Commun Nonlinear Sci Numer Simulat 73 (2019) 403–424

Contents lists available at ScienceDirect

Commun Nonlinear Sci Numer Simulat

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

Research paper

Surrogate-based uncertainty and sensitivity analysis for bacterial invasion in multi-species biofilm modeling

A. Trucchia[a,*], M.R. Mattei[b], V. Luongo[b], L. Frunzo[b], M.C. Rochoux[c]

[a] BCAM – Basque Center for Applied Mathematics, Alameda de Mazarredo 14, Bilbao 48009, Basque Country, Spain
[b] Department of Mathematics and Applications “R. Caccioppoli”, via Cintia 1, Naples 91126, Italy
[c] CECI, Université de Toulouse, CNRS, CERFACS, 42 Avenue Gaspard Coriolis, Toulouse cedex 1, 31057, France

ARTICLE INFO

Article history:
Received 14 November 2018
Revised 22 February 2019
Accepted 28 February 2019
Available online 4 March 2019

Keywords:
Biofilm modeling
Surrogate modeling
Uncertainty quantification
Sensitivity analysis

ABSTRACT

In this work, we present a probabilistic analysis of a detailed one-dimensional biofilm model that explicitly accounts for planktonic bacterial invasion in a multi-species biofilm. The objective is (1) to quantify and understand how the uncertainty in the parameters of the invasion submodel impacts the biofilm model predictions (here the microbial species volume fractions); and (2) to spot which parameters are the most important factors enhancing the biofilm model response. An emulator (or “surrogate”) of the biofilm model is trained using a limited experimental design of size $N = 216$ and corresponding to a Halton’s low-discrepancy sequence in order to optimally cover the uncertain space of dimension $d = 3$ (corresponding to the three scalar parameters newly introduced in the invasion submodel). A comparison of different types of emulator (generalized Polynomial Chaos expansion – gPC, Gaussian process model – GP) is carried out; results show that the best performance (measured in terms of the Q_2 predictive coefficient) is obtained using a Least-Angle Regression (LAR) gPC-type expansion, where a sparse polynomial basis is constructed to reduce the problem size and where the basis coordinates are computed using a regularized least-square minimization. The resulting LAR gPC-expansion is found to capture the growth in complexity of the biofilm structure due to niche formation. Sobol’ sensitivity indices show the relative prevalence of the maximum colonization rate of autotrophic bacteria on biofilm composition in the invasion submodel. They provide guidelines for orienting future sensitivity analysis including more sources of variability, as well as further biofilm model developments.

1. Introduction

Recent experimental activity has highlighted that both in natural and artificial environments, microorganisms preferentially exist in the form of self-organized assemblages termed “biofilms”, consisting of surface-associated communities embedded in an exopolysaccharide matrix and organized into microcolonies [1,2]. The exopolysaccharide matrix corresponds to extracellular polymeric substances that are secreted by microorganisms into their environment and that play an important role in the cell attachment to a given surface and therefore in the biofilm formation. Bacteria in biofilms differ substantially from free-living bacterial cells through a set of emerging properties, including the formation of physical and social interac-

[*] Corresponding author.
E-mail address: atrucchia@bcamath.org (A. Trucchia).

https://doi.org/10.1016/j.cnsns.2019.02.024
1007-5704/© 2019 Elsevier B.V. All rights reserved.

Nomenclature

Abbreviation	
GP	Gaussian Process
gPC	generalized Polynomial Chaos
LAR	Least-Angle Regression
PDF	Probability Density Function
RBF	Radial Basis Function
SLS	Standard Least Squares
STD	STandard Deviation
Model quantities	*Units*
$\mathcal{F}$,biofilm model operator	–
z, space variable	[L]
t, time variable	[T]
$X_i = \rho_i f_i$, concentration of ith microbial species	$[\mathrm{M\ L}^{-3}]$
ρ_i, biomass density for ith microbial species	$[\mathrm{ML}^{-3}]$
S_j, concentration of jth substrate	$[\mathrm{ML}^{-3}]$
$r_{S,j}$ conversion rate of S_j	$[\mathrm{ML}^{-3}\mathrm{T}^{-1}]$
$r_{M,i}$, specific growth rate of X_i	$[\mathrm{T}^{-1}]$
r_i, specific growth rate of X_i due to invasion	$[\mathrm{T}^{-1}]$
f_i, volume fraction of X_i	$[-]$
ψ_i, concentration of ith planktonic microbial species	$[\mathrm{ML}^{-3}]$
$r_{\psi,i}$, conversion rate of ψ_i	$[\mathrm{ML}^{-3}\mathrm{T}^{-1}]$
$u(z,\ t)$, advective biomass velocity	$[\mathrm{LT}^{-1}]$
$L(t)$, biofilm thickness at time t	[L]
Experiment quantities	*Units*
f_1, heterotrophic bacteria volume fraction	$[-]$
f_2, autotrophic bacteria volume fraction	$[-]$
f_3, inert material volume fraction	$[-]$
S_1, organic carbon concentration	$[\mathrm{g}_{\mathrm{CODm}}^{-3}]$
S_2, ammonia concentration	$[\mathrm{g}_{\mathrm{Nm}}^{-3}]$
S_3, oxygen concentration	$[\mathrm{g}_{\mathrm{O}_2}\mathrm{m}^{-3}]$
$k_{col,2}$, maximum colonization rate of autotrophic bacteria	$[\mathrm{d}^{-1}]$
$k_{\psi,2}$, affinity-type constant for ψ_2	$[\mathrm{g}_{\mathrm{CODm}}^{-3}]$
$Y_{\psi,2}$, yield of X_2 on ψ_2	$[-]$
Uncertainty analysis variables	*Definition*
d	Uncertain space dimension
$\boldsymbol{\theta}$	Vector of input parameters of dimension d
$\mathbf{y}$	Vector of quantities of interest of dimension n
N	Size of the training set

tions, the enhanced rate of gene exchange and the increased tolerance to antimicrobials [1]. Such complex microbial communities drive biogeochemical cycling processes of most elements in water, soil, sediment and subsurface environments. They have been extensively used in biotechnological applications such as waste-water and solid waste treatment, drinking water filtration, biofuel production. Conversely, biofilms can cause persistent infections and contamination of medical devices and implants; they are also responsible for biofouling and process water contamination, quality deterioration of drinking water and microbially influenced corrosion.

Many biofilm models have been proposed in the literature over the last decades [3,4]. Some of them have been derived in the framework of continuum mechanics and formulated as differential equations based on (mass, volume, momentum, energy) conservation principles [5–10]. Others have been introduced as bottom-up models and assume biofilms to be inherently stochastic living systems [11–15]. Still, biofilm modeling remains a challenge, in particular since the biological processes involved in biofilm formation and growth are highly nonlinear and since there is no agreed-upon methodology to guide the user in the selection of the most appropriate model(s) and in the choice of the input parameters. For instance, no reference values have been defined for these inputs [16], while they may affect the nonlinear system in unpredictable ways.

In this context, studying the sensitivity of the biofilm model predictions to the variability in the inputs provides a way to better understand the response of the model to an arbitrary choice of parameters and to highlight new insights into the underlying biological processes. To this aim, for each set of input parameters $\boldsymbol{\theta} = \{\theta_1 \ldots \theta_d\}$, the output of the model is codified into a set of quantities of interest $\mathbf{y} = \{y_1 \ldots y_n\}$, leading to the definition of the functional relation $\mathcal{F}$

$$\boldsymbol{\theta} \in \mathbb{R}^d \quad \rightarrow \quad \mathbf{y} = \mathcal{F}(\boldsymbol{\theta}) \in \mathbb{R}^n. \tag{1}$$

In the framework of uncertainty quantification [17,18], the set of input parameters $\boldsymbol{\theta}$ is considered uncertain and the objective is to propagate the input uncertainties through the numerical model and to estimate the subsequent uncertainties in the quantities of interest $\mathbf{y}$. In complement, global sensitivity analysis methods [19,20] provide valuable ways to characterize the input-output model dependency $\mathcal{F}$: they are helpful to derive a relevant screening of the input parameters, spot unimportant parameters and focus the attention on the most relevant ones. These methods can be classified in at least three categories: variance-based sensitivity analysis [21,22], derivative-based sensitivity analysis [23–26], and moment-independent sensitivity measures [27–29].

Although the parameters involved in biofilm models may vary considerably and interact with each other to determine the model output, only few attempts have been made in the past years to apply uncertainty quantification [30,31] and sensitivity analysis to biofilm models at both local and global levels [32–38]. Most of these studies refer to an application of the original Wanner–Gujer model [5], which is currently the most widely used biofilm model in engineering applications. This model has been integrated in AQUASIM [39], a computer program designed for simulating aquatic systems and also for performing parameter estimation and sensitivity analysis, see Refs. [33,35,36] related to global sensitivity analysis: Ref. [33] presents a comparison between the qualitative Morris screening method and the quantitative variance-based Fourier amplitude sensitivity test for a two-step nitrification biofilm model; Ref. [35] presents variance-based sensitivity analysis applied to a one-dimensional biofilm model for ammonium and nitrite oxidation for varying biofilm reactor geometry; and Ref. [36] calculates sensitivity by performing model output linear regression for a complete autotrophic nitrogen removal biofilm.

However, Wanner–Gujer-type biofilm modeling is not detailed enough to study bacterial invasion mechanisms, which frequently occur and are crucial in most of engineering applications. To overcome this modeling limitation, a new class of continuum models for multi-species biofilm formation and growth, which explicitly accounts for invasion mechanisms, has been recently introduced [40,41]. The novelty in such biofilm modeling class relates to the introduction of a new state variable, which describes the concentration of planktonic species within the biofilm. In this framework, the diffusion of the free cells from the bulk liquid into the biofilm and inversely is described by a diffusion-reaction equation; the growth processes are governed by a system of nonlinear hyperbolic partial differential equations; and substrate dynamics are governed by a system of semi-linear parabolic partial differential equations. All equations are mutually connected so that the resulting system of differential equations corresponds to a free boundary value problem, where the free boundary is represented by the biofilm thickness. This model formulation aims at reproducing the colonization of new species diffusing from bulk liquid to biofilm and the development of latent microbial species within the biofilm, without explicitly prescribing boundary conditions for the invading species at the free boundary. Such boundary conditions are determined self-consistently by the model, instead of being set arbitrarily [42].

This new class of continuum models can handle any number of microbial species, both in sessile and planktonic states, as well as dissolved substrates. One difficulty is that this type of model involves parameters related to species invasion that are rather new in the literature and whose reference values are not obvious to specify. To overcome this issue, we present in this study, a variance-based sensitivity analysis approach that makes use of the well known Sobol' indices [21,43] to identify the most important parameters related to bacterial invasion mechanisms. These Sobol' indices derived from variance decomposition quantify the contribution of each uncertain parameter to the variance of the quantities of interest. One non-intrusive way to compute them is to build a Monte Carlo random sample of inputs and simulated outputs [44]. While this approach is generic and robust, it is computationally expensive due to a slow rate of convergence with respect to the sample size. Due to the complexity of the biofilm model, this would require the order of 10^4–10^5 biofilm model simulations and this is therefore far out of the available computational budget. An alternative is to derive (or "train") an emulator of the biofilm model using a limited sample of inputs and simulated outputs (or "training set") and taking advantage of the regularity of the model response $\mathcal{F}$. Stated differently, the objective is to fit the emulator (or "surrogate") over a dataset of biofilm model simulations and then to mimic in an accurate and efficient way, the model response $\mathcal{F}$ for any set of parameters $\boldsymbol{\theta}$ without solving the original system of differential equations. Statistical information on the quantities of interest and Sobol' indices can then be computed using the emulator. Emulating can be regarded as a supervised learning procedure and belongs to the field of machine learning [45].

In this study, the objective is to build a surrogate that accurately represents bacterial invasion as described by a recent multi-species biofilm model and use it to perform uncertainty quantification and global sensitivity analysis. In order to provide results that are not algorithm-dependent, we compare two families of popular surrogate models, namely generalized Polynomial Chaos (gPC) [46–51] and Gaussian Processes (GP) [52–56]. Comparison of gPC-expansion and GP-model have been reported in the literature [57–60]; Ref. [59] highlights that one approach does not systematically outclass the other in terms of surrogate accuracy and computational efficiency, the best surrogate being application-dependent. It is therefore of interest to compare gPC and GP approaches for biofilm applications. The training step of the surrogate requires a sampling of the uncertain input space. The GP approach is known to be more accurate for less structured design than tensor grid when performing sensitivity analysis [61]. Consistently, the sampling is performed here using a low-discrepancy Halton's sequence with a given budget $N = 216$. Due to the nonlinearities of the biological processes involved, we investigate the impact of different choices of the gPC polynomial basis (full or sparse) on the surrogate performance for a fixed sample size

N. Using a sparse polynomial basis may reduce the size of the stochastic problem by only selecting the most significant basis components, and help to better capture a complex model response to variations in the input parameters [62]. We consider here the least-angle regression (LAR) approach to build a sparse gPC basis [63,64], which was found to provide the best performance among several sparse methods in Ref. [62].

The biofilm model we use folds into the category of hyperbolic partial differential equations, meaning that the quantities of interest may feature sharp variations, possibly discontinuities, for certain part of the input stochastic space. In this situation, building an accurate surrogate that covers the whole input space when dealing with model nonlinearities remains a challenge [47,49,50,65]. One way to overcome this issue is to partition the input space, to build local surrogates and combine them into a mixture-of-experts model [66]. It is thus of primary interest to investigate if building a global surrogate is feasible for biofilm applications before moving to more advanced settings such as mixture of experts.

In this work, the target problem represents a typical microbial interaction occurring in waste-water treatment plants. Initially, the biofilm is only made of heterotrophic bacteria and latent autotrophic bacteria are present in the bulk liquid; then autotrophic bacteria infiltrate the biofilm, switch their state from planktonic to sessile mode and start to proliferate, where they meet the best environmental conditions for their growth. The gPC and GP surrogates are exploited to quantity the uncertainties in the microbial species volume fractions and analyze their dependency with respect to three parameters related to the autotrophic bacterial invasion (the problem dimension is $d = 3$ in Eq. (1)). Note that in the literature, global sensitivity analysis and uncertainty quantification mostly deal with scalar outputs, while the biofilm model output here is functional with spatial and temporal discretizations, $n > 1$ in Eq. (1). Our approach consists here in building a surrogate at each time step of interest, over the spatial grid associated to the model output [20,67,68].

The paper is organized as follows. The biofilm model is described in Section 2. Section 3 presents the uncertain input parameters, the quantities of interest, the stochastic framework and the experimental design to build the training set. Section 4 presents the key ideas of the gPC and GP surrogates. Uncertainty quantification and global sensitivity analysis results are presented in Section 5. Conclusions and perspectives are outlined in Section 6.

2. Biofilm model

We present the recent continuum model [40] describing in a quantitative and deterministic way, the bacterial invasion in multi-species biofilms [3]. This model essentially consists of a modified Wanner–Gujer formulation accounting for the dynamics of the invading planktonic species as well as substrate diffusion, attachment, detachment, microbial growth and biomass spreading. Note that this model has been derived in one dimension and then generalized to three dimensions [4]. In the present study, we consider the one-dimensional model.

2.1. Free boundary value problem

The invasion model is formulated as a free boundary value problem for the three state variables: (1) the concentration of microbial species in sessile form $X_i(z, t)$, $i = 1, \dots, N_s$, $\mathbf{X} = X_1, \dots, X_{N_s}$; (2) the concentration of planktonic species $\psi_i(z, t)$, $i = 1, \dots, N_s$, $\boldsymbol{\psi} = \psi_1, \dots, \psi_{N_s}$; and (3) the concentration of dissolved substrates $S_j(z, t)$, $j = 1, \dots, N_m$, $\mathbf{S} = S_1, \dots, S_{N_m}$, including the substrates provided by the bulk liquid and the metabolic waste products related to microbial metabolism. Note that the state variables are functions of time t and space z, with z denoting the one-dimensional spatial coordinate assumed perpendicular to the substratum surface located at $z = 0$. Note also that for generality, both the microbial species in sessile and planktonic states are in number of N_s, although in most of applications N_s denotes the number of all particulate components, such as extracellular polymeric substance, inert material and all the phenotype variants of the microbial species.

In this model, the concentration of the ith microbial species in sessile form $X_i(z, t)$ reads

$$\begin{cases} \dfrac{\partial X_i}{\partial t}(z,t) + \dfrac{\partial}{\partial z}(u(z,t)X_i(z,t)) = \rho_i\, r_{M,i}(z,t,\mathbf{X},\mathbf{S}) + \rho_i\, r_i(z,t,\mathbf{S},\boldsymbol{\psi}), \\ X_i(z,0) = \varphi_i(z),\ t = 0,\ 0 \le z \le L(0). \end{cases} \tag{2}$$

Eq. (2) describes the growth of the ith microbial species constituting the biofilm and derives from mass conservation. Biofilm expansion is driven by biomass accumulation. In particular, biomass spreading is modeled as an advective mass flux of each species. The reaction terms $r_{M,i}$ describe the growth of sessile cells (which is controlled by the local availability of nutrients and which is usually described as standard Monod kinetics) and the natural death of cells. The terms r_i represent the growth rates of the ith microbial species due to colonization, which induces the switch of planktonic cells to a sessile growth mode. This phenotypic alteration is catalyzed by the formation within the biofilm matrix of specific environmental niches. Note that Eq. (2) can be written in terms of volume fractions

$$f_i = X_i/\rho_i,\ \sum_{i=1}^{N_s} f_i = 1, \tag{3}$$

where f_i represents the volume fraction at a particular location that is occupied by the ith species, and where ρ_i denotes the biomass density for the ith species, usually assumed the same for all microbial species. Note that $\varphi_i(z)$ in Eq. (2) represents the initial distribution of biofilm particulate components at initial time; for invading microbial species, $\varphi_i(z) = 0$. Note also that the advective biomass velocity $u(z,\ t)$ corresponding to the velocity at which the microbial mass is displaced with respect to the film-support interface is computed as

$$\begin{cases} \dfrac{\partial u}{\partial z}(z,t) = \displaystyle\sum_{i=1}^{N_s} \big(r_{M,i}(z,t,\mathbf{X},\mathbf{S}) + r_i(z,t,\mathbf{S},\boldsymbol{\psi})\big), \\ u(0,t) = 0,\ z = 0,\ t \geq 0. \end{cases} \tag{4}$$

$u(z,\ t)$ is determined by the mean observed specific growth rate of the biomass; it is assumed identical for all considered species. $u(z,\ t)$ also depends on the specific growth rates related to invasion process. The boundary condition at $z=0$ is derived from a no-flux condition at the substratum surface.

Moreover, the biofilm extent (or "thickness") changes with time, i.e. $L \equiv L(t)$. Eq. (5) governs the evolution of the free boundary, which depends on the displacement velocity of microbial biomass as well as on the attachment and detachment fluxes:

$$\begin{cases} \dfrac{dL}{dt}(t) = u(L(t),t) + \sigma_a(t) - \sigma_d(L(t)),\ t > 0, \\ L(0) = L_0,\ t = 0, \end{cases} \tag{5}$$

where L_0 corresponds to the initial biofilm thickness. Eq. (5) is derived from conservation principles at global scale.

The concentration of the ith planktonic species $\psi_i(z,\ t)$ is governed by the following diffusion-reaction equation:

$$\begin{cases} \dfrac{\partial \psi_i}{\partial t}(z,t) - \dfrac{\partial}{\partial z}\left(D_{M,i}\dfrac{\partial \psi_i}{\partial z}(z,t)\right) = r_{\psi,i}(z,t,\mathbf{S},\boldsymbol{\psi}), \\ \psi_i(z,0) = \psi_{i,0}(z),\ t = 0,\ 0 \leq z \leq L(0), \\ \dfrac{\partial \psi_i}{\partial z}(0,t) = 0, z = 0,\ t > 0, \\ \psi_i(L(t),t) = \psi_i^*(t),\ z = L(t),\ t > 0. \end{cases} \tag{6}$$

Eq. (6) governs the movement of planktonic cells within the biofilm matrix. The reaction terms $r_{\Psi,i}$ represent a loss term for invading species when biofilm colonization occurs. $D_{M,i}$ denotes the diffusion coefficient of the ith planktonic species within the biofilm. For all considered microbial species, the initial concentration of planktonic cells within the biofilm is usually set to 0 (implying that invasion occurs at initial time) or using a spatially-distributed specific function $\psi_{i,0}(z)$. Homogeneous Neumann conditions are adopted on the substratum surface at $z=0$ due to a no-flux condition. Dirichlet boundary conditions are prescribed at the free boundary $z=L(t)$. The functions $\psi_i^*(t)$ represent the concentrations of planktonic cells within the bulk liquid; they can be prescribed or derived from mass conservation within the bulk liquid.

The concentration of the jth dissolved substrate $S_j(z,\ t)$ is also governed by a reaction-diffusion equation

$$\begin{cases} \dfrac{\partial S_j}{\partial t}(z,t) - \dfrac{\partial}{\partial z}\left(D_j\dfrac{\partial S_j}{\partial z}(z,t)\right) = r_{S,j}(z,t,\mathbf{X},\mathbf{S}), \\ S_j(z,0) = S_{j,0}(z),\ t = 0,\ 0 \leq z \leq L(0), \\ \dfrac{\partial S_j}{\partial z}(0,t) = 0,\ z = 0,\ t > 0, \\ S_j(L(t),t) = S_j^*(t),\ t > 0, \end{cases} \tag{7}$$

where the term $r_{S,j}$ represents the jth substrate production or consumption rate due to microbial metabolism, and where D_j denotes the diffusion coefficient of the jth substrate within the biofilm. The initial concentration of the jth dissolved substrate is prescribed using the function $S_{j,0}(z)$. As for the concentrations of planktonic species $\psi_i(z,\ t)$, homogeneous Neumann conditions are adopted for $S_j(z,\ t)$ on the substratum surface at $z=0$ due to a no-flux condition, and Dirichlet boundary conditions $S_j^*(t)$ are prescribed at the free boundary $z=L(t)$.

2.2. Autotrophic colonization

In the present study, we consider the following target problem: the biofilm is constituted by three particulate components, heterotrophic bacteria X_1, autotrophic bacteria X_2, and inert material X_3 (X_3 directly results from the decay of the two active microbial species X_1 and X_2).

At initial time, we assume that the biofilm is only composed of heterotrophic bacteria and we enhance autotrophic colonization. We consider heterotrophic-autotrophic competition with oxygen as common substrate as in Ref. [5]. Three dissolved substrates are taken into account: organic carbon S_1, ammonia S_2, and oxygen S_3. Oxygen is used for both organic carbon oxidation and nitrification. Note that the waste products of the metabolic reactions are not explicitly modeled. The establishment and proliferation of X_2 strictly depend on the formation of an environmental niche, where the growth of heterotrophic bacteria X_1 is limited by the low concentration in organic carbon. Planktonic cells ψ_2 are considered for X_2 as the biofilm model is aimed at simulating the invasion of a constituted biofilm by autotrophic bacteria after the establishment of a favorable environmental niche.

The stoichiometry and the process rates required to close the model equations (Eqs. (2)–(7)), including the expressions for $r_{M,i}$, $r_{S,j}$, r_i and $r_{\psi,i}$, are taken from Refs. [40,42].

The biomass growth rates $r_{M,i}$ in Eq. (2) are given by

$$r_{M,1} = \left(\mu_{\max,1} \frac{S_1}{K_{1,1}+S_1} \frac{S_3}{K_{1,3}+S_3} - k_{d,1}\right) X_1, \tag{8}$$

$$r_{M,2} = \left(\mu_{\max,2} \frac{S_2}{K_{2,2}+S_2} \frac{S_3}{K_{2,3}+S_3} - k_{d,2}\right) X_2, \tag{9}$$

$$r_{M,3} = k_{d,1} X_1 + k_{d,2} X_2, \tag{10}$$

where $\mu_{\max,i}$ denotes the maximum net growth rate for the ith biomass, $K_{i,j}$ is the affinity constant of the jth substrate for the ith biomass, and $k_{d,i}$ represents the decay constant for the ith biomass. The specific growth rates induced by the switch of the planktonic cells to the sessile mode of growth, also required as inputs to Eq. (2), are defined as

$$r_1 = r_3 = 0, \tag{11}$$

$$r_2 = k_{col,2} \frac{\psi_2}{k_{\psi,2}+\psi_2} \frac{S_2}{K_{2,2}+S_2} \frac{S_3}{K_{2,3}+S_3}, \tag{12}$$

where $k_{col,2}$ corresponds to the maximum colonization rate of autotrophic bacteria, and where $k_{\psi,2}$ corresponds to the affinity-type constant for ψ_2.

The conversion rates for the three substrates required as inputs to Eq. (7) can be written as

$$r_{S,1} = -\frac{1}{Y_1} \mu_{\max,1} \frac{S_1}{K_{1,1}+S_1} \frac{S_3}{K_{1,3}+S_3} X_1, \tag{13}$$

$$r_{S,2} = -\frac{1}{Y_2} \mu_{\max,2} \frac{S_2}{K_{2,2}+S_2} \frac{S_3}{K_{2,3}+S_3} X_2, \tag{14}$$

$$\begin{aligned} r_{S,3} = &-\frac{1-Y_1}{Y_1} \mu_{\max,1} \frac{S_1}{K_{1,1}+S_1} \frac{S_3}{K_{1,3}+S_3} X_1 \\ &-\frac{4.57-Y_2}{Y_2} \mu_{\max,2} \frac{S_2}{K_{2,2}+S_2} \frac{S_3}{K_{2,3}+S_3} X_2, \end{aligned} \tag{15}$$

with Y_i denoting the yield of biomass i.

The conversion rate of the planktonic cells associated with the ith species, required as input to Eq. (6), is formulated as

$$r_{\psi,i} = -\frac{1}{Y_{\psi,i}} r_i, \tag{16}$$

with $Y_{\psi,i}$ being the yield of sessile species on planktonic ones. The terms $r_{\psi,i}$ represent the consumption rates of planktonic cells due to invasion process. $r_{\psi,i}$ are assumed proportional to r_i, meaning that they are described using the same Monod kinetics.

2.3. Simulation settings

To numerically solve the free boundary problem presented in Section 2.1 and 2.2, we use a straightforward extension of the numerical method proposed in Ref. [69]. The method of characteristics is used to track the biofilm expansion. Finite difference method is adopted to solve the diffusion-reaction equations. We extend this method to account for the new independent variables $\{\psi_i\}$, which account for invasion processes and which satisfy Eq. (6); $\{\psi_i\}$ are treated similarly as the variables $\{S_j\}$ characterizing dissolved substrates in Eq. (7). The solver is implemented in MATLAB.

In the present work, simulations are run for the target simulation time $T = 15$ days. The initial and boundary conditions associated with the free boundary problem are reported in Table 1.

3. Sources of uncertainty, quantities of interest and experimental designs

3.1. Functional output

The state of the biofilm evolves in time $t \in [0, T]$ and space $z \in [0, L(t)]$. The biofilm is characterized by biomass volume fractions, $f_i, i \in \{1, \dots, N_s\}$, and substrates $S_j, j \in \{1, \dots, N_m\}$, with $N_s = 3$ and $N_m = 3$ (see Section 2). Since the objective here is to analyze invasion mechanisms, we focus our attention on the species volume fractions f_i defined in Eq. (3).

Table 1
Initial-boundary conditions for biofilm growth, Eqs. (2)–(7).

Variable	Symbol	Value	Unit
Initial volume fraction of f_1	$\varphi_1(z)$	1.0	–
Initial volume Fraction of f_2	$\varphi_2(z)$	0.0	–
Initial volume Fraction of f_3	$\varphi_3(z)$	0.0	–
Bulk liquid concentration of S_1	S_1^*	3	$g_{COD}m^{-3}$
Bulk liquid concentration of S_2	S_2^*	13	$g_N m^{-3}$
Bulk liquid concentration of S_3	S_3^*	5	$g_{O_2}m^{-3}$
Bulk liquid concentration of ψ_1	ψ_1^*	0.0	$g_{COD}m^{-3}$
Bulk liquid concentration of ψ_2	ψ_2^*	1.0	$g_{COD}m^{-3}$
Initial biofilm thickness	L_0	300	μm
Initial concentration of S_1	$S_1(z, 0)$	0.0	$g_{COD}m^{-3}$
Initial concentration of S_2	$S_2(z, 0)$	0.0	$g_N m^{-3}$
Initial concentration of S_3	$S_3(z, 0)$	0.0	$g_{O_2}m^{-3}$
Initial concentration of ψ_1	$\psi_1(z, 0)$	0.0	$g_{COD}m^{-3}$
Initial concentration of ψ_2	$\psi_2(z, 0)$	0.0	$g_{COD}m^{-3}$

The quantities of interest could be in principle formulated as

$$y_i(t) = \frac{\displaystyle\int_0^{L(t)} f_i\,dz}{L(t)}, \quad i \in \{1, \ldots, N_s\}. \tag{17}$$

However, this choice would not show the spatial variability of the biofilm properties and would lead to an analysis of the different species as if the biofilm were concentrated in a single point. To better explore the spatial distribution of the biofilm species, the following discretization of the biofilm is proposed:

$$y_{ijk} = f_i(z_j, t_k), \quad i \in \{1, \ldots, N_s\}, \tag{18}$$

where the spatial discretization is given by $z_j = j\,\Delta z,\ \Delta z = L(t)/N_z$ and $j \in \{0, \ldots, N_z - 1\}$, and where the time discretization is given by $t_k = k\,\Delta t,\ \Delta t = T/N_t$ and $k \in \{0, \ldots, N_t - 1\}$.

In particular, we consider $N_t = 4$ times at which the biofilm extension is discretized into $N_z = 5$ locations. Note that the inert volume fraction f_3 is retrieved by mass conservation (Eq. (3)). Hence, the model output $\mathbf{y}$ is of functional type and includes the elements y_{ijk} with $i = \{1, 2\}$; $j = 1, \ldots, 5$; and $k = 1, \ldots, 4$ ($\mathbf{y} \in \mathbb{R}^n$ with $n = 40$) in the present study. This functional output is referred to as the "quantities of interest".

Note that the quantities of interest are considered as Lagrangian markers assigned to a relative position of the biofilm, whose spatial extent $L \equiv L(t)$ depends on time and on the biofilm model parameters (see Section 3.2).

3.2. Sources of uncertainty

In biological applications, a major source of uncertainty resides in the parameters associated with species or substrates. In the present modeling approach, parameters such as $\mu_{\max,i}$, $k_{d,i}$, $K_{i,j}$ and Y_i $(i = 1, \ldots, N_s,\ j = 1 \ldots N_m)$ are well characterized in Ref. [5] and are therefore assigned to reference values. We thus shift our attention to the parameters related to autotrophic bacteria biofilm invasion: $k_{col,2}$ and $k_{\psi,2}$ involved in r_2 in Eq. (12) to model the growth rate of autotrophic bacteria in sessile mode on the one hand, and $Y_{\psi,2}$ involved in Eq. (16) to model the consumption rate of planktonic cells denoted by $r_{\psi,2}$ on the other hand. Hereafter, $k_{col,2}$, $k_{\psi,2}$ and $Y_{\psi,2}$ are respectively denoted by k_{col}, k_ψ and Y_ψ for clarity purposes. The uncertain input vector $\boldsymbol{\theta}$ is thus defined as

$$\boldsymbol{\theta} = \left(k_{\mathrm{col}}, k_\psi, Y_\psi\right) \in \mathbb{R}^3. \tag{19}$$

such that the problem dimension is $d = 3$, see Table 2.

These parameters are not well characterized in literature and their determination still requires an accurate experimental activity based on ad-hoc techniques. In this work, we consider stochastic methods to represent input uncertainty. Thus,

Table 2
Uniform marginal PDF associated with k_{col}, k_ψ and Y_ψ. Note that $\mathcal{U}(a, b)$ stands for the uniform distribution with a the minimum value of the parameter and b the maximum one.

Parameter	Uniform distribution
k_{col}	$\mathcal{U}(10^{-4}, 10^{-2})$
k_ψ	$\mathcal{U}(10^{-5}, 10^{-2})$
Y_ψ	$\mathcal{U}(10^{-5}, 10^{-3})$

the uncertain input parameters are modeled by a random vector $\mathbf{\Theta}$, meaning that their values are supposed to depend on a random parameter ω such that $\mathbf{\Theta} \equiv \mathbf{\Theta}(\omega)$. ω is to be taken from the set of all outcomes Ω, which is equipped with a $\sigma-$algebra $\mathcal{S}$ and a probability measure $\mathcal{P}$. The triplet $(\Omega, \mathcal{S}, \mathbb{P})$ forms a probabilistic space [31].

The functional output $\mathbf{y}$ is considered as an element of $L^2(\Omega, \mathcal{S}, \mathbb{P})$ and is therefore represented as a vector of stochastic process, i.e.

$$\mathbf{Y}(\omega) = \mathcal{F}(\mathbf{\Theta}(\omega)), \tag{20}$$

with $\mathcal{F}$ the mapping of the input parameters onto the space of the functional output given by the biofilm model (see Eq. (1)).

Stochastic methods require to characterize the probability density function (PDF) associated with the input random vector $\mathbf{\Theta}$ denoted by $\boldsymbol{\rho}_{\mathbf{\Theta}}$. We need to introduce some assumptions on the nature of such uncertainty sources. First, we assume the components of $\mathbf{\Theta}$ are independent. Second, we consider uniform marginal PDF for each random variable Θ_i $(i = 1, \ldots, d)$ in $\mathbf{\Theta}$, denoted by ρ_{Θ_i}. The following restrictions apply: $k_{\text{col}} > 0$, $k_\psi > 0$ and $Y_\psi \in [0;\ 1]$; see Table 2. The objective here is to analyze under uncertainty, the relation between inputs $\mathbf{\Theta}$ and outputs $\mathbf{Y}$ and to build an emulator of the relation $\mathcal{F}$ in Eq. (20).

3.3. Experimental designs and databases

A design of experiments refers to the way of discretizing the uncertainty space (or “hypercube”) $Z_\Theta \in \mathbb{R}^d$ $(d = 3)$, in which the three parameters k_{col}, k_ψ and Y_ψ evolve. It is a way to define the N realizations of parameters $\boldsymbol{\theta}$, for which the biofilm model is integrated as a “black-box” to obtain the ensemble of N functional outputs $\mathbf{y}$ from which statistics can be derived. This ensemble forms a database $\mathcal{D}_N$:

$$\mathcal{D}_N = \left\{ \left(\boldsymbol{\theta}^{(l)}, \mathbf{y}^{(l)} \right)_{1 \le l \le N} \right\}, \tag{21}$$

where $\mathbf{y}^{(l)} = \mathcal{F}\left(\boldsymbol{\theta}^{(l)}\right)$ stands for the integration of the biofilm model $\mathcal{F}$ associated with the lth set of input parameters $\boldsymbol{\theta}^{(l)}$.

In the present study, two databases of size $N = 216$ are compiled using quasi-Monte Carlo sampling methods. They rely on low-discrepancy sequences to explore the hyperspace given by the support of the three PDFs without any bias and to capture most of the variance [70]. The first database built using Halton’s sampling serves as a training set and corresponds to the ensemble of simulations over which the surrogates are trained (Fig. 1(a)). The second database built using Faure’s sampling serves as a validation set and corresponds to the ensemble of simulations that is not part of the experimental design and that is used to evaluate the surrogate accuracy (Fig. 1(b)).

Note that the biofilm model features high nonlinearities. Fig. 2 presents 10 representative biofilm model snapshots at different times, (a) 5 days, (b) 10 days and (c) 15 days. The spatial distribution of the heterotrophic bacterial volume fraction f_1 is represented for each time, each line corresponds to a different realization of input parameters $\boldsymbol{\theta} = \left(k_{\text{col}}, k_\psi, Y_\psi\right)$ that is a point of the Halton’s low-discrepancy sequence presented in Fig. 1(a) and each line is colored with respect to the autotrophic bacterial volume fraction f_2. The biofilm length $L(t)$ effectively varies with time from 0.0010 to 0.0016 m.

4. Surrogate modeling

We present now the methodology to build an emulator of the biofilm model, using gPC-expansion or GP-model. The common idea of both approaches is to design for each quantity of interest Y in the vector $\mathbf{Y}$ $(Y \equiv Y_{ijk})$ a surrogate by means of a weighted (finite) sum of basis functions:

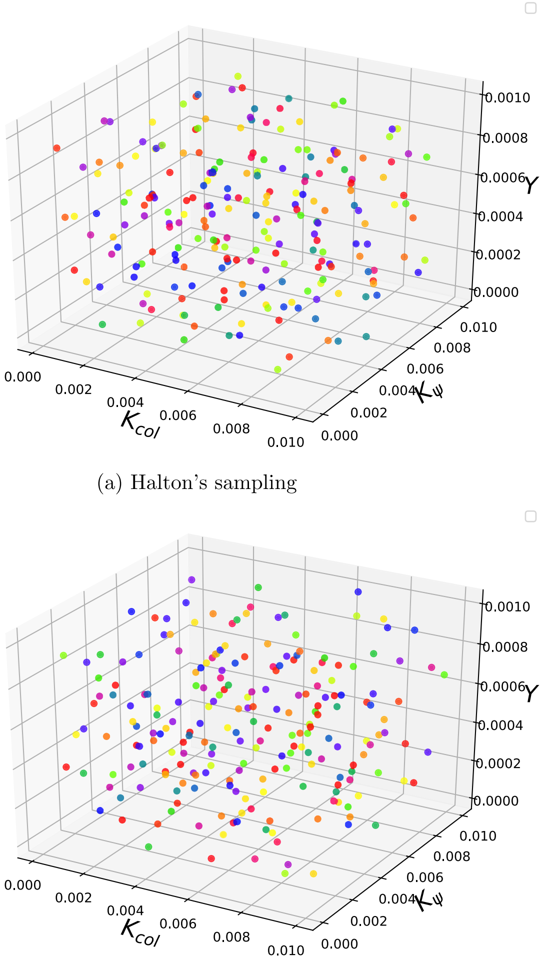
$$Y = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} \gamma_{\boldsymbol{\alpha}}\, \Psi_{\boldsymbol{\alpha}}, \tag{22}$$

where the coefficients $\{\gamma_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathcal{A}}$ and the basis functions $\{\Psi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathcal{A}}$ are calibrated using the information provided by the Halton’s training set $\mathcal{D}_N$ with $N = 216$ (see Section 3.3).

4.1. Generalized polynomial chaos (gPC) expansion

4.1.1. Standard probabilistic space

$\mathbf{\Theta}$ is defined in the input physical space and its counterpart in the standard probabilistic space is noted $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_d)$, with ζ_i the random variable associated with the ith uncertain parameter Θ_i in $\mathbf{\Theta}$ and characterized by a uniform marginal PDF ρ_{Θ_i}. The reduced variable ζ_i is therefore a uniform variable on $[-1;\ 1]$. The gPC-framework applies to the standard probabilistic space. The equivalent of $\boldsymbol{\rho}_{\mathbf{\Theta}}$ in the standard probabilistic space is denoted by $\boldsymbol{\rho}_{\boldsymbol{\zeta}}$. Since all input random variables are assumed independent (see Section 3.2), the joint PDF $\boldsymbol{\rho}_{\boldsymbol{\zeta}}$ is the product of the marginal PDFs $\{\rho_{\zeta_i}\}_{i=1,\ldots,d}$.



(a) Halton's sampling

(b) Faure's sampling

Fig. 1. Cloud representation of the two databases $\mathcal{D}_N$ with $N = 216$, corresponding to different sets of the three parameters k_{col} (x-axis), k_ψ (y-axis) and Y_ψ (z-axis). The two databases correspond to low-discrepancy sequences, (a) Halton's sampling (training set) and (b) Faure's sampling (validation set).

4.1.2. Polynomial basis

$\boldsymbol{\Theta}$ is projected onto a stochastic space spanned by the multivariate orthonormal polynomial functions $\{\Psi_{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\}_{\boldsymbol{\alpha}\in\mathcal{A}}$, with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$ a multi-index. This basis of polynomials is built with respect to the input joint PDF $\boldsymbol{\rho}_{\boldsymbol{\zeta}}$. The corresponding inner product is defined as

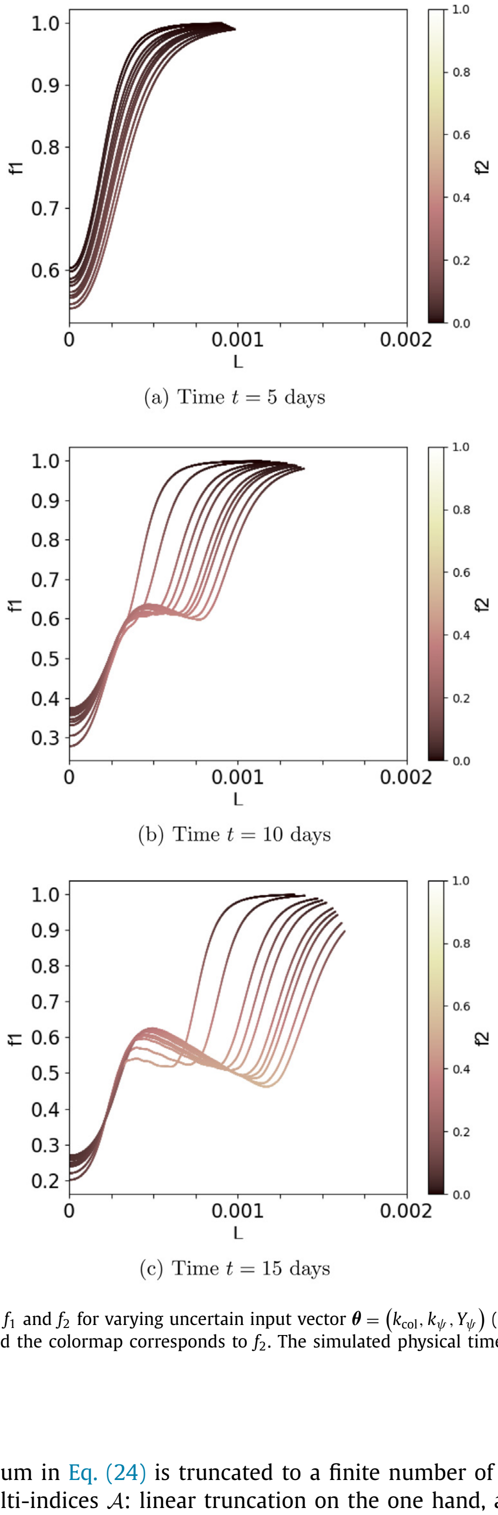
$$\langle \Psi_{\boldsymbol{\alpha}}(\boldsymbol{\zeta}), \Psi_{\boldsymbol{\beta}}(\boldsymbol{\zeta}) \rangle = \int_{Z_\zeta} \Psi_{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\, \Psi_{\boldsymbol{\beta}}(\boldsymbol{\zeta})\, \boldsymbol{\rho}_{\boldsymbol{\zeta}}\, d\boldsymbol{\zeta} = \delta_{\boldsymbol{\alpha}\boldsymbol{\beta}}, \tag{23}$$

with $\delta_{\boldsymbol{\alpha}\boldsymbol{\beta}}$ the Kronecker delta-function and $Z_\zeta \subseteq \mathbb{R}^d$ the normalized space in which $\boldsymbol{\zeta}$ evolves. In practice, the orthogonal basis is built using the tensor product of univariate polynomial functions, $\Psi_{\boldsymbol{\alpha}} = \psi_{\alpha_1} \ldots \psi_{\alpha_d}$ with ψ_{α_i} the one-dimensional polynomial function associated with ζ_i.

We assume the model outputs are of finite variance. Hence, Y can be cast as a function of the reduced variables and expanded as

$$Y(\omega) = \mathcal{F}_{\rm pc}(\boldsymbol{\Theta}) = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \gamma_{\boldsymbol{\alpha}}\, \Psi_{\boldsymbol{\alpha}}(\boldsymbol{\zeta}(\omega)), \tag{24}$$

where $\{\Psi_{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\}_{\boldsymbol{\alpha}\in\mathcal{A}}$ correspond to Legendre polynomials (this is the optimal choice for uniform PDFs according to Askey's scheme [71]); the total polynomial order is noted P. A truncation strategy is required to determine the appropriate size of the polynomial basis. Then $\{\gamma_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathcal{A}}$ are the unknowns to determine using a projection strategy to derive the emulator $\mathcal{F}_{\rm pc}$.



(a) Time $t = 5$ days

(b) Time $t = 10$ days

(c) Time $t = 15$ days

Fig. 2. Time-evolving species volume fractions f_1 and f_2 for varying uncertain input vector $\boldsymbol{\theta} = \left(k_{\text{col}}, k_\psi, Y_\psi\right)$ (Eq. (19)). The x-axis corresponds to the biofilm thickness $L(t)$; the y-axis corresponds to f_1; and the colormap corresponds to f_2. The simulated physical time is (a) 5 days, (b) 10 days and (c) 15 days.

4.1.3. Truncation strategy

For computational purposes, the sum in Eq. (24) is truncated to a finite number of terms r. We compare two truncation strategies to obtain a finite set of multi-indices $\mathcal{A}$: linear truncation on the one hand, and sparse truncation strategy on the other hand.

Linear truncation strategy. The standard truncation strategy consists in retaining in the gPC-expansion all polynomials involving the d random variables of total degree less or equal to P. Hence, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_d) \in \{0, 1, \cdots, P\}^d$. The number of terms is therefore constrained by the number of input random variables d and by the total polynomial order P so that

$$r_{\text{lin}} = (d + P)!/(d!\ P!). \tag{25}$$

The corresponding set of multi-indices $\mathcal{A}_{\text{lin}}$ is defined as

$$\mathcal{A}_{\text{lin}} \equiv \mathcal{A}_{\text{lin}}(d, P) = \{\boldsymbol{\alpha} \in \mathbb{N}^d : |\boldsymbol{\alpha}| \leq P\}, \tag{26}$$

where $|\boldsymbol{\alpha}| = ||\boldsymbol{\alpha}||_1 = \alpha_1 + \cdots + \alpha_d$ is the total order of the multi-index. In this case, we refer to the basis as the “full basis” for a given order P.

Sparse truncation strategy. A sparse truncation strategy consists in reducing the number of terms in the gPC-expansion for a given total polynomial order P. One way to build a “sparse basis” (by opposition to the “full basis” obtained when considering a linear truncation strategy) is the LAR approach. The key idea of the LAR approach is to select at each iteration, a polynomial among the r terms of the full basis based on the correlation of the polynomial term with the current residual; the selected term is added to the active set of polynomials. The coefficients of the active basis are computed so that every active polynomial is equicorrelated with the current residual until convergence is reached. Thus, LAR builds a collection of surrogates that are less and less sparse along the iterations. Iterations stop either when the full basis has been looked through or when the maximum size of the training set N has been reached. More details can be found in Refs. [64,72,73].

4.1.4. Projection strategy

In this work, for a given basis, we compute the coefficients $\{\gamma_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathcal{A}}$ through least-square minimization in a non-intrusive way, using the N-snapshots from the training set $\mathcal{D}_N$. The key idea of least-square minimization is to minimize the mean square error, i.e. the approximation error between the (exact) biofilm model evaluations and the gPC-surrogate estimations at the points of the training set [74].

The unknown coefficients are gathered into a vector $\widehat{\boldsymbol{\gamma}} = \{\gamma_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathcal{A}}$. $\widehat{\boldsymbol{\gamma}}$ is the solution of the following problem:

$$\widehat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}\in\mathbb{R}^r}{\text{argmin}} \, \frac{1}{N} \sum_{l=1}^{N} \left(y^{(l)} - \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \gamma_{\boldsymbol{\alpha}} \, \Psi_{\boldsymbol{\alpha}}\left(\boldsymbol{\xi}^{(l)}\right) \right)^2, \tag{27}$$

which is solved through classical linear algebra algorithms, i.e.

$$\widehat{\boldsymbol{\gamma}} = (\boldsymbol{\Psi}^T \boldsymbol{\Psi})^{-1} \, \boldsymbol{\Psi}^T \, \mathcal{Y}, \tag{28}$$

with $\boldsymbol{\Psi}$ the information matrix corresponding to the evaluation of the basis polynomials at each point of the experimental design $\mathcal{D}_N$, i.e. $\boldsymbol{\Psi} = \{\Psi_{\boldsymbol{\alpha}}(\boldsymbol{\zeta}^{(l)})\}_{\alpha\in\mathcal{A},1\leq l\leq N}$, and with $\mathcal{Y}$ the corresponding biofilm model evaluations.

When using non-sparse truncation, this projection method is referred to as the standard least-square (SLS) approach. In the LAR sparse method, least-square minimization is used to compute the set of active coefficients. Note that LAR allows the gPC-expansion to include high-order polynomials in the basis without generating an ill-posed problem and provides a way to explore the possible nonlinearity of the model response to the input parameters.

4.1.5. Workflow

The algorithm relative to the construction of the gPC-expansion can be described as follows:

1. choose the polynomial basis $\{\Psi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathcal{A}}$ according to the prescribed marginal PDFs of the inputs $\boldsymbol{\theta} = \left(k_{\text{col}}, k_{\psi}, Y_{\psi}\right) \in \mathbb{R}^3$ ($d = 3$);
2. choose the total polynomial order P according to the complexity of the biological processes;
3. truncate the gPC-expansion to r_{lin} terms corresponding to the multi-index set $\mathcal{A}_{\text{lin}}$ using linear truncation according to the problem dimension d and the total polynomial order P;
4. in the specific case of LAR, find a suitable set of multi-indices $\mathcal{A} \subset \mathcal{A}_{\text{lin}}$ with a cardinality $r \leq r_{\text{lin}}$, otherwise $\mathcal{A} = \mathcal{A}_{\text{lin}}$ and $r = r_{\text{lin}}$;
5. apply least-square minimization to compute the coefficients $\{\gamma_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathcal{A}}$ using $N = 216$ snapshots from the simulation database $\mathcal{D}_N$ (the experimental design is based on Halton’s low-discrepancy sequence);
6. formulate the surrogate $\mathcal{F}_{\text{pc}}$, which can be evaluated for any new pair of parameters $\boldsymbol{\theta}^* = \left(k^*_{col}, k^*_{\psi}, Y^*_{\psi}\right)$.

4.2. Gaussian Process (GP) surrogate

4.2.1. Principles

A surrogate model using GP regression can be cast as follows:

$$Y\left(\boldsymbol{\theta}\right) = \mathcal{F}_{\text{gp}}(\boldsymbol{\theta}) = \sum_{\boldsymbol{\alpha}=1}^{r} \gamma_{\boldsymbol{\alpha}} \, \Psi_{\boldsymbol{\alpha}}(\boldsymbol{\theta}), \tag{29}$$

where $\Psi_{\boldsymbol{\alpha}}$ is a GP calibrated from the training set $\mathcal{D}_N$. This GP is a random process indexed over the domain $\mathbb{R}^3$ (here $d = 3$), for which any finite collection of process values, $\{\Psi_{\boldsymbol{\alpha}}(\boldsymbol{\theta}^{(l)})\}_{1\leq l\leq N}$, share a joint Gaussian distribution [75]. Let $\widetilde{\Psi}_{\boldsymbol{\alpha}}$ be a GP fully described by its zero mean and its correlation $\pi_{\boldsymbol{\alpha}}$:

$$\widetilde{\Psi}_{\boldsymbol{\alpha}}(\boldsymbol{\theta}) \sim \text{GP}\left(0, \sigma_{\boldsymbol{\alpha}}^2 \, \pi_{\boldsymbol{\alpha}}(\boldsymbol{\theta}, \boldsymbol{\theta}')\right), \tag{30}$$

with $\pi_{\boldsymbol{\alpha}}(\boldsymbol{\theta},\boldsymbol{\theta}') = \mathbb{E}\left[\widetilde{\Psi}_{\boldsymbol{\alpha}}(\boldsymbol{\theta})\widetilde{\Psi}_{\boldsymbol{\alpha}}(\boldsymbol{\theta}')\right]$. In the present study, the correlation function π (or kernel) is chosen as a squared exponential (also known as radial basis function – RBF):

$$\pi_{\boldsymbol{\alpha}}(\boldsymbol{\theta},\boldsymbol{\theta}') = \exp\left(-\frac{\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^2}{2\,\ell_{\boldsymbol{\alpha}}^2}\right), \tag{31}$$

where $\ell_{\boldsymbol{\alpha}}$ is a length-scale describing the model dependency between the input vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, and where $\sigma_{\boldsymbol{\alpha}}^2$ is the model output variance. In this framework, the surrogate is obtained as the mean of the GP resulting of conditioning $\widetilde{\Psi}_{\boldsymbol{\alpha}}$ by the training set $\{\Psi_{\boldsymbol{\alpha}}(\boldsymbol{\theta}^{(l)})\}_{1\le l\le N}$. For any $\boldsymbol{\theta}^* \in \mathbb{R}^d$, the prediction of the GP-model can be obtained using Eq. (29) based on the following formulation for the basis function $\Psi_{\boldsymbol{\alpha}}$:

$$\Psi_{\boldsymbol{\alpha}}(\boldsymbol{\theta}^*) = \sum_{l=1}^{N} \boldsymbol{\beta}_{l,\boldsymbol{\alpha}}\,\pi_{\boldsymbol{\alpha}}\left(\boldsymbol{\theta}^*,\boldsymbol{\theta}^{(l)}\right), \tag{32}$$

where

$$\boldsymbol{\beta}_{l,\boldsymbol{\alpha}} = \left(\boldsymbol{\Pi}_{\boldsymbol{\alpha}} + \tau^2\,\mathbf{I}_N\right)^{-1}\left(\Psi_{\boldsymbol{\alpha}}\left(\boldsymbol{\theta}^{(1)}\right)\ldots\Psi_{\boldsymbol{\alpha}}\left(\boldsymbol{\theta}^{(N)}\right)\right)^T, \tag{33}$$

$$\boldsymbol{\Pi}_{\boldsymbol{\alpha}} = \left(\pi_{\boldsymbol{\alpha}}\left(\boldsymbol{\theta}^{(l)},\boldsymbol{\theta}^{(m)}\right)\right)_{1\le l,m\le N}, \tag{34}$$

and where τ (nugget effect) avoids ill-conditioning issues for the matrix $\boldsymbol{\Pi}_{\boldsymbol{\alpha}}$. The hyperparameters $\{\ell_{\boldsymbol{\alpha}},\ \sigma_{\boldsymbol{\alpha}},\ \tau\}$ are optimized by maximum likelihood applied to the dataset $\mathcal{D}_N$ using the DIRECT (the DIviding RECTangles) algorithm for global optimization [76].

4.2.2. Workflow

The algorithm relative to the construction of the GP-model can be described as follows:

1. choose the kernel function $\pi_{\boldsymbol{\alpha}}$ suitable for the input vector $\boldsymbol{\theta} = \left(k_{\text{col}}, k_{\psi}, Y_{\psi}\right) \in \mathbb{R}^3$ $(d=3)$ – we consider RBF in the present study, see Eq. (31);
2. optimize the GP-hyperparameters $\{\ell_{\boldsymbol{\alpha}},\ \sigma_{\boldsymbol{\alpha}},\ \tau\}$ associated with the kernel $\pi_{\boldsymbol{\alpha}}$ using maximum likelihood;
3. formulate the surrogate $\mathcal{F}_{\text{gp}}$, which can be evaluated for any new pair of parameters $\boldsymbol{\theta}^* = \left(k_{col}^*, k_{\psi}^*, Y_{\psi}^*\right)$ using Eqs. (29) and (32).

4.3. Numerical implementation

In practice, the implementation of the gPC-expansion and GP-model relied on the *OpenTURNS* [77] Python package (see http://www.openturns.org); *batman* [78] was used to build Halton's and Faure's datasets.

5. Results

5.1. A posteriori error estimation of the surrogate models

The construction of the surrogate model eventually introduces an approximation error, which can be computed *a posteriori* as

$$\epsilon_{\text{emp}} = \frac{1}{N_{\text{halton}}}\sum_{l=1}^{N_{\text{halton}}}\left(y^{(l)} - \widehat{y}^{(l)}\right), \tag{35}$$

with $y^{(l)}$ the lth element of the Halton's training set, $\widehat{y}^{(l)}$ the corresponding prediction by the (gPC or GP) surrogate, and $N_{\text{halton}} = 216$. This error estimator suffers from overfitting issues and may severely underestimate the actual mean square error [63]. Moreover, the GP-model can be regarded as an interpolator method at the points of the training set and will always achieve $\epsilon_{\text{emp}} = 0$ (when no noise is considered in the kernel). Note that in the following, for any tested configuration, we have $\epsilon_{\text{emp}} < 10^{-4}$.

To overcome these issues, we validate the surrogates using the Q_2 predictive coefficient that corresponds to a cross-validation error metric using the independent dataset based on Faure's low discrepancy sequence:

$$Q_2 = 1 - \frac{\sum_{l=1}^{N_{\text{faure}}}\left(y^{(l)} - \widehat{y}^{(l)}\right)^2}{\sum_{l=1}^{N_{\text{faure}}}\left(y^{(l)} - \bar{y}\right)^2}, \tag{36}$$

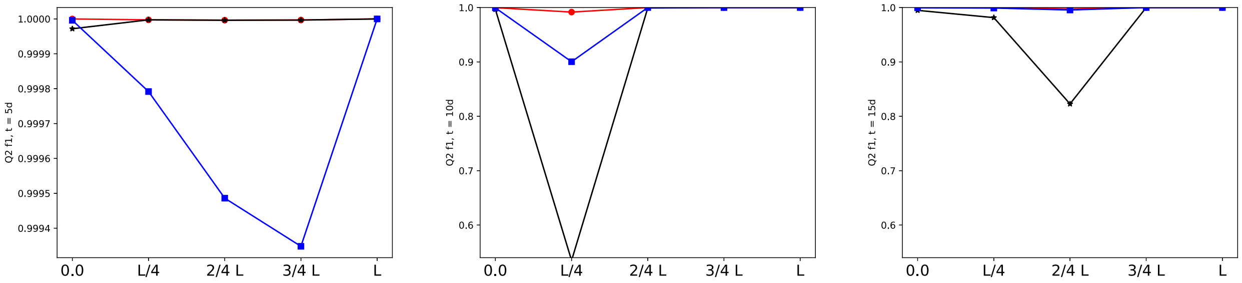


Fig. 3. Q_2 predictive coefficient along the biofilm thickness $L \equiv L(t)$ at three different time steps: 5 days, 10 days and 15 days (from left to right panels); Halton's experimental design is used as the training set with $N = 216$. Comparison of SLS-based gPC-expansion (black star line), LAR-based gPC-expansion (red dotted line), and RBF-based GP-model (blue squared line) for the species volume fraction f_1 associated with heterotrophic bacteria. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

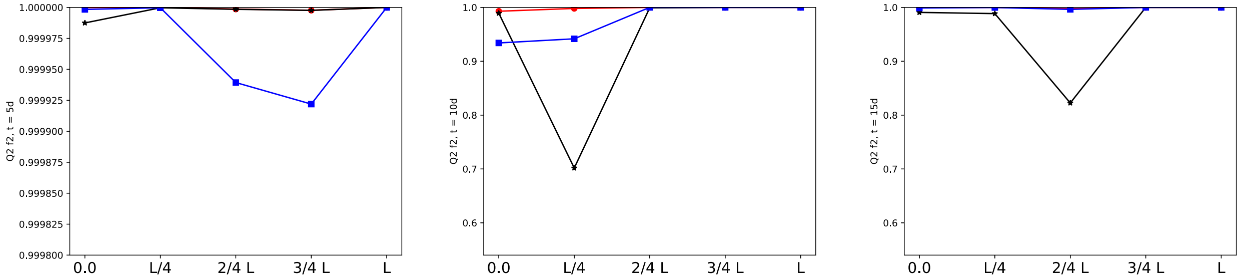


Fig. 4. Similar caption as Fig. 3 but for the species volume fraction f_2 associated with autotrophic bacteria.

with $\bar{y}$ the empirical mean over the Faure's validation set ($N_{\text{faure}} = 216$). Thus, Q_2 provides a normalized estimate of the generalization error, i.e. the error of the surrogate when considering points outside of the Halton's training set [53]. The target value for Q_2 is 1.

Figs. 3 and 4 present the Q_2 predictive coefficient along the biofilm after 5 days, 10 days and 15 days for three different surrogates: SLS-based gPC-expansion (black-star line); LAR-based gPC-expansion (red-dotted line); and RBF-based GP-model (blue-squarred line). Fig. 3 is obtained when considering the species volume fraction f_1 – heterotrophic bacteria – as model output; Fig. 4 is the counterpart of Fig. 3 for f_2 – autotrophic bacteria. Results show that the LAR gPC-expansion features the best performance with a Q_2 close to 1 over the whole time period and all along the biofilm thickness. The SLS gPC-expansion is subject to significant error after 10 days and 15 days, when the biological processes at play become more complex. Note that the minimum value for Q_2 moves along the biofilm over time, with Q_2 going down to 0.6 at $z \approx L/4$ after 10 days and 0.82 at $z = 2/4L$ after 15 days. The GP-model achieves intermediate accuracy between LAR-based gPC-expansion and SLS-based gPC-expansion; the corresponding Q_2 being at minimum equal to 0.9 when it reaches 0.6 for SLS-based gPC-expansion after 10 days. After 15 days both LAR-based gPC-expansion and GP-model feature similar performance.

Fig. 5 presents the polynomial terms that are retained in the LAR gPC-expansion built to emulate the species volume fraction f_1 at a particular location of the biofilm ($z = L(t)/4$); time evolution of these polynomial terms is presented (after 5, 10, 15 days). Note that we consider the case $z = L(t)/4$ since the LAR gPC-surrogate tends to outperform the SLS gPC-surrogate and the GP model at this location (see Fig. 3). Each active polynomial $\Psi_{\boldsymbol{\alpha}}$ is associated with a colored symbol, where the color represents the magnitude of the coefficient $\gamma_{\boldsymbol{\alpha}}$. The x-/y-/z-axis of the plots represent the degree of the polynomial. We observe that LAR offers some flexibility (due to the sparse structure of the polynomial basis) to integrate high-order polynomial terms in the gPC-expansion, in particular along the direction associated with the parameter k_{col} (x-axis), where polynomial degrees go up to 14 after 10 days. The full basis considered in the SLS gPC-surrogate cannot include these terms due to the limited size of the training set ($N = 216$, implying that $P \leq 5$). The increase in complexity of the biofilm structure with respect to time is evidenced by the increasing number of terms retained in the gPC-expansion over time.

In summary, the sparse truncation strategy underlying the LAR-based gPC-expansion seems to provide a clear advantage to build an emulator of the biofilm model. The magnitude and number of LAR gPC-coefficients give insight into the complexity of the biological processes occurring in multi-species biofilm; this complexity growing over time. The latter can only be captured by a flexible adaptative surrogate approach that identifies inline the required polynomial degree to accurately capture the system dynamics. The following analysis is therefore carried out using the standalone LAR approach.

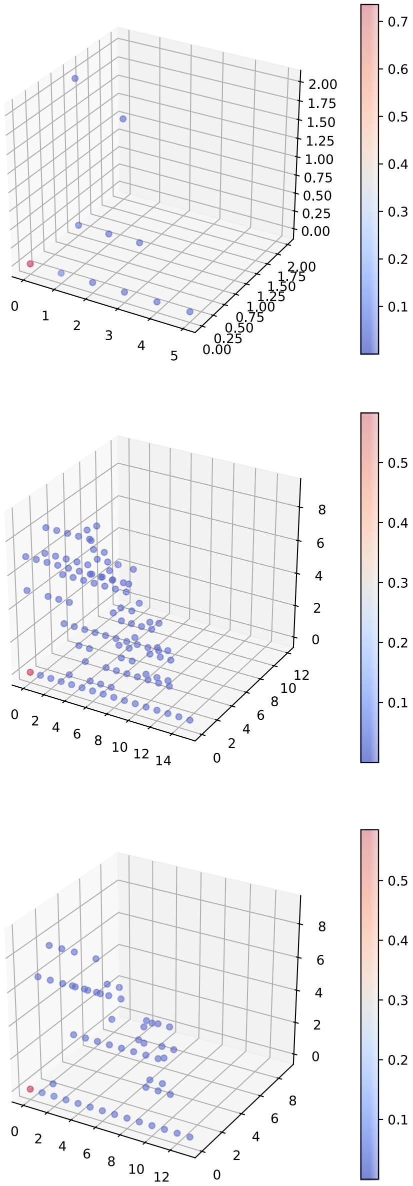


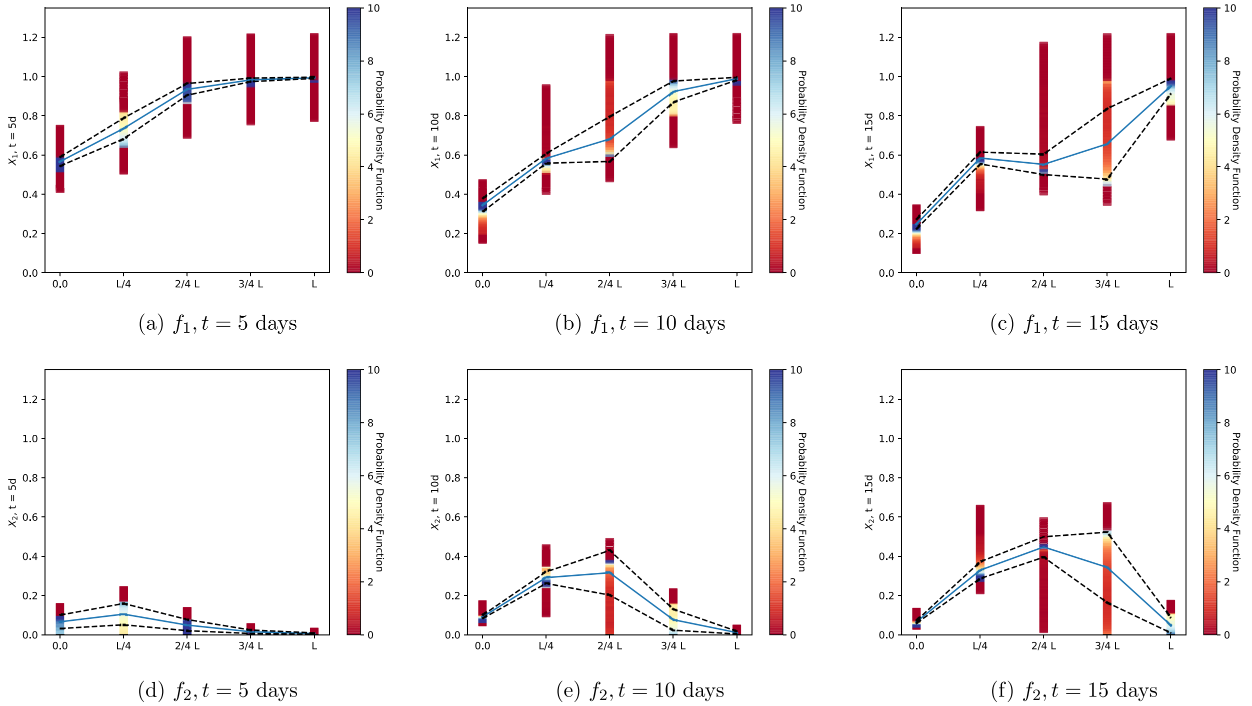
Fig. 5. Sparsity plots representing the magnitude of the LAR gPC-coefficients $\{\gamma_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathcal{A}}$ with respect to the three-dimensional input space, $\boldsymbol{\theta} = (k_{\text{col}}, k_{\psi}, Y_{\psi})$ ($d = 3$) and time evolution from 5 to 15 days (from top to bottom panels). x-, y- and z- axis correspond to the polynomial degrees of the gPC-expansion terms associated with k_{col}, k_{ψ} and Y_{ψ}, respectively. The gPC-expansion under consideration represents the model response for the species volume fraction f_1 (heterotrophic bacteria) at $z = L(t)/4$. The color of the symbols indicates the magnitude of the gPC-coefficients.

5.2. Uncertainty quantification of the biofilm model predictions

Using the LAR gPC-expansion, the statistics of each quantity of interest y can be derived analytically from the coefficients $\{\gamma_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathcal{A}}$. The mean value μ_y and STD σ_y of y can be estimated as

$$\mu_y = \gamma_0, \tag{37}$$

$$\sigma_y = \sqrt{\sum_{\substack{\boldsymbol{\alpha}\in\mathcal{A}\subset\mathbb{N}^d \\ \boldsymbol{\alpha}\neq 0}} \gamma_{\boldsymbol{\alpha}}^2}. \tag{38}$$



(a) $f_1, t = 5$ days

(b) $f_1, t = 10$ days

(c) $f_1, t = 15$ days

(d) $f_2, t = 5$ days

(e) $f_2, t = 10$ days

(f) $f_2, t = 15$ days

Fig. 6. Statistical moments and PDF of each model output $y_{ijk} = f_i(x_j, t_k)$ where i corresponds to the species index, j corresponds to the space index and k corresponds to the time index. The colormap represents the model output PDF at each location and time step. The solid line represents the mean value computed using Eq. (37). The dashed lines represent the STD computed using Eq. (38), $\mu_{\mathrm{y}} \pm \sigma_{\mathrm{y}}$.

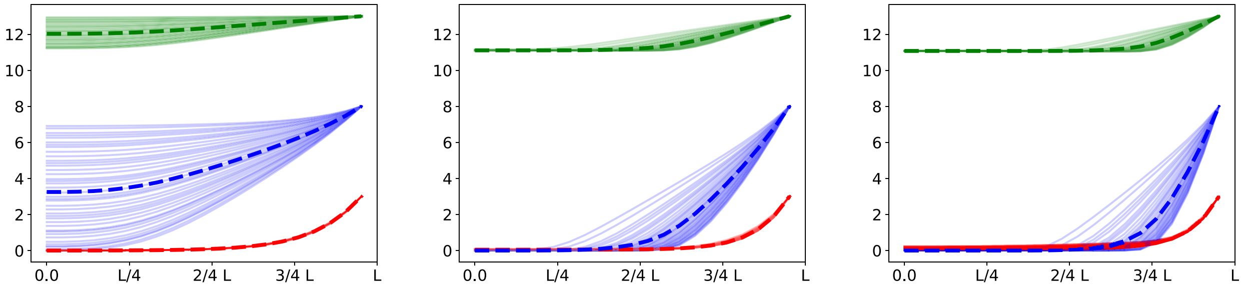


Fig. 7. Spatial and temporal evolution of the three substrates S_1 (red), S_2 (green) and S_3 (blue) from $z = 0$ μm to $z = L(t)$ after $t = 5, 10, 15$ days (from left to right panels). The thin solid lines correspond to 40 representative simulations of the biofilm model from Halton's training database. The dashed thick lines correspond to the sample means. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The PDF of each quantity of interest is retrieved through kernel smoothing techniques by sampling the uncertain input space Z_Θ using 10,000 members based on Monte Carlo random sampling and by evaluating the LAR gPC-expansion for all these points.

Fig. 6 presents the PDF of the species volume fractions f_1 and f_2 with respect to the biofilm thickness $L(t)$, along with the mean (solid line) and STD (dashed lines); each panel from left to right corresponds to a different time step over the 15-day time period under consideration. Results show that the uncertainty on the model output is driven rightwards as the simulation runs forward in time: after 5 days the largest variance is observed near $z = L(t)/4$ and moves to $z = 3/4, L(t)$ after 15 days. The same trend is observed for both species volume fractions f_1 and f_2.

The fact that the central part of the biofilm is subject to the highest level of uncertainty can be interpreted as the increase in complexity of the biofilm structure, which is correlated to the establishment of the invading species, is essentially due to the niche formation occurring far from the biofilm boundaries (substratum surface on the left and bulk liquid on the right). Recall that the adopted boundary conditions refer to a fixed bulk liquid concentration at $z = L(t)$ as well as a no-flux condition at $z = 0$ (see Table 1). Fig. 7 shows the trends for the three substrates S_j $(j = 1, \dots, 3)$ over time; the organic carbon S_1 and the oxygen S_3 feature a significantly reduced spread at the bottom of the biofilm, independently of the choice of the input vector $\boldsymbol{\theta}$. This is due to a combined effect of substrate diffusion and microbial metabolism, which leads to the

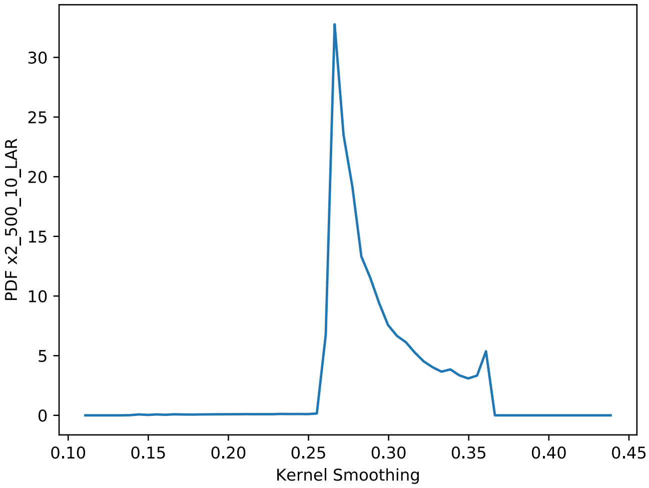


Fig. 8. Bimodal PDF of the autotrophic species mass fraction f_2 at location $z = L/4$ after 10 days obtained through kernel smoothing.

decrease of substrate concentration with respect to the constant value prescribed at the bulk liquid interface. More specifically, S_1 is mainly consumed in the outermost part of the biofilm and tends to become zero in the central part of the biofilm where the invading species finds favorable environmental conditions for its growth. Moreover, S_3 is completely depleted in the inner part of the biofilm and thus the microbial complexity due to the invasion process is significantly reduced at the bottom of the biofilm. Note that all the results have been obtained for a specific case study, reproducing a typical microbial interaction occurring in waste-water treatment plants, which is of relevant interest for engineering applications. Diverse boundary conditions may lead to different invasion processes and thereby to different uncertainty quantification results.

It is worth mentioning that some PDFs associated with f_1 and f_2 have more than one mode, see for instance Fig. 8 corresponding to the PDF of the autotrophic species volume fraction f_2 at $z = L/4$ after 10 days. This bimodal PDF has a physical explanation: for the given range of the input parameters under consideration, the autotrophic invasion at some location features two distinct behaviors, either a successful or unsuccessful niche formation. Ad-hoc simulations (data not shown) confirmed this switch from unsuccessful to successful colonization, mainly due to the adopted value of k_{col}.

5.3. Analysis of the biofilm structure

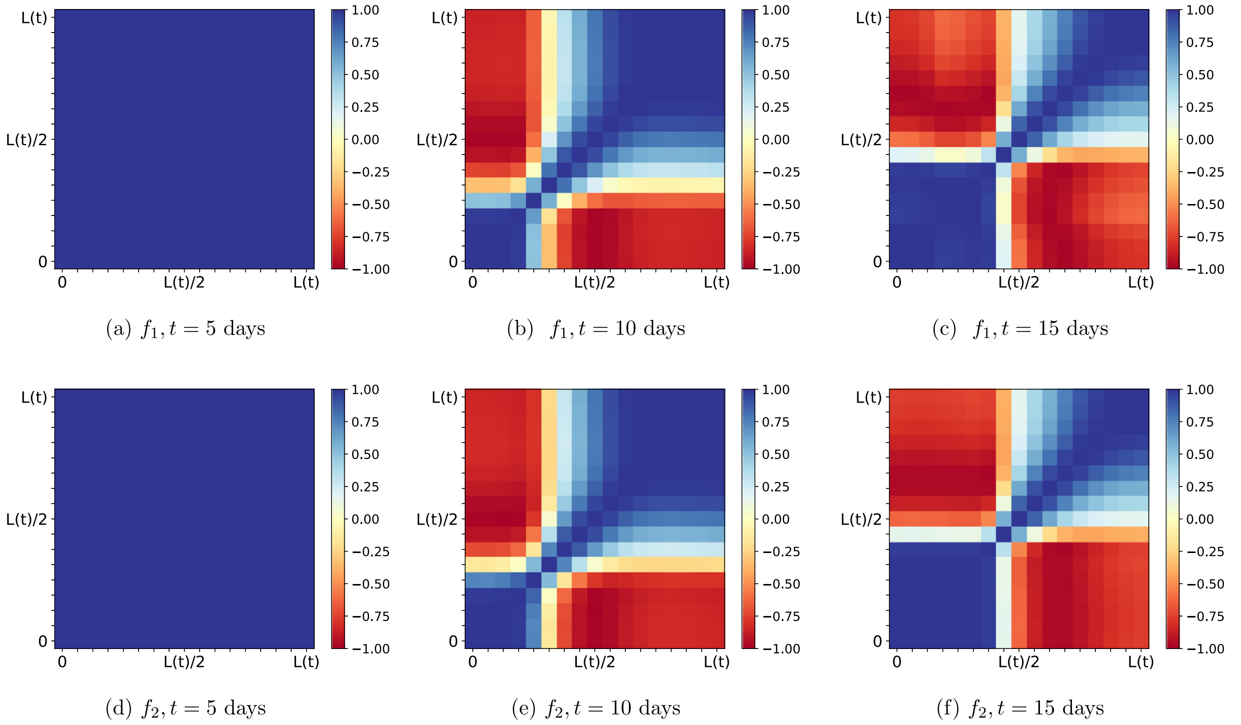
Using the Halton's training set, we can compute the covariance matrix $\mathbf{C}_{yy} \in \mathbb{R}^{N_z \times N_z}$, also known as dispersion matrix, to characterize the covariance between the model state $\mathbf{y}$ at different locations $z \in [0, L(t)]$ at a given time. $\mathbf{C}_{yy}$ can be empirically estimated as

$$\mathbf{C}_{yy} = \sum_{l=1}^{N} \frac{\left(\mathbf{y}_{ik}^{(l)} - \bar{\mathbf{y}}_{ik}\right)\left(\mathbf{y}_{ik}^{(l)} - \bar{\mathbf{y}}_{ik}\right)^T}{N-1}, \tag{39}$$

where $\mathbf{y}_{ik}^{(l)} = \{y_{ijk}^{(l)}\}_{j=1,\cdots,N_z}$ is the vector containing the ith quantity of interest y_{ijk} at a given time index k for the ensemble member l. In this matrix, the diagonal terms correspond to the variance of the model state variable at a given location j. The off-diagonal terms represent the covariances in the model state variable between two locations along the z-axis. The covariance matrix is symmetric by definition. By normalizing the covariance matrix by the variance, we can derive the correlation matrix shown in Fig. 9 (by definition diagonal terms are equal to 1). One column of the correlation matrix therefore provides the correlation function of a particular point with the rest of the z-axis.

Fig. 9 presents the evolution of the correlation matrix over the 15-day time period for both f_1 and f_2 state variables. Results show that at early times (after 5 days), the biofilm can be considered as a single entity with respect to its internal structure since the correlation factor is very high (above 0.99 for both f_1 and f_2). At later times, the internal structure becomes more complex and decorrelates. This evolution is due to the growth in spatial complexity of the biofilm, with the mechanism of autotrophic invasion that alters the species composition of the biofilm in a non-linear way via niche formation. This is inline with the complex structure of the LAR polynomial basis presented in Fig. 5, which includes for instance high-order polynomial terms in the three directions k_{col}, k_ψ and Y_ψ.

In summary, the spatial structure of the biofilm after 10 days seems to be organized as two main clusters: one related to the lack of substrates at $z = 0$ (the blue cluster at the bottom-left corner of the correlation matrix in Fig. 9), a second one related to the fixed bulk concentration of substrates at $z = L(t)$ (the blue cluster at the top-right of the correlation matrix in Fig. 9).



(a) $f_1, t = 5$ days

(b) $f_1, t = 10$ days

(c) $f_1, t = 15$ days

(d) $f_2, t = 5$ days

(e) $f_2, t = 10$ days

(f) $f_2, t = 15$ days

Fig. 9. Spatial correlation matrices for species volume fractions f_1 (top panels) and f_2 (bottom panels) evolving over time (5 days to 15 days from left to right panels) and computed using Halton's training set with $N = 216$.

5.4. Input-output sensitivity analysis

Sobol' indices [21,43] are commonly used for global sensitivity analysis based on variance decomposition. They provide the quantification of how much of the variance in the quantity of interest is due to the spread in the uncertain input parameters assuming these random variables are independent. The variance of the output random variable Y denoted by $\mathbb{V}[Y]$ can be decomposed as

$$\mathbb{V}[Y] = \sum_{i=1}^{d} \mathbb{V}_i(Y) + \sum_{j=i+1}^{d} \mathbb{V}_{ij}(Y) + \cdots + \mathbb{V}_{1,2,\ldots,d}(Y), \tag{40}$$

where $\mathbb{V}_i(Y) = \mathbb{V}[\mathbb{E}(Y|\Theta_i)]$, $\mathbb{V}_{ij}(Y) = \mathbb{V}\big[\mathbb{E}(Y|\Theta_i, \Theta_j)\big] - \mathbb{V}_i(Y) - \mathbb{V}_j(Y)$ and more generally,

$$\mathbb{V}_I(Y) = \mathbb{V}[\mathbb{E}(Y|\Theta_I)] - \sum_{J\subset I \,\text{s.t.}\, J\neq I} \mathbb{V}_J(Y), \ \forall I \subset \{1, \ldots, d\} \tag{41}$$

Based on this variance decomposition, the first-order Sobol' index S_i associated with the ith parameter of $\boldsymbol{\Theta}$ is given by

$$S_i = \frac{\mathbb{V}_i(Y)}{\mathbb{V}(Y)}, \tag{42}$$

and corresponds to the ratio of the output variance $\mathbb{V}(Y)$ that is uniquely due to the ith input parameter; S_i ranges between 0 and 1. The corresponding total Sobol' index S_{T_i} measures the whole contribution of the ith input parameter (including interaction with other parameters of $\boldsymbol{\Theta}$) on the output variance, with the following definition:

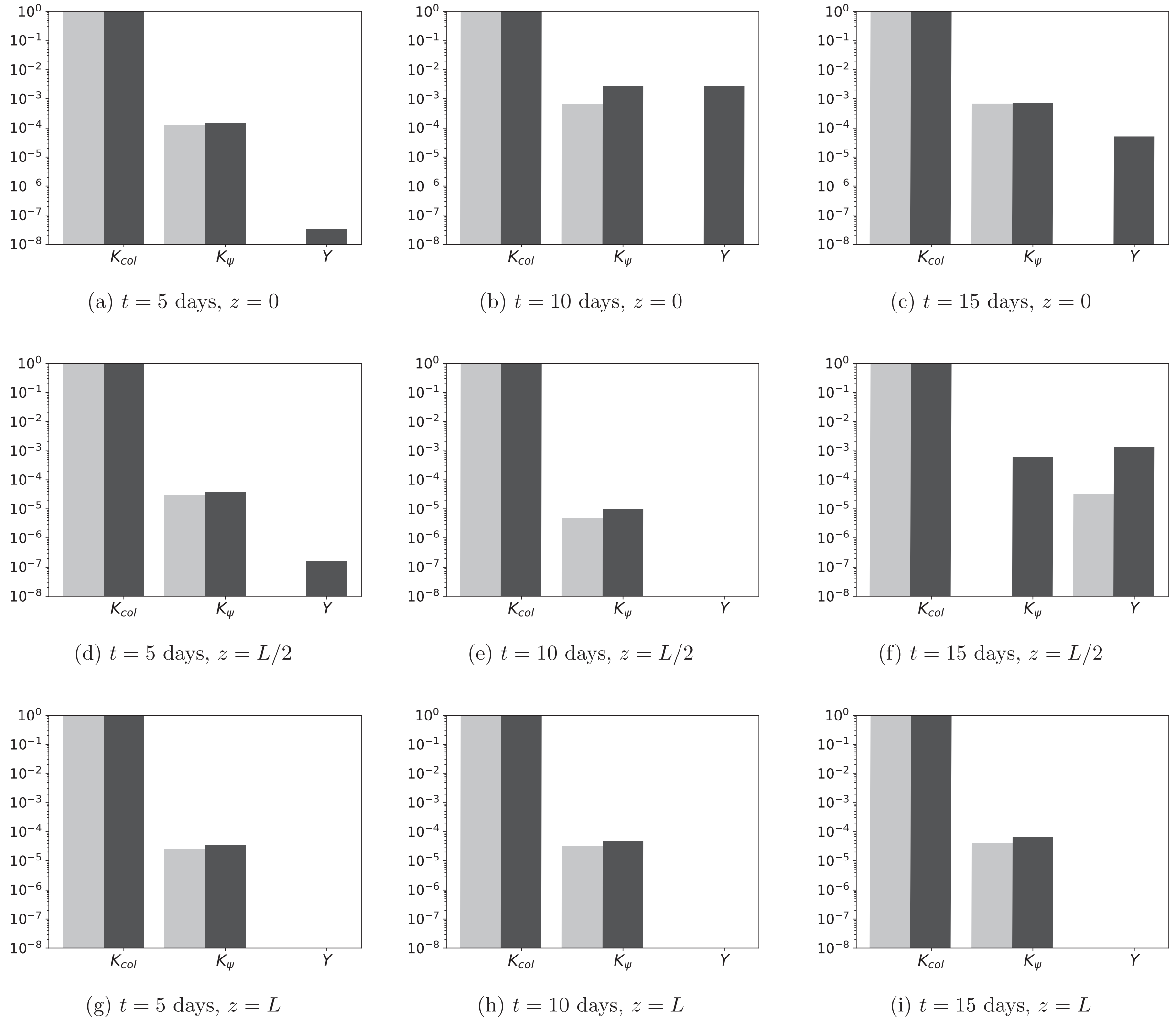
$$S_{T_i} = \sum_{\substack{I\subset\{1,\ldots,d\} \\ I\ni i}} S_I. \tag{43}$$

By definition, $S_{T_i} \geq S_i$. If both first-order and total indices are not equal, this indicates that the input parameter Θ_i has some interactions with other parameters of $\boldsymbol{\Theta}$ to explain the output variance.

In practice, for the LAR gPC-expansion, the first-order and total Sobol' indices are directly derived from the gPC-coefficients, for instance the first-order Sobol index reads

$$S_{i,\text{pc}} = \frac{1}{\sigma_y^2} \sum_{\substack{\boldsymbol{\alpha}\in\mathcal{A}, \\ \alpha_i>0 \ \text{and} \ \alpha_{k\neq i}=0}} \gamma_{\boldsymbol{\alpha}}^2, \tag{44}$$

with σ_y the output STD computed using Eq. (38).



(a) $t = 5$ days, $z = 0$

(b) $t = 10$ days, $z = 0$

(c) $t = 15$ days, $z = 0$

(d) $t = 5$ days, $z = L/2$

(e) $t = 10$ days, $z = L/2$

(f) $t = 15$ days, $z = L/2$

(g) $t = 5$ days, $z = L$

(h) $t = 10$ days, $z = L$

(i) $t = 15$ days, $z = L$

Fig. 10. First-order and total Sobol' indices (in logarithmic scale) associated with uncertain parameters $\boldsymbol{\theta} = (k_{col}, k_\psi, Y_\psi)$ and species volume fraction f_2 (autotrophic bacteria). Time evolution from 5 to 15 days of biofilm growth is presented from left to right panels; spatial distribution along the biofilm thickness $(0 \le z \le L(t))$ is presented from top to bottom panels. For each panel, light gray colors correspond to first-order Sobol' indices; dark gray colors correspond to total Sobol' indices; and indices are presented in the following order from left to right bars: k_{col}, k_ψ, Y_ψ.

Fig. 10 presents the first-order and total Sobol' indices obtained with the LAR gPC-expansion related to the autotrophic bacteria volume fraction f_2. These indices are presented at different times $t \in \{5,\ 10,\ 15\,\text{days}\}$ (from left to right panels), and at different locations along the biofilm thickness $z \in \{0,\ L/2,\ L\}$ (from top to bottom panels).

Results clearly show the prevalence of the input parameter k_{col} with Sobol' indices close to 1 for all times and locations. From a physical viewpoint, k_{col} is therefore a key parameter to represent colonization by autotrophic species X_2 at the expense of heterotrophic species X_1. It reproduces the attitude of microrganisms to switch their state from planktonic to sessile. That is, k_{col} represents the key parameter for the invasion phenomenon to occur, so changes in Y_ψ and k_ψ have a negligible effect on the overall invasion process. The concurrent presence of planktonic species and specific environmental niches allows the invasion to occur only when the planktonic species are characterized by significant values of the colonization rate for the investigated simulation times. These results inform us about which measurements should be improved to use the invasion modeling for a better understanding of the colonization process overall.

This is inline with the high-order terms retained in the LAR polynomial basis in the direction of k_{col} (see Fig. 5). The total polynomial order of the sparse gPC-expansion is due to k_{col}: k_{col} is associated with polynomial terms of degrees up to $P = 14$ after 10 days and $P = 12$ after 15 days.

Note that similar sensitivity is observed along the biofilm thickness after 5 days (first column of panels in Fig. 10), which is consistent with the uniform correlation matrices obtained at the same time in Fig. 9 and the subsequent interpretation: the biofilm can be considered as a single entity at early times.

In complement, the sensitivity of the model output to the parameters k_ψ and Y_ψ is slightly higher after 15 days than after 5 days ($10^{-2}/10^{-3}$), in particular in the first portion of the biofilm ($z \geq L/2$). These results are also consistent with the two clusters observed in the correlation matrices after 15 days in Fig. 9. The biofilm is gaining in spatial complexity as time advances: more parameters with respect to the standalone k_{col} could act on the spatial distribution of the invading species. Results show that the input parameter k_ψ is usually more influential than Y_ψ, especially at $z = L$ (third row of panels in Fig. 10), even though the relevance of these parameters is of several orders of magnitude below that of k_{col} (about 10^{-4}). First-order and total Sobol' indices are not identical, implying that some interactions occur between the three parameters.

Note that at location $z = L,$ we obtain nearly constant Sobol' indices over time. This is due to the constant boundary conditions imposed at the bulk liquid interface. In contrast, in the central part of the biofilm (second row of panels in Fig. 10 corresponding to $z = L/2$), where the niche formation takes place, the sensitivity of the model output to Y_ψ becomes higher than that of k_ψ for long times.

6. Conclusions

In this work, uncertainty quantification and global sensitivity analysis non-intrusive methods were applied to a novel and promising multi-species microbial biofilm model, which explicitly accounts for bacterial invasion processes. Invasion can rapidly alter biofilm populations and could even result in the loss of the resident species. It is therefore a key biological process that requires deeper understanding to improve engineering design. For instance, the continuum biofilm model could be helpful to predict the optimal operational conditions (dilution rates, oxygen concentration, carbon addition, etc.), which favor the establishment of a specific microbial syntrophy between resident and invading species.

We considered here the invasion by autotrophic bacteria of a heterotrophic biofilm. Initially present in the bulk liquid, autotrophic bacteria infiltrate the biofilm, switch their state from planktonic mode to sessile mode and start to proliferate, where and when they meet the best environmental conditions to enhance their growth. Heterotrophic-autotrophic competition for oxygen is a well-known biological process, which occurs for instance in the aerobic units of waste-water treatment plants. Heterotrophic bacteria conventionally oxidize organic matter into carbon dioxide, while autotrophic bacteria convert ammonium into nitrite and nitrate. The successful contextual removal of organic carbon and ammonium depends on the establishment of a multi-species biofilm constituted by both the microbial species. The growth of autotrophic bacteria strongly depends on the formation of an environmental niche, where the heterotrophic bacteria are out-competed.

The simulation of these biological processes is directly affected by the choice of the biofilm boundary conditions as well as by the range of variation of the input parameters, in particular those related to the planktonic species. The present study focused on the sensitivity of the autotrophic and heterotrophic bacteria volume fractions to the parameters characterizing the colonization rate of autotrophic bacteria and the consumption rate of planktonic cells, i.e. $\boldsymbol{\theta} = (k_{col,2}, k_{\psi,2}, Y_{\psi,2}) \in \mathbb{R}^3$. This sensitivity has been measured here through the computation of spatial and temporal Sobol' indices using a cost-effective surrogate.

It is worth mentioning that Sobol' indices measure the relative contribution of a given parameter on the output variance among the perturbed parameters and of its possible interactions with other parameters. The sensitivity analysis results therefore depend on the choice of $\boldsymbol{\theta}$. The biofilm model may depend on a rather large set of parameters, even on those that were fixed to nominal values in this work. For this reason, the output variance obtained here is necessarily a fraction of the potential variance that could be measured for a fully randomized model.

We presented a detailed analysis of the surrogate performance for a given simulation budget N. Two families of surrogates, gPC-expansion and GP-model, were compared in terms of Q_2 predictive coefficient. One difficulty in building surrogates is the choice of the basis. In particular, for gPC-expansion, the choice of the total polynomial order P and of the basis components (full basis with all elements of degree less or equal to P, or sparse basis) is an essential step to insure the surrogate accurately represents the model response over the whole input parameter space. In the present test case, the LAR gPC-expansion was found to be the best emulator of the biofilm model over the different time snapshots and biofilm locations, the sparse basis providing more flexibility on the total polynomial order for each input parameter than the full basis. The sparse basis is then an asset to fit the nonlinear biological processes with a limited training set. A single global surrogate was enough to achieve the target Q_2 criterion for the LAR gPC-expansion.

This investigation carried out via the LAR gPC-expansion provided new insights into the biofilm invasion mechanisms.

First, the spatial correlation functions along the biofilm thickness highlighted the temporal changes in the biofilm structure: the young biofilm (after a few days) featured some homogeneity in its spatial structure but the mature biofilm (after ten-to-fifteen days of growth) lost spatial correlation due to the increase in complexity of the biological processes involving niche formation and ongoing resident/invading species competition.

In complement, Sobol' sensitivity indices highlighted the key role of $k_{col,2}$, which represents the maximum colonization rate of autotrophic bacteria and which outclasses by several orders of magnitude the contribution of $k_{\psi,2}$ (affinity-type constant for planktonic species associated with autotrophic bacteria) and $Y_{\psi,2}$ (yield of sessile species on planktonic ones for autotrophic bacteria). This prevalence of $k_{col,2}$ is not only related to its key role in regulating the switch from planktonic to sessile modes of growth, but also to the specific setting of the case study. A relative increase in the relevance of ($k_{\psi,2}$, $Y_{\psi,2}$) was noticed as biofilm increased in complexity over time.

Finally, the PDF and statistics of the biofilm state provided an interesting viewpoint on the biofilm structure and its temporal evolution. While the mean values retrieved autotrophic invasion trends already documented in Ref. [40], the present study found that the invading and resident species concentrated both their variance in the central part of the biofilm, far from the free boundary, where restrictive conditions on substrates have been imposed, and far from the inert surface, where lack of substrates limited the variability. The variance trends showed for both heterotrophic and autotrophic species, a shift in the location of the maximum spread towards the free boundary $L \equiv L(t)$ for increasing time t.

Uncertainty and global sensitivity analysis is found to be a promising way to identify the most influential parameters in any given regime or application scenario and to quantify their effects on the biofilm structure and evolution. More generally, this provides guidelines to orient further biofilm model developments and design in the long-term prediction capability that could answer some of the medical, environmental and industrial issues related to bacterial invasion. Further work might be related to the extension of the present analysis to more complex biological situations, which are related to the dispersal phenomenon and involve the modeling of planktonic species dynamics in multi-species biofilm.

The key idea of this work was to set a methodology to apply sensitivity analysis to biofilm modeling, with particular attention to the integration of new variables and parameters into existing models. Sparse surrogates are a way to address high-dimensional problems, in particular when the size of the training set is limited. So future work might extend the LAR-based analysis to a wider set of perturbed parameters to provide a more complete quantification of the output variance and a more general sensitivity analysis.

A meaningful follow-up will be to analyze model output sensitivity while varying the literature parameters as indicated by specific experimental and computational results, in order to assess the potential interactions among all the parameters and their effect on the invasion process. In addition, the sensitivity analysis results might be used to infer the formulation of a proper biofilm model calibration protocol for the invasion phenomenon.

Acknowledgements

This research is supported by the Basque Government through the BERC 2014–2017 and BERC 2018–2021 programs, by the Spanish Ministry of Economy and Competitiveness MINECO through BCAM Severo Ochoa accreditations SEV-2013-0323 and SEV-2017-0718 and through project MTM2016-76016-R MIP, and by the Ph.D grant La Caixa 2014.

References

[1] Flemming H-C, Wingender J, Szewzyk U, Steinberg P, Rice SA, Kjelleberg S. Biofilms: an emergent form of bacterial life. Nat Rev Microbiol 2016;14(9):563.

[2] Stoodley P, Sauer K, Davies DG, Costerton JW. Biofilms as complex differentiated communities. Annu Rev Microbiol 2002;56(1):187–209.

[3] Mattei M, Frunzo L, D'Acunto B, Pechaud Y, Pirozzi F, Esposito G. Continuum and discrete approach in modeling biofilm development and structure: a review. J Math Biol 2018;76(4):945–1003.

[4] Klapper I, Dockery J. Mathematical description of microbial biofilms. SIAM Rev 2010;52(2):221–65.

[5] Wanner O, Gujer W. A multispecies biofilm model. Biotechnol Bioeng 1986;28(3):314–28. https://onlinelibrary.wiley.com/doi/abs/10.1002/bit.260280304.

[6] Eberl HJ, Parker DF, Van Loosdrecht MCM. A new deterministic spatio-temporal continuum model for biofilm development. J Theoret Med 2001;3(3):161–75. doi:10.1080/10273660108833072.

[7] Alpkvista E, Klapper I. A multidimensional multispecies continuum model for heterogeneous biofilm development. Bull Math Biol 2007;69(2):765–89. doi:10.1007/s11538-006-9168-7.

[8] Cogan NG. Two-fluid model of biofilm disinfection. Bull Math Biol 2008;70(3):800–19. doi:10.1007/s11538-007-9280-3.

[9] Zhang T, Cogan NG, Wang Q. Phase field models for biofilms. i. theory and one-dimensional simulations. SIAM J Appl Math 2008;69(3):641–69. http://dblp.uni-trier.de/db/journals/siamam/siamam69.html#ZhangCW08.

[10] Rahman KA, Sudarsan R, Eberl HJ. A mixed-culture biofilm model with cross-diffusion. Bull Math Biol 2015;77(11):2086–124. doi:10.1007/s11538-015-0117-1.

[11] Picioreanu C, Kreft J-U, Van Loosdrecht MCM. Particle-based multidimensional multispecies biofilm model. Appl Environ Microbiol 2004;70(5):30243040. doi:10.1128/AEM.70.5.3024-3040.2004.

[12] Kreft J-U, Picioreanu C, Wimpenny JWT, van Loosdrecht MCM. Individual-based modelling of biofilms. Microbiology 2001;147(11):2897–912.

[13] Tang Y, Valocchi AJ. An improved cellular automaton method to model multispecies biofilms. Water Res 2013;47(15):5729–42. doi:10.1016/j.watres.2013.06.055.

[14] Jayathilake PG, Gupta P, Li B, Madsen C, Oyebamiji O, Gonzlez-Cabaleiro R, et al. A mechanistic individual-based model of microbial communities. PLoS ONE 2017;12(8):1–26. doi:10.1371/journal.pone.0181965.

[15] Lardon LA, Merkey BV, Martins S, Dtsch A, Picioreanu C, Kreft J-U, et al. Idynomics: next-generation individual-based modelling of biofilms. Environ Microbiol 2011;13(9):2416–34. doi:10.1111/j.1462-2920.2011.02414.x.

[16] Boltz JP, Smets BF, Rittmann BE, van Loosdrecht M, Morgenroth E, Daigger GT. From biofilm ecology to reactors: a focused review. Water Sci Technol 2017;75(8):1753–60.

[17] Le Maitre O, Knio O. Spectral methods for uncertainty quantification. Springer; 2010.

[18] Smith R. Uncertainty quantification: theory, implementation, and applications. Society for industrial and applied mathematics; 2013. ISBN 9781611973211.

[19] Iooss B, Saltelli A. Introduction to sensitivity analysis. In: Handbook of uncertainty quantification. Springer International Publishing; 2016. p. 1–20. doi:10.1007/978-3-319-11259-6_31-1.

[20] De Lozzo M, Marrel A. Sensitivity analysis with dependence and variance-based measures for spatio-temporal numerical simulators. Stoch Environ Res Risk Assess 2017;31(6):1437–53.

[21] Sobol I. Sensitivity analysis for nonlinear mathematical models. Math Model Comput Exp 1993;1(4):407–14.

[22] Homma T, Saltelli A. Importance measures in global sensitivity analysis of nonlinear models. Reliab Eng Syst Saf 1996;52(1):1–17. doi:10.1016/0951-8320(96)00002-6.

[23] Sobol I, Kucherenko S. Derivative based global sensitivity measures and their link with global sensitivity indices. Math Comput Simul 2009;79(10):3009–17. doi:10.1016/j.matcom.2009.01.023.

[24] Lamboni M, Monod H, Makowski D. Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models. Reliab Eng Syst Saf 2011;96(4):450–9. doi:10.1016/j.ress.2010.12.002.

[25] Lamboni M, Iooss B, Popelin A-L, Gamboa F. Derivative-based global sensitivity measures: general links with sobol indices and numerical tests. Math Comput Simul 2013;87:45–54. doi:10.1016/j.matcom.2013.02.002.

[26] Kucherenko S, Iooss B. Derivative-based global sensitivity measures. Cham: Springer International Publishing; 2016. p. 1–24. ISBN 978-3-319-11259-6. doi:10.1007/978-3-319-11259-6_36-1.

[27] Borgonovo E. A new uncertainty importance measure. Reliab Eng Syst Saf 2007;92(6):771–84. doi:10.1016/j.ress.2006.04.015.

[28] Borgonovo E, Iooss B. Moment-independent and reliability-based importance measures. Cham: Springer International Publishing; 2016. p. 1–23. ISBN 978-3-319-11259-6. doi:10.1007/978-3-319-11259-6_37-1.

[29] Xie X, Ohs R, Spie A, Krewer U, Schenkendorf R. Moment-independent sensitivity analysis of enzyme-catalyzed reactions with correlated model parameters. IFAC-PapersOnLine 2018;51(2):753–8. 9th Vienna International Conference on Mathematical Modelling. doi: 10.1016/j.ifacol.2018.04.004.

[30] Stanescu D, Chen-Charpentier BM. Random coefficient differential equation models for bacterial growth. Math Comput Model 2009;50(5):885–95. doi:10.1016/j.mcm.2009.05.017.

[31] Chen-Charpentier BM, Stanescu D. Biofilm growth on medical implants with randomness. Math Comput Model 2011;54(7):1682–6. doi:10.1016/j.mcm.2010.11.075.

[32] Hao X, Heijnen JJ, van Loosdrecht MCM. Sensitivity analysis of a biofilm model describing a one-stage completely autotrophic nitrogen removal (canon) process. Biotechnol Bioeng 2002;77(3):266–77. doi:10.1002/bit.10105.

[33] Brockmann D, Morgenroth E. Comparing global sensitivity analysis for a biofilm model for two-step nitrification using the qualitative screening method of morris or the quantitative variance-based fourier amplitude sensitivity test (fast). Water Sci Technol 2007;56(8):85. doi:10.2166/wst.2007.600.

[34] Boltz J, Morgenroth E, Brockmann D, Bott C, Gellner W, Vanrolleghem P. Systematic evaluation of biofilm models for engineering practice: components and critical assumptions. Water Sci Technol 2011;64(4):930–44. doi:10.2166/wst.2011.709.

[35] Lackner S, Smets B. Effect of the kinetics of ammonium and nitrite oxidation on nitritation success or failure for different biofilm reactor geometries. Biochem Eng J 2012;69:123–9. doi:10.1016/j.bej.2012.09.006.

[36] Vangsgaard AK, Mauricio-Iglesias M, Gernaey KV, Smets BF, Sin G. Sensitivity analysis of autotrophic n removal by a granule based bioreactor: influence of mass transfer versus microbial kinetics. Bioresour Technol 2012;123:230–41. doi:10.1016/j.biortech.2012.07.087.

[37] Winkler M-K, Ettwig K, Vannecke T, Stultiens K, Bogdan A, Kartal B, et al. Modelling simultaneous anaerobic methane and ammonium removal in a granular sludge reactor. Water Res 2015;73:323–31. doi:10.1016/j.watres.2015.01.039.

[38] Clarelli F, Di Russo C, Natalini R, Ribot M. A fluid dynamics multidimensional model of biofilm growth: stability, influence of environment and sensitivity. Math Med Biol 2016;33(4):371–95. doi:10.1093/imammb/dqv024.

[39] Reichert P. Aquasim - a tool for simulation and data analysis of aquatic systems. Water Sci Technol 1994;30(2):21. doi:10.2166/wst.1994.0025.

[40] DAcunto B, Frunzo L, Klapper I, Mattei M. Modeling multispecies biofilms including new bacterial species invasion. Math Biosci 2015;259:20–6. doi:10.1016/j.mbs.2014.10.009.

[41] DAcunto B, Frunzo L, Klapper I, Mattei M, Stoodley P. Mathematical modeling of dispersal phenomenon in biofilms. Math Biosci 2018. doi:10.1016/j.mbs.2018.07.009.

[42] Wanner O, Reichert P. Mathematical modeling of mixed-culture biofilms. Biotechnol Bioeng 1996;49(2):172–84.

[43] Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, et al. Global sensitivity analysis. the primer. Chichester, UK: John Wiley & Sons, Ltd; 2007. doi:10.1002/9780470725184.

[44] Emery CM, Biancamaria S, Boone A, Garambois P-A, Ricci S, Rochoux MC, et al. Temporal variance-based sensitivity analysis of the river-routing component of the large-scale hydrological model isba-trip: application on the amazon basin. J Hydrometeorol 2016;17(12):3007–27. doi:10.1175/JHM-D-16-0050.1.

[45] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference and prediction. 2nd. Springer; 2009.

[46] Sudret B. Global sensitivity analysis using polynomial chaos expansions. Reliab Eng Syst Saf 2008;93(7):964–79. doi:10.1016/j.ress.2007.04.002.

[47] Poëtte G, Després B, Lucor D. Uncertainty quantification for systems of conservation laws. J Comput Phys 2009;228(7):2443–67. doi:10.1016/j.jcp.2008.12.018.

[48] Xiu D. Numerical methods for stochastic computations: a spectral method approach. Princeton University Press; 2010.

[49] Després B, Poëtte G, Lucor D. Robust uncertainty propagation in systems of conservation laws with the entropy closure method. Springer International Publishing; 2013. p. 105–49. doi:10.1007/978-3-319-00885-1_3.

[50] Birolleau A, Poëtte G, Lucor D. Adaptive Bayesian inference for discontinuous inverse problems, application to hyperbolic conservation laws. Commun Comput Phys 2014;16:1–34.

[51] Dubreuil S, Berveiller M, Petitjean F, Salan M. Construction of bootstrap confidence intervals on sensitivity indices computed by polynomial chaos expansion. Reliab Eng Syst Saf 2014;121:263–75. doi:10.1016/j.ress.2013.09.011.

[52] Oakley J, O'Hagan A. Probabilistic sensitivity analysis of complex models: a Bayesian approach. J Royal Stat Soc 2004;66(3):751–69. doi:10.1111/j.1467-9868.2004.05304.x.

[53] Marrel A, Iooss B, Laurent B, Roustant O. Calculations of sobol indices for the gaussian process metamodel. Reliab Eng Syst Saf 2009;94(3):742–51. doi:10.1016/j.ress.2008.07.008.

[54] Lockwood B, Anitescu M. Gradient-enhanced universal Kriging for uncertainty propagation. Nucl Sci Eng 2012:1–32.

[55] Le Gratiet L, Cannamela C, Iooss B. A Bayesian approach for global sensitivity analysis of (multifidelity) computer codes. SIAM/ASA J Uncertainty Quantif 2014;2(1):336–63. doi:10.1137/130926869.

[56] Marrel A, Perot G, Mottet C. Development of a surrogate model and sensitivity analysis for spatio-temporal numerical simulators. Stoch Environ Res Risk Assess 2015;29(3):959–74.

[57] Schoebi R, Sudret B, Wiart J. Polynomial-chaos-based Kriging. Int J Uncertain Quan 2015;5(2):171–93.

[58] Le Gratiet L, Marelli S, Sudret B. Metamodel-Based sensitivity analysis: polynomial chaos expansions and gaussian processes. In: Handbook of uncertainty quantification. Springer International Publishing; 2017. p. 1–37. doi:10.1007/978-3-319-11259-6_38-1.

[59] Owen N, Challenor P, Menon PP, Bennani S. Comparison of surrogate-based uncertainty quantification methods for computationally expensive simulators. SIAM/ASA J Uncertainty Quantif 2017;5(1):403–35. doi:10.1137/15M1046812.

[60] Roy PT, El Moçayd N, Ricci S, Jouhaud J-C, Goutal N, De Lozzo M, et al. Comparison of polynomial chaos and gaussian process surrogates for uncertainty quantification and correlation estimation of spatially distributed open-channel steady flows. Stoch Environ Res Risk Assess 2018;32(6):1723–41. doi:10.1007/s00477-017-1470-4.

[61] Urban NM, Fricker TE. A comparison of latin hypercube and grid ensemble designs for the multivariate emulation of an earth system model. Comput Geosci 2010;36(6):746–55. doi:10.1016/j.cageo.2009.11.004.

[62] Trucchia A, Egorova V, Pagnini G, Rochoux MC. On the merits of sparse surrogates for global sensitivity analysis of multi-scale nonlinear problems: application to turbulence and fire- spotting model in wildland fire simulators. Commun Nonlinear Sci Numer Simul 2018.

[63] Blatman G, Sudret B. Efficient computation of global sensitivity indices using sparse polynomial chaos expansions. Reliab Eng Syst Saf 2010;95(11):1216–29. doi:10.1016/j.ress.2010.06.015.

[64] Blatman G, Sudret B. Adaptative sparse polynomial chaos expansion based on least angle regression. J Comput Phys 2011;230(6):2345–67.

[65] Pettersson P, Doostan A, Nordström J. Level set methods for stochastic discontinuity detection in nonlinear problems. J Comput Phys 2018 arXiv:1810.08607.

[66] Liem RP, Mader CA, Martins JR. Surrogate models and mixtures of experts in aerodynamic performance prediction for aircraft mission analysis. Aerosp Sci Technol 2015;43:126–51. doi:10.1016/j.ast.2015.02.019.

[67] Campbell K, McKay M, Williams B. Sensitivity analysis when model outputs are functions. Reliab Eng Syst Saf 2006;91(10–11):1468–72.

[68] Gamboa F, Janon A, Klein T, Lagnoux A. Sensitivity analysis for multidimensional and functional outputs. Electron J Stat 2014;8(1):575–603. doi:10.1214/14-EJS895.

[69] DAcunto B, Frunzo L. Free boundary problem for an initial cell layer in multispecies biofilm formation. Appl Math Lett 2012;25(1):20–6. doi:10.1016/j.aml.2011.06.032.
[70] Damblin G, Couplet M, B I. Numerical studies of space filling designs : optimization of latin hypercube samples and subprojection properties. J Simul 2013.
[71] Xiu D, Karniadakis G. The Wiener–Askey polynomial chaos for stochastic differential equations. SIAM J Scientif Comput 2002;24(2):619–44. doi:10.1137/S1064827501387826.
[72] Blatman G. Adaptative sparse polynomial chaos expansions for uncertainty propagation and sensitivity analysis. Université Blaise Pascal, Clermont-Ferrand; 2009.
[73] Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Annals Stat 2004;32(2):407–99. doi:10.1214/009053604000000067.
[74] Berveiller M, Sudret B, Lemaire M. Stochastic finite element: a non intrusive approach by regression. Eur J Comput Mech 2006;15:81–92. doi:10.3166/remn.15.81-92.
[75] Rasmussen C, Williams C. Gaussian processes for machine learning. MIT Press; 2006.
[76] Jones DR, Perttunen CD, Stuckman BE. Lipschitzian optimization without the Lipschitz constant. J Optim Theory Appl 1993;79(1):157–81. doi:10.1007/BF00941892.
[77] Baudin M, Dutfoy A, Iooss B, Popelin A-L. OpenTURNS: an industrial software for uncertainty quantification in simulation. Springer International Publishing; 2017. p. 2001–38. ISBN 978-3-319-12385-1. doi:10.1007/978-3-319-12385-1_64.
[78] Roy PT, Ricci S, Dupuis R, Campet R, Jouhaud J-C, Fournier C. Batman: statistical analysis for expensive computer codes made easy. J Open Source Softw 2018;3(21):493. doi:10.21105/joss.00493.