

1 Parameterisation of a biodiesel plant process flow sheet 2 model

3 Janusz J. Sikorski^a, George Brownbridge^b, Sushant S. Garud^c, Sebastian
4 Mosbach^a, Iftekar A. Karimi^c, Markus Kraft^{*a,d}

5 ^a*Department of Chemical Engineering and Biotechnology, University of Cambridge,*
6 *New Museums Site, Pembroke Street, Cambridge, CB2 3RA United Kingdom*
7 *corresponding author**
8 *E-mail: mk306@cam.ac.uk*

9 ^b*CMCL Innovations, Sheraton House, Castle Park,*
10 *Cambridge, CB3 0AX, United Kingdom*

11 ^c*Department of Chemical and Biomolecular Engineering,*
12 *National University of Singapore, 4 Engineering Drive 4, Singapore 117585*

13 ^d*School of Chemical and Biomedical Engineering, Nanyang Technological University,*
14 *62 Nanyang Drive, Singapore, 637459*

15 Abstract

16 This paper presents results of parameterisation of typical input-output re-
17 lations within process flow sheet of a biodiesel plant and assesses parame-
18 terisation accuracy. A variety of scenarios were considered: 1, 2, 6 and 11
19 input variables (such as feed flow rate or a heater's operating temperature)
20 were changed simultaneously, 3 domain sizes of the input variables were con-
21 sidered and 2 different surrogates (polynomial and High Dimensional Model
22 Representation (HDMR) fitting) were used. All considered outputs were heat
23 duties of equipment within the plant. All surrogate models achieved at least
24 a reasonable fit regardless of the domain size and number of dimensions.
25 Global sensitivity analysis with respect to 11 inputs indicated that only 4 or
26 fewer inputs had significant influence on any one output. Interaction terms
27 showed only minor effects in all of the cases.

28 *Keywords:* process flow sheet model, parameterisation, biodiesel, sensitivity analysis

29 **1. Introduction**

30 Every industrial actor strives towards better understanding and, ulti-
31 mately, optimisation of any and all of its activities. That applies on each
32 level beginning with workforce schedules and individual pieces of machinery,
33 through specific processes, ending with entire plants. Traditionally the main
34 objectives of such an optimisation are minimising resource use and maximis-
35 ing profit. However, as environmental concerns become ever more press-
36 ing ecologically-focused targets such as reducing pollutants, creating cleaner
37 manufacturing processes or reducing carbon footprints rise in prominence.

38 Those trends prompted significant academic and industrial interest in the
39 concepts of "sustainable development" [1], "industrial ecology" [2, 3, 4, 5] and
40 "industrial symbiosis" [6]. The latter concept brings together separate indus-
41 tries in a collective approach to competitive advantage involving physical ex-
42 change of materials, energy, water and by-products [6]. Ecological industrial
43 development based thereon is often realised as Eco-Industrial Parks (EIPs).

44 An EIP is defined as an industrial park where businesses cooperate with
45 each other and, at times, with the local community to reduce waste and
46 pollution, efficiently share resources (such as information, materials, water,
47 energy, infrastructure, and natural resources), and minimise environmental
48 impact while simultaneously increasing business success [7]. An example of
49 an EIP exists in Kalundborg, Denmark where an exchange network is centred
50 around Asnæs Power Station, a 1500MW coal-fired power plant, and linked
51 to the local community and several other companies [6, 8]. Sample exchanges

52 include selling excess steam from the plant to Novo Nordisk, a pharmaceutical
53 and enzyme manufacturer, and to Statoil power plant or using extra heat to
54 heat local homes and a nearby fish farm. Also, one of the plant's by-products,
55 gypsum, is purchased by a wallboard producer, helping to reduce the amount
56 of necessary open-pit mining [9].

57 Primary academic interest stems from EIPs' ability to create more sus-
58 tainable industrial activities through the use of localised symbiotic relation-
59 ships [10]. To this date a great number of studies concerning various aspects
60 of EIPs have been conducted. Many of them probe methods suitable for
61 optimal design, focusing primarily on employing mathematical programming
62 to create exchange networks of materials, water and energy connecting mem-
63 bers of the EIP in question [11, 12, 13, 14, 15]. Utility of such designs is
64 evaluated by monitoring environmental, social and economical impacts.

65 Holistic modelling of complex, highly interconnected networks is a non-
66 trivial and expensive task, especially for EIPs which include numerous phys-
67 ical models of disparate processes. That is why many studies apply mathe-
68 matical optimisation to simplified models of individual aspects of the parks.

69 The limitations of this approach may be overcome by exploiting key fea-
70 tures of the concept of Industry 4.0 [7]: creation of virtual copies of the phys-
71 ical world and the ability of industrial components to communicate with each
72 other. Those virtual copies could be surrogate models of physical models pro-
73 duced for a predefined range of inputs. Developing a virtual system primarily
74 based on surrogate models would significantly reduce required computation
75 time and storage space and allow for dynamic modelling and studies other-
76 wise impossible to conduct. Figure 1 presents a framework of EIP modelling

77 based on Industry 4.0.

78 A surrogate model (or a metamodel) is an approximation of experimental
79 and/or simulation data designed to provide answers when it is too expen-
80 sive to directly measure the outcome of interest [16]. Two key requirements
81 thereof are reasonable accuracy and significantly faster evaluation than the
82 original method. The models are used to:

- 83 • explore design space of a simulation or an experiment,
- 84 • calibrate predictive codes of limited accuracy and bridging models of
85 varying fidelity,
- 86 • account for noise or missing data,
- 87 • gain insight into nature of the input-output relationship (data mining,
88 sensitivity analysis and parameter estimation).

89 Producing a surrogate model involves choosing a sampling plan (an ex-
90 perimental design), choosing a type of model and fitting the model to the
91 gathered data. Numerous sampling and fitting techniques are available as
92 documented in a number of reviews. Simpson et al. [17] provides detailed
93 reviews of data sampling and metamodel generation techniques, including
94 response surfaces, kriging, Taguchi approach, artificial neural networks and
95 inductive learning. It also discusses metrics for absolute and relative model
96 assessment, including R^2 , residual plots and root mean square error. An
97 introduction to and analysis of linear regression with a focus on general-
98 ized linear mixed models with many examples and case studies is provided
99 by Ruppert et al. [18].

100 A book by Forrester et al. [16] puts the process of data sampling and
 101 generating surrogate models into engineering perspective providing numerous
 102 case studies and MATLAB code to perform associated calculations. It dis-
 103 cusses response surfaces, kriging, support vectors machines and radial basis
 104 functions. An in-depth review of kriging, its application and new extensions
 105 are provided by Kleijnen [19]. A review and assessment of various sampling
 106 techniques is provided by Crary [20]. Reich and Barai [21] focuses on assess-
 107 ment of machine learning techniques, artificial neural networks in particular,
 108 with case studies of modelling marine propeller behavior and corrosion data
 109 analysis. An example of surrogate models bridging models of varying fidelity
 110 is provided by Bakr et al. [22] where a surrogate maps data produced by
 111 fine and coarse physical models in order to accelerate optimisation of the
 112 fine model. Jin et al. [23] assesses applicability and accuracy of metamodels
 113 for optimisation under uncertainty and reports promising results noting that
 114 only a small-size analytical problem was considered. Surrogate models are
 115 widely employed in engineering and science for space exploration [24, 25],
 116 modelling [26, 27, 28], sensitivity analysis [29, 30, 24, 31, 32], parameter esti-
 117 mation [33, 34, 35], optimisation in areas ranging from circuit design through
 118 nanoparticle synthesis to flood monitoring [36, 37, 38]. A number of studies
 119 addressed application of surrogates to process flow sheet models. Caballero
 120 and Grossmann [39] replace the computationally expensive subsystems of a
 121 flow sheet with Kriging surrogates to speed up optimisation. Hasan et al.
 122 [40], First et al. [41], Hasan et al. [42], Nuchitprasittichai and Cremaschi
 123 [43], Boukouvala and Ierapetritou [44] guide sampling of an expensive rigor-
 124 ous model using Kriging surrogates to reduce computational time required

125 for optimisation. Fahmi and Cremaschi [45] optimise a design of a biodiesel
126 production plant by replacing all subsystems in a process flow sheet model
127 with surrogate models based around artificial neural networks (ANNs) and
128 solving thus defined mixed-integer non-linear problem. Henao and Maravelias
129 [46] propose a systematic method for creating surrogate models of chemical
130 engineering systems and arranging them into a solvable network (superstruc-
131 ture). The study focuses on ANNs as a base for their surrogate models and
132 describes how a superstructure can be optimised. Kong et al. [47] employ
133 some of the concepts developed in Henao and Maravelias [46] for design op-
134 timisation of a chemical plant with heat integration and an attached utility
135 plant. This paper includes a case study of non-enzymatic ethanol produc-
136 tion from biomass. However, none of the aforementioned papers presents
137 a detailed accuracy analysis of surrogate models describing a process flow
138 sheet model of a typical industrial process nor compares the performance of
139 various surrogate models when describing a process flow sheet model.

140 The main purpose of this paper is to approximate the relations between
141 11 inputs typical to a biodiesel plant and its energy requirements using sur-
142rogate models and assess accuracy of the approximations. The models are
143intended to be used in a tool [7] for online, real-time simulations of large
144scale, industrial networks. Additionally, it aims to investigate the effects of
145dimensionality, domain size and surrogate type on the accuracy and analyse
146global sensitivities of the outputs in order to identify opportunities for di-
147mensionality reduction. High Dimensional Model Representation (HDMR)
148is used to perform global sensitivity analysis.

149 This paper is structured as follows. Section 2 describes the biodiesel

150 plant model and its modelling environment. Section 3 presents sampling
151 and surrogate generation techniques prodecures and software employed to
152 perform those. Section 4 provides implementation details of the surrogate
153 models and accuracy indices used to assess them. Section 5 presents results
154 of the numerical analysis, while Section 6 summarizes the main findings.

155 2. Model

156 2.1. Aspen Plus V8.6

157 Aspen Plus is a process modelling and optimisation software used by
158 the bulk, fine, specialty, and biochemical industries, as well as the polymer
159 industry for the design, operation, and optimisation of safe, profitable man-
160 ufacturing facilities [48]. Its capabilities include:

- 161 • optimisation of processing capacity and operating conditions,
- 162 • assessment of model accuracy,
- 163 • monitoring safety and operational issues,
- 164 • identifying energy savings opportunities and reduce greenhouse gas
165 (GHG) emissions,
- 166 • performing economic evaluation,
- 167 • improving equipment design and performance.

168 The software was used to simulate the process described in Section 2.2.

169 2.2. Biodiesel plant simulation

170 The process flow sheet model under investigation includes initial stages
171 of a biodiesel production line, namely a reaction step and a separation step,
172 with auxiliary equipment as seen in Figure 2. The final fuel, fatty acid
173 methyl ester, is produced via trans-esterification pathway where triglycerides
174 react with methanol to form methyl ester and glycerin in the presence of an
175 alkaline catalyst. The flow sheet was based on an existing plant designed

176 by Lurgi GmbH. It consists of the following elements: a continuously stirred
177 tank reactor (CSTR), a flash drum, a decanter, 3 heaters and 11 material
178 streams. In the process tripalmitine oil is reacted with methanol in the CSTR
179 to produce glycerol and methylpalmitate (biodiesel) and then passed through
180 a flash drum and a decanter to separate excess methanol and glycerol. The
181 simulation is solved for steady-state operation and produces a wide variety
182 of chemical and physical information ranging from throughput to heat duties
183 of individual equipment.

184 In this study surrogate models were used to describe relations between
185 chosen inputs and outputs occurring in the process flow sheet model. The
186 choice of variables aimed to study effects of inputs typical for chemical plants
187 on energy consumption as it is desired to study interactions between chemical
188 and electrical models in the future. Three domain sizes of the input variables
189 were considered in order to assess their effect on the parametrisation accu-
190 racy. The variables' names, domain and preferred operating conditions are
191 listed in Tables 1 and 2. Plots of heat duties of various equipment against
192 molar flow of tripalmitin oil can be seen in Figure 3.

193 **3. Parameterisation**

194 *3.1. Model Development Suite*

195 Model Development Suite (MoDS) [49] is an advanced software tool de-
196 signed to analyse black-box models (e.g. executables, batch scripts). It in-
197 cludes a broad range of tools such as data-driven modelling, multi-objective
198 optimisation, generation of surrogate models, data standardisation and visu-
199 alisation, global parameter estimation [35, 50, 51, 52, 53, 54, 55, 56, 31], un-

Table 1: Input variables.

Name	Lower bounds	Upper bounds	Operating point
Molar flow of tripalmitine oil (kmol/hr)	20, 22.5, 25	40, 37.5, 35	30
Temperature of tripalmitine oil (°C)	20, 22.5, 25	40, 37.5, 35	30
Operating temperature of CSTR 10D01 (°C)	44, 49, 54	64	60
Volume of CSTR 10D01 (m ³)	40, 43, 45	50, 49, 47	45
Operating temperature of flash drum 10D02 (°C)	80, 82.5, 85	100, 97.5, 95	90
Operating temperature of heater 10E01 (°C)	60, 62.5, 65	80, 77.5, 75	70
Molar flow of methanol (kmol/hr)	150, 160, 170	210, 200, 190	180
Temperature of methanol (°C)	20, 22.5, 25	40, 37.5, 35	30
Operating temperature of decanter 10D02D (°C)	20, 22.5, 25	40, 37.5, 35	30
Operating temperature of heater 10E02 (°C)	80, 82.5, 85	100, 97.5, 95	90
Operating temperature of heater 10E03 (°C)	60, 62.5, 65	80, 77.5, 75	70

Table 2: Output variables.

Name
Heat duty of heater 10E01 (MW)
Heat duty of heater 10E02 (MW)
Heat duty of heater 10E03 (MW)
Heat duty of reactor 10D01 (MW)
Heat duty of flash drum 10D02 (MW)
Heat duty of decanter 10D02D (MW)

200 certainty propagation [57, 58], global and local sensitivity analysis [59, 60, 29],
201 and intelligent design of experiments [61, 62, 63, 64, 65, 66]. It was used to
202 sample data, produce surrogate models and compute global sensitivities.

203 Sobol sequence, a quasi-random low discrepancy sampling method, is
204 employed for sampling data and polynomial fitting and HDMR fitting are
205 used to generate surrogate models. A brief description of each is included,
206 respectively, in Sections 3.4, 4.1 and 4.2.

207 *3.2. MoDS-Aspen Plus interface - Component Object Model (COM)*

208 The data collection and parameterization process of a model can be au-
209 tomated using MoDS provided an executable file capable of reading an input
210 file, running the considered model and producing an output file (input and
211 output files need to have either .csv or .xml format).

212 For the purpose of this study a script written in Python 3.4 was used
213 to manipulate the Aspen Plus simulation via Microsoft Component Object
214 Model (COM) interface. COM is a platform-independent, binary-interface
215 standard enabling creation of objects and communication between them [67].
216 COM object (also known as COM component) is defined as a piece of com-
217 piled code that provides a service to the rest of the system. That can be a
218 script, an instance of a program e.g. an Aspen Plus simulation. A primary
219 feature of this architecture is the fact that COM components access each
220 other through interface pointers, rather than directly. It provides a number
221 of functions applicable to all components. Any additional functions need to
222 be provided by the object or the user, in both cases via a library associated
223 with the object. In this project COM interface is primarily used to launch,
224 explore data structures, access data entries and solve models simulated within

225 Aspen Plus.

226 *3.3. Data harvest and surrogate generation*

227 Data collection, processing and visualisation were done using MoDS and
228 custom-made Python 3.4 and R 3.2.2 scripts. The process of producing a
229 surrogate of existing models involves the following steps: generation of input
230 data, reception of output data from the studied model and, when both data
231 sets are complete, scanning for and excluding erroneous data points and
232 executing a parametrisation algorithm. The first two steps are critical to
233 ensure high accuracy of the surrogate model and hence a sufficient number of
234 points and a suitable sampling method are required to satisfactorily describe
235 the input-output relation for a given number of independent variables and
236 operating range. In this study the following procedure was used:

- 237 1. A Sobol sequence was used to generate input data for user-specified
238 variables within the process flow sheet model.
- 239 2. Model's input data was altered according to the generated input data.
- 240 3. The simulation was evaluated with the new inputs.
- 241 4. MoDS retrieved values of user-specified outputs.
- 242 5. Data was scanned for errors and corrected.
- 243 6. Polynomial and HDMR fitting were used to generate surrogate models
244 describing the relation between inputs and outputs.

245 The workflow of MoDS is visualized in Figure 4. A variety of scenarios
246 were considered: 1, 2, 6 and 11 input variables were changed simultaneously,
247 3 different domain sizes of the input variables were considered and 2 different
248 surrogate generation methods (polynomial and HDMR fitting) were used. To

249 ensure that there is always sufficient number of points required to generate
 250 a surrogate, each simulation produced 400 points per input variable (prior
 251 to error exclusion). They were used for fitting surrogates and calculating
 252 R^2 and \bar{R}^2 . Depending on the case, erroneous points made up to 1% of all
 253 points. They arose due to convergence and stability issues within Aspen
 254 Plus. Additionally, test sets of points (100 points per dimension) were gen-
 255 erated for calculating Root-Mean-Square Deviation (RMSD) and residuals
 256 (see Section 4.3 for further description). In this study three domain sizes
 257 of the input variables were considered in order to assess their effect on the
 258 parameterisation accuracy. The domain bounds of input variables during
 259 simulations and initial steady state values are summarised in Table 1.

260 3.4. *Sampling*

261 Data points were generated using Sobol sequences, a type of quasi-random,
 262 low-discrepancy sequences. Low discrepancy of points in such a sequence
 263 means that their proportion falling into an arbitrary set is approximately
 264 proportional to the measure of the set. This property is true on average, but
 265 not necessarily for specific samples. Their ability to cover considered domain
 266 quickly and evenly gives them advantages over purely random numbers. Also,
 267 in contrast to deterministic sequences, they do not require a predefined num-
 268 ber of samples and their coverage improves continually as more data points
 269 are added. Sobol sequences uses a base of two to form successively finer uni-
 270 form partitions of the unit interval, and then reorder the coordinates in each
 271 dimension [68]. The MoDS implementation of a Sobol sequence generator
 272 follows the description of Joe and Kuo [69].

273 4. Implementation

274 4.1. Polynomial response surfaces

275 Polynomial response surfaces are a subset of response surface methodol-
276 ogy, a group of mathematical and statistical techniques designed to facilitate
277 empirical model building [70]. Polynomials of a predefined degree are opti-
278 mized to describe an unknown relation between independent variables (input
279 variables) and responses (output variables). Input and output data sets are
280 obtained via series of tests, an experiment, in which the input variables are
281 modified in order to study the changes in the output responses. As the num-
282 ber of adjustable coefficients in a polynomial surrogate increases combinato-
283 rially with its order and number of variables so does the minimum number
284 of data points required to produce it. Hence applying high-order polynomi-
285 als to problems with many inputs may lead to overfitting and hence poorer
286 predictive power. Generally, overfitting occurs when a model describes fea-
287 tures specific to the data set on which it is trained such as random error or
288 noise. For deterministic computer experiments those are not an issue, but an
289 overfitted model will suffer from having an exaggerated set of coefficients pro-
290 viding no intuitive insight into nature of the relationship under consideration
291 and from introducing irrelevant nonlinearity.

292 *General linear least-squares fit*

293 When fitting polynomial of a given order k to a data set the objective
294 function to be minimised is the weighted sum of the squares of the differences
295 between data and model. This analysis assumes N data values $y^{(1)}, \dots, y^{(N)}$
296 obtained at the points $x^{(1)}, \dots, x^{(N)}$, and statistical weights $W^{(1)}, \dots, W^{(N)}$

297 are given. Coefficients of the polynomial are given by

$$\beta^* = \underset{\beta}{\operatorname{argmin}} \Phi(\beta)$$

298 with

$$\Phi(\beta) = \sum_{i=1}^N W^{(i)} [y^{(i)} - f_{\beta}(x^{(i)})]^2$$

299 In order to simplify the notation, multi-indices are employed. For ex-
 300 ample, if p is a multi-index of order l , that means $p \in \mathbb{N}_0^l$, where $\mathbb{N}_0 :=$
 301 $\{0, 1, 2, \dots\}$. Then,

$$|p| := \sum_{i=1}^l p_i.$$

302 The independent variable is denoted by x and it is assumed that $x \in \mathbb{R}^n$.
 303 A polynomial in x is then a sum of terms of the form

$$x_1^{p_1} x_2^{p_2} \dots x_n^{p_n},$$

304 which can be abbreviated to x^p and is of order $|p|$. Thus the polynomial
 305 f_{β} can be written as

$$f_{\beta}(x) = \sum_{|p| \leq k} \beta_p x^p.$$

306 where the β s denote the coefficients of the individual terms and k corresponds
 307 to the polynomial order.

308 The necessary condition $\frac{\partial \Phi}{\partial \beta_q} = 0$ for any multi-index q with $|q| \leq k$ for
 309 stationary points of Φ then becomes

$$\begin{aligned}
0 = \frac{\partial}{\partial \beta_q} \Phi(\beta) &= 2 \sum_{i=1}^N W^{(i)} [y^{(i)} - f_\beta(x^{(i)})] \frac{\partial}{\partial \beta_q} f_\beta(x^{(i)}) \\
&= 2 \sum_{i=1}^N W^{(i)} [y^{(i)} - f_\beta(x^{(i)})] \frac{\partial}{\partial \beta_q} \sum_{|p| \leq k} \beta_p (x^{(i)})^p \\
&= 2 \sum_{i=1}^N W^{(i)} \left[y^{(i)} - \sum_{|p| \leq k} \beta_p (x^{(i)})^p \right] (x^{(i)})^q.
\end{aligned}$$

310 Rearranging yields

$$\begin{aligned}
\sum_{i=1}^N W^{(i)} y^{(i)} (x^{(i)})^q &= \sum_{i=1}^N W^{(i)} \sum_{|p| \leq k} \beta_p (x^{(i)})^p (x^{(i)})^q \\
&= \sum_{|p| \leq k} \beta_p \left[\sum_{i=1}^N W^{(i)} (x^{(i)})^p (x^{(i)})^q \right].
\end{aligned} \tag{1}$$

311 This linear system of equations, called normal equations, consists of $\binom{n+k}{k}$
312 equations for as many unknown coefficients β .

313 4.2. High Dimensional Model Representation

314 High Dimensional Model Representation (HDMR) is a finite expansion
315 for a given multivariable function as described by Sobol [71], Rabitz and
316 Alis [72]. It allows for readily extracting global sensitivities with respect
317 to the independent variables by calculating them from the coefficients of a
318 HDMR surrogate. Also, it needs to be noted that the number of parameters
319 within HDMR fit increases far slower than within polynomial fit when high-
320 dimensional problems are considered.

321 In HDMR representation the output function y is decomposed into a sum
 322 of functions that only depend on subsets of the input variables such that:

$$y = f(x) = f_0 + \sum_{i=1}^{N_x} f_i(x_i) + \sum_{i=1}^{N_x} \sum_{j=i+1}^{N_x} f_{ij}(x_i, x_j) + \cdots + f_{12\dots N_x}(x_1, x_2, \dots, x_{N_x})$$

323 where N_x is the number of input parameters, i and j index the input
 324 parameters, and f_0 is the mean value of $f(x)$. The expansion given above
 325 has a finite number of terms and exactly represents $f(x)$, however for most
 326 practical applications terms containing functions of more than two input
 327 parameters can often be ignored due to their negligible contributions com-
 328 pared to the lower order terms [73, 72]. Hence for most models or data the
 329 truncated approximation:

$$y \approx f(x) = f_0 + \sum_{i=1}^{N_x} f_i(x_i) + \sum_{i=1}^{N_x} \sum_{j=i+1}^{N_x} f_{ij}(x_i, x_j)$$

330 is sufficient. An efficient method of evaluating each of these terms is
 331 to approximate the functions $f_i(x_i)$ and $f_{ij}(x_i, x_j)$ with analytic functions,
 332 $\phi_k(x_i)$, [73]. For data produced using random and quasi-random sampling
 333 these functions are related by:

$$f_0 = \bar{f}, \tag{2a}$$

$$f_i(x_i) = \sum_{k=1}^M \alpha_{i,k} \phi_k(x_i), \tag{2b}$$

$$f_{ij}(x_i, x_j) = \sum_{k=1}^{M'} \sum_{l=k+1}^{M'} \beta_{ij,kl} \phi_k(x_i) \phi_l(x_j). \tag{2c}$$

334 The functions, $\phi_k(x_i)$ are orthonormal obeying,

$$\int \phi_k(x_i) dx_i = 0 \quad (3a)$$

$$\int \phi_k(x_i) \phi_l(x_i) dx_i = \delta_{kl}. \quad (3b)$$

335 This leads the following equations for the coefficients:

$$f_0 = \int f(x) dx, \quad (4a)$$

$$\alpha_{i,k} = \int f(x) \phi_k(x_i) dx, \quad (4b)$$

$$\beta_{ij,kl} = \int f(x) \phi_k(x_i) \phi_l(x_j) dx, \quad (4c)$$

336 The separation of the contributions from each individual input parameter
 337 and each combination of parameters makes the process of calculating the
 338 global sensitivities almost trivial. It has been described by Rabitz and Aliş
 339 [72] that the contribution of each term in (2), $\sigma_{y,i}^2$ and $\sigma_{y,ij}^2$, to the variance
 340 of the output parameter can be related to the total variance by

$$\sigma_y^2 = \sum_{i=1}^{N_x} \int_{-1}^1 f_i^2(x_i) dx_i + \sum_{i=1}^{N_x} \sum_{j=i+1}^{N_x} \int_{-1}^1 \int_{-1}^1 f_{ij}^2(x_i, x_j) dx_i dx_j \quad (5a)$$

$$= \sum_{i=1}^{N_x} \sigma_{y,i}^2 + \sum_{i=1}^{N_x} \sum_{j=i+1}^{N_x} \sigma_{y,ij}^2. \quad (5b)$$

341 The sensitivities, S_i and S_{ij} , can then be calculated by dividing by the
 342 total variance σ_y^2 to get

$$S_i = \frac{\sigma_{y,i}^2}{\sigma_y^2} \quad \text{and} \quad S_{ij} = \frac{\sigma_{y,ij}^2}{\sigma_y^2}. \quad (6)$$

Global sensitivity analysis explores the parameter space and provides robust sensitivity measures throughout the region of interest even in the presence of nonlinearity and parameter interactions. In nonlinear cases, derivative-based local sensitivity analysis can give a false impression of sensitivity [74].

4.2.1. Basis functions

Polynomials, including Lagrange polynomials [75], orthonormal polynomials, cubic B splines, and ordinary polynomials [73], are commonly used as basis functions for HDMR construction.

In MoDS, Legendre polynomials, $P_m(x)$, are used as the basis functions, $\phi(x)$. They are normalised according to

$$\int_{-1}^1 P_m(x) P_n(x) dx = \frac{2}{2n+1} \delta_{mn}, \quad (7)$$

to satisfy (3b). The polynomials are generated at runtime according to Bonnet's recursion formula

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x), \quad (8)$$

where $P_0(x) = 1$ and $P_1(x) = x$. This means that maximum polynomial order, M^* , can be set to an arbitrary natural number. Additionally, maximum interaction order, M'^* , needs to be set to either 1 or 2.

4.2.2. Automatic order selection

Accuracy improvement due to each new term is assessed by calculating R^2 value and comparing it against a predefined minimum value R^{2*} (0.00001), before continuing on to the next one. If a term's contribution is smaller than

the threshold, the term is discarded. The algorithm terminates once maximum polynomial orders M^* and M'^* are reached. It has several advantages over employment of a raw polynomial including reduction of data processing, computational complexity and number of optimisable parameters, which greatly helps dealing with high-dimensional problems. All of the functions f_i have the same polynomial order, M^* , and the f_{ij} are all of order M'^* . Also, it is assumed that the magnitude of the coefficients decreases as the order of the basis function increases. Whilst this is valid in many situations it may not always be applicable.

4.3. Accuracy measures

There exist various accuracy measures applicable to surrogate models, but there is no single, all-encompassing index. For that reason a number of methods were used including R^2 , \bar{R}^2 , Root-Mean-Squared-Deviation (RMSD) and residual plots. The indices are defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y^{(i)} - \bar{y})^2}{\sum_{i=1}^N (y^{(i)} - f^{(i)})^2}$$

$$\bar{R}^2 = 1 - (1 - R^2) \frac{N}{N - p}$$

$$RMSD = \sqrt{\frac{\sum_{i=1}^N (y^{(i)} - f^{(i)})^2}{N}}$$

$$e^{(i)} = y^{(i)} - f^{(i)}$$

where $y^{(i)}$ is the i^{th} data point, $f^{(i)}$ is an i^{th} model predicted value, \bar{y} is the empirical mean of data points, N is the number of data points, p is the

number of adjustable parameters, $e^{(i)}$ refers to residual for i^{th} data point and $i = 1, 2, \dots, N$. The first three measures are single number indices thus more convenient, but less informative than residual plots.

R^2 (coefficient of determination) is a measure indicating fit of a statistical model to data [76]. In essence, it compares the discrepancies between the predicted data and actual data with the discrepancies between the arithmetic average and actual data.

\bar{R}^2 (adjusted R^2) is R^2 , as described above, corrected for the number of fitted parameters relative to the number of data points. This measure cannot be greater than R^2 (for $N > p$) and it decreases as $N \rightarrow p$ indicating that the model overfits the data.

RMSD is the sample standard deviation of the differences between predicted values and observed values [77]. It is a good metric for comparing predictive power of different models for a particular variable (but not between the variables due to scale dependency).

5. Numerical experiments

5.1. Polynomial versus HDMR

\bar{R}^2 values were produced using the training set and are used to assess fit of the surrogates to the training data (data sampled from the process flow sheet model used for parameterisation), while RMSD and residual plots were produced using the test set (data sampled from process flow sheet model used for testing, but not parameterisation). Values sampled from entire domain of the input variables were used unless specified otherwise. Plots comparing surrogate types include polynomial fits of order 1 through 5 (labelled as P1

403 through P5) and HDMR fits with various constraints. Label H1 corresponds
 404 to a 1st order fit, H2a to a 2nd order without interactions, H2b to 2nd order
 405 with interactions and H10 to 10th order with 2nd order interactions. Note
 406 that HDMR fits may consist of terms with powers lower than specified, but
 407 in such a case it will be explicitly mentioned.

408 A number of different behaviours were observed in the study. Most sur-
 409rogate models achieved at least a reasonable fit regardless of the domain size,
 410 number of dimensions and according to \bar{R}^2 and RMSD. Neither R^2 nor \bar{R}^2
 411 can be used to effectively differentiate between the models as most achieve
 412 values in excess of 0.98 (for an example see Figure 5(a)). However, there
 413 is noticeable increase in \bar{R}^2 due to 2nd order interaction terms (P1 to P2
 414 and H2a to H2b). Also, it needs to be noted that the number of parame-
 415ters within HDMR fit increases far slower than within polynomial fit when
 416 high-dimensional problems are considered. Even the most extensive HDMR
 417 fit H10 had far fewer parameters than polynomial fits of order > 3 , as seen
 418 on plot 5(b).

419 RMSD provides a reasonable measure for comparing accuracy of models,
 420 as seen in Figure 6. Plots 6(a) and 6(b) suggest that polynomial fit of
 421 order 3 and HDMR fit H2b (marked by green squares) minimise RMSD
 422 and hence are the best fit for the duty of reactor 10D01 with respect to all
 423 11 inputs. The aforementioned plots (marked by orange triangles) also show
 424 that increasing order of polynomial fit lead to poorer predictive powers, most
 425 likely due to overfitting the training data. Similarly, HDMR fit H10 produces
 426 larger RMSD values than H2b. It can be seen that adding interaction (H2a
 427 to H2b) effect noticeably decreases RMSD in HDMR fitting.

428 Plots 6(c) and 6(d) show how RMSD changes as the domain size of inputs
429 increases. The former plot (for 5th order polynomial fit) shows an exponential
430 increase, while the latter (for HDMR fit H10) shows decrease of RMSD from
431 smallest to intermediate size and sharp increase from intermediate to largest
432 size.

433 Residual plots are the most informative form of error measurement as
434 they show the error size and distribution helping to understand whether the
435 fit captures the true nature of the data. In most cases data does not seem
436 to follow a polynomial relation resulting in non-random distribution of the
437 residuals. Figures 8 and 9 present residual plots for 11-dimensional surrogates
438 of heat duties of reactor 10D01 and heater 10E03. Comparison of plots in
439 Figures 8 and 7 shows that for output produced by surrogates with multiple
440 input variables the non-random features are much more difficult to identify.
441 Magnitude of the residuals in most cases is relatively small indicating strong
442 predictive powers of the fits. Comparing plots 7(c) and 8(c) reveals that
443 performance of polynomial fit of order 5 drops from being the best model
444 to the worst. Plots 8(b) and 8(d) show that even though HDMR fit H10
445 produced a higher RMSD, its residual plot is as good as seemingly better P3
446 fit. Those also confirm that P3 seems to be one of the best fits. Plot 8(c)
447 confirms that P5 fit exhibits relatively low accuracy, even worse than that of
448 a simple linear fit (see plot 8(a)).

449 5.2. Global sensitivity

450 Global sensitivities of the heat duties of all equipment under considera-
451 tion with respect to the 11 inputs produced by HDMR fitted over the entire
452 domain are summarised in Figures 10 and 11. It can be seen that in all cases
453 only 4 or fewer inputs have significant influence on a given output. Addi-
454 tionally, interaction terms have only minor effect on any one output. Heat
455 duty of each device is significantly affected by its own operating temperature
456 and operating temperature of a heating device directly upstream (given such
457 exists). While molar flow of oil, main feedstock of the process, has signifi-
458 cant effect on all heat duties (except that of the flash drum), molar flow of
459 methanol only affects heat duty of heater 10E02. This is because heat capac-
460 ity of oil is around 100 higher than that of methanol (1665.0 J/mol/K [78]
461 and 79.5 J/mol/K [79]) and only in the flash drum there is significantly more
462 methanol than oil.

463 Heat duty of heater 10E01 is primarily affected by its operating temper-
464 ature and molar flow and temperature of incoming oil. Heat duty of heater
465 10E02 is mostly affected by its operating temperature, operating tempera-
466 ture of reactor 10D01 and molar flow of oil and methanol. Heat duty of
467 heater 10E03 is primarily affected by its operating temperature, operating
468 temperature of decanter 10D02D and molar flow of oil. Heat duty of reactor
469 10D01 is primarily affected by its operating temperature, operating temper-
470 ature of heater 10E01 and molar flow of oil. Heat duty of flash drum 10D02
471 is primarily affected by its operating temperature and operating tempera-
472 ture of heater 10E02. Heat duty of decanter 10D02D is primarily affected by
473 its operating temperature, operating temperature of flash drum 10D02 and

474 molar flow of oil. Global sensitivities with respect to terms and variables not
475 mentioned here were negligible.

476 These observations show that when performing multi-dimensional anal-
477 ysis of heat duties within the system many terms in the surrogate models
478 can be ignored due to insignificant influence. Thus calculation complexity
479 and computational expense can be greatly reduced. Additionally, it shows
480 which inputs are important when heat duties of the equipment needs to be
481 controlled.

482 6. Conclusions

483 This paper presents results of parameterisation of typical input-output
484 relations within process flow sheet of a biodiesel plant and assesses param-
485 eterisation accuracy. The model under investigation includes a reaction and
486 separation steps with auxiliary equipment and was solved for steady-state
487 operation. Thus produced data was used to generate surrogate models de-
488 scribing relations between chosen inputs and outputs. A variety of scenarios
489 were considered: 1, 2, 6 and 11 input variables were changed simultaneously,
490 3 different domain sizes of the input variables were considered and 2 different
491 surrogate generation methods (polynomial and HDMR fitting). Each simu-
492 lation produced 400 points per input variable used for fitting and calculating
493 R^2 and \bar{R}^2 . Test sets of points (100 points per dimension) were generated
494 for calculating RMSD and residuals.

495 A number of different behaviours were observed in the study. Most surro-
496 gates achieved at least a reasonable fit regardless of the domain size, number
497 of dimensions and according to \bar{R}^2 and RMSD. Neither R^2 nor \bar{R}^2 could be

used to effectively differentiate between the models as most achieve values in excess of 0.98. Also, it needs to be noted that the number of parameters within HDMR fit increases far slower than within polynomial fit when high-dimensional problems are considered. The most extensive HDMR fit (H10) had far fewer parameters than polynomial fits of order > 4 . RMSD provides a reasonable measure for comparing accuracy of models. Fits P3 and H2b minimised RMSD and hence are the best fit for the duty of reactor 10D01 with respect to all 11 inputs. Increasing order of polynomial fit above 3 lead to poorer predictive powers due to overfitting the training data. RMSD increases exponentially for polynomial fits as the domain size of inputs increases. For fit H10 RMSD decreases from smallest to intermediate size and sharply increases from intermediate to largest size. Inclusion of 2nd order interaction terms accounted for a noticeable, but minor accuracy improvement in terms of \bar{R}^2 and RMSD. It was observed that non-random features in residual plots are much more difficult to identify when multiple inputs were considered. Higher order polynomial fits may not be suitable for describing high dimensional, chemical data. For example, performance of polynomial fit of order 5 drops from being the best model to the worst as dimensionality increases from 1 to 11.

Global sensitivities of the heat duties of all equipment under consideration with respect to the 11 inputs were produced by HDMR fitted over the entire domain. It was observed that in all cases only 4 or fewer inputs have significant influence on a given output. Interaction terms have only minor effect on any one output. Heat duty of each device is significantly affected by its own operating temperature and operating temperature of a heating

523 device directly upstream (given such exists). While molar flow of oil, main
524 feedstock of the process, has significant effect on all heat duties (except that
525 of the flash drum), molar flow of methanol only affects heat duty of heater
526 10E02. These observations show that when performing multi-dimensional
527 analysis of heat duties within the system many terms in the surrogate mod-
528 els can be ignored due to insignificant influence. Thus calculation complexity
529 and computational expense can be greatly reduced. Additionally, it shows
530 which inputs are important when heat duties of the equipment needs to be
531 controlled.

532 In the future a more complex chemical model should be considered as the
533 simulation used in this study was relatively simple. For example a number
534 of interconnected models forming a feedback loop necessitating coupling sur-
535rogate models and solving them simultaneously. In order to further the goal
536 of modelling eco-industrial parks chemical and electrical models and their
537 interactions should be considered.

538 **Acknowledgements**

539 This project is funded by the National Research Foundation (NRF),
540 Prime Minister’s Office, Singapore under its Campus for Research Excellence
541 and Technological Enterprise (CREATE) programme.

542 **References**

- 543 [1] G. Brundtland, M. Khalid, S. Agnelli, S. Al-Athel, B. Chidzero,
544 L. Fadika, *Our Common Future: The World Commission on Environ-*
545 *ment and Development*, Oxford University Press, Oxford, 1987.
- 546 [2] C. Hoffman, *The Industrial Ecology of Small and Intermediate-sized*
547 *Technical Companies: Implications for Regional Economic Develop-*
548 *ment*, Technical Report 197411, Texas University, USA, 1971.
- 549 [3] C. Watanabe, *Industrial ecology: Introduction of Ecology into Industrial*
550 *Policy*, Technical Report, Ministry of International Trade and Industry
551 (MITI), Tokyo, 1972.
- 552 [4] B. Allenby, *Journal of Cleaner Production* 12 (2004) 833–839. doi:10.
553 1016/j.jclepro.2004.02.010.
- 554 [5] B. Allenby, *Progress in Industrial Ecology - An International Journal*
555 1-2 (2006) 28–40. doi:10.1504/PIE.2006.010039.
- 556 [6] M. Chertow, *Annual Review of Energy and Environment* 25 (2000) 313–
557 337. doi:10.1146/annurev.energy.25.1.313.
- 558 [7] M. Pan, J. Sikorski, C. A. Kastner, J. Akroyd, S. Mosbach, R. Lau,
559 M. Kraft, *Energy Procedia* 150 (2015). doi:10.1016/j.egypro.2015.
560 07.313.
- 561 [8] P. Desrochers, *Journal of Industrial Ecology* 5 (2001) 29–44. doi:10.
562 1162/10881980160084024.

- [9] J. Ehrenfeld, N. Gertler, *Journal of Industrial Ecology* 1 (1997) 67–79.
doi:10.1162/jiec.1997.1.1.67.
- [10] M. Boix, L. Montastruc, C. Azzaro-Pantel, S. Domenech, *Journal of Clean Production* 87 (2015) 303–317. doi:10.1016/j.jclepro.2014.09.032.
- [11] E. Cimren, J. Fiksel, M. Posner, K. Sikdar, *Journal of Industrial Ecology* 15 (2012) 315–332. doi:10.1111/j.1530-9290.2010.00310.x.
- [12] I. Kantor, M. Fowler, A. Elkamel, *International Journal of Hydrogen Energy* 37 (2012) 5347–5359. doi:10.1016/j.ijhydene.2011.08.084.
- [13] S. Keckler, D. Allen, *Journal of Industrial Ecology* 2 (1999) 79–92. doi:10.1162/jiec.1998.2.4.79.
- [14] Z. W. Liao, J. T. Wu, B. B. Jiang, J. D. Wang, Y. R. Yang, *Industrial & Engineering Chemistry Research* 46 (2007) 4954–4963. doi:10.1021/ie061299i.
- [15] M. Karlsson, *Applied Energy* 88 (2011) 577–589. doi:10.1016/j.apenergy.2010.08.021.
- [16] A. Forrester, A. Sobester, A. Keane, *Engineering Design via Surrogate Modelling: A Practical Guide*, 1st ed., Wiley, 2008.
- [17] T. W. Simpson, J. D. Peplinski, P. N. Koch, J. K. Allen, *Engineering with Computers* 17 (2001) 129–150. doi:10.1007/PL00007198.
- [18] D. Ruppert, M. P. Wand, R. J. Carroll, *Semiparametric Regression*, Cambridge University Press, Cambridge, 2003.

- 585 [19] J. P. C. Kleijnen, *European Journal of Operational Research* 192 (2009)
586 707–716. doi:10.1016/j.ejor.2007.10.013.
- 587 [20] S. B. Crary, *Analog Integrated Circuits and Signal Processing* 32 (2002)
588 7–16. doi:10.1023/A:1016063422605.
- 589 [21] Y. Reich, S. Barai, *Artificial Intelligence in Engineering* 13 (1999) 257–
590 272. doi:10.1016/S0954-1810(98)00021-1.
- 591 [22] M. H. Bakr, J. W. Bandler, K. Madsen, J. Soandergaard, *Optimization
592 and Engineering* 1 (2000) 241–276. doi:10.1023/A:1010000106286.
- 593 [23] R. Jin, X. Du, W. Chen, *Structural and Multidisciplinary Optimization*
594 25 (2003) 99–116. doi:10.1007/s00158-002-0277-0.
- 595 [24] W. A. Gough, W. J. Welch, *Journal of Marine Research* 52 (1994) 773–
596 796. doi:10.1357/0022240943076911.
- 597 [25] P. Geyera, A. Schlueter, *Applied Energy* 119 (2014) 537–556. doi:10.
598 1016/j.apenergy.2013.12.064.
- 599 [26] D. L. Knill, A. A. Giunta, C. A. Baker, B. Grossman, W. H. Mason,
600 R. T. Haftka, L. T. Watson, *Journal of Aircraft* 36 (1999) 75–86. doi:10.
601 2514/2.2415.
- 602 [27] S. B. Crary, P. Cousseau, D. Armstrong, D. M. Woodcock, E. H. Mok,
603 O. Dubochet, P. Lerch, P. Renaud, *Computer Modeling in Engineering
604 and Sciences* 1 (2000) 127–140. doi:10.3970/cmcs.2000.001.127.
- 605 [28] N. Chen, K. Wang, C. Xiao, J. Gong, *Environmental Modelling & Soft-
606 ware* 54 (2014) 222–237. doi:10.1029/93JC02564.

- [29] P. Azadi, G. Brownbridge, S. Mosbach, O. R. Inderwildi, M. Kraft, Energy Procedia 61 (2014) 2767–2770. doi:10.1016/j.egypro.2014.12.302.
- [30] W. L. Chapman, W. J. Welch, K. P. Bowman, J. Sacks, J. E. Walsh, Journal of Geophysical Research 99 (1994) 919–935. doi:10.1029/93JC02564.
- [31] W. J. Menz, G. Brownbridge, , M. Kraft, Journal of Aerosol Science 76 (2014) 188–199. doi:10.1016/j.jaerosci.2014.06.011.
- [32] J. C. Jouhauda, P. Sagautb, M. Montagnaca, J. Laurenceaua, Computers & Fluids 36 (2007) 520–529. doi:10.1016/j.compfluid.2006.04.001.
- [33] C. A. Kastner, A. Braumann, P. L. W. Man, S. Mosbach, G. Brownbridge, J. Akroyd, M. Kraft, C. Himawan, Chemical Engineering Science 89 (2013) 244–257. doi:10.1016/j.ces.2012.11.027.
- [34] I. F. Bailleul, P. L. Man, M. Kraft, Society for Industrial and Applied Mathematics Journal on Numerical Analysis 48 (2010) 1064–1086. doi:10.1137/090758234.
- [35] A. Braumann, M. Kraft, P. Mort, Powder Technology 197 (2010) 196–210. doi:10.1016/j.powtec.2009.09.014.
- [36] M. C. Bernardo, R. Buck, L. Liu, W. A. Nazaret, J. Sacks, W. J. Welch, IEEE Transactions on Computer-Aided Design 11 (1992) 361–372. doi:10.1109/43.124423.

- [37] R. Aslett, R. J. Buck, S. G. Duvall, J. Sacks, W. J. Welch, *Journal of the Royal Statistical Society: Series C* 47 (1998) 31–48. doi:10.1111/1467-9876.00096.
- [38] E. Roux, P. Bouchard, *Journal of Materials Processing Technology* 213 (2013) 1038–1047. doi:10.1016/j.jmatprotec.2013.01.018.
- [39] J. A. Caballero, I. E. Grossmann, *AIChE Journal* 54 (2008) 2633–2650. URL: <http://onlinelibrary.wiley.com/doi/10.1002/aic.11579/full>. doi:10.1002/aic.11579.
- [40] M. M. F. Hasan, R. C. Baliban, J. A. Elia, C. A. Floudas, *Industrial and Engineering Chemistry Research* 51 (2012) 15665–15682. doi:10.1021/ie301572n.
- [41] E. L. First, M. M. F. Hasan, C. A. Floudas, *AIChE Journal* 60 (2014) 1767–1785. URL: <http://onlinelibrary.wiley.com/doi/10.1002/aic.14441/full>. doi:10.1002/aic.14441.
- [42] M. M. F. Hasan, E. L. First, C. a. Floudas, *Physical chemistry chemical physics : PCCP* 15 (2013) 17601–18. URL: <http://www.ncbi.nlm.nih.gov/pubmed/24037279>. doi:10.1039/c3cp53627k.
- [43] A. Nuchitprasittichai, S. Cremaschi, *Industrial & Engineering Chemistry Research* 52 (2013) 10236–10243. URL: <http://pubs.acs.org/doi/abs/10.1021/ie3029366>. doi:10.1021/ie3029366.
- [44] F. Boukouvala, M. G. Ierapetritou, *Journal of Pharmaceutical Innovation* 8 (2013) 131–145. doi:10.1007/s12247-013-9154-1.

- 651 [45] I. Fahmi, S. Cremaschi, Computers & Chemical Engineering 46 (2012)
652 105–123. URL: [http://www.sciencedirect.com/science/article/
653 pii/S0098135412001822](http://www.sciencedirect.com/science/article/pii/S0098135412001822). doi:10.1016/j.compchemeng.2012.06.006.
- 654 [46] C. A. Henao, C. T. Maravelias, AIChE Journal 57 (2011) 1216–
655 1232. URL: [http://onlinelibrary.wiley.com/doi/10.1002/aic.
656 12341/full](http://onlinelibrary.wiley.com/doi/10.1002/aic.12341/full). doi:10.1002/aic.12341.
- 657 [47] L. Kong, S. Murat Sen, C. A. Henao, J. A. Dumesic, C. T. Mar-
658 avelias, Computers and Chemical Engineering (2016). doi:10.1016/j.
659 compchemeng.2016.02.013.
- 660 [48] AspenTech, aspentech - Aspen Plus v8.6, 2015. URL: [http://www.
661 aspentech.com/products/engineering/aspen-plus/](http://www.aspentech.com/products/engineering/aspen-plus/), date accessed:
662 12.01.2016.
- 663 [49] CMCL Innovations, Model Development Suite (MoDS), 2015. URL:
664 <http://www.cmclinnovations.com/mods/>, date accessed: 12.01.2015.
- 665 [50] A. Braumann, P. L. Man, M. Kraft, Industrial and Engineering Chem-
666 istry Research 49 (2010) 428–438. doi:10.1021/ie901230u.
- 667 [51] P. L. Man, A. Braumann, M. Kraft, Industrial and Engineering Chem-
668 istry Research 65 (2010) 4038–4045. doi:10.1016/j.ces.2010.03.042.
- 669 [52] A. Braumann, P. Man, M. Kraft, AIChE Journal 57 (2011) 3105–3121.
670 doi:10.1002/aic.12526.
- 671 [53] S. Shekar, A. J. Smith, A. Braumann, P. L. Man, M. Kraft, Journal of

- 672 Aerosol Science 44 (2012) 93–98. doi:10.1016/j.jaerosci.2011.09.
673 004.
- 674 [54] W. J. Menz, S. Shekar, G. Brownbridge, S. Mosbach, R. Krmer,
675 W. Peukert, M. Kraft, Journal of Aerosol Science 44 (2012) 46–61.
676 doi:10.1016/j.jaerosci.2011.10.005.
- 677 [55] S. Shekar, M. Sander, R. Shaw, A. J. Smith, A. Braumann, M. Kraft,
678 Chemical Engineering Science 70 (2012) 54–66. doi:10.1016/j.ces.
679 2011.06.010.
- 680 [56] W. J. Menz, M. Kraft, Combustion and Flame 160 (2013) 947–958.
681 doi:10.1016/j.combustflame.2013.01.014.
- 682 [57] P. Azadi, G. Brownbridge, S. Mosbach, A. J. Smallbone, A. Bhave, O. R.
683 Inderwildi, M. Kraft, Applied Energy 113 (2014) 1632–1644. doi:10.
684 1016/j.apenergy.2013.09.027.
- 685 [58] G. Brownbridge, P. Azadi, A. J. Smallbone, A. Bhave, B. J. Taylor,
686 M. Kraft, Bioresource Technology 151 (2014) 166–173. doi:10.1016/j.
687 biortech.2013.10.062.
- 688 [59] A. Vikhansky, M. Kraft, Journal of Computational Physics 200 (2004)
689 50–59. doi:10.1016/j.jcp.2004.03.006.
- 690 [60] A. Vikhansky, M. Kraft, Chemical Engineering Science 61 (2006) 4966–
691 4972. doi:10.1016/j.ces.2006.03.009.
- 692 [61] A. J. Smallbone, A. Bhave, A. Braumann, M. Kraft, A. Dris,

- 693 R. McDavid SAE Paper No. 2010-01-0152 (2010). doi:10.4271/
694 2010-01-0152.
- 695 [62] G. Brownbridge, A. J. Smallbone, W. Phadungsukanan, M. Kraft,
696 B. Johansson SAE Paper No. 2011-01-0237 (2011). doi:10.4271/
697 2011-01-0237.
- 698 [63] J. E. Etheridge, S. Mosbach, M. Kraft, H. Wu, N. Collings SAE Paper
699 No. 2010-01-1241 (2010). doi:10.4271/2010-01-1241.
- 700 [64] A. M. Aldawood, S. Mosbach, M. Kraft, A. A. Amer SAE Paper No.
701 2011-01-1783 (2011). doi:10.4271/2011-01-1783.
- 702 [65] P. Azadi, G. Brownbridge, I. Kemp, S. Mosbach, J. S. Dennis, M. Kraft,
703 ChemCatChem 7 (2015) 137–143. doi:10.1002/cctc.201402662.
- 704 [66] E. K. Y. Yapp, R. I. A. Patterson, J. Akroyd, S. Mosbach, E. M. Adkins,
705 J. H. Miller, M. Kraft, Combustion and Flame (2016). doi:10.1016/j.
706 combustflame.2016.01.033.
- 707 [67] Microsoft, COM technical overview, 2015. URL: [https://msdn.](https://msdn.microsoft.com/en-us/windows/desktop)
708 [microsoft.com/en-us/windows/desktop](https://msdn.microsoft.com/en-us/windows/desktop), date accessed: 02.06.2015.
- 709 [68] I. M. Sobol, USSR Computational Mathematics and Mathematical
710 Physics 7 (1967) 86–112. doi:10.1016/0041-5553(67)90144-9.
- 711 [69] S. Joe, F. Y. Kuo, SIAM Journal on Scientific Computing 30 (2008)
712 2635–2654. doi:10.1137/070709359.

- 713 [70] R. H. Myers, D. C. Montgomery, C. M. Anderson-Cook, Response Sur-
714 face Methodology: Process and Product Optimization Using Designed
715 Experiments, 3rd ed., Wiley-Blackwell, 2009.
- 716 [71] I. M. Sobol, Matematicheskoe Modelirovanie 2 (1990) 112–118.
- 717 [72] H. Rabitz, Ö. F. Alis, Journal of Mathematical Chemistry 25 (1999)
718 197–233. doi:10.1023/A:1019188517934.
- 719 [73] G. Li, S.-W. Wang, H. Rabitz, The Journal of Physical Chemistry A
720 106 (2002) 8721–8733. doi:10.1021/jp014567t.
- 721 [74] H. M. Wainwright, S. Finsterle, Y. Jung, Q. Zhou, J. T. Birkholzer,
722 Computers and Geosciences 65 (2014) 84–94. doi:10.1016/j.cageo.
723 2013.06.006.
- 724 [75] M. Baran, L. Bieniasz, Applied Mathematics and Computation 258
725 (2015) 206–219. doi:10.1016/j.amc.2015.02.007.
- 726 [76] N. R. Draper, H. Smith, Applied Regression Analysis, 3rd ed., Wiley-
727 Interscience, 1998.
- 728 [77] R. J. Hyndman, A. B. Koehler, International Journal of Forecasting 22
729 (2006) 679–688. doi:10.1016/j.ijforecast.2006.03.001.
- 730 [78] V. Filatov, V. Afanas'ev, Khimiya i Tekhnoliya Vody 35 (1992) 97–100.
- 731 [79] B. Freedman, M. Bagby, H. Khoury, Journal of the American Oil
732 Chemists' Society 66 (1989) 595–596. doi:10.1007/BF02885455.

733 List of Figures

734	1	<i>Framework of EIP modelling based on Industry 4.0. Adopted</i>	
735		<i>from Pan et al. [7].</i>	38
736	2	Graphical representation of the process flow sheet model of a	
737		biodiesel production line.	39
738	3	Plots of heat duties of various equipment against molar flow	
739		of tripalmitin oil.	40
740	4	<i>Model Development Suite work flow.</i>	41
741	5	Plots of RMSD and number of parameters for the considered	
742		surrogates produced for heat duty of reactor 10D01 with re-	
743		spect to all 11 inputs. Labels P1 through P5 correspond to	
744		polynomial fits of order 1 through 5. Label H1 corresponds to	
745		a 1 st order fit, H2a to a 2 nd order without interactions, H2b	
746		to 2 nd order with interactions and H10 to 10 th order with 2 nd	
747		order interactions.	42
748	6	Plots of RMSD for the considered surrogates and domain sizes	
749		produced for heat duty of reactor 10D01 with respect to all	
750		11 inputs. Labels P1 through P5 correspond to polynomial	
751		fits of order 1 through 5. Label H1 corresponds to a 1 st or-	
752		der fit, H2a to a 2 nd order without interactions, H2b to 2 nd	
753		order with interactions and H10 to 10 th order with 2 nd order	
754		interactions. Green squares indicate models (one per type)	
755		with lowest RMSD, while red triangles indicate models (one	
756		per type) with suffering most from overfitting.	43
757	7	Plot of residuals against molar flow of tripalmitin oil for heat	
758		duty of reactor 10D01 produced for 1 input.	44
759	8	Plot of residuals against molar flow of tripalmitin oil for heat	
760		duty of reactor 10D01 produced for 11 inputs.	45
761	9	Plot of residuals against molar flow of tripalmitin oil for heat	
762		duty of heater 10E03 produced for 11 inputs.	46
763	10	Global sensitivities produced by 11-dimensional HDMR fit	
764		over the entire domain.	47
765	11	Global sensitivities produced by 11-dimensional HDMR fit	
766		over the entire domain.	48

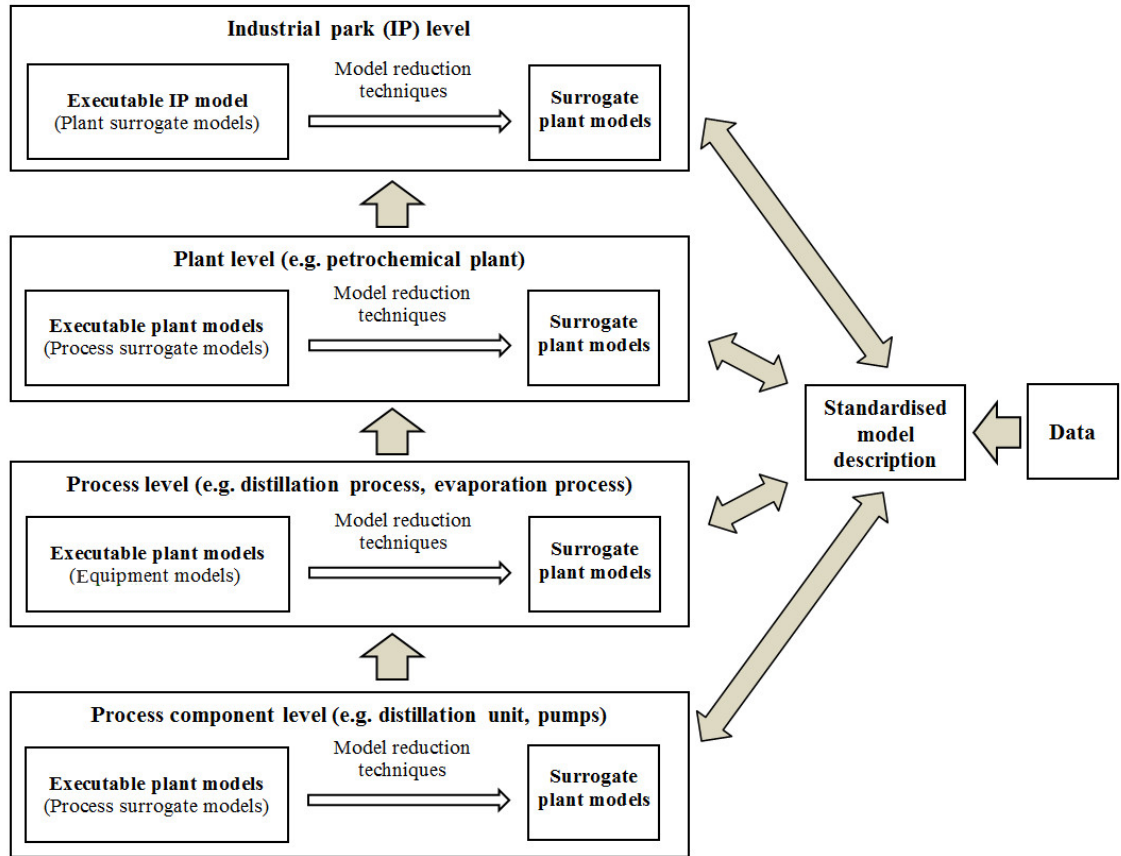


Figure 1: *Framework of EIP modelling based on Industry 4.0. Adopted from Pan et al. [7].*

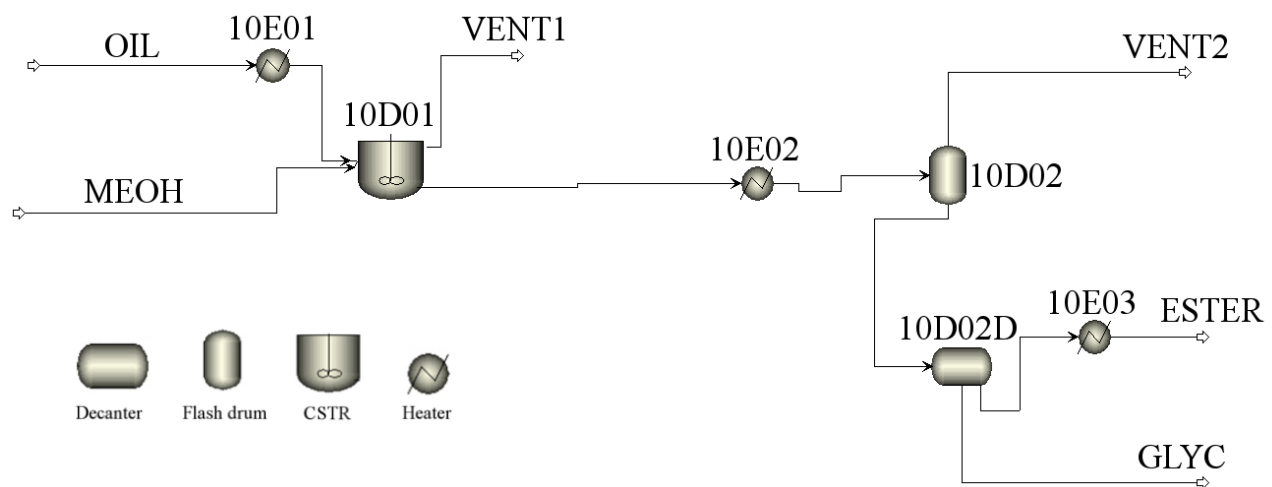
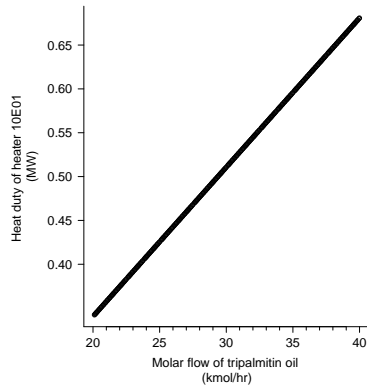
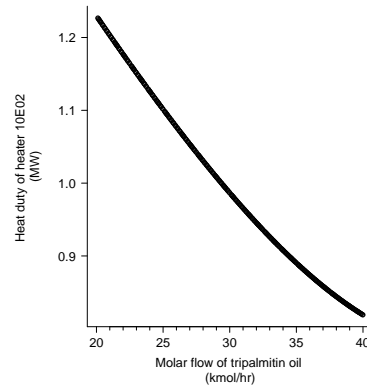


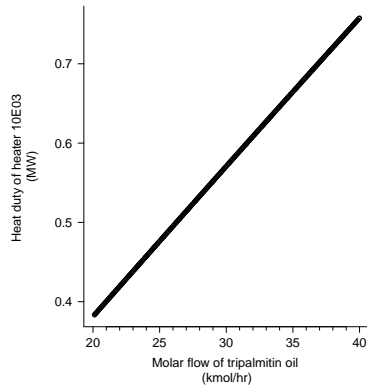
Figure 2: Graphical representation of the process flow sheet model of a biodiesel production line.



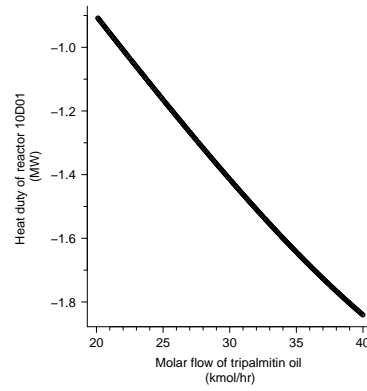
(a) Heater 10E01



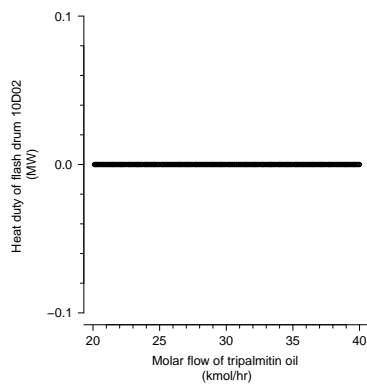
(b) Heater 10E02



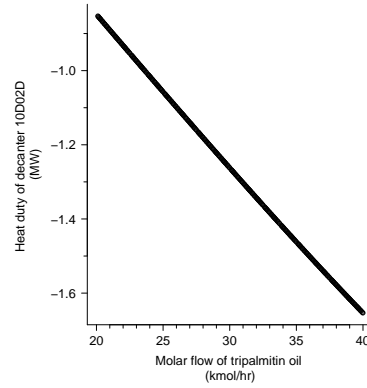
(c) Heater 10E03



(d) Reactor 10D01



(e) Flash drum 10D02



(f) Decanter 10D02D

Figure 3: Plots of heat duties of various equipment against molar flow of tripalmitin oil.

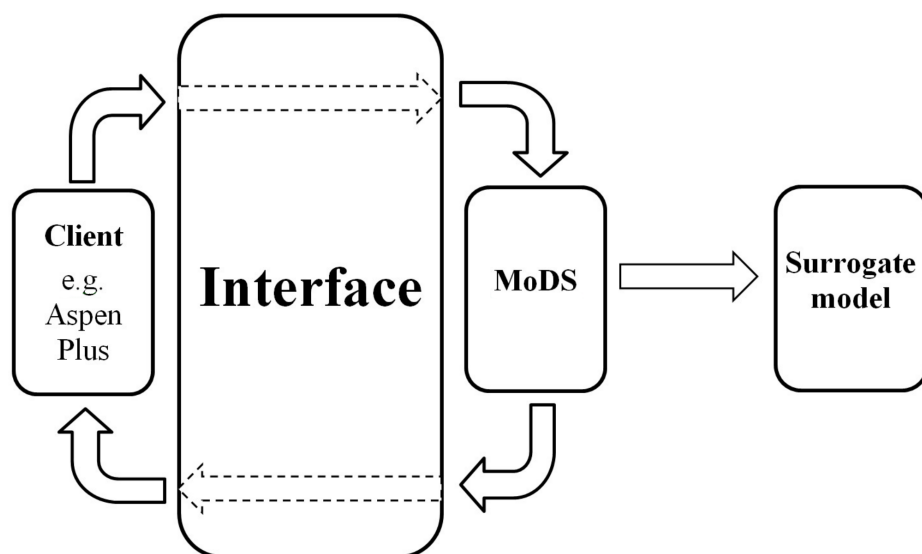
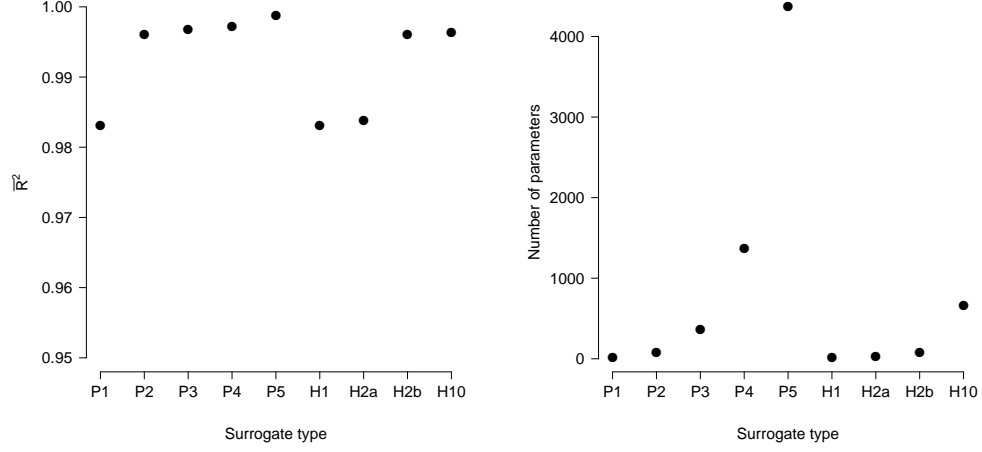
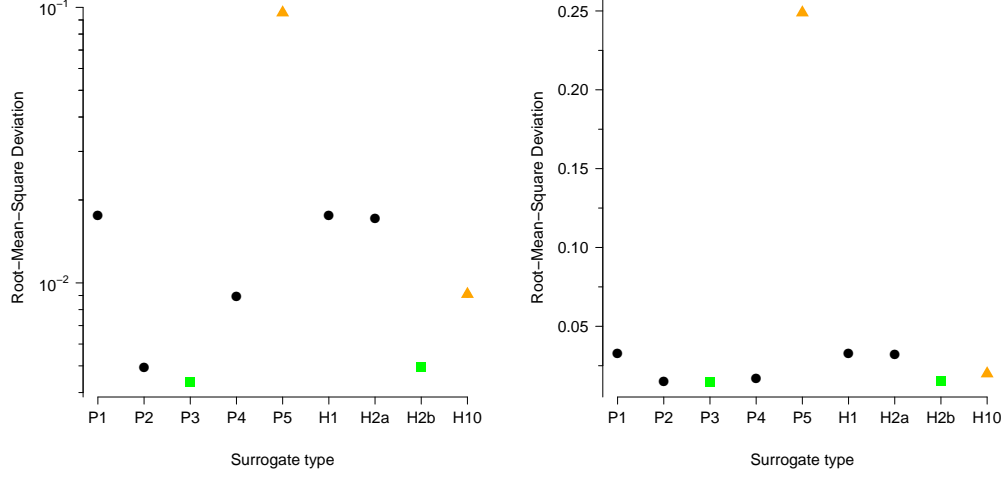


Figure 4: *Model Development Suite work flow.*

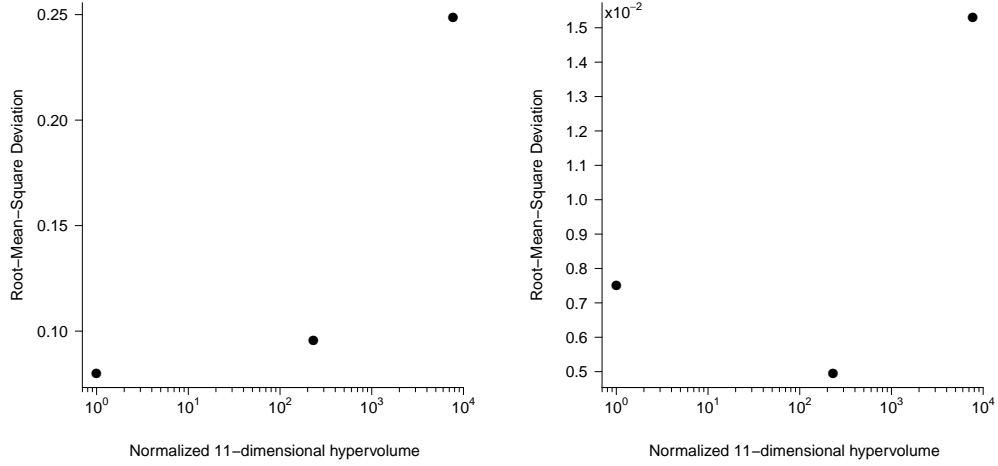


(a) Plot of \bar{R}^2 for the considered surrogates. (b) Plot of number of parameters for the considered surrogates.

Figure 5: Plots of RMSD and number of parameters for the considered surrogates produced for heat duty of reactor 10D01 with respect to all 11 inputs. Labels P1 through P5 correspond to polynomial fits of order 1 through 5. Label H1 corresponds to a 1st order fit, H2a to a 2nd order without interactions, H2b to 2nd order with interactions and H10 to 10th order with 2nd order interactions.

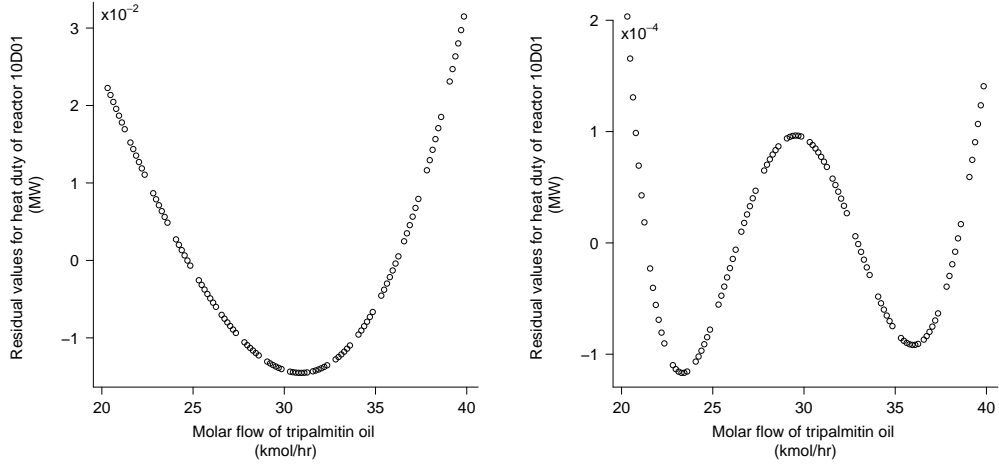


(a) RMSD for the considered surrogates for medium domain size. (b) RMSD for the considered surrogates.

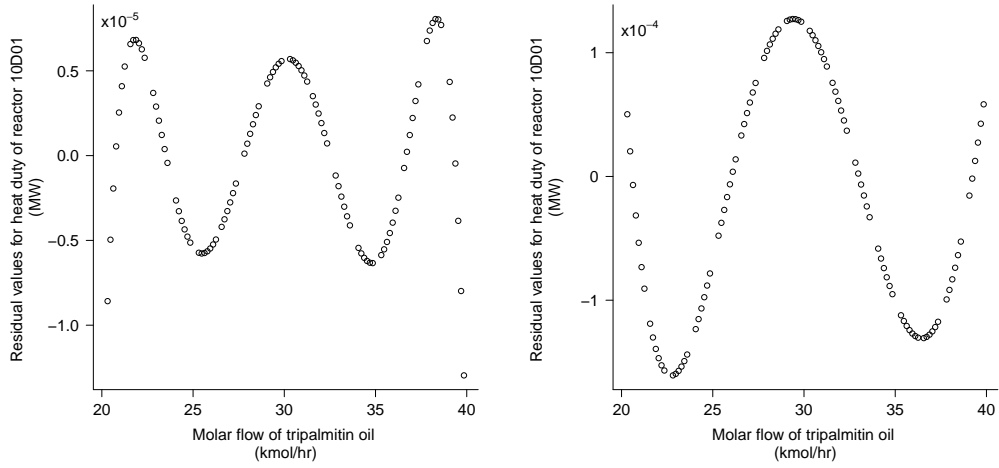


(c) RMSD against domain sizes for polynomial fit of order 5 (for boundaries see Table 1). (d) RMSD against domain sizes for HDMR fit (for boundaries see Table 1).

Figure 6: Plots of RMSD for the considered surrogates and domain sizes produced for heat duty of reactor 10D01 with respect to all 11 inputs. Labels P1 through P5 correspond to polynomial fits of order 1 through 5. Label H1 corresponds to a 1^{st} order fit, H2a to a 2^{nd} order without interactions, H2b to 2^{nd} order with interactions and H10 to 10^{th} order with 2^{nd} order interactions. Green squares indicate models (one per type) with lowest RMSD, while red triangles indicate models (one per type) with suffering most from overfitting.

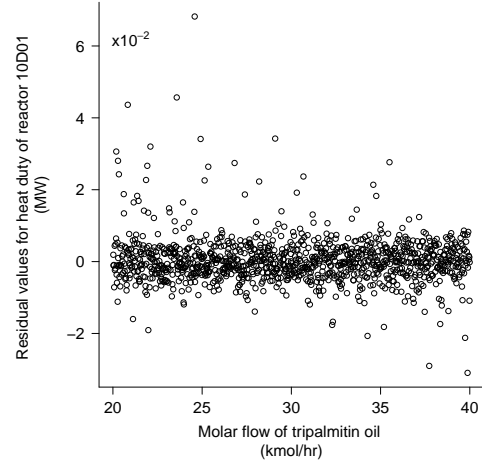
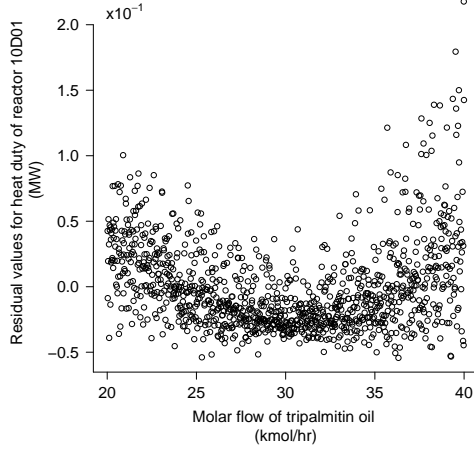


(a) Plot of residuals for 1st order polynomial fit. (b) Plot of residuals for 3rd order polynomial fit.

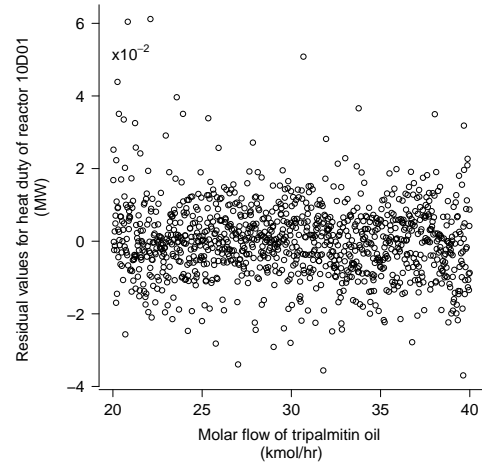
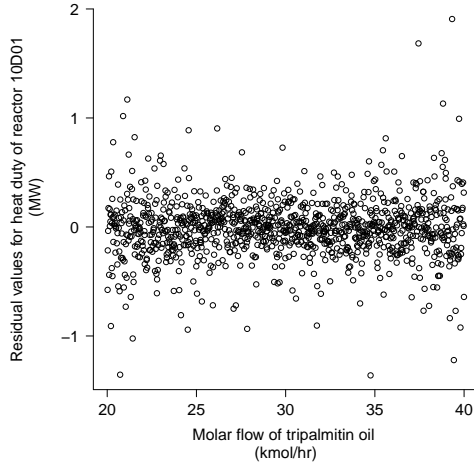


(c) Plot of residuals for 5th order polynomial fit. (d) Plot of residuals for HDMR fit H10 (3rd order polynomial).

Figure 7: Plot of residuals against molar flow of tripalmitin oil for heat duty of reactor 10D01 produced for 1 input.

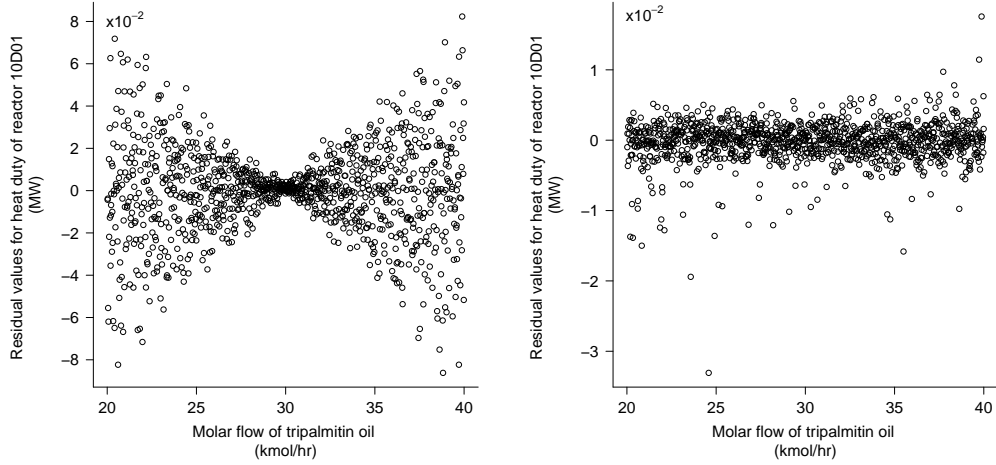


(a) Plot of residuals for 1st order polynomial fit. (b) Plot of residuals for 3rd order polynomial fit.

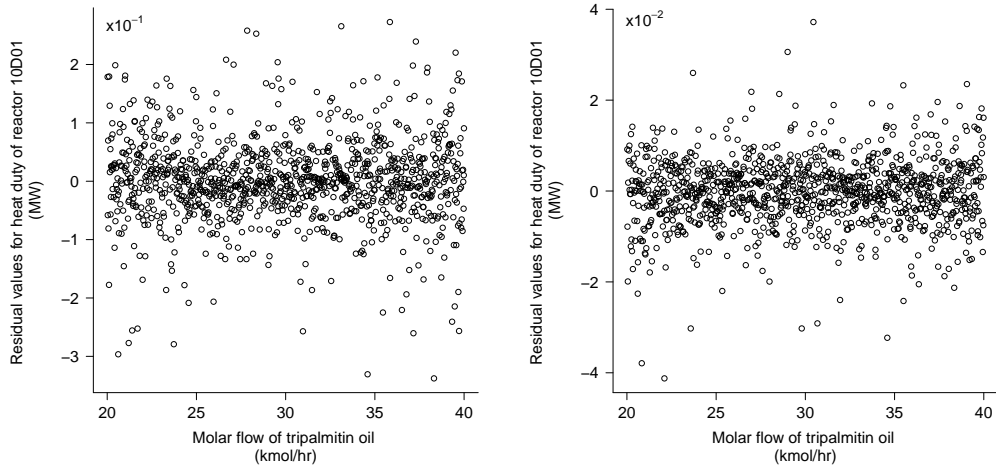


(c) Plot of residuals for 5th order polynomial fit. (d) Plot of residuals for HDMR fit H10.

Figure 8: Plot of residuals against molar flow of tripalmitin oil for heat duty of reactor 10D01 produced for 11 inputs.

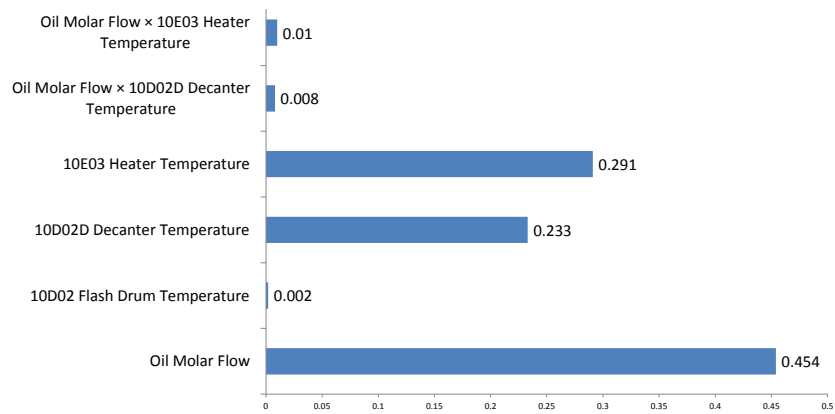


(a) Plot of residuals for 1st order polynomial fit. (b) Plot of residuals for 3rd order polynomial fit.

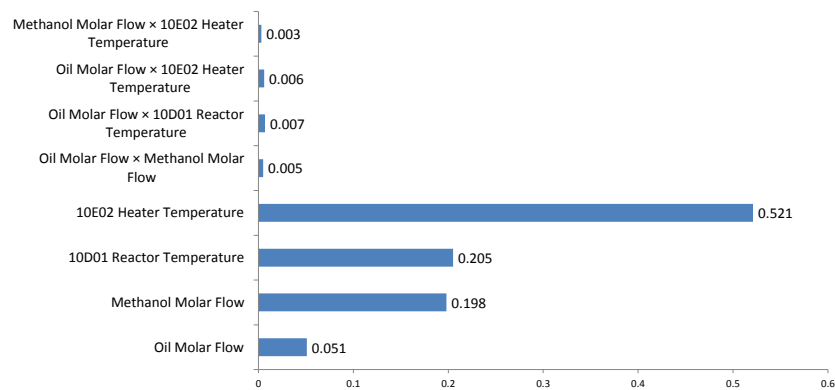


(c) Plot of residuals for 5th order polynomial fit. (d) Plot of residuals for HDMR fit H10.

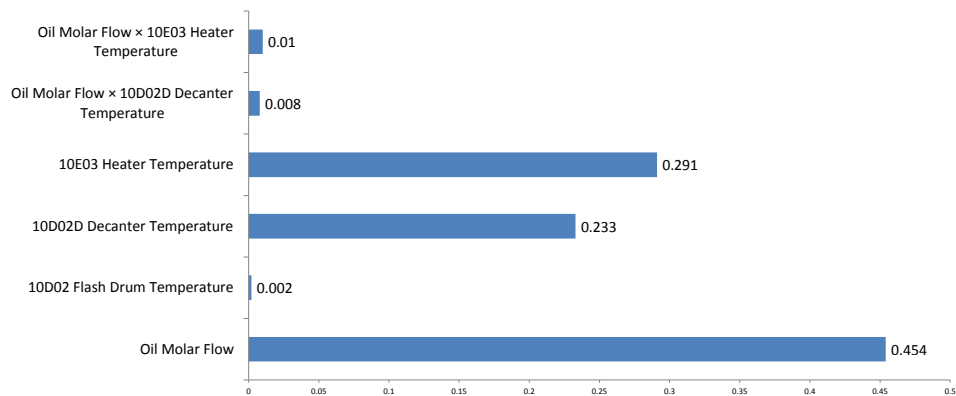
Figure 9: Plot of residuals against molar flow of tripalmitin oil for heat duty of heater 10E03 produced for 11 inputs.



(a) 10E01 Heater - Heat Duty.

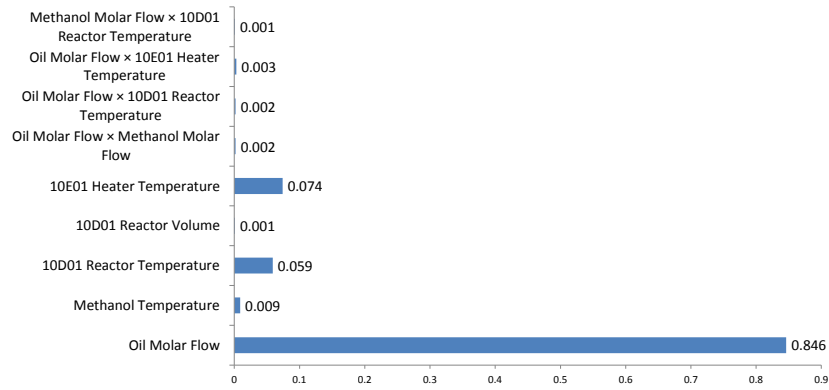


(b) 10E02 Heater - Heat Duty.

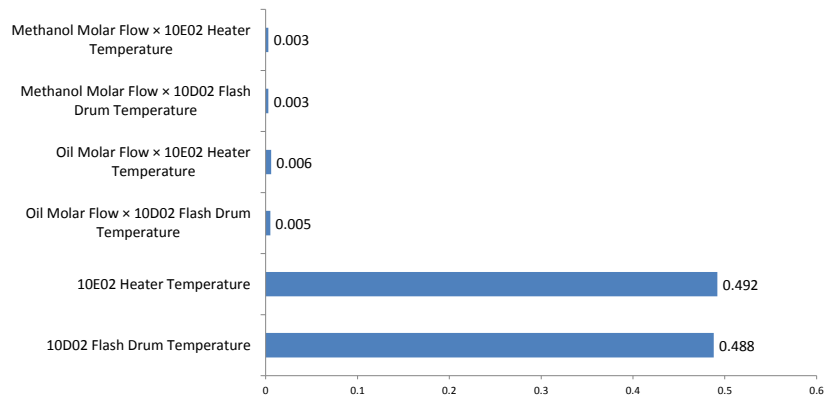


(c) 10E03 Heater - Heat Duty.

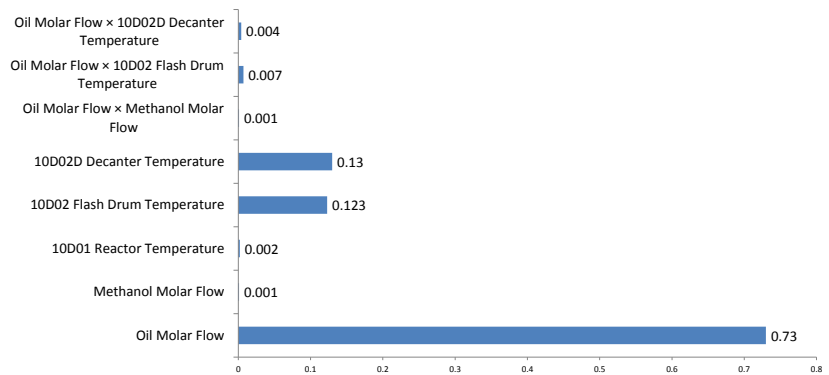
Figure 10: Global sensitivities produced by 11-dimensional HDMR fit over the entire domain.



(a) 10D01 Reactor - Heat Duty.



(b) 10D02 Flash Drum - Heat Duty.



(c) 10D02D Decanter - Heat Duty.

Figure 11: Global sensitivities produced by 11-dimensional HDMR fit over the entire domain.