Parameterisation of a biodiesel plant process flow sheet model

Janusz J. Sikorski[a], George Brownbridge[b], Sushant S. Garud[c], Sebastian Mosbach[a], Iftekar A. Karimi[c], Markus Kraft*[a,d]

[a]*Department of Chemical Engineering and Biotechnology, University of Cambridge, New Museums Site, Pembroke Street, Cambridge, CB2 3RA United Kingdom*
*corresponding author**
E-mail: mk306@cam.ac.uk
[b]*CMCL Innovations, Sheraton House, Castle Park, Cambridge, CB3 0AX, United Kingdom*
[c]*Department of Chemical and Biomolecular Engineering, National University of Singapore, 4 Engineering Drive 4, Singapore 117585*
[d]*School of Chemical and Biomedical Engineering, Nanyang Technological University, 62 Nanyang Drive, Singapore, 637459*

Abstract

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

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

1. Introduction

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

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

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

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

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

Holistic modelling of complex, highly interconnected networks is a non-trivial and expensive task, especially for EIPs which include numerous physical models of disparate processes. That is why many studies apply mathematical optimisation to simplified models of individual aspects of the parks.

The limitations of this approach may be overcome by exploiting key features of the concept of Industry 4.0 [7]: creation of virtual copies of the physical world and the ability of industrial components to communicate with each other. Those virtual copies could be surrogate models of physical models produced for a predefined range of inputs. Developing a virtual system primarily based on surrogate models would significantly reduce required computation time and storage space and allow for dynamic modelling and studies otherwise impossible to conduct. Figure 1 presents a framework of EIP modelling

based on Industry 4.0.

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

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

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

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

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

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

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

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

2. Model

2.1. Aspen Plus V8.6

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

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

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

2.2. Biodiesel plant simulation

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

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

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

3. Parameterisation

3.1. Model Development Suite

Model Development Suite (MoDS) [49] is an advanced software tool designed to analyse black-box models (e.g. executables, batch scripts). It includes a broad range of tools such as data-driven modelling, multi-objective optimisation, generation of surrogate models, data standardisation and visualisation, 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)

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

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

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

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

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

Aspen Plus.

3.3. Data harvest and surrogate generation

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

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

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

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

3.4. Sampling

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

4. Implementation

4.1. Polynomial response surfaces

Polynomial response surfaces are a subset of response surface methodology, a group of mathematical and statistical techniques designed to facilitate empirical model building [70]. Polynomials of a predefined degree are optimized to describe an unknown relation between independent variables (input variables) and responses (output variables). Input and output data sets are obtained via series of tests, an experiment, in which the input variables are modified in order to study the changes in the output responses. As the number of adjustable coefficients in a polynomial surrogate increases combinatorially with its order and number of variables so does the minimum number of data points required to produce it. Hence applying high-order polynomials to problems with many inputs may lead to overfitting and hence poorer predictive power. Generally, overfitting occurs when a model describes features specific to the data set on which it is trained such as random error or noise. For deterministic computer experiments those are not an issue, but an overfitted model will suffer from having an exaggerated set of coefficients providing no intuitive insight into nature of the relationship under consideration and from introducing irrelevant nonlinearity.

General linear least-squares fit

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

are given. Coefficients of the polynomial are given by

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

with

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

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

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

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

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

which can be abbreviated to x^p and is of order $|p|$. Thus the polynomial f_β can be written as

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

where the βs denote the coefficients of the individual terms and k corresponds to the polynomial order.

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

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

Rearranging yields

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

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

4.2. High Dimensional Model Representation

High Dimensional Model Representation (HDMR) is a finite expansion for a given multivariable function as described by Sobol [71], Rabitz and Alış [72]. It allows for readily extracting global sensitivities with respect to the independent variables by calculating them from the coefficients of a HDMR surrogate. 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.

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

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

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

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

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

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

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

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

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

$$\int \phi_k\left(x_i\right) \mathrm{d}x_i = 0 \tag{3a}$$

$$\int \phi_k\left(x_i\right) \phi_l\left(x_i\right) \mathrm{d}x_i = \delta_{kl}\,. \tag{3b}$$

This leads the following equations for the coefficients:

$$f_0 = \int f(x)\mathrm{d}x\,, \tag{4a}$$

$$\alpha_{i,k} = \int f(x)\phi_k\left(x_i\right) \mathrm{d}x\,, \tag{4b}$$

$$\beta_{ij,kl} = \int f(x)\phi_k\left(x_i\right) \phi_l\left(x_j\right) \mathrm{d}x\,, \tag{4c}$$

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

$$\sigma^2_{\overline{y}} = \sum_{i=1}^{N_x} \int_{-1}^{1} f_i^2\left(x_i\right) \mathrm{d}x_i + \sum_{i=1}^{N_x} \sum_{j=i+1}^{N_x} \int_{-1}^{1} \int_{-1}^{1} f_{ij}^2\left(x_i, x_j\right) \mathrm{d}x_i \mathrm{d}x_j \tag{5a}$$

$$= \sum_{i=1}^{N_x} \sigma^2_{\overline{y},i} + \sum_{i=1}^{N_x} \sum_{j=i+1}^{N_x} \sigma^2_{\overline{y},ij}\,. \tag{5b}$$

The sensitivities, S_i and S_{ij}, can then be calculated by dividing by the total variance $\sigma^2_{\overline{y}}$ to get

$$S_i = \frac{\sigma^2_{\overline{y},i}}{\sigma^2_{\overline{y}}} \quad \text{and} \quad S_{ij} = \frac{\sigma^2_{\overline{y},ij}}{\sigma^2_{\overline{y}}}\,. \tag{6}$$

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

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

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

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

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 $i = 1, 2, .., N$. 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 $N \to p$ 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

through P5) and HDMR fits with various constraints. 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. Note that HDMR fits may consist of terms with powers lower than specified, but in such a case it will be explicitly mentioned.

A number of different behaviours were observed in the study. Most surrogate models achieved at least a reasonable fit regardless of the domain size, number of dimensions and according to $\bar{R}^2$ and RMSD. Neither R^2 nor $\bar{R}^2$ can be used to effectively differentiate between the models as most achieve values in excess of 0.98 (for an example see Figure 5(a)). However, there is noticeable increase in $\bar{R}^2$ due to 2nd order interaction terms (P1 to P2 and H2a to H2b). 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. Even the most extensive HDMR fit H10 had far fewer parameters than polynomial fits of order > 3, as seen on plot 5(b).

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

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

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

5.2. Global sensitivity

Global sensitivities of the heat duties of all equipment under consideration with respect to the 11 inputs produced by HDMR fitted over the entire domain are summarised in Figures 10 and 11. It can be seen that in all cases only 4 or fewer inputs have significant influence on a given output. Additionally, 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 device directly upstream (given such exists). While molar flow of oil, main feedstock of the process, has significant effect on all heat duties (except that of the flash drum), molar flow of methanol only affects heat duty of heater 10E02. This is because heat capacity of oil is around 100 higher than that of methanol (1665.0 J/mol/K [78] and 79.5 J/mol/K [79]) and only in the flash drum there is significantly more methanol than oil.

Heat duty of heater 10E01 is primarily affected by its operating temperature and molar flow and temperature of incoming oil. Heat duty of heater 10E02 is mostly affected by its operating temperature, operating temperature of reactor 10D01 and molar flow of oil and methanol. Heat duty of heater 10E03 is primarily affected by its operating temperature, operating temperature of decanter 10D02D and molar flow of oil. Heat duty of reactor 10D01 is primarily affected by its operating temperature, operating temperature of heater 10E01 and molar flow of oil. Heat duty of flash drum 10D02 is primarily affected by its operating temperature and operating temperature of heater 10E02. Heat duty of decanter 10D02D is primarily affected by its operating temperature, operating temperature of flash drum 10D02 and

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

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

6. Conclusions

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

A number of different behaviours were observed in the study. Most surrogates achieved at least a reasonable fit regardless of the domain size, number 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

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

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

Acknowledgements

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

References

[1] G. Brundtland, M. Khalid, S. Agnelli, S. Al-Athel, B. Chidzero, L. Fadika, Our Common Future: The World Commission on Environment and Development, Oxford University Press, Oxford, 1987.

[2] C. Hoffman, The Industrial Ecology of Small and Intermediate-sized Technical Companies: Implications for Regional Economic Development, Technical Report 197411, Texas University, USA, 1971.

[3] C. Watanabe, Industrial ecology: Introduction of Ecology into Industrial Policy, Technical Report, Ministry of International Trade and Industry (MITI), Tokyo, 1972.

[4] B. Allenby, Journal of Cleaner Production 12 (2004) 833–839. doi:10.1016/j.jclepro.2004.02.010.

[5] B. Allenby, Progress in Industrial Ecology - An International Journal 1-2 (2006) 28–40. doi:10.1504/PIE.2006.010039.

[6] M. Chertow, Annual Review of Energy and Environment 25 (2000) 313–337. doi:10.1146/annurev.energy.25.1.313.

[7] M. Pan, J. Sikorski, C. A. Kastner, J. Akroyd, S. Mosbach, R. Lau, M. Kraft, Energy Procedia 150 (2015). doi:10.1016/j.egypro.2015.07.313.

[8] P. Desrochers, Journal of Industrial Ecology 5 (2001) 29–44. doi:10.1162/10881980160084024.

[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.

[19] J. P. C. Kleijnen, European Journal of Operational Research 192 (2009) 707–716. doi:10.1016/j.ejor.2007.10.013.

[20] S. B. Crary, Analog Integrated Circuits and Signal Processing 32 (2002) 7–16. doi:10.1023/A:1016063422605.

[21] Y. Reich, S. Barai, Artificial Intelligence in Engineering 13 (1999) 257–272. doi:10.1016/S0954-1810(98)00021-1.

[22] M. H. Bakr, J. W. Bandler, K. Madsen, J. Soandergaard, Optimization and Engineering 1 (2000) 241–276. doi:10.1023/A:1010000106286.

[23] R. Jin, X. Du, W. Chen, Structural and Multidisciplinary Optimization 25 (2003) 99–116. doi:10.1007/s00158-002-0277-0.

[24] W. A. Gough, W. J. Welch, Journal of Marine Research 52 (1994) 773–796. doi:10.1357/0022240943076911.

[25] P. Geyera, A. Schlueter, Applied Energy 119 (2014) 537–556. doi:10.1016/j.apenergy.2013.12.064.

[26] D. L. Knill, A. A. Giunta, C. A. Baker, B. Grossman, W. H. Mason, R. T. Haftka, L. T. Watson, Journal of Aircraft 36 (1999) 75–86. doi:10.2514/2.2415.

[27] S. B. Crary, P. Cousseau, D. Armstrong, D. M. Woodcock, E. H. Mok, O. Dubochet, P. Lerch, P. Renaud, Computer Modeling in Engineering and Sciences 1 (2000) 127–140. doi:10.3970/cmes.2000.001.127.

[28] N. Chen, K. Wang, C. Xiao, J. Gong, Environmental Modelling & Software 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.

[45] I. Fahmi, S. Cremaschi, Computers & Chemical Engineering 46 (2012) 105–123. URL: http://www.sciencedirect.com/science/article/pii/S0098135412001822. doi:10.1016/j.compchemeng.2012.06.006.

[46] C. A. Henao, C. T. Maravelias, AIChE Journal 57 (2011) 1216–1232. URL: http://onlinelibrary.wiley.com/doi/10.1002/aic.12341/full. doi:10.1002/aic.12341.

[47] L. Kong, S. Murat Sen, C. A. Henao, J. A. Dumesic, C. T. Maravelias, Computers and Chemical Engineering (2016). doi:10.1016/j.compchemeng.2016.02.013.

[48] AspenTech, aspentech - Aspen Plus v8.6, 2015. URL: http://www.aspentech.com/products/engineering/aspen-plus/, date accessed: 12.01.2016.

[49] CMCL Innovations, Model Development Suite (MoDS), 2015. URL: http://www.cmclinnovations.com/mods/, date accessed: 12.01.2015.

[50] A. Braumann, P. L. Man, M. Kraft, Industrial and Engineering Chemistry Research 49 (2010) 428–438. doi:10.1021/ie901230u.

[51] P. L. Man, A. Braumann, M. Kraft, Industrial and Engineering Chemistry Research 65 (2010) 4038–4045. doi:10.1016/j.ces.2010.03.042.

[52] A. Braumann, P. Man, M. Kraft, AIChE Journal 57 (2011) 3105–3121. doi:10.1002/aic.12526.

[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,

R. McDavid SAE Paper No. 2010-01-0152 (2010). doi:10.4271/2010-01-0152.

[62] G. Brownbridge, A. J. Smallbone, W. Phadungsukanan, M. Kraft, B. Johansson SAE Paper No. 2011-01-0237 (2011). doi:10.4271/2011-01-0237.

[63] J. E. Etheridge, S. Mosbach, M. Kraft, H. Wu, N. Collings SAE Paper No. 2010-01-1241 (2010). doi:10.4271/2010-01-1241.

[64] A. M. Aldawood, S. Mosbach, M. Kraft, A. A. Amer SAE Paper No. 2011-01-1783 (2011). doi:10.4271/2011-01-1783.

[65] P. Azadi, G. Brownbridge, I. Kemp, S. Mosbach, J. S. Dennis, M. Kraft, ChemCatChem 7 (2015) 137–143. doi:10.1002/cctc.201402662.

[66] E. K. Y. Yapp, R. I. A. Patterson, J. Akroyd, S. Mosbach, E. M. Adkins, J. H. Miller, M. Kraft, Combustion and Flame (2016). doi:10.1016/j.combustflame.2016.01.033.

[67] Microsoft, COM technical overview, 2015. URL: https://msdn.microsoft.com/en-us/windows/desktop, date accessed: 02.06.2015.

[68] I. M. Sobol, USSR Computational Mathematics and Mathematical Physics 7 (1967) 86–112. doi:10.1016/0041-5553(67)90144-9.

[69] S. Joe, F. Y. Kuo, SIAM Journal on Scientific Computing 30 (2008) 2635–2654. doi:10.1137/070709359.

[70] R. H. Myers, D. C. Montgomery, C. M. Anderson-Cook, Response Surface Methodology: Process and Product Optimization Using Designed Experiments, 3rd ed., Wiley-Blackwell, 2009.

[71] I. M. Sobol, Matematicheskoe Modelirovanie 2 (1990) 112–118.

[72] H. Rabitz, Ö. F. Alış, Journal of Mathematical Chemistry 25 (1999) 197–233. doi:10.1023/A:1019188517934.

[73] G. Li, S.-W. Wang, H. Rabitz, The Journal of Physical Chemistry A 106 (2002) 8721–8733. doi:10.1021/jp014567t.

[74] H. M. Wainwright, S. Finsterle, Y. Jung, Q. Zhou, J. T. Birkholzer, Computers and Geosciences 65 (2014) 84–94. doi:10.1016/j.cageo.2013.06.006.

[75] M. Baran, L. Bieniasz, Applied Mathematics and Computation 258 (2015) 206–219. doi:10.1016/j.amc.2015.02.007.

[76] N. R. Draper, H. Smith, Applied Regression Analysis, 3rd ed., Wiley-Interscience, 1998.

[77] R. J. Hyndman, A. B. Koehler, International Journal of Forecasting 22 (2006) 679–688. doi:10.1016/j.ijforecast.2006.03.001.

[78] V. Filatov, V. Afanas'ev, Khimiya i Tekhnoliya Vody 35 (1992) 97–100.

[79] B. Freedman, M. Bagby, H. Khoury, Journal of the American Oil Chemists' Society 66 (1989) 595–596. doi:10.1007/BF02885455.

List of Figures

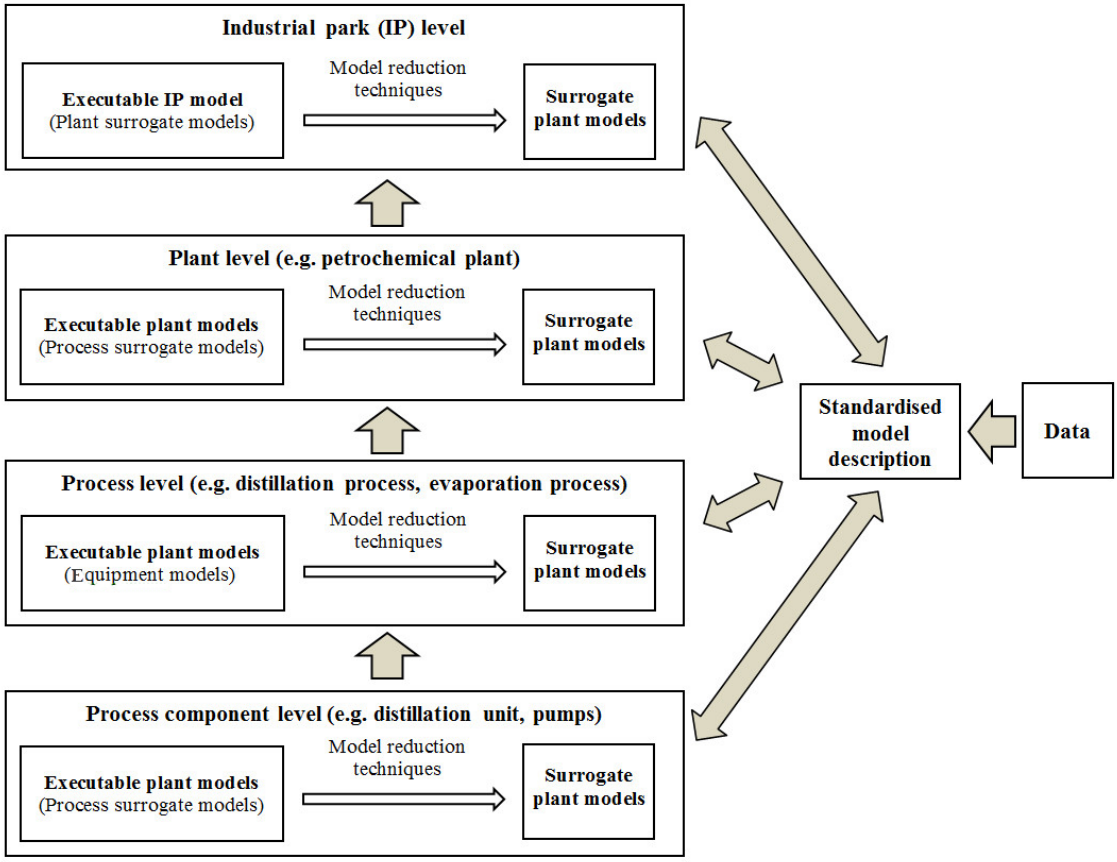


Figure 1: *Framework of EIP modelling based on Industry 4.0. Adopted from Pan et al. [7].*

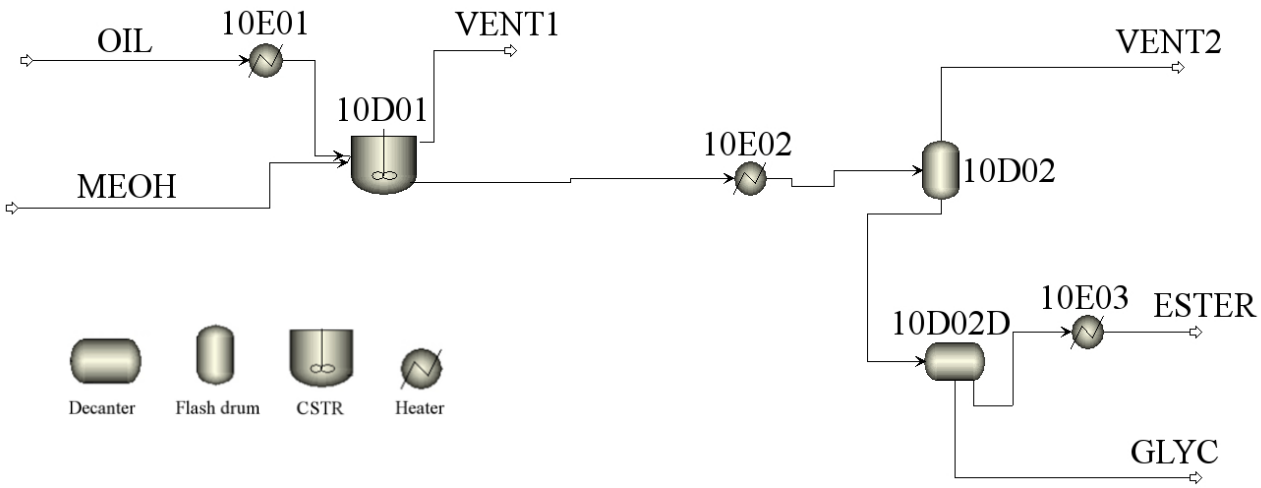
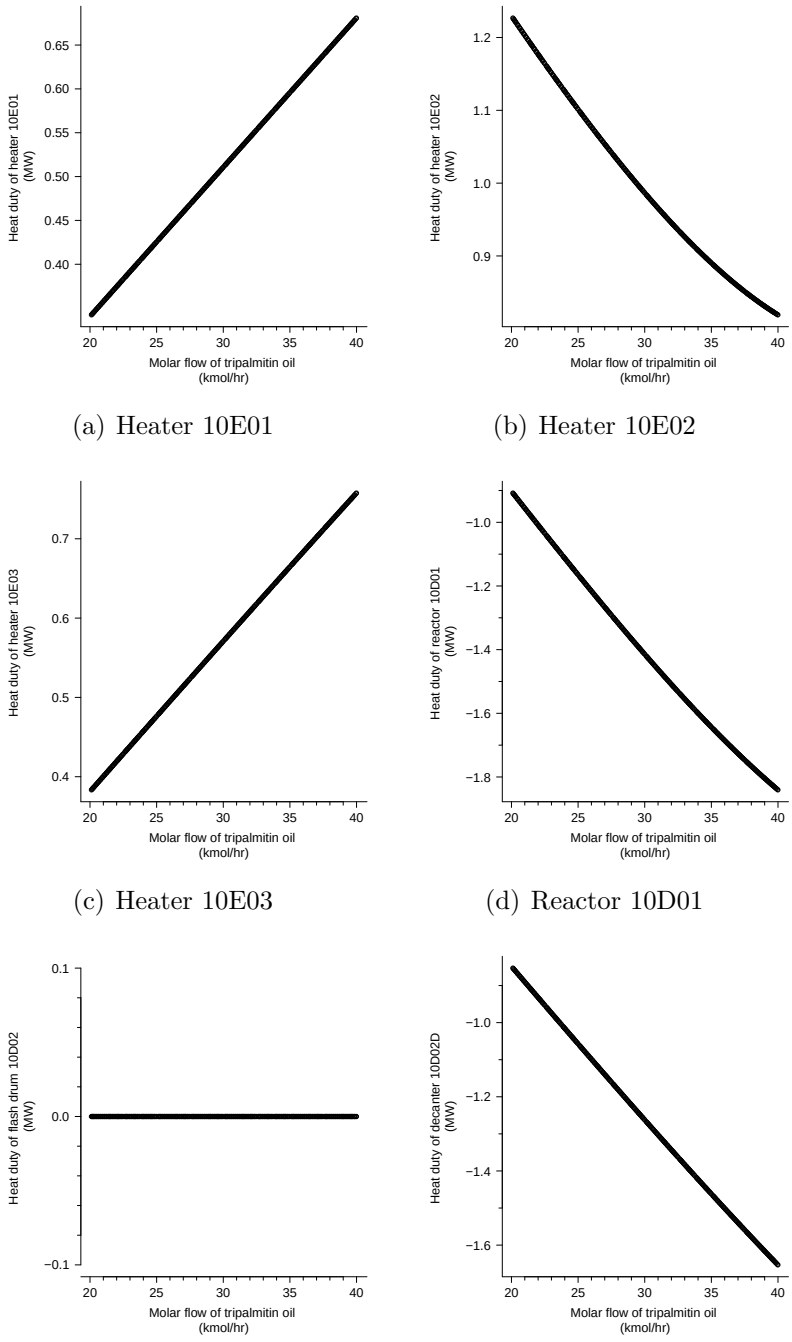


Figure 2: Graphical representation of the process flow sheet model of a biodiesel production line.



(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.

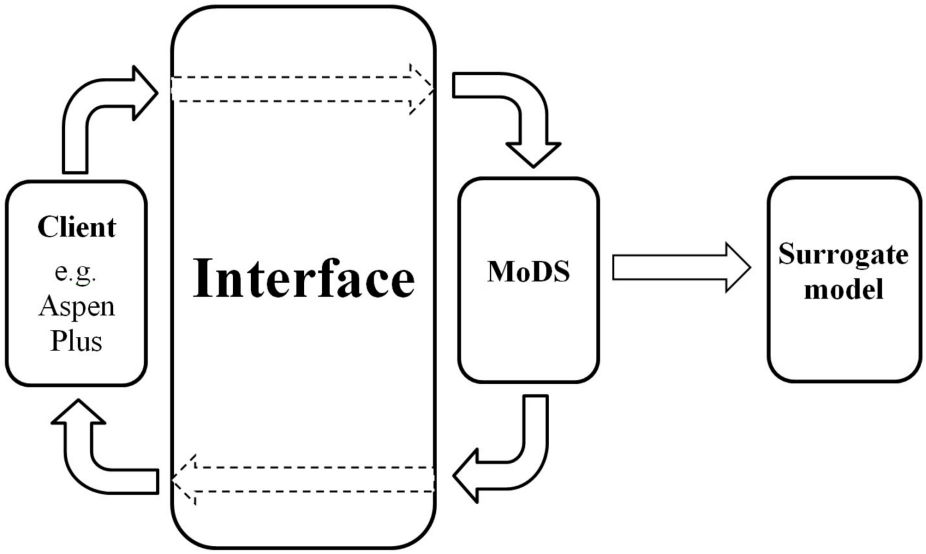
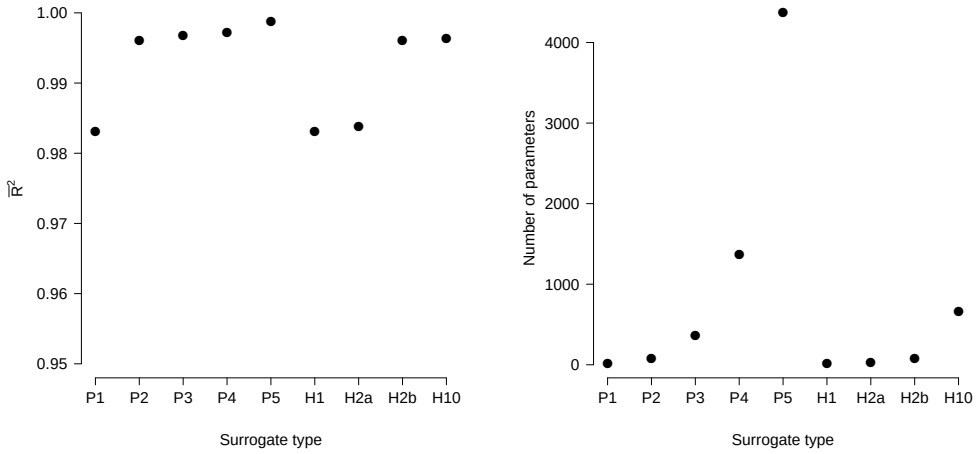


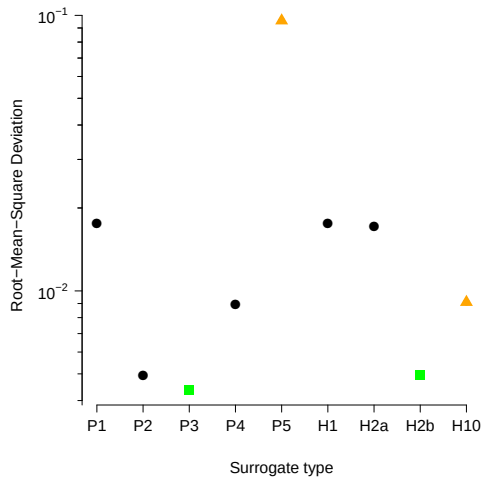
Figure 4: *Model Development Suite work flow.*



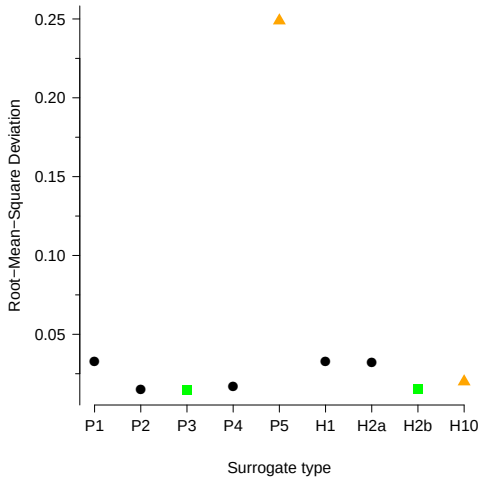
(a) Plot of $\bar{R}^2$ for the considered surrogates.

(b) Plot of number of parameters for the considered surrogates.

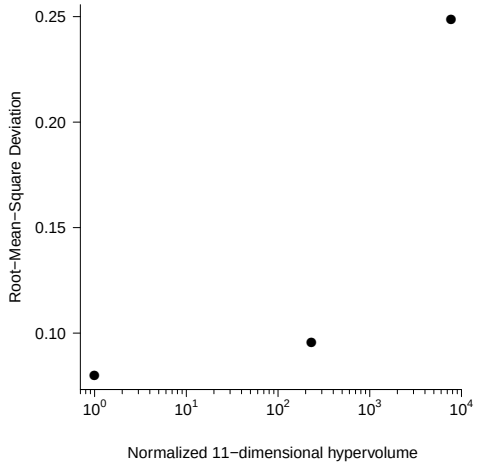
Figure 5: Plots of RMSD and number of parameters for the considered surrogates produced for heat duty of reactor 10D01 with respect to all 11 inputs. Labels P1 through P5 correspond to polynomial fits of order 1 through 5. Label H1 corresponds to a 1^{st} order fit, H2a to a 2^{nd} order without interactions,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



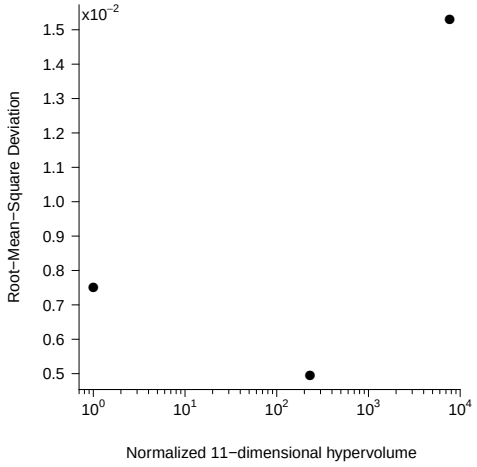
(a) RMSD for the considered surrogates for medium domain size.



(b) RMSD for the considered surrogates.

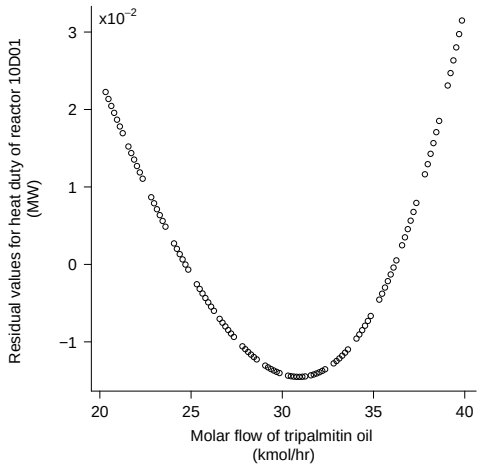


(c) RMSD against domain sizes for polynomial fit of order 5 (for boundaries see Table 1).

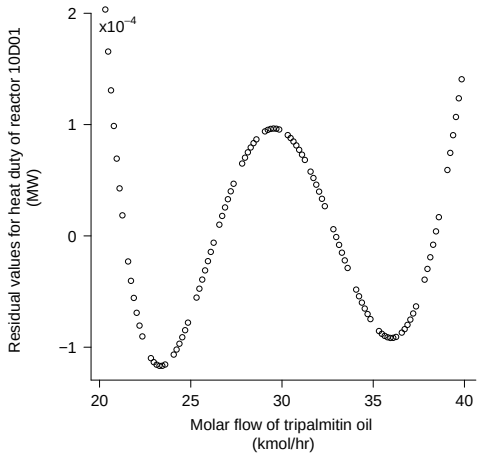


(d) RMSD against domain sizes for HDMR fit (for boundaries see Table 1).

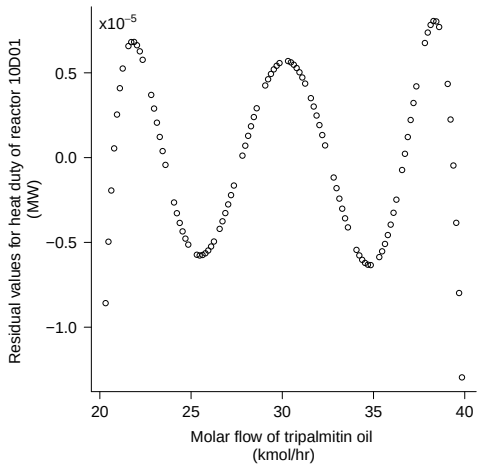
Figure 6: Plots of RMSD for the considered surrogates and domain sizes produced for heat duty of reactor 10D01 with respect to all 11 inputs. Labels P1 through P5 correspond to polynomial fits of order 1 through 5. Label H1 corresponds to a 1^{st} order fit, H2a to a 2^{nd} order without interactions, 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.



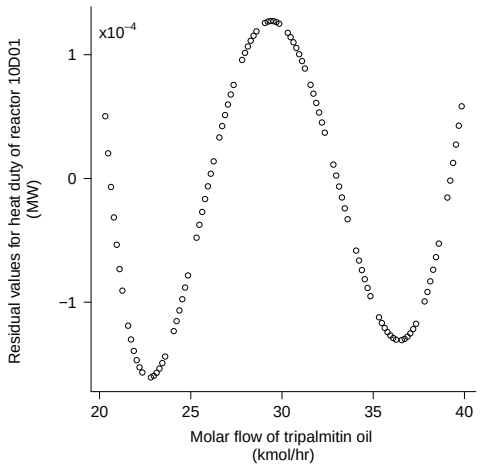
(a) Plot of residuals for 1^{st} order polynomial fit.



(b) Plot of residuals for 3^{rd} order polynomial fit.

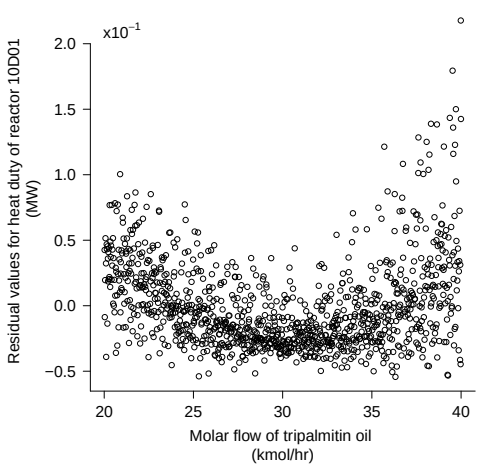


(c) Plot of residuals for 5^{th} order polynomial fit.

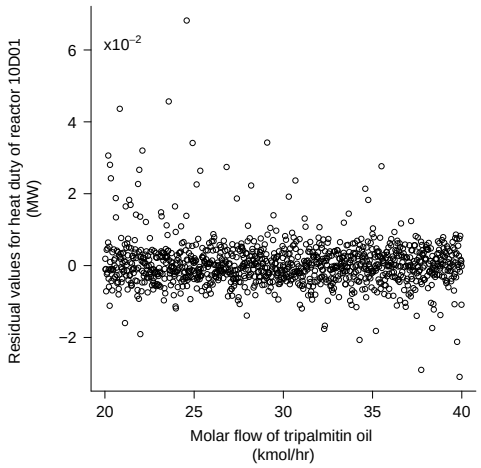


(d) Plot of residuals for HDMR fit H10 (3^{rd} order polynomial).

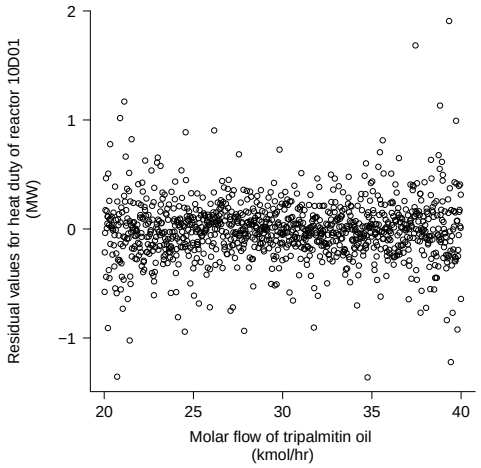
Figure 7: Plot of residuals against molar flow of tripalmitin oil for heat duty of reactor 10D01 produced for 1 input.



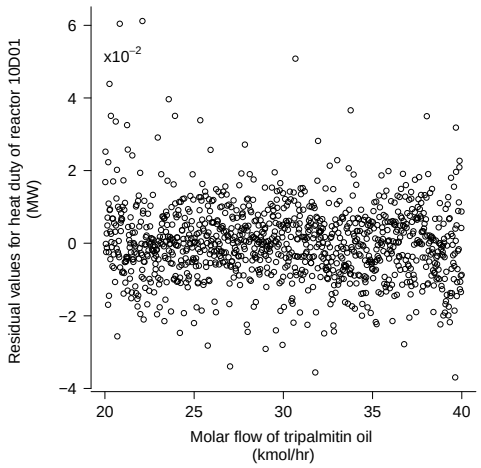
(a) Plot of residuals for 1^{st} order polynomial fit.



(b) Plot of residuals for 3^{rd} order polynomial fit.

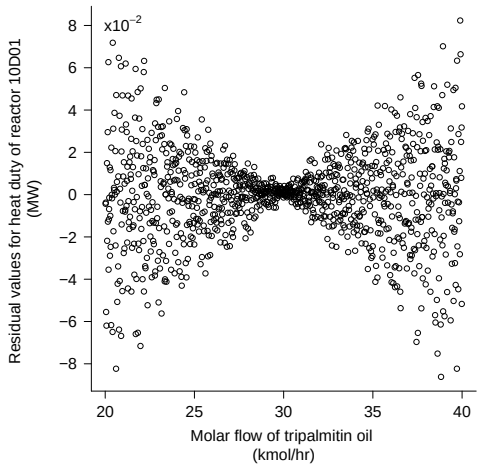


(c) Plot of residuals for 5^{th} order polynomial fit.

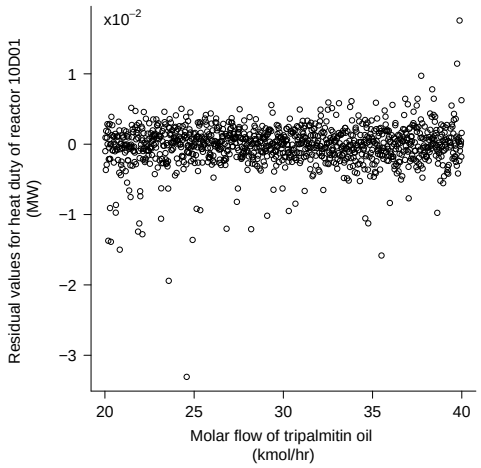


(d) Plot of residuals for HDMR fit H10.

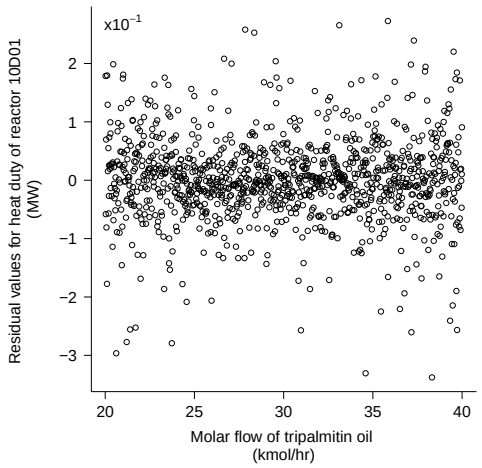
Figure 8: Plot of residuals against molar flow of tripalmitin oil for heat duty of reactor 10D01 produced for 11 inputs.



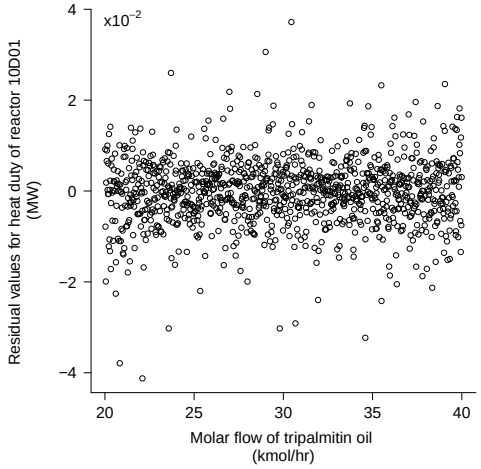
(a) Plot of residuals for 1^{st} order polynomial fit.



(b) Plot of residuals for 3^{rd} order polynomial fit.

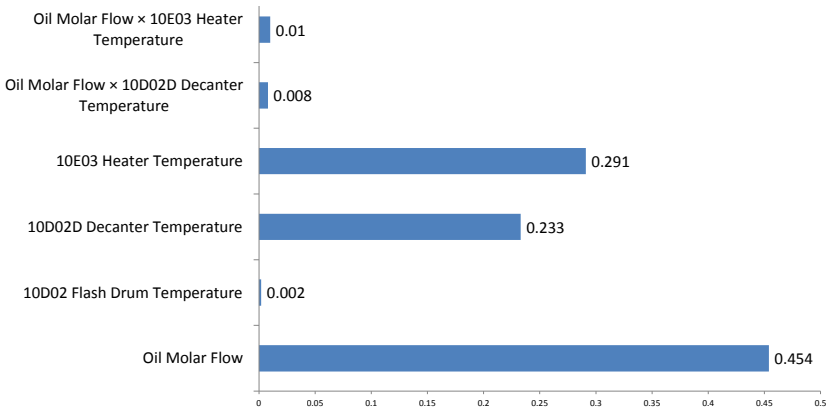


(c) Plot of residuals for 5^{th} order polynomial fit.

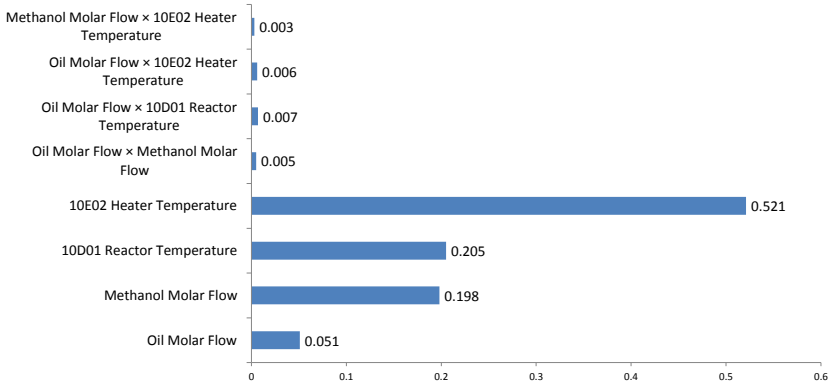


(d) Plot of residuals for HDMR fit H10.

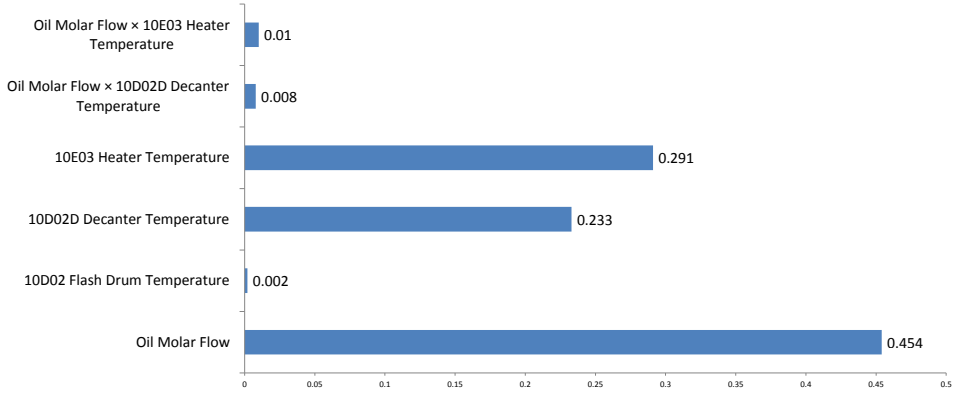
Figure 9: Plot of residuals against molar flow of tripalmitin oil for heat duty of heater 10E03 produced for 11 inputs.



(a) 10E01 Heater - Heat Duty.

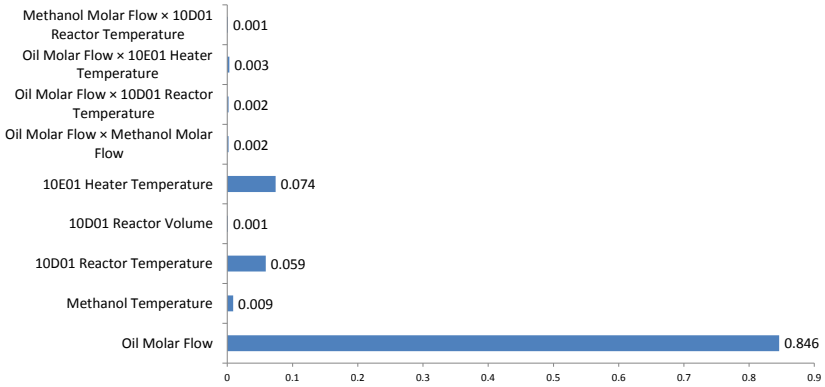


(b) 10E02 Heater - Heat Duty.

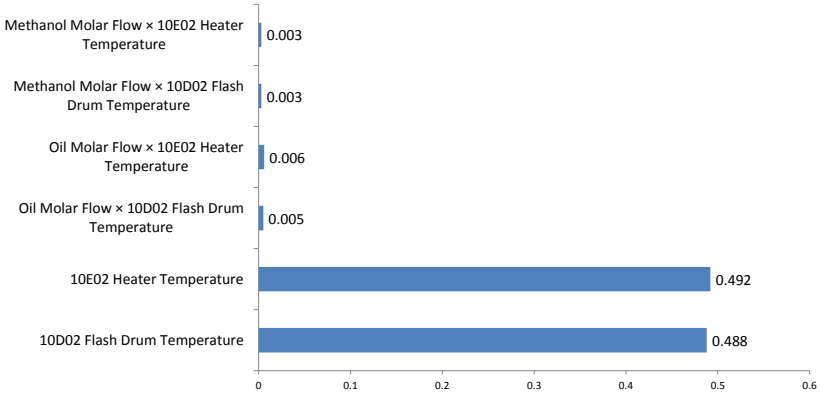


(c) 10E03 Heater - Heat Duty.

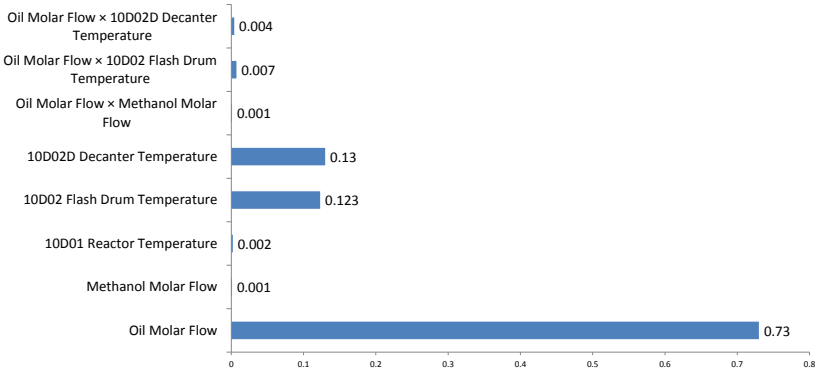
Figure 10: Global sensitivities produced by 11-dimensional HDMR fit over the entire domain.



(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.