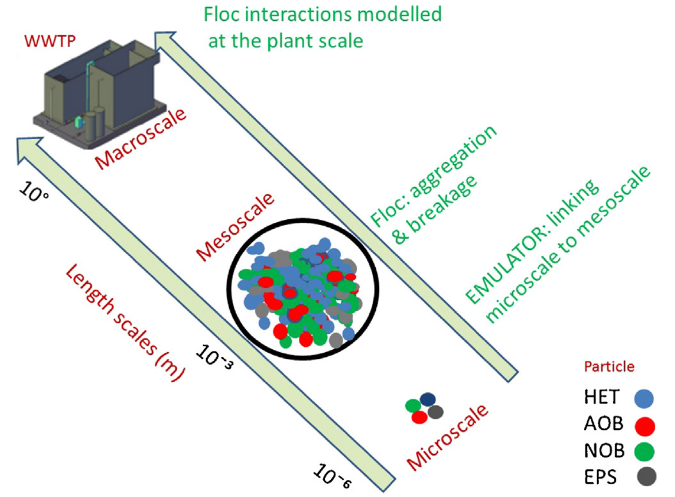
Floc interactions modelled at the plant scale
WWTP
Macroscale
Floc: aggregation & breakage
EMULATOR: linking microscale to mesoscale
Mesoscale
Length scales (m)
10°
10⁻³
10⁻⁶
Microscale
Particle
HET ●
AOB ●
NOB ●
EPS ●

Fig. 1. Schematic of different length scales for multiscale modelling of an activated sludge process based WWTP. The scale transition from a bacterium or cellular level (microscale – <micrometer size) to the floc and biofilm aggregates (mesoscale – millimetre size) to the macroscopic bulk WWTP operation as well as floc and biofilm interactions (macroscale – metre size). The emulator is linking the microscopic (bacterium/cell) to the mesoscopic (biofilm/floc) and, ultimately, to the macroscopic bulk operational parameters.

The flocs and biofilms are mixed with an adhesive material called extracellular polymeric substance (EPS). The EPS is a class of organic macromolecules such as polysaccharide, proteins, nucleic acids, lipids and other polymeric compounds which are found in the intracellular space of organic aggregates [57]. We do not model each component of EPS. In our microscale simulations, EPS particles represent the collection of different substances of EPS. The flocs and biofilms are often difficult to measure or quantify because of their irregular size and shape. For instance, a wide range of different “equivalent diameters” has been used to characterise the floc size; see [21] for further details. The floc plays a strategic role in understanding the processes involved in wastewater treatment plants.

In this study, we describe the procedure for emulating summary outputs from an IB model simulation of microbial organisms based on large-scale atomic/molecular massively parallel simulator (LAMMPS), a classical dynamical model for particle simulation [46]. The emulator constructed will be further used to transfer information to macro-level processes of wastewater treatment plants. [54] earlier reviewed some of the popular techniques for upscaling complex problems while [13] and [56] specifically focused their attention on how to use emulators for upscaling hydrological processes and land use management properties.

Due to the spatio-temporal nature of LAMMPS outputs, 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 flocs and biofilm 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 Fig. 2.

We use Gaussian process emulation (or kriging metamodels) where output data can be decomposed into a mixture of deterministic (non-random trend) and a residual random variation. In