Surrogate Based Global Sensitivity Analysis of ADM1-based Anaerobic Digestion Model

A. Trucchia[a], L. Frunzo[b]

[a]*BCAM – Basque Center for Applied Mathematics, Alameda de Mazarredo 14, 48009 Bilbao, Basque Country, Spain*
[b]*Department of Mathematics and applications "R. Caccioppoli" via Cintia 1, 91126 Naples, Italy*

Abstract

In order to calibrate the model parameters, Sensitivity Analysis routines are mandatory to rank the parameters by their relevance and fix to nominal values the least influential factors. Despite the high number of works based on ADM1, very few are related to sensitivity analysis. In this study Global Sensitivity Analysis (GSA) and Uncertainty Quantification (UQ) for an ADM1-based Anaerobic Digestion Model have been performed. The modified version of ADM-based model selected in this study was presented by *Esposito and co-authors* in 2013. Unlike the first version of ADM1, focused on sewage sludge degradation, the model of *Esposito* is focused on organic fraction of municipal solid waste digestion. It his recalled that in many applications the hydrolysis is considered the bottleneck of the overall anaerobic digestion process when the input substrate is constituted of complex organic matter. In *Esposito*'s model a surfaced based kinetic approach for the disintegration of complex organic matter is introduced. This approach allows to better model the disintegration step taking into account the effect of particle size distribution on the digestion process. Due to the large number of param-

eters to be analyzed a first preliminary screening analysis, with the Morris' Method, has been conducted. Since two quantities of interest (QoI) have been considered, the initial screening has been performed twice, obtaining two set of parameters containing the most influential factors in determining the value of each QoI. A surrogate of ADM1 model has been defined making use of the two defined quantities of interest. The output results from the surrogate model have been analyzed with Sobol' indices for the quantitative GSA. Finally, uncertainty quantification has been performed. By adopting kernel smoothing techniques, the Probability Density Functions of each quantity of interest have been defined.
Keywords: Global Sensitivity analysis, Uncertainty Quantification, ADM1-based Anaerobic Digestion Model

1. Introduction

Anaerobic Digestion (AD) technology is a bio-chemical process for the treatment of organic matrices.

During the last decades, AD has been widely applied in several industrial fields, such as the treatment of organic wastes. Such technology allows to reduce the environmental pollution as well as to generate energy. The development of the AD technology called for the introduction of specific mathematical models for the design and the management of AD reactors.

The first models were proposed from the early 1980s [1]. They were mainly focused on the modeling of biochemical processes occurring in AD reactors, based on Ordinary Differential Equation (ODE) systems. Characterized by different levels of complexity, they requested different assump-

tions and simplifications. The development of the models would follow for a couple of decades and several approaches have been consolidated during these years [2]. Moreover, in order to propose an unified approach in AD mathematical modeling in 2002 a Task group of the International Water Association (IWA) developed a comprehensive mathematical model known as ADM1 [3], which was based on the knowledge on modeling and simulation of anaerobic digestion systems emerged over the previous years. After its publication, the ADM1 became very soon a well-known and widely studied mathematical model, able to describe the conversion of complex organic compounds into methane (CH_4). ADM1 simulates the main biochemical (related to the microbial community) processes leading to the final production of CH_4. Initially based on the AD of sewage sludge of urban wastewater treatment plants, IWA's ADM1 model has undergone many modifications/manipulations aimed to introduce specific process affecting the anaerobic conversion of organic substrates, in order to simulate the degradation of more complex organic substrates than sewage sludge, such as Organic fraction of municipal solid waste (OFMSW). Important ADM1 extensions were made by *Fedovovich et al.* [4] with the introduction of sulfur degradation and kinetics, and the by Batstone and Keller [5] who took into account the $CaCO_3$ precipitations. ADM1 extensions have also been proposed to remove the discrepancies in both carbon and nitrogen balances [6] and to improve the physicochemical ADM1 framework by incorporating more inorganic components such as trace elements (TE) and phosphates. More precisely, TE dynamics and their effects on AD systems have been modeled [7, 8, 9].

With the aim to extend the ADM1 applicability to the anaerobic digestion

of organic solid wastes, *Esposito and co-authors* [10] modified the ADM1 by introducing a surfaced based kinetic. This allowed to consider the effects of particle size distribution in AD of complex organics.

The selection of the parameters in numerical simulations of the ADM1 model constitutes a topic worth investigating. Due to the high number of processes and parameters, and thus of kinetic parameters, their choice plays a key role in the simulation result. In this context the study of the sensitivity of the AD model predictions with respect to the variability in the inputs provides a way to better understand the response of the model to an arbitrary choice of parameters.

In order to calibrate the model parameters for the model to exhibit a better fitness with the experimental data, Sensitivity Analysis (SA) routines are mandatory in order to rank the parameters by their relevance and fix to nominal values the least influential factors. The need for a reliable Global Sensitivity analysis (GSA) of ADM1 model is expressed in a general framework of good practices in modeling, suggested by *Saltelli et al.* in [11]. In *Saltelli*'s work, it is pointed out that many uncertainty and sensitivity analyses still explore the input space moving along one-dimensional corridors (i.e Local Sensitivity analysis) and thus leave a vast part of the input parameter space unexplored. In their extensive systematic literature review *Saltelli et al.* show that many highly cited papers (42%, according to their analysis) fail the elementary requirement to properly explore the space of the input factors. The results (that emerged to be discipline-dependent) pointed to a strong need for recognized good practices in SA and Uncertainty Quantification (UQ) procedures.

Despite the high number of works based on ADM1, there are very few works related to sensitivity analysis. In practice, these works are totally focused on local procedures neglecting the more exhaustive global techniques.

Several examples can be found in literatures regarding procedures related to local sensitivity: *Jeong et al* in [12] introduced a local sensitivity index for the sake of ordering kinetic and stoichiometric parameters of ADM1 with respect to their influence on the simulation results. Such index had to be averaged over different simulation times. *Souza et al*, in [13], used biochemical methane potential (BMP) tests data for calibrating the Anaerobic Digestion Model No. 1 (ADM1) by the means of a preliminar screening via SA techniques adopting Sensitivity Index (SI) introduced in [12]; *Lee et al.* in [14] applied the screening of [12] for the ADM1 model in a temperature-phased anaerobic digestion (TPAD) application.

Barrera et al. in [15] adopted the so called Local Relative Sensitivity Analysis method (see Ref. [16]) for a screening phase that ultimately led to a calibration and validation of a modified version of ADM1 that accounted for sulfate reduction.

In [17] a parametric, derivative-based local sensitivity analysis was enforced with respect to the level of CH_4 production, in order to apply ADM1 to simulate biogas production from Hydrilla verticillata.

Morales et al. in [18] adopted a sensitivity analysis screening, by using a simple methodology that consisted of changing the value of each input concentrations "one at a time" (OAT) while leaving the other parameters fixed. In their work, they analyzed a continuous stirred tank reactor (CSTR) in steady state for a wastewater treatment plant.

Concerning Global Sensitivity Analysis, in [19] a GSA study has been performed on the Benchmark Simulation Model no. 2 (BSM2) model in its open loop (without control) version, by means of Monte Carlo (MC) experiments and linear regression of the MC results [20]. Such model is a rather complex *plant-wide* model, which accounts for wastewater and sludge treatment, and the main focus was not unfortunately the sole ADM1 model rather than its interactions with the other subsystems.

The aim of this study is to perform a GSA and UQ on a modified version of ADM1 with surfaced based kinetic. The performed analysis focuses on a large set of parameters, which models different physical, biological and chemical phenomena, entwined in the complex dynamics of ADM1 process. Such set of input factors took into account kinetic parameters and operational parameters. As for the outputs, the dynamics of each execution of the ADM1 model are encoded into two quantities of interest, that account for the CH_4 production history and the peak of acetic acid.

One of the main outcomes of this paper is to state, after an extensive GSA, that the two model parameters r_0 and K_{sbk}, both related to the degradation of the substrate (as explained in detail in Section 2) play a key role in the examined AD model. This confirms that the disintegration phase is one of the most important phases of the overall AD process. In the manuscript the adopted methodology and tools, concerning GSA and UQ, are described in detail. In fact, this work is not limited to a mere sensitivity analysis of an ADM1-based model since it constitutes a methodological example of for a global sensitivity study for this class of models. In the presented research it is demonstrated that a rigorous GSA procedure for a complex model of

practical interest such as ADM1 is possible, connecting several algorithms available in literature. This analysis is not in contrast with a preliminary screening, typically based on OAT techniques, that are widespread in applied sciences. On the other hand, it is shown that with a further effort a solid framework of tools is available to the scientist who is willing to have a deeper overview on the interacting parameters, in order to shed light on the structure of models that sometimes are too complex to be analyzed beforehand. In particular, the Morris Elementary Effects and the Sobol' indices (described in the following Sections) are obtained via an exploration of the whole hypercube of the uncertain model parameters, rather than exploring a finite set of segments. The surrogate models allow for a computationally cheap activity of Uncertainty Quantification on the output variables, and the built databases allow for interesting insights on the effect of input parameters on the model output, such as the cobweb plots. All the aforementioned algorithms create a more informative framework, that can help the practitioners in a second phase of the modeling process, where deeper insights are needed.

This paper is structured as follows: in Section 2, the modified version of ADM1 object of the study is described. In Section 3, the selection of the groups of variables for the SA and UQ study is outlined, as well as the choice of the main observables. The studied test case and the databases of simulation are also described in this Section. In Section 4 the UQ and SA techniques adopted in this work, namely preliminary Morris' screening and Surrogate-based UQ and SA, are introduced. The Results are presented and discussed in Section 5, and the concluding remarks as well as future perspectives are given in Section 6.

2. Mathematical Model of Anaerobic Digestion

The analysis conducted in this work is based on a modified version of ADM1 proposed by *Esposito et al.* [21, 22]. The model accounts for the effect of particle size distribution during the disintegration process by using a surface based kinetic and removes the ADM1 discrepancies in both carbon and nitrogen balances according to [6]. In particular, the use of surface based kinetic approach allows to model through the two constants K_{sbk} and a^* the degradation of the substrate due to the mechanical characteristics and to the granulometry, respectively.

As it has just been remarked, the main novelty introduced by *Esposito and co-authors*[10] lies in the different approach used in the disintegration kinetic. In simple terms, it may be stated that the disintegration constant K_{dis} used in the original version of the ADM1 has been substituted by the product of two newly introduced factors, K_{sbk} and a^*. In the considered surface based kinetic, the degradation rate of the organic biodegradable mass M is function of the available area A, namely

$$\frac{dM}{dt} = -K_{sbk}A. \tag{1}$$

It is possible to transpose the last formula in terms of the concentration of the organic biodegradable mass C,

$$\frac{dC}{dt} = -K_{sbk}a^*C, \tag{2}$$

where K_{sbk} is the surface based kinetic constant and $a^* = a^*(r_0)$ is the specific area which depends of the particle radius r_0.

Notably, K_{sbk} is independent of the granulometry of the waste and depends only from the mechanical characteristics of the substrate (i.e. the physical resistence of the waste to disintegration). On the other hand, the parameter a^* depends only from the granulometry, i.e. from the size of the waste particles which need to be anaerobically digested. In particular, K_{sbk} needs to be determined experimentally, while a^* can be calculated *a priori* knowing the granulometry. Assuming spherical particles with radius r_0, a^* reads [10]

$$a^* = \frac{3}{\delta r_0}, \tag{3}$$

where δ is the mass density.

Although the difference of the adopted model with respect to the original ADM1 from the point of view of implemented differential equations may seem little, from a modeling point of view the difference is substantial. The original ADM1 was made for anaerobic digestion of sewage sludge, a substrate characterized by homogeneity both in terms of mechanical disintegration than in terms of granulometry. In that specific case, a simple constant has been sufficient. The modified version of *Esposito et al.* has been proposed for the AD of organic fraction of municipal solid waste (OFMSW), a strongly dishomogeneous substrate, characterized by differences in granulometry. Since the aim of this work is to propose a sensitivity analysis methodology for this type of models and show the effect of disintegration on AD modeling, the original form of the ADM1 would have been limiting because in that case only a constant,K_{dis} would have been taken into account and, above all, the probabilistic description of K_{dis} and its range of variation would have been

less detailed.

The model is based on mass conservation principles and is formulated as a set of ordinary differential equations for the soluble and particulate components constituting the system. In general form, the model is formulated in terms of three groups of state variables: i) soluble components in liquid phase S_i, including the compounds deriving from the hydrolysis of the complex organic matter. ii) particulate components X_i, representing the concentration of the microbial groups involved in the biochemical reactions and the complex organic matter fed to the AD system, and the macromolecules deriving from the disintegration step; iii) gas components $S_{gas,i}$ (i.e. hydrogen, carbon dioxide, methane), in equilibrium with the corresponding components in the liquid phase.

The differential equations governing substrates and bacterial groups dynamics involved in the AD processes take the following form:

$$\frac{dV_{liq}S_i}{dt} = q_{in}S_{in} - q_{out}S_{out} + V_{liq}(\gamma_i\rho_{A,i}(t,\mathbf{S}) - \rho_{T,i}(t,\mathbf{S},\mathbf{S_{gas}}) + \\ + V_{liq}\sum_{j=1}^{m}\alpha_{i,j}\rho_j(t,\mathbf{S},\mathbf{X}), \\ i = 1,...,n_1, \quad t > 0 \tag{4}$$

$$\frac{dV_{liq}X_i}{dt} = q_{in}X_{in} - q_{out}X_{out} + V_{liq}\sum_{j=1}^{m}\alpha_{i,j}\rho_j(t,\mathbf{S},\mathbf{X}), \\ i = n_1+1,...,n_2, \quad t > 0 \tag{5}$$

$$\frac{dV_{gas}S_{gas,i}}{dt} = -q_{gas}S_{gas,i} + V_{liq}\rho_{T,i}(t,\mathbf{S},\mathbf{S_{gas}}),$$

$$i = 1, ..., n_1, \quad t > 0 \tag{6}$$

where:

n_1 denotes the number of soluble components,

$n_2 - n_1$ denotes the number of particulate components,

m_1 denotes the number of biochemical processes taken into account,

$\alpha_{i,j}$ is the stoichiometric coefficient of species i on biochemical process j,

γ_i is the stoichiometric coefficient for the acid base reaction involving the ith soluble component,

S_i denotes the ith soluble component,

X_i denotes the ith particulate component,

$S_{gas,i}$ denotes the ith component in gas form,

$\rho_j(t, \mathbf{S}, \mathbf{X})$ represents the rate of the jth biochemical process,

$\rho_{A,i}(t, \mathbf{S})$ represents the acid base kinetic rate equation for the i^{th} soluble component,

$\rho_{T,i}(t, \mathbf{S}, \mathbf{S_{gas}})$ represents the gas transfer rate for the i^{th} soluble component.

The charge balance accounting for all the ionic species is needed to evaluate the pH:

$$\sum_{i=1}^{p} \mathbf{Q_i^+} - \sum_{i=1}^{q} \mathbf{Q_i^-} = 0, \quad p + q < n_1, \tag{7}$$

where:

p defines the number of cationic components,

q defines the number of anionic components,

$\mathbf{Q_i^+}$ represents the cationic equivalent concentration of species i^{th},

$\mathbf{Q_i^-}$ represents the anionic equivalent concentration of species i^{th}.

In order to solve the differential algebraic system [4-7] suitable initial conditions have to be prescribed.

$$S_i(0) = S_i^0, i = 1, ..., n_1, \tag{8}$$

$$X_i(0) = X_i^0, i = n_1 + 1, ..., n_2, \tag{9}$$

$$S_{gas,i}(0) = S_{gas,i}^0, i = 1, ..., n_1, \tag{10}$$

The detailed biochemical ($\rho_j(t, \mathbf{S}, \mathbf{X})$), acid/base ($\rho_{A,i}(t, \mathbf{S})$) and gas transfer ($\rho_{T,i}(t, \mathbf{S}, \mathbf{S_{gas}})$) reaction rates expression adopted in the model are reported in the following sections. In Appendix B the model equations in matrix form (Petersen matrices) are shown.

2.1. Biochemical reaction rates

According to the the ADM1 approach, the AD process is composed by five main degradation steps (Fig):

i) the *disintegration* of complex organic matter X_c in readily and slowly degradable particulate organic macromolecules (X_{Ch}, X_{Pr}, X_{Li}, X_I) and the contextual release of inorganic carbon (X_{IC}) and inorganic nitrogen (X_{IN});

ii)the *hydrolysis* of the particulate macromolecules in soluble monomers (S_{su}, S_{ss}, S_{fa});

iii) the degradation of soluble monomers in organic volatile acids (S_{va}, S_{pr}, S_{bu}), this step is usually named *acidogenesis*;

iv) the formation of the acetic acid (S_{ac}) and hydrogen gas (S_{h2}) from the degradation of volatile acids and partially from the hydrolisis of soluble monomers (i.e *acetogenesis*), and

229 v) the formation of Methane gas (S_{ch4}) through acetoclastic and hy-
230 drogenotrophic *methanogenesis*.

231 These processes have been mediated by seven microbial groups: Sugar
232 degraders (X_{su}), amino acid degraders (X_{aa}), LCFA degraders (X_{fa}), Valer-
233 ate and Buryrate degraders (X_{c4}), propionate degraders (X_{pro}), acetate de-
234 graders (X_{ac}), hydrogen degraders (X_{h2}).

235 The kinetic rate equation $\rho_j(t, \mathbf{S}, \mathbf{X})$, have been following listed:

$$\rho_1 = K_{sbk} Ca^* \tag{11}$$

$$\rho_2 = K_{hyd,Ch} X_{Ch} \tag{12}$$

$$\rho_3 = K_{hyd,Pr} X_{Pr} \tag{13}$$

$$\rho_4 = K_{hyd,Li} X_{Li,R} \tag{14}$$

$$\rho_5 = \nu_{max,su} \frac{S_{su}}{K_{S,su} + S_{su}} X_{su} I_1 \tag{15}$$

$$\rho_6 = \nu_{max,aa} \frac{S_{aa}}{K_{S,aa} + S_{aa}} X_{aa} I_1 \tag{16}$$

$$\rho_7 = \nu_{max,fa} \frac{S_{fa}}{K_{S,fa} + S_{fa}} X_{fa} I_{2,fa} \tag{17}$$

$$\rho_8 = \nu_{max,c4} \frac{S_{va}}{K_{S,va} + S_{va}} \frac{1}{1 + S_{bu}/S_{va}} X_{c4} I_{2,va} \tag{18}$$

$$\rho_9 = \nu_{max,c4} \frac{S_{bu}}{K_{S,bu} + S_{bu}} \frac{1}{1 + S_{va}/S_{bu}} X_{c4} I_{2,bu} \tag{19}$$

$$\rho_{10} = \nu_{max,pro} \frac{S_{pro}}{K_{S,pro} + S_{pro}} X_{pro} I_{2,pro} \tag{20}$$

$$\rho_{11} = \nu_{max,ac} \frac{S_{ac}}{K_{S,ac} + S_{ac}} X_{ac} I_3 \tag{21}$$

$$\rho_{12} = \nu_{max,H2} \frac{S_{h2}}{K_{S,h2} + S_{h2}} X_{h2} I_1 \tag{22}$$

$$\rho_{13} = K_{dec,X_{su}} X_{su} \tag{23}$$

$$\rho_{14} = K_{dec,X_{aa}} X_{aa} \tag{24}$$

$$\rho_{15} = K_{dec,X_{fa}} X_{fa} \tag{25}$$

$$\rho_{16} = K_{dec,X_{c4}} X_{c4} \tag{26}$$

$$\rho_{17} = K_{dec,X_{pro}} X_{pro} \tag{27}$$

$$\rho_{18} = K_{dec,X_{ac}} X_{ac} \tag{28}$$

$$\rho_{19} = K_{dec,X_{h2}} X_{h2} \tag{29}$$

The inhibition terms in eqs. [18-25], defined according to the ADM1, are reported here:

$$I_1 = I_{pH} I_{IN,lim}$$

$$I_{2,i} = I_{pH} I_{IN,lim} I_{H2,i},\ i = (ac, va, bu, pro)$$

$$I_3 = I_{pH} I_{IN,lim} I_{NH3}$$

Where:

$$I_{pH} = \begin{cases} exp\left(-3\left(\frac{pH-pH_{UL}}{pH-pH_{LL}}\right)^2\right), & pH<pH_{UL}, \\ 0, & pH>pH_{UL}. \end{cases}$$

$$I_{IN,lim} = \frac{1}{1 + K_{S,IN}/S_{IN}}$$

$$I_{H2,i} = \frac{1}{1 + S_{H2}/K_{I,H2,i}},\ i = (ac, va, bu, pro)$$

$$I_{NH3} = \frac{1}{1 + S_{NH3}/K_{I,NH3}}$$

2.2. Acid-base process rate

The kinetic rate equations for all the acid base reactions implemented in the model are here reported:

$$\rho_{A,i+1} = -\rho_{A,i}, \tag{30}$$

$$\rho_{A,1} = K_{A/B,va}(S_{va^-}(S_{H^+}K_{a,va}) - S_{va}) \tag{31}$$

$$\rho_{A,3} = K_{A/B,bu}(S_{bu^-}(S_{H^+}K_{a,bu}) - S_{bu}) \tag{32}$$

$$\rho_{A,5} = K_{A/B,pro}(S_{pro^-}(S_{H^+}K_{a,pro}) - S_{pro}) \tag{33}$$

$$\rho_{A,7} = K_{A/B,ac}(S_{ac^-}(S_{H^+}K_{a,ac}) - S_{ac}) \tag{34}$$

$$\rho_{A,9} = K_{A/B,co2}(S_{hco3^-}(S_{H^+}K_{a,co2}) + K_{a,co2}S_{co3^{2-}} - S_{IC}) \tag{35}$$

$$\rho_{A,11} = K_{A/B,nh4}(S_{nh3}(S_{H^+}K_{a,nh3}) - S_{nh4}) \tag{36}$$

2.3. Gas-transfer process rate

According to the ADM1 model the liquid-gas transfer processes for the variables S_{H2}, S_{CH4}, S_{IC} in the liquid phase have been considered.

$$\rho_{T,1} = K_{La}(S_{h2} - 16K_{H,h2}p_{gas,h2}), \tag{37}$$

$$\rho_{T,2} = K_{La}(S_{ch4} - 64K_{H,ch4}p_{gas,ch4}), \tag{38}$$

$$\rho_{T,3} = K_{La}(S_{co2} - 16K_{H,co2}p_{gas,co2}), \tag{39}$$

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

3.1. Sources of uncertainty

In ADM1 Model, the state variables integrated through eqns. (4-7) depend on a rather large set of parameters. In this study we selected 37 parameters, belonging to different groups according to their specific biological significance. The groups are: inhibition constants (K_{IH2fa}, K_{IH2c4}, K_{IH2pro}, K_{INH3}); half-saturation values (Ks_{su}, Ks_{aa}, Ks_{fa}, Ks_{c4}, Ks_{pro}, Ks_{ac}, Ks_{h2}); yields of biomass on substrates (Y_{su}, Y_{aa}, Y_{fa}, Y_{c4}, Y_{pro}, Y_{ac}, Y_{h2}); first order decay rates for substrates ($k_{dec,su}, k_{dec,aa}, k_{dec,fa}, k_{dec,c4}, k_{dec,pro}, k_{dec,ac}, k_{dec,h2}$); Monod maximum specific growth rate ($\nu_{max,su}$, $\nu_{max,aa}$, $\nu_{max,fa}$, $\nu_{max,c4}$, $\nu_{max,pro}$, $\nu_{max,ac}$, $\nu_{max,H2}$); first order parameters for hydrolisis ($k_{hyd,ch}$, $k_{hyd,pr}$, $k_{hyd,li}$) and the parameters related to the granularity of the composite particulate material, namely r_0 (from which the specific area a^* depends, see Eq. 3) and K_{sbk}. Uniform distributions have been selected for the 37 input parameters, reported in Table 4 (second column).

3.2. Quantities of Interest

The state of the digestion model described by Equations (4-7) evolves in time $t \in (0, T)$ and it is characterized through the state variable (S_i, X_i, $S_{gas,i}$).

Since the latter set of variables is particularly vast, it is mandatory to focus on a small set of scalar outputs in order to better catch the relation between the uncertain inputs and the behavior of the Digestion Model.

We shall concentrate here on two quantity of interest: the first one, from now on y_1 is the integral of the net CH_4 production profile along the whole time range $(0, T)$:

$$y_1 = \int_0^T S_{gas,CH_4}(\tau)d\tau \qquad [molCH_4d] \tag{40}$$

The second quantity of interest, y_2, is the S_{ac} peak that could alter significantly the pH of the reactor, potentially implying the abortion of the whole process.

In formulas, y_2 reads

$$y_2 = \max_{t\in(0,T)} S_{ac} \qquad [gCOD/l] \tag{41}$$

The choice of this two quantities is due to the necessity to consider the effects of the parameters variability in therms of i) methane production with the aim to optimize the performance of AD plant; ii) acetic acid production, since this represents the main intermediate of reaction and play a key role in evaluation of pH and thus on the occurrences of undesired acidification.

Notably, separate analysis are performed for the two different objectives, isolating different set of parameters that share an influence on the variability of the two quantities of interest.

3.3. Description of test case

Bio-methane potential (BMP) tests are robust and reliable experimental methods, mainly thanks to their easy set up and conduction as well as the useful information they can provide [23]. The BMP tests are conducted in batch conditions, and are finalized on the measuring the maximum amount of bio-methane produced per unit of substrate (e.g. COD or VS basis) used

for the anaerobic digestion process. The average length of 30 days, the ease of conduction, the reproducibility and the relatively low implematation costs, make the BMP tests the most used in studies concerning the anaerobic digestion of organic matrices. Moreover, BMP tests give significant information about the bio-methanation of specific substrates and provide experimental results essential to calibrate and validate mathematical models [23]. It is for the aforementioned reasons that BMP experiments were simulated. Standard one Liter glass digesters have been considered, fed with known concentration of substrate, expressed as COD (Xc) and Inoculum, the latter expressed as concentrations of the six microbial species taken into account into the model (i.e. X_{su}, X_{aa}, X_{fa}, X_{c4}, X_{pro}, X_{ac}, X_{h2}).

The choice of input substrate plays a key role. In the proposed research, the substrate has been selected in order to represent Organic Fraction of Municipal Solid Waste (OFMSW). It is for this reason that we adopt a modified version of ADM1, the model proposed by *Esposito and co-authors*. Our goal is to focus on anaerobic digestion process of organic waste rather than on sewage sludge. In this perspective, the role of the parameters r_0 and K_{sbk} is expected to be crucial.

Batch conditions have been assumed. Initial conditions are reported in Table 1 all the kinetic constant and stoichiometric parameters have been taken according to [3, 10]. Based on experimental results, a total elapsed simulation time of forty days is deemed necessary for the complete depletion of the substrates and the achievement of the steady state condition.

Table 1: Initial condition used in numerical simulations

Parameter	Description	Dimension	value
Xc_0	OFMSW	$kgCODm^{-3}$	10
$Ssuin$	Sugar	$kgCODm^{-3}$	0
$Saain$	Amino Acids	$kgCODm^{-3}$	0
$Sfain$	Fatty acids	$kgCODm^{-3}$	0
$Svain$	Valerate	$kgCODm^{-3}$	0.001
$Sbuin$	Butyrate	$kgCODm^{-3}$	0.001
$Sproin$	Propionate	$kgCODm^{-3}$	0
$Sacin$	Acetate	$kgCODm^{-3}$	0
SH_2in	Hydrogen	$kgCODm^{-3}$	0
Sch_4in	Methane	$kgCODm^{-3}$	0
$SICin$	Inorganic Carbon	$kmolm^{-3}$	0.0001
$SINin$	Inorganic Nitrogen	$kmolm^{-3}$	0.05
$Siin$	Soluble Inert	kgm^{-3}	0
$Xchin$	Carbohydrates	$kgCODm^{-3}$	0
$Xprin$	Proteins	$kgCODm^{-3}$	0
$Xliin$	Lipids	$kgCODm^{-3}$	0
$Xsuin$	Carbohydrates degraders	$kgCODm^{-3}$	0.15
$Xaain$	Amino acids degraders	$kgCODm^{-3}$	0.10
$Xfain$	Fatty acids degraders	$kgCODm^{-3}$	0.10
$Xc4in$	Valerate and butyrate degraders	$kgCODm^{-3}$	0.01
$Xproin$	Propionate degraders	$kgCODm^{-3}$	0.033
$Xacin$	Acetate degraders	$kgCODm^{-3}$	0.9
$Xh2in$	Hydrogen degraders	$kgCODm^{-3}$	0.1
$Xiin$	Particulate inert	kgm^{-3}	0
$fchxc$	Fraction of carbohydrates from composites	-	0.20
$fprxc$	Fraction of proteins from composites	-	0.20
$flixc$	Fraction of lipids from composites	-	0.25
$fxixc$	Fraction of particulate inerts from composites	-	0.25
$fsixc$	Fraction of soluble inerts from composites	-	0.10

3.4. Experimental designs and databases

A design of experiments refers to the way of discretizing the space of the uncertain parameters (also referred to as "hypercube") $Z_\Theta \in \mathbb{R}^d$ (in this work, $d = d_i \leq 37$), in which the parameters $\boldsymbol{\theta}_i$ evolve. It is a way to define the N realizations of parameters $\boldsymbol{\theta}_i$, for which the ADM1-based model is integrated as a "black-box" in order to obtain the ensemble of N functional outputs $\mathbf{y}$ from which useful statistics can be extracted. For each $\boldsymbol{\theta}_i$, the

ensemble forms a database $\mathcal{D}_{N_i}$:

$$\mathcal{D}_{N_i} = \left\{ \left(\boldsymbol{\theta}^{(l)}, \mathbf{y}^{(l)} \right)_{1 \leq l \leq N} \right\}, \tag{42}$$

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

It is pointed out that in this work the formalism of SA is adopted, and thus the word "input" stands for the set of uncertain parameters whose effect on the model output is investigated, and not for the feeding characteristics of the reactor. In the present work, the parameter set corresponding to the first (screening) stage $\boldsymbol{\theta}_{\text{Morris}}$ has a cardinality of $N_M = 380$ and is compiled by the randomized algorithm proposed in [24]. On the other hand, concerning each second-stage parameter set $\boldsymbol{\theta}_{CH_4,S_{ac}}$, two databases of size $N = 2^{10}$ are compiled by making use of quasi-Monte Carlo sampling methods. They rely on low-discrepancy sequences to explore the hyperspace given by the support of the d_i Probability Density Functions (PDFs) without any bias with the aim of capturing most of the variance, see e.g. [25]. The first database of each pair is built using Halton's sampling and is used as a training set; it corresponds to the ensemble of simulations over which the different surrogates are trained. The second database of each pair is built using Faure's sampling, to be used as a validation set, i.e. it is made of the ensemble of simulations that is not part of the experimental design. The validation set is used in a subsequent stage to evaluate the accuracy of the different surrogate techniques. The compiled databases are listed in Table 2.

It is remarked that that the considered digestion model features nonlinearities for both QoI $y_{1,2}$. Figure 1 presents 40 representative ADM1 snapshots

sampled from Morris Database. In particular, the CH_4 cumulative profiles in Figure 1 (a) and S_{ac} profiles in Figure 1 (b) are represented.

Table 2: Datasets $\mathcal{D}_{N_i}$ of ADM1-based model simulations used in this work whether for the sake of performing Morris screening, or building surrogates ("training"), or for validating them ("validation").

Sampling Strategy	Purpose	Sample size
	$\boldsymbol{\theta} = \boldsymbol{\theta}_{CH_4}$	
Randomized algorithm of [24]	Morris Screening	380
Halton's sequence	Surrogate Training	2^{10}
Faure's sequence	Surrogate Validation	2^{10}
	$\boldsymbol{\theta} = \boldsymbol{\theta}_{S_{ac}}$	
Randomized algorithm of [24]	Morris Screening	380
Halton's sequence	Surrogate Training	2^{10}
Monte Carlo random sampling	Surrogate Validation	2^{10}

In practice, the model equations 4–39 are integrated with the aid of the `OCTAVE` programming language [26] using `LSODE` solver for the system of ordinary differential equations [27].

4. Surrogate-Based Sensitivity Analysis

For the Sensitivity Analysis of the model presented in Section 2, the problem represented by the large size of the parameter set $\boldsymbol{\theta}$ is outflanked by making use of a preliminary screening analysis adopting the well known *Morris Method*, also called the *Elementary Effect Test* (EET). We remind the reader that in Section 3 we defined two quantities of Interest concerning

the output of ADM model. The initial screening is performed thus twice, i.e. determining the screened set of variables for each observable. This way we obtain two set of parameters, $\boldsymbol{\theta}_{CH_4}$ and $\boldsymbol{\theta}_{S_{ac}}$. For each $\boldsymbol{\theta_i} \in \{\boldsymbol{\theta}_{CH_4}, \boldsymbol{\theta}_{S_{ac}}\}$, the variables θ_i contained in $\boldsymbol{\theta} \setminus \boldsymbol{\theta}_i$ are set to nominal values, namely the average value of the uniform distribution related to θ_i. $\boldsymbol{\theta}_{CH_4}$ and $\boldsymbol{\theta}_{S_{ac}}$ will undergo an exhaustive surrogate-based GSA, see Subsection 4.2.

4.1. Screening of influential parameters via Morris' Scheme

In [24] Morris proposed an effective screening sensitivity measure in order to identify the most important factors in models characterized by many input parameters. Such method consists on the computation for each input of a set of incremental ratios, namely elementary effects, which are then averaged to determine the overall importance of each input parameter.

Here, the mean of r *Elementary Effects* (or EEs) is taken as a measure of global sensitivity. The experimental plan is built by making use of randomized One-At-Time (OAT) experiments. In the following, input parameters are assumed to be uniformly distributed in $[0, 1]$ and then transformed from the unit hypercube to their respective distribution.

For a given value of $\theta_i \in \boldsymbol{\theta}$, the associated elementary effect EE_i reads

$$EE_i = \frac{y(\theta_1^*, \ldots, \theta_i^* + \delta_i, \ldots, \theta_d^*) - y(\theta_1^*, \ldots \theta_i^*, \ldots, \theta_d^*)}{\delta_i}, \tag{43}$$

where $\delta \in \left\{ \frac{1}{n_l - 1}, 1 - \frac{1}{n_l - 1} \right\}$, n_l is the number of levels, $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_d^*)$ is a randomly selected value in the hypercube Z_θ such that the point $(\boldsymbol{\theta}^* + \boldsymbol{e}_i \delta)$ still maps to a point in Z_θ for each $i \in 1, \ldots, d$ and $\boldsymbol{e}_i$ is a vector of zeros except for its i-th component $e_i = 1$.

The empirical distribution of elementary effects EE_i for each input parameter θ_i is obtained with a random sampling of $\boldsymbol{\theta}$, s.t. $EE_i \sim F_i$.

The sensitivity measures proposed by *Morris* [24], μ_i and σ_i, are respectively the mean and the standard deviation of the distributions F_i . To estimate these quantities, Morris suggested sampling r elementary effects from each F_i via an efficient design that constructs r trajectories of $(d+1)$ points in the input space, each providing d elementary effects, one per input factor. This algorithm would thus require $r(d+1)$ model evaluations. An alternative measure proposed by Campolongo and Saltelli [28] consists of taking instead of the mean μ_i the absolute value of the EEs to avoid that differences of different signs would cancel out,

$$S_i = \mu^*_{\text{Morris}} = \frac{1}{n}\sum_{j=1}^{n} EE^j = \frac{1}{n}\sum_{j=1}^{n} \left| \frac{y(\theta_1^j, \dots \theta_i^j + \delta_i^j, \dots, \theta_d^j) - y(\theta_1^j, \dots \theta_i^j, \dots, \theta_d^j)}{\delta_i^j} \right| c_i \tag{44}$$

Besides the above sensitivity measure, as already mentioned it is common practice to also compute the standard deviation of the EEs, which provides information on the degree of interaction of the i-th input factor with the others, and on the non-linearity of the forward model $\mathcal{F}$. A high standard deviation indicates that a factor is interacting with others because its sensitivity changes across the variability space due to the different values assumed buy the other θ_is.

4.2. Surrogate Modeling

We present now the methodology to build an emulator of the ADM1 based model of Section 2, adopting two distinct families of algorithms, namely

generalized polynomial chaos (gPC) expansion or Gaussian Process (GP) model [29]. The key idea of both approaches is to express for each quantity of interest $y = y_1, y_2$ a surrogate by making use of a (finite) sum of basis functions. In formulas, we have:

$$y = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} \gamma_{\boldsymbol{\alpha}} \, \Psi_{\boldsymbol{\alpha}}. \tag{45}$$

In the last formula, the coefficients $\{\gamma_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathcal{A}}$ and the basis functions $\{\Psi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathcal{A}}$ need to be calibrated using the information provided by the Halton's training set $\mathcal{D}_N$ with $N = 2^{10}$ (see Section 3.4).

The coefficients are obtained in different ways depending on the adopted methodology. gPC-expansion and GP model are explained in detail in the Appendix, see Appendix A.1 and Appendix A.2. In this work, three algorithms are tested: two variations of gPC expansion (with linear and sparse truncation scheme via Least Angle Regression) and an implementation of GP.

4.2.1. Workflow for gPC-expansion

The algorithm to compute a gPC-expansion can be resumed as follows:

1. choose the polynomial basis $\{\Psi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathcal{A}}$ according to the prescribed input marginal PDFs of the inputs $\boldsymbol{\theta}_i \in \{\boldsymbol{\theta}_{CH_4}, \boldsymbol{\theta}_{S_{ac}}\} \in \mathbb{R}^d$ $(d = 6, 8)$;
2. choose the total polynomial order P according to the complexity of the digestion 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. if LAR sparse truncation is selected, compute a 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. compute the coefficients $\{\gamma_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathcal{A}}$ with least-square minimization, using $N = 2^{10}$ 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 $\boldsymbol{\theta}^*$.

4.2.2. Workflow for GP surrogate

The scheme of the construction of the GP surrogate is summarized in the following:

1. choose the kernel function $\pi_{\boldsymbol{\alpha}}$ suitable for the input vector $\boldsymbol{\theta}_i \in \{\boldsymbol{\theta}_{CH_4}, \boldsymbol{\theta}_{S_{ac}}\} \in \mathbb{R}^{d_i}$ $(d = 6, 8)$ – we consider RBF in the present study, see Eq. (A.9);
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 $\boldsymbol{\theta}^*$ using Eq. (A.7) and Eq. (A.10).

4.3. Numerical implementation

In practice, the implementation of Morris Screening, gPC-expansion and GP-model relied on the Python package *OpenTURNS* [30] (see www.openturns.org).

5. Results

5.1. Initial Screening via Morris Method

Figure 2 shows the output of the Morris screening procedure for both QoIs. For both observables, parameters with a large value of μ^* are more

likely to have a large σ. That means that shifting the attention towards high μ^* variables would allow in principle to concentrate not only on the parameters that share a more pronounced effect on the observables, but also on the more nonlinear ones, and on the ones that are more prone to interactions with other factors. The ranking between parameters given by this screening method can be found on table 4.

Fixing a threshold value μ_T^* for both y_1 and y_2 cases (0.01 and 0.2, respectively) and retaining only the parameters that trespass such value allow for the definition of the *second-stage* parameter vectors, namely $\boldsymbol{\theta}_{CH_4} \equiv (r_0,\ k_{hyd,ch},\ k_{hyd,pr},\ K_{sbk},\ k_{hyd,li},\ \nu_{max,CH_4})$ and $\boldsymbol{\theta}_{S_{ac}} \equiv (\ \nu_{max,ac},\ r_0,\ k_{hyd,li},\ K_{sbk},\ k_{hyd,pr},\ k_{m,h2},\ K_{I,NH3},\ k_{hyd,ch})$.

The Morris screening procedure provided interesting results. Such findings are in line with the fact that when the Anaerobic Digestion process is applied to a more complex substrate, the bottleneck of the process is represented by the disintegration and hydrolysis steps. The simple first order kinetics approach used in the first version of ADM1 model is sufficient to model the disintegration and hydrolysis of sewage sludge, but it is not appropriate when confronting a more complex and heterogeneous substrate like the organic fraction of municipal solid waste.

Similarly when the QoI is the S_{ac}, with the obvious exception of $\nu_{max,ac}$, which direct affect the bio degradation of the acetic acid, the kinetics parameters r_0, and K_{sbk} are the more sensitive.

5.2. A posteriori error estimation of the surrogate models

The construction of the surrogate model may introduce an approximation error, which can be computed in an *a posteriori* fashion as

$$\epsilon_{\text{emp}} = \frac{1}{N_{\text{halton}}} \sum_{l=1}^{N_{\text{halton}}} \left(y^{(l)} - \widehat{y}^{(l)} \right), \tag{46}$$

with $y^{(l)}$ the lth element of the training set, $\widehat{y}^{(l)}$ the corresponding prediction by the surrogate model (gPC or GP), and $N = 2^{10}$ (see Table 2. However, this estimator for the metamodel error suffers from *overfitting* issues and may severely understimate the mean square error [31]. In addition, the GP-model can be considered an interpolator method at the training set points and therefore it will always achieve $\epsilon_{\text{emp}} = 0$ (when no noise is added to the kernel). In the following, for any tested metamodel algorithm and configuration, we have $\epsilon_{\text{emp}} < 2.0 \times 10^{-3}$.

To circumvent these shortcomings, the surrogates are validated using the so called Q_2 predictive coefficient, that corresponds to a cross-validation error metric using the independent dataset based on Faure's low discrepancy sequence (again, see Table 2):

$$Q_2 = 1 - \frac{\sum_{l=1}^{N_{\text{faure}}} \left(y^{(l)} - \widehat{y}^{(l)} \right)^2}{\sum_{l=1}^{N_{\text{faure}}} \left(y^{(l)} - \overline{y} \right)^2}, \tag{47}$$

with $\overline{y}$ the empirical mean over the Faure's validation set ($N_{\text{faure}} = 2^{10}$). Thus, Q_2 furnishes a normalized estimate of the generalization error, i.e. the surrogate error when considering points outside of the Halton's training set [32]. The target value for Q_2 is 1: the closer the result to unity, the better is the surrogate.

The Q_2 indicator performs thus the function of ranking the surrogates by their effectiveness in reproducing the dynamics of the studied ADM1-based model. In particular, when gPC techniques is applied, we consider the results of the surrogate with total polynomial order P that gives the best results. For $\boldsymbol{\theta}_{CH_4}$ ($d = 6$), P varied from 1 to 14, while for $\boldsymbol{\theta}_{S_{ac}}$ ($d = 8$) P varied from 1 to 8.

Figure 3, first panel, shows the adequacy plots, i.e. the plots of metamodel computed over points of the DOE over actual forward model $\mathcal{F}$ runs. In the second panel, the robustness of LAR-gPC algorithm with respect to the choice of P is given by the plots of Q_2 values over the tested values for the maximum polynomial degree. In Table 3, the different error estimators for the adopted surrogate techniques are tabulated.

Table 3: Errors relative to built surrogates. For LAR-gPC and SLS-gPC, the best results for the spanned values for P are reported.

Metamodel	Q_2	ϵ_{emp}
	$\boldsymbol{\theta} = \boldsymbol{\theta}_{CH_4}, y = y_1$	$d = 6$
SLS-gPC ($P = 5$)	0.938	$5.63e-05$
LAR-gPC ($P = 13$)	0.996	$9.57e-06$
GP (RBF Kernel)	0.910	0.
	$\boldsymbol{\theta} = \boldsymbol{\theta}_{S_{ac}}, y = y_2$	$d = 8$
SLS-gPC ($P = 4$)	0.982	$1.14e-03$
LAR-gPC ($P = 5$)	0.988	$1.21e-03$
GP (RBF Kernel)	0.985	0.

5.3. Quantitative SA with Sobol' indices

Sobol' indices [33, 34] are commonly used for variance decomposition-based global sensitivity analysis. They provide the quantification of how much of the variance in the quantity of interest y is caused by the spread in a single uncertain input parameter (or a group of them), assuming these random variables to be independent. In this framework, the variance of the output random variable Y denoted by $\mathbb{V}[Y]$ is decomposed as

$$\mathbb{V}[Y] = \sum_{i=1}^{d} \mathbb{V}_i(Y) + \sum_{j=i+1}^{d} \mathbb{V}_{ij}(Y) + \cdots + \mathbb{V}_{1,2,\ldots,d}(Y), \tag{48}$$

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

$$\mathbb{V}_I(Y) = \mathbb{V}\left[\mathbb{E}(Y|\Theta_I)\right] - \sum_{J \subset I \text{ s.t. } J \neq I} \mathbb{V}_J(Y), \ \forall I \subset \{1, \ldots, d\} \tag{49}$$

By making use of this variance decomposition, the first-order Sobol' index S_i associated with the ith parameter of $\mathbf{\Theta}$ reads

$$S_i = \frac{\mathbb{V}_i(Y)}{\mathbb{V}(Y)}. \tag{50}$$

It corresponds to the ratio of the output variance $\mathbb{V}(Y)$ that is *uniquely* linked to the variability in the ith input parameter; S_i ranges between 0 and 1. The corresponding total Sobol' index S_{T_i}, on the other hand, measures the whole contribution of the ith input parameter (including this time interactions with other parameters of $\mathbf{\Theta}$) on the output variance. It is defined as follows:

$$S_{T_i} = \sum_{\substack{I \subset \{1,\ldots,d\} \\ I \ni i}} S_I. \tag{51}$$

By definition, $S_{T_i} \geq S_i$. If first-order and total indices do not coincide, this means that the input parameter θ_i has some interactions with other parameters of $\boldsymbol{\Theta}$ to describe the output variance. For the GP-surrogate approach, Sobol' indices are estimated stochastically adopting Martinez' formulation as a stable estimator [35]. 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

$$S_{i,\mathrm{pc}} = \frac{1}{\sigma_y^2} \sum_{\substack{\boldsymbol{\alpha}\in\mathcal{A}, \\ \alpha_i>0 \text{ and } \alpha_{k\neq i}=0}} \gamma_{\boldsymbol{\alpha}}^2, \tag{52}$$

with σ_y the output STD computed using Eq. (54).

Figure 4 presents the first-order and total Sobol' indices obtained with the three adopted algorithms, related to the two different set of input parameters $\boldsymbol{\theta}_{S_ac,CH_4}$. However, since the best performing algorithm with respect to Q_2 error has been LAR-gPC for both sets of input variables, in the following we shall discuss only the SA and UQ results concerning this surrogate.

For both input parameters studies, it is worth noting that first-order and total Sobol' indices are not identical, implying that some interactions take place between the factors.

Regarding CH_4 production, K_{sbk} and r_0 are the two most influential parameters, while for S_{ac} peak the most important parameter is $\nu_{max,ac}$, followed by K_{sbk}, r_0 and $K_{I,NH3}$. As mentioned by *Esposito and co-authors* in[23] and experimentally demonstrated in [36], with complex substrate such as the OFMSW, disintegration and hydrolysis represent the bottlenecks of the AD process. About the second QoI the acetate peak while $\nu_{max,ac}$ it is to be expected as sensitive parameter, the presence of $K_{I,NH3}$ in the set of the

relevant parameters is rather unexpected. It is worth interest noting that r_0 and k_{sbk} occupy high rank position in this global Sensitivity Analysis for two different outputs, suggesting the fact that in a general process of calibrating an ADM1-based model they play an overall paramount role.

5.4. Graphical insights through Cobweb Plots

When dealing with large sets of sensitivity indices, the interpretation of results that come from a SA procedure can be enhanced by ad hoc visualization tools, see [37].

In Figure 5, a *cobweb plot*, also known as *parallel coordinate plot* of the Halton databases for the two input factor sets of the second phase of SA is portrayed. The range on the y axis is normalized. Every configuration of input parameters $\boldsymbol{\theta}$ contained in $\mathcal{D}_{N_i}$ is reported in grey, and the ones that give rise to an observable $y_{1,2}$ belonging to bottom 10% (top panels for each color) or top 10% (bottom panels for each color) are reported in red-blue color.

If highlighted lines cover the entire range of a certain factor, the latter has a negligible influence on determining the most extreme behavior of the considered observable. On the contrary, if they concentrate on a subrange, the sensitivity of the parameters for the extreme (high or low, depending on the selected threshold criteria) observable response is high.

In particular, Figure 5 shows that concerning CH_4 production, K_{sbk} is very influential in determining a low output, while in the case of high CH_4 production, low r_0 is almost always observed. K_{sbk} models the disintegration capability of the substrate. The results reported in Figure 5(a) state that low productions of methane can follow only from substrates characterized by

low value of K_{sbk}. As expected, high methane productions can be obtained only with low values of r_0 5 (b). This means that with the trituration (i.e. granulometry reduction) of the OFMSW it is possible to increase the methane production of the substrate in agreement with experimental evidences.

Shifting the attention to S_{ac} peak, a low value of $\nu_{max,ac}$ is somewhat mandatory for a high peak of S_{ac}; on the other hand, for a low value of S_{ac} peak, K_{sbk} is almost always needed to lie in the lowest part of its range, and $\nu_{max,ac}$ is required to be in the upper half of its range. According with experimental results where the hydrolysis represents the bottleneck of the process, with low hydrolysis rates the system is characterized by low acid concentration values.

The results of this graphical method are not only in agreement with the more quantitative results given by Sobol' Indices, but also complementary to them, since we could spot the ranges of high rank input parameters that produce a significant high (or low) output in the observables.

5.5. Uncertainty Quantification

For the sake of Uncertainty Quantification, we restrict the study at the surrogate that behaved best with respect to the Q_2 error estimator, that is LAR-gPC. In the framework of LAR gPC-expansion, the statistics of the two quantities of interest $y_{1,2}$ 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 respectively read

$$\mu_y = \gamma_0, \tag{53}$$

$$\sigma_y = \sqrt{\sum_{\substack{\boldsymbol{\alpha}\in\mathcal{A}\subset\mathbb{N}^d \\ \boldsymbol{\alpha}\neq 0}} \gamma_{\boldsymbol{\alpha}}^2}. \tag{54}$$

The PDF of each quantity of interest is computed using kernel smoothing techniques by sampling the uncertain input space Z_Θ (with sample size of 10,000 members) adopting a Monte Carlo random sampling and by evaluating the LAR gPC-expansion for all these points.

Figure 6 presents the PDF of CH_4 net production (first panel) and of S_{ac} peak magnitude (second panel), while the moments of the latter PDFs are shown in Table 5. Some information can be deducted from the shape of such PDFs (and more concretely from the higher order moments of their respective distributions). Both distributions are characterized by a non-zero skewness. In particular, CH_4 distribution exhibits a sharp drop after its maximum value. This corresponds to the fact that there is a sharp restriction for CH_4 production in the batch setting of the proposed test case, the one dictated by mass conservation: it is impossible to generate more CH_4 than the theoretical maximum which is strictly dependent on the initial amount of mass in the BMP reactor. The fact that the obtained PDF of quantity y_1 is heavily peaked around the value of 0.30 is in line with the plot of 40 trajectories of Fig. 1 (first panel), with the majority of trajectories exhibiting a sharp increase in the first days of elapsed time.

On the other hand, S_{ac} peak distribution exhibits a positive skewness and a rather large support. The right tail shows that the occurrence of a high peak should not be underestimated for a wide set of input parameters configurations. This is inline with experimental evidence that failures in anaerobic digestive processes are often due to high concentration in volatile acids.

6. Conclusions

In this paper, uncertainty quantification and global sensitivity analysis non-intrusive methods were applied to a modified version of the ADM-1 model for a test case of engineering relevance that depended on a rather large set of input parameters of different nature and lacked up to now a GSA and UQ for two main observables that are of great interest for the practitioners, related to CH_4 production and volatile acid peaks in the reactor.

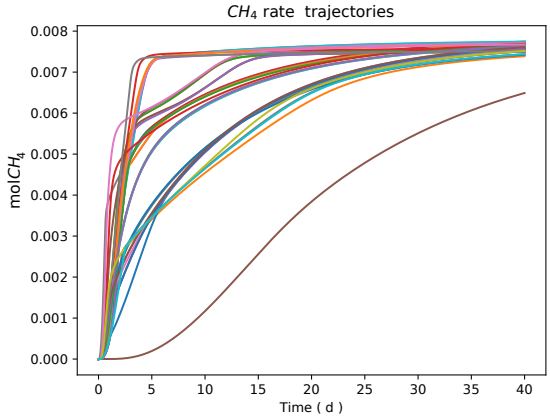
The overall relevance of the present work is two-fold: on the one hand, it gives useful insights of the bio-physical and chemical processes that play a major role in AD. On the other hand, on a methodological level, it gives an example of successful GSA and UQ procedures with preliminary screening and subsequent Surrogate-Based analysis, which received help from graphical methods in order to give further insights.

Spotting the most sensitive parameters provides information on the most important phenomena of the process and would allow the experimental researchers to focus the efforts on their calibration. In particular, the role of the parameters r_0 and K_{sbk} resulted to be crucial for the whole set of QoI adopted. Such parameters are related respectively to mechanical and chemical pre-processing of the municipal solid waste before the start of the AD process. Their importance in the ADM model thus confirms that the designer of an AD procedure has at his disposal several strategies to otpimize the overall process.

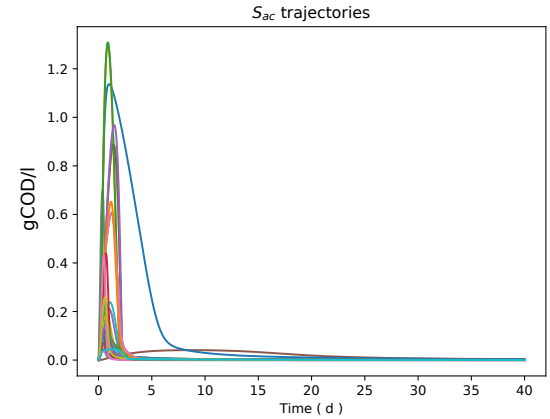
Ultimately, this GSA and UQ procedure may also inspire further efforts in order to reduce the number of model equations, obtaining simplified models.

Acknowledgements

AT has been supported by the Basque Government through the BERC 2018–2021 program, by the Spanish Ministry of Economy and Competitiveness (MINECO) through BCAM Severo Ochoa excellence accreditation and SEV-2017-0718, and by the PhD Grant "La Caixa 2014". LF has been supported by Fondazione Cariplo, progetto VOLAC Grant numbers: 2017-0977.This paper has been performed under the auspices of the G.N.F.M. of I.N.d.A.M.



(a) S_{gas,CH_4}



(b) S_{ac}

Figure 1: An ensemble of 40 different profiles extracted from Halton sampling database, with different values of θ.

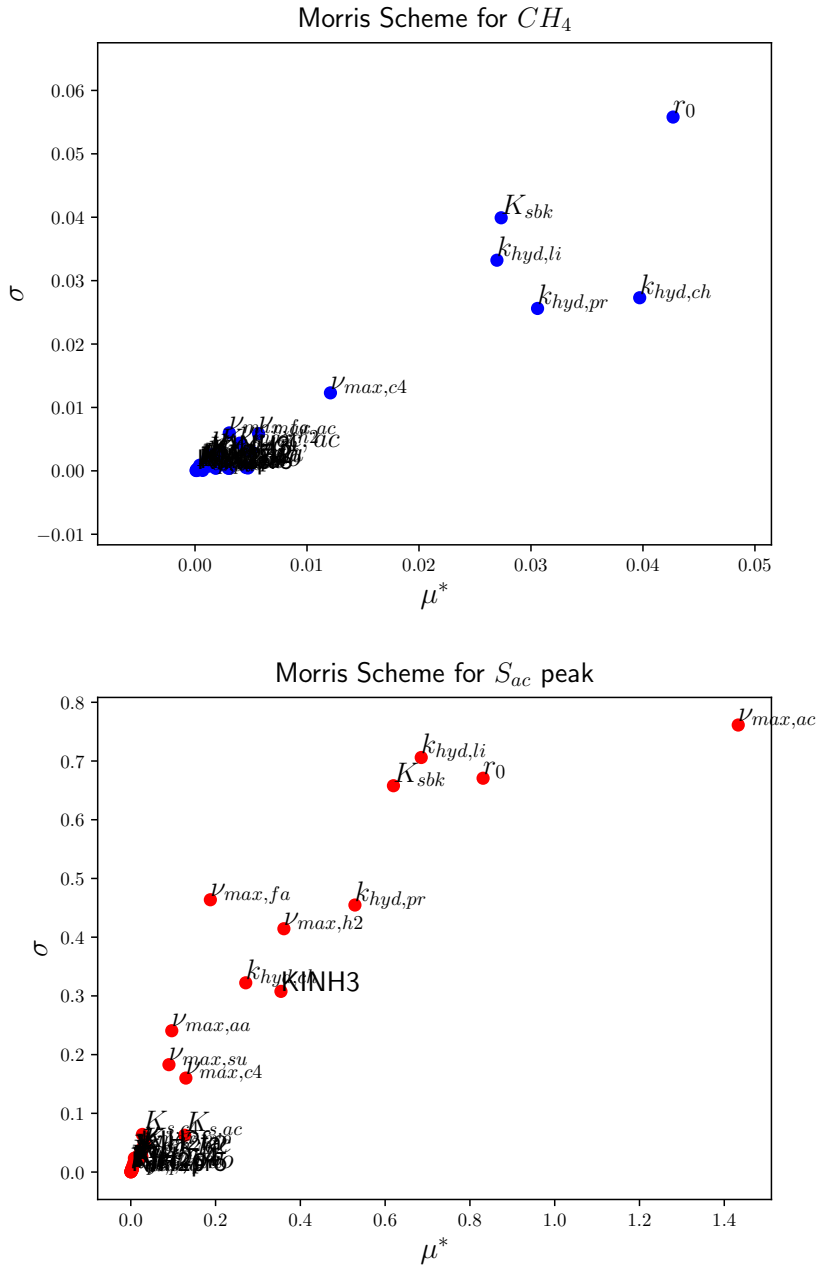
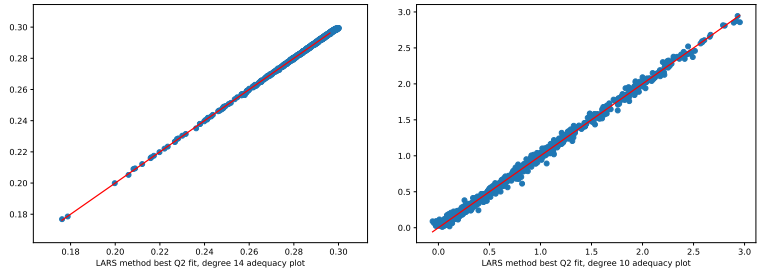
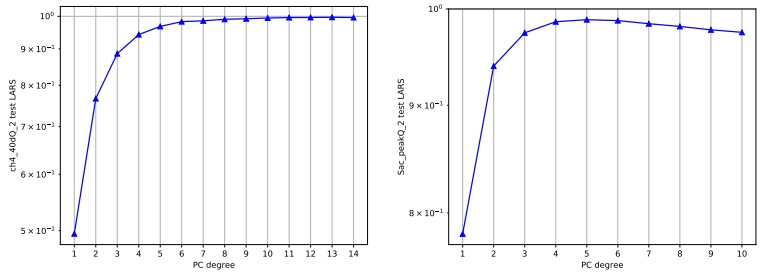


Figure 2: Morris algorithm applied with respect to y_1 (top) and y_2 (bottom) observables.

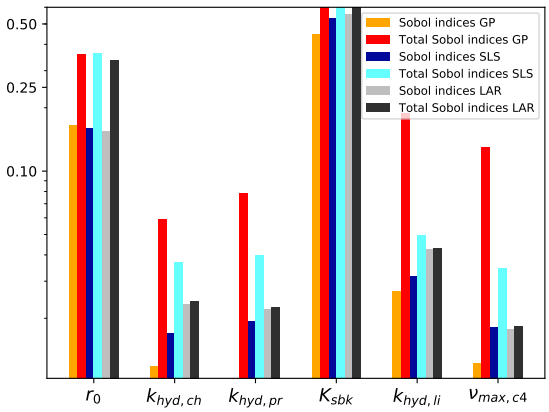


(a) $\boldsymbol{\theta}_{CH_4}$, $y = y_1$ (left),$\boldsymbol{\theta}_{S_{ac}}$, $y = y_2$(right), adequacy plot for LAR gPC algorithm

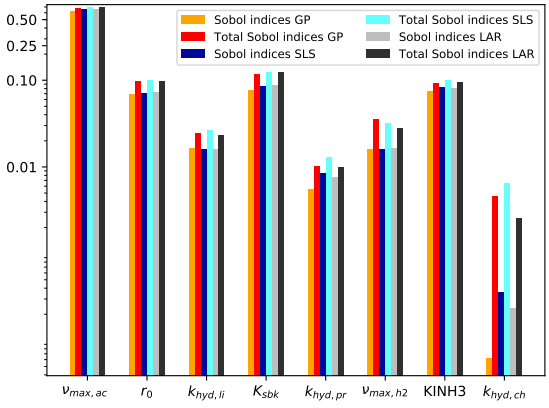


(b) $\boldsymbol{\theta}_{CH_4}$, $y = y_1$ (left),$\boldsymbol{\theta}_{S_{ac}}$, $y = y_2$(right) Q_2 test for varying maximum order of gPC from $p_0 = 1$ to $p_1 = 8, 14$

Figure 3: Adequacy plots for both $\boldsymbol{\theta} \in \{\boldsymbol{\theta}_{S_{ac}}, \boldsymbol{\theta}_{CH_4}\}$ parameter sets for the quantities of interest related to each set, namely y_1 the CH_4 net production at $t = 40d$ and y_2 the magnitude of S_ac peak. For the Q_2 test, we showed the results of several maximum degrees p of LAR-based gPC algorithm.

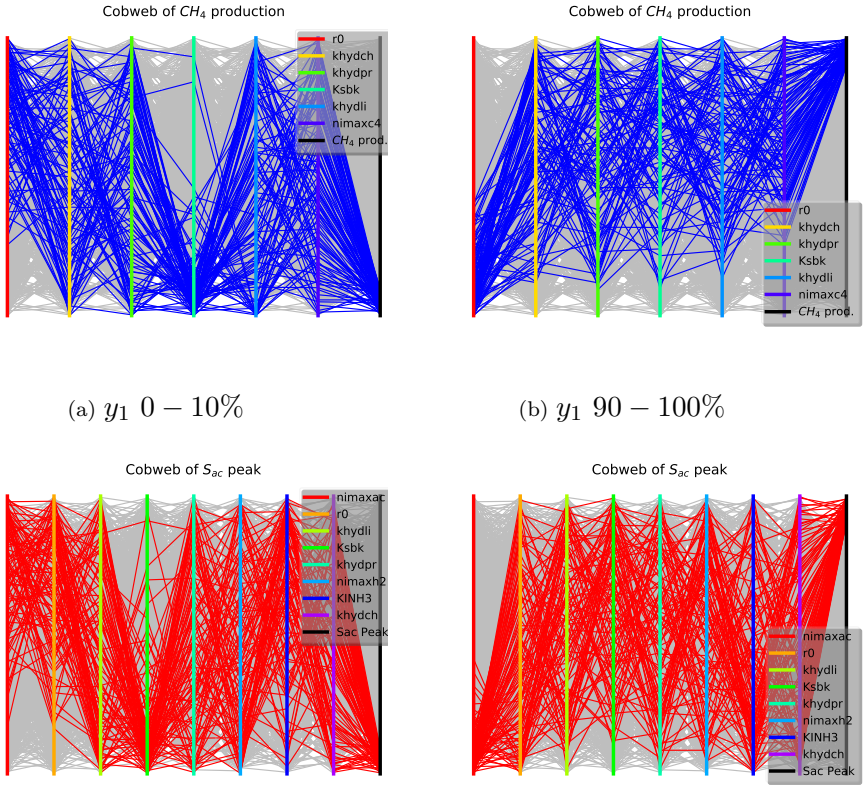


(a) $\boldsymbol{\theta}_{CH_4}$, $y = y_1$



(b) $\boldsymbol{\theta}_{S_{ac}}$, $y = y_2$

Figure 4: First-order and total Sobol' indices (in logarithmic scale) associated with uncertain parameters $\boldsymbol{\theta} \in \{\boldsymbol{\theta}_{S_{ac}}, \boldsymbol{\theta}_{CH_4}\}$ for the quantities of interest related to each set, namely y_1 the CH_4 net production at $t = 40d$ and y_2 the magnitude of S_ac peak. For each panel, the three different tested algorithm are presented. For GP, orange colors correspond to first-order Sobol' indices; red colors correspond to total Sobol' indices. For SLS-gPC, light blue colors correspond to first-order Sobol' indices; dark blue colors correspond to total Sobol' indices. Finally, for LAR-gPC, gray colors correspond to first-order Sobol' indices; dark gray colors correspond to total Sobol' indices.



(a) y_1 $0-10\%$

(b) y_1 $90-100\%$

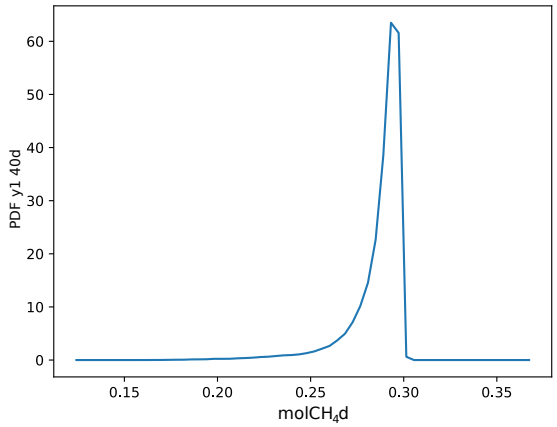
(c) y_2 $0-10\%$

(d) y_2 $90-100\%$

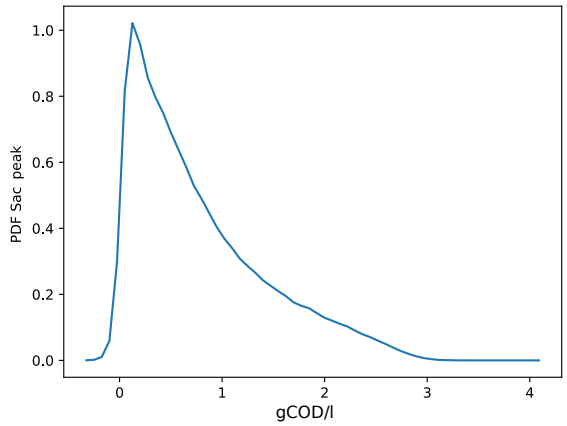
Figure 5: Cobweb plot related to Halton Databases for $\boldsymbol{\theta} = \boldsymbol{\theta}_{CH_4}, \boldsymbol{\theta}_{S_{ac}}$ respectively. In the left column we have cobweb plots where the parameter combinations that give a bottom 10% of their respective quantity of interest $y_{1,2}$ are reported in blue (y_1) and red (y_2), while in the right hand side panels the colored lines correspond this time to top 10% of the related quantity of interest $y_{1,2}$.

Table 4: In the second column the complete list of Uniform marginal PDFs associated with vector θ_{Morris} is reported. Note that $\mathcal{U}(a, b)$ stands for the uniform distribution with a the minimum value of the parameter and b the maximum one. The last four columns show the ranking of the parameters according to different QoI and different stages of the SA procedure.

Parameter	Uniform distribution	Morris rank (CH_4)	Morris rank (S_{ac})	gPC rank (CH_4)	gPC rank (S_{ac})
r_0	$\mathcal{U}(0.001, 0.05)$	1	2	2	3
$Ksbk$	$\mathcal{U}(1.0, 20.0)$	4	4	1	2
$k_{hyd,ch}$	$\mathcal{U}(0.1, 10.0)$	2	8	4	8
$k_{hyd,pr}$	$\mathcal{U}(0.1, 10.0)$	3	5	5	7
$k_{hyd,li}$	$\mathcal{U}(0.1, 10.0)$	5	3	3	6
$\nu_{max,su}$	$\mathcal{U}(5.0, 100.0)$	16	13	-	-
$\nu_{max,aa}$	$\mathcal{U}(5.0, 100.0)$	24	12	-	-
$\nu_{max,fa}$	$\mathcal{U}(5.0, 100.0)$	12	9	-	-
$\nu_{max,c4}$	$\mathcal{U}(5.0, 100.0)$	6	10	6	-
$\nu_{max,pro}$	$\mathcal{U}(5.0, 100.0)$	31	14	-	-
$\nu_{max,ac}$	$\mathcal{U}(5.0, 100.0)$	7	1	-	1
$\nu_{max,h2}$	$\mathcal{U}(5.0, 100.0)$	11	6	-	5
$k_{dec,su}$	$\mathcal{U}(0.001, 0.1)$	15	33	-	-
$k_{dec,aa}$	$\mathcal{U}(0.001, 0.1)$	23	37	-	-
$k_{dec,fa}$	$\mathcal{U}(0.001, 0.1)$	21	62	-	-
$k_{dec,c4}$	$\mathcal{U}(0.001, 0.1)$	22	31	-	-
$k_{dec,pro}$	$\mathcal{U}(0.001, 0.1)$	33	36	-	-
$k_{dec,ac}$	$\mathcal{U}(0.001, 0.1)$	8	16	-	-
$k_{dec,h2}$	$\mathcal{U}(0.001, 0.1)$	28	21	-	-
Y_{su}	$\mathcal{U}(0.08, 0.12)$	10	24	-	-
Y_{aa}	$\mathcal{U}(0.064, 0.096)$	14	29	-	-
Y_{fa}	$\mathcal{U}(0.048, 0.072)$	13	28	-	-
Y_{c4}	$\mathcal{U}(0.048, 0.072)$	20	23	-	-
Y_{pro}	$\mathcal{U}(0.032, 0.048)$	26	35	-	-
Y_{ac}	$\mathcal{U}(0.04, 0.06)$	9	30	-	-
Y_{h2}	$\mathcal{U}(0.048, 0.072)$	18	20	-	-
$K_{s,su}$	$\mathcal{U}(0.25, 0.75)$	34	22	-	-
$K_{s,aa}$	$\mathcal{U}(0.15, 0.45)$	36	27	-	-
$K_{s,fa}$	$\mathcal{U}(0.2, 0.6)$	29	19	-	-
$K_{s,c4}$	$\mathcal{U}(0.1, 0.3)$	17	15	-	-
$K_{s,pro}$	$\mathcal{U}(0.05, 0.15)$	37	32	-	-
$K_{s,ac}$	$\mathcal{U}(0.075, 0.225)$	30	11	-	-
$K_{s,h2}$	$\mathcal{U}(3.5e-06, 1.05e-05)$	25	17	-	-
$KIH2_{fa}$	$\mathcal{U}(2.5e-06, 7.5e-06)$	32	18	-	-
$KIH2_{c4}$	$\mathcal{U}(5.0e-06, 1.5e-05)$	27	25	-	-
$KIH2_{pro}$	$\mathcal{U}(1.75e-06, 5.25e-06)$	35	34	-	-
$KINH3$	$\mathcal{U}(0.0009, 0.0027)$	19	7	-	4



(a) $\boldsymbol{\theta}_{CH_4}$, $y = y_1$



(b) $\boldsymbol{\theta}_{S_{ac}}$, $y = y_2$

Figure 6: Probability density functions for $\boldsymbol{\theta} \in \{\boldsymbol{\theta}_{S_{ac}}, \boldsymbol{\theta}_{CH_4}\}$ for the quantities of interest related to each set, namely y_1 the time integral of CH_4 net production up to $t = 40d$ and y_2 the magnitude of S_ac peak.

Table 5: Statistical moments for the PDFs of the two QoI $y_i, i = 1, 2$ subject to variation of $\boldsymbol{\theta} \in \{\boldsymbol{\theta}_{CH_4}, \boldsymbol{\theta}_{S_{ac}}\}$, respectively.

Moment	y_1 (CH_4)	y_2 (S_{ac})
Mean	0.286	0.766
St. Deviation	0.0149	0.647
Skewness	-2.934	1.026
Kurtosis	14.354	3.372

Appendix A. Surrogate Modeling

In this Section, the two approaches adopted in the manuscript to produce surrogate models are described. In Subsection Appendix A.1, generalized Polynomial Chaos is presented, while in Subsection Appendix A.2 Gaussian Process surrogate model is described.

Appendix A.1. Generalized polynomial chaos (gPC) expansion

Appendix A.1.1. Standard probabilistic space

The random vector $\boldsymbol{\Theta}$ is defined in the input physical space. We refer to its counterpart in the standard probabilistic space as $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_d)$, with ζ_i the random variable associated with the ith uncertain parameter Θ_i in $\boldsymbol{\Theta}$ and characterized by a uniform marginal PDF ρ_{Θ_i}. The reduced variable ζ_i is then a uniform variable with support $[-1; 1]$. The gPC-framework applies naturally to the standard probabilistic space. The equivalent of $\boldsymbol{\rho_\Theta}$ in the standard probabilistic space is denoted by $\boldsymbol{\rho_\zeta}$. Since all input random variables are assumed independent (see Section 3.1), the joint PDF $\boldsymbol{\rho_\zeta}$ is the product of the marginal PDFs $\{\rho_{\zeta_i}\}_{i=1,\ldots,d}$.

Appendix A.1.2. Polynomial Basis

The random vector $\boldsymbol{\Theta}$ 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 $\boldsymbol{\rho_\zeta}$. The corresponding inner product reads:

$$\langle \Psi_{\boldsymbol{\alpha}}(\boldsymbol{\zeta}), \Psi_{\boldsymbol{\beta}}(\boldsymbol{\zeta}) \rangle = \int_{Z_\zeta} \Psi_{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\, \Psi_{\boldsymbol{\beta}}(\boldsymbol{\zeta})\, \boldsymbol{\rho_\zeta}\, d\boldsymbol{\zeta} = \delta_{\boldsymbol{\alpha}\boldsymbol{\beta}}, \tag{A.1}$$

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

We assume the model outputs are of finite variance. Hence, $Y \in \{y_1, y_2\}$ can be cast as a function of the reduced variables and expanded as

$$Y(\omega) = \mathcal{F}_{\mathrm{pc}}(\boldsymbol{\Theta}) = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} \gamma_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}} (\boldsymbol{\zeta}(\omega)), \tag{A.2}$$

where $\{\Psi_{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\}_{\boldsymbol{\alpha} \in \mathcal{A}}$ correspond to Legendre polynomials (the latter constitute the the optimal choice for uniform PDFs following the *Askey's scheme* [38]); we define the total polynomial order as P. Since we deal with a finite sum, a truncation strategy is required to determine the appropriate size of the polynomial basis. $\{\gamma_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathcal{A}}$ are the unknowns to determine with a suitable projection strategy to finally obtain the surrogate $\mathcal{F}_{\mathrm{pc}}$.

Appendix A.1.3. Truncation strategy

In practice, the sum in Eq. (A.2) is truncated to a finite number of terms r. In this work, two truncation strategies are compared 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

$$r_{\text{lin}} = (d + P)!/(d!\ P!). \tag{A.3}$$

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

$$\mathcal{A}_{\text{lin}} \equiv \mathcal{A}_{\text{lin}}(d, P) = \{\boldsymbol{\alpha} \in \mathbb{N}^d : |\boldsymbol{\alpha}| \leq P\}, \tag{A.4}$$

where $|\boldsymbol{\alpha}| = ||\boldsymbol{\alpha}||_1 = \alpha_1 + \cdots + \alpha_d$ is the total order of the multi-index. We will refer in this case to the basis as the "full basis" for a given order P.

Sparse truncation strategy. A *sparse truncation* strategy aims at reducing the number of terms in the gPC-expansion for a given total polynomial order P. One method to build a "sparse basis" (by opposition to the "full basis" obtained when considering a linear truncation strategy) is the Least Angle Reduction (LAR) approach. The key argument of the LAR approach is to choose 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 then added to the active set of polynomials. The coefficients of the active basis are computed so that every active polynomial needs to be equicorrelated with the current residual, until convergence. LAR method builds thus a collection of surrogates that are less and less sparse along the iterations. The method stops either when the full basis has been looked through or when the maximum size of the training set N has been reached. Further details are given in Refs. [39, 40, 41].

Appendix A.1.4. Projection strategy

For a given basis, the coefficients $\{\gamma_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathcal{A}}$ are computed through least-square minimization in a non-intrusive way, by making use of the N-snapshots

of the training set $\mathcal{D}_N$. The principal idea of least-square minimization is to minimize the mean square error, i.e. the error of approximation between the ADM1-based model evaluations and the estimations given by the gPC-surrogate at the points of the training set [42].

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

$$\widehat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}\in\mathbb{R}^r}{\operatorname{argmin}} \frac{1}{N} \sum_{l=1}^{N} \left(y^{(l)} - \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \gamma_{\boldsymbol{\alpha}} \, \Psi_{\boldsymbol{\alpha}} \left(\boldsymbol{\xi}^{(l)} \right) \right)^2 , \tag{A.5}$$

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

$$\widehat{\boldsymbol{\gamma}} = \left(\boldsymbol{\Psi}^T \boldsymbol{\Psi} \right)^{-1} \boldsymbol{\Psi}^T \, \mathcal{Y}, \tag{A.6}$$

with $\boldsymbol{\Psi}$ the *information matrix*, which corresponds 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 evaluations of ADM1 model.

If non-sparse truncation is adopted, this projection method is the standard least-square (SLS) approach. In the LAR sparse method, least-square minimization is used to retrieve the set of active coefficients. It is worth noting that LAR allows the gPC-expansion to include high-order polynomials in the basis without leading to an ill-posed problem, providing a way to explore the possible nonlinearity of the model response to the input parameters with a limited budget of simulations.

Appendix A.2. Gaussian Process (GP) surrogate

Appendix A.2.1. Principles

A surrogate of the ADM1-based model using GP regression can be conceived as follows:

$$y\left(\boldsymbol{\theta}\right) = \mathcal{F}_{\mathrm{gp}}(\boldsymbol{\theta}) = \sum_{\boldsymbol{\alpha}=1}^{r} \gamma_{\boldsymbol{\alpha}}\, \Psi_{\boldsymbol{\alpha}}(\boldsymbol{\theta}), \tag{A.7}$$

where $\Psi_{\boldsymbol{\alpha}}$ is a GP calibrated with the data of the training set $\mathcal{D}_N$. This GP is a random process indexed over the domain $\mathbb{R}_i^d$ (here $d_i \in \{d_{S_{ac}}, d_{CH_4}\}$, $d_i \leq 8$), for which any finite collection of process values, $\left\{\Psi_{\boldsymbol{\alpha}}(\boldsymbol{\theta}^{(l)})\right\}_{1\leq l\leq N}$, share a joint Gaussian distribution [43]. Let $\widetilde{\Psi}_{\boldsymbol{\alpha}}$ be a GP fully characterized by its zero mean and its correlation $\pi_{\boldsymbol{\alpha}}$:

$$\widetilde{\Psi}_{\boldsymbol{\alpha}}(\boldsymbol{\theta}) \sim \mathrm{GP}\left(0, \sigma_{\boldsymbol{\alpha}}^2\, \pi_{\boldsymbol{\alpha}}(\boldsymbol{\theta}, \boldsymbol{\theta}')\right), \tag{A.8}$$

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 work, the correlation function π (also named *kernel*) is a squared exponential (also known as radial basis function – RBF), which reads:

$$\pi_{\boldsymbol{\alpha}}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \exp\left(-\frac{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2}{2\,\ell_{\boldsymbol{\alpha}}^2}\right), \tag{A.9}$$

where $\ell_{\boldsymbol{\alpha}}$ is a length-scale which describes the model dependency between the input vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, and where $\sigma_{\boldsymbol{\alpha}}^2$ is the variance of the model output. In practice, the surrogate is obtained as the mean of the GP resulting of conditioning $\widetilde{\Psi}_{\boldsymbol{\alpha}}$ by the training set $\left\{\Psi_{\boldsymbol{\alpha}}\left(\boldsymbol{\theta}^{(l)}\right)\right\}_{1\leq l\leq N}$. For any $\boldsymbol{\theta}^* \in \mathbb{R}^{d_i}$, the prediction of the GP-model is obtained using Eq. (A.7) with the following formulation for the basis function $\Psi_{\boldsymbol{\alpha}}$:

$$\Psi_{\boldsymbol{\alpha}}(\boldsymbol{\theta}^*) = \sum_{l=1}^{N} \boldsymbol{\beta}_{l,\boldsymbol{\alpha}}\, \pi_{\boldsymbol{\alpha}}\left(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(l)}\right), \tag{A.10}$$

where

$$\boldsymbol{\beta}_{l,\boldsymbol{\alpha}} = \left(\boldsymbol{\Pi}_{\boldsymbol{\alpha}} + \tau^2\,\mathbf{I}_N\right)^{-1}\left(\Psi_{\boldsymbol{\alpha}}\left(\boldsymbol{\theta}^{(1)}\right)\ldots\Psi_{\boldsymbol{\alpha}}\left(\boldsymbol{\theta}^{(N)}\right)\right)^T, \tag{A.11}$$

$$\boldsymbol{\Pi}_{\boldsymbol{\alpha}} = \left(\pi_{\boldsymbol{\alpha}}\left(\boldsymbol{\theta}^{(l)},\boldsymbol{\theta}^{(m)}\right)\right)_{1\leq l,m\leq N}, \tag{A.12}$$

and where τ (nugget effect parameter) prevents ill-conditioning issues for the matrix $\boldsymbol{\Pi}_{\boldsymbol{\alpha}}$. The hyperparameters $\{\ell_{\boldsymbol{\alpha}}, \sigma_{\boldsymbol{\alpha}}, \tau\}$ are the result of an optimization by maximum likelihood applied to the dataset $\mathcal{D}_N$ using the Tunrcated-Newton method non-linear optimizer [44].

Appendix B. Petersen Matrix

	Components →	1	2	3	4	5	6	7	8	9	10	11	12	Rate ρ_j
	Process ↓	S_{su}	S_{aa}	S_{fa}	S_{va}	S_{bu}	S_{pro}	S_{ac}	S_{H2}	S_{CH4}	S_{IC}	S_{IN}	S_I	$[\frac{kgCODm^{-3}d^{-1}}{kmolm^{-3}d^{-1}}]$
1	Disintegration												f_{x_i,x_c}	ρ_1
2	Hydrolysis of Carbohydrates	1												ρ_2
3	Hydrolysis of Proteins		1											ρ_3
4	Hydrolysis of Lipids	$1-f_{fa,li}$		$f_{fa,li}$										ρ_4
5	Uptake of Sugars	-1				$(1-Y_{su})f_{bu,su}$	$(1-Y_{su})f_{pro,su}$	$(1-Y_{su})f_{ac,su}$	$(1-Y_{su})f_{H2,su}$		$-\sum_{i=1-9,11-24} C_i\nu_{i,5}$	$-Y_{su}N_{bac}$		ρ_5
6	Uptake of AminoAcids		-1		$(1-Y_{aa})f_{va,aa}$	$(1-Y_{aa})f_{bu,aa}$	$(1-Y_{aa})f_{pro,aa}$	$(1-Y_{aa})f_{ac,aa}$	$(1-Y_{aa})f_{H2,aa}$		$-\sum_{i=1-9,11-24} C_i\nu_{i,6}$	$N_{aa}-Y_{su}N_{bac}$		ρ_6
7	Uptake of LCFA			-1				$(1-Y_{fa}0.7$	$(1-Y_{fa}0.3$		$-\sum_{i=1-9,11-24} C_i\nu_{i,7}$	$-Y_{fa}N_{bac}$		ρ_7
8	Uptake of Valerate				-1		$(1-Y_{C4})0.54$	$(1-Y_{C4})0.31$	$(1-Y_{C4})0.15$		$-\sum_{i=1-9,11-24} C_i\nu_{i,8}$	$-Y_{C4}N_{bac}$		ρ_8
9	Uptake of Butyrate					-1		$(1-Y_{C4}0.8$	$(1-Y_{C4}0.2$		$-\sum_{i=1-9,11-24} C_i\nu_{i,9}$	$-Y_{C4}N_{bac}$		ρ_9
10	Uptake of Propionate						-1	$(1-Y_{pro}0.57$	$(1-Y_{pro}0.43$		$-\sum_{i=1-9,11-24} C_i\nu_{i,10}$	$-Y_{pro}N_{bac}$		ρ_{10}
11	Uptake of Acetate							-1		$(1-Y_{ac})$	$-\sum_{i=1-9,11-24} C_i\nu_{i,11}$	$-Y_{ac}N_{bac}$		ρ_{11}
12	Uptake of Hydrogen								-1	$(1-Y_{ac})$	$-\sum_{i=1-9,11-24} C_i\nu_{i,12}$	$-Y_{H2}N_{bac}$		ρ_{12}
13	Decay of X_{su}										$-\sum_{i=1-9,11-24} C_i\nu_{i,13}$	$N_{bac}-N_{X_c}$		ρ_{13}
14	Decay of X_{aa}										$-\sum_{i=1-9,11-24} C_i\nu_{i,14}$	$N_{bac}-N_{X_c}$		ρ_{14}
15	Decay of X_{fa}										$-\sum_{i=1-9,11-24} C_i\nu_{i,15}$	$N_{bac}-N_{X_c}$		ρ_{15}
16	Decay of X_{C4}										$-\sum_{i=1-9,11-24} C_i\nu_{i,16}$	$N_{bac}-N_{X_c}$		ρ_{16}
17	Decay of X_{pro}										$-\sum_{i=1-9,11-24} C_i\nu_{i,17}$	$N_{bac}-N_{X_c}$		ρ_{17}
18	Decay of X_{ac}										$-\sum_{i=1-9,11-24} C_i\nu_{i,18}$	$N_{bac}-N_{X_c}$		ρ_{18}
19	Decay of X_{H2}										$-\sum_{i=1-9,11-24} C_i\nu_{i,19}$	$N_{bac}-N_{X_c}$		ρ_{19}
		Monosaccarides $kgCODm^{-3}$	Amino Acids $kgCODm^{-3}$	LCFA $kgCODm^{-3}$	Total Valerate $kgCODm^{-3}$	Total Butyrate $kgCODm^{-3}$	Total Propionate $kgCODm^{-3}$	Total Acetate $kgCODm^{-3}$	Liquid Hydrogen $kgCODm^{-3}$	Liquid Methane $kgCODm^{-3}$	Inorganic Carbon $kmolCm^{-3}$	Inorganic Nitrogen $kmolNm^{-3}$	Soluble Inerts $kmolPm^{-3}$	

Table B.6: Petersen Matrix, part *a*.

	Components →	13	14	15	16	17	18	19	20	21	22	23	24	Rate ρ_j
	Process ↓	X_C	X_{ch}	X_{pr}	X_{li}	X_{su}	X_{aa}	X_{fa}	X_{c4}	X_{pro}	X_{ac}	X_{H2}	X_I	$[\frac{kgCODm^{-3}d^{-1}}{kmolm^{-3}d^{-1}}]$
1	Disintegration	1	$f_{ch,xc}$	$f_{pr,xc}$	$f_{li,xc}$								f_{x_i,x_c}	ρ_1
2	Hydrolysis of Carbohydrates		-1											ρ_2
3	Hydrolysis of Proteins			-1										ρ_3
4	Hydrolysis of Lipids				-1									ρ_4
5	Uptake of Sugars					Y_{su}								ρ_5
6	Uptake of AminoAcids						Y_{aa}							ρ_6
7	Uptake of LCFA							Y_{fa}						ρ_7
8	Uptake of Valerate								Y_{C4}					ρ_8
9	Uptake of Butyrate								Y_{C4}					ρ_9
10	Uptake of Propionate									Y_{pro}				ρ_{10}
11	Uptake of Acetate										Y_{ac}			ρ_{11}
12	Uptake of Hydrogen											Y_{H2}		ρ_{12}
13	Decay of X_{su}	0.25				-1							0.75	ρ_{13}
14	Decay of X_{aa}	0.25					-1						0.75	ρ_{14}
15	Decay of X_{fa}	0.25						-1					0.75	ρ_{15}
16	Decay of X_{C4}	0.20											0.80	ρ_{16}
17	Decay of X_{pro}	0.20								-1			0.80	ρ_{17}
18	Decay of X_{ac}	0.20									-1		0.80	ρ_{18}
19	Decay of X_{H2}	0.25										-1	0.75	ρ_{19}
		Monosaccarides $kgCODm^{-3}$	Amino Acids $kgCODm^{-3}$	LCFA $kgCODm^{-3}$	Total Valerate $kgCODm^{-3}$	Total Butyrate $kgCODm^{-3}$	Total Propionate $kgCODm^{-3}$	Total Acetate $kgCODm^{-3}$	Liquid Hydrogen $kgCODm^{-3}$	Liquid Methane $kgCODm^{-3}$	Inorganic Carbon $kmolCm^{-3}$	Inorganic Nitrogen $kmolNm^{-3}$	Soluble Inerts $kmolPm^{-3}$	

Table B.7: Petersen Matrix, part *b*.

References

[1] J. A. Eastman, J. F. Ferguson, Solubilization of particulate organic carbon during the acid phase of anaerobic digestion, Journal (Water Pollution Control Federation) 53 (3) (1981) 352 – 366.

[2] M. Mattei, L. Frunzo, B. D'acunto, Y. Pechaud, F. Pirozzi, G. Esposito, Continuum and discrete approach in modeling biofilm development and structure: a review, Journal of mathematical biology 76 (4) (2018) 945–1003.

[3] D. Batstone, J. Keller, I. Angelidaki, S. Kalyuzhnyi, S. Pavlostathis, A. Rozzi, W. Sanders, H. Siegrist, V. Vavilin, IWA Publishing, London, 2002.

[4] V. Fedorovich, P. Lens, S. Kalyuzhnyi, Extension of anaerobic digestion model no. 1 with processes of sulfate reduction, Applied biochemistry and biotechnology 109 (2003) 33–45. doi:10.1385/ABAB:109:1-3:33.

[5] B. D.J., K. J., Industrial applications of the iwa anaerobic digestion model no. 1 (adm1)., Water Sci Technol. 47 (12) (2003) 199–206.

[6] F. Blumensaat, J. Keller, Modelling of two-stage anaerobic digestion using the iwa anaerobic digestion model no. 1 (adm1), Water Research 39 (1) (2005) 171 – 183. doi:https://doi.org/10.1016/j.watres.2004.07.024.
URL http://www.sciencedirect.com/science/article/pii/S0043135404004403

[7] L. Frunzo, F. G. Fermoso, V. Luongo, M. Mattei, G. Esposito, Adm1-based mechanistic model for the role of trace elements in

anaerobic digestion processes, Journal of Environmental Management-doi:10.1016/j.jenvman.2018.11.058.

[8] B. Chandra Maharaj, M. Rosaria Mattei, L. Frunzo, E. van Hullebusch, G. Esposito, Adm1 based mathematical model of trace element complexation in anaerobic digestion processes, Bioresource Technology 276. doi:10.1016/j.biortech.2018.12.064.

[9] B. C. Maharaj, M. R. Mattei, L. Frunzo, E. D. van Hullebusch, G. Esposito, Adm1 based mathematical model of trace element precipitation/dissolution in anaerobic digestion processes, Bioresource technology 267 (2018) 666–676.

[10] G. Esposito, L. Frunzo, A. Panico, F. Pirozzi, Modelling the effect of the olr and ofmsw particle size on the performances of an anaerobic co-digestion reactor, Process Biochemistry 46 (2011) 557 – 565. doi:10.1016/j.procbio.2010.10.010.

[11] A. Saltelli, K. Aleksankina, W. Becker, P. Fennell, F. Ferretti, N. Holst, S. Li, Q. Wu, Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices, Environmental Modelling & Software 114 (2019) 29 – 39. doi:https://doi.org/10.1016/j.envsoft.2019.01.012.
URL http://www.sciencedirect.com/science/article/pii/S1364815218302822

[12] H.-S. Jeong, C.-W. Suh, J.-L. Lim, S.-H. Lee, H.-S. Shin, Analysis and application of adm1 for anaerobic methane production, Bioprocess and Biosystems Engineering 27 (2) (2005) 81–89. doi:10.1007/s00449-004-

711 0370-4.
712 URL https://doi.org/10.1007/s00449-004-0370-4

713 [13] T. S. Souza, A. Carvajal, A. Donoso-Bravo, M. Peña, F. Fdz-
714 Polanco, Adm1 calibration using bmp tests for modeling the effect
715 of autohydrolysis pretreatment on the performance of continu-
716 ous sludge digesters, Water Research 47 (9) (2013) 3244 – 3254.
717 doi:https://doi.org/10.1016/j.watres.2013.03.041.
718 URL http://www.sciencedirect.com/science/article/pii/S0043135413002650

719 [14] M.-Y. Lee, C.-W. Suh, Y.-T. Ahn, H.-S. Shin, Variation of adm1
720 by using temperature-phased anaerobic digestion (tpad) op-
721 eration, Bioresource Technology 100 (11) (2009) 2816 – 2822.
722 doi:https://doi.org/10.1016/j.biortech.2008.12.025.
723 URL http://www.sciencedirect.com/science/article/pii/S0960852408010821

724 [15] E. L. Barrera, H. Spanjers, K. Solon, Y. Amerlinck, I. Nopens, J. Dewulf,
725 Modeling the anaerobic digestion of cane-molasses vinasse: Extension
726 of the anaerobic digestion model no. 1 (adm1) with sulfate reduction
727 for a very high strength and sulfate rich wastewater, Water Research
728 71 (2015) 42 – 54. doi:https://doi.org/10.1016/j.watres.2014.12.026.
729 URL http://www.sciencedirect.com/science/article/pii/S0043135414008586

730 [16] D. Dochain, P. Vanrolleghem, Dynamical modelling & estimation
731 in wastewater treatment processes, Water Intelligence Online 4.
732 doi:10.2166/9781780403045.

733 [17] X. Chen, Z. Chen, X. Wang, C. Huo, Z. Hu, B. Xiao, M. Hu, Ap-

plication of adm1 for modeling of biogas production from anaerobic digestion of hydrilla verticillata, Bioresource Technology 211 (2016) 101 – 107. doi:https://doi.org/10.1016/j.biortech.2016.03.002.
URL http://www.sciencedirect.com/science/article/pii/S0960852416302875

[18] L. A. Morales, . A. D. RodrÃguez, H. E. Rojas, Assessment of the Input Substrate Characteristics Included in the Anaerobic Digestion Model No. 1 (ADM1), IngenierÃa 22 (2017) 269 – 282.

[19] L. Benedetti, D. Batstone, B. De Baets, I. Nopens, P. Vanrolleghem, Global sensitivity analysis of biochemical, design and operational parameters of the benchmark simulation model no. 2, Proceedings of the 4th International Congress on Environmental Modelling and Software - Barcelona, Catalonia, Spain.
URL https://scholarsarchive.byu.edu/iemssconference/2008/all/133

[20] A. Saltelli, S. Tarantola, F. Campolongo, Sensitivity analysis as an ingredient of modeling, Statistical Science 15 (4) (2000) 377–395.
URL http://www.jstor.org/stable/2676831

[21] G. Esposito, L. Frunzo, A. Panico, G. d'Antonio, Mathematical modelling of disintegration-limited co-digestion of ofmsw and sewage sludge, Water science and technology : a journal of the International Association on Water Pollution Research 58 (2008) 1513–9. doi:10.2166/wst.2008.509.

[22] G. Esposito, L. Frunzo, A. Panico, F. Pirozzi, Model calibration and validation for ofmsw and sewage sludge co-digestion re-

actors, Waste management (New York, N.Y.) 31 (2011) 2527–35. doi:10.1016/j.wasman.2011.07.024.

[23] G. Esposito, L. Frunzo, A. Giordano, F. Liotta, A. Panico, F. Pirozzi, Anaerobic co-digestion of organic wastes, Reviews in Environmental Science and Bio/Technology 11 (4) (2012) 325–341.

[24] M. D. Morris, Factorial sampling plans for preliminary computational experiments, Technometrics 33 (2) (1991) 161–174. arXiv:https://www.tandfonline.com/doi/pdf/10.1080/00401706.1991.10484804, doi:10.1080/00401706.1991.10484804.
URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1991.10484804

[25] G. Damblin, M. Couplet, I. B., Numerical studies of space filling designs : optimization of latin hypercube samples and subprojection properties, Journal of Simulation.

[26] J. W. Eaton, D. Bateman, S. Hauberg, R. Wehbring, GNU Octave version 4.2.0 manual: a high-level interactive language for numerical computations (2016).
URL http://www.gnu.org/software/octave/doc/interpreter

[27] K. Radhakrishnan, Description and use of lsode : the livermore solver for ordinary differential equations, Tech. rep., National Aeronautics and Space Administration, Office of Management, Scientific and Technical Information Program (Springfield, Va.), National Technical Information Service (distributor), n94- 35259 NAS 1.61:1327 ID = 999753844002121

(1993).
URL https://search.library.wisc.edu/catalog/999753844002121

[28] F. Campolongo, A. Saltelli, Sensitivity analysis of an environmental model: an application of different analysis methods., Reliab. Eng. Syst. Saf. 57 (1) (1997) 49–69.

[29] A. Trucchia, M. Mattei, V. Luongo, L. Frunzo, M. Rochoux, Surrogate-based uncertainty and sensitivity analysis for bacterial invasion in multi-species biofilm modeling, Communications in Nonlinear Science and Numerical Simulation 73 (2019) 403–424.

[30] M. Baudin, A. Dutfoy, B. Iooss, A.-L. Popelin, OpenTURNS: An Industrial Software for Uncertainty Quantification in Simulation, Springer International Publishing, 2017, pp. 2001–2038. doi:10.1007/978-3-319-12385-1_64.

[31] G. Blatman, B. Sudret, Efficient computation of global sensitivity indices using sparse polynomial chaos expansions, Reliability Engineering & System Safety 95 (11) (2010) 1216–1229. doi:10.1016/j.ress.2010.06.015.

[32] A. Marrel, B. Iooss, B. Laurent, O. Roustant, Calculations of sobol indices for the gaussian process metamodel, Reliability Engineering & System Safety 94 (3) (2009) 742–751. doi:10.1016/j.ress.2008.07.008.

[33] I. Sobol, Sensitivity analysis for nonlinear mathematical models, Mathematical Modeling and Computational Experiment 1 (4) (1993) 407–414.

[34] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, S. Tarantola, Global Sensitivity Analysis. The Primer, John Wiley & Sons, Ltd, Chichester, UK, 2007. doi:10.1002/9780470725184.

[35] M. Baudin, K. Boumhaout, T. Delage, B. Iooss, J.-M. Martinez, Numerical stability of Sobol' indices estimation formula, in: 8th International Conference on Sensitivity Analysis of Model Output,, Réunion Island, 2016.

[36] G. Esposito, L. Frunzo, A. Panico, F. Pirozzi, Model calibration and validation for ofmsw and sewage sludge co-digestion reactors, Waste management 31 (12) (2011) 2527–2535.

[37] F. Pianosi, K. Beven, J. Freer, J. W. Hall, J. Rougier, D. B. Stephenson, T. Wagener, Sensitivity analysis of environmental models: A systematic review with practical workflow, Environmental Modelling & Software 79 (2016) 214 – 232. doi:https://doi.org/10.1016/j.envsoft.2016.02.008. URL http://www.sciencedirect.com/science/article/pii/S1364815216300287

[38] D. Xiu, G. Karniadakis, The wiener–askey polynomial chaos for stochastic differential equations, SIAM Journal on Scientific Computing 24 (2) (2002) 619–644. doi:10.1137/S1064827501387826.

[39] G. Blatman, B. Sudret, Adaptative sparse polynomial chaos expansion based on least angle regression, Journal of Computational Physics 230 (6) (2011) 2345–2367.

[40] G. Blatman, Adaptative sparse polynomial chaos expansions for un-

certainty propagation and sensitivity analysis, Ph.D. thesis, Université Blaise Pascal, Clermont-Ferrand (2009).

[41] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, The Annals of Statistics 32 (2) (2004) 407–499. doi:10.1214/009053604000000067.

[42] M. Berveiller, B. Sudret, M. Lemaire, Stochastic finite element: a non intrusive approach by regression, European Journal of Computational Mechanics 15 (2006) 81–92. doi:10.3166/remn.15.81-92.

[43] C. Rasmussen, C. Williams, Gaussian processes for machine learning, MIT Press, 2006.

[44] S. G. Nash, A survey of truncated-newton methods, Journal of Computational and Applied Mathematics 124 (1) (2000) 45 – 59, numerical Analysis 2000. Vol. IV: Optimization and Nonlinear Equations. doi:https://doi.org/10.1016/S0377-0427(00)00426-X.
URL http://www.sciencedirect.com/science/article/pii/S037704270000426X