Computers and Chemical Engineering 145 (2021) 107162

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

AutoMoG: Automated data-driven Model Generation of multi-energy systems using piecewise-linear regression

Andreas Kämper[a], Ludger Leenders[a,c], Björn Bahl[a], André Bardow[a,b,c,*]

[a] Institute of Technical Thermodynamics, RWTH Aachen University, Aachen 52062, Germany
[b] Institute of Energy and Climate Research, Energy Systems Engineering (IEK-10), Forschungszentrum Jülich GmbH, Jülich 52425, Germany
[c] Energy & Process Engineering, Department of Mechanical and Process Engineering, ETH Zurich, Zurich 8092, Switzerland

ARTICLE INFO

Article history:
Received 29 June 2020
Revised 4 October 2020
Accepted 8 November 2020
Available online 13 November 2020

Keywords:
Regression analysis
Mixed-integer linear programming
Energy system optimization
Information criterion

ABSTRACT

Operational optimization of multi-energy systems requires a mathematical model that is accurate and computationally efficient. A model can be generated in a data-driven way if measured data is available. Commonly, data is then used to model each component of the multi-energy system independently. However, independent modeling of each component may lead to models that are unnecessarily complicated and, thus, inefficient in practice.

In this work, we propose the method AutoMoG for Automated data-driven Model Generation of multi-energy systems using piecewise-linear regression. AutoMoG provides Mixed-Integer Linear Programming models of multi-energy systems. To accurately model the overall multi-energy system, AutoMoG balances the errors caused by each component. Model accuracy is measured in terms of operating cost.

In a case study, AutoMoG provides a multi-energy system model with less linear sections than single-component regression Still, AutoMoG retains high accuracy. Thereby, AutoMoG enables efficient data-driven modeling as the basis for multi-energy system optimization.

1. Introduction

Multi-energy systems usually consist of multiple components such as combined-heat-and-power (CHP) engines or absorption chillers. The components may be very different. The interaction of these different components leads to complex behavior on the system level. Due to the complex behavior of multi-energy systems, optimal operation usually requires mathematical optimization (Mancarella, 2014). The resulting operational optimization problem is challenging. Goderbauer et al. (2019) show that the operational optimization problem of multi-energy systems is weakly NP-hard, even for a single load case. As these challenging operational optimization problems have to be solved frequently in practice, the underlying models of the multi-energy systems need to be computationally efficient. However, it is challenging to generate multi-energy system models that are both accurate and computationally efficient (Mitsos et al., 2018).

In general, two approaches are followed to generate multi-energy system models: first-principles modeling and data-driven modeling. First-principles models are derived from theory with the aim to represent the real physical behavior of a system or component (Smolin et al., 2019). However, solving full first-principles models often is computationally demanding (McBride and Sundmacher, 2019). Furthermore, frequently the physical behavior of the system is partly unknown, which prevents full first-principles modeling.

The other possibility to generate a multi-energy system model is data-driven modeling. Data-driven models are derived from data with the aim to represent the input-output relationship of a system (McBride and Sundmacher, 2019). Measured data is increasingly available in multi-energy systems, in particular, due to the implementation of energy management systems according to ISO 50001:2018 (2018). Thus, data-driven model generation for multi-energy systems becomes increasingly promising.

In a multi-energy system, measured input and output data can be used to generate a data-driven model of each component. The measured data is used to regress the input-output relationship of each component. The regression approximates a functional relationship between independent input variables and output vari-

* Corresponding author.
E-mail address: abardow@ethz.ch (A. Bardow).

https://doi.org/10.1016/j.compchemeng.2020.107162

ables from a given data set (Yang et al., 2016). For regression, many approaches are available such as linear regression, kriging (Kleijnen and Beers, 2004), support-vector regression (Smola and Schölkopf, 2004) or neural networks (Huang et al., 2010).

In operational optimization, the generated model will usually be solved repeatedly. Thus, the regression analysis should yield a model that is computationally efficient. At the same time, the model has to be sufficiently accurate. To generate accurate and computationally efficient models, Cozad et al. (2014) and Wilson and Sahinidis (2017) presented a framework for automated learning of algebraic models (ALAMO). ALAMO provides black-box models from data obtained by simulation or experiments. Information criteria are used to find simple models with sufficient accuracy. However, the resulting model is in general nonlinear. Thus, if this nonlinear model is used for operational optimization, the problem will usually be a Mixed-Integer Nonlinear Program (MINLP). In practice, MINLPs are still challenging to solve to global optimality (Mitsos et al., 2018).

Commonly, nonlinearities are therefore approximated by piecewise-linear models leading to Mixed-Integer Linear Programs (MILPs) (Zhang et al., 2016; Gao et al., 2018; Voll et al., 2013). MILPs can be efficiently solved to global optimality with commercial state-of-the-art solvers. Methods are available to generate piecewise-linear models from measured data: Zhang et al. (2016) proposed a data-driven algorithm to generate surrogate models of process systems. The generated surrogate models are piecewise linear in convex regions and, thus, can be used in MILPs. Yang et al. (2016) and Gkioulekas and Papageorgiou (2018) provided a mathematical programming approach for piecewise-linear regression. The piecewise-linear regression models are obtained by solving MILP regression problems. MILP regression problems minimize the least distances between data and model to retain linearity. Recently, Kong and Maravelias (2020) and Rebennack and Krasko (2020) proposed formulations to model continuous piecewise-linear regression problems as MILPs. The provided models are piecewise linear and, thus, can be used in MILPs.

The reviewed methods could solve the piecewise-linear regression problem for any component in a multi-energy system. However, modeling each component of a multi-energy system independently may lead to an overall model of the multi-energy system that is unnecessarily complicated and, thus, computationally inefficient.

Thus, in this work, we propose the method AutoMoG for Automated data-driven Model Generation of multi-energy systems. AutoMoG solves a data-driven model generation problem for multi-energy systems, while balancing the errors caused by each component's model in the overall model of the multi-energy system. The model of the multi-energy system is assumed to be used for an economic optimization. Thus, cost-based weighting factors are used to determine the impact of each component's model error on the error of the multi-energy system model. AutoMoG terminates once a predefined accuracy of the multi-energy system model is achieved. However, if the predefined accuracy is not achievable, AutoMoG avoids overfitting by using the Corrected Akaike Information Criterion AIC_C (Hurvich and Tsai, 1993). AutoMoG provides an MILP model of the multi-energy system with continuous representation of the components' input-output relationship.

In Section 2, we formulate the data-driven model generation problem for multi-energy systems. In Section 3, we describe the proposed method AutoMoG. In Section 4, we apply AutoMoG to a case study for a decentralized multi-energy system from literature. In Section 5, we conclude with the key findings.

2. Data-driven model generation for multi-energy systems

The data-driven model generation problem for multi-energy systems shall provide a sufficiently accurate and computationally efficient MILP model of the multi-energy system. In the provided MILP model, the input-output relationship of each component $s \in S$ in the multi-energy system has to be represented by a piecewise-linear model.

In general, the functional relationship between input I_s (e.g. gas or electricity) and output O_s (e.g. heating or cooling) of a component s (e.g. boiler or compression chiller) is nonlinear. For MILP optimization models, nonlinear functional relationships are approximated by piecewise-linear functions $I_s^{\text{Model}}(O_{s,n})$. Here, we choose to model the input I_s^{Model} as linear function of the output $O_{s,n}$, because we can easily convert the input to operating cost. This conversion is crucial for the AutoMoG method; more details are given in Section 3.2. However, AutoMoG can be easily adapted to model the output as function of the input.

$$I_s^{\text{Model}} = \sum_n \eta_{s,n} \cdot (a_{s,n} \cdot O_{s,n} + b_{s,n}) \tag{1}$$

$$O_{s,n} \geq o_{s,n-1}^{\text{UB}} \cdot \eta_{s,n} \qquad \forall\ n \in N_s \tag{2}$$

$$O_{s,n} \leq o_{s,n}^{\text{UB}} \cdot \eta_{s,n} \qquad \forall\ n \in N_s \tag{3}$$

$$\sum_n \eta_{s,n} \leq 1 \tag{4}$$

with I_s^{Model} being the modeled input of component s and N_s being the number of piecewise-linear sections n of component s. The parameter $a_{s,n}$ denotes the gradient of linear section n and the parameter $b_{s,n}$ denotes the intercept of linear section n. The binary variable $\eta_{s,n}$ is equal to 1, if and only if the output $O_{s,n}$ lies in between the upper bound $o_{s,n}^{\text{UB}}$ of the linear section n and the upper bound $o_{s,n-1}^{\text{UB}}$ of the lower linear section $n-1$. Eq. (4) ensures that the output $O_{s,n}$ lies on maximum one linear section n.

We assume that an MILP model with fewer binary variables can be more efficiently solved. This assumption is often made in practice (Katz et al., 2020). The number of binary variables in the MILP model rises with the number of piecewise-linear sections. Thus, the objective of the data-driven model generation is to identify the minimal number of piecewise-linear sections N for a multi-energy system model with a given accuracy.

The resulting structure of the data-driven model-generation problem for multi-energy systems is the following:

min number of piecewise-linear sections in multi-energy system model (Eq. (5))

s.t. the multi-energy system model fulfills a given accuracy (Eq. (6))
the multi-energy system model is fitted to measured data (Eq. (7))
the component models are piecewise linear (Eq. (8)-(9))
the piecewise-linear models are continuous (Eq. (10)-(13))
2 equations to count all piecewise-linear sections (Eq. (14)-(15))

The mathematical formulation of the data-driven model-generation problem for multi-energy systems is given in Eq. (5)-(19):

$$\min \quad N = \sum_s^S N_s \tag{5}$$

Table 1
List of all variables and parameters of the actual problem given in Eq. (5)-(19).

Variables	$N, N_s, \Delta C^{\text{System}}, I^{\text{Model}}_{s,d}, \gamma_{s,n,d}, A_{s,n}, B_{s,n}, O^{\text{UB}}_{s,n}, \kappa_{s,n}$
Parameters	$\delta^{\text{rel}}, c^{\text{Input}}_s, I^{\text{Data}}_{s,d}, O^{\text{Data}}_{s,d}, \lvert D\rvert, m$

$$\text{s.t.} \quad \Delta C^{\text{System}} \leq \delta^{\text{rel}} \cdot \sum_{s\in S}\sum_{d\in D} c^{\text{Input}}_s \cdot I^{\text{Data}}_{s,d} \tag{6}$$

$$\Delta C^{\text{System}} = \sum_{s\in S} c^{\text{Input}}_s \cdot \sqrt{\sum_{d\in D}(I^{\text{Data}}_{s,d} - I^{\text{Model}}_{s,d})^2} \tag{7}$$

$$I^{\text{Model}}_{s,d} = \sum_{n\in N^{\max}_s} \gamma_{s,n,d} \cdot (A_{s,n} \cdot O^{\text{Data}}_{s,d} + B_{s,n}), \qquad \forall\, s \in S, d \in D \tag{8}$$

$$\sum_{n\in N^{\max}_s} \gamma_{s,n,d} = 1, \qquad \forall\, s \in S, d \in D \tag{9}$$

$$0 = \kappa_{s,n+1} \cdot \left[(A_{s,n+1} - A_{s,n}) \cdot O^{\text{UB}}_{s,n} + B_{s,n+1} - B_{s,n}\right], \quad \forall\, s \in S, n \in N^{\max}_s \tag{10}$$

$$O^{\text{UB}}_{s,n} \geq O^{\text{Data}}_{s,d} \cdot \gamma_{s,n,d}, \qquad \forall\, s \in S, n \in N^{\max}_s, d \in D \tag{11}$$

$$O^{\text{UB}}_{s,n} \cdot \gamma_{s,n+1,d} \leq O^{\text{Data}}_{s,d} \cdot \gamma_{s,n+1,d}, \qquad \forall\, s \in S, n \in N^{\max}_s, d \in D \tag{12}$$

$$0 \leq (O^{\text{UB}}_{s,n+1} - O^{\text{UB}}_{s,n}) \cdot \kappa_{s,n+1}, \qquad \forall\, s \in S, n \in N^{\max}_s \tag{13}$$

$$\kappa_{s,n} \geq \frac{1}{|D|} \cdot \sum_{d\in D} \gamma_{s,n,d}, \qquad \forall\, s \in S, n \in N^{\max}_s \tag{14}$$

$$N_s = \sum_{n\in N^{\max}_s} \kappa_{s,n}, \qquad \forall\, s \in S \tag{15}$$

$$I^{\text{Model}}_{s,d} \geq 0, \qquad \forall\, s \in S, d \in D \tag{16}$$

$$O^{\text{UB}}_{s,n} \geq 0, \qquad \forall\, s \in S, n \in N^{\max}_s \tag{17}$$

$$A_{s,n}, B_{s,n} \leq m \cdot \kappa_{s,n}, \qquad \forall\, s \in S, n \in N^{\max}_s \tag{18}$$

$$A_{s,n}, B_{s,n} \geq -m \cdot \kappa_{s,n}, \qquad \forall\, s \in S, n \in N^{\max}_s \tag{19}$$

In the following, we refer to the data-driven model generation problem for multi-energy systems (Eq. (5)-(19)) as the actual problem. We state the variables and parameters of the actual problem in Table 1.

$d \in D$ is a measured data point. c^{Input}_s is a cost-based weighting factor and depends on the type of input for component s. Different components of the multi-energy system may have different forms of input (e.g., gas for a boiler, but electricity for a compression chiller). By using cost-based weighting factors c^{Input}_s, we convert the different forms of input to operating costs. We show in Section 3.2 how we determine cost-based weighting factors c^{Input}_s for a typical multi-energy system. $N^{\max}_s$ is the maximum number of piecewise-linear section allowed to model component s. The binary variable $\kappa_{s,n}$ denotes whether linear section n is used to model component s. The binary variable $\gamma_{s,n,d}$ denotes whether data point d is assigned to linear section n of component s. $A_{s,n}$ is the gradient and $B_{s,n}$ is the intercept of linear section n in the piecewise-linear model of component s.

The objective of the actual problem is to minimize the number of piecewise-linear sections N within the multi-energy system model (Eq. (5)). Constraints (6)-(7) restrict the sum of squared residuals of all data points d and components s to be smaller than the product of the predefined relative error of the multi-energy system δ^{rel} and the sum of the cost of all measured input data. Here, the sum of squared residuals of all data points d of component s are weighted by the cost-based weighting factor c_s. Constraints (8) evaluate the piecewise-linear models at each data point d for each component s. Constraints (9) ensure that each data point d is assigned to exactly one linear section n for each component s. Constraints (10) force the piecewise-linear models to be continuous at the breakpoints. Constraints (11)-(13) define the variables for the upper bound $O^{\text{UB}}_{s,n}$ of each linear section n and arrange the linear sections in ascending order. Constraints (14) ensure that the linear section n is chosen to model component s, if at least one data point d is assigned to the linear section n. Constraints (15) sum up the number of chosen linear sections for each component s. Constraints (16)-(17) ensure $I^{\text{Model}}_{s,d}$ and $O^{\text{UB}}_{s,n}$ to be positive variables. Constraints (18)-(19) assign the value 0 to the variables $A_{s,n}$ and $B_{s,n}$ if linear section n is not selected for component s. The constrains use a Big-M formulation with the Big-M value m.

The actual problem is an MINLP problem. The nonlinear character of the actual problem results from Eq. (7), (8), (10) and (13). Solving the actual problem is computationally demanding. We implemented the actual problem in GAMS (GAMS Development Corporation, 2016) and tried to solve the actual problem with state-of-the-art MINLP solvers (SCIP (Gleixner et al., 2018), BARON (Tawarmalani and Sahinidis, 2005), DICOPT (Kocis and Grossmann, 1989) and BONMINH (COIN-OR (Project Manager P. Bonami), 2016)). The MINLP solvers could not even find a feasible solution for a typically sized industrial multi-energy system (Section 4.5). The MINLP solvers ran without a time limit. 2 solvers wrongly considered the problem infeasible, 1 solver terminated without a solution and 1 solver reached an iteration limit. Thus, solving the actual problem is impractical in applications.

However, the actual problem can be rendered computationally feasible by 2 possibilities: One possibility is to linearize the actual problem (MINLP) to an MILP, based on the formulation by Yang et al. (2016). This linearization introduces a few shortcomings: Squared residuals can no longer be employed in an MILP (Eq. (7)). Absolute residuals are calculated instead. Furthermore, the resulting piecewise-linear models are in general not continuous, because the nonlinear continuity constraint cannot be considered in an MILP (Eq. (10)). Recently, Kong and Maravelias (2020) and Rebennack and Krasko (2020) reformulated the nonlinear continuity constraint into a set of linear constraints. However, we show in Section 4.5 that the performance of the linearized problem is not always satisfying for practical applications even if the continuity constraint is ignored. Thus, in general, the solution of the linearized problem is not a feasible solution of the actual problem. Therefore, solving the linearized problem is not always satisfying for practical applications. Thus, there is a need for a solution method that provides a solution of the actual problem in a short time.

This need leads to the second possibility: decomposing the actual problem and solving the decomposed problem. For this purpose, we propose the decomposition method AutoMoG in this work. AutoMoG provides a solution of the actual problem in a short time and, thus, is suitable for practical applications.

3. AutoMoG: Automated data-driven model generation of multi-energy systems

AutoMoG decomposes the actual problem (Eq. (5)-(19)) to piecewise-linear regression problems for each component in a

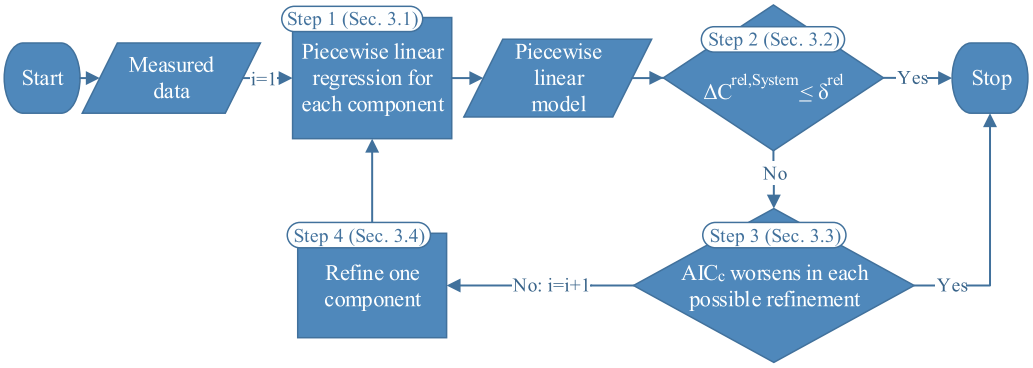


Fig. 1. Proposed method AutoMoG for automated model generation using measured data. $\Delta C^{\text{rel,System}}$ is the relative error of the multi-energy system model (Section 3.2). δ^{rel} is the allowed relative error of the multi-energy system model. The Corrected Akaike Information Criterion AIC_c is checked to avoid overfitting.

multi-energy system. AutoMoG iteratively increases the accuracy of the multi-energy system model by increasing the number of linear sections N (Fig. 1).

In the following, we briefly outline the AutoMoG method, before we explain the steps of AutoMoG in detail.

Step 1: For each component s, AutoMoG performs a least-squares regression between measured input and output data with a predefined number of linear sections $N_s^{\max}$ (Section 3.1). The resulting piecewise-linear models $I_s^{\text{Model}}(O_s)$ are suitable for an MILP model of the multi-energy system.

Step 2: The error of the multi-energy system model is evaluated (Section 3.2). For evaluation, AutoMoG uses a cost-based weighting factor c_s^{Input} to determine the impact of a component's error on the error of the multi-energy system model. The cost-based weighting factor c_s^{Input} depends on the input of a component, because AutoMoG models the input as a function of the output (Eq. (1)). We describe how to determine the cost-based weighting factors c_s^{Input} in Section 3.2. AutoMoG terminates if the relative error of the multi-energy system model $\Delta C^{\text{rel,System}}$ is smaller than or equal to the allowed relative error δ^{rel}.

Step 3: If the relative error of the multi-energy system model $\Delta C^{\text{rel,System}}$ exceeds the allowed relative error δ^{rel}, AutoMoG checks the Corrected Akaike Information Criterion AIC_c (Hurvich and Tsai, 1993). The Corrected Akaike Information Criterion AIC_c is used for model selection by capturing the trade-off between model accuracy and model complexity. If the information criterion AIC_c worsens for the refinement of a component, overfitting might occur. Thus, AutoMoG does not refine any component for which the information criterion AIC_c worsens. AutoMoG terminates if the information criterion AIC_c worsens for the refinement of all possible components. If AutoMoG terminates in Step 3, the allowed relative error δ^{rel} is not reached, but the measured data do not allow a more accurate model of the multi-energy system without the risk of overfitting.

Step 4: AutoMoG chooses one component to be refined based on the maximum error reduction in the multi-energy system model. For this purpose, AutoMoG calculates the error reduction in the multi-energy system model for the refinement of each component (Section 3.4). For the chosen component, AutoMoG increases the number of piecewise-linear sections by 1 and applies step 1 to the chosen component.

If the allowed relative error δ^{rel} is reached, AutoMoG provides a fully parameterized MILP model of the multi-energy system. AutoMoG aims to find the minimal number of piecewise-linear sections N to accurately represent the multi-energy system. However, AutoMoG cannot guarantee to provide the model with the minimal number of piecewise-linear sections N.

In the following, we explain the steps of AutoMoG in detail.

3.1. Step 1: Piecewise-linear regression for each component

AutoMoG solves a piecewise-linear regression problem for each component $s \in S$ separately, minimizing the sum of squared residuals ε_s. The squared residuals between the modeled Input $I_{s,d}^{\text{Model}}$ and the measured Input $I_{s,d}^{\text{Data}}$ are summed up for all measured data points $d \in D$:

$$\min_{A_{s,n},B_{s,n},O_{s,n}^{\text{UB}}} \varepsilon_s = \sum_{d\in D} \left(I_{s,d}^{\text{Data}} - I_{s,d}^{\text{Model}}\right)^2 \tag{20}$$

$$\text{s.t.} \quad I_{s,d}^{\text{Model}} = \sum_{n\in N_s} \gamma_{s,n,d} \cdot (A_{s,n} \cdot O_{s,d}^{\text{Data}} + B_{s,n}), \qquad \forall d \in D \tag{8}$$

$$\sum_{n\in N_s} \gamma_{s,n,d} = 1, \qquad \forall d \in D \tag{9}$$

$$0 = (A_{s,n+1} - A_{s,n}) \cdot O_{s,n}^{\text{UB}} + B_{s,n+1} - B_{s,n}, \qquad \forall n \in N_s \tag{10}$$

$$O_{s,n}^{\text{UB}} \geq O_{s,d}^{\text{Data}} \cdot \gamma_{s,n,d}, \qquad \forall n \in N_s, d \in D \tag{11}$$

$$O_{s,n}^{\text{UB}} \cdot \gamma_{s,n+1,d} \leq O_{s,d}^{\text{Data}} \cdot \gamma_{s,n+1,d}, \qquad \forall n \in N_s, d \in D \tag{12}$$

$$I_{s,d}^{\text{Model}} \geq 0, \qquad \forall d \in D \tag{16}$$

$$O_{s,n}^{\text{UB}} \geq 0, \qquad \forall n \in N_s \tag{17}$$

The constraints (8)-(12), (16) and (17) of the piecewise-linear regression problem are the same constraints as in the actual problem. The objective function (20) of the piecewise-linear regression problem is the sum of squared residuals ε_s for each component s, derived from constraint (7) of the actual problem. From this regression problem, AutoMoG obtains the parameters of the piecewise-linear model $A_{s,n}, B_{s,n}$ and the positions of the breakpoints $O_{s,n}^{\text{UB}}$ of each component s. The number of piecewise-linear sections N_s is fixed for each piecewise-linear regression problem.

Methods for solving the piecewise-linear regression problem are available in the literature (Camponogara and Nazari, 2015; Kong and Maravelias, 2020; Rebennack and Krasko, 2020; Yang et al., 2016; Zhang et al., 2016). Here, the piecewise-linear regression problem is an MINLP that is solved by applying an existing Matlab-Toolbox (D'Errico, 2009). The Matlab-Toolbox reformulates the MINLP into NLP subproblems and solves the subproblems with a local Nonlinear Programming (NLP) solver. The Matlab-Toolbox initializes the positions of the breakpoints for the piecewise-linear functions equidistantly and calculates the sum of squared residuals ε_s. After the initialization, the Matlab-Toolbox minimizes the sum

of squared residuals ε_s by iteratively changing the positions of the breakpoints. The piecewise-linear model of each component s is constrained to be continuous at the breakpoints (Eq. (10)). However, AutoMoG cannot guarantee to find the globally optimal positions of the breakpoints as it uses a local NLP solver for the NLP subproblems. For proof of optimality, the approaches of Kong and Maravelias (2020) and Rebennack and Krasko (2020) can be used instead of the Matlab-Toolbox.

Thus, step 1 of AutoMoG provides a continuous piecewise-linear model of each component in the multi-energy system.

3.2. Step 2: Accuracy measure on system level

After AutoMoG generated a model of each component in step 1, all component models are merged to a model of the multi-energy system. The accuracy of the multi-energy system model is then assessed. For a sufficiently accurate multi-energy system model, the relative error $\Delta C^{\text{rel,System}}$ shall be smaller than the allowed relative error δ^{rel} of the multi-energy system model:

$$\Delta C^{\text{rel,System}} \leq \delta^{\text{rel}}. \tag{21}$$

The allowed relative error δ^{rel} is a user-specified parameter. The allowed relative error can be set to $\delta^{\text{rel}} = 0$ if the user is not able to specify an appropriate value. In this case, AutoMoG provides an MILP model of the multi-energy system that is as accurate as possible without overfitting the input-output relationships of the components due to the use of the Corrected Akaike Information Criterion AIC_C (cf. Section 3.3).

The only information about the actual multi-energy system is the measured input and output data of each component. Thus, AutoMoG calculates the relative error of the multi-energy system model $\Delta C^{\text{rel,System}}$ as

$$\Delta C^{\text{rel,System}} = \frac{\Delta C^{\text{System}}}{\sum_{s\in S}\sum_{d\in D} c_s^{\text{Input}} \cdot I_{s,d}^{\text{Data}}}. \tag{22}$$

ΔC^{System} is the error of the multi-energy system model. $I_{s,d}^{\text{Data}}$ is the measured input data of component s and c_s^{Input} is the cost-based weighting factor of component s. The error of the multi-energy system model ΔC^{System} is the sum of the component model errors ΔC_s:

$$\Delta C^{\text{System}} = \sum_s \Delta C_s, \tag{23}$$

$$\text{with} \quad \Delta C_s = c_s^{\text{Input}} \cdot \sqrt{\varepsilon_s}, \qquad \forall\, s \in S. \tag{24}$$

In the following, we explain why we use the sum of squared residuals ε_s and the cost-based weighting factors c_s^{Input} to calculate the component model error ΔC_s (Eq. (24)).

3.2.1. Sum of squared residuals ε_s

AutoMoG uses the sum of squared residuals ε_s to calculate the component model error ΔC_s of component s (Eq. (24)). As a result, components with many data points tend to have a higher component model error ΔC_s and, thus, have a higher impact on the error of the multi-energy system model ΔC^{System} (Eq. (23)). Thereby, AutoMoG takes into account that frequently used components are more important for the operation of the actual multi-energy system than rarely used components. However, components with many data points are not inherently more important. Thus, using the sum of squared residuals is only meaningful if the number of data points reflects the importance of the component compared to other components and not, e.g., only a lack of measured data. Preferentially, the data of all components is measured at the same time interval, using the same time step for the measurements. Alternative error measures could be used, e.g., the mean squared error. If the number of data points is known not to reflect the importance of a component, alternative error measures could be used, e.g., the mean squared error where the sum of squared residuals is divided by the number of data points for each component.

3.2.2. Cost-based weighting factors c_s^{Input}

The obtained multi-energy system model is assumed to be used for economic optimization. To obtain a targeted model for economic optimization, AutoMoG assesses the component model errors ΔC_s in terms of operating costs.

However, measured data are commonly not available as operating costs. Instead, the consumed and produced energy is measured. Thus, AutoMoG converts the consumed and produced energy to operating costs using cost-based weighting factors c_s^{Input} (Eq. (24)). The cost-based weighting factors c_s^{Input} enable balancing the component model errors ΔC_s in terms of costs. Thereby, AutoMoG incorporates the purpose of the model into the modeling process.

However, AutoMoG is not limited to generate models for economic optimization. Other weighting factors (e.g., primary energy factors or CO_2-eq.) can be implemented easily to obtain an optimization model targeted for other objective functions.

3.2.3. Determination of cost-based weighting factors c_s^{Input} for a multi-energy system

The user has to provide a cost-based weighting factor for each energy form that is an input of at least one component in the multi-energy system. In the following, we illustrate the determination of cost-based weighting factors for a multi-energy system with gas-driven boilers and CHP engines, heat-driven absorption chillers, and electricity-driven compression chillers. Thus, cost-based weighting factors are required for gas (input for boilers and CHP engines), heat (input for absorption chillers), and electricity (input for compression chillers).

For components that are driven by energy forms which are purchased from an external grid (e.g., gas and electricity), we propose to choose the specific prices of the energy forms (c^{gas} and c^{el}) as cost-based weighting factors.

For components that are driven by energy forms which are not purchased from an external grid (e.g., heat), we need to determine a cost-based weighting factor that approximates the specific cost for this energy form in the multi-energy system. In the given example, heat is produced by different components in the multi-energy system, e.g., by CHP engines or boilers. We want to calculate one cost-based weighting factor for heat. For this purpose, we average the cost of all heat-producing components. This procedure needs to be applied for every energy form that cannot be purchased directly but is used within the energy system. The procedure can also be adapted when using AutoMoG to generate models of other systems, for example, for any intermediate chemical that is transformed into a desired fuel in chemical plants.

For heat produced by boiler b, the component-specific cost c_b^{heat} is taken from the operation of nominal load:

$$c_b^{\text{heat}} = \frac{c^{\text{gas}}}{\eta_b^{\text{nominal}}}, \qquad \forall\, b \in B. \tag{25}$$

η_b^{nominal} is the nominal efficiency of boiler b extracted from measured data. For this extraction, we search the data point with the maximum heat output. This maximum heat output is divided by the corresponding gas input to calculate the nominal efficiency η_b^{nominal}.

For heat produced by CHP engine chp, the component-specific cost c_{chp}^{heat} is calculated using the energetic method from The Association of German Engineers (2008). The energetic method allocates the cost of purchased gas to the produced heat and the produced electricity of the CHP engine based on the amount of pro-

duced thermal and electrical energy:

$$c_{chp}^{\text{heat}} = \frac{c^{\text{gas}}}{\eta_{chp}^{\text{heat}} + \eta_{chp}^{\text{el}}}, \qquad \forall\ chp \in CHP. \tag{26}$$

The thermal efficiency η_{chp}^{heat} and the electrical efficiency η_{chp}^{el} of the CHP engine are extracted from measured data in the same manner as the nominal efficiency η_b^{nominal} of boiler b.

The overall cost-based weighting factor c^{heat} for heat in the multi-energy system is calculated from the component-specific cost for heat produced by each boiler and CHP engine:

$$c^{\text{heat}} = \sum_b c_b^{\text{heat}} \cdot \frac{Q_b}{Q^{\text{System}}} + \sum_{chp} c_{chp}^{\text{heat}} \cdot \frac{Q_{chp}}{Q^{\text{System}}}, \tag{27}$$

with Q_b and Q_{chp} being the heat amount produced by boiler b and CHP engine chp, respectively. Q^{System} is the overall heat amount produced in the multi-energy system. Q_b, Q_{chp} and Q^{System} are extracted from measured data. To calculate the cost-based weighting factor c^{heat} for heat, the component-specific cost for heat are weighted by the amount of produced heat.

With the determined cost-based weighting factors, we can calculate the relative error of the multi-energy system $\Delta C^{\text{rel,System}}$ in Eq. (22). However, the proposed procedure to determine the cost-based weighting factors is not exact. Still, we find that the cost-based weighting factors are important to consider (Section 4.4).

Now, AutoMoG is able to check the accuracy of the multi-energy system model using Eq. (21). If the multi-energy system model does not fulfill the desired accuracy measure in Eq. (21), AutoMoG increases the number of piecewise-linear sections N in the multi-energy system model and, thus, refines one component to decrease the relative error of the multi-energy system model $\Delta C^{\text{rel,System}}$.

3.3. Step 3: Avoid overfitting with the corrected akaike information criterion AIC_C

In Step 3, AutoMoG aims to avoid overfitting. For this purpose, AutoMoG checks the Corrected Akaike Information Criterion AIC_C for each component, before refining a component.

Information criteria have been developed to select the most suitable model of a data set. The model with the lowest value of the used information criterion is selected. Widely known information criteria are, e.g., the Akaike Information Criterion AIC (Akaike, 1974) or the Bayesian Information Criterion BIC (Stoica and Selén, 2004). In AutoMoG, we use the Corrected Akaike Information Criterion AIC_C (Hurvich and Tsai, 1993) since it is an extension of the AIC suitable for small sample sizes. However, other information criteria can be implemented easily in AutoMoG.

AutoMoG uses the Corrected Akaike Information Criterion $AIC_{C,s,i}$ to compare the model of component s from iteration i to its refined model from iteration $i+1$. Thus, if for component s

$$AIC_{C,s,i+1} \geq AIC_{C,s,i} \tag{28}$$

holds, the improvement in model accuracy does not overcome the increase in model complexity from iteration i to iteration $i+1$. Thus, further model refinement of component s would probably risk overfitting. Consequently, AutoMoG does not refine any components for which the Corrected Akaike Information Criterion $AIC_{C,s,i}$ increases. Instead, AutoMoG refines only one of the components for which the Corrected Akaike Information Criterion $AIC_{C,s,i}$ decreases.

The Corrected Akaike Information Criterion AIC_C has been proposed by Burnham and Anderson (2003) with the assumption of normally distributed errors in the measured data as follows:

$$AIC_{C,s,i} = d \cdot \ln\left(\frac{\varepsilon_s}{d}\right) + 2K_{s,i} + \frac{2K_{s,i} \cdot (K_{s,i}+1)}{d - K_{s,i+1} - 1} \qquad \forall s \in S, \tag{29}$$

with d being the number of data points and ε_s being the sum of squared residuals. $K_{s,i}$ is the total number of regression parameters used to describe the model. The first term of the sum in Eq. (29) rewards model accuracy, whereas the other terms of the sum penalize model complexity.

AutoMoG terminates if the Corrected Akaike Information Criterion $AIC_{C,s,i}$ increases for every component, even if the allowed relative error δ^{rel} is not reached (Fig. 1). If there is at least one component for which the Corrected Akaike Information Criterion $AIC_{C,s,i}$ decreases, AutoMoG proceeds with step 4.

3.4. Step 4: Refine one component

In the first iteration, AutoMoG solves the piecewise-linear regression problem for each component s with 1 linear section (Step 1). In each subsequent iteration i, AutoMoG refines one component s by allowing one more linear section in the piecewise-linear regression problem of component s:

$$N_{s,i+1} = N_{s,i} + 1 \tag{30}$$

with $N_{s,i}$ being the number of piecewise-linear sections used to model component s in iteration i.

In step 4, the component to be refined is chosen by identifying the maximum error reduction. The maximum error reduction is the maximum difference between the component model error $\Delta C_s(N_{s,i})$ and the component model error in the next refinement $\Delta C_s(N_{s,i+1})$:

$$\max_s \left(\Delta C_s(N_{s,i}) - \Delta C_s(N_{s,i+1})\right). \tag{31}$$

Thus, the component model error of the refinement $\Delta C_s(N_s+1)$ has to be known already. Consequently, in the first iteration, AutoMoG has to solve the piecewise-linear regression problem with 2 linear sections for each component. In all subsequent iterations, only one additional piecewise-linear regression problem has to be solved for the component s that was refined in the previous iteration.

After choosing the component to be refined in step 4, AutoMoG refines the chosen component by applying step 1 (Section 3.1, Fig. 1). AutoMoG terminates once either the allowed relative error δ^{rel} is reached or all components in the multi-energy system would be overfitted when further refined.

4. Case study

In this Section, we apply AutoMoG to a case study based Goderbauer et al. (2016). Section 4.1 describes the case study. Section 4.2 presents the results AutoMoG. As benchmark approach, the common approach is employed where each component is modeled independently. In Section 4.3, we test the performance of the generated multi-energy system model in operational optimization. Section 4.4 shows a sensitivity analysis for the cost-based weighting factors used in the case study. Section 4.5 compares the performance of the actual problem and the linearized problem to AutoMoG.

4.1. Description of the case study

The case study is based on Goderbauer et al. (2016), who study a real world multi-energy system (Fig. 2).

The multi-energy system consists of 2 identical boilers, a small and a large CHP engine, a small and a large absorption chiller, and 2 identical compression chillers. In their model, Goderbauer et al. (2016) use nonlinear input-output relationships for all components. The boilers, absorption chillers and compression chillers are modeled with 1 input and 1 output each: The boilers are driven by gas and produce heat; the absorption chillers are

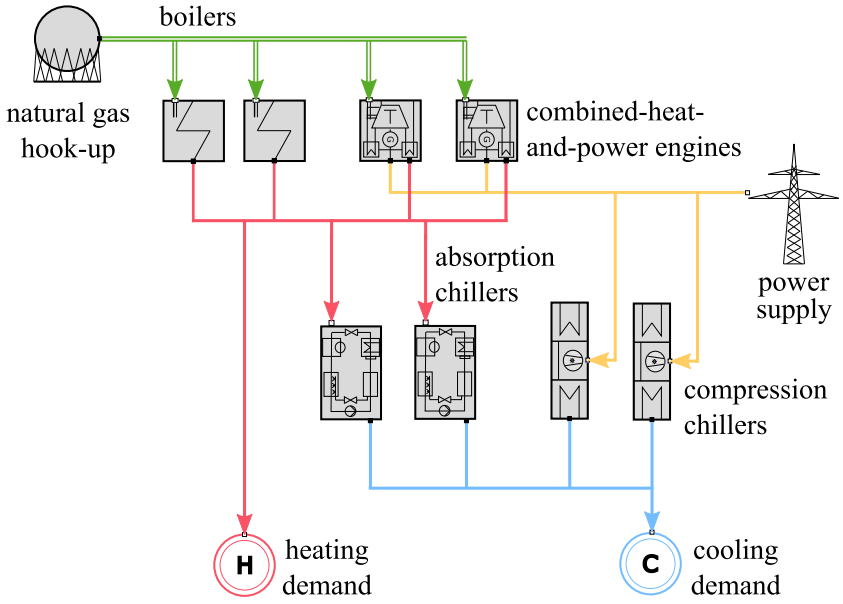


Fig. 2. Flowsheet of the multi-energy system in the case study. This multi-energy system is modeled by the proposed method AutoMoG.

driven by heat and produce cooling; the compression chillers are driven by electricity and produce cooling. Thus, for each of these components, 1 input-output relationship is required.

However, the CHP engines have 1 input (gas) and 2 outputs (heat and electricity). Thus, 2 input-output relationships are required to model a CHP engine. Goderbauer et al. (2016) model both, gas input and electricity output, as function of heat output:

$$I^{\text{Model}}_{chp,d} = f\left(O^{\text{heat}}_{chp,d}\right), \qquad \forall\ chp \in CHP,\ d \in D, \tag{32}$$

$$O^{\text{el,Model}}_{chp,d} = f\left(O^{\text{heat}}_{chp,d}\right), \qquad \forall\ chp \in CHP,\ d \in D. \tag{33}$$

Following this modeling approach for CHP engines, AutoMoG uses as cost-based weighting factors the specific gas price c^{gas} for the gas input $I^{\text{Model}}_{chp,d}$ and the specific electricity price c^{el} for the electricity output $O^{\text{el,Model}}_{chp,d}$.

In total, 10 piecewise-linear models are required to describe the input-output relationships of the 8 components in the case study (1 for each boiler, absorption chiller and compression chiller, 2 for each CHP engine).

To generate a multi-energy system model, AutoMoG requires measured input and output data of all modeled components. To obtain the required input and output data in the case study, we simulate 100 load cases of the multi-energy system with the nonlinear input-output relationships from Goderbauer et al. (2016). The 100 load cases are created by aggregating the demand time-series of heat, cooling and electricity from Goderbauer et al. (2016). The demand time-series of one year with a resolution of 1 h is aggregated to 100 typical time steps, using k-medoids (Kaufman and Rousseeuw, 1987). The simulation of the 100 typical time steps provides the required input and output data of each component for 100 load cases. We use the input and output data of each component from the simulation and add normally distributed noise in a range of ±5 % of the simulated values. The thus obtained noisy data are used as measured input and output data in the case study.

To apply AutoMoG, we have to choose a value for the allowed relative error δ^{rel} of the multi-energy system model. For this purpose, we calculate the relative error of the multi-energy system model $\Delta C^{\text{rel,System}}$ with the nonlinear input-output relationships from Goderbauer et al. (2016) compared to the obtained noisy data. We choose this relative error of the multi-energy system model as the allowed relative error δ^{rel}. Aiming for a higher accuracy than the actual functional relationship is not reasonable. However, if a meaningful allowed relative error δ^{rel} is not available, we recommend to set the allowed relative error $\delta^{\text{rel}} = 0$. In such cases, AutoMoG still provides a meaningful model as the Corrected Akaike Information Criterion AIC_C discourages overfitting (Section 3.3).

4.2. Results of the case study

The AutoMoG method is applied to the case study (Section 4.1). AutoMoG provides a multi-energy system model for the case study that reaches the allowed relative error δ^{rel}. To reach the allowed relative error δ^{rel}, AutoMoG needs 4 iterations and less than one minute. The multi-energy system model uses 14 piecewise-linear sections to describe the input-output relationships of all components. The solution provided by AutoMoG is a feasible solution for the actual problem (Eq. (5)-(19)).

In common practice, each component of a multi-energy system is modeled independently. Thus, we compare the results of AutoMoG to the independent modeling of each component. In independent modeling of each component, the relative error of the multi-energy system $\Delta C^{\text{rel,System}}$ cannot be evaluated during model generation. Thus, to ensure that the relative error of the multi-energy system $\Delta C^{\text{rel,System}}$ is lower than the allowed relative error δ^{rel}, the relative error of each component ΔI^{rel}_s has to be lower or equal than the allowed relative error δ^{rel}:

$$\Delta I^{\text{rel}}_s = \frac{\sqrt{\varepsilon_s}}{\sum_d I^{\text{Data}}_{s,d}} \leq \delta^{\text{rel}}, \qquad \forall s \in S, \tag{34}$$

with ε_s being the sum of squared residuals of component s and $I^{\text{Data}}_{s,d}$ being the input of component s in data point d. In independent modeling, the only way to guarantee the fulfillment of Eq. (34) is to ignore the Corrected Akaike Information Criterion AIC_C. However, ignoring the Corrected Akaike Information Criterion AIC_C may lead to overfitting. Thus, we generate 2 models of the multi-energy system with independent modeling:

- Independent modeling: we ignore the Corrected Akaike Information Criterion AIC_C
- Independent modeling (AIC_C): we apply the Corrected Akaike Information Criterion AIC_C.

Table 2
Solution times of operational optimization for all 5 instances and all 3 models. We implemented the operational optimization problems in Python and solved them with Gurobi 9.0.0 (Gurobi Optimization, LLC, 2020).

Instance	1	2	3	4	5	∅
AutoMoG	32 s	31 s	32 s	31 s	31 s	31 s
Independent modeling	1595 s	1674 s	1566 s	1546 s	1505 s	1577 s
Ind. modeling (AIC_C)	438 s	533 s	603 s	452 s	588 s	523 s

Note that the cost-based weighting factors c_s are not used for independent modeling because the relative errors of the components ΔI_s^{rel} are not compared to each other.

In the case study, independent modeling needs 31 piecewise-linear sections to reach $\Delta I_s^{rel} \leq \delta^{\text{rel}}$ for each component s. Independent modeling (AIC_C) uses 23 piecewise-linear sections and does not reach $\Delta I_s^{rel} \leq \delta^{\text{rel}}$ for 2 components. Thus, AutoMoG provides an MILP model of the multi-energy system that needs significantly less piecewise-linear sections than both models from independent modeling (14 vs. 31 and 23, respectively). Still, the multi-energy system model provided by AutoMoG is sufficiently accurate and satisfies the error criterion for the multi-energy system $\Delta C^{\text{rel,System}} \leq \delta^{\text{rel}}$.

Fig. 3 exemplary shows the models of AutoMoG and independent modeling for a compression chiller and the large CHP engine. The models are compared to the original models from Goderbauer et al. (2016) and the measured data. AutoMoG uses 2 linear sections to represent the input-output relationship of the compression chiller (cf. Fig. 3(a)), whereas independent modeling uses 3 linear sections in both models (cf. Fig. 3(c) and (e)). The 3 models seem appropriate to represent the input-output relationship of the compression chiller, and from visual inspect none of the models is obviously the better model. However, the accuracy measured on the system level in AutoMoG reveals that 2 linear sections are sufficient for the model of the compression chiller.

For the model of the large CHP engine, AutoMoG uses 1 linear section, whereas independent modeling uses 10 linear sections and, thereby, obviously overfits the input-output relationship of the large CHP engine (cf. Fig. 3(d)). The independent modeling is overfitting due to the need to reach the allowed relative error δ^{rel} for each component. If we discourage overfitting by the Corrected Akaike Information Criterion AIC_C in independent modeling, the model of the large CHP engine is modeled with 3 linear sections (cf. Fig. 3(f)). This finding shows that the Corrected Akaike Information Criterion AIC_C is powerful to prevent overfitting. Still, even with the Corrected Akaike Information Criterion AIC_C independent modeling increases model complexity strongly compared to AutoMoG (1 vs. 3 linear sections).

4.3. Model performance in operational optimization

The multi-energy system model is generated for efficient operational optimization. Thus, we test the model performance by analyzing the solution times of the models.

For this purpose, we solve 5 instances of operational optimization problems with the models from AutoMoG, independent modeling and independent modeling (AIC_C).

The original demand time-series of heat, cooling and electricity from Goderbauer et al. (2016) consist of 1 year with an hourly resolution. Thus, the original demand time-series contains 8760 time steps. We create 5 instances of the original demand time-series with latin-hypercube sampling (McKay et al., 2000) and variations of ±5% of the original demands.

The AutoMoG model is the fastest model for all instances (cf. Table 2). As expected, the model from independent modeling is the slowest for all instances as it contains the most piecewise-linear sections and thus the most binary variables compared to the other two models. On average, the AutoMoG model solves the operational optimization problems more than 50 times faster than independent modeling. Even compared to independent modeling (AIC_C), the AutoMoG model solves more than 15 times faster on average. The solution times show that AutoMoG generates efficient models for operational optimization.

Additionally, we test the model performance by analyzing how well the models predict the operating cost *OPEX*. For this purpose, we perform a 5-fold cross-validation for the already used 100 operating points.

For the 5-fold cross-validation, the 100 operating points are split into 5 data sets of equal size. Afterwards, 4 data sets are used for model generation and the remaining data set is used as the test set. Thus, each data set is used once as test set.

As accuracy measure for the cross-validation, we use the relative difference in operating cost for each operating point d $\Delta OPEX_d^{\text{rel}}$:

$$\Delta OPEX_d^{\text{rel}} = \frac{OPEX_d^{\text{Data}} - OPEX_d^{\text{Model}}}{OPEX_d^{\text{Data}}} \quad \forall d \in D. \tag{35}$$

with $OPEX_d^{\text{Data}}$ being the operating cost of the multi-energy system according to the measured data and $OPEX_d^{\text{Model}}$ being the operating cost of the multi-energy system according to the tested model.

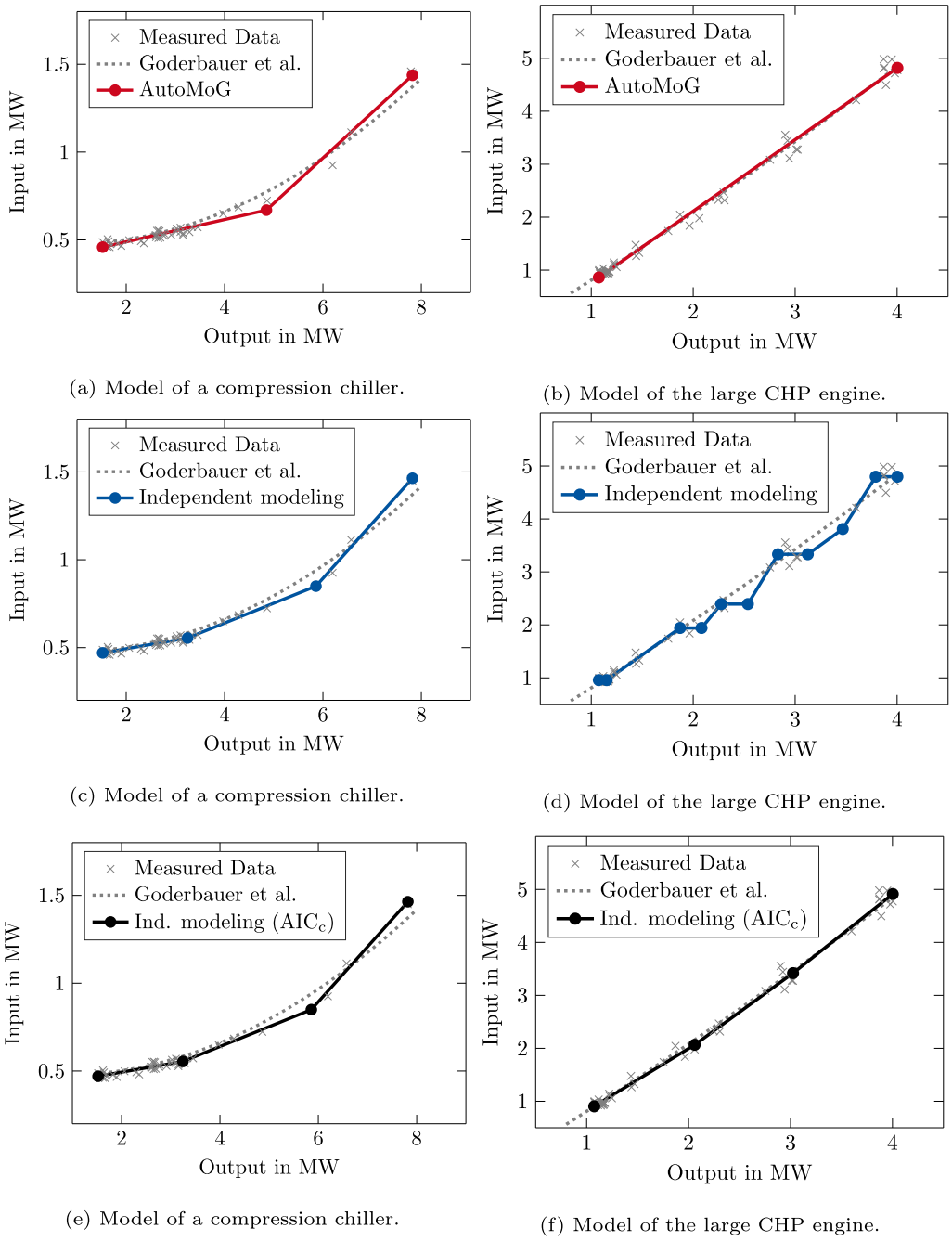
The 5-fold cross-validation is applied to the AutoMoG model and both models from independent modeling of each component (cf. Table 3).

The results of the 5-fold cross-validation show that AutoMoG uses less piecewise-linear sections in each test set (16 linear sections on average in the models of AutoMoG compared to 37.2 on average in independent modeling and 24.6 on average in independent modeling (AIC_C)). However, still, AutoMoG reduces the mean relative difference in operating cost between model and data $\overline{\Delta OPEX^{\text{rel}}}$ by approx. 20% from 0.98% to 0.78%. Furthermore, AutoMoG reaches the same mean relative difference in operating cost as independent modeling (AIC_C), although AutoMoG uses less piecewise-linear sections. Thus, in the case study, it is not necessary to model each component as accurate as possible to reach an accurate representation of the multi-energy system.

4.4. Sensitivity analysis of the cost-based weighting factors

In this section, we perform a sensitivity analysis of the cost-based weighting factors to show that they are important. At the same time, we find that the cost-based weighting factors do not need to be calculated with high accuracy.

The cost-based weighting factors essentially sort the components by importance. In the case study (Section 4.1 and 4.2), we use as weighting factors for gas $c^{\text{gas}} = 0.06$ €/kWh, for electricity $c^{\text{el}} = 0.16$ €/kWh and for heat $c^{\text{heat}} = 0.071$ €/kWh. Thus, the errors of components that use electricity as input are weighted more than two times higher than the errors of components that use gas or heat as an input. AutoMoG needs 4 iterations and, thus, performs 4 refinements in total for the 10 piecewise-linear models. The 4 refinements are: 1. large absorption chiller, 2. small absorption chiller, 3. a compression chiller, 4. a boiler.



(a) Model of a compression chiller.

(b) Model of the large CHP engine.

(c) Model of a compression chiller.

(d) Model of the large CHP engine.

(e) Model of a compression chiller.

(f) Model of the large CHP engine.

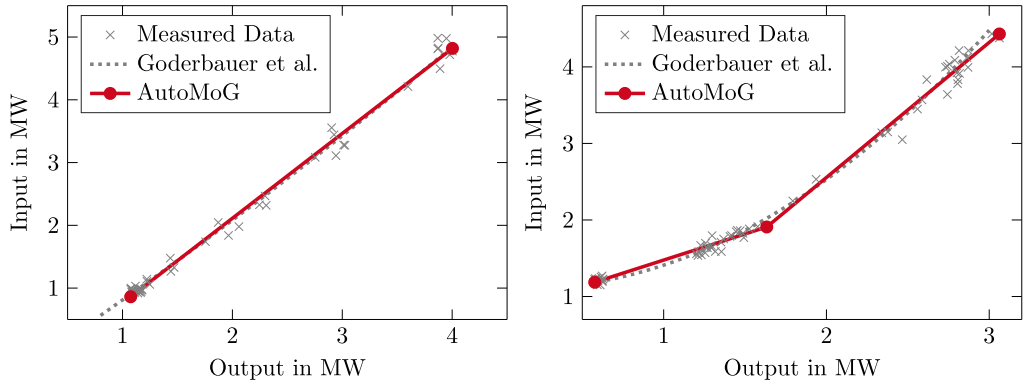
Fig. 3. Derived models of a compression chiller (a,c,e) and the large CHP engine (b,d,f) in the case study. The models from AutoMoG are shown in red (a), (b) and the models from independent modeling of each component are shown in blue (c), (d) and black, respectively (e), (f). The models are compared to the original model from Goderbauer et al. (2016) (grey dashed line) and the measured data (grey crosses). Independent modeling of each component shows overfitting for the large CHP engine (d). If the Corrected Akaike Information Criterion AIC_C is applied to independent modeling, overfitting can be prevented. Still, AutoMoG uses less linear sections to model the large CHP engine. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Results of the 5-fold cross-validation for the case study. $\overline{\Delta OPEX^{rel}}$ is the mean relative difference in operating cost between model and data. N is the number of piecewise-linear sections in the multi-energy system model.

Test set		1	2	3	4	5	⌀
$\overline{\Delta OPEX^{rel}}$/%	AutoMoG	0.93	0.91	0.76	0.55	0.74	0.78
	Independent modeling	0.95	0.95	1.11	0.71	1.19	0.98
	Ind. modeling (AIC_C)	0.79	0.78	0.86	0.77	0.70	0.78
N	AutoMoG	16	15	17	16	16	16
	Independent modeling	34	35	40	32	45	37.2
	Ind. modeling (AIC_C)	26	21	24	27	25	24.6

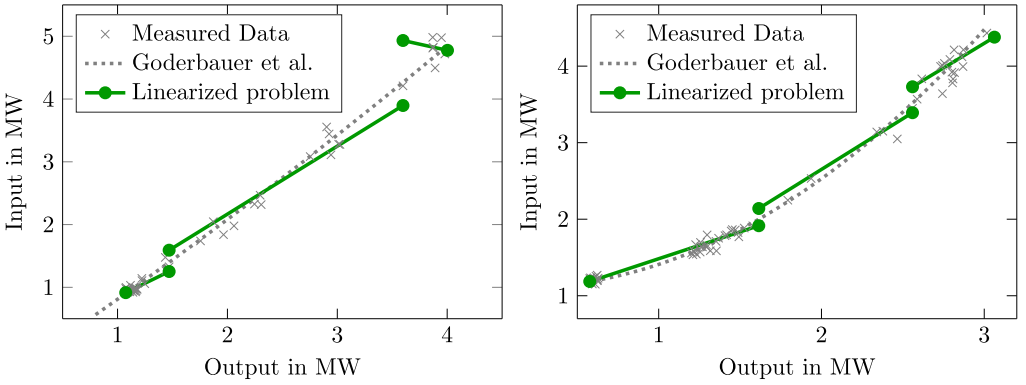
Table 4
Results of solving the actual problem with different MINLP solvers without starting solution and with the AutoMoG solution as starting solution.

Solver	No starting solution	Starting solution from AutoMoG
SCIP	MINLP is infeasible	Starting solution is feasible Starting solution is optimal
BARON	MINLP is infeasible	Starting solution is feasible Starting solution is optimal
DICOPT	No feasible solution found within Iteration Limit of 200	Starting solution is feasible No better solution found within Iteration Limit of 200
BONMINH	Terminated by solver	-



(a) Model of the large CHP engine.

(b) Model of an absorption chiller.



(c) Model of the large CHP engine.

(d) Model of an absorption chiller.

Fig. 4. Derived model of the large CHP engine and an absorption chiller from AutoMoG (red) and from the linearized problem (green). Furthermore, the original models from Goderbauer et al. (2016) (grey dashed line) and the measured data (grey crosses) are presented. The solution of the linearized problem uses more linear sections and shows discontinuities at the breakpoints. Thus, the solution of the linearized problem is not a feasible solution of the actual problem. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

For a sensitivity analysis of the cost-based weighting factors, we change each of the 3 weighting factors by ±20% and run AutoMoG once for each changed weighting factor. In each of the 6 additional runs, AutoMoG refines the same 4 components. However, the order of refinement changes, when we increase the weighting factor for electricity to $c^{\text{el}} = 0.192$ € /kWh or reduce the weighting factor for heat to $c^{\text{heat}} = 0.057$ € /kWh. In these 2 cases, a compression chiller is refined before the small absorption chiller is refined.

In another additional run, we ignore the weighting factors. Thus, the errors of all components have the same weight, independently of the input of the components. In this case, AutoMoG refines the models of the following 4 components in the following order: 1. large absorption chiller, 2. small absorption chiller, 3. a boiler, 4. another boiler. We can see that the compression chiller is refined only when we use the cost-based weighting factors. Thus, the cost-based weighting factors show that economically it makes sense to refine the compression chiller since electricity is more valuable than gas. If we ignore the weighting factors, different components are refined. Refining different components leads to a different model of the overall multi-energy system.

The sensitivity analysis shows that the weighting factors decide which components are refined in which order. However, changing the weighting factors by ±20% does not fundamentally change the results of the case study, which leads us to the conclusion that a rough estimation of the weighting factors is sufficient. In summary, we conclude that cost-based weighting factors are important if their values significantly differ for the input energy forms.

4.5. Computational study of the actual problem

AutoMoG is essentially a primal heuristic for the actual problem (Eq. (5)-(19)) using problem decomposition to find a good feasible solution. However, there is no proof that the AutoMoG solution is optimal. Furthermore, the AutoMoG solution is only feasible for the actual problem if AutoMoG terminates because the allowed relative error δ^{rel} is reached. The allowed relative error δ^{rel} is not reached if AutoMoG stops because all components would be overfitted when further refined. In this case, the AutoMoG solution is infeasible for the actual problem because the allowed relative error δ^{rel} cannot be reached, and the corresponding constraint (Eq. (6)) is violated.

To assess the solution quality and the performance of the AutoMoG method, we implemented the actual problem (Eq. (5)-(19)) and the linearization of the actual problem (cf. Section 2) for the case study in GAMS. In the linearization of the actual problem, we followed the approach of Yang et al. (2016), i.e., we used absolute errors instead of squared residuals and ignored the constraints that enforce continuity at the breakpoints of the piecewise-linear functions. Thus, the linearization does not guarantee continuity at the breakpoints, which leads to different component models. Still, we learn about the performance of the MILP approach compared to AutoMoG.

To solve the actual problem, we used the MINLP solvers SCIP (Gleixner et al., 2018), BARON (Tawarmalani and Sahinidis, 2005), DICOPT (Kocis and Grossmann, 1989) and BONMINH (COIN-OR (Project Manager P. Bonami), 2016). All MINLP solvers were used with default subsolvers and default settings, if not stated otherwise. The default subsolvers and default settings can be found in the solver manuals of the GAMS documentation (GAMS Development Corporation, 2016). Each run of the actual problem with an MINLP solver was executed without a time limit.

None of the used MINLP solvers were able to find any feasible solution for the actual problem (cf. Table 4). The solvers SCIP and BARON even identified the actual problem as infeasible. However, we showed that the AutoMoG solution is feasible for the actual problem by using the AutoMoG solution as starting solution for SCIP, BARON and DICOPT. Even with the AutoMoG solution as starting solution, none of the used MINLP solvers was able to find another solution than the AutoMoG solution (cf. Table 4).

Thus, solving the actual problem with standard MINLP solvers is impractical and even impossible for the present case study.

To solve the linearized problem, we used CPLEX 12.6.0.1 (IBM Cooperation, 2016) with a time limit of 48 h. CPLEX reached the time limit with a best feasible solution of 17 piecewise-linear sections for the multi-energy system model (the AutoMoG solution uses 14 piecewise-linear sections).

In Fig. 4, we compare the models of the large CHP engine and an absorption chiller derived by AutoMoG and by the linearized problem. Within the time limit of 48 h, the best model found by the linearized problem has 3 linear sections for the large CHP engine and shows discontinuities (cf. Fig. 4 (c)), whereas AutoMoG finds a model for the large CHP engine with 1 linear section within 1 min (cf. Fig. 4 (a)). The AutoMoG model of the large CHP engine is obviously more appropriate to represent the input-output relationship of the large CHP engine than the model derived by the linearized problem. For the model of the absorption chiller, the solution of linearized problem uses 3 linear sections (cf. Fig. 4 (d)), whereas AutoMoG uses 2 linear sections (cf. Fig. 4 (b)).

In summary, even with the stated simplifications, we could not find a better solution of the linearized problem within 48 h than the solution AutoMoG finds within 1 min. In contrast to the AutoMoG solution, the solution of the linearized problem is not feasible for the actual problem, because the solution of the linearized problem shows discontinuities at the breakpoints of the piecewise-linear models (cf. Fig. 4). If we used the approaches of Kong and Maravelias (2020) and Rebennack and Krasko (2020) for the linearized problem, we could overcome the problem of discontinuities at the breakpoints. Since these methods further contrain the solutions presented in this work, we do not expect a significantly improved performance for the linearized problem. However, it would certainly be interesting to explore these methods in future work. Neither the actual problem nor the linearized problem is applicable in practice. In summary, we showed that the method AutoMoG provides an accurate and computationally efficient model of the multi-energy system in short time.

5. Conclusions

The AutoMoG method is proposed for automated data-driven model generation of multi-energy systems using piecewise-linear regression. AutoMoG decomposes the data-driven model generation problem of a multi-energy system to one model generation problem of each component. Still, the error of model generation is evaluated for the overall multi-energy system. For this purpose, AutoMoG uses cost-based weighting factors to balance the errors caused by each component model. Through the decomposition, AutoMoG provides an accurate solution for the data-driven model generation problem of the multi-energy system in short time. The solution provided by AutoMoG is an MILP model of the multi-energy system that is usable for optimization.

In the case study, AutoMoG needs significantly less piecewise-linear sections (57% on average) to reach an allowed model error compared to the commonly employed independent modeling of each component. As a result, the model of AutoMoG solves the operational optimization more than 50 times faster on average. Still, the models from AutoMoG are more accurate (20% on average) in terms of operating cost than the models provided by independent modeling of each component.

The results of the case study show that it is not mandatory to model each component with high accuracy to reach an accurate multi-energy system model. Instead, each component's modeling error should be seen in context of the multi-energy system model.

The proposed method AutoMoG is only applicable if measured input and output data of the components are available. In its present form, AutoMoG is limited to systems that contain components with one independent variable. For example, variable-speed pumps cannot be modeled, since the consumed power of a variable-speed pump depends on two independent variables, i.e., its rotational speed and volume flow. However, in principle, AutoMoG should be extendable to handle components with more than one independent variable, which is currently explored.

AutoMoG is an easy-to-use method to generate efficient MILP models. Furthermore, AutoMoG is not limited to generate models of multi-energy systems but can generate models of any engineering system. AutoMoG drastically decreases the effort for data-driven model generation, enabling a wider spread of optimization models in real-world applications.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Andreas Kämper: Writing - original draft, Conceptualization, Methodology, Software, Investigation, Visualization, Data curation, Project administration, Funding acquisition. **Ludger Leenders:**

Writing - review & editing, Conceptualization, Methodology, Visualization. **Björn Bahl:** Conceptualization, Methodology. **André Bardow:** Conceptualization, Methodology, Writing - review & editing, Supervision, Resources, Funding acquisition.

Acknowledgements

This study is funded by the German Federal Ministry of Economic Affairs and Energy (ref. no.: 03ET4068A). The support is gratefully acknowledged.

Appendix A. Nomenclature

variables

symbol	explanation	unit
$A_{s,n}$	gradient of linear section n	-
$B_{s,n}$	intercept of linear section n	kW
$I_{s,d}^{\text{Model}}$	input value of the model of component s for the data point d	kW
N	number of linear sections in the overall multi-energy system model	-
N_s	number of linear sections used to model component s	-
$N_s^{\max}$	maximum number of linear sections to model component s	-
$O_{s,n}^{\text{UB}}$	upper bound for the output value of linear section n	kW
ΔC^{System}	error of the multi-energy system model	€
ΔC_s	component model error	€
ΔI_s^{rel}	relative error of component s	-
$\gamma_{s,n,d}$	binary variable to assign data points to linear sections	-
ϵ_s	sum of squared residuals	kW^2
$\kappa_{s,n}$	binary variable to denote if the linear section n is used	-

parameters

symbol	explanation	unit
$a_{s,n}$	gradient of the linear section n	-
AIC_C	Corrected Akaike Information Criterion	-
$b_{s,n}$	intercept of the linear section n	kW
c_s^{Input}	cost-based weighting factor for input of component s	€ /kW
c^{el}	cost-based weighting factor for electricity	€ /kW
c^{gas}	cost-based weighting factor for gas	€ /kW
c^{heat}	cost-based weighting factor for heat	€ /kW
c_b^{heat}	specific cost for heat produced by boiler b	€ /kW
c_{chp}^{heat}	specific cost for heat produced by CHP engine chp	€ /kW
$\|D\|$	number of data points	-
$I_{s,d}^{\text{Data}}$	input value of the data point d of the component s	kW
$K_{s,i}$	total number of regression parameters to model component s	-
m	Big-M parameter	-
$O_{s,d}^{\text{Data}}$	output value of the data point d of the component s	kW
$OPEX$	operating cost	€
$OPEX_d^{\text{Data}}$	operating cost according to measured data	€
$OPEX_d^{\text{Model}}$	operating cost according to tested model	€
Q_b	amount of heat produced by boiler b	kWh
Q_{chp}	amount of heat produced by CHP engine chp	kWh
Q^{System}	amount of heat produced in the multi-energy system	kWh
δ^{rel}	predefined relative error of the multi-energy system	-
$\Delta OPEX_d^{\text{rel}}$	relative difference in operating cost	-
$\overline{\Delta OPEX^{\text{rel}}}$	mean relative difference in operating cost	-
η_b^{nominal}	nominal efficiency of boiler b	-
η_{chp}^{el}	electrical efficiency of CHP engine chp	-
η_{chp}^{heat}	thermal efficiency of CHP engine chp	-

sets and elements

symbol	explanation
$b \in B$	boiler
$chp \in CHP$	combined heat and power engine
$d \in D$	measured data point
i	number of iteration in AutoMoG
$n \in N$	linear section
$s \in S$	component

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19 (6), 716–723.

Burnham, K.P., Anderson, D.R., 2003. Model selection and multimodel inference: A practical information-theoretic approach. Springer Science & Business Media.

COIN-OR (Project Manager P. Bonami), 2016. Basic Open-source Nonlinear Mixed INteger programming.

Camponogara, E., Nazari, L.F., 2015. Models and algorithms for optimal piecewise–linear function approximation. Mathematical Problems in Engineering 2015, 9.

Cozad, A., Sahinidis, N.V., Miller, D.C., 2014. Learning surrogate models for simulation-based optimization. AIChE J. 60 (6), 2211–2227.

D'Errico, J., 2009. SLM - shape language modeling.

GAMS Development Corporation, 2016. General Algebraic Modeling System (GAMS) Release 24.7.3. Washington DC, USA.

Gao, X., Feng, Z., Wang, Y., Huang, X., Huang, D., Chen, T., Lian, X., 2018. Piecewise linear approximation based MILP method for PVC plant planning optimization. Industrial & Engineering Chemistry Research 57 (4), 1233–1244.

Gkioulekas, I., Papageorgiou, L.G., 2018. Piecewise regression through the akaike information criterion using mathematical programming. IFAC-PapersOnLine 51 (15), 730–735.

Gleixner, A., Bastubbe, M., Eifler, L., Gally, T., Gamrath, G., Gottwald, R.L., Hendel, G., Hojny, C., Koch, T., Lübbecke, M.E., Maher, S.J., Miltenberger, M., Müller, B., Pfetsch, M.E., Puchert, C., Rehfeldt, D., Schlösser, F., Schubert, C., Serrano, F., Shinano, Y., Viernickel, J.M., Walter, M., Wegscheider, F., Witt, J.T., Witzig, J., 2018. The SCIP Optimization Suite 6.0. Technical Report. Optimization Online.

Goderbauer, S., Bahl, B., Voll, P., Lübbecke, M.E., Bardow, A., Koster, A.M., 2016. An adaptive discretization MINLP algorithm for optimal synthesis of decentralized energy supply systems. Computers & Chemical Engineering 95, 38–48.

Goderbauer, S., Comis, M., Willamowski, F.J., 2019. The synthesis problem of decentralized energy systems is strongly NP-hard. Computers & Chemical Engineering 124, 343–349.

Gurobi Optimization, LLC, 2020. Gurobi Optimizer Reference Manual.

Huang, X., Xu, J., Wang, S., 2010. Identification algorithm for standard continuous piecewise linear neural network. In: Proceedings of the 2010 American Control Conference, pp. 4931–4936.

Hurvich, C.M., Tsai, C.-L., 1993. A corrected akaike information criterion for vector autoregressive model selection. Journal of Time Series Analysis 14 (3), 271–279.

IBM Cooperation, 2016. IBM ILOG and CPLEX Optimization and Studio and CPLEX User's and Manual and Version 12 and Release 7.

ISO 50001:2018, 2018. Energy management systems–Requirements with guidance for use.

Katz, J., Pappas, I., Avraamidou, S., Pistikopoulos, E.N., 2020. Integrating deep learning models and multiparametric programming. Computers & Chemical Engineering 136, 106801.

Kaufman, L., Rousseeuw, P., 1987. Clustering by means of medoids. North-Holland: Reports of the Faculty of Mathematics and Informatics.

Kleijnen, J.P.C., Beers, W.C.M.v., 2004. Application-driven sequential designs for simulation experiments: kriging metamodelling. Journal of the Operational Research Society 55 (8), 876–883.

Kocis, G., Grossmann, I., 1989. Computational experience with dicopt solving MINLP problems in process systems engineering. Computers & Chemical Engineering 13 (3), 307–315.

Kong, L., Maravelias, C.T., 2020. On the derivation of continuous piecewise linear approximating functions. INFORMS J. Comput. 32 (3), 531–546.

Mancarella, P., 2014. MES (Multi-energy systems): an overview of concepts and evaluation models. Energy 65, 1–17.

McBride, K., Sundmacher, K., 2019. Overview of surrogate modeling in chemical process engineering. Chem. Ing. Tech. 91 (3), 228–239.

McKay, M.D., Beckman, R.J., Conover, W.J., 2000. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 42 (1), 55–61.

Mitsos, A., Asprion, N., Floudas, C.A., Bortz, M., Baldea, M., Bonvin, D., Caspari, A., Schäfer, P., 2018. Challenges in process optimization for new feedstocks and energy sources. Computers & Chemical Engineering 113, 209–221.

Rebennack, S., Krasko, V., 2020. Piecewise linear function fitting via mixed-Integer linear programming. INFORMS J. Comput. 32 (2), 507–530.

Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14 (3), 199–222.

Smolin, Y.Y., Lau, K.K., Soroush, M., 2019. First-principles modeling for optimal design, operation, and integration of energy conversion and storage systems. AIChE J. 65 (7), e16482.

Stoica, P., Selén, Y., 2004. Model-order selection: a review of information criterion rules. IEEE Signal Process Mag 21 (4), 36–47.

Tawarmalani, M., Sahinidis, N.V., 2005. A polyhedral branch-and-cut approach to global optimization. Math. Program. 103 (2), 225–249.

The Association of German Engineers, 2008. VDI 4608 Part 2: Energy systems - Combined heat and power - Allocation and evaluation.

Voll, P., Klaffke, C., Hennen, M., Bardow, A., 2013. Automated superstructure-based synthesis and optimization of distributed energy supply systems. Energy 50, 374–388.

Wilson, Z.T., Sahinidis, N.V., 2017. The ALAMO approach to machine learning. Computers & Chemical Engineering 106, 785–795. ESCAPE-26

Yang, L., Liu, S., Tsoka, S., Papageorgiou, L.G., 2016. Mathematical programming for piecewise linear regression analysis. Expert Syst. Appl. 44, 156–167.

Zhang, Q., Grossmann, I.E., Sundaramoorthy, A., Pinto, J.M., 2016. Data-driven construction of convex region surrogate models. Optimization and Engineering 17 (2), 289–332.