A Bayesian approach to modelling the impact of hydrodynamic shear stress on biofilm deformation

Journal:	*Technometrics*
Manuscript ID	Draft
Manuscript Type:	Original Article
Keywords:	Biofilms, Dynamic linear models, Individual-based models, Markov chain Monte Carlo, Poisson regression, Surrogate models

A Bayesian approach to modelling the impact of hydrodynamic shear stress on biofilm deformation

Abstract

We investigate the feasibility of using a surrogate-based method to emulate the deformation and detachment behaviour of a biofilm in response to hydrodynamic shear stres. The influence of shear force and growth rate parameters on the patterns of growth, structure and resulting shape of microbial biofilms was examined. We develop a novel statistical modelling approach to this problem, using a combination of Bayesian Poisson regression and dynamic linear models for the emulation. We observe that the hydrodynamic shear force affects biofilm deformation in line with some literature. Sensitivity results also showed that the shear flow and yield coefficient for heterotrophic bacteria are the two principal mechanisms governing the bacteria detachment in this study. The sensitivity of the model parameters is temporally dynamic, emphasising the significance of conducting the sensitivity analysis across multiple time points. The surrogate models are shown to perform well, and produced ≈ 480 fold increase in computational efficiency. We conclude that a surrogate-based approach is effective, and resulting biofilm structure is determined primarily by a balance between bacteria growth and applied shear stress.

Keywords: Biofilms, Dynamic linear models, Individual-based models, Markov chain Monte Carlo, Poisson regression, Surrogate models.

1 Introduction

Water is crucial for life on earth and is valuable also for its supporting role in ecosystem function. Water that is safe for drinking is scarce partly due to an increase in wastewater. Biofilm technology is being deployed in the management and treatment of wastewater. A model is required that describes the individual processes in the wastewater treatment system. The simulation of microbial communities has important application in wastewater treatment studies. Wastewater treatment plants are open systems that depend on many species of bacteria to form a microbial community for the transformation of waste into biomass and other substances. According to Merkey et al. (2011), biofilms are regarded as the commonest form of bacteria on earth.

It has been established that the growth, structure and performance of bacteria biofilms are stongly affected by the hydrodynamic shear force. It is increasingly being recognised that hydrodynamic shear stress has a significant role to play on the deformation of biofilms and detachment of bacteria. There has been a large number of research projectss dealing with the assessment of the impact of hydrodynamic stress on biofilms deformation. Liu and Tay (2002) observed that steady state structures of biofilms are strongly affected by the hydrodynamic shear stress. The general understanding of the influence of bacteria detachment is documented in (Li et al., 2015; Bryers, 1988; Xavier et al., 2005).

The detachment process is an essential mechanism for removal of biomass from biofilm thereby controlling the biofilm key processes like growth, development and performance (Choi and Morgenroth, 2003; Rittmann et al., 1992; Liu and Tay, 2002). Moreover, Kommedal and Bakke (2003) and Bryers (1988) identified five different mechanisms of biomass detachment in biofilm while recently Li et al. (2015) focus on just only three out of these processes. The first category is the shear detachment which occurs as a result of flow fluid in the bacteria compartment, and an erosion detachment which is breakage of small particles from the surface of biofilm into bulk fluid. The third type is the nutrient-limited detachment which is associated with insufficient nutrient effects. However, Choi and Morgenroth (2003) and Picioreanu et al. (2001) limit their attentions to erosion (small-particle loss) and sloughing detachment of relatively large portions of the biofilm. In a similar vein, Picioreanu et al. (2001) noted that the biofilm simulation subjected to an erosion type of

detachment event has the potential for making the biofilm surface smoother while sloughing type detachment can cause an increase biofilm surface roughness. In other words, the detachment phenomenon is a significant determinant of the shape, composition and structure of emerging biofilm.

We know that biofilm growth simulation is computationally demanding, and most of the available studies on biomass detachment usually base their inference on a relatively small sample of simulation data. For instance, Li et al. (2015) consider only three parameter values for shear, nutrient-limited and erosion detachment coefficients in their experiments while the consideration of only six parameter values is used as detachment parameters in (Xavier et al., 2005). The limitation of these studies is their lack of sufficient data to make a rigorous validation for testing how shear forces influence bacteria detachment. Similarly, Xavier et al. (2005) and Li et al. (2015) use only the simple detachment rate and probabilities in their studies but fail to incorporate the knowledge of mechanical interactions among the particles in their simulations, while Picioreanu et al. (2000) and Kreft et al. (1998) completely neglect the effect of biomass deformation in their studies.

The simulation of the impacts of shear flow on the size and structure of biofilm is undertaken in this paper such that the shear force from the fluid flow on mature biofilm leads to deformation and eventual breakage as an emergent event. The approach is computationally demanding as noted in some literature eg Xavier et al. (2005) because of the partial differential equation (PDE) necessary to model the mechanical forces in the system.

To gain deeper insights into how the shear force affects the deformation of biofilm, we develop a novel surrogate-based technique, often referred to as a "statistical emulator" Oyebamiji et al. (2015) for predicting the shearing behaviour of model systems without having to rely on an expensive simulator. We are interested in strengthening the study of the essential role of shear stress breakage of microbial aggregates. We note that the crude parameter scans used in Li et al. (2015); Xavier et al. (2005) is not an ideal approach to fully understand the emerging properties of the microbial organism, and could be improved using a properly designed experiment. Our approach is to use advanced statistical techniques based on a large ensemble of simulation data to make a rigorously tested and validated assessment of the effect of shear force on microbial deformation. This approach will provide

new insights into how quantitative statistical techniques can be used to simplify and study this complex problem.

The simulation data we analysed in this paper are from an expensive dynamic model. There have been a large number of studies that have examined data from a dynamic simulator. For instance, Oyebamiji et al. (2017) emulated the characterized outputs of large linear dynamic models using statistical principles of dynamic emulation while Young and Ratto (2011) focus on a low-order dynamic model that approximates the response of the high-order dynamic simulator at a low computational cost. Oakley and O'Hagan (2004) described a Bayesian method for quantification of uncertainty in complex computer models while Kennedy and O'Hagan (2001) applied a Bayesian technique to calibrate computer models. Also, Shi et al. (2003), focusing more on heterogeneous data, used a hybrid Markov chain Monte-Carlo method.

Bayesian computations have extended and broadened the scope of statistical models that can be handled in practice due to the development of Markov chain Monte Carlo (MCMC). However, if the model errors have a Gaussian distribution and a given form of the prior distribution is assumed, then the posterior distributions of the model's parameters can sometimes be obtained analytically, without MCMC. A major benefit of using Bayesian regression is the provision of a measure of uncertainty in its analysis. The disadvantage is that, computationally, it can be very demanding. This is not a serious drawback for the problems we addressed in this paper.

The two outputs we consider to model in this paper are the expected number of detachment events (count data) and volume of detached clusters. The traditional approach for modelling event-count data is to use a Poisson model. In the earlier studies of Doss and Narasimhan (1994), a Bayesian Poisson regression model based on the Gibbs sampler was used while Chan and Vasconcelos (2009); Ma and Kockelman (2006) apply a similar Bayesian Poisson regression for modelling the crowd counting and injury count data respectively. We use an individual-based model simulation of microbial organisms incorporating a fluid flow that is based on LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator), a classical dynamical model for particle simulation. This simulator was enhanced to incorporate biological and physical processes to model bacterial growth, decay

and mechanical interactions among bacteria cells (Jayathilake et al., 2017).

We know from the available literature that the impact of hydrodynamic shearing force on the biofilm fragmentation has not been thoroughly studied using quantitative statistical techniques. The primary objective of this work is to investigate the effect of shearing force on the biofilm deformation and bacteria detachment using a surrogate-based method. Firstly, we assume that the biofilm fragmentation occurs as an event and we proceed by examining the extent to which shearing force impacts on the hazard of a bacteria detached from a parent biofilm.

Secondly, we quantify the relationship between the average number of shearing events per unit time and some covariates like biofilm height, mass and EPS composition, using a novel combination of Bayesian Poisson regression and dynamic linear models. We then predict the total volume of detached clusters per unit time as a continuous function of the predicted number of shearing events, shear rates and other covariates using a dynamic linear model. This modular approach will enable us to predict the distribution of detached clusters over time. We describe the models and simulation data utilised for the analysis in Section 2. In Section 3, we describe the Bayesian methods including the dynamic linear models and Poisson regression. Section 4 provides the results of the analysis including the sensitivity analysis. Section 5 presents the discussion and concluding comments.

2 Simulation model

2.1 Individual-based modelling of microbial communities

The present study models the biofilm that might be found in a wastewater treatment plant (WWTP) at the individual microbe level since pilot scale plants and laboratory scale experiments of wastewater treatment plants WWTP are expensive, cumbersome, non-invasive and often cannot provide information at the micro-scale, which is required for operational optimisation of WWTP. The mathematical models used for biological treatment can be mainly divided into two general classes according to the way the biomass is represented: continuous and discrete models. In the present work, a discrete IB Model is used. Biofilms are the aggregated microbial communities attached to surfaces. Their typical size is around

500 μm. Appendix B describes further details about the simulation model and Jayathilake et al. (2017) for details about their biological and chemical functions.

2.2 Simulation data

The IB model was run with a series of growth parameters and shearing forces over an extended period using a Latin hypercube design to generate biofilms of various size. The LHD technique provides good coverage of the input space with a relatively small number of design points (Santner et al., 2013). The parameters are varied within the range of $\pm 50\%$ of the standard values given in Table 1 for 120 training points and five replicates at each design point, due to the expense of this computer model. The parameters are $\mu_{m,HET}$ which is the maximum specific growth rate for HET, $K_{s,HET}$ is the substrate affinity for HET, Y_{HET} is the yield coefficient for HET growth, and γ is the hydrodynamic shear rate; see Table 1 below, other simulation parameters are held constant; see Table A.1 in the Appendix C.

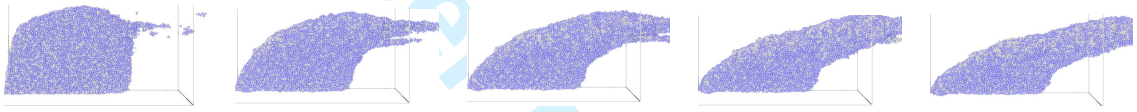
The design matrix is denoted as $\mathbf{X}_{120\times 4} = (\theta_p^i, p = 1, \ldots, 4; i = 1, \ldots, 120)$; where the subscript p represents the 4 model parameters that are varied in Table 1. The superscript i denotes the 120 different realisations. The simulator was run for 40000 s to grow the biofilm to a certain height without flow and then subjected the resulting biofilms to shear flow for an additional 200000 s where the biofilm detachment events occur. For each i, the output data is recorded at every 2000 s giving 120 time slices, $t = 1, \ldots, 120$. The number of particles and volume of detached clusters lost from the parent biofilm was recorded for different shearing forces. We compute the biofilm height using equation **??** below. Other morphological characteristics like biofilm mass, total number of particles and EPS composition are also calculated for predicting expected number of shearing events over time.

To compute average biofilm height ($metre$), the biofilms are partitioned into several smaller blocks. Each sub-block has dimension $d^{max} \times d^{max} \times d^{max}$. We compute the Euclidean distances between the center of each particle and the lattice blocks along the baseline (plane $z = 0$) to identify the occupied blocks. We, therefore, marked as "occupied" every block with one or more particle centers contained within it while the others are marked

Table 1: IB model parameters

Index	Parameters	Values	Units	References
1	$K_{s,HET}$	0.000035	$\mathrm{kg}m^{-3}$	Wanner and Reichert (1996)
2	$\mu_{m,HET}$	1.000000	h^{-1}	Schluter et al. (2015)
3	Y_{HET}	0.610000	$gCOD/gCOD$	Ni et al. (2009)
4	γ	0.250000	s^{-1}	chosen

as "vacant". The height $h_t(x, y)$ of the biofilm above each base block is defined as the maximum of the particle $z-$ values of the occupied blocks. The biofilm mean height at time t is then given as $\bar{h}(t) = \frac{1}{L_x L_y} \int_i \int_j h_t(x, y) dxdy$, where $L_x = 10^{-4}$, $L_y = 4 \times 10^{-5}$ are dimensions of simulation box.



$\gamma = 0.25$

$\gamma = 0.37$

Figure 1: Biofilm structures showing temporal evolution of effect of shear rate on biofilm deformation and bacteria detachment for different shear rates for the period the flow is applied. Leftplot: γ=0.25 and rightplot: γ=0.37 for 10,000, 60,000, 110,000, 160,000 and 200,000 seconds respectively.

2.3 The scope of the problem

The problem we are addressing in this paper is the simulation of biofilm under shear stress, where the fluid force of appropriate magnitude flow on individual biofilms leads to the weakening of biofilms. This can result in eventual deformation and bacteria loss from the surface. Our focus in this paper is to be able to treat each shear phenomenon as an

event, and we are interested in testing the feasibility of using a surrogate-based model for predicting expected number of shear events and size of detached clusters. The knowledge about the emerging composition and structure of biofilms subjected to shear flow is useful to improve the performance and stability of wastewater reactors. Figure 1 represents a typical simulation of biofilm under two different shear rates at $\gamma = 0.25s^{-1}$ and $0.37s^{-1}$ respectively, and we can see diverging temporal behaviours and structures when the shear flow is applied on a mature biofilm, the influence of cell detachment from the surface begins to emerge. The detachment phenomenon occurs when cohesive failures happen due to hydrodynamic shear force. At 10,000 s, the frequency of detachment events is higher for $0.37s^{-1}$ rate than for $0.25s^{-1}$. The detached clusters are moving out in the opposite direction to the bulk flow (shear flow from left-hand) as expected. The two shear rates give rise to different detachment patterns. We also see a gradual decreasing and flattening of the biofilms over time. In particular, the elongated filamentous cell clusters (streamers and clumps of cells) at a later time is obvious.

The density plots for the two outputs we are considering are given in Figure 2 (column 1). This enables us to look closely at the distribution of the two outputs to be analysed. It is apparent that each density plot has relatively different distribution under different shear rates and are highly skewed and nonnormal. The number of shear events is modelled using a Poisson distributed random variable because they are count data. We modelled the expected number of shear events function using a Poisson regression having a log-mean parameter that is a quadratic function of the explanatory variables. The density of detached clusters follows a log-normal distribution suggesting the use of a log-transformation.

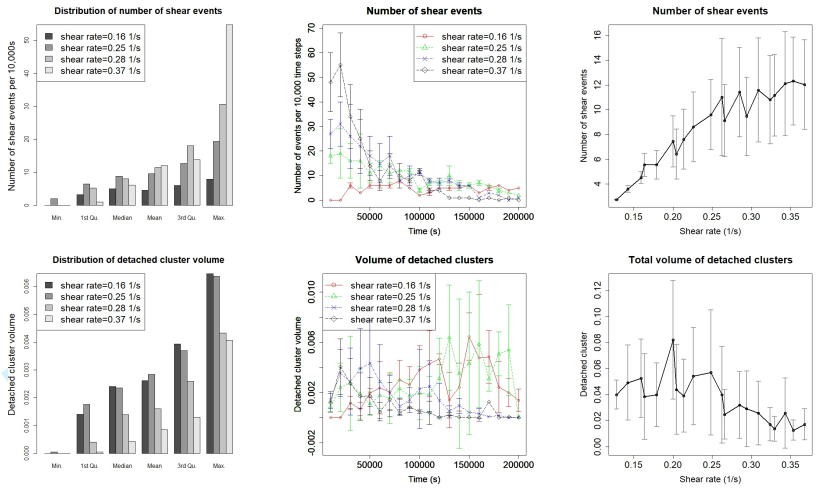


Figure 2: Density and time series plots for expected number of shear events and volume of detached clusters for different shear rates. The error bars show ± 1 standard deviation calculated from five replicates. Note: volumes are normalized by their initial biofilm volume.

The corresponding time-series plots are displayed in Figure 2 (middle column) for different shear rates. There is a reduction in the number of shear events after reaching the maximum values for shear rates $0.37s^{-1}$, $0.28s^{-1}$ and $0.25s^{-1}$ respectively. A phenomenon which can be attributed to a more frequent detachment of smaller cells from biofilm surface at the beginning of experiment $< 30,000s$, often called erosion detachment. At a low shear rate of $0.16s^{-1}$, the number of shear events is relatively constant over time. We can conclude that the number of shear events under different shear rates has slightly varying patterns.

Even though the number of events is increasing between $< 30,000s$, the corresponding values of detached clusters are very low Figure 2 (middle column) because the larger aggregates occasionally detached. Similar to the number of events, the detached volume has a different trend under different shear rates which increase slowly over the time. For instance, at shear rates of $0.37s^{-1}$ and $0.28s^{-1}$ there is a gradual increment in the volume of detached clusters due to top surface cells sheared off quite early after which there is a reduction and relatively constant detachment. This trend could be attributed to the particle at the top surface growing to a larger size because of better access to nutrients from the

bulk medium. The effect of stochastic variation is pronounced because of large standard deviations around each output.

Figure 2 (column 3) shows the expected number of shear events averaged over all time and total volume of the detached cluster over all time. The growth in the number of events at higher shear rates agrees with the works of Stoodley et al. (2002) and Walter et al. (2013) who observed that doubling the shear stress frequency from 21.8 to 43.6mPa resulted in a multiple fold increase in detachment rate for both erosion and sloughing detachments.

It is apparent that while shear events increase linearly with an increase in shear rates, the total volume of detached clusters also increases nonlinearly until a threshold value of around $0.20s^{-1}$ before a decline. This suggests that a critical shear stress exists, above which the volume of detached clusters is negatively correlated with shear stress. For shear rate $> 0.20s^{-1}$, it can be inferred that the total volume of detached clusters is gradually decreasing as the shear rates increases. We do not observe as large a reduction as reported in some literature, eg Walter et al. (2013) recorded that there will be a decreasing in mean size of eroded clumps as shear rate increases.

Other summary outputs for the simulation data which are used as explanatory variables in the modelling are displayed in Figure 3. There is a general decreasing trend over the time, those with higher shear rates (eg $0.37s^{-1}$) decline more rapidly than those with lower shear rate (eg $0.16s^{-1}$). At shear rate of $0.37s^{-1}$, the biofilm height approach a minimum threshold value of $2.0e^{-05}$. The EPS composition plots are relatively constant in the first 50,000 s before they gradually decrease.

The EPS composition denotes the number of EPS particles in the simulation. EPS is a gel-like material that keeps bacteria together in the biofilms. Therefore high EPS composition rate will favour attachment of bacteria. EPS composition declines at a slower rate than biofilm height and mass. The biofilm mass and total number of particles have a relatively similar trend which declines rapidly as expected because the clusters are being continuously detached from the surface. Paul et al. (2012) also observed an exponential and asymptotic decrease of the biofilm thickness and mass when exposed to high shear stress. The effect of stochastic variation is considerably larger for the shear rate of $0.37s^{-1}$ and increases with time.

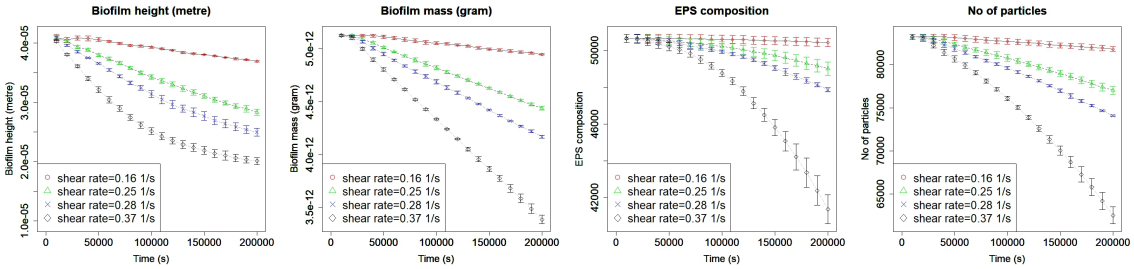


Figure 3: Dynamics of characterized outputs under different shear rates for biofilm height, mass, EPS composition and total number of particles at every 10,000s respectively. The error bars show ±1 standard deviation calculated from five replicates. These summary outputs are used as explanatory variables for predcting the expected number of shear events using Poisson regression. Note: EPS composition is the number of EPS particles.

3 Methods

The Bayesian framework involves combining observed data with a likelihood, and a prior distribution on unknown parameters to obtain the posterior distribution of the parameter given the data. It is often difficult to derive the posterior distribution in a closed form in most applications, and this often results in the use of Markov chain Monte Carlo simulation methods as an alternative. Markov chain Monte Carlo has been widely used in many complex applications for parameter estimation. Geyer (1991) describes MCMC as a general tool for simulation of complex stochastic processes useful for making statistical inference. MCMC produces a sequence of random variables which can be used to approximate the true posterior distribution.

Gibbs sampling, for instance, is based on the principle that the knowledge of the conditional distributions is often sufficient to determine a joint distribution (Casella and George, 1992). On the other hand, the Metropolis-Hastings algorithm allows one to make random draws from such a non-standard posterior distribution using proposal distributions. Moreover, Casella and George (1992) gives a detailed explanation of the theory behind Gibbs sampling while Chib and Greenberg (1995) highlights the theoretical background behind the Metropolis-Hastings algorithm.

3.1 Dynamic linear models

Let us consider bacterial biofilms transported in a fluid flow with the following morphological characteristics measured on them. The parent biofilm size ($\mathbf{Z}_t$), the volume of the detached cluster ($\mathbf{Y}_t$), a hydrodynamic shear stress (γ), biofilm mean height (H) and time to a detachment event (t). We are interested in developing a surrogate model for predicting the volume of the detached cluster. We propose to use a dynamic linear model for modelling the log-transformed volume of the detached cluster because of the time series nature of our data. The dynamic linear model is an extension of standard linear regression models with incorporation of time-varying regression coefficients (Petris et al., 2009). Therefore, a dynamic estimation of our model parameters will enable us to have a better understanding of the complex problem we are addressing. A dynamic model is usually given as a pair of equations such that for $t > 0$, we have

$$\begin{cases} \mathbf{Y}_t = \mathbf{F}_t\boldsymbol{\beta}_t + v_t, \qquad v_t \sim N(0, \mathbf{V}_t), \qquad (a) \\ \boldsymbol{\beta}_t = \mathbf{G}_t\boldsymbol{\beta}_{t-1} + w_t, \qquad w_t \sim N_2(0, \mathbf{W}_t), \qquad (b) \end{cases} \tag{1}$$

where $\mathbf{F}_t$ is an $m \times p$ dynamic regression matrix (explanatory variables) and $\mathbf{G}_t$ is an $p \times p$ state evolution matrix. v_t and w_t are two independent Gaussian random vectors with mean 0 and variances $\mathbf{V}_t$ and $\mathbf{W}_t$, respectively. $\mathbf{W}_t$ is the evolution variance matrix for $\boldsymbol{\beta}_t$ and $\mathbf{V}_t$ is the observation variance matrix while $\boldsymbol{\beta}_t$ is an $m \times 1$ vector of regression parameters. Equations 1(a) and 1(b) are usually called observation and state equations, respectively. Let

$$\boldsymbol{\beta}_0 \sim N_p(m_0, C_0), \tag{2}$$

Combining these two equations above, we can easily infer that $\mathbf{Y}_t|\boldsymbol{\beta}_t \sim N(\mathbf{F}_t\boldsymbol{\beta}_t, \mathbf{V}_t)$ and $\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1} \sim N(\mathbf{G}_t\boldsymbol{\beta}_{t-1}, \mathbf{W}_t)$. The two equations above can be applied together for making one-step ahead predictions of our output of interest, ie volume of the detached particle. It is evident that we require the state estimation of parameters in making such a prediction or forecasting. For instance, to predict observations $\mathbf{Y}_{t+1}$ based on data $\mathbf{Y}_{1:t}$, we can first estimate $\boldsymbol{\beta}_{t+1}$ of the state vector and use the estimated values for making predictions $\mathbf{Y}_{t+1}$. In other words, we can obtain the one-step ahead observation predictive density $\pi(\mathbf{y}_{t+1}|\mathbf{y}_{1:t})$, from one-step ahead state predictive density $\pi(\boldsymbol{\beta}_{t+1}|\mathbf{y}_{1:t})$. See Appendix A for

further details and summary of Gibbs sampling algorithm.

3.2 Poisson regression

The Poisson model is employed for modelling event count data, eg the number of shearing events or organisms in an experiment. One of the key properties of count data is that they must be non-negative integers. A Poisson regression model expresses the logarithm of a response or dependent variable (count or rate data) as a linear function of a set of predictor variables. Such a log-linear Poisson model is often adopted to describe a time series of counts or rates. The model assumes the outcome $\mathbf{y}_k$ to be Poisson with mean λ_k, so that for a univariate predictor variable x_i (eg biofilm height), the model is

$$\mathbf{y}_k \sim Poisson(\lambda_k), \tag{3}$$

$$\lambda_k(x) = \exp\left(\sum_{i=1}^{p} \mathbf{x}_{ik}\boldsymbol{\beta}_i\right), \tag{4}$$

where $k = 1, \ldots, n$, $\lambda_k(x)$ is the log-mean function and $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)$ is the vector of unknown parameters and $\mathbf{x}_{ik}$ are the explanatory variables. A discrete random variable $\mathbf{Y}$ with the probability mass function of $\mathbf{Y}$ given as

$$f(k;\lambda) = \Pr(\mathbf{Y} = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \tag{5}$$

with parameter $\lambda > 0$, for $k = 0, 1, 2, \ldots$ is regarded as a Poisson distribution. The mean and variance of a Poisson-distributed random variable are both equal to λ. Parameters $\mathbf{B}$ are unknown and need to be estimated. Having seen earlier in Figure 3 that the number of shear events have complex patterns, we will need to use a sophisticated technique like MCMC to efficiently estimate our parameters. Adopting a fully Bayesian approach, the Poisson likelihood function is given by

$$L(B) = \frac{1}{\prod_{k=1}^{n} \mathbf{y}_i!} \exp\left[-\sum_{k=1}^{n} \exp\left(\sum_{i=1}^{p} \mathbf{x}_{ik}\boldsymbol{\beta}_i\right) + \sum_{i=1}^{p} \boldsymbol{\beta}_i \sum_{k=1}^{n} \mathbf{x}_{ik}\mathbf{y}_k\right]. \tag{6}$$

Let the prior π for parameter $\boldsymbol{\beta}_i$ be given as an independent normal distribution with mean m_i and variance v_i, ie $\boldsymbol{\beta}_i \sim N(m_i, v_i)$. Under this procedure we will have a joint density of

B given as

$$\pi(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p) = \prod_{i=1}^{p} \frac{1}{\sqrt{(2\pi v_i)}} \exp\left(-\frac{1}{2m_i}(\boldsymbol{\beta}_i - m_i)^2 \right). \quad (7)$$

Using Bayes theorem, the posterior distribution is proportional to the product of the likelihood function and the joint prior of all parameters. Here, the posterior distribution of $\boldsymbol{\beta}_i$ conditioning on the given data can be obtained by combining equations 6 and 7 above as

$$\pi_{\mathbf{y}}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p) \propto \exp\left[-\sum_{i=1}^{p} \frac{1}{2m_i}\boldsymbol{\beta}_i^2 \sum_{i=1}^{p} \alpha_i \boldsymbol{\beta}_i - \sum_{k=1}^{n} \exp\left(\sum_{i=1}^{p} \mathbf{x}_{ik}\boldsymbol{\beta}_i \right) \right], \quad (8)$$

where $\alpha_i = \frac{m_i}{vi} + \sum_{k=1}^{n} \mathbf{x}_{ik}\mathbf{y}_k$. We note that there is no conjugacy between the Poisson likelihood and normal prior distribution which makes exact inference analytically infeasible (Chan and Vasconcelos, 2009). We use Poisson regression to model expected number of shear events per unit time and apply Bayesian MCMC to estimate the parameters of the model by assigning a prior distribution on the regression parameters. The expected value for the Poisson model can be derived from the posterior draws of **B** based on the MCMC iterations and is given by

$$E(\mathbf{y}|\mathbf{X}) = \lambda_k = \exp(\mathbf{x}_k\mathbf{B}), \quad (9)$$

while the predictions are draws from the Poisson distribution with parameter λ_k (Martin et al., 2005).

4 Results

4.1 Procedure for modelling outputs

We use data from the LAMMPS model simulation output. We consider two different simulation datasets in this paper. The first dataset is the expected number of shearing events per unit time. The second dataset is the volume of detached biofilm clusters per unit time. The input variables to the simulator are the four parameters listed in Table 1. They are $K_{s,HET}$, $\mu_{m,HET}$, Y_{HET} and shear rate γ. Auxilliary variables of total number of particles, EPS composition, biofilm height and mass (Figure 3) are summary statistics computed and used for predicting the expected number of events.

Here, we present the results of our analysis. We based our analysis on the last 200,000 s, corresponding only to the period when the shear flow was applied. We chose to further reduce the dimension of our data by averaging at every 10,000 s which made handling and processing of the data much easier. In other words, our outputs of interest are given as the number of events per 10,000s and detached volume per 10,000s. We have 120 simulations with five replicates for each of them. Our data are averaged and taken to be deterministic. We subdivided our data into two groups. We use 100 data points as a training dataset and use the remaining 20 data points as the test data to verify the performance of our surrogate models.

4.2 Bayesian Poisson results

To proceed with our analyses, we first fitted a Bayesian Poisson regression to the number of shear events as a quadratic function of time, initial biofilm size, EPS composition, biofilm height and mass. We used a quadratic model of the form

$$\sum_{i=0}^{p}\sum_{k=0}^{p}\boldsymbol{\beta}_i\mathbf{x}_i\mathbf{x}_k, \tag{10}$$

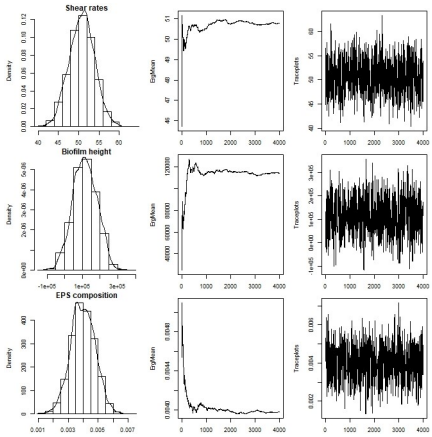
where each $\mathbf{x}_i$ represents an explanatory variable ($p = 5$). In the Bayesian context, the data are augmented by a prior distribution. This prior information given over the parameters is then combined with the likelihood function using Bayes theorem to give the posterior distribution for the parameters.

We used MCMC to estimate the unknown $\boldsymbol{\beta}$ parameters. Our prior for each variable is taken as a normal distribution. To initialize our algorithm we used the maximum likelihood estimates of $\boldsymbol{\beta}$ as the starting values.

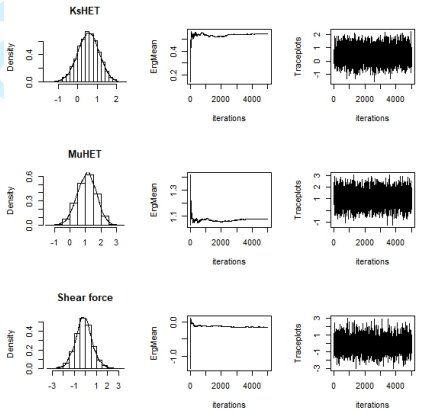
We generate samples from the posterior distribution of Poisson regression given in equation 8 using a Metropolis algorithm. We run the algorithm for 600,000 MCMC iterations, with a burn-in of 1000 samples discarded. We kept every 100 iterations (thin= 120) to reduce the autocorrelation in the saved MCMC samples. Figure 4(top) is the diagnostic plots for examination of MCMC samples for convergence. We provide the estimates for the shear rate, biofilm height and EPS composition. The posterior density plots provide information about the shape of posterior distributions of parameters. The ergodic means

(middle column) from MCMC samples are relatively stable after 1000 simulations. We can conclude that the convergence of the MCMC has been reached to estimate the posterior means of unknown parameters. The plots for other parameters are not shown in this paper but also indicated that those parameters have converged.

Now, we test the performance of fitted Poisson regression models on the 20 left-out observations. Figure 5 is the cross-validation plot that compares the expected number of shear events under four different shear rates from simulation and the predictive model. Each output is plotted against time. As we earlier observed in Figure 3, there is a moderate linear increase in the number of shear events until a threshold value is reached and then gradually declines afterwards. The patterns are consistent with different shear rates. Overall, the four results in this Figure 5 are well predicted as most of the simulated values lie within the 95% probability intervals. We assess the overall performance of Poisson model by computing the root mean squared value (RMSE= 2.885) and percentage of variance explained for the left out data points ($\rho = 89.6\%$).



(a) Poisson model



(b) DLM model

Figure 4: (a): Diagnostic plots showing the convergence of some of the estimated θ parameters of Bayesian Poisson model for expected number of events. The first column shows the posterior density of the state variances. The middle column is the running ergodic means of MCMC samples. The third colum is the trace plots for the MCMC samples. The plots (b) showing the convergence of the three randomly chosen regression parameters θ of the Bayesian dynamic linear model at time 10000s.

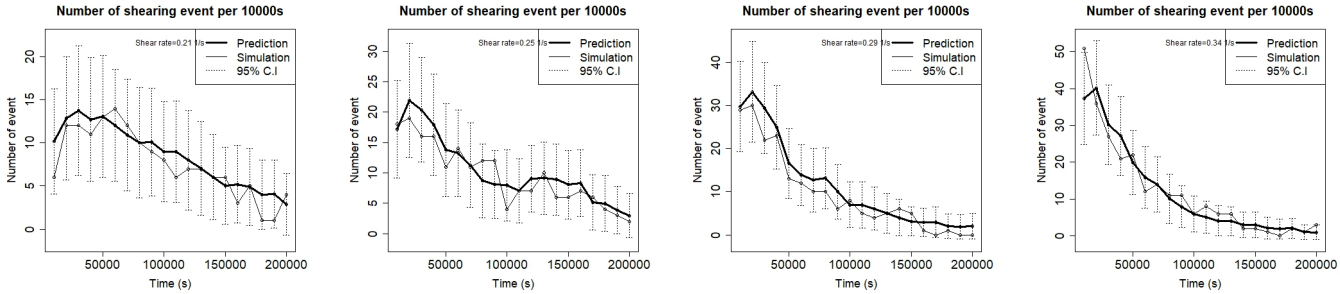


Figure 5: Comparison between the simulation and prediction for expected number of shearing events per 10000s for different shear forces

4.3 Dynamic linear model results

We next fit the dynamic linear model to the volume of detached clusters using equation 1 where $m = 100$, $p = 5$ and $k = p$. The predicted number of shear events from the Poisson model are used as an additional explanatory variable. We standardize our input data to range over $[0, 1]$. This transformation will eliminate the effect of different measurement units and will enable us to get better parameter estimates. We assume that matrices of unknown parameters are time-invariant, $\mathbf{G}_t = \mathbf{G}$, $\mathbf{V}_t = \mathbf{V}$ and $\mathbf{W}_t = \mathbf{W}$. Suppose further that the matrix of the explanatory variable is also time-invariant, then we have $\mathbf{F}_t = \mathbf{F}$. We initialized the Markov chain sampler with the following prior hyperparameters; $\psi_0 = 0$; $\tau_0 = 1$; $\alpha_{\mathbf{y},i} = 4$; $b_{\mathbf{y},i} = 0.01$; $\alpha_{\boldsymbol{\beta},j} = 1000$; $b_{\boldsymbol{\beta},j} = 10$. We run the DLM algorithm for 1,000,000 MCMC iterations, keeping every 100 iterations (thin=100) in order to reduce the autocorrelation in the saved MCMC samples where a burn of 5000 samples is removed before making the diagnostic plots.

Figure 4 (bottom panel) shows the diagnostic plots obtained from the MCMC outputs for the three randomly selected regression parameters. The posterior density plots provide information about the shape of posterior distributions of parameters. The ergodic means (middle column) from MCMC samples are relatively stable after 1000 iterations. We can conclude that the convergence of the MCMC has been reached to estimate the posterior means. Diagnostic plots for three randomly selected values from the state and observation variance ($\mathbf{V}$ and $\mathbf{W}$) parameters are displayed in Figures (A.1 and A.2) of Appendix C. These plots are similar in pattern to Figure 4 in term of convergence while the diagnostic

plots for evolution matrix G are not shown but also indicated convergence.

Now, we test the performance of fitted models on the left-out observations. Figure A.3 of Appendix C is the cross-validation plot that compares the detached volume under four different shear rates for the simulator and emulator predictions. The plot for each shear rate has a slightly different pattern. Similar to what we earlier saw in Figure 2 (middle panel), the detached volume grows linearly over time for shear rates 0.15 and 0.21 s^{-1} and a rapid increase followed by a moderately decreasing trend for 0.25 and 0.32 s^{-1} respectively. There is a consistency in the pattern observed for the four selected shear rates. Overall, the simulated output values and that of predictions are relatively close. The degree of closeness reflects the accuracy of our DLM model. The uncertainty levels are a little bit higher compared to the number of shear events in Figure 3. The large confidence bands could be attributed to the significant noise in the data.

4.4 Sensitivity analysis

To further understand the dynamics of the system we are modelling, the relative contribution of each variable to the total output variance is explored. We perform dynamic sensitivity analysis because of time-dependent nature of our data. We examine how sensitive the log-transformed volume of detached clusters are to changes in parameters over time. We use the Sobol method which calculates indices by variance decomposition. We compute the first order and total indices. Suppose our model is represented by $\mathbf{y}_t = f(\mathbf{x}_{1,t}, \ldots, \mathbf{x}_{p,t})$. The first order index is given as $S_{i,t} = \frac{Var[E(\mathbf{y}|x_{i,t})]}{Var(\mathbf{y}_t)}$, where $Var[E(\mathbf{y}|x_{i,t})]$ is partial variance or the main effect of variable x_i, and Var($\mathbf{y}$) is the total variance of the response $\mathbf{y}$ (Saltelli, 2002).

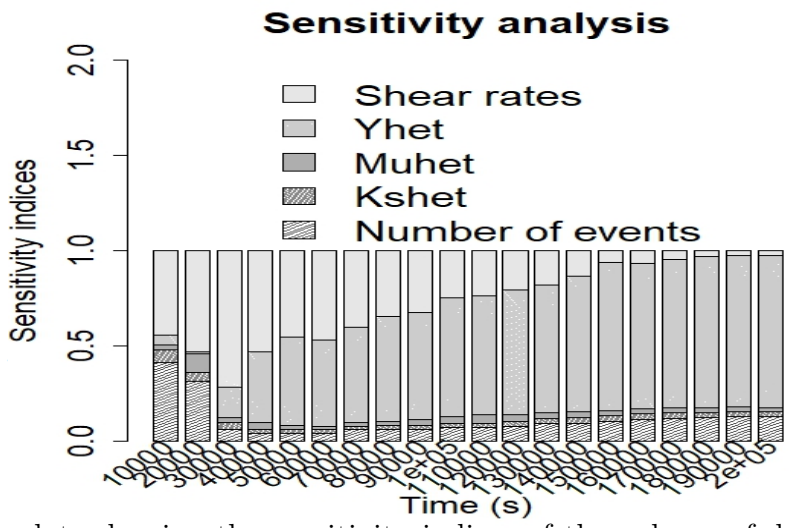


Figure 6: Barplots showing the sensitivity indices of the volume of detached clusters for the five variables over time.

We sampled 5,000 observations from a uniform distribution for each of the five input variables. The relative importance of each parameter is shown in Figure 6. We observed that detached cluster volume is mostly sensitive to yield coefficient for heterotrophic bacteria Y_{HET} growth and shear rate γ. At earlier time points, the sensitivity of shear rate is high and gradually becomes less sensitive at the later time. On the contrary, the sensitivity of Y_{HET} grows over time. Overall, the Y_{HET} and shear stress are the two principal determinants of the volume of detached clusters.

There are no significant changes in the behaviour of the number of shearing events over time (except at the first two time points), an indication that the volume of the detached cluster does not react greatly to a change in this parameter value. $\mu_{m,HET}$ and $K_{s,HET}$ have very low indices. It is obvious that sensitivity of the model parameters is temporally dynamic, emphasising the significance of conducting the sensitivity analysis across multiple time points.

5 Discussion and conclusion

There is a significant change in the morphology and dynamics of biofilm formation when a shear flow is applied on a mature biofilm as seen in Figure 1. Also, it is obvious that shear

force affects the biofilm structure in line with Liu and Tay (2002) observations. Moreover, at higher shear rates, a more dense and stable biofilm is likely to be produced because of stronger adherence from EPS matrix than those subjected to lower shear forces.

The role of interactive effects of shear force and other factors like pH and temperature on the biofilm fragmentation should be explored. We also remark that biofilms belong to viscoelastic materials and this property determines to some extent the deformation behaviour of biofilm growing under shear flow. However, the present work does not investigate the influence of this viscoelastic property of biofilms but we would like to implement this in the future work.

In this study, a shear flow is applied to a pre-grown biofilm of certain height to explore the detachment event. It is also possible to simultaneously model both the growth (attachment) and detachment, but only the dominant process will be explicitly modelled as noted in Wanner and Reichert (1996). This implies that if the detached velocity is greater the attached velocity only the net detachment will be modelled and there will be no particle attachment.

In summary, the influence of hydrodynamic shear force on biofilm fragmentation has been examined. We have developed a novel surrogate-based model for quantifying the effect of shear stress on the volume of detached clusters and number of shear events. This paper provides new insights on how advanced statistical techniques can be used to simplify and study biofilm deformation and bacteria detachment. We note that it is essential to develop a cheaper predictive model of biofilm deformation and bacteria detachment in response to mechanical forces and growth parameters because the knowledge can advance the performance and operational stability of wastewater reactors.

The biofilm simulation was initialized and grown for 40000 s without flow then subjected to shear flow to reduce the biofilms size because of the predominance of the breakup process. This results in biofilm of smaller size than the original size due to the shearing event. The volume of biofilm that gets sheared-off and the number of shear event over time are recorded for different shear rates. This study examines the extent to which the shear force affect the number of shear events and volume of detached clusters using a cheaper surrogate model. The joint impact of shear stress and other covariates are examined on biofilm of different

sizes. We assume that each occurrence of shearing can be modelled in terms of an event.

We used a 10,000s averaging as a strategy to condense the time series data. It will be interesting to assess the effect of this averaging on our predictive models. In our analyses, we have used normal and gamma priors because they are flexible and widely employed in various applications for modelling with Bayesian MCMC. The limitation with the MCMC algorithm is that the computational cost of a large parameter space is high. We compute the average number of shear events and volume of detached clusters that occur over time. We observe that the number of shear events increases until maximum values after which there is a gradual reduction. We used a Bayesian Poisson log-linear model to relate the expected number of shear events to characterize output summaries from the simulation.

The sensitivity analysis indicated that the Y_{HET} and shear stress γ are the two primary variables for predicting the volume of detached clusters and are less affected by the expected number of shear events. We can conclude that the growth, structure and performance of bacteria biofilms are highly related to the hydrodynamic shear force.

The IB model simulation implemented within LAMMPS is computationally expensive, and our surrogate models are much faster to run than the simulator. Under different parameter combinations, it takes an average of between 8-11 hours to simulate both the growth and detachment patterns for about 3 days at 2000s timestep on a Linux cluster machine. Apart from the computational time required to estimate the necessary parameters, the emulator produces the required outputs within $\approx 60s$. This approximately 480-fold increase in computational efficiency is particularly useful as a computational tool for the simulation and analysis of multiscale biological systems. This novel combination of advanced statistical techniques for modelling biofilm detachment behaviour using a surrogate-based approach is capable of greatly reducing the computational cost of modelling across large spatial and temporal scales. This study provides a significant step towards improving the performance, robustness and stability of biofilm-based wastewater treatment plant.

Supplementary Materials

Additional details: All appendices mentioned in this article (.pdf file)

R-code and dataset: The volume of detached clusters and number of shear events datasets used in Section 4 and R codes to fit the DLM, Bayesian Poisson regression models and sensitivity analysis (.zip file)

Acknowledgements

References

Bryers, J. (1988). Modeling biofilm accumulation. *Physiological models in microbiology*, 2:109–144.

Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174.

Chan, A. B. and Vasconcelos, N. (2009). Bayesian poisson regression for crowd counting. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 545–551. IEEE.

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335.

Choi, Y. and Morgenroth, E. (2003). Monitoring biofilm detachment under dynamic changes in shear stress using laser-based particle size analysis and mass fractionation. *Water Science and Technology*, 47(5):69–76.

Doss, H. and Narasimhan, B. (1994). Bayesian poisson regression using the gibbs sampler: Sensitivity analysis through dynamic graphics. Technical report, Technical report, Citeseer.

Geyer, C. J. (1991). Markov chain monte carlo maximum likelihood. pages 156–163.

Jayathilake, P. G., Gupta, P., Li, B., Madsen, C., Oyebamiji, O., González-Cabaleiro, R., Rushton, S., Bridgens, B., Swailes, D., Allen, B., et al. (2017). A mechanistic individual-based model of microbial communities. *PloS one*, 12(8):e0181965.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464.

Kommedal, R. and Bakke, R. (2003). Modeling pseudomonas aeruginosa biofilm detachment. page 3.

Kreft, J.-U., Booth, G., and Wimpenny, J. W. (1998). Bacsim, a simulator for individual-based modelling of bacterial colony growth. *Microbiology*, 144(12):3275–3287.

Li, C., Zhang, Y., and Yehuda, C. (2015). Individual based modeling of pseudomonas aeruginosa biofilm with three detachment mechanisms. *RSC Advances*, 5(96):79001–79010.

Liu, Y. and Tay, J.-H. (2002). The essential role of hydrodynamic shear force in the formation of biofilm and granular sludge. *Water Research*, 36(7):1653–1665.

Ma, J. and Kockelman, K. (2006). Bayesian multivariate poisson regression for models of injury count, by severity. *Transportation Research Record: Journal of the Transportation Research Board*, (1950):24–34.

Martin, A. D., Quinn, K. M., and Park, J. H. (2005). Markov chain monte carlo (mcmc) package. *http://mcmcpack. wustl. edu*.

Merkey, B. V., Lardon, L. A., Seoane, J. M., Kreft, J.-U., and Smets, B. F. (2011). Growth dependence of conjugation explains limited plasmid invasion in biofilms: an individual-based modelling study. *Environmental microbiology*, 13(9):2435–2452.

Ni, B.-J., Fang, F., Xie, W.-M., Sun, M., Sheng, G.-P., Li, W.-H., and Yu, H.-Q. (2009). Characterization of extracellular polymeric substances produced by mixed microorganisms in activated sludge with gel-permeating chromatography, excitation–emission matrix fluorescence spectroscopy measurement and kinetic modeling. *Water Research*, 43(5):1350–1358.

Oakley, J. E. and O'Hagan, A. (2004). Probabilistic sensitivity analysis of complex models: a bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):751–769.

Oyebamiji, O., Wilkinson, D., Jayathilake, P., Curtis, T., Rushton, S., Li, B., and Gupta, P. (2017). Gaussian process emulation of an individual-based model simulation of microbial communities. *Journal of Computational Science*, 22:69–84.

Oyebamiji, O. K., Edwards, N. R., Holden, P. B., Garthwaite, P. H., Schaphoff, S., and Gerten, D. (2015). Emulating global climate change impacts on crop yields. *Statistical Modelling*, 15(6):499–525.

Paul, E., Ochoa, J. C., Pechaud, Y., Liu, Y., and Liné, A. (2012). Effect of shear stress and growth conditions on detachment and physical properties of biofilms. *Water Research*, 46(17):5499–5508.

Petris, G., Petrone, S., and Campagnoli, P. (2009). Dynamic linear models. In *Dynamic Linear Models with R*, pages 31–84. Springer.

Picioreanu, C., Van Loosdrecht, M. C., Heijnen, J. J., et al. (2000). Effect of diffusive and convective substrate transport on biofilm structure formation: a two-dimensional modeling study. *Biotechnology and bioengineering*, 69(5):504–515.

Picioreanu, C., Van Loosdrecht, M. C., Heijnen, J. J., et al. (2001). Two-dimensional model of biofilm detachment caused by internal stress from liquid flow. *Biotechnology & Bioengineering*, 72(2):205–218.

Rittmann, B., Trinet, F., Amar, D., and Chang, H. (1992). Measurement of the activity of a biofilm: Effects of surface loading and detachment on a three-phase, liquid-fluidized-bed reactor. *Water Science and Technology*, 26(3-4):585–594.

Saltelli, A. (2002). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, 145(2):280–297.

Santner, T. J., Williams, B. J., and Notz, W. I. (2013). *The design and analysis of computer experiments*. Springer Science & Business Media.

Schluter, J., Nadell, C. D., Bassler, B. L., and Foster, K. R. (2015). Adhesion as a weapon in microbial competition. *The ISME journal*, 9(1):139.

Shi, J. Q., Murray-Smith, R., and Titterington, D. (2003). Bayesian regression and classification using mixtures of gaussian processes. *International Journal of Adaptive Control and Signal Processing*, 17(2):149–161.

Stoodley, P., Cargo, R., Rupp, C., Wilson, S., and Klapper, I. (2002). Biofilm material properties as related to shear-induced deformation and detachment phenomena. *Journal of Industrial Microbiology and Biotechnology*, 29(6):361–367.

Walter, M., Safari, A., Ivankovic, A., and Casey, E. (2013). Detachment characteristics of a mixed culture biofilm using particle size analysis. *Chemical engineering journal*, 228:1140–1147.

Wanner, O. and Reichert, P. (1996). Mathematical modeling of mixed-culture biofilms. *Biotechnology and bioengineering*, 49(2):172–184.

Xavier, J. d. B., Picioreanu, C., and van Loosdrecht, M. (2005). A general description of detachment for multidimensional modelling of biofilms. *Biotechnology and bioengineering*, 91(6):651–669.

Young, P. C. and Ratto, M. (2011). Statistical emulation of large linear dynamic models. *Technometrics*, 53(1):29–43.

A Bayesian approach to modelling the impact of hydrodynamic shear stress on biofilm deformation

This supplementary documents include further details about the dynamic linear model in the original article. Appendix A presents the theoretical details of dynamic linear model and summary of Gibbs sampling algorithm. Appendix B contains the details of the simulation model. Appendix C gives further diagnostic plots of DLM for the volume of detached clusters and Table of other parameters in the simulation model.

Appendix A: Derivation of DLMs

Let $\boldsymbol{\beta}_t$ be a Markov chain under some regularity conditions, and suppose that each $\mathbf{Y}_t'^s$ are independent conditionally on $\boldsymbol{\beta}_t$, if we also assume further that $\mathbf{Y}_t$ depends only on $\boldsymbol{\beta}_t$. Therefore, equation A.1 follows directly from equations **2** and **3** in the original article which completely determines any given state space model, and other distributions can be easily derived using this equation

$$\pi(\boldsymbol{\beta}_{0:t}, \mathbf{y}_{1:t}) = \pi(\boldsymbol{\beta}_0).\prod_{j=1}^{t}\pi(\boldsymbol{\beta}_j|\boldsymbol{\beta}_{j-1})\pi(\mathbf{y}_j|\boldsymbol{\beta}_j), \qquad t > 0 \tag{A.1}$$

where $\pi(\boldsymbol{\beta}_0)$ is the initial distribution and $\pi(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1})$ and $\pi(\mathbf{y}_t|\boldsymbol{\beta}_t)$ are conditional densities; see details in Petris et al. (2009) and Petris et al. (2011). For a general state space model, we use the relations given in equation A.2 below

$$\begin{cases} \pi(\boldsymbol{\beta}_t|\mathbf{y}_{1:t-1}) = \int \pi(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1})\pi(\boldsymbol{\beta}_{t-1}|\mathbf{y}_{1:t-1})d\boldsymbol{\beta}_{t-1} & \text{state predictive density} \\ \pi(\mathbf{y}_t|\mathbf{y}_{1:t-1}) = \int \pi(\mathbf{y}_t|\boldsymbol{\beta}_t)\pi(\boldsymbol{\beta}_t|\mathbf{y}_{1:t-1})d\boldsymbol{\beta}_t & \text{observation predictive density} \\ \pi(\boldsymbol{\beta}_t|\mathbf{y}_{1:t}) = \dfrac{\pi(\mathbf{y}_t|\boldsymbol{\beta}_t)\pi(\boldsymbol{\beta}_t|\mathbf{y}_{1:t-1})}{\pi(\mathbf{y}_t|\mathbf{y}_{1:t-1})} & \text{filtering density,} \end{cases} \tag{A.2}$$

and suppose further that the random vectors $(\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_t, \mathbf{Y}_1, \ldots, \mathbf{Y}_t)$ are normally distributed, and for any time $t > 0$, then the marginal and conditional distributions are also normally distributed which are completely specified by their means and variances. Therefore, using a normality assumption in conjunction with dynamic linear models defined in equations **2** and A.2, and suppose that $\boldsymbol{\beta}_t|\mathbf{y}_{1:t-1} \sim N(m_{t-1}, C_{t-1})$, we can deduce easily that the one-step ahead predictive distributions for $\boldsymbol{\beta}_t|\mathbf{y}_{1:t-1}$, $\mathbf{Y}_t|\mathbf{y}_{1:t-1}$ and filtering distribution of $\boldsymbol{\beta}_t|\mathbf{y}_{1:t}$ are given respectively as

$$\begin{cases} \boldsymbol{\beta}_t|\mathbf{y}_{1:t-1} \sim N(a_t, \mathbf{R}_t), & \text{where } a_t = \mathbf{G}_t m_{t-1}, \quad \mathbf{R}_t = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t' + \mathbf{W}_t, \\ \mathbf{y}_t|\mathbf{y}_{1:t-1} \sim N(f_t, \mathbf{Q}_t), & \text{where } f_t = \mathbf{F}_t a_t, \quad \mathbf{Q}_t = \mathbf{F}_t\mathbf{R}_t\mathbf{F}_t' + \mathbf{V}_t, \\ \boldsymbol{\beta}_t|\mathbf{y}_{1:t} \sim N(m_t, \mathbf{C}_t), & \text{where } m_t = a_t + \mathbf{A}e_t, \quad \mathbf{C}_t = \mathbf{R}_t + \mathbf{A}\mathbf{F}_t\mathbf{R}_t + \mathbf{W}_t, \end{cases} \tag{A.3}$$

where $e_t = \mathbf{Y}_t - f_t$ and $\mathbf{A} = \mathbf{R}_t + \mathbf{F}_t'\mathbf{Q}_t^{-1}$. We already know that the joint distribution of states and observations is Gaussian, when all the parameters of a dynamic regression are known it is relatively easy to derive conditional distributions of states or future observations conditional on

the given data. These parameters are unknown, and a major problem in time series analysis is their estimation. One popular approach is to use the maximum likelihood estimation (MLE) framework by maximising the likelihood function for the statistical parameter estimation. The limitations of MLE are the presence of many local maximal and flat likelihood and failure to always account for the uncertainty associated with the estimation of such parameters. Here, we describe a Bayesian method where the unknown parameters are treated as random variables.

In our case, the unknown parameters that are required to be estimated are the state evolution matrix $\mathbf{G}_t$ and the evolution and observation variance matrices $\mathbf{W}_t$ and $\mathbf{V}_t$. To simplify our approach and make the problem identifiable, we assume they are diagonal matrices and constant over time. Therefore, we define the matrices of unknown parameters as below

$$\begin{cases} \mathbf{G}_t = \mathbf{G} = diag(\psi_1, \ldots, \psi_k), \\ \mathbf{W}_t = \mathbf{W} = diag(\phi_{\mathbf{y},1}^{-1}, \ldots, \phi_{\mathbf{y},m}^{-1}), \\ \mathbf{V}_t = \mathbf{V} = diag(\phi_{\boldsymbol{\beta},1}^{-1}, \ldots, \phi_{\boldsymbol{\beta},p}^{-1}), \end{cases} \tag{A.4}$$

and take the parameters of the evolution matrix $\mathbf{G}$ to be normally independent and identically distributed such that

$$\psi_j \sim N(\psi_0, \tau_0), \qquad j = 1, \ldots, k, \tag{A.5}$$

and $\phi_{\mathbf{y},i}^{-1}$ and $\phi_{\boldsymbol{\beta},j}^{-1}$ to have an independent gamma distribution

$$\begin{cases} \phi_{\mathbf{y},i}^{-1} = Ga(\alpha_{\mathbf{y},i}, b_{\mathbf{y},i}), & i = 1, \ldots, m, \\ \phi_{\boldsymbol{\beta},j}^{-1} = Ga(\alpha_{\boldsymbol{\beta},i}, b_{\boldsymbol{\beta},i}), & j = 1, \ldots, p. \end{cases} \tag{A.6}$$

Therefore, given the observations $\mathbf{y}_{1:T}$, we can obtain the joint posterior density of the states $\boldsymbol{\beta}_{0:T}$ and unknown parameters $\boldsymbol{\Psi} = (\psi_j, \phi_{\mathbf{y},i}^{-1}, \phi_{\boldsymbol{\beta},j}^{-1})$ which is proportional to their joint density such that

$$\begin{cases} \pi(\mathbf{y}_{1:T}, \boldsymbol{\beta}_{0:T}, \boldsymbol{\Psi}) = \prod_{t=1}^{T} \pi(\mathbf{y}_t|\boldsymbol{\beta}_t, \phi_{\mathbf{y}}). \prod_{t=1}^{T} \pi(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}, \\ \phi_{\boldsymbol{\beta},1}, \ldots, \phi_{\boldsymbol{\beta},p})\pi(\boldsymbol{\beta}_0) \prod_{i=1}^{m} \pi(\phi_{\mathbf{y},i}) \prod_{j=1}^{p} \pi(\phi_{\boldsymbol{\beta},i}). \end{cases} \tag{A.7}$$

It is always difficult to obtain the posterior distribution defined above in a closed form and as a result we can result to numerical approximation using MCMC methods. In particular, the joint posterior of the states and model parameters can be approximated by Gibbs sampling. The full conditionals of $\phi_{\mathbf{y},i}^{-1}$ and $\phi_{\boldsymbol{\beta},j}^{-1}$ can be derived from their joint density and are both Gamma distributions given below as

$$\begin{cases} \phi_{\mathbf{y},i}|\ldots \sim Ga\left(\alpha_{\mathbf{y},i} + \frac{T}{2}, b_{\mathbf{y},i} + \frac{1}{2}S_{\mathbf{y},i}\right), & i = 1, \ldots, m, \\ \phi_{\boldsymbol{\beta},j}|\ldots \sim Ga\left(\alpha_{\boldsymbol{\beta},j} + \frac{T}{2}, b_{\boldsymbol{\beta},j} + \frac{1}{2}S_{\boldsymbol{\beta},j}\right), & j = 1, \ldots, p, \end{cases} \tag{A.8}$$

where $S_{\mathbf{y},i} = \sum_{t=1}^{T}(\mathbf{y}_{i,t} - (\mathbf{F}\boldsymbol{\beta}_t)i)^2$ and $S_{\boldsymbol{\beta},j} = \sum_{t=1}^{T}(\boldsymbol{\beta}_{j,t} - (\mathbf{G}\boldsymbol{\beta}_{t-1})_j)^2$.

Therefore, the sampler can be run to draw a sample from both the full conditional distributions of the states $(\phi_{\mathbf{y},i}|\ldots)$ and $(\phi_{\boldsymbol{\beta},j}|\ldots)$. The Gibbs sampler is an efficient technique for approximating

the joint distribution. The mechanism of Gibbs sampling involves iteratively simulating from full conditional distributions. Lastly, the full conditional of ψ_j is derived by combining the prior distribution with the theory of normal distributions such that

$$\psi_j|\ldots \sim N(\psi_{j,T}, \tau_{j,T}), \qquad j = 1, \ldots, k, \tag{A.9}$$

with $\psi_{j,T} = \tau_{j,T}\left[\phi_{\beta,j}\sum_{t=1}^{T}\boldsymbol{\beta}_{j,t-1}\boldsymbol{\beta}_{j,t} + \frac{1}{\tau_{j,0}\psi_0}\right]$ and $\tau_{j,T} = \left[\frac{1}{\tau_0} + \phi_{\beta,j}\sum_{t=1}^{T}\boldsymbol{\beta}_{j,t-1}^2\right]^{-1}$. See (Harrison and West, 1999; Doss and Narasimhan, 1994; Casella and George, 1992) for further background on Gibbs sampling. The summary of this algorithm is given below.

Summary of Gibbs sampling algorithm

Derive the posterior conditionals for each of the random variables in the model from the full density.
Simulate samples from the posterior joint distribution based on the posterior conditionals.

- Initialize: set $\phi_{\mathbf{y}} = \phi_{\mathbf{y}}^{(0)}$; $\phi_{\boldsymbol{\beta}} = \phi_{\boldsymbol{\beta}}^{(0)}$; $\psi = \psi^{(0)}$ For $k = 1, \ldots, N$:
- Draw $\boldsymbol{\beta}_{0:T}^{(k)}$ from $\pi(\boldsymbol{\beta}_{0:T}|\mathbf{y}_T, \phi_{\mathbf{y}} = \phi_{\mathbf{y}}^{(k-1)}, \phi_{\boldsymbol{\beta}} = \phi_{\boldsymbol{\beta}}^{(k-1)}, \psi = \psi^{(k-1)})$ with FFBS
- Draw $\phi_{\mathbf{y}}^{(k)}$ from $\pi(\phi_{\mathbf{y}}|\mathbf{y}_T, \boldsymbol{\beta}_{0:T} = \boldsymbol{\beta}_{0:T}^{(k)}, \phi_{\boldsymbol{\beta}} = \phi_{\boldsymbol{\beta}}^{(k-1)}, \psi = \psi^{(k-1)})$
- Draw $\phi_{\boldsymbol{\beta}}^{(k)}$ from $\pi(\phi_{\boldsymbol{\beta}}|\mathbf{y}_T, \boldsymbol{\beta}_{0:T} = \boldsymbol{\beta}_{0:T}^{(k)}, \phi_{\mathbf{y}} = \phi_{\mathbf{y}}^{(k)}, \psi = \psi^{(k-1)})$
- Draw $\psi^{(k)}$ from $\pi(\psi|\mathbf{y}_T, \boldsymbol{\beta}_{0:T} = \boldsymbol{\beta}_{0:T}^{(k)}, \phi_{\mathbf{y}} = \phi_{\mathbf{y}}^{(k)}, \phi_{\boldsymbol{\beta}} = \phi_{\boldsymbol{\beta}}^{(k)})$

Appendix B: Simulation model

We have one functional group of microorganism and one dormant state as soft agents within the present model. The microorganisms are heterotrophs (HET) which consume organic carbon source and oxygen. The inert state is extracellular polymeric substance (EPS), secreted by some heterotrophs. The EPS can also be regarded as 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 (Wingender et al., 1999). The dead agents are represented by soft spheres (labelled DEAD). Agents have four state variables: position, mass, radius, and type. This model consists of two principal submodels; one deals with the growth and behaviour of individual bacteria as autonomous agents (i.e., biological processes); the other deals with the substrate and product diffusion and reaction and fluid flow (i.e., physical processes). Each cell grows by consuming the substrate and divides when a certain mass is reached. When agents grow and split, the system deviates from its mechanical equilibrium due to some extra pressure built-up in the biomass.

Depending on the net force acting on each agent, resulting from its spatial interaction with other local agents, the position of each agent is updated using the Discrete Element Method (DEM). In DEM, contact, EPS adhesion, shear force are considered, and the position of agents are

updated by solving Newton's second law equation. For the substrates, Chemical Oxygen Demand (COD), oxygen, ammonia, nitrite, and nitrate are considered. The diffusion-reaction equation governs the substrate concentrations, and this transport equation is solved in a fixed Cartesian grid using a Finite Difference Method. This model extends the traditional IB model by incorporating mechanical interactions among bacteria. The model is implemented in LAMMPS, an open-source C^{++} molecular dynamics code (http://lammps.sandia.gov/) (Plimpton, 1995). More details about the model can be found in Jayathilake et al. (2017).

Appendix C: Diagnostic plots for DLM of volume of detached clusters

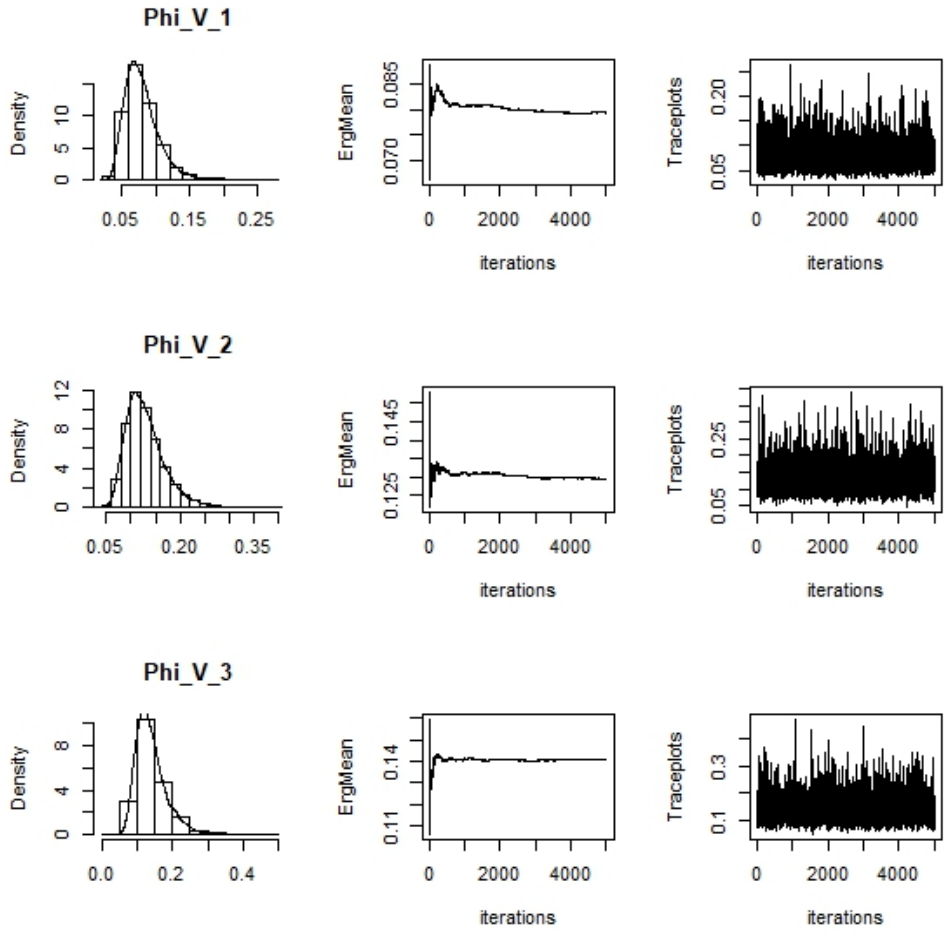


Figure A.1: Diagnostic plots showing the convergence of the three V_1, V_2 and V_3 randomly chosen observation variance parameters of the Bayesian dynamic linear model. The first column shows the posterior density of the observation variances. The middle column is the running ergodic means of MCMC samples. The third colum is the traceplots for the MCMC samples.

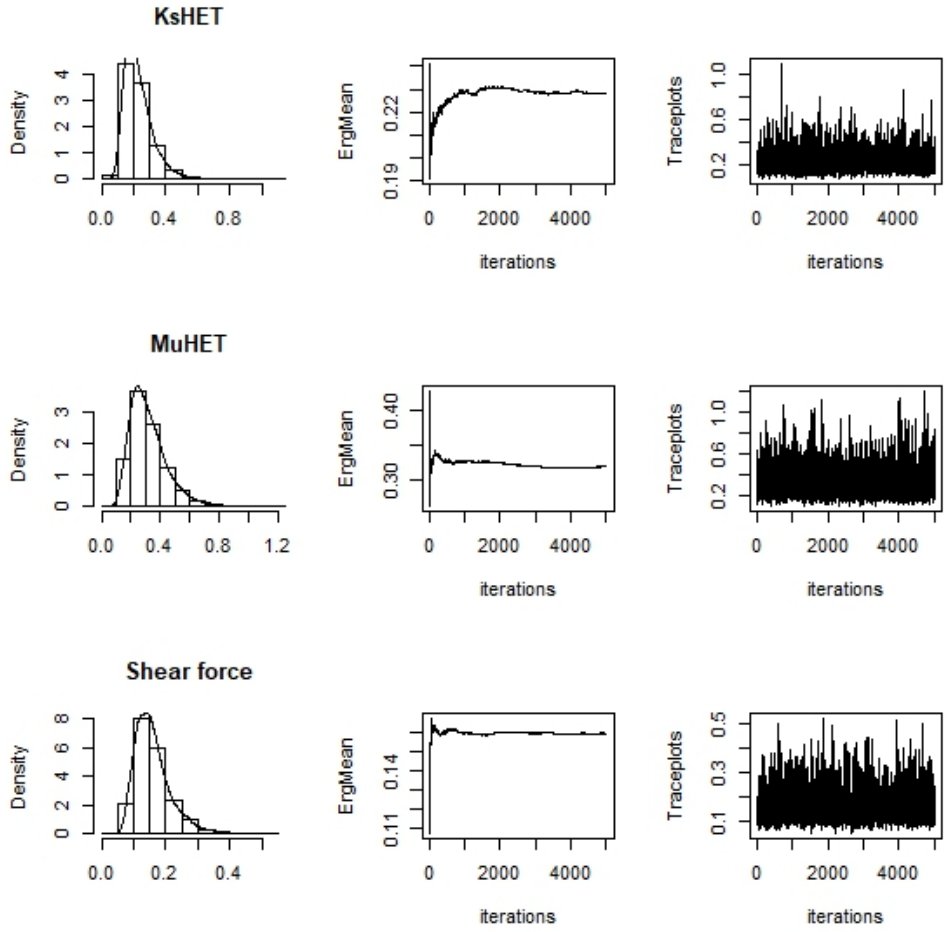


Figure A.2: Diagnostic plots showing the convergence of the three randomly chosen W_1, W_2 and W_3 state variance parameters of the Bayesian dynamic linear model. The first column shows the posterior density of the state variances. The middle column is the running ergodic means of MCMC samples. The third colum is the trace plots for the MCMC samples.

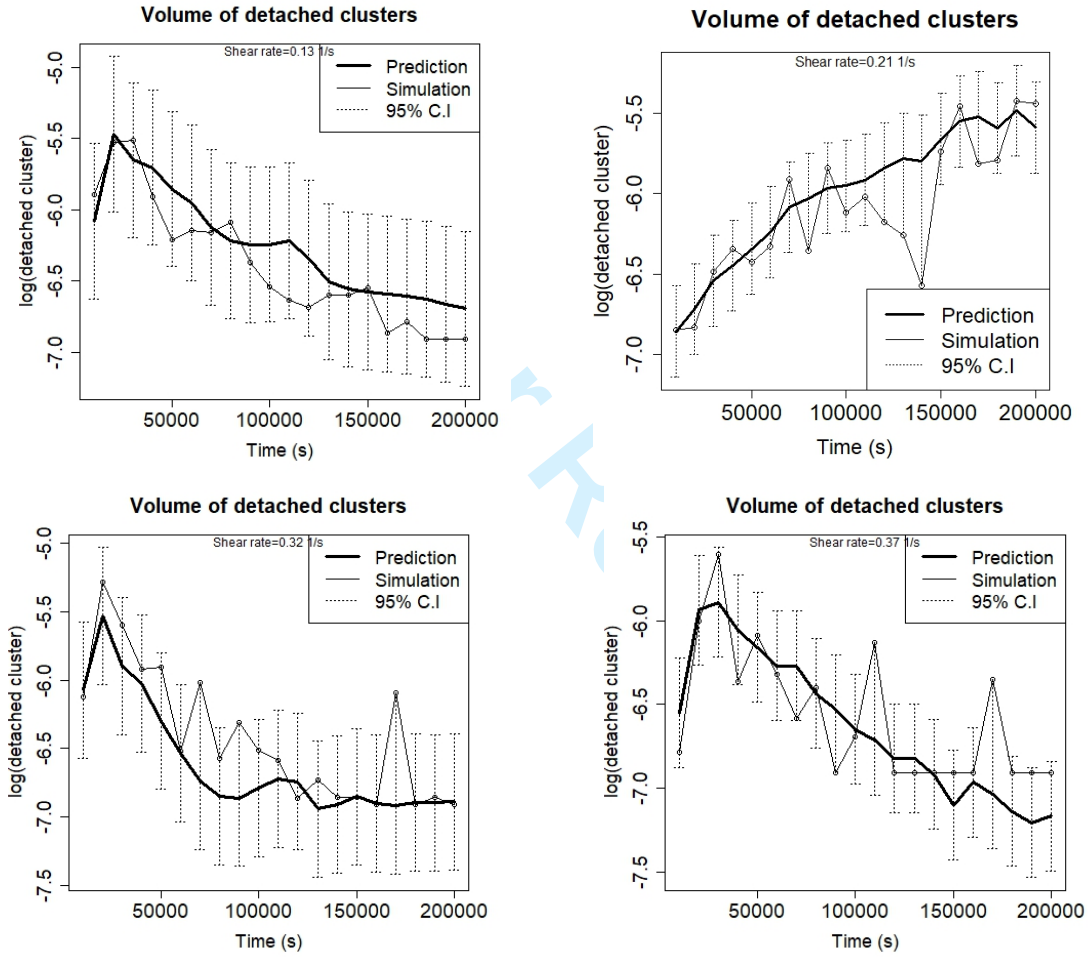


Figure A.3: Comparison between the simulation and prediction for log-transformed detached volume over time for different shear forces. The results are normalized by initial biofilm volume.

Table A.1: List of parameters that remain unchanged for all simulation experiments

Index	Parameters	Values	Units	References
	Affinity variables			
1	Kno2HET	0.0003	kgm^{-3}	Ofiţeru et al. (2014)
2	Kno3HET	0.0003	kgm^{-3}	Ofiţeru et al. (2014)
3	Knh4AOB	0.001	kgm^{-3}	Bruce and Perry (2001)
4	Ko2AOB	0.0005	kgm^{-3}	Bruce and Perry (2001)
5	Kno2NOB	0.0013	kgm^{-3}	Bruce and Perry (2001)
6	Ko2NOB	0.00068	kgm^{-3}	Bruce and Perry (2001)
	Yield coefficient variables			
8	YAOB	0.33	$gCOD/gN$	Bruce and Perry (2001)
9	YNOB	0.083	$gCOD/gN$	Bruce and Perry (2001)
10	YEPS	0.18	$gCOD/gN$	Ni et al. (2009)
	Diffusion coefficient variables			
11	D_{o2}	0.0000000023	$m^2 s^{-1}$	Alpkvist et al. (2006)
12	D_{nh4}	0.00000000115	$m^2 s^{-1}$	Alpkvist et al. (2006)
13	D_{no2}	0.00000000115	$m^2 s^{-1}$	Alpkvist et al. (2006)
14	D_{no3}	0.00000000115	$m^2 s^{-1}$	Alpkvist et al. (2006)
15	D_s	0.0000000016	$m^2 s^{-1}$	Alpkvist et al. (2006)
	Critical diameter of death			
16	deadDia	0.0000008	-	-
17	factor	1.5	-	-
	Boundary concentrations (nutrients)			
18	sub	0.00010	kgCOD m^{-3}	Chosen
19	no2	0.00010	kgN m^{-3}	Chosen
20	no3	0.00010	kgN m^{-3}	Chosen
21	o2	0.00010	kg m^{-3}	Chosen
22	nh4	0.00010	kgN m^{-3}	Chosen
	Mechanics			
23	Spring coefficient for collision k_n	1×10^{-4}	Nm^{-1}	Celler et al. (2014)
24	Viscous coefficient for collision γ^n	1×10^{-5}	s^{-1}	Chosen
25	EPS stiffness per unit EPS mass ke	5×10^{9}	s^{-2}	Head (2013)
26	Dynamic viscosity μ	1×10^{-3}	Pas	(For water)

References

Alpkvist, E., Picioreanu, C., van Loosdrecht, M., and Heyden, A. (2006). Three-dimensional biofilm model with individual cells and continuum eps matrix. *Biotechnology and bioengineering*, 94(5):961–979.

Bruce, E. R. and Perry, L. M. (2001). Environmental biotechnology: principles and applications. *New York: McGrawHill*, 400.

Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174.

Celler, K., Hödl, I., Simone, A., Battin, T., and Picioreanu, C. (2014). A mass-spring model unveils the morphogenesis of phototrophic diatoma biofilms. *Scientific reports*, 4.

Doss, H. and Narasimhan, B. (1994). Bayesian poisson regression using the gibbs sampler: Sensitivity analysis through dynamic graphics. Technical report, Technical report, Citeseer.

Harrison, J. and West, M. (1999). *Bayesian forecasting & dynamic models*. Springer New York.

Head, D. (2013). Linear surface roughness growth and flow smoothening in a three-dimensional biofilm model. *Physical Review E*, 88(3):032702.

Jayathilake, P. G., Gupta, P., Li, B., Madsen, C., Oyebamiji, O., González-Cabaleiro, R., Rushton, S., Bridgens, B., Swailes, D., Allen, B., et al. (2017). A mechanistic individual-based model of microbial communities. *PloS one*, 12(8):e0181965.

Ni, B.-J., Fang, F., Xie, W.-M., Sun, M., Sheng, G.-P., Li, W.-H., and Yu, H.-Q. (2009). Characterization of extracellular polymeric substances produced by mixed microorganisms in activated sludge with gel-permeating chromatography, excitation–emission matrix fluorescence spectroscopy measurement and kinetic modeling. *Water Research*, 43(5):1350–1358.

Ofiţeru, I. D., Bellucci, M., Picioreanu, C., Lavric, V., and 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.

Petris, G., Petrone, S., and Campagnoli, P. (2009). Dynamic linear models. In *Dynamic Linear Models with R*, pages 31–84. Springer.

Petris, G., Petrone, S., et al. (2011). State space models in r. *Journal of Statistical Software*, 41(4):1–25.

Plimpton, S. (1995). Fast parallel algorithms for short-range molecular dynamics. *Journal of computational physics*, 117(1):1–19.

Wingender, J., Neu, T. R., and Flemming, H.-C. (1999). What are bacterial extracellular polymeric substances? In *Microbial extracellular polymeric substances*, pages 1–19. Springer.