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

A. Trucchia\textsuperscript{a,∗}, M.R. Mattei\textsuperscript{b}, V. Luongo\textsuperscript{b}, L. Frunzo\textsuperscript{b}, M.C. Rochoux\textsuperscript{c}

\textsuperscript{a} BCAM – Basque Center for Applied Mathematics, Alameda de Mazarredo 14, Bilbao 48009, Basque Country, Spain
\textsuperscript{b} Department of Mathematics and Applications “R. Caccioppoli”, via Cistia 1, Naples 81126, Italy
\textsuperscript{c} CECI, Université de Toulouse, CNRS, CERFACS, 42 Avenue Gaspard Coriolis, Toulouse cedex 1, 31057, France

\textbf{A R T I C L E   I N F O}

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

\textbf{A B S T R A C T}

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.

© 2019 Elsevier B.V. All rights reserved.

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-

\textsuperscript{∗} Corresponding author.

\textit{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**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>GP</td>
<td>Gaussian Process</td>
</tr>
<tr>
<td>gPC</td>
<td>generalized Polynomial Chaos</td>
</tr>
<tr>
<td>LAR</td>
<td>Least-Angle Regression</td>
</tr>
<tr>
<td>PDF</td>
<td>Probability Density Function</td>
</tr>
<tr>
<td>RBF</td>
<td>Radial Basis Function</td>
</tr>
<tr>
<td>SLS</td>
<td>Standard Least Squares</td>
</tr>
<tr>
<td>STD</td>
<td>Standard Deviation</td>
</tr>
</tbody>
</table>

**Model quantities**

| $\mathcal{F}$, biofilm model operator | Units |
| $z$, space variable | [L] |
| $t$, time variable | [T] |
| $X_i = \rho_i f_i$, concentration of $i$th microbial species | [M L$^{-3}$] |
| $\rho_i$, biomass density for $i$th microbial species | [ML$^{-3}$] |
| $S_j$, concentration of $j$th substrate | [ML$^{-3}$] |
| $r_{sj}$, conversion rate of $S_j$ | [ML$^{-3}$T$^{-1}$] |
| $r_{Mi}$, specific growth rate of $X_i$ | [T$^{-1}$] |
| $r_i$, specific growth rate of $X_i$ due to invasion | [T$^{-1}$] |
| $f_i$, volume fraction of $X_i$ | [-] |
| $\psi_i$, concentration of $i$th planktonic microbial species | [ML$^{-3}$] |
| $r_{\psi,i}$, conversion rate of $\psi_i$ | [ML$^{-3}$T$^{-1}$] |
| $u(z, t)$, advective biomass velocity | [LT$^{-1}$] |
| $L(t)$, biofilm thickness at time $t$ | [L] |

**Experiment quantities**

| $f_1$, heterotrophic bacteria volume fraction | [-] |
| $f_2$, autotrophic bacteria volume fraction | [-] |
| $f_3$, inert material volume fraction | [-] |
| $S_1$, organic carbon concentration | [g$_{\text{CODm}}$] |
| $S_2$, ammonia concentration | [g$_{\text{N_m}}$] |
| $S_3$, oxygen concentration | [g$_{\text{O_m}}$] |
| $k_{\text{coli}}$, maximum colonization rate of autotrophic bacteria | [d$^{-1}$] |
| $k_{\text{af}}$, affinity-type constant for $\psi_2$ | [g$_{\text{CODm}}$] |
| $Y_{\psi,2}$, yield of $X_2$ on $\psi_2$ | [-] |

**Uncertainty analysis variables**

| $d$, uncertain space dimension | | |
| $\theta$, vector of input parameters of dimension $d$ | | |
| $y$, vector of quantities of interest of dimension $n$ | | |
| $N$, size of the training set | | |

...
codified into a set of quantities of interest $y = \{y_1 \ldots y_n\}$, leading to the definition of the functional relation $F$

$$\theta \in \mathbb{R}^d \rightarrow y = F(\theta) \in \mathbb{R}^n.$$  

In the framework of uncertainty quantification [17,18], the set of input parameters $\theta$ 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 $y$. In complement, global sensitivity analysis methods [19,20] provide valuable ways to characterize the input-output model dependency $F$: 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 $F$. 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 $F$ for any set of parameters $\theta$ 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 quantify 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 \(X_i(z, t), i = 1, \ldots, N_s\), \(X = X_1, \ldots, X_{N_s}\); (2) the concentration of planktonic species \(\psi_j(z, t), i = 1, \ldots, N_p, \psi = \psi_1, \ldots, \psi_{N_p}\); and (3) the concentration of dissolved substrates \(S_j(z, t), j = 1, \ldots, N_m, S = S_1, \ldots, S_{N_m}\), 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 \(N_i\), although in most of applications \(N_i\) denotes the number of all particular components, such as extracellular polymeric substance, inert material and all the phenotype variants of the microbial species.

In this model, the concentration of the \(i\)th microbial species in sessile form \(X_i(z, t)\) reads

\[
\begin{align*}
\frac{\partial X_i}{\partial t}(z, t) + \frac{\partial}{\partial z}(u(z, t)X_i(z, t)) &= \rho_i r_{Mi}(z, t, X, S) + \rho_i r_i(z, t, S, \psi), \\
X_i(z, 0) &= \varphi_i(z), & 0 \leq z \leq L(0).
\end{align*}
\]

Eq. (2) describes the growth of the \(i\)th 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 \(r_{Mi}\) 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 \(i\)th 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

\[
f_i = X_i/\rho_i, \quad \sum_{i=1}^{N_i} f_i = 1.
\]

where \(f_i\) represents the volume fraction at a particular location that is occupied by the \(i\)th species, and where \(\rho_i\) denotes the biomass density for the \(i\)th 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
\[
\begin{align*}
\frac{\partial u}{\partial z}(z, t) &= \sum_{i=1}^{N_i} \left( r_{M,i}(z, t, \mathbf{X}, \mathbf{S}) + r_i(z, t, \mathbf{S}, \psi) \right), \\
u(0, t) &= 0, \quad z = 0, \quad t \geq 0.
\end{align*}
\] (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. $L = L(t)$. 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:
\[
\begin{align*}
\frac{dL}{dt}(t) &= u(L(t), t) + \sigma_a(t) - \sigma_d(L(t)), \quad t > 0, \\
L(0) &= L_0, \quad t = 0,
\end{align*}
\] (5)

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

The concentration of the $i$th planktonic species $\psi_i(z, t)$ is governed by the following diffusion-reaction equation:
\[
\begin{align*}
\frac{\partial \psi_i}{\partial t}(z, t) - \frac{\partial}{\partial z} \left( D_{M,i} \frac{\partial \psi_i}{\partial z}(z, t) \right) &= r_{\psi,i}(z, t, \mathbf{S}, \psi), \\
\psi_i(z, 0) &= \psi_{i,0}(z), \quad t = 0, \quad 0 \leq z \leq L(0), \\
\frac{\partial \psi_i}{\partial z}(0, t) &= 0, \quad z = 0, \quad t > 0, \\
\psi_i(L(t), t) &= \psi^*_i(t), \quad z = L(t), \quad t > 0.
\end{align*}
\] (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 $i$th 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 $j$th dissolved substrate $S_j(z, t)$ is also governed by a reaction-diffusion equation
\[
\begin{align*}
\frac{\partial S_j}{\partial t}(z, t) - \frac{\partial}{\partial z} \left( D_j \frac{\partial S_j}{\partial z}(z, t) \right) &= r_{S,j}(z, t, \mathbf{X}, \mathbf{S}), \\
S_j(z, 0) &= S_{j,0}(z), \quad t = 0, \quad 0 \leq z \leq L(0), \\
\frac{\partial S_j}{\partial z}(0, t) &= 0, \quad z = 0, \quad t > 0, \\
S_j(L(t), t) &= S^*_j(t), \quad t > 0.
\end{align*}
\] (7)

where the term $r_{S,j}$ represents the $j$th substrate production or consumption rate due to microbial metabolism, and where $D_j$ denotes the diffusion coefficient of the $j$th substrate within the biofilm. The initial concentration of the $j$th 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 $\psi_2$ 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 \( r_{M,1}, r_{M,2}, r_i \) and \( r_{\psi,i} \), are taken from Refs. [40,42].

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

\[
\begin{align*}
    r_{M,1} &= \left( \mu_{\text{max},1} \frac{S_1}{K_{1,1} + S_1} + \frac{S_2}{K_{1,2} + S_2} + \frac{S_3}{K_{1,3} + S_3} - k_{d,1} \right) X_1, \\
    r_{M,2} &= \left( \mu_{\text{max},2} \frac{S_2}{K_{2,2} + S_2} + \frac{S_3}{K_{2,3} + S_3} - k_{d,2} \right) X_2, \\
    r_{M,3} &= k_{d,1} X_1 + k_{d,2} X_2,
\end{align*}
\]

where \( \mu_{\text{max},i} \) denotes the maximum net growth rate for the \( i \)th biomass, \( K_{i,j} \) is the affinity constant of the \( j \)th substrate for the \( i \)th biomass, and \( k_{d,i} \) represents the decay constant for the \( i \)th 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

\[
\begin{align*}
    r_1 &= r_3 = 0, \\
    r_2 &= k_{\text{col},2} \frac{\psi}{\psi_2} \frac{S_2}{K_{2,2} + S_2} \frac{S_3}{K_{2,3} + S_3},
\end{align*}
\]

where \( k_{\text{col},2} \) corresponds to the maximum colonization rate of autotrophic bacteria, and where \( k_{\psi,2} \) corresponds to the affinity-type constant for \( \psi_2 \).

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

\[
\begin{align*}
    r_{S,1} &= -Y_1 \mu_{\text{max},1} \frac{S_1}{K_{1,1} + S_1} \frac{S_3}{K_{1,3} + S_3} X_1, \\
    r_{S,2} &= -Y_2 \mu_{\text{max},2} \frac{S_2}{K_{2,2} + S_2} \frac{S_3}{K_{2,3} + S_3} X_2, \\
    r_{S,3} &= -Y_1 \mu_{\text{max},1} \frac{S_1}{K_{1,1} + S_1} \frac{S_3}{K_{1,3} + S_3} X_1 - 4.57 \frac{Y_2}{Y_1} \mu_{\text{max},2} \frac{S_2}{K_{2,2} + S_2} \frac{S_3}{K_{2,3} + S_3} X_2,
\end{align*}
\]

with \( Y_i \) denoting the yield of biomass \( i \).

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

\[
    r_{\psi,i} = -\frac{1}{Y_{\psi,i}} r_i,
\]

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 \( \{S_i\} \) 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, \( f_i, i \in \{1, \ldots, N_1\} \), and substrates \( S_j, j \in \{1, \ldots, N_m\} \), with \( N_1 = 3 \) 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).
The quantities of interest could be in principle formulated as

\[ y_i(t) = \frac{\int_0^{L(t)} f_i(z) \, dz}{L(t)}, \quad i \in \{1, \ldots, N_i\}. \tag{17} \]

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:

\[ y_{ijk} = f_i(z_j, t_k), \quad i \in \{1, \ldots, N_i\}, \tag{18} \]

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 \( N_z = 4 \) times at which the biofilm extension is discretized into \( N_z = 5 \) locations. Note that the inert volume fraction \( f_2 \) is retrieved by mass conservation (Eq. (3)). Hence, the model output \( y \) is of functional type and includes the elements \( y_{ijk} \) with \( i = \{1, 2\} \), \( j = 1, \ldots, 5 \), and \( k = 1, \ldots, 4 \) (\( y \in \mathbb{R}^n \) 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 \( \mu_{\text{max},i} \), \( k_{di,j} \), \( K_{ij} \) 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: \( k_{\text{col},2} \) 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 \( r_{g,2} \) on the other hand. Hereafter, \( k_{\text{col},2} \), \( k_{\Psi,2} \) and \( Y_{\Psi,2} \) are respectively denoted by \( k_{\text{col}} \), \( k_{\Psi} \) and \( Y_{\Psi} \) for clarity purposes. The uncertain input vector \( \theta \) is thus defined as

\[ \theta = (k_{\text{col}}, k_{\Psi}, Y_{\Psi}) \in \mathbb{R}^3. \tag{19} \]

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>
<thead>
<tr>
<th>Parameter</th>
<th>Uniform distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k_{\text{col}} )</td>
<td>( U(10^{-4}, 10^{-2}) )</td>
</tr>
<tr>
<td>( k_{\Psi} )</td>
<td>( U(10^{-3}, 10^{-2}) )</td>
</tr>
<tr>
<td>( Y_{\Psi} )</td>
<td>( U(10^{-3}, 10^{-2}) )</td>
</tr>
</tbody>
</table>
the uncertain input parameters are modeled by a random vector \( \Theta \), meaning that their values are supposed to depend on a random parameter \( \omega \) such that \( \Theta = \Theta(\omega) \). \( \omega \) is to be taken from the set of all outcomes \( \Omega \), which is equipped with a \( \sigma \)-algebra \( S \) and a probability measure \( \mathbb{P} \). The triplet \( (\Omega, S, \mathbb{P}) \) forms a probabilistic space [31].

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

\[
Y(\omega) = F(\Theta(\omega)),
\]

with \( F \) 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 \( \Theta \) denoted by \( \rho_{\Theta} \). We need to introduce some assumptions on the nature of such uncertainty sources. First, we assume the components of \( \Theta \) are independent. Second, we consider uniform marginal PDF for each random variable \( \Theta_i \) \((i = 1, \ldots, d)\) in \( \Theta \), denoted by \( \rho_{\Theta_i} \). 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 \( \Theta \) and outputs \( Y \) and to build an emulator of the relation \( F \) 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 \( k_{\text{col}} \), \( k_\psi \) and \( Y_\psi \) evolve. It is a way to define the \( N \) realizations of parameters \( \Theta \), for which the biofilm model is integrated as a “black-box” to obtain the ensemble of \( N \) functional outputs \( y \) from which statistics can be derived. This ensemble forms a database \( D_N \):

\[
D_N = \left\{ (\theta^{(l)}, y^{(l)}) \mid 1 \leq l \leq N \right\},
\]

where \( y^{(l)} = F(\theta^{(l)}) \) stands for the integration of the biofilm model \( F \) associated with the \( l \)th set of input parameters \( \theta^{(l)} \).

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 \( \theta = (k_{\text{col}}, k_\psi, Y_\psi) \) 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 \( Y \equiv Y_{ijk} \) a surrogate by means of a weighted (finite) sum of basis functions:

\[
Y = \sum_{\alpha \in \mathcal{A}} \gamma_\alpha \Psi_\alpha,
\]

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

4.1. Generalized polynomial chaos (gPC) expansion

4.1.1. Standard probabilistic space

\( \Theta \) is defined in the input physical space and its counterpart in the standard probabilistic space is noted \( \zeta = (\zeta_1, \ldots, \zeta_d) \), with \( \zeta_i \) the random variable associated with the \( i \)th uncertain parameter \( \Theta_i \) in \( \Theta \) and characterized by a uniform marginal PDF \( \rho_{\Theta_i} \). The reduced variable \( \zeta_i \) is therefore a uniform variable on \([-1; 1]\). The gPC-framework applies to the standard probabilistic space. The equivalent of \( \rho_{\Theta} \) in the standard probabilistic space is denoted by \( \rho_{\zeta} \). Since all input random variables are assumed independent (see Section 3.2), the joint PDF \( \rho_{\zeta} \) is the product of the marginal PDFs \( \{\rho_{\zeta_i}\}_{i=1,\ldots,d} \).
Fig. 1. Cloud representation of the two databases \( D_N \) with \( N = 216 \), corresponding to different sets of the three parameters \( k_{col} \) (x-axis), \( k_\psi \) (y-axis) and \( Y_\psi \) (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

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

\[
\langle \psi_\alpha(\zeta), \psi_\beta(\zeta) \rangle = \int_{Z_\zeta} \psi_\alpha(\zeta) \psi_\beta(\zeta) \rho_\zeta d\zeta = \delta_{\alpha\beta},
\]

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

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

\[
Y(\omega) = F_{pc}(\Theta) = \sum_{\alpha \in A} \gamma_\alpha \psi_\alpha(\zeta(\omega)),
\]

where \( \{ \psi_\alpha(\zeta) \}_{\alpha \in A} \) 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 \( \{ \gamma_\alpha \}_{\alpha \in A} \) are the unknowns to determine using a projection strategy to derive the emulator \( F_{pc} \).
Fig. 2. Time-evolving species volume fractions $f_1$ and $f_2$ for varying uncertain input vector $\theta = (k_{\text{col}}, k_{\psi}, Y_{\psi})$ (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, $\alpha = (\alpha_1, \ldots, \alpha_d) \in \{0, 1, \ldots, P\}^d$. The number of terms is therefore constrained by the number of input random variables $d$ and by the total polynomial order $P$ so that

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

(25)
The corresponding set of multi-indices $\mathcal{A}_{\text{lin}}$ is defined as
\begin{equation}
\mathcal{A}_{\text{lin}} = \mathcal{A}_{\text{lin}}(d, P) = \{ \alpha \in \mathbb{N}^d : |\alpha| \leq P \},
\end{equation}
where $|\alpha| = |\alpha_1| + \cdots + |\alpha_d|$ 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 $\{ \gamma_{\alpha} \}_{\alpha \in \mathcal{A}}$ 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 $\tilde{y} = \{ \gamma_{\alpha} \}_{\alpha \in \mathcal{A}}$. $\tilde{y}$ is the solution of the following problem:
\begin{equation}
\tilde{y} = \arg\min_{y \in \mathbb{R}^r} \frac{1}{N} \sum_{i=1}^{N} \left( y^{(i)} - \sum_{\alpha \in \mathcal{A}} \gamma_{\alpha} \Psi_{\alpha} \left( \xi^{(i)} \right) \right)^2 ,
\end{equation}
which is solved through classical linear algebra algorithms, i.e.
\begin{equation}
\tilde{y} = (\Psi^T \Psi)^{-1} \Psi^T y ,
\end{equation}
with $\Psi$ the information matrix corresponding to the evaluation of the basis polynomials at each point of the experimental design $\mathcal{D}_N$, i.e. $\Psi = \{ \Psi_{\alpha} (\xi^{(i)}) \}_{\alpha \in \mathcal{A}, 1 \leq i \leq N}$ and with $y$ 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 $\{ \Psi_{\alpha} \}_{\alpha \in \mathcal{A}}$ according to the prescribed marginal PDFs of the inputs $\theta = (k_{\text{col}}, k_{\psi}, Y_\psi) \in \mathbb{R}^3$ ($d = 3$);
2. choose the total polynomial order $P$ according to the complexity of the biological processes;
3. truncate the gPC-expansion to $r_{\text{lin}}$ 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 $A \subset \mathcal{A}_{\text{lin}}$ with a cardinality $r \leq r_{\text{lin}}$. otherwise $A = \mathcal{A}_{\text{lin}}$ and $r = r_{\text{lin}}$;
5. apply least-square minimization to compute the coefficients $\{ \gamma_{\alpha} \}_{\alpha \in A}$ 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 $\mathcal{F}_{\text{GPs}}$ which can be evaluated for any new pair of parameters $\theta' = (k'_{\text{col}}, k'_{\psi}, Y'_{\psi})$.

4.2. Gaussian Process (GP) surrogate
4.2.1. Principles
A surrogate model using GP regression can be cast as follows:
\begin{equation}
Y(\theta) = \mathcal{F}_{\text{GPs}}(\theta) = \sum_{\alpha=1}^{r} \gamma_{\alpha} \Psi_{\alpha}(\theta) ,
\end{equation}
where $\Psi_{\alpha}$ 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, $\{ \Psi_{\alpha}(\theta^{(i)}) \}_{1 \leq i \leq N}$, share a joint Gaussian distribution [75]. Let $\Psi_{\alpha}$ be a GP fully described by its zero mean and its correlation $\pi_{\alpha}$:
\begin{equation}
\tilde{\Psi}_{\alpha}(\theta) \sim \text{GP}(0, \sigma_{\alpha}^2 \pi_{\alpha}(\theta, \theta')) ,
\end{equation}
with \( \pi_\alpha(\theta, \theta') = \mathbb{E}\left[ \tilde{\Psi}_\alpha(\theta) \tilde{\Psi}_\alpha(\theta') \right] \). In the present study, the correlation function \( \pi \) (or kernel) is chosen as a squared exponential (also known as radial basis function – RBF):

\[
\pi_\alpha(\theta, \theta') = \exp\left( -\frac{||\theta - \theta'||^2}{2 \ell^2_\alpha} \right).
\]  
(31)

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

\[
\Psi_\alpha(\theta^*) = \sum_{l=1}^{N} \beta_{l, \alpha} \pi_\alpha(\theta^*, \theta^{(l)}),
\]  
(32)

where

\[
\beta_{l, \alpha} = (\Pi_\alpha + \tau^2 I_N)^{-1} \left( \Psi_\alpha(\theta^{(1)}) \cdots \Psi_\alpha(\theta^{(N)}) \right)^T,
\]

\[
\Pi_\alpha = \left( \pi_\alpha(\theta^{(l)}, \theta^{(m)}) \right)_{1 \leq l, m \leq N},
\]  
(34)

and where \( \tau \) (nugget effect) avoids ill-conditioning issues for the matrix \( \Pi_\alpha \). The hyperparameters \( \{\ell_\alpha, \sigma_\alpha, \tau\} \) 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 \( \pi_\alpha \) suitable for the input vector \( \theta = (k_{col}, k_{\psi}, Y_\psi) \in \mathbb{R}^3 \) (\( d = 3 \)) – we consider RBF in the present study, see Eq. (31);
2. optimize the GP-hyperparameters \( \{\ell_\alpha, \sigma_\alpha, \tau\} \) associated with the kernel \( \pi_\alpha \) using maximum likelihood;
3. formulate the surrogate \( \mathcal{F}_{\text{GP}} \), which can be evaluated for any new pair of parameters \( \theta^* = (k^*_{col}, k^*_{\psi}, Y^*_\psi) \) 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

\[
\epsilon_{\text{emp}} = \frac{1}{N_{\text{halton}}} \sum_{l=1}^{N_{\text{halton}}} \left( y^{(l)} - \hat{y}^{(l)} \right),
\]

(35)

with \( y^{(l)} \) the \( l \)th element of the Halton’s training set, \( \hat{y}^{(l)} \) 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:

\[
Q_2 = 1 - \frac{\sum_{l=1}^{N_{\text{indep}}} \left( y^{(l)} - \hat{y}^{(l)} \right)^2}{\sum_{l=1}^{N_{\text{indep}}} \left( y^{(l)} - \bar{y} \right)^2},
\]

(36)
with 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-squared 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 \(z = L/4\) after 10 days and 0.82 at \(z = 2L/4\) 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 \(\Psi_{\alpha}\) is associated with a colored symbol, where the color represents the magnitude of the coefficient \(c_{\alpha}\). 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 \(k_{\text{col}}\) (\(x\)-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 \((N = 216, \text{implying } P \leq 5)\). 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 adaptive 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 \( \{ \gamma_{\alpha} \}_{\alpha \in A} \) with respect to the three-dimensional input space, \( \theta = (k_{\text{col}}, k_{\psi}, Y_{\psi}) \) \((d = 3)\) 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 \( k_{\text{col}}, k_{\psi} \) and \( Y_{\psi} \), 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 \( \{ \gamma_{\alpha} \}_{\alpha \in A} \). The mean value \( \mu_y \) and STD \( \sigma_y \) of \( y \) can be estimated as

\[
\mu_y = \gamma_0, \tag{37}
\]

\[
\sigma_y = \sqrt{\sum_{\alpha \in A : \alpha \neq 0} \gamma_{\alpha}^2}, \tag{38}
\]
The PDF of each quantity of interest is retrieved through kernel smoothing techniques by sampling the uncertain input space \( Z_\theta \) 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 \( S_j \) \((j = 1, \ldots, 3)\) 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 \( \theta \). This is due to a combined effect of substrate diffusion and microbial metabolism, which leads to the
decrease of substrate concentration with respect to the constant value prescribed at the bulk liquid interface. More specifically, $S_2$ 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 $C_{yy} \in \mathbb{R}^{N_r \times N_r}$, also known as dispersion matrix, to characterize the covariance between the model state $y$ at different locations $z \in [0, L(t)]$ at a given time. $C_{yy}$ can be empirically estimated as

$$C_{yy} = \frac{1}{N - 1} \sum_{i=1}^{N} (y_k^{(i)} - \bar{y}_k)(y_k^{(i)} - \bar{y}_k)^T,$$

where $y_k^{(i)} = (y_{jk}^{(i)})_{j=1,\ldots,N_r}$ is the vector containing the $i$th quantity of interest $y_{jk}$ 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 $k_{col}$, $k_{\psi}$ and $y_{\psi}$.

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).
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

\[
\mathbb{V}[Y] = \sum_{i=1}^{d} \mathbb{V}(Y) + \sum_{j=i+1}^{d} \mathbb{V}_{ij}(Y) + \cdots + \mathbb{V}_{1,2,...,d}(Y),
\]

(40)

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

\[
\mathbb{V}_I(Y) = \mathbb{V}[E(Y|\Theta_I)] - \sum_{j \in I, k \notin I} \mathbb{V}_j(Y), \quad \forall I \subset \{1, \ldots, d\}
\]

(41)

Based on this variance decomposition, the first-order Sobol’ index \( S_i \) associated with the \( i \)th parameter of \( \Theta \) is given by

\[
S_i = \frac{\mathbb{V}(Y)}{\mathbb{V}(Y)}.
\]

(42)

and corresponds to the ratio of the output variance \( \mathbb{V}(Y) \) that is uniquely due to the \( i \)th input parameter; \( S_i \) ranges between 0 and 1. The corresponding total Sobol’ index \( S_{\text{T}_i} \) measures the whole contribution of the \( i \)th input parameter (including interaction with other parameters of \( \Theta \)) on the output variance, with the following definition:

\[
S_{\text{T}_i} = \sum_{I \subset \{1, \ldots, d\} \setminus i} S_i.
\]

(43)

By definition, \( S_{\text{T}_i} \geq S_i \). If both first-order and total indices are not equal, this indicates that the input parameter \( \Theta_i \) has some interactions with other parameters of \( \Theta \) 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

\[
S_{i, \text{gPC}} = \frac{1}{\sigma^2} \sum_{n_i=0}^{\infty} \sum_{n_{ij}=0}^{\infty} \gamma^{2}_{n_i}.
\]

(44)

with \( \sigma^2 \) the output STD computed using Eq. (38).
Fig. 10. First-order and total Sobol’ indices (in logarithmic scale) associated with uncertain parameters \( \theta = (k_{col}, k_p, Y_p) \) 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 \( (0 \leq z \leq L(t)) \) 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: \( k_{col}, k_p, Y_p \).

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 \( t \in \{5, 10, 15\} \) 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 microorganisms 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 \( Y_p \) and \( k_p \) 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 $k_\psi$ and $Y_\psi$ 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 $k_\psi$ is usually more influential than $Y_\psi$, 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 $Y_\psi$ becomes higher than that of $k_\psi$ 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. Heterotroph-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. $\theta = (k_{col,2}, k_\psi, Y_\psi) \in \mathbb{R}^3$. 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 $\theta$. 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 ensure 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 $k_\psi$ (affinity-type constant for planktonic species associated with autotrophic bacteria) and $Y_\psi$ (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 $(k_\psi, Y_\psi)$ 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 \( L=\ell(t) \) 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
