Computers and Chemical Engineering 122 (2019) 275–292

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

# A framework for model reliability and estimability analysis of crystallization processes with multi-impurity multi-dimensional population balance models

D. Fysikopoulos[a], B. Benyahia[a,*], A. Borsos[a], Z.K. Nagy[a,b], C.D. Rielly[a]

[a] EPSRC Centre in Continuous Manufacturing and Crystallization at the Department of Chemical Engineering, Loughborough University, Loughborough, Leicestershire LE11 3TU, UK
[b] School of Chemical Engineering, Purdue University, West Lafayette, IN 47907, USA

ARTICLE INFO

*Article history:*
Received 29 November 2017
Revised 17 July 2018
Accepted 3 September 2018
Available online 8 September 2018

*Keywords:*
Parameter estimability
Variance-based sensitivity analysis
Sequential Orthogonalization
Multi-dimensional population balance modelling
Parameter estimation
Crystal growth modifiers

ABSTRACT

The development of reliable mathematical models for crystallization processes may be very challenging due the complexity of the underlying phenomena, the inherent Population Balance Models (PBMs) and the large number of parameters that need to be identified from experimental data. Due to the poor information content of the experiments, the structure of the model itself and correlation between model parameters, the mathematical model may contain more parameters than can be accurately and reliably identified from the available experimental data. A novel framework for parameter estimability for guaranteed optimal model reliability is proposed then validated by a complex crystallization process. The latter is described by a differential algebraic system which involves a multi-dimensional population balance model that accounts for the combined effects of different crystal growth modifiers/impurities on the crystal size and shape distribution of needle-like crystals. Two estimability methods were combined: the first is based on a sequential orthogonalization of the local sensitivity matrix and the second is Sobol, a variance-based global sensitivities technic. The framework provides a systematic way to assess the quality of two nominal sets of parameters: one obtained from prior knowledge and the second obtained by simultaneous identification using global optimization. A cut-off value was identified from an incremental least square optimization procedure for both estimability methods, providing the required optimal subset of model parameters. The implemented methodology showed that, although noisy aspect ratio data were used, the 8 most influential and least correlated parameters could be reliably identified out of twenty-three, leading to a crystallization model with enhanced prediction capability.



## 1. Introduction

Crystallization is an important separation process, extensively used in most chemical industries, either as a method of production or as a method of purification or recovery of solids. Many substances of scientific, technological, and commercial importance are crystalline, ranging from large-tonnage commodity materials to high-value specialty chemicals, such as active pharmaceutical ingredients (APIs). The pharmaceutical industry relies heavily on crystallization as 70% of all the pharmaceuticals formulation and 90% of APIs involve at least one crystallization step during the manufacturing process (Pena et al., 2015; Alvarez and Myerson, 2010). Besides, crystallization is one of the key steps in the production of pharmaceutical tablets which are the most popular dosage form. Hence, the crystallization step has a considerable impact on tuning the critical quality attributes (CQA), such as crystal size and shape distribution (CSSD), purity and polymorphic form, which impact the final product quality performance indicators and inherent end-use properties (e.g. bioavailability, tablet stability, dissolution, dosage form etc.), along with the downstream processability (e.g. filtration, drying etc.). As such, an effective control and design of the crystallization processes can lead to more robust and efficient manufacturing processes and consequently to higher product quality (Nagy et al., 2013; Rawlings et al., 1993).

With the recent advances in online process analytical technology (PAT) tools, more reliable and real-time data can be made available for process understanding and manipulation (Nagy et al., 2013; Yu, 2004). Therefore, high fidelity models and model-based approaches received considerable attention in many different areas,

* Corresponding author.
*E-mail address:* B.Benyahia@lboro.ac.uk (B. Benyahia).

https://doi.org/10.1016/j.compchemeng.2018.09.007

**Nomenclature**

| | |
|---|---|
| A | Inverse covariance matrix |
| $a_{i,\,i}$ | Area of the crystal per unit, [$m^2$] |
| $B_p$ | Primary nucleation, [#/s] |
| $C$ | Concentration of KDP crystals in the solution, [g/g solvent] |
| $C_{CGM,\,j}$ | Concentration of the $j$th crystal growth modifier, [g/g solvent] |
| $C_{sat}$ | Saturation concentration of KDP crystals in solution, [g/g solvent] |
| $d(N-p,\ {}^{a_d}/_2)$ | t-distribution, [–] |
| $E_p$ | Kinetic energy of primary nucleation, [kJ/mol] |
| $g$ | Aggregate vector of the variables |
| $\Delta G_{ads,\,i,\,j,\,k}$ | Adsorption energy, [kJ/mol] |
| $\Delta G_{des,\,i,\,j,\,k}$ | Desorption energy, [kJ/mol] |
| $g_i$ | Exponent of growth kinetic equitation of the $i$th characteristic facet, [–] |
| $G_i$ | Crystal growth rate of the $i$th characteristic facet, [m/s] |
| $G_{min,\,i}$ | Specific growth rate when distribution does not occur, [m/s] |
| $J(p)$ | Minimum sum of squared errors, [–] |
| $k_B$ | Boltzmann factor, [$m^2$ kg $s^{-2}$ $K^{-1}$] |
| $K_{d,\,i,\,j}$ | Distribution coefficient, [–] |
| $k_{ads,\,0,\,i,\,j,\,k}$ | - Adsorption rate constant of the $j$th crystal growth modifier on the $k^{th}$ type site at $i^{th}$ characteristic crystal facet, [–] |
| $k_{des,\,0,\,i,\,j,\,k}$ | - Desorption rate constant of the $j$th crystal growth modifier on the $k^{th}$type site at $i^{th}$ characteristic crystal facet, [–] |
| $k_e$ | Kinetic constant of Primary nucleation, [–] |
| $K_{e,\,j}$ | Thermodynamic distribution coefficient, [–] |
| $k_{gi}$ | Growth kinetic constant, [m/s] |
| $K_{i,\,j,\,k}$ | Langmuir constant of $j$th CGM on the $k$th active site on $i$th characteristic facet |
| $k_{m,\,i,\,j}$ | Mass transfer coefficient with crystal growth [m/s] |
| $k_{m0}$ | Mass transfer coefficient without crystal growth [$m/s$] |
| $k_{p,\,0}$ | Coefficient of primary nucleation [$m^{-3}\ s^{-1}$] |
| $L_{i,\,k}$ | Average distance between $k^{th}$type of sites [$m$] |
| $M_{CGM,\,j}$ | Molecular weight of CGM [$g/mol$] |
| $M_c$ | Molecular weight of KDP [g] |
| $n$ | Size and shape distribution [$\#/m^2$] |
| $P_{imp,\,i}$ | Impurity factor of the growth rate of $i$th characteristic facet |
| $N_p$ | Number of the model parameters [#] |
| $Ny$ | Number of measured outputs [#] |
| $Nm$ | Number of measurements of sampling times [#] |
| $Ne$ | Number of measurements [#] |
| $p$ | Vector of the input parameters (estimated parameters) |
| $p_{1-ad}$ | Vector of confidence domain boundaries |
| $R$ | Ideal gas constant [$Pa\ m^3\ mol^{-1}\ K^{-1}$] |
| $r_i$ | Orthogonal projection of $Z$, [–] |
| $s_i$ | First – order sensitivity index, [–] |
| $s_t$ | Total – order sensitivity index, [–] |
| $s_{ij}$ | Second – order sensitivity index, [–] |
| $S_{ij}$ | Sensitivity coefficients, [–] |
| $T$ | Temperature, [$K$] |
| $t$ | Time, [$s$] |
| $t_{ij}$ | $j^{th}$ sampling time of the $i$th output, [$s$] |
| $V_i$ | Variance, [–] |
| $x$ | Vector of the differential state variables |
| $\hat{y}_{ij}$ | Vector of numerically calculated aspect ratio at $k$th point in time, [–] |
| $y_{ij}$ | Vector of measured aspect ratio at $k$th point in time, [–] |
| $z$ | Vector of the algebraic state variables |
| $Z$ | Sensitivity Matrix, [–] |

*Greek letters*

| | |
|---|---|
| $\alpha_{i,\,k}$ | Effectiveness factor of the adsorption on the $k$th site on $i$th characteristic facet, [–] |
| $\beta_{i,\,k}$ | Constant of the effectiveness factor, [$m/K$] |
| $\gamma_i$ | Edge free energy on the $i$th crystal face per unit length, [$J/m$] |
| $\varepsilon_{ij}$ | Stochastic measurement error, [–] |
| $\eta_{ij}$ | Time spent by a particle in the presence of impurities, [$s$] |
| $\theta$ | Angle between {101} and {100} surfaces, [$rad$] |
| $\lambda$ | Cut-off value, [–] |
| $\mu_{m,\,r}$ | $m$, $r$ order joint moment |
| $\sigma$ | Relative supersaturation [–] |
| ${\sigma_{ij}}^2$ | Variance [–] |
| $\rho_c$ | Density of the KDP crystals, 2.338 [$kg/m^3$] |
| $\tau_{i,\,j,\,k}$ | Adsorption time constant [$s$] |
| $\chi_{c,\,j}$ | Mole fraction of the CGM in the crystal phase |
| $\Omega^{\kappa}$ | Sample Space |
| $\Omega_{sz}$ | Size Space |

such as process design, control, real time optimization and Quality-by-Design (Su et al., 2015; Nagy, 2009; Mascia et al., 2013; Lakerveld et al., 2013; Aamir et al., 2009; Benyahia et al., 2012; Ramin et al., 2018; Benyahia 2018). A prerequisite to apply model-based control strategies is the availability of a predictable mathematical model. The most fundamental approach for modelling particulate processes, such as crystallization, is the population balance model (PBM) framework coupled with kinetic expressions, mass and energy balances, which yields a set of nonlinear integro-partial differential equations. The set provides a rigorous approach to model the dynamic evolution of the dispersed phase system's properties, such as CSSD (Majumder et al., 2012; Sato et al., 2008; Borsos and Lakatos, 2014; Kumar et al., 2008). Although the PBM framework is based on first principles, a general theoretical mathematical expression for the determination of the crystallization kinetics doesn't exist and hence, empirical or semi-empirical expressions (e.g. power law etc.) are used, that in most of the cases account the supersaturation as the key variable (Rawlings et al., 1993; Cao et al., 2012).

Although the benefits of the mathematical models are widely accepted, setting a unified rigorous framework for building reliable and predictable models is still an open subject, particularly for pharmaceutical processes. In order to obtain accurate model predictions, identification of the unknown model parameters is often required. However, in many cases, first-principles models comprise a large number of parameters which often cannot be estimated reliably from the available experimental data. In addition, the quality and the information content of the available experimental data can be affected by many factors such as noisy measurements, limited number of data points, poor design of experiments (DoE) and limited range of operating conditions (Benyahia et al., 2013; Chu and Hahn, 2011). Furthermore, strong influence of a parameter on one

or more of the measured responses, high correlation between the parameters' effects and/or the effects of a parameter on model predictions can also lead to unreliable and inaccurate identification of the unknown parameter values, which in turn degrades the prediction capability of the mathematical model (Kravaris et al., 2013 Benyahia et al., 2013; Eghtesadi and McAuley, 2014). Of course, mismatch could also arise from the model structure itself, since several assumptions are commonly made in order to simplify the numerical representations of the system and reduce its complexity with the risk of neglecting some of the key underlying phenomena and consequently reducing the prediction capabilities of the model.

Several approaches have been developed to cope with some of these problems, such as modifying the model structure, incorporating additional measured outputs (e.g. using different PAT) and improving the information content of the experimental data by utilizing DoE approaches. However, before deciding whether the mathematical equations should be modified, or supplementary experiments should be designed and performed, one key step is to investigate whether the available experimental data contain enough information to identify uniquely and reliably the overall model parameters, or alternatively, the subset of the model parameters that could be identified reliably and lead to the most predictable mathematical model. This could be achieved by evaluating the structural identifiability and estimability (i.e. practical identifiability) of the model parameters (McLean and McAuley, 2011; Sin et al., 2010). The structural identifiability approach evaluates whether the parameters are locally or globally identifiable based utterly on the model structure, while the estimability appraises whether the parameters can be identified uniquely by using the available experimental data or data from a proposed set of experiments (McLean and McAuley, 2011; Walter and Pronzato, 1997). The estimability or practical identifiability methodology depends on the domain of variability of model parameters and experimental conditions whereas the structural identifiability is totally independent from both. The objective of the estimability analysis is to identify how many of the model parameters can be estimated accurately from the available data, while the ones with low estimability potential can be set to certain nominal values without degrading the prediction capability of the model (Benyahia et al., 2013; Chu and Hahn, 2011). Consequently, the estimability potential can be defined as a measure of the effects of parameters on the experimental outputs and/or correlation among the model parameters. In this work, only the estimability of the model parameters is evaluated.

Different approaches have been developed and proposed to help identify the most appropriate subset of parameters for estimation based on the estimability approach. Degenring et al., (2004) proposed a method for parameter selection based on principal component analysis (PCA), which is a statistical procedure that converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables. A parameter selection was obtained by using different PCA methods (Jolliffe, 1972), which provided different parameter ranking outcomes. The PCA-based approach was applied in more recent investigations (Schittkowski, 2007; Quaiser and Mönnigmann, 2009) and was proven to be less robust compared to the orthogonalization and the eigenvalue method discussed below. The eigenvalue method, introduced by Vajda et al., (1989) and was improved independently by other researchers (Schittkowski, 2007; Quaiser and Mönnigmann, 2009), determines the most estimable subset of parameters based on the eigenvector and eigenvalues of the fisher information matrix (FIM). Although, the method shown better accuracy compared to other methods, it becomes challenging sometimes to match eigenvalues with specific parameters (McLean and McAuley, 2011). The singular value decomposition (Velez-Reyes and Verghese, 1995) and the correlation and collinearity methods (Brun et al., 2002; Sin et al., 2010) were also proposed for the estimability analysis. The main drawback of the correlation and collinearity techniques is the fact that only the directions of the sensitivity vectors are considered without taking into account the magnitude of the sensitivities (McLean and McAuley, 2011). This limitation led Brun and coauthors (2002) to propose a robust approach that combined a method based on a scalar measure of the FIM and the collinearity method. A more robust approach for performing the estimability analysis is based on the orthogonalization of the sensitivity matrix which was initially introduced by Yao et al., (2003) then improved by Lund and Foss (2008) and Thomson et al. (2009). The method ranks the parameters according to both their individual effect on the measured responses and the correlation between the parameters. Due to the efficiency of this forward-selection method, it has been employed widely in complex chemical and biochemical systems (Benyahia et al., 2013; Surisetty et al., 2010; Thomson et al., 2009; Jayasankar et al., 2009; Onyemelukwe et al., 2018).

Despite the popularity of the estimability analysis in numerous scientific areas, such as polymer science, environmental engineering and biology, this class of methods is still novel in the area of crystallization and its inherent benefits are not well understood, as only very limited number of studies have been reported in the literature. Chen et al. (2004) presented a model-discrimination for model-based design by using the D-optimal criterion for the parameter set selection. However, only four parameters were considered making the benefits of the method unclear. Some of the benefits of the parameters selection methods were discussed by Czapla et al. (2009) who used an approach proposed by Brun et al. (2002) to select the most sensitive model parameters of a preferential batch crystallization of enantiomers. However, both studies utilized an arbitrary cut-off value for the parameter selection. A more comprehensive study was presented by Samad et al. (2013a,b) where two global sensitivity analysis techniques, Morris screening and the standardized coefficients, were utilized to identify the most significant parameters. Although, these techniques may be useful for the classification of the parameters in terms of sensitivity, the correlation of the parameters was not considered during the ranking procedure but estimated afterwards.

Considering all the challenges inherent to parameter selection and identification discussed above and with the scope of improving the current methodology for parameter identification for crystallization processes, a new framework (Fig. 1– see next section) is proposed for a systematic and optimal selection of the parameter subset with the highest estimability potential for guaranteed model reliability. As a case study, a batch cooling crystallization process is considered under the presence of multiple impurities, more specifically crystal growth modifiers (CGM), which can affect, besides product purity, the growth and potentially the nucleation kinetics and hence the size and shape distribution of the final crystals. A novel morphological multi-dimensional population balance model that incorporates mechanisms for multisite competitive adsorption of the impurities on the crystal faces, coupled with mass balance equations is used (Borsos et al., 2016).

To the best of our knowledge, it is the first time that the modified Gram Schmidt orthogonalization algorithm and Sobol analysis are combined and applied in the area of crystallization and equally the first time that the estimability analysis in general is being applied to assess the model reliability of a PBM that takes into consideration the presence of impurities. The complexity of the case study provides an opportunity to show the capabilities of the methodology with the scope of building more reliable and high-fidelity models for the pharmaceutical industry for process design, optimization and advanced control that would enhance the implementation of model-based Quality-by-Design (QbD).

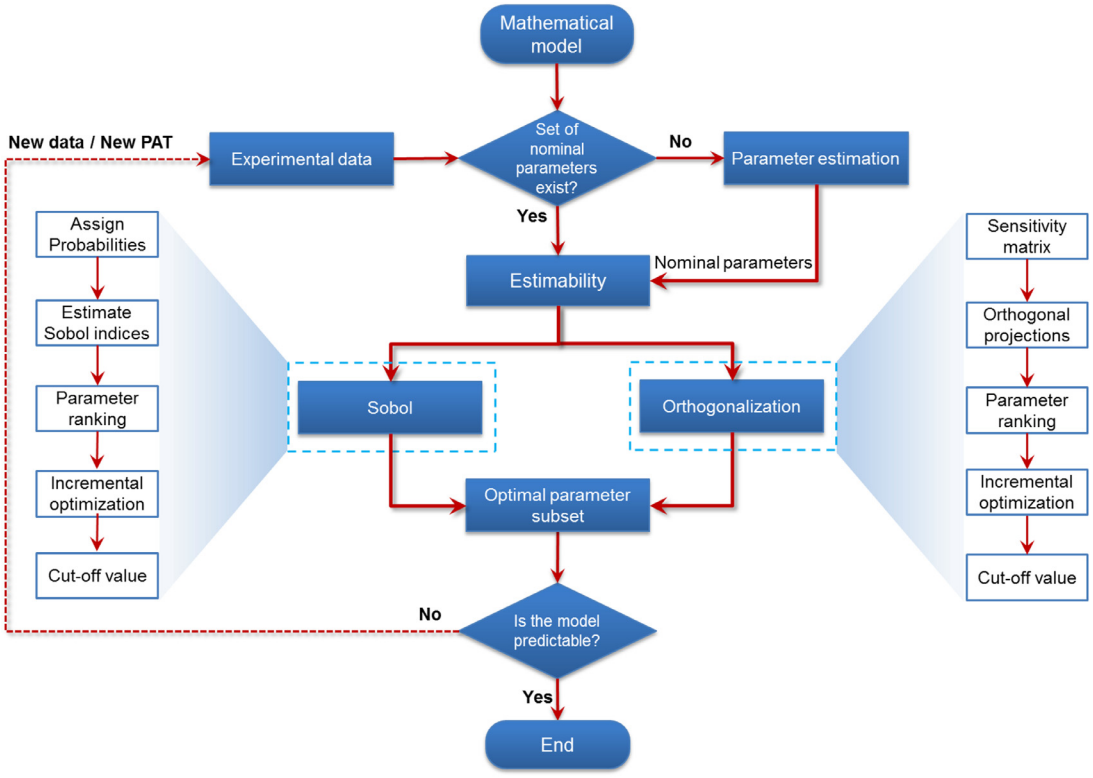


**Fig. 1.** Schematic of the parameter identification and estimability analysis framework.

## 2. Method

The proposed methodology (Fig. 1) combines a sequential orthogonalization method, which takes into account the overall magnitude of the local sensitivities and the correlation between the parameters, with a variance-based global sensitivity ranking method (Sobol). In both cases, a rigorous approach is used to identify the cut-off values based on the minimization of the maximum likelihood criterion. To assess the consistency and quality of the methodology, the correlation coefficients are calculated and the parameter estimates are assessed against their confidence domains. The proposed methodology enables a more robust classification of the model parameters based on the estimability potential, and provides a key tool to analyze the information content of the experimental data and consequently the quality of the measurements and the employed sensors (PAT). As such, the method helps identify the parameters that could be estimated accurately from the available data and informs whether additional data are required to identify a specific model parameter (e.g. new experiments or additional data from another sensor).

The estimability approach aims at identifying more reliably the model parameters from the existing data and requires an initial nominal vector of model parameters, commonly obtained from prior knowledge of the process. In the case of lack of prior knowledge or uncertain model parameters (extremely poor estimates or broad confidence intervals), the estimability framework described in the paper provides a methodology to help identify the set of the nominal parameters. Although a variety of different approaches has been applied to estimate model parameters of complex and highly nonlinear chemical processes, nonlinear optimization algorithms has been widely adopted due to their accuracy and efficiency (Rawlings et al., 1993). In this paper, a hybrid global optimization technique that combines a genetic algorithm and local deterministic method (sequential quadratic programming) was used to identify the unknown parameters.

To maximize the benefits of the methodology, the estimability approach was implemented in both cases: the case where the initial nominal vector of parameters exists form prior process knowledge and the case where all parameters of the nominal vector should be identified globally and simultaneously by minimizing the weighted least square error. Both estimability approaches, the sequential orthogonalization and Sobol (variance-based method), rank the model parameters by order of importance. The ultimate objective of the estimability approach is then to find the optimal subset of model parameters that guarantee maximum model reliability. As a consequence, an estimability threshold or cut-off value is required to identify the subset of parameters that should be subject to re-estimation, to maximize model accuracy, and the subset of parameters that should be kept at nominal values, without degrading the prediction capability of the model. An optimal subset of parameters can be obtained by running a sequential parameter estimation procedure by identifying the top $i$th parameters (where $i = 1, 2, \ldots$) each time and calculating the corresponding objective function value. The cut-off value can be obtained when the improvement in the objective function due to an additional parameter becomes insignificant. If the model prediction capability with the optimal parameter subset is unsatisfactory, the method suggest to run additional experiments, redesign the experiments (e.g. optimal experimental designs) or/and select additional or alternative PAT tools with the scope of increasing the information content of the data.

### 2.1. Process model

The Multi-Impurity Adsorption Model (MIAM) was developed by Borsos et al. (2016) as a novel mathematical model for crystallization processes considering multi-impurity adsorption mechanisms with the purpose of process design, optimization and control. The model was built to predict the dynamic evolution of size and shape distribution during crystallization under the presence of

**Table 1**
Complete set of differential-algebraic equations (DAEs) that represent the Multi-Impurity Adsorption Model (MIAM).

General form of the moment – based PBEs

$\frac{\partial \mu_{0,0}}{\partial t} = B_p$ $\quad \frac{\partial \mu_{m,r}}{\partial t} = m\ G_1\ \mu_{m-1,r} + r\ G_2\ \mu_{m,r-1},\ m, r = 0, 1, 2, \ldots$

Component Mass Balance – Solute Concentration Component Mass Balance – Impurities Concentration

$\frac{dC(t)}{dt} = -\rho_c \frac{d\mu_{1,2}}{dt}$ $\quad \frac{dC_{CGM,j}}{dt} = \frac{\chi_{c,j}}{1-\sum_j \chi_{c,j}} \frac{M_{CGM,j}}{M_c} \frac{dC}{dt}$

Primary nucleation rate Crystal growth kinetic rate

$B_p = k_{p,0} \exp(-\frac{E_p}{RT}) exp(-k_e\ ln^{-2}(\frac{C}{C_{sat}}))$ $\quad G_i = k_{g,i}\ (\frac{C-C_{sat}}{C_{sat}})^{g_i} \{1 - (a_{i,i}\ \frac{K_{i,CGMi,i}\ C_{i,CGMi,i}}{1+K_{i,CGMi,i}\ C_{i,CGMi,i}})\}$

Mole fraction of the $j^{th}$ CGM Thermodynamic distribution coefficient of the $j^{th}$ CGM

$\chi_{c,j} = \sum_i K_{d,i,j} \frac{C_{CGM,j}}{M_{CGM,j}} (\frac{C}{M_c} + \sum_j \frac{C_{CGM,j}}{M_{CGM,j}})^{-1}$ $\quad K_{d,i,j} = 1 - (1 - K_{e,j}) \sqrt{\frac{G_{min,i}\ k_{m,i,j}}{G_i\ k_{min,i,j}}}$

Langmuir constant of the $j^{th}$ CGM on the $k^{th}$ site on $i^{th}$ characteristic face

$K_{i,j,k} = \frac{k_{ads,i,j,k}}{k_{des,i,j,k}} = \frac{k_{ads,0,i,j,k}}{k_{des,0,i,j,k}} exp(\frac{\Delta G_{des,i,j,k} - \Delta G_{ads,i,j,k}}{RT})$

Absorption effectiveness factor of the k$^{th}$ site on the i$^{th}$ characteristic face

$a_{i,k} = \frac{\gamma_i}{k_B T (\frac{C-C_{sat}}{C_{sat}}) L_{i,k}} = \frac{\beta_i}{T (\frac{C-C_{sat}}{C_{sat}})}$

Mass transfer coefficient when impurity distribution does and does NOT occur, respectively

$k_{m,i,j} = G_i\ [1 - exp(-\frac{G_i}{k_{m0,j}})]^{-1}\ k_{min,i,j} = G_{min,i}\ [1 - exp(-\frac{G_{min,i}}{k_{m0,j}})]^{-1}$

Unknown Parameters for Primary Nulceation & Crystal Growth in each characteristic face

$\boldsymbol{p} = [k_{ads,0,CGM1}\ ,\ k_{des,0,CGM1}\ , \beta_1, G_{min,1}, k_{m,0,CGM1}\ ,\ K_{e,CGM1}, \Delta G_{des,1}, \Delta G_{ads,1} \cdots$
$k_{ads,0,CGM2}\ ,\ k_{des,0,CGM2}\ , \beta_2, G_{min,2}, k_{m,0,CGM2}\ ,\ K_{e,CGM2}, \Delta G_{des,2}, \Delta G_{ads,2} \cdots$
$g_1,\ k_{g1},\ g_2,\ k_{g2},\ k_{p,0}\ ,\ E_p,\ k_e];$

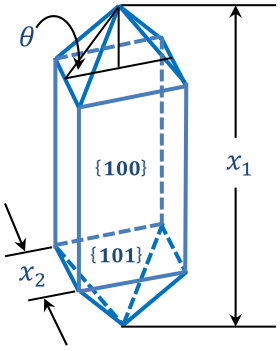


**Fig. 2.** Graphical representation of the morphology of the KDP crystal.

impurities. The effect of the crystal growth modifiers was monitored in real time by using an in-situ video imaging probe: Lasentec Particle Vision and Measurement V819 (PVM). Images were automatically obtained with a frequency equal to six images per second and analyzed by Lasentec's image and stat acquisition software, where blob analysis was utilized to monitor the aspect ratio. In more detail, the cooling crystallization of pure potassium dihydrogen phosphate (KDP) in deionized water was investigated under the presence of aluminum sulfate (Crystal Growth Modifier: CGM1) and sodium hexametaphosphate (CGM2) and aspect ratio (AR) measurements were obtained as experimental outputs. As it was presented by Borsos et al. (2016), divalent and trivalent metal ions preferably adsorb onto the {100} KDP crystal facet hindering the crystal growth in that facet, while anionic growth modifiers prefer to adsorb onto the {101} KDP crystal facet inhibiting the crystal growth of the corresponding length. Hence, CGM1 is likely to adsorb onto {100} facet leading to more needle-like shaped crystals, while CGM2 tends to adsorb onto the {101} facet causing an opposite effect by generating crystals with lower aspect ratio. Thus, the CGMs considered in this case have competing effects.

Multidimensional population balance equations (PBEs) with two characteristic lengths $\boldsymbol{x_l} = \{x_1, x_2\}$ were considered to model the evolution of the crystal shape distribution, (Fig. 2). To integrate /solve the PBM model, the concentrations of the solute and the impurities are also required, which are calculated by coupling the corresponding mass balances. The overall model is summarized in Table 1 and it consists of a set of differential-algebraic equations (DAEs) combined with algebraic equations that describe the kinetics and thermodynamics.

The mathematical model requires 23 parameters and can be represented, for notational expediency, by the following general form of differential-algebraic equations (DAEs):

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{z}(t), \boldsymbol{p}, t),\ \boldsymbol{x}(t=0) = \boldsymbol{x}_0,\ z(t) = \boldsymbol{g}(\boldsymbol{x}(t), \boldsymbol{z}(t), \boldsymbol{p}, t) \tag{1}$$

where $\boldsymbol{x}$ is the vector of the differential state variables, $\boldsymbol{z}$ is the vector of the algebraic state variables, and $\boldsymbol{p}$ is the vector of the parameters.

A multi-variate nonlinear dynamic regression model can be considered for the mathematical illustration of the interaction between the model prediction and measured output:

$$\boldsymbol{y_{ij}} = \hat{\boldsymbol{y}}_{\boldsymbol{ij}}(\boldsymbol{p}, t_{ij}) + \varepsilon_{ij} \tag{2}$$

where $\boldsymbol{y_{ij}}$ is the $j^{\text{th}}$ measurement of the $i^{\text{th}}$ experimental output, $\hat{\boldsymbol{y}}_{\boldsymbol{ij}}$ is the corresponding model prediction, $t_{ij}$ is the $j^{\text{th}}$ sampling time of the $i^{\text{th}}$ output and $\varepsilon_{ij}$ is the measurement error assumed to be uncorrelated, Gaussian distributed, with zero mean.

### 2.2. Estimability analysis

In the current study, the estimability analysis consists of three main steps. In the first step, the relative effect of each model parameter on the measured outputs is determined through local sensitivity analysis of the dynamic system. The Sensitivity analysis is a fundamental study that can determine how the variations of the outputs could be related to certain variations of the input variables. The second step is to apply the orthogonalization algorithm with the scope of ranking the parameters in descending order, in terms of impact on the outputs and minimum correlation between the parameters. Finally, a parameter estimation procedure is performed incrementally and sequentially in order to identify the threshold (cut-off value) on the objective function, which in turn helps select of the optimum most estimable subset. These steps are thoroughly described and discussed below.

#### 2.2.1. Ranking the model parameters - orthogonalization method

The development of an effective solution to the parameter selection problem requires the quantification of the influence of each parameter on the measured outputs. This approach indicates which parameters are the most important and most likely to affect the model predictions. The first step of the estimability analysis method is the evaluation of the sensitivity coefficients which can be calculated analytically or numerically. The numerical approach

consists in applying a perturbation to the nominal values of the parameters according to the backward finite differences method as follows

$$s_{ij} = \frac{\partial \hat{y}_i}{\partial p_j} \approx \frac{\hat{y}_i(t, p_j) - \hat{y}_i(t, p_j - \Delta p_j)}{\Delta p_j}, \quad j = 1, 2, \ldots, N_p \tag{3}$$

where $N_p$ is the number of the parameters.

It should be mentioned that the relative perturbation applied to the nominal values of the parameters was equal to −2% (i.e. $\Delta p_j/p_j$). As such the local sensitivity can be calculated for each sampling or measurement time. As the model parameters and outputs have different units and numerical values that could span several orders of magnitude, a normalization of the local sensitivities is often applied with respect to the parameters' nominal values and corresponding model output in order to make a more reliable comparison between the inherent effects of the parameters. The normalized sensitivity coefficients are given by the following equation:

$$s_{ij}\big|_{t=t_k} = \frac{\overline{p_j}}{\overline{y_i}\big|_{t=t_k}} \frac{\partial \hat{y}_i}{\partial p_j} \approx \frac{\overline{p_j}}{\overline{y_i}\big|_{t=t_k}} \frac{\hat{y}_i(t, p_j) - \hat{y}_i(t, p_j - \Delta p_j)}{\Delta p_j} \tag{4}$$

where $\overline{p_j}$ is the nominal value of the $j^{th}$ parameter, $\overline{y_i}|_{t=t_k}$ is the model prediction of the $i^{th}$ output, evaluated at a sampling time $t_k$ using the nominal vector of parameters $\overline{\boldsymbol{p_j}}$ and $j = 1, 2, \ldots, N_p$.

After the sensitivity coefficients have been calculated, a sensitivity matrix $\boldsymbol{Z}$ is constructed as follows:

$$\boldsymbol{Z} = \begin{bmatrix} s_{11}\big|_{t=t_1} & \cdots & s_{1N_p}\big|_{t=t_1} \\ \vdots & \ddots & \vdots \\ s_{N_y1}\big|_{t=t_1} & \cdots & s_{N_yN_p}\big|_{t=t_1} \\ s_{11}\big|_{t=t_2} & \cdots & s_{1N_p}\big|_{t=t_2} \\ \vdots & \ddots & \vdots \\ s_{N_y1}\big|_{t=t_{N_m}} & \cdots & s_{N_yN_p}\big|_{t=t_{N_m}} \end{bmatrix} \tag{5}$$

The sensitivity matrix has a dimension $N_y \times (N_p \times N_m)$, where $N_y$ is the number of the measured outputs, $N_m$ is the number of the measurements or sampling times, while $N_p$ is the number of model parameters. Hence, each column represents the sensitivity coefficients with respect to one particular parameter, while each row captures the sensitivities of a specific output to the whole set of parameters at a particular sampling time.

The orthogonalization method provides an efficient forward-selection method that has been applied extensively for parameter ranking and selection. The technique is relatively simple to implement and most importantly it ranks the parameters more reliably, as both the magnitude of the effect of each model parameter on the outputs and the correlations between the effects of different parameters are considered simultaneously. This is paramount since both phenomena are critical for the parameter selection and discrimination, and consequently, to the prediction capabilities of the mathematical model. If a perturbation of a model parameter has minor effect on the outputs, then the parameter cannot be identified accurately from the data. This can be mathematically determined by calculating the norm of the sensitivity vectors (the norm of the columns $Z_i$). Conversely, large magnitudes/norms indicate significant effects on the outputs. At the same time, if a disturbance of two or more model parameters have similar trends/effects on the outputs, and then the parameters are highly correlated. As a result, the impact of one parameter overlaps with the impact of the other, and hence these parameters cannot be reliably and uniquely identified from the data (Benyahia et al., 2013). It should be noted that the orthogonalization method selects sequentially the least correlated and most influential parameters. The correlation can also be evaluated using the FIM (e.g. linear dependency of the sensitivity vectors) as described later. In this work, a modified Gram-Schmidt orthogonalization algorithm (Yao et al., 2003) is used to help rank sequentially the model parameters according to the magnitude of the sensitivities and the least correlation effect. The sequential orthogonalization algorithm is presented in Table 2. The first parameter is selected then all vectors of the scaled sensitivity matrix are sequentially projected onto an orthogonal basis (the sensitivity vectors with the highest magnitude).

**Table 2**
Orthogonalization algorithm for the estimability analysis (Benyahia et al., 2013).

$\boldsymbol{Z_i}$ : sensitivity vector corresponding to the parameter; $p_i$ : $\lambda$ : cut – off value;
$\boldsymbol{r_i}$: orthogonal projection of $\boldsymbol{Z_i}$; $\boldsymbol{P_j}$: set of estimable parameters;
$\boldsymbol{X_j}$: the matrix of the selected parameters vectors at the $j^{th}$ stage;

**1. Select the parameter with the highest effect: find the index k such that:**

$$\boldsymbol{k = arg\,max_i (Z_i)^T Z_i},\ \boldsymbol{i \in I_0 = \{1, \ldots, n_p\}}$$
$$if\ (\boldsymbol{Z_k})^T \boldsymbol{Z_k} \geq \lambda\ set\ \boldsymbol{P_1 = \{p_k\}}\ and\ \boldsymbol{X_1 = Z_k}$$
otherwise stop

**2. Orthogonalization: Compute the orthogonal projection of the matrix Z:**

$$\boldsymbol{R^j = (I - X_j (X_j{}^T X_j)^{-1} X_j{}^T) Z}$$

**3. Select the next parameter with the highest effect:**

$$\boldsymbol{l = arg\,max_i (r_i^j)^T r_i},\ \boldsymbol{i \in I_j = (I_{j-1} - \{k, \ldots\})}$$
$$if\ (\boldsymbol{r_l^j})^T\, \boldsymbol{r_l^j} \geq \lambda\ set\ \boldsymbol{P_j = \{P_{j-1}, p_l\}}\ and\ \boldsymbol{X_{j+1} = \{X_j, Z_l\}}$$

**Return to step 2**
**Otherwise Stop**

The orthogonalization method makes it possible to rank the parameters according to their estimability potential. However, the development of a reliable and rebut methodology for the selection of the optimal subset of parameters remains an open subject in the literature, since arbitrary cut-off values are applied in most cases. In this work, an optimization based approach is utilized for the optimum parameter selection based on the maximum likelihood approach:

$$J(p) = \min_{\boldsymbol{p}} \left\{ \sum_{i=1}^{N_y} N_e \bullet ln\left( \sum_{j=1}^{N_e} \left[ \left( \boldsymbol{y_{ij}}(\boldsymbol{p}, t) - \boldsymbol{\hat{y}_{ij}}(\boldsymbol{p}, t) \right)^2 \right] \right) \right\} \tag{6}$$

$$\boldsymbol{s.t.}\ \boldsymbol{\dot{x}} = \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{z}(t), \boldsymbol{p}, t),\ \boldsymbol{x}(t=0) = \boldsymbol{x}_0,\ \boldsymbol{g}(\boldsymbol{x}(t), \boldsymbol{z}(t), \boldsymbol{p}, t) = \boldsymbol{\hat{y}_{ij}}(t)$$

where $y_{ij}(p, t)$ is the experimental measurement; $N_y$ is the number of outputs and $N_e$ is the number of the experiments.

The maximum likelihood criterion was also used in the parameter estimation problem to identify the initial set of model parameters (i.e. nominal set).

### 2.2.2. Global Sensitivity Analysis (GSA)

The sensitivity analysis has been extensively applied as a technique for model simplification, model calibration and process understanding through computer-aided design (Varma et al., 2005; Saltelli et al., 2004). The LSA are widely accepted by the research community due to the low computational cost. However, the LSA techniques can only determine the sensitivity of each input separately, without taking into account the overall contributions of the input variables to the output predictions. In GSA methods, a simultaneous perturbation of all parameters (inputs) is performed within specific bounds, as opposed to LSA techniques where the parameters inputs are varied once at a time. Hence, the GSA approaches are capable of measuring not only the relative impact of each input variable, but also the interactions between them. The variance-based global sensitivity techniques, which are used here, depend on the calculation of the following ratio:

$$\frac{Var_p\left[E\left(y_{ij}(\boldsymbol{p}, t)|\boldsymbol{p}\right)\right]}{Var\left(y_{ij}(\boldsymbol{p}, t)\right)} \tag{7}$$

where $E(y_{ij}(\boldsymbol{p}, t)|\boldsymbol{p})$ denotes the expectation of the output y on a fixed value, and the variance is calculated over all possible values

of the inputs. In our case the inputs are the vector of the unknown parameters $\boldsymbol{p}$.

Many GSA techniques have been developed with the most well established being the method of Sobol′ (2001). The Sobol method decomposes the output function $y(p_1, \ldots, p_k)$ into terms of increasing degrees of interactions between the model inputs as follows:

$$\hat{y}\left(p_1, \ldots, p_{N_p}\right) = \hat{y}_0 + \sum_{i=1}^{N_p} \hat{y}_i(p_i) + \sum_{i=1}^{N_p-1} \sum_{j=i+1}^{N_p} \hat{y}_{ij}\left(p_i, p_j\right) + \ldots + \hat{y}_{1,2,\ldots,N_p}\left(p_1, p_2, \ldots, p_{N_p}\right) \quad (8)$$

In general, there are infinite ways to decompose the function $\hat{y}(p_1, \ldots, p_{N_p})$. However, for independent factors, the decomposition based on orthogonal terms becomes unique (Sobol′, 2001) and the functions can be calculated through multidimensional integrals as follows:

$$f_0 = E(y) = \int_{\Omega^k} f(p)dp \quad (9)$$

$$f_i = E(y|p_i) - E(y) = -f_0 + \int_0^1 \ldots \int_0^1 f(p)dp_{\sim i} \quad (10)$$

$$f_{ij} = E\left(y|p_i, p_j\right) - E(y|p_i) - E(y) = \int_0^1 \ldots \int_0^1 f(p)dp_{\sim ij} - f_0 - f_i \quad (11)$$

where $dp_{\sim i}$ and $dp_{\sim ij}$ denote the integration over all variables except $p_i$ and $p_j$ respectively, $\Omega^k$ is the sampling space. Sobol's method employs Monte Carlo approximations to calculate the integrals described in Eqs. (9)–(11) (Sobol′, 2001; Saltelli et al., 2005).

Similarly, Eq. (8) can be re-written as a variance (Eq. (12)) and sensitivity (Eq. (13)) respectively:

$$V(y) = \sum_{i=1}^{k} V_i + \sum_{1 \le i \le j \le k} V_{ij} + \ldots + V_{1,2,\ldots,k} \quad (12)$$

$$\sum_{i=1}^{k} s_i + \sum_{1 \le i \le j \le k} s_{ij} + \ldots + s_{1,2,\ldots,k} = 1 \quad (13)$$

where $V_i$ is the contribution of the parameter $p_i$ to the total variance $V(y)$, while $V_{ij}$ is the contribution inherent to the interactions between two parameters $p_i$ and $p_j$.

Hence, these contributions (i.e. partial variances) can be used to calculate the first-order sensitivity index for the parameter $p_i$ which evaluates the main effects of $p_i$ on the output (i.e. partial variance of $p_i$ to the total variance):

$$s_i = \frac{V_i}{V(y)} \quad (14)$$

In a similar way, the second-order $s_{ij}$ and the total order sensitivity indices $s_{Ti}$ can be determined from:

$$s_{ij} = \frac{V_{ij}}{V(y)} \quad (15)$$

$$s_{Ti} = 1 - \frac{V_{\sim j}}{V(y)} \quad (16)$$

The total sensitivity index $s_{Ti}$ determines the total contribution of the parameter $p_i$ considering both direct and indirect effects. Hence, the difference between $s_{Ti}$ and $s_i$ indicates the degree of interaction. More information regarding the method of Sobol and other variance-based method techniques can be found in Saltelli et al. (2008).

## 3. Results and discussions

Here, a rigorous selection procedure of the optimal subset of parameters, based on the estimability approach, is developed and implemented to the MIAM. The method combines two estimability methods: the first associates local sensitivities to a sequential orthogonalization procedure and the second uses a variance-based global sensitivity selection. Only the optimal subset or parameters require identification (the rest of the parameters can be set to their nominal values) with a guaranteed minimum model mismatch (i.e. high prediction capability).

It should be emphasized that the parameters of the novel Multi-Impurity PBM model were previously identified by Borsos and co-authors (2016) using a sequential identification methodology and attempted to identify decoupled kinetic parameters while taking several parameters from literature. For instance, the parameters associated with the primary nucleation and the crystal growth of the two different facets were obtained from the literature, while the kinetic parameters inherent to the two crystal growth modifiers (i.e. impurities) were estimated from on-line image analysis data. However, the addition of additives/impurities might affect the kinetics of nucleation and growth (Epstein, 1982; Kubota, 2001), and hence the nominal parameter vector might not be reliable enough.

Commonly, the estimability approach requires a set of nominal parameter values which represent a reasonable initial guess, usually obtained from literature or prior process knowledge. To guarantee a generic robust framework for parameter estimation and extend the discussions above, the estimability approach is developed for two case scenarios: the nominal parameter values were obtained by Borsos and coauthors (2016), in conjunction with the literature, and the case where no prior knowledge of the model parameters is available. In the latter case, a simultaneous identification approach, based on a hybrid global optimization approach that combines a genetic algorithm and a local deterministic approach was used to identify the nominal parameter values presented in Table 3. In both cases the estimability approach will help identify the best parameter estimates of the optimal subset that guarantees maximum prediction capability, given the available experimental data. It is worth mentioning that this forward estimability method can also be used for the optimum design of experiments for improved parameter estimation (Benyahia, 2009; Benyahia et al., 2011).

One crucial step after the identification of the 23 unknown parameters is the evaluation of the uncertainty of the model estimates, since it could provide information regarding the robustness and the predictive capability of the model. One way of assessing these uncertainties is through confidence domains. In this study, a method based on the FIM is used to estimate the 95% confidence intervals of the model-parameters. The mean values of the identified model parameters for the simultaneous approach and the corresponding confidence intervals are presented in Table 3.

Although the confidence interval associated with some of the nominal parameter estimates are reasonably narrow, most of the model parameters as highlighted in red present broad confidence intervals. Statistically speaking, this indicates that the parameters are unidentifiable and consequently the parameter estimates are not reliable. Hence, a good fit should be combined with the estimation of confidence regions and the corresponding correlation matrix. In this way, the reliability of the estimated unknown parameters can be assessed (Table 3). In most of the cases, the cause of these broad confidence domains is associated with the existence of strong correlations amongst the parameters. The correlation effects will be thoroughly discussed in conjunction with the estimability analysis in the next sections.

**Table 3**
Nominal vector of the model parameters and their confidence intervals (C.I.).

| Parameter | Nominal value from Borsos et al. (2016) | Estimated nominal values Current work Value ± C.I. | Units | Corresponding Number |
|---|---|---|---|---|
| $k_{ads,\ 0,\ CGM1}$ | 27.3 | 6.131 ± 5.56 | - | 1 |
| $k_{des,\ 0,\ CGM1}$ | 0.56562 | 0.630 ± 0.62 | - | 2 |
| $\beta_1$ | 4.6 | 10.626 ± 9.30 | $m/K$ | 3 |
| $G_{min,\ 1}$ | $4.5 \times 10^{-4}$ | $2.812 \times 10^{-4}$ ± 0.000281 | $\mu m/s$ | 4 |
| $k_{m,\ 0,\ CGM1}$ | 389.348 | 19.135 ± 0.000268 | $m/s$ | 5 |
| $K_{e,\ CGM1}$ | 0.999 | 1.669 ± 1.114 | - | 6 |
| $\Delta G_{des,\ 1}$ | $2.436 \times 10^{+3}$ | $1.578 \times 10^{+3}$ ± 35.35 | $kJ/mol$ | 7 |
| $\Delta G_{ads,\ 1}$ | $2.2994 \times 10^{+4}$ | $2.17 \times 10^{+4}$ ± 514.29 | $kJ/mol$ | 8 |
| $k_{ads,\ 0,\ CGM2}$ | 11.24 | 4.046 ± 3.51 | - | 9 |
| $k_{des,\ 0,\ CGM2}$ | 0.49127 | 0.4566 ± 0.45 | - | 10 |
| $\beta_2$ | 5.15 | 5.1164 ± 4.45 | $m/K$ | 11 |
| $G_{min,\ 2}$ | 246.952 | 487.034 ± 1.12 | $\mu m/s$ | 12 |
| $k_{m,\ 0,\ CGM2}$ | 61.1286 | 79.096 ± 0.78 | $m/s$ | 13 |
| $K_{e,\ CGM2}$ | 0.994 | 0.997 ± 0.99 | - | 14 |
| $\Delta G_{des,\ 2}$ | $5.301 \times 10^{+3}$ | $6.386 \times 10^{+3}$ ± 69.6074 | $kJ/mol$ | 15 |
| $\Delta G_{ads,\ 2}$ | $2.4181 \times 10^{+4}$ | $2.709 \times 10^{+4}$± 238.392 | $kJ/mol$ | 16 |
| $g_1$ | 1.4776 | 1.553 ± 1.55 | - | 17 |
| $k_{g1}$ | 12.2063 | 21.028 ± 15.58 | $\mu m/s$ | 18 |
| $g_2$ | 1.692 | 1.692 ± 1.67 | - | 19 |
| $k_{g2}$ | 1.7412 | 98.109 ± 23.69 | $\mu m/s$ | 20 |
| $k_{p,\ 0}$ | 100.751 | 334.331 ± 1.17 | $m^{-3}s^{-1}$ | 21 |
| $E_p$ | $2.814 \times 10^{+3}$ | 0.001 ± 0.000643 | $kJ/mol$ | 22 |
| $k_e$ | $1.576 \times 10^{-3}$ | 4.895 ± 0.11 | - | 23 |

### *3.1. Local estimability analysis: Local sensitivity analysis (LSA) and orthogonalization algorithm*

Although the estimability aims in essence at improving the parameter estimates in order to enhance the model prediction capability, the initial parameter estimates, used as nominal values, can play a crucial role in the quality of the sensitivity analysis, both LSA and GSA, and consequently may determine the outcomes of the estimability analysis. Poor nominal model parameters would potentially lead to inaccurate parameter ranking that may lead to a degradation of the predictive capability of the mathematical model (Benyahia, 2009). The investigation of the estimability analysis is carried out, as explained before, in three steps: a local sensitivity analysis is performed in order to evaluate the relative effect of the parameters on the process outputs, then the model parameters are ranked in descending order, in terms of sensitivity magnitude and correlation, by using the orthogonalization algorithm (Table 2). Finally, an incremental optimization approach that consists in a sequential identification of the top $i$th parameters (where $i = 1,2,\ldots$) is utilized to determine the threshold or cut-off value and identify the optimal subset of model parameters .

In Fig. 3, the variation of the dynamic sensitivity of some model parameters is presented. The first selected parameter (Fig. 3(a)), $g_1$, which is the exponent of the growth kinetic equation in the $x_1$ dimension (i.e. along the length of the crystal), shows very high sensitivities at all times. This means that $g_1$, has a strong effect on the model predictions (outputs) and consequently its estimability potential may be very high depending the concurrent correlation effects. The same stands for the second selected parameter (Fig. 3(b)), $k_{e,\ CGM2}$, which describes the thermodynamic mass distribution coefficient for the CGM2 (i.e. sodium hexametaphosphate). These two parameters are likely to be ranked high in terms of estimability potential meaning that the information obtained from the measurements in the considered time window will be adequate for their accurate estimation. On the other hand, $k_{p,\ 0}$ and $k_{m,\ 0,\ CGM1}$ show very week sensitivities at all times. For instance, the sensitivities associated with $k_{m,\ 0,\ CGM1}$ are always below $6 \times 10^{-7}$ which indicates that these model parameters are likely to be practically unidentifiable or inestimable.

Besides the relative effect of the parameters on the outputs, the sensitivity analysis can give a very good indication of the existence of correlations between the parameters. Similar sensitivity trajectories indicate strong correlation as seen in the sensitivity trajectories of $k_{ads,\ 0,\ CGM1}$ and $k_{des,\ 0,\ CGM1}$ (Fig. 4(a)) and $\Delta G_{des,\ 1}$ and $k_{ads,\ 0,\ CGM2}$ (Fig. 4(b)). These outcomes are also consistent with the correlogram (correlation matrix) depicted in Fig. 6.

Similar results could be drawn from Fig. 5, where the whole parameter set is presented in a box plot. This diagram illustrates the variation of the estimated model parameters. The parameters may be classified in three different discrete subgroups according to their contribution to the output: high, moderate and low. As such, some of the parameters such as $\{k_{e,\ CGM2}, \Delta G_{ads,\ 2}, \Delta G_{des,\ 2}, g_1, k_{g1}, g_2, \Delta G_{des,\ 1}\}$ may be classified as parameters with high impact. $\{k_{ads,\ 0,\ CGM1},\ k_{des,\ 0,\ CGM1},\ \beta_1,\ \Delta G_{ads,\ 1},\ k_{ads,\ 0,\ CGM2},\ k_{des,\ 0,\ CGM2},\ \beta_2,\ k_{g2}\}$ and $\{G_{min,\ 1},\ k_{m,\ 0,\ CGM1},\ K_{e,\ CGM1},\ G_{min,\ 2},\ k_{m,\ 0,\ CGM2},\ k_{p,\ 0},\ E_p,\ k_e\}$, on the other hand, seem to present moderate and low sensitivity to the imposed perturbation respectively. Hence a considerable number of the parameters showed low sensitivity values. This lack of sensitivity suggests that the model is over-parametrized (Saltelli et al., 2008). However, model discrimination is beyond the scope of this paper and all parameters are considered essential for other physical aspects of the model performance and may be set to their nominal values without degrading the prediction capabilities of the model. These observations advocate that the vector of the model parameters, as a whole, is practically unidentifiable (from the available data) and further analysis should be done to select an optimal subset of parameters. It should be highlighted that this classification is based utterly on observation of the variation of the sensitivities and is presented as a preliminary qualitative analysis of the results. The formal implementation of the estimability approach and identification of the cut-off value will be discussed in the subsequent sections.

A robust approach for ranking the model parameters according to their estimability potential is based on the orthogonalization algorithm (Table 2), which takes into account both the sensitivity magnitude (i.e. Euclidean norm) and correlation during the sequential selection of the most estimable parameter. The results obtained based on modified Gram-Schmidt orthogonalization algorithm are presented in Table 4. The exponents of the growth kinetic equations in the $\{x_1, x_2\}$ dimensions (i.e. $g_1$ and $g_2$) indicate high estimability potential. This was expected since these parame-

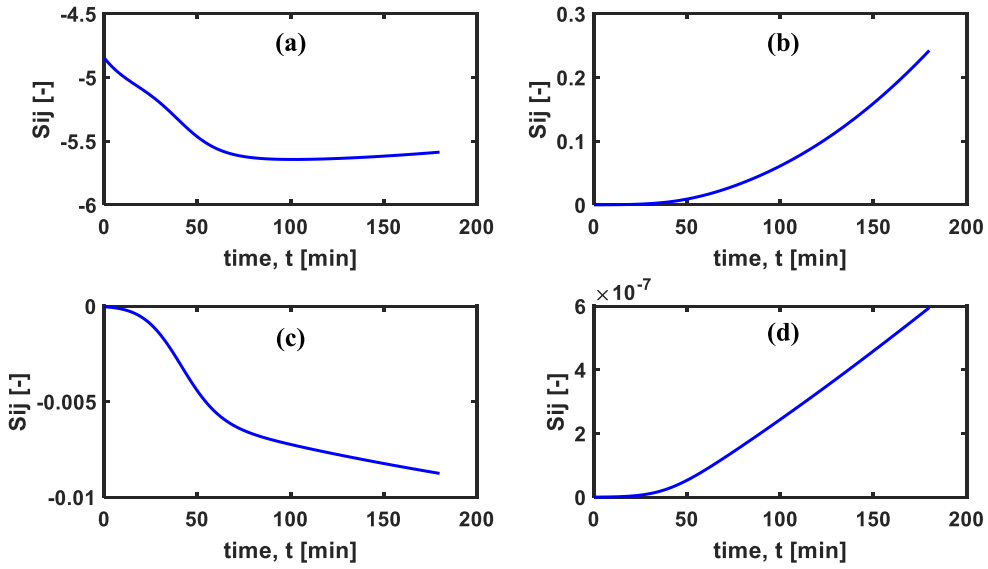


**Fig. 3.** Comparison of the dynamic sensitivity of selected model parameters: (a) $g_1$ (b) $K_{e,\,CGM2}$ (c) $k_{p,\,0}$ (d) $k_{m,\,0,\,CGM1}$,.

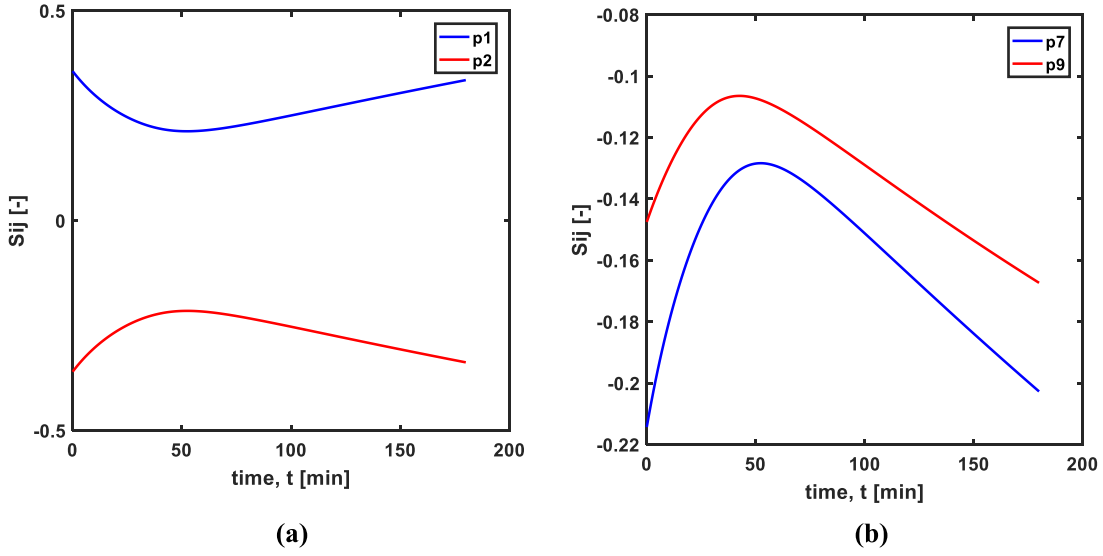


**Fig. 4.** Comparison of the sensitivity profiles of selected model parameters: (a) $p_1 : k_{ads,0,CGM1}$ & $p_2 : k_{des,0,CGM1}$ and (b) $p_7 : \Delta G_{des,1}$ & $p_9 : k_{ads,0,CGM2}$.

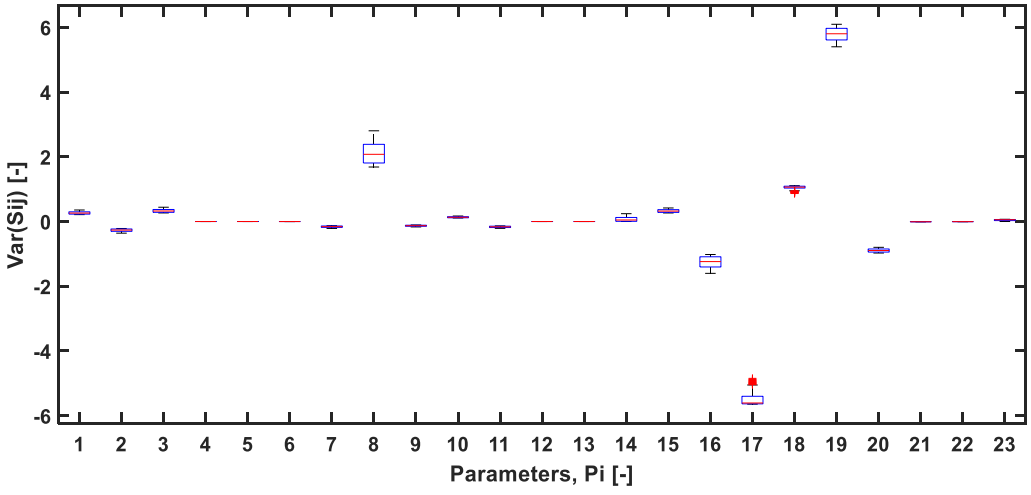


**Fig. 5.** Box plot illustrating the variation of the estimated model parameters.

**Table 4**
Ranking of parameters with the highest estimability potential.

| Method | Parameter ranking |
|---|---|
| Orthogonalization algorithm | $g_1$, $k_{e,\ CGM2}$, $\Delta G_{ads,\ 1}$, $g_2$, $\Delta G_{ads,\ 2}$, $k_{g1}$, $\beta_1$, $\Delta G_{des,\ 2}$, $k_{ads,\ 0,\ CGM1}$, $k_{des,\ 0,\ CGM1}$, $k_{g2}$, $\Delta G_{des,\ 1}$, $\beta_2$, $k_{des,\ 0,\ CGM2}$, $k_{ads,\ 0,\ CGM2}$, $k_e$, $K_{e,\ CGM1}$, $k_{p,\ 0}$, $k_{m,\ 0,\ CGM2}$, $G_{min,\ 1}$, $E_p$, $G_{min,\ 2}$, $k_{m,\ 0,\ CGM1}$<br>**High Estimability Potential** → **Low Estimability Potential**<br>17, 14, 8, 19, 16, 18, 3, 15, 1, 2, 20, 7, 11, 10, 9, 23, 6 , 21, 13, 4, 22, 12, 5 |

ters represent the exponential factors of the crystal growth rates used in the model algebraic equations (i.e. empirical power law expressions). The absorption energy of the impurities (i.e. $\Delta G_{ads,\ 1}$ and $\Delta G_{ads,\ 2}$) also appear to have a significant impact on the outputs since they are highly ranked on the list illustrating high estimability potential. It is also evident, that the kinetic parameters corresponding to the nucleation mechanism are ranked quite low $\{k_{p,\ o}, E_p, k_e\}$ because of their weak sensitivity coefficients at the sampling times. This reveals how critical is the incorporation of the estimability analysis in the development of the design of experiment and consequently in mathematical modelling. Moreover, it is known that the AR measurements can provide negligible information regarding the nucleation phenomena. This limitation maybe overcome by incorporating additional PAT to measure the concentration and number of counts (focussed beam reflectance (FBRM)) and considering these two variables as process outputs in the estimability framework.

The estimability analysis revealed that the data are not adequate to estimate accurately the nucleation kinetics. This is key, especially in systems utilizing different sensors. As such, the information content of each sensor may be assessed and consequently the number of parameters that can be estimated from each individual PAT or from their combination (e.g. sensors providing different outputs) may be determined. This may also be employed for the evaluation of the accuracy of the measuring method with the scope of collecting more accurate measurements. Hence, the estimability analysis can be utilized for the selection of the appropriate PAT and consequently the most efficient strategy to collect the experimental data to improve the accuracy of the model parameters, which in turns enhances the model reliability in key applications such as process design and control.

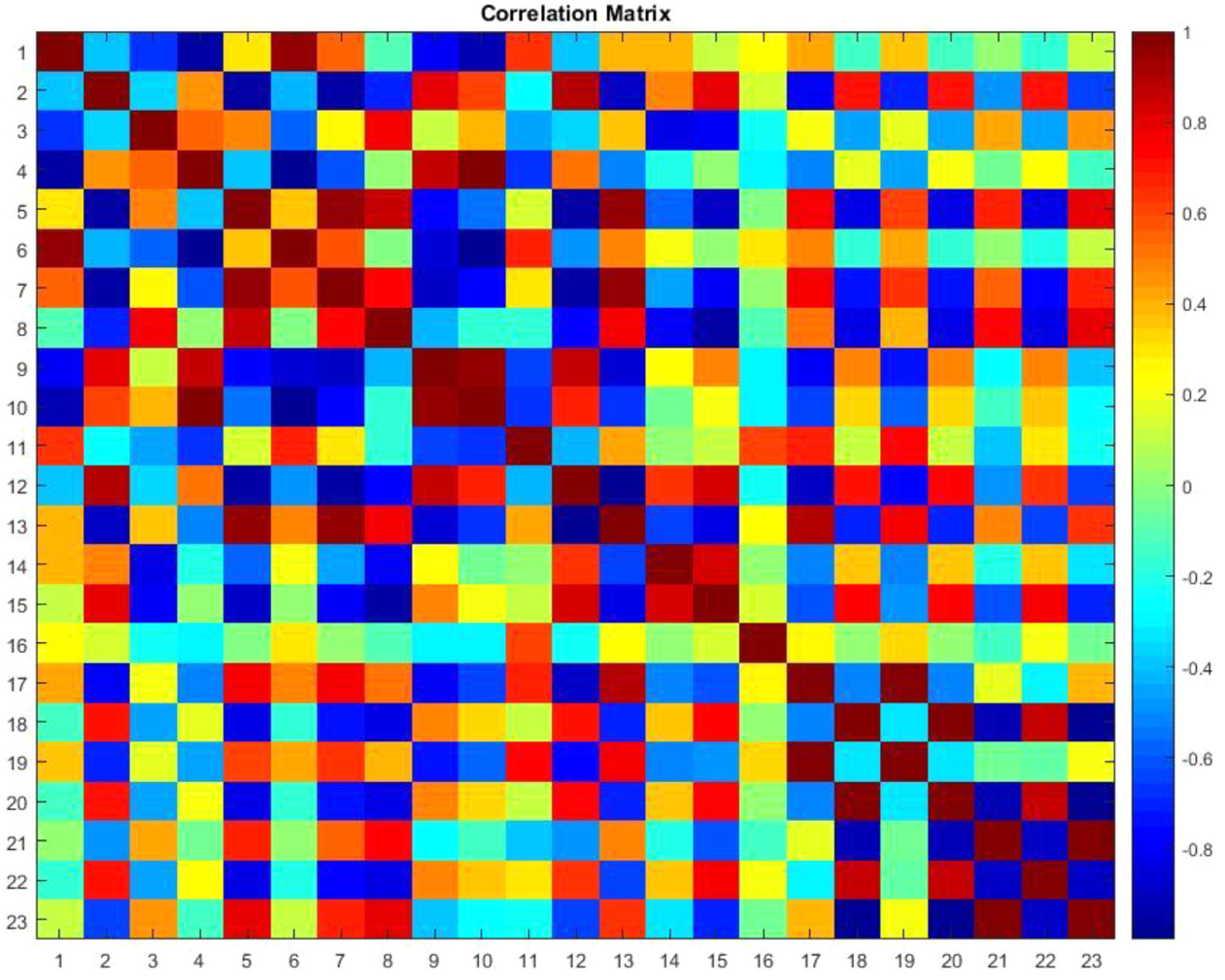


**Fig. 6.** Correlation matrix for the estimated nominal parameter set.

The sequential orthogonalization approach helps select the parameters with the highest sensitivity and least correlation. To provide a more rigorous insight into the correlation effects, a correlation matrix, can be computed by using the Pearson method (Kendall et al., 1977). The $23 \times 23$ correlation matrix obtained from the covariance matrix (inverse of the FIM) is presented in Fig. 6. As shown, strong positive or negative correlations exist between some parameters. For instance, the parameters $p_{20}$ and $p_{23}$ present very strong negative correlation. Hence, any change that occurs in $p_{20}$ can be compensated by an inverse change in $p_{23}$. Similar interaction patterns do occur between the kinetics describing the nucleation phenomenon (i.e. $p_{21}, p_{22}, p_{23}$), which present strong positive and negative correlation. In general, the presence of high correlations, especially if it involves many parameters, can make the identification process difficult and inaccurate (not unique).

In order to identify the optimal subset of parameters that maximize model reliability, a cut-off value should be set as a boundary between the parameters with high estimability potential, that can be identified reliably, and the remaining parameters that are poorly identifiable and should be set to their nominal value without degrading the prediction capability of the model. As such, the cut-off value is critical as it affects both the cost and quality of the estimability approach and consequently the prediction capability of the model.

To identify the cut-off value and consequently the optimal subset of parameters, an iterative approach is performed. It consists in identifying incrementally the subsets of model parameters, from the experimental data, according to their estimability potential (Table 4) staring with the top ranked parameter, then the top two parameters and so forth. This approach will help identify the optimal objective function threshold (i.e. cut-off value), beyond which all improvements are significant, and consequently the optimum identifiable subset of parameters. The results of the iterative incremental approach are depicted in Fig. 7. Typically, when a mean square error approach is considered, the objective function, $J(p)$, decreases until a plateau is reached. The initial point of the plateau can be considered as the cut-off value as no significant improvement can be achieved from that point onwards, which consequently sets the limit of the optimal identifiable parameter set. Fig. 7(a) indicates that the top 7 ranked parameters, in the case of the nominal values obtained from Borsos et al. (2016), and the top 8 ranked parameters, in the case of the simultaneous optimization approach, are sufficient to capture the information contained in the experimental data. Despite the fact that using more parameters may lead to a slight decrease in the objective function, as depicted in Fig. 7, the estimability approach guarantees the best trade-off between model reliability and minimum set of parameters to be identified. Fig. 7 also confirms that different nominal parameter values, as clearly shown in Fig. 7(a) and (b), lead to different threshold values (340 in case a and 290 in case b). In this particular case, the estimability approach implemented with the nominal vector inherent to a simultaneous identification approach out forms the quality of the one carried out with Borsos and co-authors' nominal value obtained sequentially. It should be noted that the objective functions show non-smoothness in both cases which is likely due to the high nonlinearity and stiffness of the set of ODEs and the increased correlation between the parameters as more parameters are being added. This non-smooth behaviour may also indicate that the local solver got stuck in local optima.

### 3.2. Global sensitivity analysis

The global sensitivity analysis (GSA) is utilized here in order to assess the performance of the model itself and to cross-validate the local estimability analysis approach discussed earlier. The method provides another alternative to rank the model parameters and identify the optimal set of parameters that could be estimated from the experimental data. In this case, the total order sensitivity index will be used to rank the parameters, followed by an incremental optimization-based selection approach whose performances will be compared against the previously described estimability approach, associated with the local sensitivities.

The Sobol analysis is performed as follows. Firstly, a nominal set of model parameters is defined followed by the definition of the probability distributions for each individual parameter. In this work, a Gaussian distribution was assigned for every parameter by considering 2% variance. Narrow limits are applied since the population balance models for crystallization processes present, in general, high stiffness, which might have a considerable effect on the computational burden. Random combinations of the parameter values are generated from the assigned probability distribution functions. Thus, the output of the model is evaluated for different parameter sets along with the uncertainties. Consequently, the sensitivity indices are calculated in order to assess the effect of the parameters and rank them accordingly.

The global sensitivity analysis is performed here by taking into account two different scenarios. In the first scenario, the effects of the parameters are analyzed considering only the model predicted outputs inherent to the set of the DAEs representing the studied system. Hence, the impact of the parameters on the joint moments and on the concentration of the solution and impurities is investigated based exclusively on simulations (i.e. without considering the sampling times). The second scenario considers the mean AR measurements. Hence, Sobol analysis is applied for the decomposition of the variance which is associated with the difference between experimental and simulation data (i.e. global estimability analysis). In more detail, the computation of the root mean square error can provide information with the scope of parameter ranking and model selection (cut-off value determination). The 23 unknown model parameters estimated in this work and defined in Table 3 are used as inputs for the sensitivity and estimability analysis.

It was demonstrated that a tradeoff between computational accuracy and efficiency of the first and total order sensitivity indices can be achieved at a cost of $(N_p + 2)$ $N$ model evaluations (Saltelli et al., 2005), where $N$ is the number of samples that should be between $5 \times 10^2$ and $1 \times 10^3$ and $N_p$ is the number of parameters (23 in our case). In this analysis, a conservative approach is adopted by considering $N = 1 \times 10^3$ and the total number of evaluations as $25 \times 10^3$ for both scenarios. The results were validated using different numbers of samples ($N$) to ensure consistency and robustness.

The results are summarized in Figs. 8 and 9, where the first and total order Sobol sensitivity indices are presented in descending order for the two scenarios. Fig. 8(a) and (b) indicate that both first and total order Sobol sensitivity indices yield the same order of priority for the first scenario which illustrates that certain parameters have a considerable impact on the output variable (i.e. AR) both directly (relative impact of each input variable) and indirectly (interaction among the input parameters). In a similar way to the discussion above, the effect of randomly generated subsets of parameters on the mean square error between the measured and predicted mean AR is analyzed for the second scenario.

The greater the sensitivity indices are, the more critical the parameters are for the model. Figs. 8 and 9 show that the parameters $g_1$ and $g_2$, which are the exponents of the growth kinetic equations in the $x_1$ and $x_2$ dimension respectively, possess the highest total sensitivity indexes. This was expected since a growth dominated physical system is under investigation. The analysis also demonstrates that $\Delta G_{ads,\,2}$, $\Delta G_{des,\,2}$ and $K_{e,\,CGM2}$, which represent the adsorption, desorption kinetics and the thermodynamic mass distribution coefficient for CGM2 respectively, can be reliably identified.

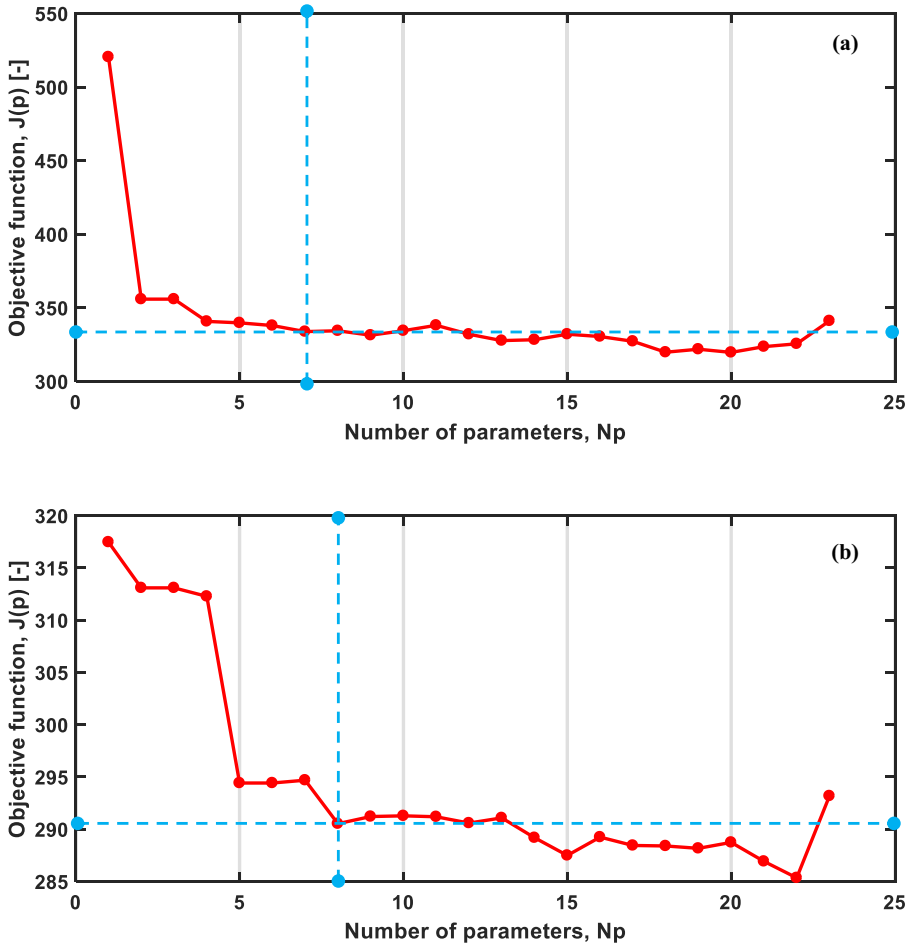


**Fig. 7.** Maximum likelihood error vs the number of selected parameters for: (a) nominal set of parameters estimated by Borsos et al., (2016) – sequential approach and (b) nominal set of parameters estimated in this work – simultaneous approach.

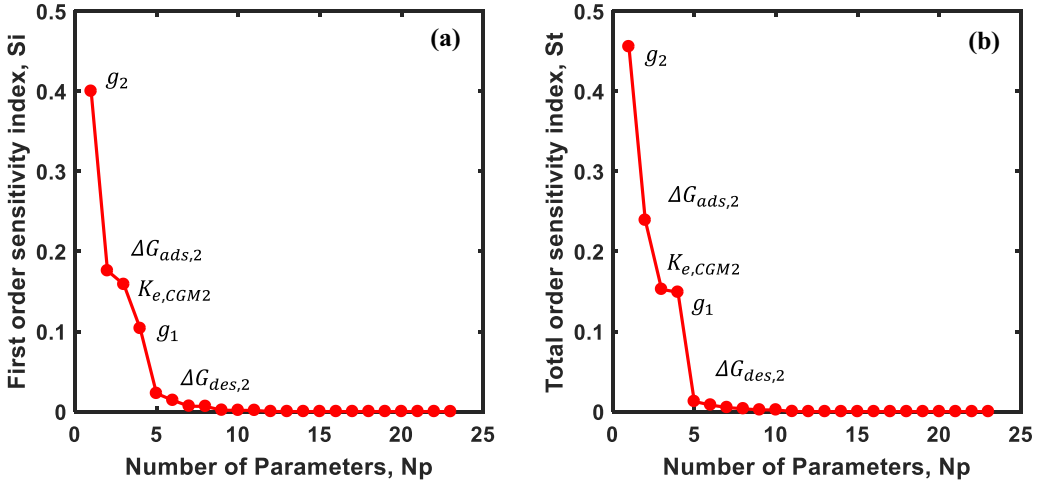


**Fig. 8.** Sobol analysis for the 1st case scenario: (a) first order sensitivity indices of the 23 parameters in descending order (b) total order sensitivity indices in descending order.

This can be anticipated as well since it was experimentally proven (Borsos et al., 2016) that the CGM2 (i.e. sodium hexametaphosphate) has a more prominent effect compared to CGM1 (i.e. aluminum sulfate). When both growth modifiers are present in the system, the AR decreases which is caused by CGM2, even when lower amounts of CGM2 are used. The nucleation kinetics present low sensitivity values, which is consistent with the outcomes of the estimability analysis based on local sensitivities. By comparing the two scenarios, the majority of the parameters show significant lack of sensitivity. However, in both scenarios interesting patterns emerge. Sensitive parameters (high $s_i$ values) affect the output through both direct and indirect effects (high $s_t$ values). Thus, the parameters with moderate and low sensitivity values cannot affect the system even indirectly (i.e. through interactions) from a sensi-

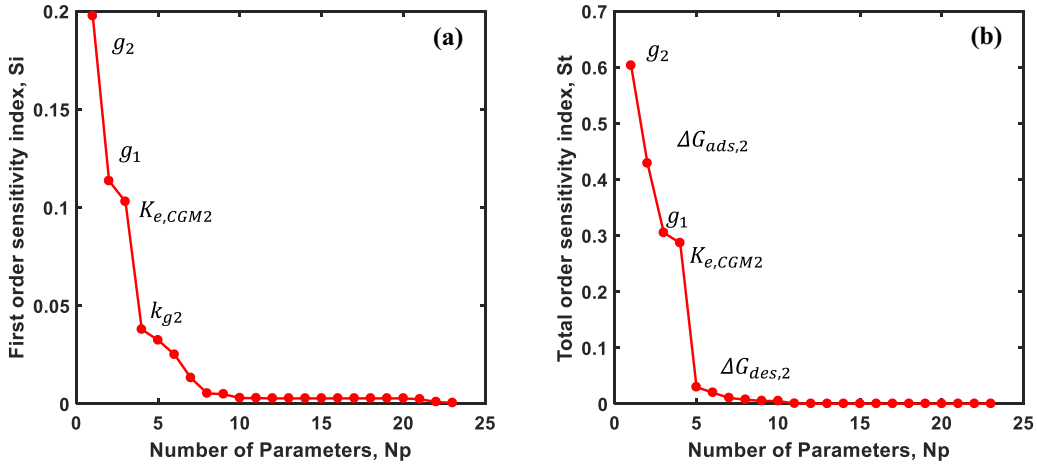


**Fig. 9.** Sobol analysis for the 2nd case scenario: (a) first order sensitivity indices of the 23 parameters in descending order (b) total order indices in descending order.

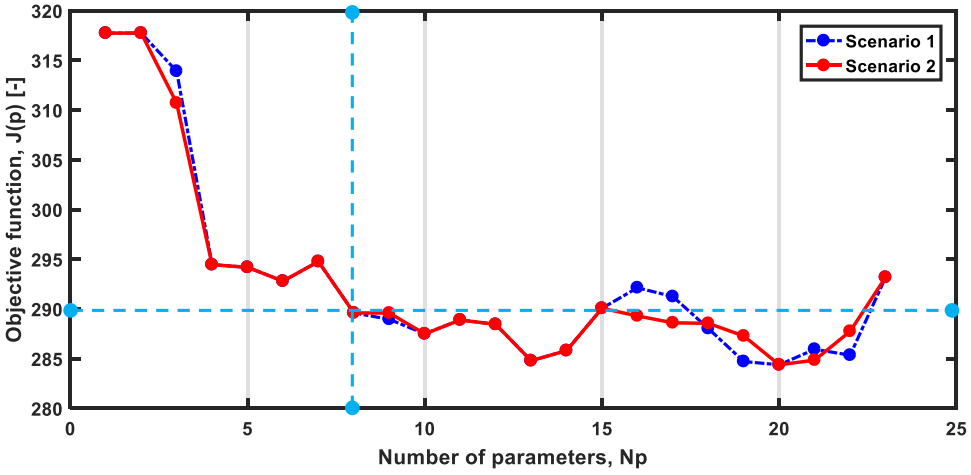


**Fig. 10.** Maximum likelihood error vs the number of selected parameters for both Sobol scenarios.

tivity point of view. Overall, the Sobol analysis demonstrates that a large number of parameters can be set to nominal values without degrading the model prediction capabilities.

The total order indices, presented in Figs. 8(b) and 9(b), are used to identify the cut-off value for the selection of the optimal subset of model parameters. As it can be seen, the values of the total order indices are reduced until a plateau is reached. The initial point of the plateau can be considered as the cut-off value since the addition of more parameters, from that point onwards, does not improve the prediction capability of the model

The Sobol analysis indicates that a cut-off value can be identified directly from the total order indices and accordingly the top 7 and 8 parameters are sufficient to build a reliable model for the 1st and 2nd scenarios respectively. However, for the sake of consistency and in order to enable a reliable comparison between the two estimability methods, the cut-off value will be identified from the profile of the objective function associated with the parameter identification problem (minimization of the error between the model predictions and the experimental data). The profile of the objective functions for the two scenarios, obtained by an incremental iterative approach as described above for the case of LS-based estimability, are depicted in Fig. 10. As noticed in the previous case, the objective function decreases significantly with the introduction of the top few parameters. The diagram confirms that the selection of the top 8 parameters can be sufficient enough to maximize the prediction capabilities of the model. Despite these consistent outcomes, the selection process through an incremental iterative parameter estimation procedure is highly advised as it is more reliable compared to the selection based on the magnitudes of the total order sensitivity index. Fig. 10 also shows non-smooth behavior similar to Fig. 7.

To make a reliable and effective comparison between the two methods described in the paper (the estimability method based on LS and Sobol method with two scenarios), the parameter ranking and optimal parameter sets are summarized in Table 5. Although, each method yields a different classification. As expected, some consistency was achieved as the same four parameters, highlighted in red, which were identified by both methods as the ones with the most prominent effects. The inconsistencies can be explained by the fact that the methods use essentially different approaches, LS and GL, besides, the quality of the nominal vector of parameters can play a key role in both cases. Although, both technics can be used separately, the outcomes of the analysis show that their combination can provide a more systematic and robust selection of the subset of parameters that provide guaranteed optimal model prediction capabilities, based on the available data. In addition, the methodology can provide a basis to assess the quality and quantity of the experimental data or alternatively inform or help design the required experiments and/or measurements (DoE) that could improve the estimability potential of a specific parameter, which in turn helps improve the prediction capabilities of the mathematical model, particularly in the case of multi-dimensional popula-

**Table 5**
Summary of the parameter ranking based on Orthogonalization algorithm and the Sobol analysis.

| Orthogonalization (LSA) | | Sobol Analysis (GSA) Estimated nominal vector of parameters | | | |
|---|---|---|---|---|---|
| Nominal parameters from Borsos et al., 2016 | Estimated nominal vector of parameters | 1st Scenario | | 2nd Scenario | |
| | | $S_i$ | $S_t$ | $S_i$ | $S_t$ |
| $\boldsymbol{g_1\ (p_{17})}$ | $g_1\ (p_{17})$ | $g_2\ (p_{19})$ | $g_2\ (p_{19})$ | $g_2\ (p_{19})$ | $g_2\ (p_{19})$ |
| $\boldsymbol{k_{e,\ CGM2}\ (p_{14})}$ | $k_{e,\ CGM2}\ (p_{14})$ | $\Delta G_{ads,\ 2}\ (p_{16})$ | $\Delta G_{ads,\ 2}\ (p_{16})$ | $g_1\ (p_{17})$ | $\Delta G_{ads,\ 2}\ (p_{16})$ |
| $\boldsymbol{g_2\ (p_{19})}$ | $\Delta G_{ads,\ 1}\ (p_8)$ | $k_{e,\ CGM2}\ (p_{14})$ | $k_{e,\ CGM2}\ (p_{14})$ | $k_{e,\ CGM2}\ (p_{14})$ | $g_1\ (p_{17})$ |
| $\boldsymbol{\Delta G_{ads,\ 1}\ (p_8)}$ | $g_2\ (p_{19})$ | $g_1\ (p_{17})$ | $g_1\ (p_{17})$ | $k_{g2}\ (p_{20})$ | $k_{e,\ CGM2}\ (p_{14})$ |
| $\boldsymbol{\Delta G_{ads,\ 2}\ (p_{16})}$ | $\Delta G_{ads,\ 2}\ (p_{16})$ | $\Delta G_{des,\ 2}\ (p_{15})$ | $\Delta G_{des,\ 2}\ (p_{15})$ | $\Delta G_{ads,\ 2}\ (p_{16})$ | $\Delta G_{des,\ 2}\ (p_{15})$ |
| $\boldsymbol{k_{g1}\ (p_{18})}$ | $k_{g1}\ (p_{18})$ | $k_{g2}\ (p_{20})$ | $k_{g2}\ (p_{20})$ | $\Delta G_{des,\ 2}\ (p_{15})$ | $k_{g2}\ (p_{20})$ |
| $\boldsymbol{\beta_1\ (p_3)}$ | $\beta_1\ (p_3)$ | $k_{g1}\ (p_{18})$ | $\beta_2\ (p_{11})$ | $k_{ads,\ 0,\ CGM2}\ (p_9)$ | $\beta_2\ (p_{11})$ |
| $\boldsymbol{k_{g2}\ (p_{20})}$ | $\Delta G_{des,\ 2}\ (p_{15})$ | $\beta_2\ (p_{11})$ | $k_{g1}\ (p_{18})$ | $k_{g1}\ (p_{18})$ | $k_{g1}\ (p_{18})$ |
| $\boldsymbol{\Delta G_{des,\ 2}\ (p_{15})}$ | $k_{ads,\ 0,\ CGM1}\ (p_1)$ | $k_e\ (p_{23})$ | $k_{des,\ 0,\ CGM2}(p_{10})$ | $k_{des,\ 0,\ CGM2}(p_{10})$ | $k_{ads,\ 0,\ CGM2}(p_9)$ |
| $\boldsymbol{\beta_2\ (p_{11})}$ | $k_{des,\ 0,\ CGM1}\ (p_2)$ | $k_{des,\ 0,\ CGM2}\ (p_{10})$ | $k_{ads,\ 0,\ CGM2}\ (p_9)$ | $k_{p,\ 0}\ (p_{21})$ | $k_{des,\ 0,\ CGM2}(p_{10})$ |
| $\boldsymbol{k_{ads,\ 0,\ CGM1}\ (p_1)}$ | $k_{g2}\ (p_{20})$ | $k_{ads,\ 0,\ CGM2}\ (p_9)$ | $k_e\ (p_{23})$ | $E_p\ (p_{22})$ | $k_e\ (p_{23})$ |
| $\boldsymbol{k_{des,\ 0,\ CGM1}\ (p_2)}$ | $\Delta G_{des,\ 1}\ (p_7)$ | $E_p\ (p_{22})$ | $k_{m,\ 0,\ CGM2}\ (p_{13})$ | $G_{min,\ 2}\ (p_{12})$ | $k_{m,\ 0,\ CGM2}\ (p_{13})$ |
| $\boldsymbol{\Delta G_{des,\ 1}\ (p_7)}$ | $\beta_2\ (p_{11})$ | $G_{min,\ 2}\ (p_{12})$ | $k_{p,\ 0}\ (p_{21})$ | $k_{m,\ 0,\ CGM1}\ (p_5)$ | $k_{p,\ 0}\ (p_{21})$ |
| $\boldsymbol{k_{des,\ 0,\ CGM2}\ (p_{10})}$ | $k_{des,\ 0,\ CGM2}\ (p_{10})$ | $G_{min,\ 1}\ (p_4)$ | $E_p\ (p_{22})$ | $K_{e,\ CGM1}\ (p_6)$ | $E_p\ (p_{22})$ |
| $\boldsymbol{k_{ads,\ 0,\ CGM2}\ (p_9)}$ | $k_{ads,\ 0,\ CGM2}\ (p_9)$ | $\Delta G_{ads,\ 1}\ (p_8)$ | $G_{min,\ 2}\ (p_{12})$ | $k_{des,\ 0,\ CGM1}\ (p_2)$ | $G_{min,\ 2}\ (p_{12})$ |
| $\boldsymbol{k_{p,\ 0}\ (p_{21})}$ | $k_e\ (p_{23})$ | $\beta_1\ (p_3)$ | $\beta_1\ (p_3)$ | $\Delta G_{des,\ 1}\ (p_7)$ | $k_{m,\ 0,\ CGM1}\ (p_5)$ |
| $\boldsymbol{k_e\ (p_{23})}$ | $K_{e,\ CGM1}\ (p_6)$ | $\Delta G_{des,\ 1}\ (p_7)$ | $k_{ads,\ 0,\ CGM1}\ (p_1)$ | $\beta_1\ (p_3)$ | $k_{ads,\ 0,\ CGM1}\ (p_1)$ |
| $\boldsymbol{K_{e,\ CGM1}\ (p_6)}$ | $k_{p,\ 0}\ (p_{21})$ | $K_{e,\ CGM1}\ (p_6)$ | $k_{des,\ 0,\ CGM1}\ (p_2)$ | $G_{min,\ 1}\ (p_4)$ | $\Delta G_{ads,\ 1}\ (p_8)$ |
| $\boldsymbol{k_{m,\ 0,\ CGM2}\ (p_{13})}$ | $k_{m,\ 0,\ CGM2}\ (p_{13})$ | $k_{ads,\ 0,\ CGM1}\ (p_1)$ | $K_{e,\ CGM1}\ (p_6)$ | $\Delta G_{ads,\ 1}\ (p_8)$ | $\beta_1\ (p_3)$ |
| $\boldsymbol{E_p\ (p_{22})}$ | $G_{min,\ 1}\ (p_4)$ | $k_{des,\ 0,\ CGM1}\ (p_2)$ | $\Delta G_{ads,\ 1}\ (p_8)$ | $k_{ads,\ 0,\ CGM1}\ (p_1)$ | $G_{min,\ 1}\ (p_4)$ |
| $\boldsymbol{G_{min,\ 2}\ (p_{12})}$ | $E_p\ (p_{22})$ | $k_{m,\ 0,\ CGM1}\ (p_5)$ | $G_{min,\ 1}\ (p_4)$ | $k_{m,\ 0,\ CGM2}\ (p_{13})$ | $k_{des,\ 0,\ CGM1}\ (p_2)$ |
| $\boldsymbol{G_{min,\ 1}\ (p_4)}$ | $G_{min,\ 2}\ (p_{12})$ | $k_{p,\ 0}\ (p_{21})$ | $\Delta G_{des,\ 1}\ (p_7)$ | $\beta_2\ (p_{11})$ | $K_{e,\ CGM1}\ (p_6)$ |
| $\boldsymbol{k_{m,\ 0,\ CGM1}\ (p_5)}$ | $k_{m,\ 0,\ CGM1}\ (p_5)$ | $k_{m,\ 0,\ CGM2}\ (p_{13})$ | $k_{m,\ 0,\ CGM1}\ (p_5)$ | $k_e\ (p_{23})$ | $\Delta G_{des,\ 1}\ (p_7)$ |

tion balance models. Although Sobol analysis provides one of the most accurate methods for calculating the sensitivities of the parameters, the method doesn't consider the correlation between the parameters systematically during the ranking process as opposed the local estimability which addresses quite effectively the correlations issue, as the sequential orthogonalization approach is used precisely to exclude the parameters showing high correlations from being selected amongst the optimal top ranked set.

Finally, to further demonstrate the benefits of the estimably analysis and appraise the prediction capability of the model built with the optimal subset of parameters, the model predictions are compared against the experimental data as well as the predictions of the model built without the estimability approach (Borsos et al., 2016). Three different experiments associated with mean AR measurements are used, as shown in Fig. 11. It should be noted that obtaining accurate AR data is very challenging. The PVM is currently the main technique available to measure real time the AR despite the inherent noisy and non-smooth data, as clearly seen in Fig. 11, which is commonly associated with most of image monitoring tools.

Although both model seem to provide a good fitting, Fig. 11(a) and (b) show that the mathematical model with estimability approach demonstrates better prediction capability. This outcome is consistent with the incremental objective function analysis (Fig. 7). The results show that the model build by identifying the 8 most estimable parameters outperforms the one build by identifying all parameters sequentially, as can be seen also in Fig. 12. It becomes clear that the estimability approach makes the identification process more accurate and less laborious, as a reduced set of parameters is identified while the rest of the parameters are kept to their nominal values without compromising the prediction capability of the mathematical model.

## 4. Conclusions

Parameter estimability is essential to assess whether the model parameters can be reliably identified from existing data, which consequently provides a key step towards more predictable and robust mathematical models. Within this perspective, a novel estimability framework that combines a sequential orthogonalization of the local sensitivity matrix and Sobol, a variance-based global sensitivities technic, was proposed. The estimability analysis requires an initial or nominal vector of model parameters. When either of the two situations occurs: a nominal vector of parameters is not available or the initial parameter estimates are considered as highly uncertain, the framework suggests a simultaneously identification of the whole set of parameters using a hybrid global optimization approach. The estimability procedure can be then conducted using the nominal vector of parameters in conjunction with the available experimental data. The systematic combination of two different estimability methods guarantees a robust selection of the optimal subset of parameters; the set that can be identified more reliably with a guaranteed maximum model prediction capability. As such, both parameters significance and correlations should be considered to rank the model parameters. The framework suggests a systematic methodology, based on the parameter identification objective function, to identify the cut-off value which indicates the boundary between the parameters that can be reliably identified (the optimal subset) and those who should be set to their nominal value. When the resulting model prediction capability is not satisfactory or/and very limited number of parameters can be identified reliably, the method suggests extracting additional experimental data that can be based on appropriate design of experiments.

As a validation step, the methodology was implemented to a complex multi-dimensional morphological population balance for batch crystallization processes, which combines the effects of different crystal growth modifiers/ impurities on the crystal size and shape distribution of the population of needle-like crystals. Initially, two situations were considered regarding the nominal vector of parameters: parameters obtained from literature and those identified using a simultaneous global optimization. The first evaluation of the quality of the nominal vector of parameters revealed that most of the nominal parameters are inherently uncertain, based on the confidence domains, which justifies the need for the estimability analysis. The 23 model parameters were ranked accordingly in terms of highest local sensitivity magnitude and least correlation, in the case of the sequential orthogonalization method, and total

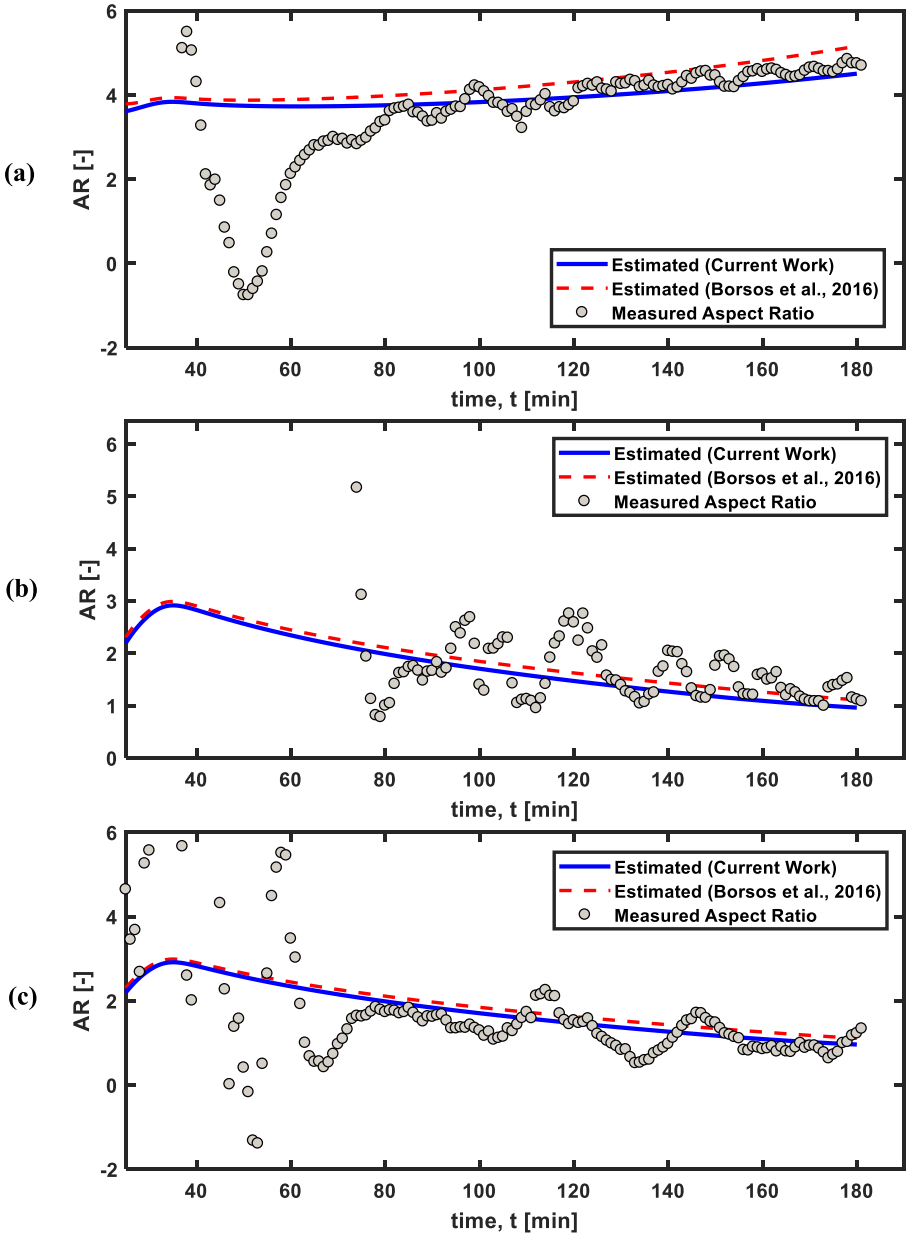


**Fig. 11.** Comparison between the experimental and simulated mean AR dynamic evolution: (a) Experiment 1 (400 g $H_2O$; 150 gr KDP; 12.5 ppm CGM1; 0.0 ppm CGM2), (b) experiment 2 (400 g $H_2O$; 150 gr KDP; 12.5 ppm CGM1; 7.5 ppm CGM2) and (c) experiment 3 (400 g $H_2O$; 150 gr KDP; 0.0 ppm CGM1; 5.0 ppm CGM2).

order sensitivity indices, in the case of Sobol. The correlation patterns confirmed the existence of strong correlation between some parameters, which helped explain the resulting parameter ranking. The least square incremental parameter identification procedure helped determine the cut-off value and consequently the optimal subset of parameters which turned out to be 8 parameters using both methods. Despite some slight parameter ranking differences, the two different estimability methods managed to capture consistently the most significant parameters. However, it is highly recommended to run both methods to maximize the benefits of the estimability approach and minimize the least square value at the cut off value, which guarantees maximum model prediction capability. The case study showed that although noisy AR data with low information content were used, a set of the most influential and the least correlated parameters could be identified, providing enhanced prediction capabilities of the dynamic model of the studied crystallization process. As a consequence, the framework can be extremely valuable in complex model systems when a large number of parameters needs be identified from low information content data, which is commonly encountered in real systems. The proposed framework can also embed an optimization of the experimental campaign to maximize the information content and reduce the cost inherent to redundant experimental information. In the case of systems utilizing different sensors, the information content of each sensor can be assessed and consequently the number of parameters that can be estimated from each individual PAT or from their combination (e.g. sensors providing different outputs) can be determined, which helps select the most appropriate PAT depending on the targeted level of prediction capability and application (e.g. process control).

## Acknowledgments

The authors would like to thank EPSRC and the Doctoral Training Centre in Continuous Manufacturing and Crystallization(grant

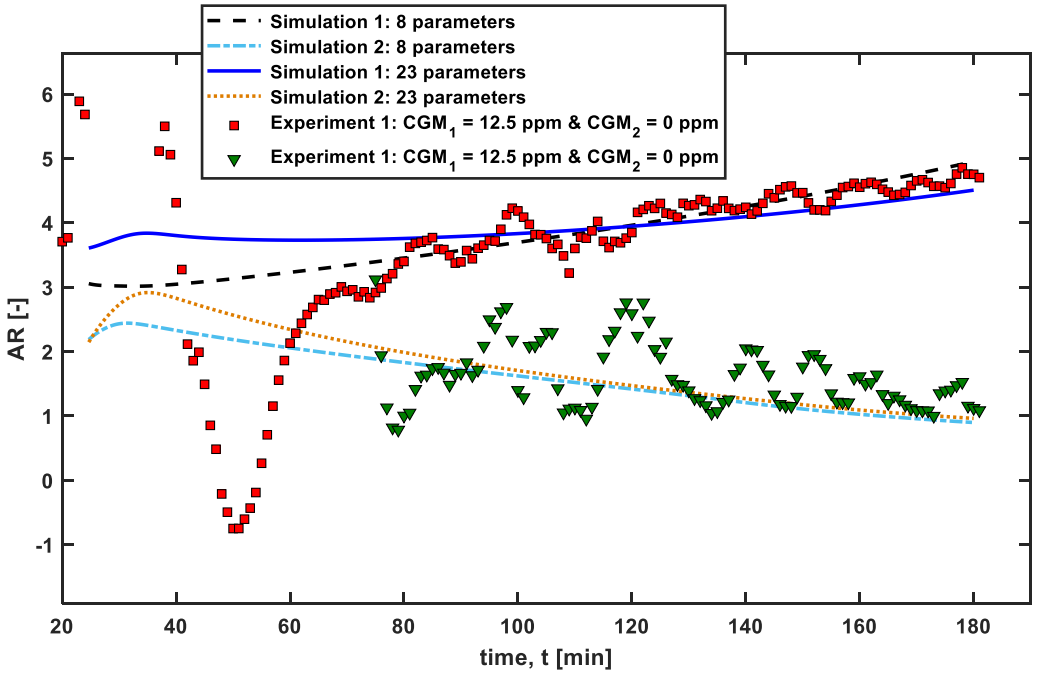


**Fig. 12.** Comparison between the experimental and simulated mean AR dynamic evolution: Experiment 1 (400 g $H_2O$; 150 gr KDP; 12.5 ppm CGM1; 0.0 ppm CGM2) and experiment 2 (400 g $H_2O$; 150 gr KDP; 12.5 ppm CGM1; 7.5 ppm CGM2) by considering 8 and 23 parameters.

ref: EP/K503289/1) for funding D.F. and the European Research Council (grant ref: 200106-CrySys) for funding A.B.

## Appendix A. Multidimensional PBM model with multi-impurity adsorption model (MIAM)

To describe the dynamic evolution of the crystal shape distribution, a multidimensional population balance equation (PBEs) with two characteristic lengths $\boldsymbol{x_l} = \{x_1, x_2\}$ was considered that can be written as:

$$\frac{\partial n(t,x)}{\partial t} + \frac{\partial n[G_1\ n(t,x)]}{\partial x_1} + \frac{\partial n[G_2 n(t,x)]}{\partial x_2} = B_p\ \delta(x_1 - x_{1,0})\ \delta(x_2 - x_{2,0}) \tag{A.1}$$

where $n(t, x)$ is the number density function, $\delta(x - x_0)$ is the delta distribution that characterizes the formation of the nuclei, $B_p$ is the primary nucleation rate and $G_i$ is the crystal growth rate of the $i^{th}$ characteristic crystal facet. The initial and boundary conditions of the PBE are respectively:

$$n(\boldsymbol{x_l}, t = 0) = n_0(\boldsymbol{x_l}) \tag{A.2}$$

$$G_i n(\boldsymbol{x_l}, t) = 0,\ \ \boldsymbol{x_l} \in\ \partial\Omega_{sz} \tag{A.3}$$

where $\partial\Omega_{sz}$ is the space boundary of the particle size.

The model can be reduced from a partial differential equation (PDE) to a set of ordinary differential equations (ODEs) by using the standard method of moments (SMOM). Since only average properties are needed for the determination of the mean crystal AR, the SMOM method can provide an efficient and accurate method for the estimation of the key characteristics of the crystal population.

The joint moments of internal variables can be calculated as:

$$\mu_{k,m}(t) = \int_0^\infty \int_0^\infty {x_1}^k {x_2}^m\ n(x_1, x_2, t) dx_1 dx_2\ ,\ \ m, k = 0, 1, 2 \ldots \tag{A.4}$$

Hence, by applying the moment transformation rule to the PBE (Eq. (A.1)), considering the initial (Eq. (A.2)) and boundary conditions (Eq. (A.3)), a finite set of ODEs can be acquired:

$$\frac{\partial \mu_{0,0}}{\partial t} = B_p;\ \ \frac{\partial \mu_{m,r}}{\partial t} = m\ G_1\ \mu_{m-1,r} + r\ G_2\ \mu_{m,r-1},\ \ m, r = 0, 1, 2, \ldots \tag{A.5}$$

This set of ODEs coupled with the component mass balances, for the solute and impurities, describes a comprehensive moment-based model for crystallization processes under the presence of one or multiple impurities/ additives. The interpretation of the most critical joint moments is as follows: $\mu_{0,0}$ is the total number of crystals ($\#/m^3$) and $\mu_{2,1}$ represents the crystal volume in a unit volume of suspension ($m^3/m^3$). However, although these are the only joint moments that have a physical meaning, other ones can be used to determine other key properties of the crystal population. Furthermore, the moments can be utilized to determine the mean crystal sizes (Eqs. (A.6) and (A.7)) of the total population of each characteristic length, while the mean AR of the crystals (Eq. (A.8)) can be estimated by the division of the mean sizes as illustrated below:

$$\overline{x_1} = \frac{\mu_{0,1}}{\mu_{0,0}} \tag{A.6}$$

$$\overline{x_2} = \frac{\mu_{1,0}}{\mu_{0,0}} \tag{A.7}$$

$$\overline{AR} = \frac{\overline{x_1}}{\overline{x_2}} \tag{A.8}$$

It should be noted the model is based on several assumptions mainly:

- All the new formed crystals have a nominal size $L_{x_1,n} \approx L_{x_2,n} \geq 0$. Hence, it can be considered that the initial nuclei size

is $L_n \approx 0$ (in most of the modelling studies, describing crystallization processes, the initial nucleus size is set to zero for practical purposes).
- The process operates under well-mixed conditions, so it could be assumed that the system is perfectly mixed. Hence a lumped parameter model is developed since the dependent variable does not change with spatial location (e.g. the density function, the concentration of the different chemical compounds and the moments are functions of time and not space).
- Only primary nucleation and crystal growth is considered since only these phenomena were detected experimentally. Thus, agglomeration and breakage can be neglected.
- Size-independent growth rates are assumed for the two characteristic faces since the SMOM is applied for the identification of the parameters
- Two different impurities and two different active sites are taken into account, which are located on two different crystal facets.
- There is no interaction between the active sites.
- Impurity effect on the nucleation is insignificant and hence it is considered negligible.

Equilibrium adsorption model is considered.

## References


Aamir, E., Nagy, Z.K., Rielly, C.D., Kleinert, T., Judat, B., 2009. Combined quadrature method of moments and method of characteristics approach for efficient solution of population balance models for dynamic modeling and crystal size distribution control of crystallization processes. *Ind. Eng. Chem. Res.* 48 (18), 8575–8584. doi:10.1021/ie900430t.

Alvarez, A.J., Myerson, A.S., 2010. Continuous plug flow crystallization of pharmaceutical compounds. *Cryst. Growth Des.* 10 (5), 2219–2228. doi:10.1021/cg901496s.

Benyahia, B., Lakerveld, R., Barton, P.I., 2012. A plant-wide dynamic model of a continuous pharmaceutical process. Ind. Eng. Chem. Res. 51 (47), 15393–15412. doi:10.1021/ie3006319.

Benyahia, B., 2018. Applications of a plant-wide dynamic model of an integrated continuous pharmaceutical plant: design of the recycle in the case of multiple impurities. In: Singh, R, Yuan, Z (Eds.). In: Process Systems Engineering for Pharmaceutical Manufacturing: From Product Design to Enterprise-Wide Decisions, Computer Aided Chemical Engineering, 41, pp. 141–157. doi:10.1016/B978-0-444-63963-9.00006-3, doi.org.

Benyahia, B., Latifi, M.A., Fonteix, C., Pla, F., 2013. Emulsion copolymerization of styrene and butyl acrylate in the presence of a chain transfer agent. Part 2: Parameters estimability and confidence regions. *Chem. Eng. Sci.* 90, 110–118. doi:10.1016/j.ces.2012.12.013.

Benyahia, B., Latifi, M.A., Fonteix, C., Pla, F., 2011. Modeling of a batch emulsion copolymerization reactor in the presence of a chain transfer agent: Estimability analysis, parameters identification and experimental validation. Comput. Aided Chem. Eng. 29, 126–130. doi:10.1016/B978-0-444-63234-0.50121-4.

Benyahia, B, 2009. Modélisation, expérimentation et optimisation multicritère d'un procédé de copolymérisation en émulsion en présence d'un agent de transfert de chaîne. Institut National Polytechnique de Lorraine, France.

Borsos, A., Majumder, A., Nagy, Z.K., 2016. Multi-impurity adsorption model for modeling crystal purity and shape evolution during crystallization processes in impure media. *Crys. Growth Des.* 16 (2), 555–568. doi:10.1021/acs.cgd.5b00320.

Borsos, Á, Lakatos, B.G., 2014. Investigation and simulation of crystallization of high aspect ratio crystals with fragmentation. *Chem. Eng. Res. Des.* 92 (6), 1133–1141. doi:10.1016/j.cherd.2013.08.020.

Brun, R., Kühni, M., Siegrist, H., Gujer, W., Reichert, P., 2002. Practical identifiability of ASM2d parameters—systematic selection and tuning of parameter subsets. *Water Res.* 36 (16), 4113–4127. doi:10.1016/s0043-1354(02)00104-5.

Cao, Y., Kariwala, V., Nagy, Z.K., 2012. Parameter estimation for crystallization processes using Taylor method. *IFAC Proc. Vol.* 45 (15), 880–885. doi:10.3182/20120710-4-sg-2026.00079.

Chu, Y., Hahn, J., 2011. Generalization of a parameter set selection procedure based on orthogonal projections and the d-optimality criterion. *AIChE J.* 58 (7), 2085–2096. doi:10.1002/aic.12727.

Chen, B.H., Bermingham, S., Neumann, A.H., Kramer, H.J., Asprey, S.P., 2004. On the design of optimally informative experiments for dynamic crystallization process modeling. *Ind. Eng. Chem. Res.* 43 (16), 4889–4902. doi:10.1021/ie030649n.

Czapla, F., Haida, H., Elsner, M., Lorenz, H., Seidel-Morgenstern, A., 2009. Parameterization of population balance models for polythermal auto seeded preferential crystallization of enantiomers. *Chem. Eng. Sci.* 64 (4), 753–763. doi:10.1016/j.ces.2008.05.008.

Degenring, D., Froemel, C., Dikta, G., Takors, R., 2004. Sensitivity analysis for the reduction of complex metabolism models. *J. Process Control* 14 (7), 729–745. doi:10.1016/j.jprocont.2003.12.008.

Eghtesadi, Z., McAuley, K.B., 2014. Mean square error based method for parameter ranking and selection to obtain accurate predictions at specified operating conditions. *Ind. Eng. Chem. Res.* 53 (14), 6033–6046. doi:10.1021/ie5002444.

Epstein, M.A. (1982). Nucleation, growth, and impurity effects in crystallization process engineering. New York, NY: American Institute of Chemical Engineers.

Jayasankar, B.R., Ben-Zvi, A., Huang, B., 2009. Identifiability and estimability study for a dynamic solid oxide fuel cell model. *Comput. Chem. Eng.* 33 (2), 484–492. doi:10.1016/j.compchemeng.2008.11.005.

Jolliffe, I.T., 1972. Discarding variables in a principal component analysis. I: Artificial data. *Appl. Stat.* 21 (2), 160. doi:10.2307/2346488.

Kendall, M., Keith, J., Stuart, A., 1977. The Advanced Theory of Statistics. Griffin and Company, London.

Kravaris, C., Hahn, J., Chu, Y., 2013. Advances and selected recent developments in state and parameter estimation. *Comput. Chem. Eng.* 51, 111–123. doi:10.1016/j.compchemeng.2012.06.001.

Kubota, N., 2001. Effect of impurities on the growth kinetics of crystals. *Cryst. Res. Technol.* 36 (8-10), 749–769 doi:10.1002/1521-4079(200110)36:8/10<749:aid-crat749>3.0.co;2-#.

Kumar, J., Peglow, M., Warnecke, G., Heinrich, S., 2008. The cell average technique for solving multi-dimensional aggregation population balance equations. *Comput. Chem. Eng.* 32 (8), 1810–1830. doi:10.1016/j.compchemeng.2007.10.001.

Lakerveld, R., Benyahia, B., Braatz, R.D., Barton, P.I., 2013. Model-based design of a plant-wide control strategy for a continuous pharmaceutical plant. AIChE J. 59, 3671–3685. doi:10.1002/aic.14107.

Lund, B.F., Foss, B.A., 2008. Parameter ranking by orthogonalization—applied to nonlinear mechanistic models. Automatica 44 (1), 278–281. doi:10.1016/j.automatica.2007.04.006.

Majumder, A., Kariwala, V., Ansumali, S., Rajendran, A., 2012. Lattice Boltzmann method for multi-dimensional population balance models in crystallization. *Chem. Eng. Sci.* 70, 121–134. doi:10.1016/j.ces.2011.04.041.

Mascia, S., Heider, P.L., Zhang, H., Lakerveld, R., Benyahia, B., Barton, P.I., Trout, B.L., 2013. End-to-end continuous manufacturing of pharmaceuticals: integrated synthesis, purification, and final dosage formation. *Angew. Chem.* 125 (47), 12585–12589. doi:10.1002/ange.201305429.

McLean, K.A.P., McAuley, K.B., 2011. Mathematical modelling of chemical processes-obtaining the best model predictions and parameter estimates using identifiability and estimability procedures. *Can. J. Chem. Eng.* 90 (2), 351–366. doi:10.1002/cjce.20660.

Nagy, Z.K., 2009. Model based robust batch-to-batch control of particle size and shape in pharmaceutical crystallisation. *IFAC Proc. Vol.* 42 (11), 195–200. doi:10.3182/20090712-4-tr-2008.00029.

Nagy, Z.K., Fevotte, G., Kramer, H., Simon, L.L., 2013. Recent advances in the monitoring, modelling and control of crystallization systems. *Chem. Eng. Res. Des.* 91 (10), 1903–1922. doi:10.1016/j.cherd.2013.07.018.

Onyemelukwe, I., Benyahia, B., Reis, N.M., Nagy, Z.K., Rielly, C.D., 2018. The heat transfer characteristics of a mesoscale continuous oscillatory flow crystalliser with smooth periodic constrictions. Int. J. Heat Mass Transf. 123, 1109–1119. doi:10.1016/j.ijheatmasstransfer.2018.03.015.

Quaiser, T., Mönnigmann, M., 2009. Systematic identifiability testing for unambiguous mechanistic modeling – application to JAK-STAT, MAP kinase, and NF-$\kappa$B signaling pathway models. *BMC Syst. Biol.* 3 (1), 50. doi:10.1186/1752-0509-3-50.

Peña, R., Nagy, Z.K., 2015. Process intensification through continuous spherical crystallization using a two-stage mixed suspension mixed product removal (MSMPR) system. *Cryst. Growth Des.* 15 (9), 4225–4236. doi:10.1021/acs.cgd.5b00479.

Ramin, P., Mansouri, S.S., Udugama, I.A., Benyahia, B., Gernaey, K.V., 2018. Modelling continuous pharmaceutical and bio-based processes at plant-wide level: a roadmap towards efficient decision-making. *Chem. Today* 36 (2), 26–30.

Rawlings, J.B., Miller, S.M., Witkowski, W.R., 1993. Model identification and control of solution crystallization processes: a review. *Ind. Eng. Chem. Res.* 32 (7), 1275–1296. doi:10.1021/ie00019a002.

Saltelli, A., Chan, K., Scott, M.E., Saltelli, A., Chan, K., Scott, E.M., Saltelli, S., 2008. Sensitivity Analysis. Wiley-Blackwell (an imprint of John Wiley & Sons Ltd), United Kingdom.

Saltelli, A., Ratto, M., Tarantola, S., Campolongo, F., 2005. Sensitivity analysis for chemical models. ChemInform 36 (42). doi:10.1002/chin.200542290.

Saltelli, A., Tarantola, S., Campolongo, F., Saltelli, S.T., Ratto, M., 2004. Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models. Wiley, John & Sons, United Kingdom.

Samad, N.A., Sin, G., Gernaey, K.V., Gani, R., 2013a. Introducing uncertainty analysis of nucleation and crystal growth models in process analytical technology (PAT) system design of crystallization processes. *Eur. J. Pharm. Biopharm.* 85 (3), 911–929. doi:10.1016/j.ejpb.2013.05.016.

Samad, N.A., Sin, G., Gernaey, K.V., Gani, R., 2013b. A systematic framework for design of process monitoring and control (PAT) systems for crystallization processes. *Comput. Chem. Eng.* 54, 8–23. doi:10.1016/j.compchemeng.2013.03.003.

Sato, K., Nagai, H., Hasegawa, K., Tomori, K., Kramer, H., Jansens, P., 2008. Two-dimensional population balance model with breakage of high aspect ratio crystals for batch crystallization. *Chem. Eng. Sci.* 63 (12), 3271–3278. doi:10.1016/j.ces.2008.03.013.

Schittkowski, K., 2007. Experimental design tools for ordinary and algebraic differential equations. Ind. Eng. Chem. Res. 46 (26), 9137–9147. doi:10.1021/ie0703742.

Sin, G., Meyer, A.S., Gernaey, K.V., 2010. Assessing reliability of cellulose hydrolysis models to support biofuel process design—Identifiability and uncertainty analy-

sis. *Comput. Chem. Eng.* 34 (9), 1385–1392. doi:10.1016/j.compchemeng.2010.02.012.

Sobol′, I.M., 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math. Comput. Simul* 55 (1-3), 271–280. doi:10.1016/s0378-4754(00)00270-6.

Su, Q., Nagy, Z.K., Rielly, C.D., 2015. Pharmaceutical crystallisation processes from batch to continuous operation using MSMPR stages: modelling, design, and control. *Chem. Eng. Process.* 89, 41–53. doi:10.1016/j.cep.2015.01.001.

Surisetty, K., Hoz Siegler, H.D.la, McCaffrey, W.C., Ben-Zvi, A., 2010. Model reparameterization and output prediction for a bioreactor system. *Chem. Eng. Sci.* 65 (16), 4535–4547. doi:10.1016/j.ces.2010.04.024.

Thompson, D.E., McAuley, K.B., McLellan, P.J., 2009. Parameter estimation in a simplified MWD model for HDPE produced by a Ziegler-Natta catalyst. *Macromol. React. Eng.* 3 (4), 160–177. doi:10.1002/mren.200800052.

Vajda, S., Rabitz, H., Walter, E., Lecourtier, Y., 1989. Qualitative and quantitative identifiability analysis of nonlinear chemical kinetic models. *Chem. Eng. Commun.* 83 (1), 191–219. doi:10.1080/00986448908940662.

Varma, A., Morbidelli, M., Hua, W., Wu, H., 2005. Parametric Sensitivity in Chemical Systems. Cambridge University Press, United Kingdom.

Velez-Reyes, M., Verghese, G.C., 1995. Subset selection in identification, and application to speed and parameter estimation for induction machines. In: Proceedings of the 4th IEEE conference on Control Applications, pp. 991–997. doi:10.1109/CCA.1995.555890.

Walter, E., Pronzato, L., 1997. Identification of Parametric Models from Experimental Data. Springer, Berlin.

Yao, K.Z., Shaw, B.M., Kou, B., McAuley, K.B., Bacon, D.W., 2003. Modeling ethylene/butene copolymerization with multi-site catalysts: parameter estimability and experimental design. *Polym. React. Eng.* 11 (3), 563–588. doi:10.1081/pre-120024426.

Yu, L., 2004. Applications of process analytical technology to crystallization processes. *Adv. Drug. Deliv. Rev.* 56 (3), 349–369. doi:10.1016/j.addr.2003.10.012.