Preprint Cambridge Centre for Computational Chemical Engineering ISSN 1473 – 4273

Outlier analysis for a silicon nanoparticle population balance model

Sebastian Mosbach [1,2], William J. Menz [1], and Markus Kraft [1,2,3]

released: 04 December 2015

[1] Department of Chemical Engineering
and Biotechnology
University of Cambridge
New Museums Site
Pembroke Street
Cambridge CB2 3RA
UK

[2] Cambridge Centre for Carbon Reduction
in Chemical Technologies (CARES C4T)
#05-05 CREATE Tower
1 CREATE Way
Singapore 138602

[3] School of Chemical
and Biomedical Engineering
Nanyang Technological University
62 Nanyang Drive
Singapore 637459
Email: mk306@cam.ac.uk

Preprint No. 162

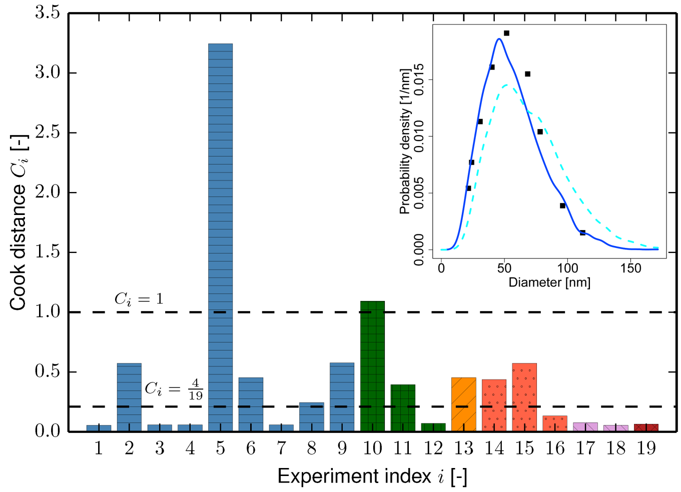
Keywords: Silicon, nanoparticles, population balance, regression influence diagnostics

Edited by
Computational Modelling Group
Department of Chemical Engineering and Biotechnology
University of Cambridge
New Museums Site
Pembroke Street
Cambridge CB2 3RA
United Kingdom

Fax: + 44 (0)1223 334796
E-Mail: c4e@cam.ac.uk
World Wide Web: http://como.cheng.cam.ac.uk/

Abstract

We assess the impact of individual experimental observations on a multivariate population balance model for the formation of silicon nanoparticles from the thermal decomposition of silane by means of basic regression influence diagnostics. The nanoparticle model includes morphological and compositional details which allow representation of primary particles within aggregates, and of coagulation, surface growth, and sintering processes. Predicted particle size distributions are optimised against 19 experiments across ranges of initial temperature, pressure, residence time, and initial silane mass fraction. The influence of each experimental observation on the model parameter estimates is then quantified using the Cook distance and DFBETA measures. Seven model parameters are included in the analysis, with five Arrhenius pre-exponential factors in the gas-phase kinetic rate expressions, and two kinetic rate constants in the population balance model. The analysis highlights certain experimental conditions and kinetic parameters which warrant closer inspection due to large influence, thus providing clues as to which aspects of the model require improvement. We find the insights provided can be useful for future model development and planning of experiments.



Highlights:

- Silicon nanoparticle synthesis is modelled using a detailed population balance model.
- An omission-based regression influence analysis is carried out.
- The impact of individual experimental observations on parameter estimates is quantified.
- Outliers are identified, suggesting areas for model improvement.

Contents

1 Introduction **3**

2 Background **4**

2.1 Population balance model for silicon nanoparticle formation 4

2.1.1 Gas phase 4

2.1.2 Particulate phase 5

2.2 Omission-based regression influence diagnostics 6

2.2.1 Parameter estimation 6

2.2.2 Influence measures 6

2.2.3 Outlier detection 8

3 Experimental data **9**

4 Results and discussion **10**

5 Conclusions **15**

References **16**

1 Introduction

Gas-phase synthesis in hot-wall reactors and shock-tubes is a common way in which silicon nanoparticles are manufactured. Typically, these processes begin with silane (SiH_4) as a precursor, which is transformed into the eventual nanoparticle product at high temperatures. A variety of models have been proposed to describe this transformation [26]. These models usually contain unknown or low-confidence (kinetic) parameters with large uncertainties associated to them. Systematic parameter estimation techniques can then be employed to arrive at better values for these quantities, based on available experimental data. One of the most elementary parameter estimation methods is least-squares optimisation, *i.e.* minimising the distance between experimental observations and model prediction as measured by a sum-of-squares objective function. The result of such an optimisation is a set of values, called ('best') estimates, for the selected model parameters. Not all experimental data points may equally inform the optimal value of the parameters, though – different parameters may be determined to a varying extent by different observations. In order to assess which experiments are the most relevant in the optimisation, one can conduct what may be called an omission-based regression influence analysis [30]: Firstly, optimise the model against the full data set, and then repeat the optimisation with one of the data points removed, for each of the data points. Based on the difference between the parameter estimates of the full optimisation and the optimisations with an omitted data point, it is then possible to quantify the influence of individual observations on the model overall or on individual parameters. Several such measures have been proposed [10, 37], the most widely-used one being Cook's distance [8], and applied to detect influential data points, high-leverage points, and statistical outliers [6, 12].

An alternative approach to quantifying influence of experimental observations is uncertainty propagation [43], part of which is concerned with how experimental measurement errors propagate into model parameters and responses. Some of these methods allow calculating the relative contribution of each data point (and its error bar) to the uncertainty in each of the parameters. In particular, the Data Collaboration framework [15] exploits the pairwise consistency of data set units to identify outliers.

Yet another approach, called perturbation of the optimum, has been developed for constrained optimisation [16, p. 34] and unconstrained least-squares optimisation [14], which has found application in chemical kinetics [30, 36, 44]. These methods allow calculating sensitivities of parameter estimates with respect to any other quantity in the objective function (or constraints), including in particular experimental data.

The purpose of this paper is to conduct an omission-based outlier analysis of a selection of experimental data for silicon nanoparticles produced from a silane precursor in hot-wall flow reactors and shock tubes which are modelled using a detailed population balance model. A main aim is to identify those experimental conditions which are the most challenging for the model. We apply a technique established in the field of regression influence diagnostics to quantify the influence of individual experimental observations on kinetic parameter estimates for this purpose. We determine the influence of the measurements on estimates of some Arrhenius pre-exponential factors in the gas-phase kinetic mechanism as well as the population balance model for the particle phase. Using a threshold for the influence values, specific measurements are then highlighted for further

Table 1: *The gas-phase kinetic mechanism. Values in bold correspond to parameters chosen for the influence analysis. Units for the Arrhenius pre-exponential factors are cm, mol, and s.*

Idx.	Reaction	A	β [-]	E [kcal/mol]	Ref.
1	SiH_4 (+M)$\rightleftharpoons$$SiH_2$ + H_2 (+M)	3.12×10^{9}	1.7	54.71	[23]
	Low pressure limit:	$\mathbf{3.96\times10^{12}}$	0	45.10	[26, 35][1]
2	Si_2H_6 (+M)$\rightleftharpoons$$SiH_4$ + SiH_2 (+M)	1.81×10^{10}	1.7	50.20	[23]
	Low pressure limit:	$\mathbf{5.09\times10^{53}}$	−10.37	56.03	[23]
3	Si_2H_6 (+M)$\rightleftharpoons$$Si_2H_4$B + H_2 (+M)	9.09×10^{9}	1.8	54.20	[23]
	Low pressure limit:	$\mathbf{7.79\times10^{40}}$	−7.77	59.02	[23, 26][2]
4	Si_3H_8 (+M)$\rightleftharpoons$$SiH_2$ + Si_2H_6 (+M)	6.97×10^{12}	1.0	52.68	[23]
	Low pressure limit:	1.73×10^{69}	−15.07	60.49	[23]
5	Si_3H_8 (+M)$\rightleftharpoons$$Si_2H_4$B + SiH_4 (+M)	3.73×10^{12}	1.0	50.85	[23]
	Low pressure limit:	$\mathbf{4.36\times10^{76}}$	−17.26	59.30	[23]
6	Si_2H_4B (+M)$\rightleftharpoons$$Si_2H_4$A (+M)	2.54×10^{13}	−0.2	5.38	[23]
	Low pressure limit:	1.10×10^{33}	−5.76	9.15	[23]
7	Si_2H_4B + $H_2$$\rightleftharpoons$$SiH_4$ + SiH_2	9.41×10^{13}	0	4.09	[23]
	Reverse coefficients:	9.43×10^{10}	1.1	5.79	[23]
8	Si_2H_4B + $SiH_4$$\rightleftharpoons$$Si_2H_6$ + SiH_2	1.73×10^{14}	0.4	8.90	[23]
	Reverse coefficients:	$\mathbf{2.65\times10^{15}}$	0.1	8.47	[23]

[1] A is from [26], β and E are from [35]. [2] A is from [26], β and E are from [23].

analysis, providing further insight into the model and potential improvements, as well as suggestions for future experiments.

2 Background

We firstly describe the model, provide some background on omission-based regression influence diagnostics, and how it can be used to identify outliers.

2.1 Population balance model for silicon nanoparticle formation

We briefly summarise the main features of the model here. Full details can be found in [26–29, 39, 39, 40]. It consists of two main parts, a gas-phase model, and a particulate phase model.

2.1.1 Gas phase

The gas-phase chemical kinetic reaction mechanism used is a modified version of the one proposed by [23], and is summarised in Table 1 (more details can be found in [26]). Two isomers of Si_2H_4 are included: silene, *i.e.* H_2SiSiH_2, denoted by the suffix “A”, and silylene, *i.e.* $HSiSiH_3$, denoted by the suffix “B”.

2.1.2 Particulate phase

The particle phase is described by a detailed, high-dimensional population balance model covering aggregate morphology and chemical composition [26]. In this model, each nanoparticle is represented as a list of primary particles, together with a (triangular) matrix, called connectivity matrix, each entry of which represents the common surface area for the corresponding pair of primary particles. For each primary particle, the number of silicon and the number of hydrogen atoms are stored.

The following processes which create or transform particles, or account for interaction of the particles with the gas phase, are represented in the model:

Inception: Any two molecules of any of the three species SiH_2, Si_2H_4A, and Si_2H_4B can collide to form a new particle. The rate at which this happens is assumed to be non-zero only if the diameter of the resulting particle exceeds a temperature- and pressure-dependent critical nucleus diameter. If the latter is the case, the inception rate is proportional to the product of the concentrations of the collision partners and the transition regime coagulation kernel.

Condensation: An existing particle can grow through (barrier-free) deposition of SiH_2, Si_2H_4A, or Si_2H_4B molecules from the gas phase onto its surface. It is assumed that the collision efficiency, *i.e.* the probability of sticking, is unity. The rate is given by a free-molecular collision kernel.

Surface reaction: Apart from simply condensing, gas-phase species can also react heterogeneously on the particle surface. Specifically, silanes (SiH_4, Si_2H_6, and Si_3H_8) can be integrated into the particle, with each step releasing one, two, and three molecules of hydrogen, respectively. The rate is proportional to the particle surface area and an Arrhenius expression with non-zero activation energy. Rounding of adjacent primary particles caused by this process is also taken into account.

Hydrogen release: In order to attain a stable crystal structure, particles need to release some of the hydrogen acquired through each of the above processes. The rate of desorption is proportional to an Arrhenius expression and the coverage of hydrogen on the particle surface, which is approximated by the ratio of hydrogen to silicon atoms within the particle. It is assumed that the sintering level of adjacent primaries is unaffected by this process, *i.e.* the connectivity matrix remains unchanged.

Coagulation: Two particles can collide and stick to each other at their point of contact. The rate is given by transition regime coagulation kernel, which is the harmonic mean of the slip-flow and free-molecular kernels.

Sintering: The sintering of any pair of adjacent primary particles is modelled by an exponential decay of the excess of the joint surface area of the primaries compared to the surface area of their equivalent sphere. In other words, the corresponding entry in the connectivity matrix decreases exponentially towards the equivalent spherical area of the primary particle pair.

2.2 Omission-based regression influence diagnostics

2.2.1 Parameter estimation

Given a set of N experimental observations η_n^{exp}, with $n = 1, \ldots, N$. For example, these could be, as in this work, means or modes of the particle size distribution at given temperatures and pressures. Assuming we have a model which depends on a vector ϑ of P model parameters, we denote its response for the conditions of the n^{th} experiment by $\eta_n(\vartheta)$. For simplicity, we restrict ourselves in this work to a single response, but the generalisation of all that follows to multiple responses is straightforward.

In order to quantify agreement between experiment and model, a measure of the distance between the model response and experimental results needs to be defined. We use the ordinary least-squares objective function

$$\Phi(\vartheta) := \sum_{n=1}^{N} \left[\eta_n(\vartheta) - \eta_n^{\text{exp}}\right]^2 \tag{1}$$

for this purpose. The term ‘ordinary’ refers to the fact that the covariance matrix of the responses is the unity matrix, *i.e.* the responses are assumed to be uncorrelated and are subject to the same or very similar uncertainties, meaning all the terms in the sum are equally weighted.

The vector $\hat{\vartheta}$ of parameter values which are optimal with respect to the objective function can be obtained by minimising (1):

$$\hat{\vartheta} := \underset{\vartheta}{\operatorname{argmin}}\, \Phi(\vartheta) \tag{2}$$

The best estimate of the model responses is then defined as $\hat{\eta} := \eta(\hat{\vartheta})$.

2.2.2 Influence measures

The basic idea underlying omission-based regression influence diagnostics is to analyse the effect of deleting a single observation from the considered set of data. In the following, we use a subscript “$-i$” to denote quantities based on the data set with the i^{th} observation removed. In particular, the objective function (1) becomes

$$\Phi_{-i}(\vartheta) := \sum_{n=1,\ldots,i-1,i+1,\ldots,N} \left[\eta_n(\vartheta) - \eta_n^{\text{exp}}\right]^2, \tag{3}$$

with the corresponding best parameter estimate

$$\hat{\vartheta}_{-i} := \underset{\vartheta}{\operatorname{argmin}}\, \Phi_{-i}(\vartheta) \tag{4}$$

and response estimate $\hat{\eta}_{-i} := \eta(\hat{\vartheta}_{-i})$.

There are numerous ways of assessing how the optimum, *i.e.* the best estimate of the parameters, is affected by removing a data point [6]. The most elementary statistic is obtained by considering the difference between the best estimate of the parameters and the best estimate with the i^{th} data point removed:

$$D^*_{ij} := \hat{\vartheta}_j - \hat{\vartheta}_{-i,j}, \tag{5}$$

where $\hat{\vartheta}_{-i,j}$ is the value of the j^{th} parameter obtained from the optimisation with the i^{th} experiment omitted. In the literature this is usually referred to as DFBETA_i [2, p. 13].

We note that such an analysis requires $\hat{\vartheta}_{-i}$ to be calculated for all $i = 1, \ldots, N$, each requiring one optimisation. This can become computationally prohibitively expensive if the model itself is expensive or there are many experimental observations. If the considered model is linear, at least approximately, then it is possible to derive a formula which allows calculating the entire set of D^*_{ij} based on only a single optimisation [30]. This, however, is not an option if the model responses are strongly non-linear or are subject to numerical or statistical noise. The model considered in this work is by nature a stochastic model and its responses do exhibit non-negligible noise.

In order to compare or rank different parameters against each other with respect to their influence, due to different physical dimensions and/or orders of magnitude, it is essential to consider non-dimensionalised diagnostic measures. Belsley et al. [2, p. 13] recommend to normalise by the square root of an estimate of the variance of each parameter (with the i^{th} data point removed). This allows assessing the influence of data points on each parameter in relation to their uncertainty. Specifically, they propose to measure the influence of the i^{th} experiment upon the j^{th} parameter using $\mathsf{DFBETAS}_{ij} := D^*_{ij}/(\operatorname{Var}\hat{\vartheta}_j)^{1/2}$ (see also [9]), where $\operatorname{Var}\hat{\vartheta}_j$ refers to the variance of the j^{th} parameter. In some situations, the parameter variance may not be readily available, such as in this work where we directly optimise the model while progressively excluding experiments. Hence, we simply use here parameters which are normalised by (logarithmically) mapping them to the interval $[-1, 1]$.

Cook's distance [8], one of the most widely-used influence diagnostics, can be a useful tool for assessing the influence of an experimental data point during an optimisation. In the special case we consider in this work, *i.e.* that of uncorrelated responses with similar uncertainty, it can be defined as [6]

$$C_i := \frac{\sum_{n=1}^{N} \left[\hat{\eta}_n - \hat{\eta}_{-i,n}\right]^2}{P s^2}, \tag{6}$$

where $\hat{\eta}_{-i,n}$ is the value of the model response for the conditions of the n^{th} experiment obtained using the best parameter value estimates determined through optimisation with the i^{th} observation omitted (*i.e.* $\hat{\vartheta}_{-i}$), and where s^2 is an estimate of the mean square error, given by

$$s^2 = \frac{1}{N-P} \sum_{n=1}^{N} \left(\eta_n^{\text{exp}} - \hat{\eta}_n\right)^2. \tag{7}$$

Large values of Cook's distance C_i occur if deleting case i causes large differences in the parameter estimates.

The motivation for definition (6) stems from the notion of joint confidence regions for the parameters. Joint $100(1-\alpha)\%$ confidence ellipsoids for the model responses can be defined as

$$(\hat{\eta} - \eta)^\top \Sigma^{-1} (\hat{\eta} - \eta) \leq P s^2 F(P, N-P, 1-\alpha), \tag{8}$$

with s given by (7), and $F(P, N-P, 1-\alpha)$ the $1-\alpha$ point of the F-distribution (consult [13, pp. 94 & 108] and [19] for more details). Σ is the covariance matrix of the responses. Cook introduced his original measure [8] for ordinary least squares, *i.e.* unity covariance matrix, and later generalised it to weighted least squares [10, p. 209]. As mentioned above, if the responses are uncorrelated, of equal dimension, and of similar order of magnitude and uncertainty, Σ can be assumed to be the unit matrix.

Definition (6), like (5), involves one optimisation per experimental data point. As mentioned above, in situations where this is too computationally expensive, there may be the option of conducting a linearised analysis. For linear models, one can derive an expression for Cook's distance (6) which requires only a single regression for all observations. Whether or not a linear approximation is appropriate can be decided for example by means of local curvature [1, 38], but this is beyond the scope of the paper.

It is noted that Cook's distance measures only the *overall* influence of an observation, in contrast to (5), which assesses parameters individually. More generally, while in this work we consider the influence of single observations only on either single parameters or the model as a whole, this can be generalised to the influence of subsets of observations on subsets of parameters in the model (see for example [6, 9]). As the original notions, however, the measures tend to be applicable to linear models only, and may require additional regressions.

It is furthermore noted that, unlike (5), the Cook distance (6) is dimensionless by definition – a necessary property in order to achieve a generic classification of data points.

2.2.3 Outlier detection

One way of identifying potential outliers is by means of a threshold: A data point is deemed to require further attention if the corresponding value of the chosen diagnostic measure exceeds the threshold. Naturally, the choice of any such threshold is ultimately arbitrary, which is reflected in the fact that a range of them has been suggested in the literature. For example, Bollen and Jackman [3] propose

$$C_i \geq 4/N. \tag{9}$$

This threshold is very conservative in the sense that it tends to highlight too many points as outliers. On the other hand, Cook and Weisberg [11, p. 345] suggest

$$C_i \geq 1, \tag{10}$$

i.e. approximately the median of the F distribution with P and $N-P$ degrees of freedom (see Eqn. (8)). Irrespective of which value is chosen, it needs to be emphasised that this method can give only a rough indication, which should be interpreted merely as a suggestion of which data points warrant closer investigation. The main reason for this is that the method does not automatically distinguish between errors and highly influential points which potentially point towards genuine model improvements. Therefore, highlighted points should not necessarily be excluded from the analysis, as one may lose valuable information. Furthermore, whether or not a data point is deemed an ‘outlier’ by this method, is by definition dependent on the chosen model. That is, a data point labelled an outlier with respect to one model, may or may not appear as an outlier with respect to another (possibly better) model. As there is no consensus in the literature as to which cut-off should be used, in this work we consider both (9) and (10).

3 Experimental data

Table 2: *Experimental data sets with process conditions used to model them. X_{SIH_4} denotes the initial silane mole fraction, and τ denotes the residence time.*

Idx. i	Reference	Reactor type	Bath gas	$X_{\mathrm{SiH_4}}$ [%]	T [K]	P [kPa]	τ [ms]	d type	μ type	μ_i^{exp} [nm]
1	Körmer *et al.* [24]	Hot-wall flow reactor	Ar	4.0	873-1373	2.5	80	d_{pri}	Mode	26.7
2				4.0	873-1373		192		Mean	26.0
3				12.0	873-1373		192		Mean	38.0
4				12.8	873-1373		80		Mode	31.0
5				2.0	873-1373		80		Mode	41.0
6				8.0	873-1373		80		Mode	24.0
7				4.0	873-1373		420		Mode	32.5
8				4.0	673-1173		420		Mode	21.2
9				4.0	773-1273		420		Mode	28.5
10	Frenklach *et al.* [21]	Shock tube	Ar	3.3	1089	49	2.6	d_{pri}	Mode	11.0
11					1320		2.1			11.0
12					1580		1.8			15.0
13	Wu *et al.* [45]	Hot-wall flow reactor	N_2	1.0	770-1520	101	1000	d_{mob}	Mode	127
14	Flint *et al.* [18]	Laser-driven flow reactor	Ar	21.4	923-1270	20	5.2	d_{pri}	Mean	43.4
15				9.0	1023-1483		18			55.4
16				0.6	1023-1400		53			23.0
17	Nguyen and Flagan [33]	Hot-wall flow reactor	N_2	0.1	770-1080	101	900	d_{mob}	Mode	89.0
18				0.04						51.0
19	Onischuk *et al.* [34]	Hot-wall flow reactor	Ar	5.0	853	39	870	d_{pri}	Mean	52.0

As in previous work [26, 29], a total of nineteen experimental data points were selected from six different studies, spanning a range of process conditions and reactor configurations. Reactor types include hot-wall flow reactors and a shock tube, for each of which different temperatures, pressures, residence times, and initial silane mole fractions are covered. The particular selection of studies, though ultimately arbitrary amongst large amounts of literature, was motivated by covering a range of conditions. An overview of the chosen datasets is given in Table 2.

The study of Körmer et al. [24] is focused on synthesising silicon nanoparticles with narrow size distributions in a hot-wall flow reactor. In this setup, it turns out that most of the precursor is lost to deposits on the reactor wall, and therefore the initial composition is adjusted to account for this particle deposition [25]. As in [22], an initial silane mass of about 6×10^{-5} kg/m^3 is assumed. The amount of mass expected for a partial pressure of 1 mbar of silane at 1024 K is about 3.8×10^{-4} kg/m^3 indicating that only about 16% of the precursor are available to form particles. The initial silane fractions listed in Table 2 for this data subset are adjusted accordingly for our simulations.

The Flint et al. [18] data refers to their cases $630S$, $631S$, and $654S$, respectively. The experiment is described in detail in [4, 5, 17], including how to convert flow rates into residence times and initial compositions.

4 Results and discussion

Table 3: *The seven model parameters considered in the influence analysis, all Arrhenius pre-exponential factors (see Table 1), with optimal values resulting from optimisation against the complete data set.*

Idx.	Parameter	Optimal value	Unit	Phase	Role
1	$A_{1,\mathrm{LP}}$	2.87×10^{12}	cm^3/mol/s	Gas	Low-pressure limit of reaction #1
2	$A_{2,\mathrm{LP}}$	2.11×10^{35}			Low-pressure limit of reaction #2
3	$A_{3,\mathrm{LP}}$	4.90×10^{39}			Low-pressure limit of reaction #3
4	$A_{5,\mathrm{LP}}$	2.98×10^{68}			Low-pressure limit of reaction #5
5	$A_{8,\mathrm{rev}}$	1.48×10^{14}			Reverse of reaction #8
6	$A_{\mathrm{SR,SiH_4}}$	4.47×10^{33}	cm/mol/s	Particle	Surface reac.: silane addition, H_2-release
7	$A_{\mathrm{H_2}}$	1.88×10^{18}	1/s		H_2-release from particle

Both reactor types occurring in the set of experiments (Table 2), *i.e.* flow reactor and shock tube, are modelled as homogenous batch reactors. The shock tube is modelled as a constant temperature, constant pressure reactor. For the flow reactors, plug-flow is assumed, and the experimentally measured temperature profile, where available, is imposed. In case 19 [34], no temperature profile is available, so a constant temperature is assumed, and the residence time given refers to the approximate time spent in the 'hot-zone', *i.e.* at that temperature.

As software to carry out the necessary optimisations, we use the Model Development Suite (MoDS) [7] – a software tool for conducting various generic tasks to develop black-box models. Such tasks include parameter estimation and uncertainty quantification [32], Design of Experiments (DoE) [31], and global sensitivity analysis [29].

Each optimisation involved in the Cook distance and DFBETA analysis is performed in two stages: Firstly, a quasi-random global search is conducted using Sobol low-discrepancy sequences [41]. Secondly, starting from the best point identified in the first stage, a local optimisation is carried out using the Simultaneous Perturbation Stochastic Approximation (SPSA) [42] algorithm. The SPSA method estimates the local gradient based on only two objective function evaluations, and can be shown to obey the traditional gradient descent

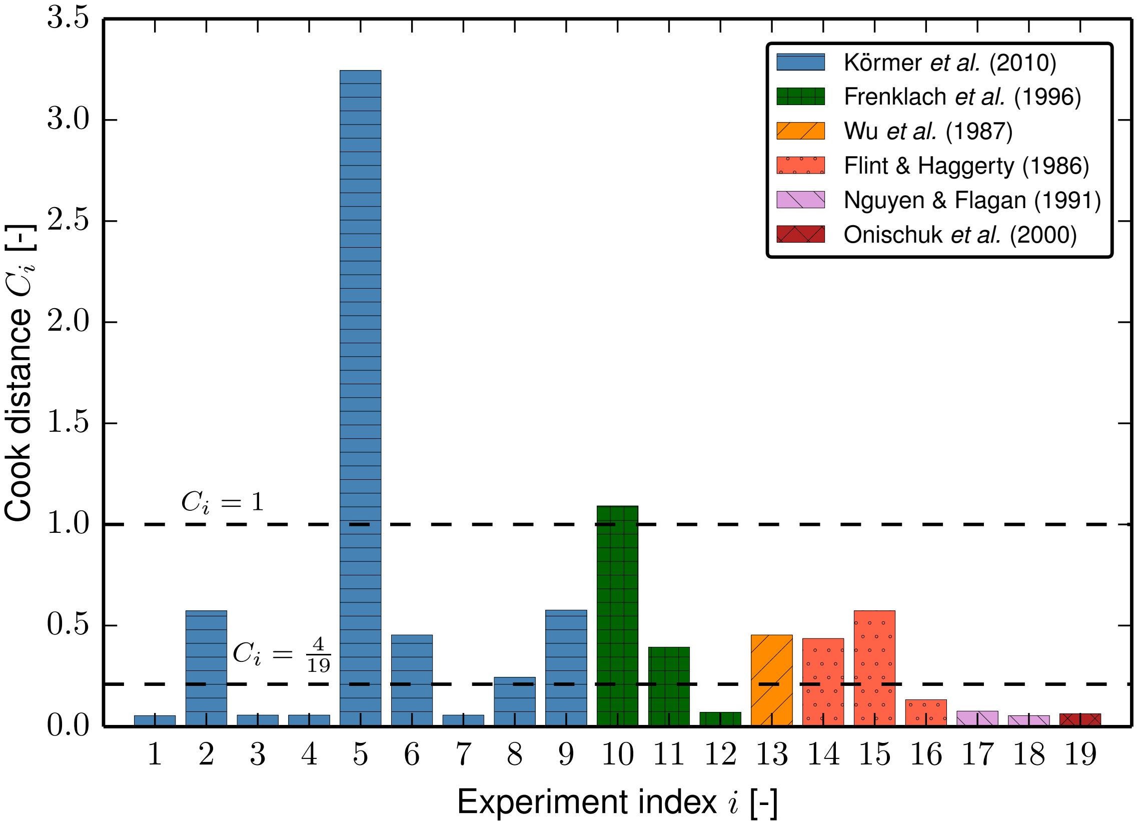


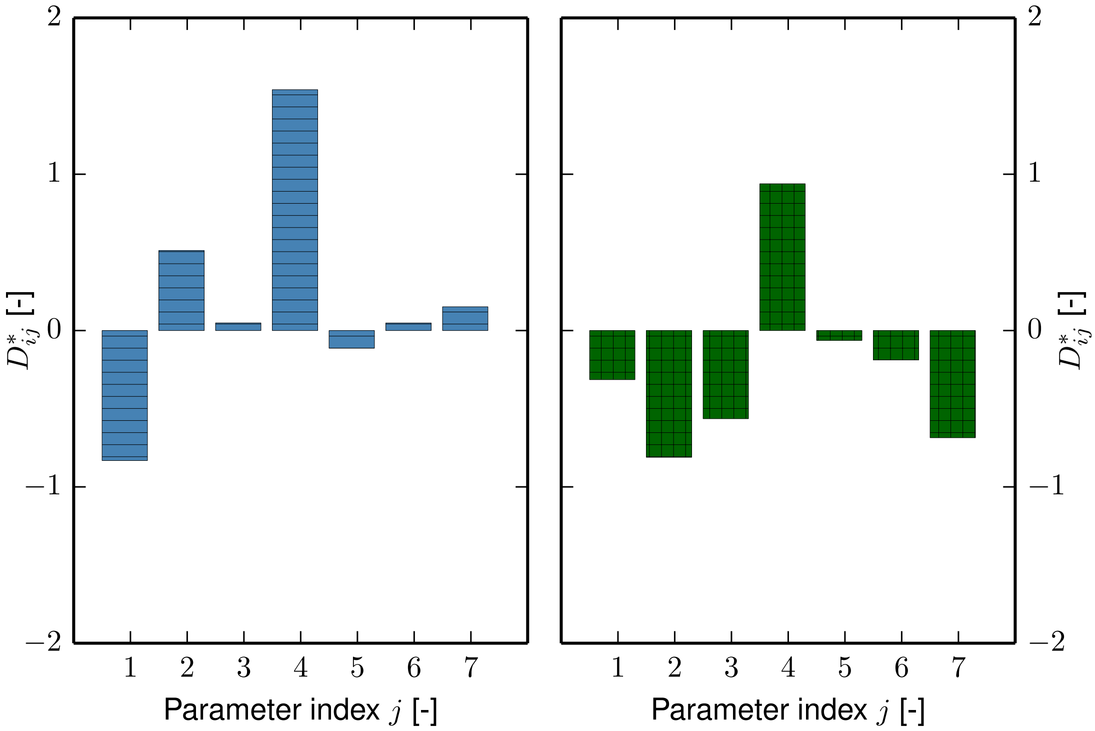
Figure 1: *Overall influence of each of the experimental observations in Table 2 as measured by Cook's distance C_i (Eqn. 6). Each of the thresholds (9) and (10) are indicated through dashed horizontal lines.*

on average. It is designed for problems where stochastic noise is present. The motivation for the first stage is to avoid becoming trapped in local minima or valleys on the objective function surface, which could happen with a method purely based on the local gradient. Chemical kinetic objective functions are widely reported to exhibit a complex, highly structured surface with multiple local minima and/or valleys (see for example [19]). Regarding the second stage, the reason for not choosing a more conventional method utilising the local Jacobi matrix or Hessian is the stochastic noise in the model response. While the procedure adopted here cannot guarantee to find the global minimum, based on previous experience [32], a low-lying minimum can be found at a manageable computational expense. On an objective function surface with multiple local minima, there is then of course the risk of selecting the 'wrong' optimum, *i.e.* not the global one. Any conclusions derived from perturbations such as those induced by omission of data points may change depending on the chosen minimum and the local geometry surrounding it.

Here, seven parameters were adjusted which represent key gas-phase and heterogeneous growth rates identified through sensitivity analysis [26]. Details are given in Table 3. Thus, the vector of model parameters to be optimised is given by

$$\vartheta = (A_{1,\mathrm{LP}}, A_{2,\mathrm{LP}}, A_{3,\mathrm{LP}}, A_{5,\mathrm{LP}}, A_{8,\mathrm{rev}}, A_{\mathrm{SR,SiH_4}}, A_{\mathrm{H_2}}).$$

The optimal values for the parameters resulting from optimisation against the full data set



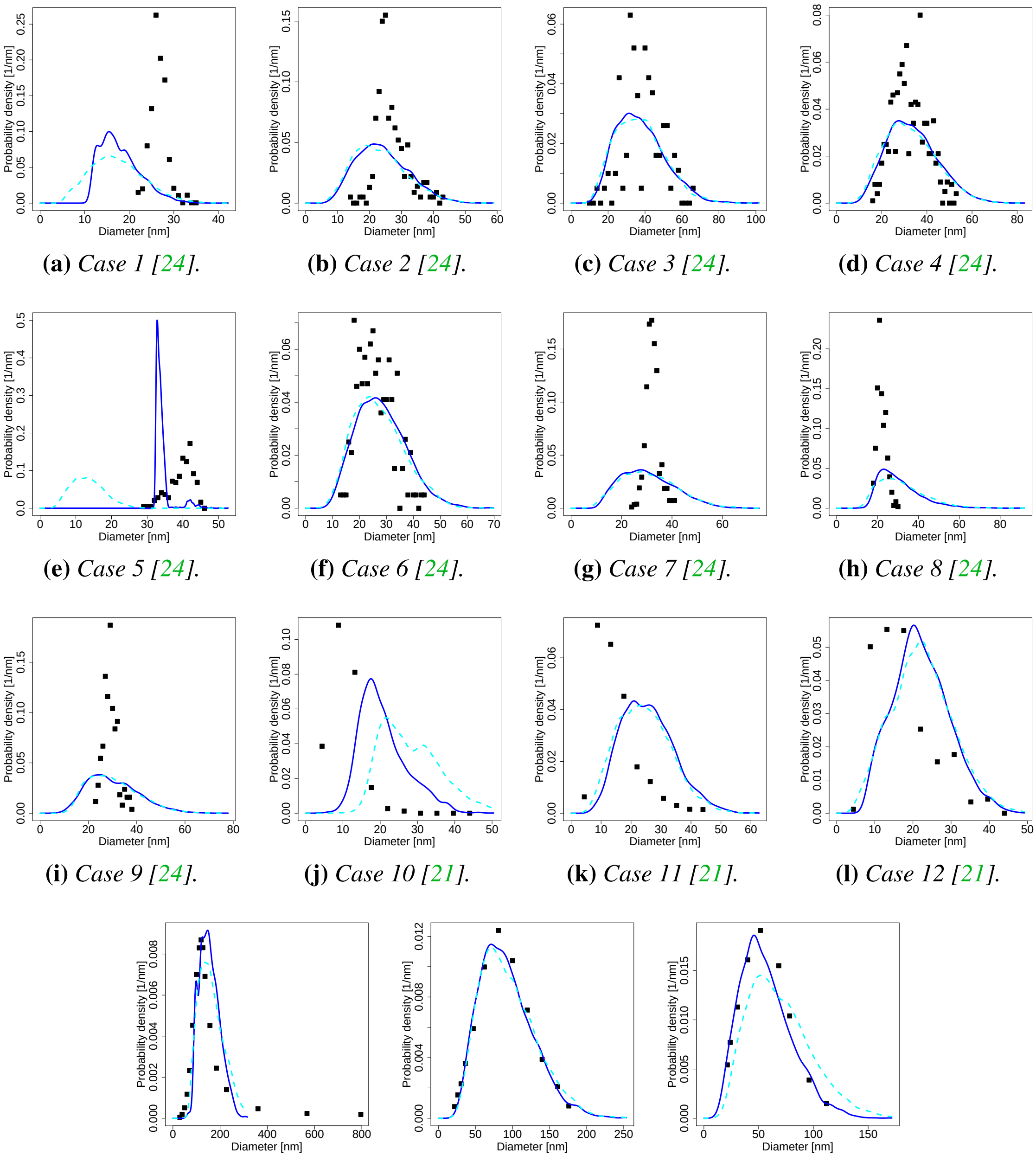
(a) *Influence of observation $i = 5$ by Körmer et al. [24] on each of the considered model parameters.* **(b)** *Influence of observation $i = 10$ by Frenklach et al. [21] on each of the considered model parameters.*

Figure 2: *DFBETA D_{ij}^* (Eqn. 5), for the two most influential experimental observations as identified in Fig. 1 (see also Table 2), for each of the parameters in Table 3.*

are also given in Table 3. The differences in these values as compared to [26] and [29] are due to the fact that different sets of responses are being considered.

For the optimisation against the complete data set, 800 Sobol points were generated, followed by 240 SPSA points. Recall that each point involves one evaluation of the objective function (Eqn. 1), and that every objective function evaluation involves 19 model evaluations. For all subsequent optimisations, *i.e.* those of Φ_{-i} (Eqn. 3) with $i = 1, \ldots, 19$, the model evaluations performed as part of the original set of Sobol points can be re-used, as all that is required is for each i to calculate the different objective function Φ_{-i} for all of the points. For each of the Φ_{-i} optimisations, 120 SPSA points were used. In total, this corresponds to about 3300 CPU-hours of computation.

The Cook distance analysis was conducted for all of the 19 experiments in Table 2, and results are shown in Fig. 1. In this figure, the responses are grouped by the particular experimental papers from which they were obtained. Both of the two outlier thresholds, Eqn. (9) and Eqn. (10), are shown. While several of the observations exceed the lower threshold (9), only two of them exceed the upper one (10) (with one of them only marginally). This is consistent with reports that (9) is too conservative in that it has a tendency to highlight too many observations, as mentioned in subsection 2.2.3. We conclude that observation $i = 5$ requires further attention, as its Cook distance exceeds both



(a) *Case 1 [24].* **(b)** *Case 2 [24].* **(c)** *Case 3 [24].* **(d)** *Case 4 [24].*

(e) *Case 5 [24].* **(f)** *Case 6 [24].* **(g)** *Case 7 [24].* **(h)** *Case 8 [24].*

(i) *Case 9 [24].* **(j)** *Case 10 [21].* **(k)** *Case 11 [21].* **(l)** *Case 12 [21].*

(m) *Case 13 [45].* **(n)** *Case 17 [33].* **(o)** *Case 18 [33].*

Figure 3: *Particle size distributions for those experiments for which they were measured. Solid lines: model optimised against the complete data set. Dashed lines: model optimised against the data set with the 5th experiment omitted. Points: experiment.*

thresholds and is significantly larger than all the others. This indicates that this experimental point most strongly affects the objective function Φ (Eqn. 1), which in turn affects the parameter estimates, *i.e.* the optimal values $\hat{\vartheta}$ of the parameters (Eqn. 2). It could furthermore suggest that this particular observation might be an outlier with respect to the present model, or, more likely, that the model describes it inadequately.

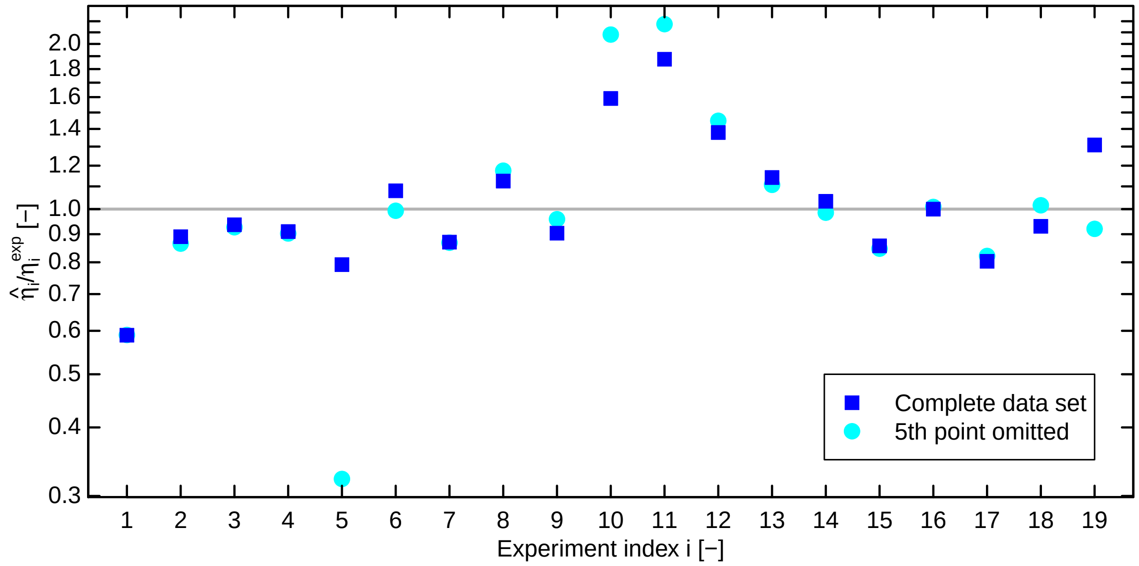


Figure 4: *Ratios of model responses to experimental values for each of the 19 experiments in Table 2. Two sets are shown: squares – optimised against the complete data set, and circles – optimised against the data set with the 5th experiment omitted.*

Additionally, a DFBETA analysis was conducted to assess how the experimental observations affect the values of the parameters which are determined through the optimisation (Fig. 2). In terms of highlighting individual observations, the DFBETA analysis agrees with the Cook distance analysis: The values of D^*_{ij} for $i = 5$ and $i = 10$ are at least two orders of magnitude larger than those obtained for any other experiment. The DFBETA values for these experiments are shown in Figs. 2a and 2b respectively. We notice that the best estimate of parameter 4, *i.e.* the pre-exponential factor in the low-pressure limit of reaction 5 (Table 1), is influenced most by both of the considered experimental observations.

In principle, there are two possible reasons for why an observation stands out in a Cook distance or DFBETA analysis: errors associated with the experiments, and errors associated with the model. Regarding experimental errors, we assume here that all experimental data are both correct and accurate. Considering model errors, these can be further categorised into the following: errors arising from the solution methodology, *i.e.* numerical algorithms, and flaws in the model. Specifically in this case, the latter include reactor model errors, and deficiencies in the gas or particulate phase sub-models.

Figure 3 shows particle size distributions for those experiments in Table 2 for which they have been measured. Figure 4 shows ratios of model responses to experimental ones for all experiments. Recall that only the means or modes of the distributions are optimised, not the widths or any other characteristic. Both figures show two sets of results, one for the optimisation against the complete data set, and one for the data set with the 5th experiment omitted. As expected, if the 5th experiment is omitted, the corresponding response deteriorates.

The $i = 5$ experiment refers to the lowest silane pressure case (0.5 mbar) reported by Körmer et al. [24]. In this hot-wall reactor experiment, a modal size of 42 nm was obtained for particles, larger than that obtained for a higher pressure (at 1 mbar yielded 26 nm particles). This inverse proportionality is not captured by the model, thus indicating clearly that this aspect requires further development.

5 Conclusions

We determined optimal values of seven parameters in a population balance model for the formation of silicon nanoparticles by means of least-squares optimisation against a set of 19 experiments. The influence of each of those measurements on the values of the considered kinetic model parameters was then quantified using Cook's distance and DFBETA – two basic omission-based measures popular in the field of regression influence diagnostics. An outlier analysis was then conducted by applying standard thresholds in order to identify the most important experimental datasets in the optimisation. This highlighted one particular experimental condition for further scrutiny. We emphasise again that, in general, a particular measurement exceeding an outlier threshold does not necessarily imply that there is a problem with that measurement or more generally the experiment. In the first instance, one should thoroughly examine whether there are shortcomings in the model which are responsible for the disagreement with the measurement. This informs future model development [31] by helping to identify aspects of the model which require improvement. Furthermore, if one regards the model as a formal representation of the best current knowledge about the experiment or system under consideration [20], then the methods can be thought of as giving an indication as to which measurements are most informative.

Acknowledgements

This work was partly funded by the Cambridge Australia Trust, by the National Research Foundation (NRF), Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme, and by the European Union Horizon 2020 Research and Innovation Programme under grant agreement 646121.

References

[1] D. M. Bates and D. G. Watts. Relative curvature measures of nonlinearity. *Journal of the Royal Statistical Society B*, 42(1):1–25, 1980.

[2] D. A. Belsley, E. Kuh, and R. E. Welsch. *Regression Diagnostics : Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons, 1980.

[3] K. A. Bollen and R. W. Jackman. Regression diagnostics: An expository treatment of outliers and influential cases. In J. Fox and J. S. Long, editors, *Modern Methods of Data Analysis*, pages 257–291. Sage Publications, Newbury Park, 1990.

[4] W. R. Cannon, S. C. Danforth, J. S. Flint, J. S. Haggerty, and R. A. Marra. Sinterable ceramic powders from laser-driven reactions: I, Process description and modeling. *Journal of the American Ceramic Society*, 65(7):324–330, 1982. doi:10.1111/j.1151-2916.1982.tb10464.x.

[5] W. R. Cannon, S. C. Danforth, J. S. Haggerty, and R. A. Marra. Sinterable ceramic powders from laser-driven reactions: II, Powder characteristics and process variables. *Journal of the American Ceramic Society*, 65(7):330–335, 1982. doi:10.1111/j.1151-2916.1982.tb10465.x.

[6] S. Chatterjee and A. S. Hadi. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, 1(3):379–393, 1986. doi:10.1214/ss/1177013622.

[7] cmcl innovations. MoDS (Model Development Suite), version 0.2.3, 2015. http://www.cmclinnovations.com/mod-suite/.

[8] R. D. Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18, 1977. Stable URL: http://www.jstor.org/stable/1268249.

[9] R. D. Cook and S. Weisberg. Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22(4):495–508, 1980. doi:10.1080/00401706.1980.10486199. Stable URL: http://www.jstor.org/stable/1268187.

[10] R. D. Cook and S. Weisberg. *Residuals and Influence in Regression*. Chapman and Hall, New York, 1982.

[11] R. D. Cook and S. Weisberg. Criticism and influence analysis in regression. In S. Leinhardt, editor, *Sociological Methodology*, pages 313–316. Jossey-Bass, San Francisco, 1982.

[12] N. R. Draper and J. A. John. Influential observations and outliers in regression. *Technometrics*, 23(1):21–26, 1981. doi:10.1080/00401706.1981.10486232. Stable URL: http://www.jstor.org/stable/1267971.

[13] N. R. Draper and H. Smith. *Applied Regression Analysis*. John Wiley & Sons, New York, 2nd edition, 1981.

[14] L. Eno, J. G. B. Beumee, and H. Rabitz. Sensitivity analysis of experimental data. *Applied Mathematics and Computation*, 16(2):153–163, 1985. doi:10.1016/0096-3003(85)90005-0.

[15] R. Feeley, P. Seiler, A. Packard, and M. Frenklach. Consistency of a reaction dataset. *Journal of Physical Chemistry A*, 108(44):9573–9583, 2004. doi:10.1021/jp047524w.

[16] A. V. Fiacco and G. P. McCormick. *Nonlinear Programming – Sequential Unconstrained Minimization Techniques*. Classics in Applied Mathematics. SIAM, 1990.

[17] J. Flint and J. Haggerty. A model for the growth of silicon particles from laser-heated gases. *Aerosol Science and Technology*, 13(1):72–84, 1990. doi:10.1080/02786829008959425.

[18] J. H. Flint, R. A. Marra, and J. S. Haggerty. Powder temperature, size, and number density in laser-driven reactions. *Aerosol Science and Technology*, 5(2):249–260, 1986. doi:10.1080/02786828608959091.

[19] M. Frenklach. Modeling. In W. C. Gardiner, editor, *Combustion Chemistry*, chapter 7, pages 423–453. Springer Verlag, New York, 1984.

[20] M. Frenklach. Transforming data into knowledge – Process Informatics for combustion chemistry. *Proceedings of the Combustion Institute*, 31(1):125–140, 2007. doi:10.1016/j.proci.2006.08.121.

[21] M. Frenklach, L. Ting, H. Wang, and M. J. Rabinowitz. Silicon particle formation in pyrolysis of silane and disilane. *Israel Journal of Chemistry*, 36(3):293–303, 1996. doi:10.1002/ijch.199600041.

[22] M. Gröschel, R. Körmer, M. Walther, G. Leugering, and W. Peukert. Process control strategies for the gas phase synthesis of silicon nanoparticles. *Chemical Engineering Science*, 73:181–194, 2012. doi:10.1016/j.ces.2012.01.035.

[23] P. Ho, M. E. Coltrin, and W. G. Breiland. Laser-induced fluorescence measurements and kinetic analysis of Si atom formation in a rotating disk chemical vapor deposition reactor. *The Journal of Physical Chemistry*, 98(40):10138–10147, 1994. doi:10.1021/j100091a032.

[24] R. Körmer, M. P. M. Jank, H. Ryssel, H.-J. Schmid, and W. Peukert. Aerosol synthesis of silicon nanoparticles with narrow size distribution – Part 1: Experimental investigations. *Journal of Aerosol Science*, 41(11):998–1007, 2010. doi:10.1016/j.jaerosci.2010.05.007.

[25] R. Körmer, H.-J. Schmid, and W. Peukert. Aerosol synthesis of silicon nanoparticles with narrow size distribution – Part 2: Theoretical analysis of the formation mechanism. *Journal of Aerosol Science*, 41(11):1008–1019, 2010. doi:10.1016/j.jaerosci.2010.08.002.

[26] W. J. Menz and M. Kraft. A new model for silicon nanoparticle synthesis. *Combustion and Flame*, 160(5):947–958, 2013. doi:10.1016/j.combustflame.2013.01.014.

[27] W. J. Menz, S. Shekar, G. P. E. Brownbridge, S. Mosbach, R. Körmer, W. Peukert, and M. Kraft. Synthesis of silicon nanoparticles with a narrow size distribution: A theoretical study. *Journal of Aerosol Science*, 44:46–61, 2012. doi:10.1016/j.jaerosci.2011.10.005.

[28] W. J. Menz, R. I. A. Patterson, W. Wagner, and M. Kraft. Application of stochastic weighted algorithms to a multidimensional silica particle model. *Journal of Computational Physics*, 248:221–234, 2013. doi:10.1016/j.jcp.2013.04.010.

[29] W. J. Menz, G. P. E. Brownbridge, and M. Kraft. Global sensitivity analysis of a model for silicon nanoparticle synthesis. *Journal of Aerosol Science*, 76:188–199, 2014. doi:10.1016/j.jaerosci.2014.06.011.

[30] S. Mosbach and M. Kraft. Influence of experimental observations on n-propylbenzene kinetic parameter estimates. *Proceedings of the Combustion Institute*, 35(1):357–365, 2015. doi:10.1016/j.proci.2014.05.061.

[31] S. Mosbach, A. Braumann, P. L. W. Man, C. A. Kastner, G. P. E. Brownbridge, and M. Kraft. Iterative improvement of Bayesian parameter estimates for an engine model by means of experimental design. *Combustion and Flame*, 159(3):1303–1313, 2012. doi:10.1016/j.combustflame.2011.10.019.

[32] S. Mosbach, J. H. Hong, G. P. E. Brownbridge, M. Kraft, S. Gudiyella, and K. Brezinsky. Bayesian error propagation for a kinetic model of n-propylbenzene oxidation in a shock tube. *International Journal of Chemical Kinetics*, 46(7):389–404, 2014. doi:10.1002/kin.20855.

[33] H. V. Nguyen and R. C. Flagan. Particle formation and growth in single-stage aerosol reactors. *Langmuir*, 7(8):1807–1814, 1991. doi:10.1021/la00056a038.

[34] A. A. Onischuk, A. I. Levykin, V. P. Strunin, K. K. Sabelfeld, and V. N. Panfilov. Aggregate formation under homogeneous silane thermal decomposition. *Journal of Aerosol Science*, 31(11):1263–1281, 2000. doi:10.1016/S0021-8502(00)00031-8.

[35] E. L. Petersen and M. W. Crofton. Measurements of high-temperature silane pyrolysis using SiH_4 IR emission and SiH_2 laser absorption. *The Journal of Physical Chemistry A*, 107(50):10988–10995, 2003. doi:10.1021/jp0302663.

[36] H. Rabitz, M. Kramer, and D. Dacol. Sensitivity analysis in chemical kinetics. *Annual Review of Physical Chemistry*, 34:419–461, 1983. doi:10.1146/annurev.pc.34.100183.002223.

[37] R. Schall and T. T. Dunne. Influential variables in linear regression. *Technometrics*, 32(3):323–330, 1990. doi:10.1080/00401706.1990.10484685. Stable URL: http://www.jstor.org/stable/1269109.

[38] G. A. F. Seber and C. J. Wild. *Nonlinear Regression*. John Wiley & Sons, 2003.

[39] S. Shekar, W. J. Menz, A. J. Smith, M. Kraft, and W. Wagner. On a multivariate population balance model to describe the structure and composition of silica nanoparticles. *Computers & Chemical Engineering*, 43:130–147, 2012. doi:10.1016/j.compchemeng.2012.04.010.

[40] S. Shekar, A. J. Smith, W. J. Menz, M. Sander, and M. Kraft. A multidimensional population balance model to describe the aerosol synthesis of silica nanoparticles. *Journal of Aerosol Science*, 44:83–98, 2012. doi:10.1016/j.jaerosci.2011.09.004.

[41] I. M. Sobol. On the systematic search in a hypercube. *SIAM Journal on Numerical Analysis*, 16(5):790–793, 1979. Stable URL: http://www.jstor.org/stable/2156633.

[42] J. C. Spall. Implementation of the Simultaneous Pertubation Algorithm for stochastic optimization. *IEEE Transactions on Aerospace and Electronic Systems*, 34(3): 817–823, 1998. doi:10.1109/7.705889.

[43] A. S. Tomlin. The role of sensitivity and uncertainty analysis in combustion modelling. *Proceedings of the Combustion Institute*, 34(1):159–176, 2013. doi:10.1016/j.proci.2012.07.043.

[44] T. Turányi. Sensitivity analysis of complex kinetic systems. Tools and applications. *Journal of Mathematical Chemistry*, 5(3):203–248, 1990. doi:10.1007/BF01166355.

[45] J. J. Wu, H. V. Nguyen, and R. C. Flagan. A method for the synthesis of submicron particles. *Langmuir*, 3(2):266–271, 1987. doi:10.1021/la00074a021.