


RESEARCH ARTICLE

Process Systems Engineering

Parameter estimation and estimability analysis in pharmaceutical models with uncertain inputs

Iman Moshiritabrizi¹ | Kaveh Abdi¹ | Jonathan P. McMullen² |
Brian M. Wyvratt² | Kimberley B. McAuley¹ 

¹Department of Chemical Engineering,
Queen's University, Kingston, Ontario, Canada

²Process Research & Development, Merck &
Co., Inc., Rahway, New Jersey, USA

Correspondence

Kimberley B. McAuley, Department of
Chemical Engineering, Queen's University,
Kingston, ON K7L 3N6, Canada.
Email: kim.mcauley@queensu.ca

Funding information

Merck Sharp and Dohme

Abstract

A methodology is proposed to aid parameter estimation in fundamental models of pharmaceutical processes. This methodology addresses situations with insufficient data to reliably estimate all parameters, when the estimation is complicated by uncertain independent variables. The proposed method uses an augmented sensitivity matrix to rank the combined set of parameters and uncertain inputs from most estimable to least estimable. An updated mean-squared-error criterion is then used to determine the appropriate parameters and inputs that should be estimated, based on the ranked list. A model for one step in a batch pharmaceutical production process with an uncertain initial reactant concentration is used to illustrate the method, revealing that the initial reactant concentration in each batch should be estimated along with three out of six model parameters. Non-estimable parameters are fixed at their initial values to prevent overfitting. The method will aid error-in-variables parameter estimation in many situations involving limited data.

KEYWORDS

error-in-variables-model, estimability analysis, fundamental model, parameter estimation, pharmaceutical

1 | INTRODUCTION

Mathematical models are used by pharmaceutical industries for formulation development, scale-up, control, and monitoring of production processes.¹ Models are also used because they provide useful insights and reduce the experimental effort required for process and product quality improvement. Two main categories of models are fundamental (mechanistic) models and empirical models. The current study focuses on fundamental models, which can produce more reliable predictions over a wider range of operating conditions than empirical models, especially when data are limited.^{2–4} Usually there are unknown parameters

in fundamental models that need to be estimated from experimental data. Therefore, scientists and engineers employ a variety of statistical techniques to estimate these parameters.^{5,6} A summary of the fundamental modeling studies for pharmaceutical production processes that involve real experimental data and parameter estimation is given in Table 1. Several additional studies rely on simulated pharmaceutical data to illustrate statistical methods.^{32–34} In all the studies listed in Table 1, model inputs (independent variables) were assumed to be perfectly known during parameter estimation and all of the experimental uncertainty was assigned to the model outputs (dependent variables). This assumption enabled modelers to use either least squares (LS)^{12,15,27} or weighted least

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *AIChE Journal* published by Wiley Periodicals LLC on behalf of American Institute of Chemical Engineers.

TABLE 1 Modeling studies of pharmaceutical production involving parameter estimation with real data.

References	Process modeled	Process type	Number of unknown parameters	All parameters estimated?
Kuu et al. ⁷	Primary drying of 5% mannitol and 5% povidone	Batch	2	Y
Sadikoglu and Liapis ⁸	Drying stages in bulk solution freeze-drying of pharmaceuticals in trays	Batch	30	Y
Togkalidou et al. ⁹	Cooling crystallization of a drug compound	Batch	4	Y
Hermanto et al. ¹⁰	Crystallization of L-glutamic acid polymorphs	Batch	21	N
Velardi et al. ¹¹	Freeze-drying of bovine serum albumin, sucrose, and mannitol	Batch	3	Y
Mortier et al. ¹²	Drying behavior of single pharmaceutical granules	Continuous	1	Y
Barrasso et al. ¹³	Manufacturing of microcrystalline cellulose and tablets	Continuous	24	N
Barrasso et al. ¹⁴	Twin screw granulation process	Continuous	10	Y
Selisteanu et al. ¹⁵	Monoclonal antibody production using mammalian cell culture process	Batch	23	Y
Gagnon et al. ¹⁶	Drying of pharmaceutical particles containing calcium carbonate	Batch	22	N
Garcia-Munoz et al. ¹⁷	Direct compression process for pharmaceutical tablets	Continuous	8	N
Garg et al. ¹⁸	Antisolvent crystallization for production of dexlansoprazole	Batch	15	Y
Montes et al. ¹⁹	Hoescht process for the synthesis of ibuprofen	Batch	14	Y
Wang et al. ²⁰	Chinese hamster ovary cell metabolism to produce antibody	Fed-batch	51	N
Cuthbertson et al. ²¹	Enzymatic synthesis of amoxicillin	Batch	14	Y
Lee et al. ²²	Kinetics, evaporation, and crystallization in the manufacturing of active pharmaceutical ingredients (API)	Batch, continuous	5	Y
Maloney et al. ²³	Carfilzomib drug substance intermediate manufacturing	Batch, continuous	58	N
Schenk et al. ²⁴	Multistage solid-liquid pharmaceutical process for urea compound synthesis	Fed-batch	8	N
Diab et al. ²⁵	Flow synthesis kinetics for an anticancer API called Lomustine	Continuous	5	Y
Grimard et al. ²⁶	Hot-melt extrusion process for the manufacturing of itraconazole tablets	Continuous	14	N
Pal et al. ²⁷	Spherical agglomeration processes for a drug containing benzoic acid	Semi batch	1	Y
Sen et al. ²⁸	Methylation of heteroatom-containing molecules	Batch	12	N
Szilagy et al. ²⁹	Pharmaceutical crystallization processes for Indomethacin	Batch	8	N
Diab et al. ³⁰	Kinetics, distillation, and crystallization for one step in amine production	Batch	15	N
Dos Santos et al. ³¹	Adsorption of Praziquantel enantiomers	—	5	Y

squares (WLS) estimation,^{7,9,13,14,16,20,24,25,29,31} which is applied when there are multiple dependent variables with different levels of variability. Sometimes, however, uncertainties in independent variables can be large due to measurement errors in process inputs or other difficulties in achieving the desired experimental settings. In such cases, neglecting the input uncertainties can adversely affect the accuracy of parameter estimates and associated model predictions.^{35,36}

Although input uncertainties were not considered during parameter estimation in the studies shown in Table 1, they have been considered in other types of fundamental chemical process models.³⁷ The main approach used to account for input uncertainty in chemical engineering literature is called error-in-variables-model (EVM) parameter estimation. WLS and EVM are similar, except that the objective function for EVM

parameter estimation is more complicated because true values of the uncertain inputs are estimated along with the model parameters.^{37,38} Abdi and McAuley³⁷ recently reviewed the EVM literature and showed that EVM has been used in a diverse array of models for polymerization reactions,³⁹ vapor-liquid equilibrium,^{40–42} gas-solid adsorption,⁴³ liquid-liquid diffusion,⁴⁴ and ion-exchange equilibrium.⁴⁵ In all of these EVM studies, the authors assumed that the available data contained sufficient information to estimate all the unknown parameters.

Notice, however, that in 11 of the 25 studies shown in Table 1 (see right-most column), the authors determined that only a subset of the model parameters should be estimated from the available data, either to avoid numerical problems or parameter overfitting. In 6 of the 11 studies where only a subset of the parameters was estimated, the

authors decided which parameters should be fixed at nominal values and which should be estimated based on their scientific or engineering judgment.^{10,13,23,26,29,30} The authors for the remaining five studies, used formal statistical methods for subset selection with sensitivity-based methods being most popular.^{16,17,20,24,28} For example, Garcia-Munoz et al. and Sen et al. used a popular orthogonalization-based algorithm to rank their model parameters from most estimable to least estimable.^{17,28,46} This parameter ranking method has been used along with a mean-squared-error (MSE)-based criterion^{46,47} for parameter subset selection in a wide variety of chemical process models where input uncertainties are neglected (e.g., Refs. 33,48–51). To our knowledge, statistical methods for parameter subset selection have not been developed that account for uncertain inputs until now.

The main objective of the current study is to extend the orthogonalization-based algorithm and associated MSE criterion so they can be applied to pharmaceutical models with input uncertainties. We believe that it is important for modelers to select appropriate parameters for estimation when datasets are too limited to reliably estimate all of the model parameters, especially when unknown inputs are considered as additional parameters for estimation. Our goal is to help developers of chemical and pharmaceutical products, especially drug substances, to tune their mechanistic models and use them to obtain preliminary information about proposed production processes based on a few initial experiments. The proposed methods will be applied for parameter estimability analysis and estimation in a pharmaceutical case study involving uncertain addition of a reactant to a batch reactor. The associated experimental data were obtained by Merck & Co., Inc. (also known as Merck Sharp & Dohme [MSD] outside of the United States and Canada) during experiments aimed at understanding the key steps in the production of an intermediate in the manufacturing of a drug substance to treat the human immunodeficiency virus (HIV).

The remainder of this article is organized as follows. In Section 2, background information on WLS and EVM parameter estimation, and parameter estimability analysis are presented. In Section 3, we propose extensions to the estimability ranking algorithm and MSE-based criterion so that they can be used when inputs are uncertain. In Section 4, the pharmaceutical case study is used to illustrate the proposed methods. We investigate the number of parameters that can be estimated from the available data, and we show that improved model predictions are obtained when input uncertainties are considered.

2 | BACKGROUND INFORMATION

2.1 | Parameter estimation using weighted least squares and error-in-variables models

Consider the following multiresponse non-linear model in which the independent variables are assumed to be perfectly known:

$$Y = g(x, \theta) + \varepsilon_Y \quad (1)$$

In Equation (1), $Y \in \mathbb{R}^{N_Y}$ is a measurement vector and $g(x, \theta) \in \mathbb{R}^{N_Y}$ is the solution of nonlinear equations, which may be

differential equations that are solved numerically. If the model predicts the values of N_d different response variables at several different times during multiple runs, the corresponding measured values are stacked together in the Y vector. For example, if all of the N_d dependent variables are measured at N_s sampling times per run in N_r runs, then the dimensionality of Y is $N_Y = N_d N_s N_r$. In Equation (1), $x \in \mathbb{R}^{N_x \times N_r}$ is a matrix of experimental settings for the N_x independent variables, $\theta \in \mathbb{R}^{N_\theta}$ is the vector of N_θ unknown model parameters and $\varepsilon_Y \in \mathbb{R}^{N_Y}$ is a vector of random measurement noise.

Assuming that the model equations are correct, and that the measurement noise is independent, identically, and normally distributed, the following WLS objective function can be used to estimate the parameters^{6,52,53}:

$$J_{WLS} = \sum_{i=1}^{N_r} (y_{m,i} - g(x_i, \theta))^T \Sigma_Y^{-1} (y_{m,i} - g(x_i, \theta)) \quad (2)$$

where $y_{m,i} \in \mathbb{R}^{N_{Yi}}$ is a vector of N_{Yi} measured data values for the i th run, $g(x_i, \theta) \in \mathbb{R}^{N_{Yi}}$ is the corresponding model predictions, $x_i \in \mathbb{R}^{N_x}$ is a vector of experimental settings for the i th run, and $\Sigma_Y \in \mathbb{R}^{N_{Yi} \times N_{Yi}}$ is a diagonal covariance matrix associated with the independent measurement noise in the responses.

Unfortunately, the assumption of perfectly known inputs is not always applicable. Considering uncertainties in some of the independent variables, the model becomes:

$$Y = g(x, u, \theta) + \varepsilon_Y \quad (3)$$

$$U = u + \varepsilon_U \quad (4)$$

where $U \in \mathbb{R}^{N_U \times N_U}$ is a matrix of measurements of uncertain inputs, $u \in \mathbb{R}^{N_U \times N_U}$ is a matrix containing unknown true values of these inputs, and $\varepsilon_U \in \mathbb{R}^{N_U \times N_U}$ contains the random input uncertainties. In the parameter-estimation literature, the model in Equations (3) and (4) is referred to as an “error-in-variables” model because it accounts for random errors in both types of variables (i.e., independent variables and dependent variables).⁵⁴

Using Equations (3) and (4) along with a maximum likelihood approach results in the following EVM objective function³⁸:

$$J_{EVM} = \sum_{i=1}^{N_r} (y_{m,i} - g(x_i, u_i, \theta))^T \Sigma_Y^{-1} (y_{m,i} - g(x_i, u_i, \theta)) + (u_{m,i} - u_i)^T \Sigma_U^{-1} (u_{m,i} - u_i) \quad (5)$$

where $u_i \in \mathbb{R}^{N_U}$ are the true values of uncertain inputs used in the i th run, $u_{m,i} \in \mathbb{R}^{N_U}$ is a vector of measured values for the corresponding uncertain inputs, and $\Sigma_U \in \mathbb{R}^{N_U \times N_U}$ is a diagonal covariance matrix for the random error in the uncertain inputs.

Notice that the EVM objective function contains an additional term compared to the WLS objective function in Equation (2) to account for the unknown inputs u_i that are estimated along with the unknown model parameters.^{35,36,38}

TABLE 2 Orthogonalization algorithm for parameter estimability ranking when inputs are perfectly known.^{55,56}

- 1 Compute the magnitude (i.e., the Euclidean norm) of each column in the Z matrix. Select the column with the largest magnitude as the most estimable parameter. Set $k = 1$.
- 2 Put the k selected columns from Z that correspond to parameters that have been ranked in the matrix X_k .
- 3 Use X_k to predict columns in Z using ordinary least squares:
 $\hat{Z}_k = X_k (X_k^T X_k)^{-1} X_k^T Z$ (2.1)
 and calculate the residual matrix:
 $R_k = Z - \hat{Z}_k$ (2.2)
- 4 Calculate the magnitude of each column in R_k . The $(k + 1)$ th-most estimable parameter corresponds to the column in R_k with the largest magnitude
- 5 Increase the iteration counter k by one and repeat Steps 2–4, until all parameters are ranked or until it is impossible to perform the least-squares calculation in Step 3 due to matrix singularity.

TABLE 3 MSE-based algorithm to determine optimal number of parameters to estimate.^{46,47}

- 1 Rank model parameters from most estimable to least estimable using the estimability algorithm in Table 2.
- 2 Use WLS regression to estimate the first parameter from the list, with all others fixed at initial guesses. Next, estimate the top two parameters, followed by the top three parameters and so on, until all the ranked parameters have been estimated. Denote the value of the objective function with the top k parameters estimated and the remaining $N_\theta - k$ parameters held fixed as J_k . Weighting factors used in parameter estimation should be consistent with measurement uncertainties s_{y_j} used for scaling during parameter ranking.
- 3 Compute the critical ratio:
 $r_{C,k} = (J_k - J_{N_\theta}) / (N_\theta - k)$ (3.1)
 for $k = 1, 2, \dots, N_\theta - 1$.
- 4 For each value of k , compute the corrected critical ratio:
 $r_{CC,k} = \frac{(N_\theta - k)}{N_\theta} (r_{CKub,k} - 1)$ (3.2)
 where
 $r_{CKub,k} = \max \left(r_{C,k} - 1, \frac{2}{N_\theta - k + 2} r_{CC,k} \right)$ (3.3)
- 5 Select the value of k corresponding to the lowest value of $r_{CC,k}$ as the appropriate number of parameters to estimate

2.2 | Orthogonalization based method for parameter subset selection

Parameter subset selection methods are used to select appropriate parameters for estimation when there is insufficient information in the available data to reliably estimate all the model parameters.^{46,55,56} In the current article, we extend the parameter subset selection methods shown in Tables 2 and 3 which were developed assuming that all the model inputs are perfectly known. Using the algorithm in Table 2, parameters with strong and independent influence on one or more model predictions appear near the top of the list. Other less-important parameters appear near the bottom of the list.^{55–57} The

algorithm in Table 3 is used to determine the appropriate number of parameters to estimate from the ranked list.

The orthogonalization method in Table 2 relies on a sensitivity matrix $S \in \mathbb{R}^{N_y \times N_\theta}$ containing partial derivatives of the model predictions with respect to the model parameters:

$$S = \begin{bmatrix} \frac{\partial g_{11}}{\partial \theta_1} \Big|_{x_1} & \dots & \frac{\partial g_{11}}{\partial \theta_p} \Big|_{x_1} & \dots & \frac{\partial g_{11}}{\partial \theta_{N_\theta}} \Big|_{x_1} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial g_{jl}}{\partial \theta_1} \Big|_{x_i} & \dots & \frac{\partial g_{jl}}{\partial \theta_p} \Big|_{x_i} & \dots & \frac{\partial g_{jl}}{\partial \theta_{N_\theta}} \Big|_{x_i} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial g_{N_d N_s}}{\partial \theta_1} \Big|_{x_{N_r}} & \dots & \frac{\partial g_{N_d N_s}}{\partial \theta_p} \Big|_{x_{N_r}} & \dots & \frac{\partial g_{N_d N_s}}{\partial \theta_{N_\theta}} \Big|_{x_{N_r}} \end{bmatrix} \quad \begin{matrix} j = 1, \dots, N_d \\ l = 1, \dots, N_s \\ i = 1, \dots, N_r \\ p = 1, \dots, N_\theta \end{matrix} \quad (6)$$

In this sensitivity matrix, each column is associated with a particular parameter and each row is associated with prediction of a particular measured value that will be used for parameter estimation. The elements of S are often approximated using finite differences:

$$\frac{\partial g_{jl}}{\partial \theta_p} \Big|_{x_i} \cong \frac{g_{jl}(x_i, \theta_p + \Delta \theta_p) - g_{jl}(x_i, \theta_p)}{\Delta \theta_p} \quad (7)$$

The indices j and l correspond to the j th response obtained at the l th sampling time and x_i indicates the experimental settings for the i th run. Note that if fewer than N_s samples are available in some of the runs for any of the measured responses, the corresponding row(s) are deleted from the sensitivity matrix.⁵⁵ The algorithm in Table 2 uses a matrix $Z \in \mathbb{R}^{N_y \times N_\theta}$ whose elements are scaled. For example, $\frac{\partial g_{jl}}{\partial \theta_p} \Big|_{x_i}$ is scaled to become $\frac{\partial g_{jl}}{\partial \theta_p} \Big|_{x_i} \frac{s_{\theta_p}}{s_{y_j}}$, which makes the elements dimensionless and permits fair comparison of the sensitivities. The scaling factor s_{y_j} accounts for uncertainty in measurements of the j th response, and the scaling factor s_{θ_p} accounts for uncertainty in the initial guess of parameter θ_p . For example, if we assume that the prior uncertainty in θ_p corresponds to a normal distribution, with six standard deviations between the lower bound lb_{θ_p} and upper bound ub_{θ_p} used for parameter estimation, we could select the scaling factor:

$$s_{\theta_p} = \frac{ub_{\theta_p} - lb_{\theta_p}}{6} \quad (8)$$

The MSE-based algorithm in Table 3 was developed to determine an appropriate number of parameters to estimate to obtain a good fit to the data while preventing overfitting.³³ As more parameters are estimated from the ranked list, the bias in the model predictions decreases while the variance increases. The algorithm selects the number of parameters that minimizes the MSE, which is the sum of the squared bias and the variance.^{47,58–60}

In Equation (3.1), J_k and J_{N_θ} are the WLS objective function values when k and all N_θ parameters are estimated, respectively. In Equation (3.2), the subscript Kub in $r_{CKub,k}$ refers to an improved estimator developed by Kubokawa et al. that Wu et al. used in their calculations.^{47,61} The parameter estimability ranking and MSE-

based parameter subset selection methods in Tables 2 and 3 have been used to aid WLS parameter estimation for models of a wide variety of chemical processes (e.g., 47,50,62–64) where all of the independent variables are assumed to be perfectly known. In the next section, these methodologies are extended for use in EVM parameter estimation.

2.3 | Proposed methodology

The main idea of the proposed parameters subset selection methodology is to construct an augmented scaled sensitivity matrix \mathbf{Z}_{EVM} that has additional columns (compared to \mathbf{Z}) to account for the unknown inputs that may be estimated along with the model parameters and additional rows to account for uncertain measurements of these unknown inputs:

Notice that the top left corner of \mathbf{Z}_{EVM} is the same as \mathbf{Z} for the corresponding WLS parameter estimation problem. The matrix \mathbf{Z}_{EVM} contains additional columns, one for each unknown input value that is considered as an extra parameter for estimation. \mathbf{Z}_{EVM} also has additional rows, one for each measurement of an unknown input. For example, if each run involves N_U unknown inputs that will require estimation and each unknown input is measured (or estimated with some uncertainty) once per run then \mathbf{Z}_{EVM} will contain $N_r N_U$ more columns and $N_r N_U$ more rows compared to \mathbf{Z} , as shown in Equation (9). In Equation (9), the scaling factors s_{u11} to $s_{uN_r N_U}$ reflect uncertainties in the corresponding input values. For example:

$$s_{u11} = \frac{ub_{u11} - lb_{u11}}{6} \quad (10)$$

The scaled sensitivity matrix in Equation (9) can be simplified as follows:

$$\mathbf{Z}_{\text{EVM}} = \begin{bmatrix} \frac{\partial g_{11}}{\partial \theta_1} \bigg|_{x_1, u_1} \frac{s_{\theta_1}}{s_{y_1}} & \dots & \frac{\partial g_{11}}{\partial \theta_{N_\theta}} \bigg|_{x_1, u_1} \frac{s_{\theta_{N_\theta}}}{s_{y_1}} & \frac{\partial g_{11}}{\partial u_{11}} \bigg|_{x_1, \theta} \frac{s_{u_{11}}}{s_{y_1}} & \dots & \frac{\partial g_{11}}{\partial u_{N_r N_U}} \bigg|_{x_1, \theta} \frac{s_{u_{N_r N_U}}}{s_{y_1}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_{j1}}{\partial \theta_1} \bigg|_{x_j, u_j} \frac{s_{\theta_1}}{s_{y_j}} & \dots & \frac{\partial g_{j1}}{\partial \theta_{N_\theta}} \bigg|_{x_j, u_j} \frac{s_{\theta_{N_\theta}}}{s_{y_j}} & \frac{\partial g_{j1}}{\partial u_{11}} \bigg|_{x_j, \theta} \frac{s_{u_{11}}}{s_{y_j}} & \dots & \frac{\partial g_{j1}}{\partial u_{N_r N_U}} \bigg|_{x_j, \theta} \frac{s_{u_{N_r N_U}}}{s_{y_j}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_{N_d N_s}}{\partial \theta_1} \bigg|_{x_{N_r}, u_{N_r}} \frac{s_{\theta_1}}{s_{y_{N_d}}} & \dots & \frac{\partial g_{N_d N_s}}{\partial \theta_{N_\theta}} \bigg|_{x_{N_r}, u_{N_r}} \frac{s_{\theta_{N_\theta}}}{s_{y_{N_d}}} & \frac{\partial g_{N_d N_s}}{\partial u_{11}} \bigg|_{x_{N_r}, \theta} \frac{s_{u_{11}}}{s_{y_{N_d}}} & \dots & \frac{\partial g_{N_d N_s}}{\partial u_{N_r N_U}} \bigg|_{x_{N_r}, \theta} \frac{s_{u_{N_r N_U}}}{s_{y_{N_d}}} \\ \frac{\partial u_{11}}{\partial \theta_1} \frac{s_{\theta_1}}{s_{u_{11}}} & \dots & \frac{\partial u_{11}}{\partial \theta_{N_\theta}} \frac{s_{\theta_{N_\theta}}}{s_{u_{11}}} & \frac{\partial u_{11}}{\partial u_{11}} \frac{s_{u_{11}}}{s_{u_{11}}} & \dots & \frac{\partial u_{11}}{\partial u_{N_r N_U}} \frac{s_{u_{N_r N_U}}}{s_{u_{11}}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial u_{N_r N_U}}{\partial \theta_1} \frac{s_{\theta_1}}{s_{u_{N_r N_U}}} & \dots & \frac{\partial u_{N_r N_U}}{\partial \theta_{N_\theta}} \frac{s_{\theta_{N_\theta}}}{s_{u_{N_r N_U}}} & \frac{\partial u_{N_r N_U}}{\partial u_{11}} \frac{s_{u_{11}}}{s_{u_{N_r N_U}}} & \dots & \frac{\partial u_{N_r N_U}}{\partial u_{N_r N_U}} \frac{s_{u_{N_r N_U}}}{s_{u_{N_r N_U}}} \end{bmatrix} \quad (9)$$

$$\mathbf{Z}_{\text{EVM}} = \begin{bmatrix} \frac{\partial g_{11}}{\partial \theta_1} \bigg|_{x_1, u_1} \frac{s_{\theta_1}}{s_{y_1}} & \dots & \frac{\partial g_{11}}{\partial \theta_{N_\theta}} \bigg|_{x_1, u_1} \frac{s_{\theta_{N_\theta}}}{s_{y_1}} & \frac{\partial g_{11}}{\partial u_{11}} \bigg|_{x_1, \theta} \frac{s_{u_{11}}}{s_{y_1}} & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_{j1}}{\partial \theta_1} \bigg|_{x_j, u_j} \frac{s_{\theta_1}}{s_{y_j}} & \dots & \frac{\partial g_{j1}}{\partial \theta_{N_\theta}} \bigg|_{x_j, u_j} \frac{s_{\theta_{N_\theta}}}{s_{y_j}} & \frac{\partial g_{j1}}{\partial u_{11}} \bigg|_{x_j, \theta} \frac{s_{u_{11}}}{s_{y_j}} & \dots & \frac{\partial g_{j1}}{\partial u_{N_r N_U}} \bigg|_{x_j, \theta} \frac{s_{u_{N_r N_U}}}{s_{y_j}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_{N_d N_s}}{\partial \theta_1} \bigg|_{x_{N_r}, u_{N_r}} \frac{s_{\theta_1}}{s_{y_{N_d}}} & \dots & \frac{\partial g_{N_d N_s}}{\partial \theta_{N_\theta}} \bigg|_{x_{N_r}, u_{N_r}} \frac{s_{\theta_{N_\theta}}}{s_{y_{N_d}}} & 0 & \dots & \frac{\partial g_{N_d N_s}}{\partial u_{N_r N_U}} \bigg|_{x_{N_r}, \theta} \frac{s_{u_{N_r N_U}}}{s_{y_{N_d}}} \\ 0 & \dots & 0 & 1 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 1 \end{bmatrix} \quad (11)$$

TABLE 4 Orthogonalization algorithm for estimability ranking when some inputs are uncertain.

1	Compute the magnitude (i.e., the Euclidean norm) of each column in the Z_{EVM} matrix. Select the column with the largest magnitude as the most estimable decision variable. Set $k = 1$.
2	Put the k selected columns from Z_{EVM} corresponding to decision variables that have been ranked in the matrix $X_{EVM,k}$.
3	Use $X_{EVM,k}$ to predict columns in Z_{EVM} using ordinary least squares: $\hat{Z}_{EVM,k} = X_{EVM,k} (X_{EVM,k}^T X_{EVM,k})^{-1} X_{EVM,k}^T Z_{EVM}$ and calculate the residual matrix: $R_{EVM,k} = Z_{EVM} - \hat{Z}_{EVM,k}$ (4.1) (4.2)
4	Calculate the magnitude of each column in $R_{EVM,k}$. The $(k + 1)$ th-most estimable decision variable corresponds to the column in $R_{EVM,k}$ with the largest magnitude
5	Increase the iteration counter k by one and repeat Steps 2–4, until all decision variables are ranked or until it is impossible to perform the least-squares calculation in Step 3 due to matrix singularity.

because uncertain inputs are independent of the model parameters (e.g., $\frac{\partial u_{11}}{\partial \theta_1} = 0$) and the predicted responses for one run are independent of uncertain inputs for other runs. Notice that the bottom right-hand corner of Z_{EVM} is the identity matrix because the true values of the uncertain inputs are independent of each other (e.g., $\frac{\partial u_{11}}{\partial u_{12}} = 1$ and $\frac{\partial u_{11}}{\partial u_{12}} = 0$).

Table 4 presents orthogonalization-based algorithms to rank decision variables when (some) model inputs are uncertain. Decision variables refer to the unknown parameters and unknown inputs that are considered as extra parameters for estimations. The algorithm in Table 4 is almost the same as Table 2; however, it uses the matrix Z_{EVM} instead of Z to rank decision variables.

Table 5 shows MSE-based algorithms to determine optimal number of decision variables when (some) model inputs are uncertain. This algorithm is very similar to Table 3; however, it uses EVM objective function instead of WLS and introduces N_D which is the number of decision variables and N_{Um} which is the number of measured values for unknown inputs. For example, if N_U unknown inputs are measured once per run, $N_D = N_\theta + N_U N_r$ and $N_{Um} = N_U N_r$.

Next section investigates the application of this proposed method in the parameter estimability and estimation in a pharmaceutical production model.

3 | CASE STUDY: EVM PARAMETER SELECTION AND ESTIMATION IN A PHARMACEUTICAL PRODUCTION MODEL

3.1 | Reactants and reaction scheme

Table 6 shows the reaction scheme for the case study. In Table 6, SM is the starting material and TMA is trimethyl amine, a gaseous material that is bubbled into the liquid solution in the reactor to start the first reaction. Because it is difficult to reproducibly add the desired initial

TABLE 5 MSE-based algorithm to determine optimal number of decision variables to estimate when (some) inputs are uncertain.

1	Rank the decision variables for EVM parameter estimation from most estimable to least estimable using the EVM estimability algorithm in Table 4.
2	Use EVM regression to estimate the first decision variables from the list, with all others fixed at initial guesses. Next, estimate the top two decision variables, followed by the top three and so on, until all decision variables have been estimated. Denote the value of the objective function with the top k decision variables estimated as $J_{EVM,k}$. Weighting factors used in EVM parameter estimation should be consistent with measurement uncertainties and input uncertainties used for scaling during parameter ranking.
3	Compute the critical ratio: $r_{C,k} = (J_{EVM,k} - J_{EVM,N_D}) / (N_D - k)$ for $k = 1, 2, \dots, N_D - 1$. (5.1)
4	For each value of k , compute the corrected critical ratio: $r_{CC,k} = \frac{(N_D - k)}{N_r + N_{Um}} (r_{CKub,k} - 1)$ where $r_{CKub,k} = \max \left(r_{C,k} - 1, \frac{2}{N_D - k + 2} r_{CC,k} \right)$ (5.2) (5.3)
5	Select the value of k corresponding to the lowest value of $r_{CC,k}$ as the appropriate number of decision variables to estimate.

TABLE 6 Reaction scheme for the case study.

Main Reaction
$SM + TMA \xrightleftharpoons[k_r]{k_f} QS1Cl$
Side Reaction
$QS1Cl \xrightarrow{k_{fs}} CIDMI + MeCl$

quantity of TMA to the reactor, the initial concentration C_0^{TMA} in each run is treated as an uncertain input. The main reaction in Table 6 is a desired reaction. The side reaction is an undesirable reaction, which consumes the quaternary chloride salt ($QS1Cl$) and produces chlorodemethylated impurity ($CIDMI$) and methyl chloride ($MeCl$). In the future, our goal is to develop a model for a more complex reaction system (see Figure 1) wherein an additional reagent allows the reaction to proceed from $QS1Cl$ to the desired product, P . This product is quenched with ammonium hydroxide to provide the isolated intermediate, 2-fluoroadenine-9-THP (not shown in Figure 1), before subsequent glycosylation to form crude Islatravir.⁶⁵ The current study involves only the reactions inside the blue dashed box, which were performed to better understand the kinetics of the side reaction before building a full kinetic model for the overall reaction scheme in Figure 1.

3.2 | Experimental methods and available data

Two experimental runs were conducted by MSD in a batch reactor, one at 33°C and one at 23°C. Both experiments were conducted in a

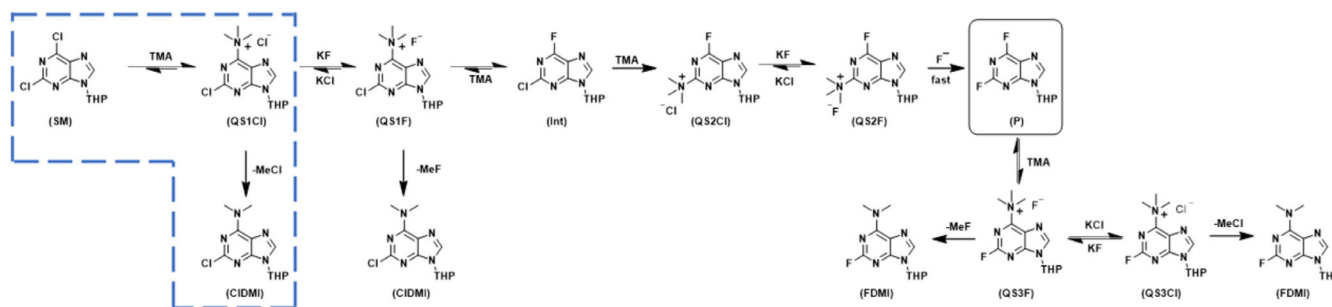


FIGURE 1 Reaction scheme used to produce 9-THP-2,6-difluoropurine. The blue box indicates the portion of the scheme considered in the current experimental and modeling study.

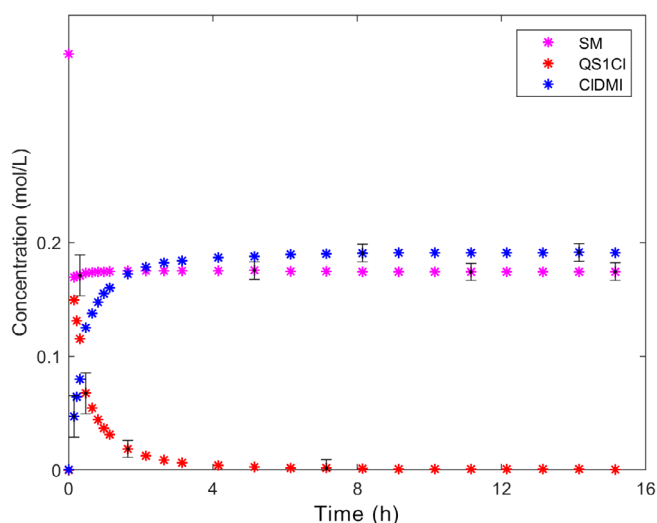


FIGURE 2 Experimental data for SM, CIDMI, and QS1Cl for Run 1 conducted at $T = 33^\circ\text{C}$.

MettlerToledo EasyMax 102 Advanced Synthesis Station equipped with an MettlerToledo Easy Control Box (ECB), using a 100 mL Hastelloy C (HC), two-piece pressure reactor equipped with a HC-22 4-blade pitched impeller, an HC thermowell, and a digital pressure gauge. A MettlerToledo EasySampler 1210 was used for automated reaction sampling to obtain the reaction profile data. All Ultra Performance Liquid Chromatography (UPLC) analyses were performed using an Agilent 1290 Infinity II equipped with a Waters Cortecs T3 column (4.6 mm \times 150 mm; 2.7 μm particle size) and a diode array detector. An Alicat mass flow controller staged on a hot plate set to 30°C was used to charge the gaseous trimethylamine reagent to the reactor to initiate reaction. The charge rate was kept constant at 30 sccm to prevent condensation in the line; the charge duration was set to ensure that the total quantity charged was near the target value.

Before each experiment, the necessary plumbing connections were established to add the pressure gauge and EasySampler probe to the reactor head, and the reactor was pressure-tested. To accomplish this, the reactor was pressurized with nitrogen to 103.421–137.895 kPaG and the pressure was monitored over ~ 10 min to evaluate leak rates; reactors are considered acceptable if the pressure dropped by < 2 kPaG over that time-period. To start an experiment,

TABLE 7 Dynamic model equations for the batch reactor.

Equation	Initial condition
$\frac{dC_{SM}}{dt} = -k_f C_{SM} C_{TMA}$ (7.1)	$C_0^{SM} = 0.366 \text{ mol/L}$
$\frac{dC_{TMA}}{dt} = -k_f C_{SM} C_{TMA} + k_r C_{QS1Cl}$ (7.2)	C_0^{TMA}
$\frac{dC_{QS1Cl}}{dt} = k_f C_{SM} C_{TMA} - k_r C_{QS1Cl} - k_{fs} C_{QS1Cl}$ (7.3)	$C_0^{QS1Cl} = 0 \text{ mol/L}$
$\frac{dC_{CIDMI}}{dt} = k_{fs} C_{QS1Cl}$ (7.4)	$C_0^{CIDMI} = 0 \text{ mol/L}$
$\frac{dC_{MeCl}}{dt} = k_{fs} C_{QS1Cl}$ (7.5)	$C_0^{MeCl} = 0 \text{ mol/L}$
$k_f = k_{f,ref} \exp\left(\frac{-E_f}{R} \left(\frac{1}{T} - \frac{1}{T_{ref}}\right)\right)$ (7.6)	—
$k_{fs} = k_{fs,ref} \exp\left(\frac{-E_{fs}}{R} \left(\frac{1}{T} - \frac{1}{T_{ref}}\right)\right)$ (7.7)	—
$k_r = \frac{k_f}{K}$ (7.8)	—
$K = K_{ref} \exp\left(\frac{-\Delta H}{R} \left(\frac{1}{T} - \frac{1}{T_{ref}}\right)\right)$ (7.9)	—

the starting material (2,6-dichloropurine-9-THP, 6 g) and dimethylformamide (DMF, anhydrous, 60 mL) solvent are added to the 100 mL HC reactor inside an inert-atmosphere glovebox due to the moisture sensitivity of the reaction; note that residual water present during the reaction would lead to water-capture impurities, which are ignored in the reaction mechanism in Figure 1. The reactor body is covered with parafilm, removed from the glovebox, and seated in the EasyMax. Very rapidly, the parafilm is removed and the reactor head connected to the body. The reactor is pressure-purged (0 to 103.421–137.895 kPaG, 5 times) to remove any air from the vessel. To start the reaction, the reactor agitation is initiated (600 rpm) and the batch equilibrated at the target reaction temperature. Subsequently, trimethylamine (TMA) is charged subsurface at a constant volumetric rate (30 sccm) for ~ 9 min to achieve the desired charge quantity (0.649 g, 0.5 equivalents) and start the reaction. The assumed TMA charge quantity is based on the volumetric flowrate set point and the time duration of the charge, with the start and end of the addition corresponding to the manual opening and closing of a valve, respectively. Of all experimental factors, the TMA charge has the most uncertainty. The TMA is highly soluble in DMF and readily dissolves in the reaction mixture so that rapid dissolution of TMA in the liquid is not a problem. The greatest uncertainty stems from the combined variability in the feed flowrate of TMA and in the TMA charge time.

TABLE 8 List of unknown parameters and inputs to estimate.

Parameter	Initial guess	Lower bound	Upper bound
$k_{f \text{ ref}}$ (L/mol h)	100,000	10,000	1,000,000
E_f (kJ/mol)	100	0	200
K_{ref} (L/mol)	100	10	10,000
ΔH (kJ/mol)	0	−200	200
$k_{fs \text{ ref}}$ (1/h)	1	0.1	10
E_{fs} (kJ/mol)	100	50	200
$C_0^{TMA,1}$ (mol/L)	$0.5C_0^{SM}$	$0.4C_0^{SM}$	$0.6C_0^{SM}$
$C_0^{TMA,2}$ (mol/L)	$0.5C_0^{SM}$	$0.4C_0^{SM}$	$0.6C_0^{SM}$

The reaction is exothermic, so that a temperature increase of up to 0.2°C is observed during the *TMA* charge, after which the temperature quickly equilibrates to the target temperature. The pressure in the reactor headspace was monitored during each run and changed by <12 kPag. As a result, temperature and pressure changes during each batch are negligible compared to other sources of uncertainty. Following the completion of the charge, the reaction is aged for 15 hours. Diluted samples are collected via the EasySampler at specified times for offline UPLC analysis.

Figure 2 shows the experimental data for the experimental run at $T = 33^\circ\text{C}$ (Run 1). Error bars, shown on only a few of the data points in Figure 2 to avoid clutter, were calculated from earlier replicate experiments involving some additional reagents (see in Figure 1) on the same reactor system. Notice that larger error bars appear on measurements made during the first 0.5 h because these replicate experiments revealed larger run-to-run variability at short reaction times.

3.3 | Model equations and unknown parameters

If all reactions in Table 6 are assumed to be elementary and the solution density is constant, mass balances on the species shown in Table 6 give the ordinary differential equations (ODEs (7.1) to (7.5)) in Table 7 where C_{SM} , C_{TMA} , C_{QS1Cl} , C_{CIDMI} , and C_{MeCl} are concentrations of *SM*, *TMA*, *QS1Cl*, *CIDMI*, and *MeCl*, respectively. As indicated in Table 7, the same known initial condition for *SM*, (i.e., $C_0^{SM} = 0.366 \text{ mol/L}$) is used in both experiments that are being modeled. No initial condition for the *TMA* concentration is provided in Table 7 because C_0^{TMA} is uncertain. The other initial concentrations are zero because the corresponding species are not present in the reactor at time zero. Algebraic Equations (7.6) to (7.9) in Table 7 are used to account for the influence of temperature on the reaction rates. In Equation (7.6), $k_{f \text{ ref}}$ is the value of the forward rate constant for the main reaction at $T_{\text{ref}} = 23^\circ\text{C} = 296.15 \text{ K}$, R is the ideal gas constant, and E_f is the corresponding activation energy. Similarly, $k_{fs \text{ ref}}$ is the rate constant for the side reaction at 296.15 K and E_{fs} is the activation energy for the side reaction. In Equations (7.8) and (7.9), K is the equilibrium constant for the main reaction and ΔH is the reaction enthalpy. Note that the model in Table 7 relies on the reasonable

TABLE 9 List of ranked unknown parameters and inputs.

Parameters and inputs	EVM rank	WLS rank
$k_{fs \text{ ref}}$	1	1
K_{ref}	2	2
$C_0^{TMA,1}$	3	—
$C_0^{TMA,2}$	4	—
E_{fs}	5	3
ΔH	6	4
$k_{f \text{ ref}}$	7	5
E_f	8	6

assumption that the highly soluble gaseous *TMA* fed to the reactor rapidly dissolves in the liquid phase. As a result, mass transfer and solubility parameters do not need to be considered in the model.

Table 8 provides initial parameter guesses, which are required to solve model equations, along with lower and upper bounds. These bounds are used to ensure that the resulting estimates are physically realistic. Initial guesses in Table 8 are based on preliminary simulations and experience from earlier Merck modeling studies on a similar system. Notice that K_{ref} and ΔH are specified as model parameters requiring estimation, rather than the reverse rate constant $k_{r \text{ ref}}$ and corresponding activation energy E_r . Our reason for selecting this formulation is to reduce the amount of correlation among the model parameters.

The last two rows in Table 8 are associated with $C_0^{TMA,1}$ and $C_0^{TMA,2}$, which are the initial concentrations of *TMA* in Run 1 (conducted at 33°C) and Run 2 (conducted at 23°C), respectively. As explained in Sections 3.1 and 3.2, due to difficulties in charging the gaseous *TMA* reproducibly, these values are treated as uncertain inputs. As described in Section 2, these uncertain inputs are ranked along with the model parameters and may or may not be selected for estimation.

4 | RESULTS AND DISCUSSION

In the current case study, both EVM and WLS methods were used for parameter ranking and estimation. The algorithm in Table 4 was used to rank the unknown parameters and inputs from most to least estimable for EVM. Similarly, the algorithm in Table 2 was used to rank the unknown parameters from most to least estimable for WLS, assuming that $C_0^{TMA,1}$ and $C_0^{TMA,2}$ are perfectly known. Table 9 compares the ranked lists for both methods. Notice that the proposed new ranking method in Table 4 and the usual ranking method in Table 2 agree that $k_{fs \text{ ref}}$ is the most estimable parameter and that E_f is the least estimable. Using the MSE-based subset selection algorithms in Tables 3 and 5, the optimal number of parameters for estimation by EVM and WLS, respectively, were determined and the estimable parameters are shown in bold in Table 9 for both methods. Details are provided in the Supporting Information

TABLE 10 WLS and EVM objective functions.

WLS	$J_{WLS} = \sum_{i=1}^2 \sum_{j=1}^4 \left(\frac{(y_{m,j}^{SM} - C_{m,j}^{SM})^2}{8.2 \times 10^{-5}} + \frac{(y_{m,j}^{QS1Cl} - C_{m,j}^{QS1Cl})^2}{8.2 \times 10^{-5}} + \frac{(y_{m,j}^{CIDMI} - C_{m,j}^{CIDMI})^2}{8.2 \times 10^{-5}} \right) + \sum_{i=1}^2 \sum_{j=5}^{24} \left(\frac{(y_{m,j}^{SM} - C_{m,j}^{SM})^2}{1.46 \times 10^{-5}} + \frac{(y_{m,j}^{QS1Cl} - C_{m,j}^{QS1Cl})^2}{1.46 \times 10^{-5}} + \frac{(y_{m,j}^{CIDMI} - C_{m,j}^{CIDMI})^2}{1.46 \times 10^{-5}} \right)$	(10.1)
EVM	$J_{EVM} = J_{WLS} + \sum_{i=1}^2 \frac{(0.5C_0^{SM} - C_0^{TMA,i})^2}{(0.0333C_0^{SM})^2}$	(10.2)

TABLE 11 EVM and WLS estimated values for model parameters and uncertain inputs.

Estimated variable	Initial guess	EVM estimated value	WLS estimated value
$k_{fs \text{ ref}}$ (1/h)	1	0.4098	0.4116
K_{ref} (L/mol)	100	9512	9905
$C_0^{TMA,1}$ (mol/L)	$0.5C_0^{SM} = 0.1831$	0.1905	—
$C_0^{TMA,2}$ (mol/L)	$0.5C_0^{SM} = 0.1831$	0.1848	—
E_{fs} (kJ/mol)	100	109.1	111.9
ΔH (kJ/mol)	0	—	—
$k_{f \text{ ref}}$ (L/mol h)	100,000	—	—
E_f (kJ/mol)	100	—	—

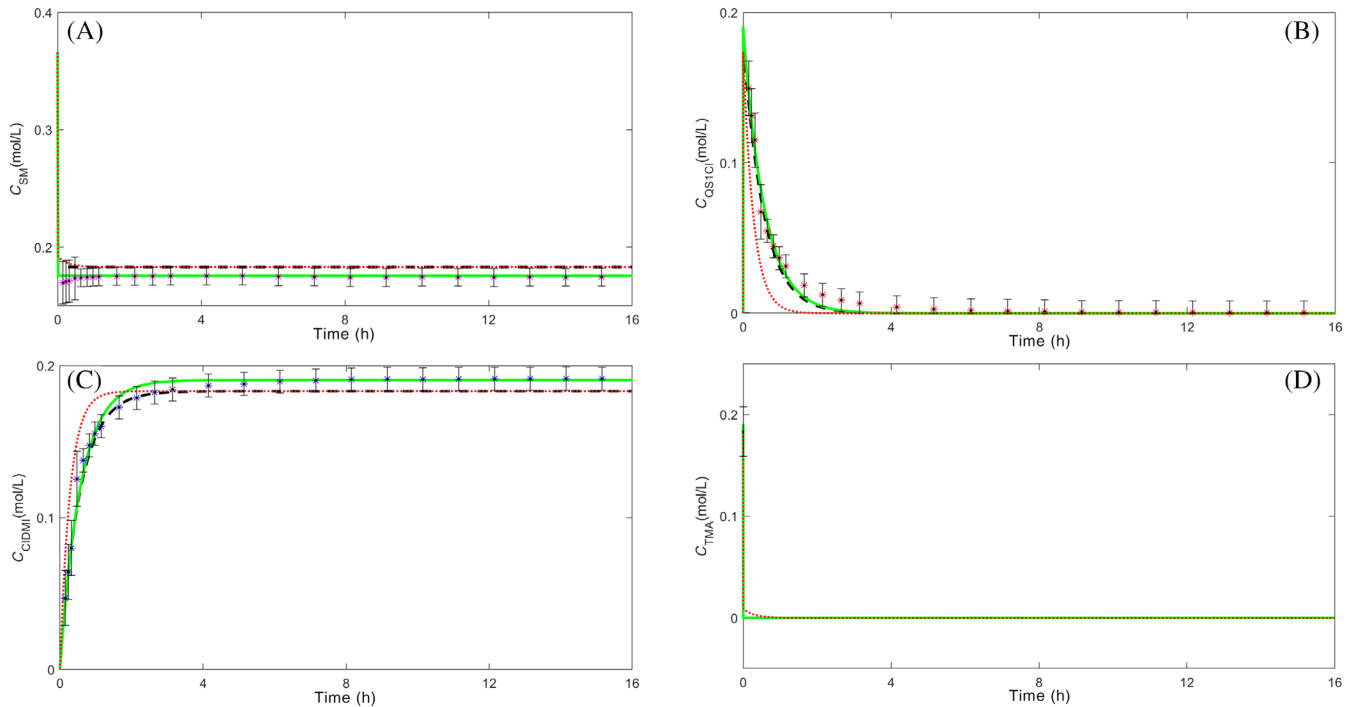


FIGURE 3 Comparison of model predictions using initial parameter guesses, EVM parameter estimates, and WLS parameter estimates ---- with experimental data for SM, QS1Cl, and CIDMI from batch experiment conducted at $T = 33^\circ\text{C}$.

Data S1. These results indicate that the 144 available data points are only sufficient for estimation of $k_{fs \text{ ref}}$, K_{ref} , $C_0^{TMA,1}$, $C_0^{TMA,2}$, and E_{fs} without overfitting using EVM. Similarly, three ($k_{fs \text{ ref}}$, K_{ref} , and E_{fs}) out of six parameters can be estimated without overfitting using WLS. Both of the parameter estimation methods result in ΔH , $k_{f \text{ ref}}$, and E_f remaining fixed at their initial guesses to prevent overfitting. Notice that $C_0^{TMA,1}$ and $C_0^{TMA,2}$ were both selected for estimation by EVM. It makes sense that $k_{f \text{ ref}}$ and E_f were not selected for estimation because the main forward reaction is very fast compared to the

reverse reaction and the side reaction. As a result, any very large value of k_f will lead to similar predictions of the available data. As such, the influences of $k_{f \text{ ref}}$ and E_f on the model predictions are small when their values are set near the initial guesses shown in Table 8. The parameter ΔH was not selected for estimation because the available data contain very little information about the influence of temperature on the equilibrium constant K . Further details about the parameter ranking and subset selection results are provided in the Supporting Information Data S1.

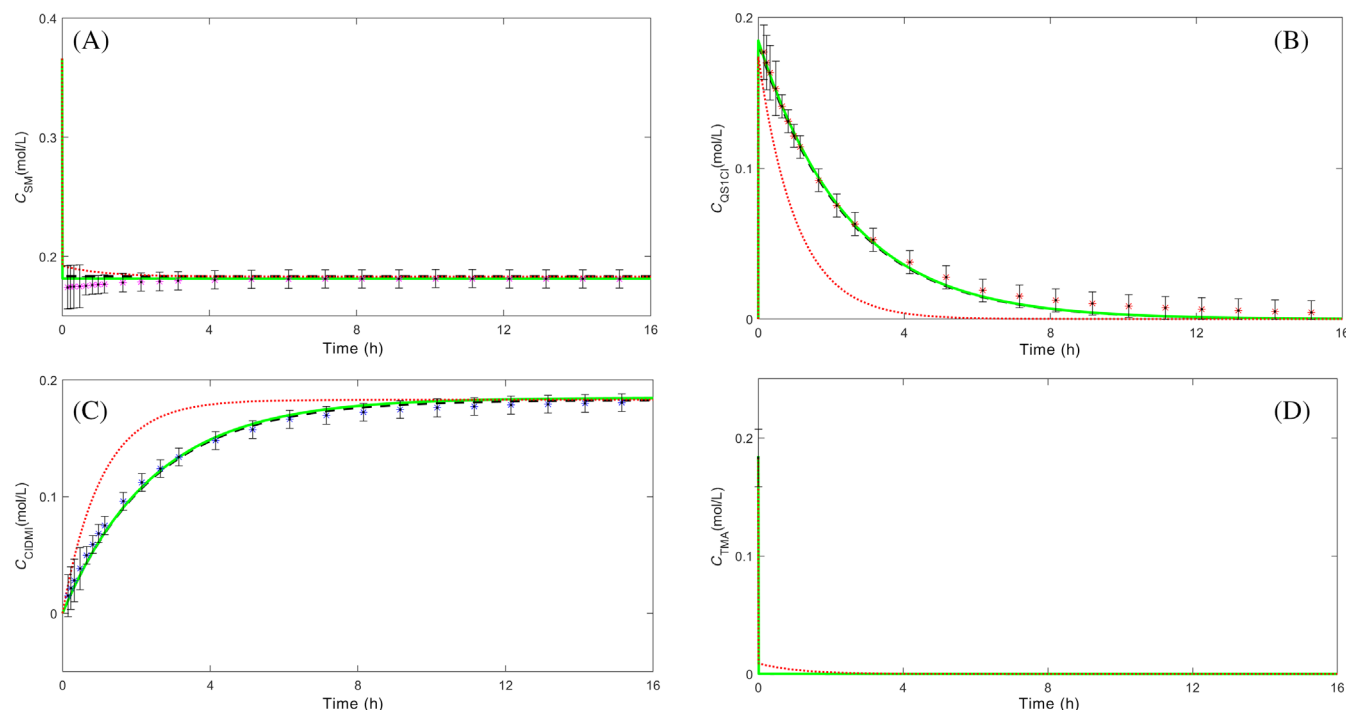


FIGURE 4 Comparison of model predictions using initial parameter guesses, EVM parameter estimates, and WLS parameter estimates ---- with experimental data for SM, QS1CI, and CIDMI from batch experiment conducted at $T = 23^{\circ}\text{C}$.

Table 10 shows the objective functions used for EVM and WLS parameter estimation where $y_{m,il}^j$ is the l th measured concentration in the i th experimental run for the j th species and C_{ij}^j is the corresponding model prediction. In J_{WLS} , the terms corresponding to values of l from 1 to 4 correspond to measurements made during the first 0.5 h of each experimental run when results are less reproducible. Larger weighting factors in the denominators are used for these terms compared to those used for terms with l ranging from 5 to 24. As shown in Equation (10.2), J_{EVM} is similar to J_{WLS} , with additional terms corresponding to the uncertain inputs. To minimize J_{WLS} and J_{EVM} , the trust region reflective algorithm in the *lsqnonlin* solver in MATLAB® (The Mathworks, Natick, MA) was used with a StepTolerance of 10^{-5} , FunctionTolerance of 10^{-7} , and OptimalityTolerance of 10^{-5} . These tolerances ensured accurate convergence of the parameter estimation. Table 11 provides the EVM estimation results and compares them to the WLS estimates.

Figures 3 and 4 show the model predictions for the experiments conducted at $T = 33^{\circ}\text{C}$ and $T = 23^{\circ}\text{C}$, respectively. As expected, predictions obtained using parameter estimates from EVM and WLS are better than the predictions obtained using the initial parameter values. Zoomed-in versions of Figures 3A,D and 4A,D are provided in the Supporting Information Data S1 to better show the details of the fit. As shown in Figure 3A,C, there is a noticeable offset between the model predictions obtained using the WLS parameter estimates and the corresponding SM and CIDMI concentration data, especially at long reaction times. This offset disappears when model predictions are made using the EVM parameter estimates, which account for uncertainty in $C_0^{\text{TMA},1}$. Notice that the EVM estimate of 0.1904 mol/L

for $C_0^{\text{TMA},1}$ is higher than the target value of 0.1831 mol/L, suggesting that more TMA than the target value was charged to the reactor at the start of Run 1, which is why more SM and CIDMI were consumed and generated, respectively, than are predicted using the WLS approach. The results for Run 2 in Figure 4 reveal that, although EVM provides somewhat better predictions of SM and QS1CI (see Figure 4A,B), both EVM and WLS methods provide good predictions for this run. This result makes sense, because value of $C_0^{\text{TMA},1} = 0.1905$ estimated using EVM is quite close to the target value of 0.183 mol/L. In summary, this case study shows that employing the proposed methodology leads to effective EVM parameter estimation results, even when some of the model parameters are not estimable from the available data. It also confirms that there are benefits to using EVM parameter estimation instead of WLS when some of the model inputs are uncertain. These results depend on the prior experience of the modeler when selecting initial guesses and bounds for the parameters and uncertain inputs. A modeler working earlier in the process development cycle, might have greater uncertainty in the initial parameter guesses, which would influence the scaling factors used in the ranking algorithm in Table 4 and might influence the parameters selected for estimation. Nevertheless, the proposed method would be useful for obtaining reasonable parameter values while preventing overfitting.

5 | CONCLUSION

New methods are proposed to aid parameter estimation in fundamental models of pharmaceutical processes when some of the

independent variables contain important uncertainties. These methods prevent parameter overfitting during EVM parameter estimation when there is not enough information in the available data to reliably estimate all the uncertain inputs and parameters. The proposed methods are extensions to previously developed techniques used to rank model parameters from most estimable to least estimable and to select an appropriate subset of parameters for estimation in models where the independent variables are perfectly known.^{46,47,55,56} The proposed methodologies rely on an augmented sensitivity matrix, which treats uncertain independent variables as both additional parameters requiring estimation and additional measured variables used for model fitting. The augmented scaled sensitivity matrix can be used in straightforward manner to simultaneously rank the parameters and uncertain inputs from the most estimable to least estimable. An extended MSE-based subset selection method is then used to determine how many parameters and inputs from the ranked list should be estimated to achieve reliable model predictions.

A pharmaceutical batch production case study is used to demonstrate the proposed methodology. This case study involves an uncertain initial concentration of trimethylamine (TMA) in two experimental runs, due to variability in the amount of TMA charged to the reactor. The proposed ranking method determined that the initial concentrations $C_0^{TMA,1}$ and $C_0^{TMA,2}$ are ranked 3rd and 4th on the combined list of parameters and inputs. The proposed MSE-based subset selection method determined that $C_0^{TMA,1}$ and $C_0^{TMA,2}$ should be estimated along with three model parameters (i.e., $k_{fs,ref}$, K_{ref} and E_{fs}). The remaining three parameters (i.e., ΔH , $k_{f,ref}$ and E_f) were not selected for estimation and were held constant at their initial guesses. Keeping these parameters at their initial guesses is consistent with assuming that the main forward reaction is very fast and is independent of temperature and that the equilibrium for the main reaction is independent of temperature. In future, additional data may make it possible to estimate these three parameters and release the corresponding simplifying assumptions.

The resulting fit to the data, obtained using EVM parameter estimates, is excellent. A comparison with WLS parameter estimation results, obtained assuming that $C_0^{TMA,1}$ and $C_0^{TMA,2}$ were perfectly known and at their target values, reveal that the EVM fit to the data is much better than the WLS fit. For example, there is noticeable offset between the model predictions obtained using the WLS parameter estimates and the corresponding SM and CIDMI concentration data, especially at long reaction times. This offset is not present in the fit to the data obtained using the proposed EVM methodology.

The proposed parameter ranking and subset selection methodology should be useful in a wide range of pharmaceutical and chemical process models in which some independent variables are uncertain and there is insufficient data to estimate all the unknown parameters and inputs. In future, we will use the proposed methodology to develop a dynamic model for a more complex pharmaceutical production process, shown in Figure 1, which involves additional reagents and reactions. In addition, we will use the proposed method as an important step when designing new experiments aimed at providing improved parameter estimates. To our knowledge, Model-Based

Design of Experiments (MBDoe) methods that account for input uncertainties have not been developed. This will be an important future topic for research.

AUTHOR CONTRIBUTIONS

Iman Moshiritabrizi: Investigation (lead); methodology (equal); software (lead); writing – original draft (lead). **Kaveh Abdi:** Methodology (equal); writing – review and editing (supporting). **Jonathan P. McMullen:** Data curation (equal); funding acquisition (equal); investigation (equal); resources (equal); writing – review and editing (supporting). **Brian M. Wyvrat:** Data curation (equal); investigation (equal); resources (equal); writing – original draft (supporting); writing – review and editing (supporting). **Kimberley B. McAuley:** Conceptualization (lead); funding acquisition (equal); investigation (equal); methodology (equal); project administration (lead); supervision (lead); writing – review and editing (equal).

DATA AVAILABILITY STATEMENT

All data shown in Figures 2, 3 and 4, which are used for parameter estimation, are tabulated in Tables S1 and S2 of the supplementary information. Error bars in Figures 2, 3 and 4 correspond to two standard deviations, where the standard deviations are pooled estimates computed from eight replicate experiments involving SM, TMA, KF and DMF solvent. Error bars are shown on all of the data points in Figures 3 and 4 but are omitted from some of the data points in Figure 2 to avoid clutter. Larger error bars appear on measurements during the first 0.5 h because replicate experiments revealed larger run-to-run variability at short reaction times.

ORCID

Kimberley B. McAuley  <https://orcid.org/0000-0002-5201-0310>

REFERENCES

- Chatterjee S, Moore CM, Nasr MM. An overview of the role of mathematical models in implementation of quality by design paradigm for drug development and manufacture. In: Reklaitis GV, Seymour C, Garcia-Munoz S, eds. *Comprehensive Quality by Design for Pharmaceutical Product Development and Manufacture*. Wiley; 2017:1.
- Scherrman JM, Bourne DWA, eds. *Mathematical Modeling of Pharmacokinetic Data*. Technomic Publishing Co., Inc; 1995.
- Clegg LE, Mac Gabhann F. Molecular mechanism matters: benefits of mechanistic computational models for drug development. *Pharmacol Res*. 2015;99:149-154.
- Destro F, Barolo M. A review on the modernization of pharmaceutical development and manufacturing-trends, perspectives, and the role of mathematical modeling. *Int J Pharm*. 2022;2022:121715.
- Maria G. A review of algorithms and trends in kinetic model identification for chemical and biochemical systems. *Chem Biochem Eng Q*. 2004;18(3):195-222.
- Beck JV, Arnold KJ. *Parameter Estimation in Engineering and Science*. James Beck; 1977.
- Kuu W-Y, McShane J, Wong J. Determination of mass transfer coefficients during freeze drying using modeling and parameter estimation techniques. *Int J Pharm*. 1995;124(2):241-252.
- Sadikoglu H, Liapis A. Mathematical modelling of the primary and secondary drying stages of bulk solution freeze-drying in trays: parameter estimation and model discrimination by comparison of theoretical results with experimental data. *Dry Technol*. 1997;15(3-4):791-810.

9. Togkalidou T, Tung H-H, Sun Y, Andrews AT, Braatz RD. Parameter estimation and optimization of a loosely bound aggregating pharmaceutical crystallization using in situ infrared and laser backscattering measurements. *Ind Eng Chem Res.* 2004;43(19):6168-6181.
10. Hermanto MW, Kee NC, Tan RB, Chiu MS, Braatz RD. Robust Bayesian estimation of kinetics for the polymorphic transformation of l-glutamic acid crystals. *AIChE J.* 2008;54(12):3248-3259.
11. Velardi SA, Rasetto V, Barresi AA. Dynamic parameters estimation method: advanced manometric temperature measurement approach for freeze-drying monitoring of pharmaceutical solutions. *Ind Eng Chem Res.* 2008;47(21):8445-8457.
12. Mortier STF, De Beer T, Gernaey KV, et al. Mechanistic modelling of the drying behaviour of single pharmaceutical granules. *Eur J Pharm Biopharm.* 2012;80(3):682-689.
13. Barrasso D, Oka S, Muliadi A, Litster JD, Wassgren C, Ramachandran R. Population balance model validation and prediction of CQAs for continuous milling processes: toward QbDin pharmaceutical drug product manufacturing. *J Pharm Innov.* 2013;8(3):147-162.
14. Barrasso D, El Hagrasy A, Litster JD, Ramachandran R. Multi-dimensional population balance model development and validation for a twin screw granulation process. *Powder Technol.* 2015;270:612-621.
15. Selișteanu D, Șendrescu D, Georgeanu V, Roman M. Mammalian cell culture process for monoclonal antibody production: nonlinear modelling and parameter estimation. *Biomed Res Int.* 2015;2015:1-16.
16. Gagnon F, Desbiens A, Poulin É, Lapointe-Garant P-P, Simard J-S. Nonlinear model predictive control of a batch fluidized bed dryer for pharmaceutical particles. *Control Eng Pract.* 2017;64:88-101.
17. García-Muñoz S, Butterbaugh A, Leavesley I, Manley LF, Slade D, Birmingham S. A flowsheet model for the development of a continuous process for pharmaceutical tablets: an industrial perspective. *AIChE J.* 2018;64(2):511-525.
18. Garg M, Roy M, Chokshi P, Rathore AS. Process development in the QbD paradigm: mechanistic modeling of antisolvent crystallization for production of pharmaceuticals. *Cryst Growth Des.* 2018;18(6):3352-3359.
19. Montes FC, Gernaey K, Sin GR. Dynamic plantwide modeling, uncertainty, and sensitivity analysis of a pharmaceutical upstream synthesis: ibuprofen case study. *Ind Eng Chem Res.* 2018;57(30):10026-10037.
20. Wang Z, Sheikh H, Lee K, Georgakis C. Sequential parameter estimation for mammalian cell model based on in silico design of experiments. *Processes.* 2018;6(8):100.
21. Cuthbertson AB, Rodman AD, Diab S, Gerogiorgis DI. Dynamic modelling and optimisation of the batch enzymatic synthesis of amoxicillin. *Processes.* 2019;7(6):318.
22. Lee BW, Peterson JJ, Yin K, Stockdale GS, Liu YC, O'Brien A. System model development and computer experiments for continuous API manufacturing. *Chem Eng Res Des.* 2020;156:495-506.
23. Maloney AJ, İçten EI, Capellades G, et al. A virtual plant for integrated continuous manufacturing of a carfilzomib drug substance intermediate, part 3: manganese-catalyzed asymmetric epoxidation, crystallization, and filtration. *Org Process Res Dev.* 2020;24(10):1891-1908.
24. Schenk C, Biegler LT, Han L, Mustakis J. Kinetic parameter estimation from spectroscopic data for a multi-stage solid-liquid pharmaceutical process. *Org Process Res Dev.* 2020;25(3):373-383.
25. Diab S, Raiyat M, Gerogiorgis DI. Flow synthesis kinetics for lomustine, an anti-cancer active pharmaceutical ingredient. *React Chem Eng.* 2021;6(10):1819-1828.
26. Grimard J, Dewasme L, Wouwer AV. Dynamic model reduction and predictive control of hot-melt extrusion applied to drug manufacturing. *IEEE Trans Control Syst Technol.* 2020;29(6):2366-2378.
27. Pal K, Szilagyi B, Burcham CL, Jarmer DJ, Nagy ZK. Iterative model-based experimental design for spherical agglomeration processes. *AIChE J.* 2021;67(5):e17178.
28. Sen M, Arguelles AJ, Stamatis SD, García-Muñoz S, Kolis S. An optimization-based model discrimination framework for selecting an appropriate reaction kinetic model structure during early phase pharmaceutical process development. *React Chem Eng.* 2021;6(11):2092-2103.
29. Szilagyi B, Wu W-L, Eren A, et al. Cross-pharma collaboration for the development of a simulation tool for the model-based digital design of pharmaceutical crystallization processes (CrySiV). *Cryst Growth Des.* 2021;21(11):6448-6464.
30. Diab S, Christodoulou C, Taylor G, Rushworth P. Mathematical modeling and optimization to inform impurity control in an industrial active pharmaceutical ingredient manufacturing process. *Org Process Res Dev.* 2022;26(10):2864-2881.
31. Dos Santos RC, Cunha FC, Marcellos CFC, et al. Adsorption of praziquantel enantiomers on chiral cellulose tris 3-chloro, 4-methylphenylcarbamate by frontal analysis: Fisherian and Bayesian parameter estimation and inference. *J Chromatogr A.* 2022;1676:463200.
32. Borg N, Westerberg K, Andersson N, von Lieres E, Nilsson B. Effects of uncertainties in experimental conditions on the estimation of adsorption model parameters in preparative chromatography. *Comput Chem Eng.* 2013;55:148-157.
33. Shahmohammadi A, McAuley KB. Using prior parameter knowledge in model-based design of experiments for pharmaceutical production. *AIChE J.* 2020;66(11):e17021.
34. Besenhard MO, Chaudhury A, Vetter T, Ramachandran R, Khinast JG. Evaluation of parameter estimation methods for crystallization processes modeled via population balance equations. *Chem Eng Res Des.* 2015;94:275-289.
35. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement Error in Nonlinear Models: A Modern Perspective.* Chapman and Hall/CRC; 2006.
36. Abdi K, Celse B, McAuley KB. Propagating input uncertainties into parameter uncertainties an model prediction uncertainties: a review. *Revision in Progress for Industrial and Engineering Chemistry Research* 2022.
37. Abdi K, McAuley KB. Estimation of output measurement variances for EVM parameter estimation. *AIChE J.* 2022;68(8):e17735.
38. Britt H, Luecke R. The estimation of parameters in nonlinear, implicit models. *Dent Tech.* 1973;15(2):233-247.
39. Keeler SE, Reilly PM. The error-in-variables model applied to parameter estimation when the error covariance matrix is unknown. *Can J Chem Eng.* 1991;69(1):27-34.
40. Sutton TL, Macgregor JF. The analysis and design of binary vapour-liquid equilibrium experiments. Part I: parameter estimation and consistency tests. *Can J Chem Eng.* 1977;55(5):602-608.
41. Duever T, Keeler S, Reilly P, Vera J, Williams P. An application of the error-in-variables model—parameter estimation from Van Ness-type vapour-liquid equilibrium experiments. *Chem Eng Sci.* 1987;42(3):403-412.
42. Kim IW, Lieberman MJ, Edgar TF. Robust error-in-variables estimation using nonlinear programming techniques. *AIChE J.* 1990;36(7):985-993.
43. High M, Danner R. Treatment of gas-solid adsorption data by the error-in-variables method. *AIChE J.* 1986;32(7):1138-1145.
44. Bardow A, Marquardt W. Identification of diffusive transport by means of an incremental approach. *Comput Chem Eng.* 2004;28(5):585-595.
45. Vamos RJ, Haas CN. Reduction of ion-exchange equilibria data using an error in variables approach. *AIChE J.* 1994;40(3):556-569.
46. McLean KA, McAuley KB. Mathematical modelling of chemical processes—obtaining the best model predictions and parameter

- estimates using identifiability and estimability procedures. *Can J Chem Eng.* 2012;90(2):351-366.
47. Wu S, McLean KA, Harris TJ, McAuley KB. Selection of optimal parameter set using estimability analysis and MSE-based model-selection criterion. *Int J Adv Mechatron Syst.* 2011;3(3):188-197.
 48. Karimi H, Cowperthwaite EV, Olayiwola B, Farag H, McAuley KB. Modelling of heat transfer and pyrolysis reactions in an industrial ethylene cracking furnace. *Can J Chem Eng.* 2018;96(1):33-48.
 49. Aiello JP, Jiang Y, Moebus JA, Greenhalgh BR, McAuley KB. Predicting polyethylene molecular weight and composition distributions obtained using a multi-site catalyst in a gas-phase lab-scale reactor. *Macromol Theory Simul.* 2021;30(3):2000079.
 50. Feng H-H, Chen X, Gu X-P, et al. Modeling of the molecular weight distribution and short chain branching distribution of linear low-density polyethylene from a pilot scale gas phase polymerization process. *Chem Eng Sci.* 2022;2022:117952.
 51. Bae J, Jeong DH, Lee JM. Ranking-based parameter subset selection for nonlinear dynamics with stochastic disturbances under limited data. *Ind Eng Chem Res.* 2020;59(50):21854-21868.
 52. Johnson ML, Faunt LM. [1] parameter estimation by least-squares methods. *Methods Enzymol.* 1992;210:1-37.
 53. Montgomery DC, Runger GC, Hubele NF. *Engineering Statistics.* John Wiley & Sons; 2009.
 54. Madansky A. The fitting of straight lines when both variables are subject to error. *J Am Stat Assoc.* 1959;54(285):173-205.
 55. Yao KZ, Shaw BM, Kou B, McAuley KB, Bacon D. Modeling ethylene/butene copolymerization with multi-site catalysts: parameter estimability and experimental design. *Polym React Eng.* 2003;11(3):563-588.
 56. Thompson DE, McAuley KB, McLellan PJ. Parameter estimation in a simplified MWD model for HDPE produced by a Ziegler-Natta catalyst. *Macromol React Eng.* 2009;3(4):160-177.
 57. Shaw BM. *Statistical Issues in Kinetic Modelling of Gas-Phase Ethylene Copolymerization.* Queen's University; 1999.
 58. Hocking RR. A biometrics invited paper. The analysis and selection of variables in linear regression. *Biometrics.* 1976;32:1-49.
 59. Rao P. Some notes on misspecification in multiple regressions. *Am Stat.* 1971;25(5):37-39.
 60. Wu S, Harris T, McAuley K. The use of simplified or misspecified models: linear case. *Can J Chem Eng.* 2007;85(4):386-398.
 61. Kubokawa T, Robert CP, Saleh AKME. Estimation of noncentrality parameters. *Can J Stat.* 1993;21(1):45-57.
 62. Schenk C, Short M, Rodriguez JS, et al. Introducing KIPET: a novel open-source software package for kinetic parameter estimation from experimental datasets including spectra. *Comput Chem Eng.* 2020;134:106716.
 63. Vo ADD, Shahmohammadi A, McAuley KB. Model-based design of experiments for polyether production from bio-based 1, 3-propanediol. *AIChE J.* 2021;67(11):e17394.
 64. Zhao YR, McAuley KB, Puskas JE. Parallel models for arborescent polyisobutylene synthesized in batch reactor. *AIChE J.* 2015;61(1):253-265.
 65. Hong CM, Xu Y, Chung JY, et al. Development of a commercial manufacturing route to 2-fluoroadenine, the key unnatural nucleobase of islatravir. *Org Process Res Dev.* 2020;25(3):395-404.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Moshiritabrizi I, Abdi K, McMullen JP, Wyvratt BM, McAuley KB. Parameter estimation and estimability analysis in pharmaceutical models with uncertain inputs. *AIChE J.* 2024;70(1):e18168. doi:[10.1002/aic.18168](https://doi.org/10.1002/aic.18168)