A learning-from-data approach with soft clustering and path relinking to the history-matching problem

Cristina C. B. Cavalcante¹ · Cid C. de Souza¹ · Célio Maschio² · Denis Schiozer² · Anderson Rocha¹

Received: 5 November 2020 / Accepted: 28 April 2021 / Published online: 14 May 2021
© The Author(s) 2021

Abstract

History matching is an important reservoir engineering process whereby the values of uncertain attributes of a reservoir model are changed to find models that have a better chance of reproducing the performance of an actual reservoir. As a typical inverse and ill-posed problem, different combinations of reservoir uncertain attributes lead to equally well-matched models and the success of a history-matching approach is usually measured in terms of its ability to efficiently find multiple history-matched models inside the search space defined by the parameterization of the problem (multiple-matched models have a higher chance of better representing the reservoir performance forecast). While studies on history-matching approaches have produced remarkable progress over the last two decades, given the uniqueness of each reservoir's history-matching problem, no strategy is proven effective for all cases, and finding alternative, efficient, and effective history-matching methodologies is still a research challenge. In this work, we introduce a learning-from-data approach with path relinking and soft clustering to the history-matching problem. The proposed algorithm is designed to learn the patterns of input attributes that are associated with good matching quality from the set of available solutions, and has two stages that handle different types of reservoir uncertain attributes. In each stage, the algorithm evaluates the data of all-available solutions continuously and, based on the acquired information, dynamically decides what needs to be changed, where the changes shall take place, and how such changes will occur in order to generate new (and hopefully better) solutions. We validate our approach using the UNISIM-I-H benchmark, a complex synthetic case constructed with real data from the Namorado Field, Campos Basin, Brazil. Experimental results indicate the potential of the proposed approach in finding models with significantly better history-matching quality. Considering a global misfit quality metric, the final best solutions found by our approach are up to 77% better than the corresponding initial best solutions in the datasets used in the experiments. Moreover, compared with previous work for the same benchmark, the proposed learning-from-data approach is competitive regarding the quality of solutions found and, above all, it offers a significant reduction (up to 30× less) in the number of simulations.

Keywords History Matching · Reservoir Simulation · Learning-from-Data Strategy · Path Relinking · Soft Clustering

Introduction

Reservoir simulation is an important tool in the petroleum industry. Combining data from different areas, such as mathematics, physics, geology, and reservoir engineering, it involves the construction of numerical reservoir models

that aim to represent a real reservoir and can be simulated to forecast the reservoir's performance in different production stages, or under various operating conditions. The main role of reservoir simulation is to support decision-making in reservoir management tasks, such as the economic analysis of the recovery project or optimization of the production strategy.

In a reservoir model, the physical space of a real reservoir is represented by a grid—a set of discrete cells (grid blocks)—and there are two main types of input data (Ertekin et al. 2001): geological data, that describe reservoir rock properties (e.g., porosity and permeability) and how they are distributed along the reservoir; and engineering data, that

✉ Cristina C. B. Cavalcante
cristina.cavalcante@ic.unicamp.br

¹ Institute of Computing, University of Campinas, Campinas, São Paulo, Brazil

² School of Mechanical Engineering, Center for Petroleum Studies, University of Campinas, Campinas, São Paulo, Brazil

provide the description of the static and dynamic properties related to the reservoir's fluids.

The reservoir model input data is gathered and validated by a multidisciplinary team. However, hydrocarbon reservoirs are complex in nature, with heterogeneous rock properties, regional variations of fluid properties, and very often located beneath thousands of feet of overburden (Oliver et al. 2008). Except for seismic data, direct measures of reservoir properties are only available at well locations, which accounts for less than 1% of the reservoir's volume (Schulze-Riegert and Ghedan 2007). The data used in the construction of a reservoir model is only the best interpretation from engineers and geologists of the available data. Many uncertainties are still present, and an initial reservoir model can rarely reproduce the real reservoir performance.

History matching is the classical reservoir engineering process whereby the values of uncertain data of a reservoir model are changed until the model performance matches existing (history) reservoir production data. It is a typical inverse problem (Oliver et al. 2008), where one knows the output (existing reservoir production data, e.g., well oil and water rates, and pressure), and must determine the input (reservoir model input data) that leads to the observed data. It plays a fundamental role in reservoir simulation studies as the better the model performance matches observed reservoir data, the higher the chance the model reproduces the actual reservoir performance, and therefore the more reliable will be the forecasts from the model regarding the reservoir future performance. The problem is also mathematically ill-posed, and different values for the reservoir model input data may result in models that equally reproduce the reservoir history data. Additionally, since the history data is itself subject to measurement errors, finding a good history-matched model is not sufficient to ensure reliable forecasts. Ideally, a history-matching process shall be capable of finding different, yet comparably good, history-matched models, inside the search space defined by the parameterization of the problem. Multiple-matched models have a higher chance of better representing the reservoir's performance forecast and can be used to perform a risk analysis while forecasting reservoir performance under (reduced) uncertainty. In the context of history-matching, the terms *model* and *solution* carry the same meaning, and throughout this work, they are used interchangeably.

Oliver and Chen (2011) reviewed the recent progress on history-matching. Although the earliest studies in the area date back to the 1960s (Jacquard and Jain 1965), the authors note a remarkable development of history-matching methods in the past decade, mainly pushed by the advances in computing techniques and geostatistical modeling. History-matching processes, that if carried out manually, could require months of intense work to handle a limited number of uncertain properties and output variables, can now

be executed using one of a wide range of assisted history-matching methods that automate time-consuming tasks (e.g., generation of new models, dispatch of simulations, and comparison of simulation models' performance with real data), and enable the solution of problems of much greater complexity in a much shorter computational time.

The diversity of methods in the literature about assisted history-matching is so vast, that there is no standard classification for them (Emerick 2017). However, regarding the search mechanism, they are usually divided into three groups (Abdollahzadeh 2014): deterministic methods, stochastic methods, and data assimilation methods.

Deterministic methods are the ones that do not contain random or statistical steps (whenever executed with a particular input data, they produce the same result), and very often use the analytical properties of an optimization's objective function to guide the search. In the context of history-matching, gradient-based methods are the oldest and most common deterministic approaches applied to solve the problem in a computer-assisted way (Anterion et al. 1989; Zhang and Reynolds 2002; Dickstein et al. 2018). They require the calculation of the derivatives of the history-matching objective function with respect to the uncertain variables and, although they are fast methods, for complex history-matching problems, with highly nonlinear objective functions, they are easily trapped within local minimum solutions, which is a limitation in the context of uncertainty assessment.

Stochastic methods are algorithms that incorporate random steps into some component with the purpose of increasing the exploration of the search space. This type of method has been extensively applied to the history-matching problem (Oliver and Chen 2011), mainly because of its straightforward implementation and success with other engineering problems. Examples include: genetic algorithms (Schulze-Riegert and Ghedan 2007); particle swarm optimization and neighborhood algorithm (Mohamed et al. 2010); ant-colony optimization (Hajizadeh et al. 2011); simulated annealing (Maschio and Schiozer 2018); estimation of distribution algorithms (Abdollahzadeh et al. 2014); hybrid differential evolution algorithm (Santosh and Sangwai 2016); and random search and evolutionary algorithm (Ranazzi and Pinto 2017). The main drawback of stochastic methods, in the case of history-matching problems, is the high number of simulations that may be required before converging into good-quality solutions.

Data assimilation methods are among notable recent developments for history-matching (Oliver and Chen 2011) and they are gaining increased attention. With this type of method, solutions are generated in a time series approach, with the observed data being used to sequentially update the best estimate of the model's uncertain parameters. The ensemble Kalman filter is perhaps the most commonly used

data assimilation method in the oil and gas industry (Aanonson et al. 2009; Evensen et al. 2007), but other examples of methods in this group, applied to the history-matching problem, include ensemble-based methods such as ensemble-smoother (Skjervheim and Evensen 2011) or ensemble-smoother with multiple data assimilation (Emerick and Reynolds 2013; Le et al. 2016; Emerick 2018; Soares et al. 2018). Although these methods seem to handle a large number of the model's uncertain variables, and generate multiple history-matched models, they tend to converge to local minimum solutions, excessively reducing the variability of the uncertain variables (Emerick 2016).

Apart from the three groups in the search-mechanism-based division, Oliver and Chen (2011) highlight another important group of history-matching methods: the ones able to quantify uncertainties. In these methods, the history-matching problem is tackled from a probabilistic point of view, and new solutions are generated through sampling from conditional distributions. Approaches that use Markov Chain Monte Carlo methods to sample the posterior distribution of the reservoir model's uncertain attributes are the most common (Elsheikh et al. 2012; Maschio and Schiozer 2014), but methods using other sampling strategies also exist, as the one developed by Maschio and Schiozer (2016), which combines a discrete Latin Hypercube sampler with a nonparametric density estimator. Although these methods may enable a successful sampling of posterior distributions, and the analysis of the reduction of uncertainties, they may be computationally expensive, especially for practical and complex history-matching problems.

Two final ramifications of history-matching methods worth mentioning are the “geologically consistent” and the “reduced physics” groups, proposed by Petrovska (2009). “Geologically consistent” methods are explicitly designed to account for geological constraints or to manipulate geostatistical properties. Examples of these methods can be found in Hoffman and Caers (2003), Caeiro et al. (2015), Maschio et al. (2015), and Oliveira et al. (2017). As the group name suggests, the main advantage of these methods is to produce history-matched models that are geologically consistent, respecting correlations between the reservoir's petrophysical properties and other features, such as variogram modeling and spatial continuity. The drawback is their usual dependence on complex geostatistical modeling software.

The “reduced physics” group encompasses history-matching methods where the flow simulator is replaced by a proxy or emulator (Cullick et al. 2006; Maschio and Schiozer 2014). Considering that simulating a single model is already a costly process, which can take hours or even days, the main purpose of these methods is to speed up the evaluation of the performance of new models generated during the process. The limitation is twofold: first, representing spatial properties through proxies is a major challenge and

still's object of research; second, generating accurate proxies, especially for nonlinear problems as is the case of history-matching, is a non-trivial task: The proxy evaluation of a given model may be very different from the one obtained by the simulator, and for this reason, the process guided by proxy evaluations may end up with solutions that are not actually well-matched solutions.

The aforementioned sets of history-matching methods are not disjoint, and most history-matching algorithmic approaches fall in more than one of the groups, as is the case of the proposal from Li et al. (2018): a design of experiment-based workflow for history-matching, that includes (i) sensitive analysis to infer the influence of the input attributes on the response variables, (ii) a proxy building to approximate the reservoir response; and (iii) Monte Carlo sampling.

It is also important to note a growing tendency in the literature of model-based reservoir management towards the integration of the different parts of the modeling process into a single, cyclic, and near-continuous workflow known as ‘closed-loop reservoir management’ (CLRM) (Emerick 2016). Designed to periodically update reservoir models as new wells are drilled and start producing, the main goal of this workflow is to improve and accelerate reservoir management decision-making processes (Jansen et al. 2005, 2009; Zachariassen et al. 2011; Hou et al. 2015). In a more generic version, the ‘closed-loop field development’ workflow (CLFD) considers, in the model updating process, not only the existing wells (as is the case of the CLRM workflow), but also optimizes the production strategy (e.g. position and type) of future wells (Shirangi and Durlofsky 2016; Schiozer et al. 2019). In CLRM or CLFD workflows, history matching is just one of the steps continuously executed. In this context, in addition to the need of providing calibrated models to support reliable forecasts of reservoir performance, it is of paramount importance that the history-matching process is both effective and efficient, so that it can be integrated into these reservoir management workflows.

If, on one hand, the different and numerous types of approaches available in the literature of history-matching indicate the importance of the problem, both from academic and industrial perspectives, on the other, each history-matching process is unique (given the singularities of the reservoir's uncertain attributes and history data), and none of the existing approaches can be proven the best for all problems. Considering the potential high computational cost to (accurately) evaluate a single model, research efforts toward developing alternative history-matching approaches capable of finding multiple and equally well-matched models, using a small number of simulations, is more than justified.

In this work, we propose a learning-from-data approach with path relinking and soft clustering to the history matching problem. Our strategy is a substantial evolution of the works presented in Cavalcante et al. (2017, 2019). We use a

similar learning-from-data approach that dynamically analyzes the data of available solutions to uncover patterns of reservoir uncertain properties that lead to good matching of the output variables considered in the history-matching process, and that can be used to guide the generation of new solutions. As stated by Cavalcante et al. (2017), the fundamental purpose of this learning-from-data approach is to give the algorithm the “*potential to become a specialist on the structure of the history-matching problem, what according to Ho and Pepyne (2002), would definitely improve the chances of it being considered superior when compared to other strategies for the same problem*”.

Compared to Cavalcante et al. (2017), there are three major differences introduced in this work:

- (i) The first concerns the type of solutions considered during the process: While the strategy from Cavalcante et al. (2017) analyzes input attribute patterns present in the best available solution for each well, our new learning-from-data approach, apart from the best solution for each well, also considers the best solutions for each variable involved in the history-matching problem. This characteristic broadens the learning spectrum of our algorithm, potentializing its ability to find solutions with better history-matching quality.
- (ii) The second difference from the approach proposed by Cavalcante et al. (2017) is related to the strategy used to delimit reservoir regions changed during the history-matching process. Both works dynamically delimit these regions' clustering reservoir grid blocks considering their spatial and petrophysical properties. However, while Cavalcante et al. (2017) adopted a hard assignment clustering approach, in this work, we propose a soft clustering strategy to delimit the reservoir regions. Our primary motivation for this proposal is to potentially consider the impact of each reservoir grid block on the performance of one or more wells.
- (iii) The third and final difference from Cavalcante et al. (2017) involves the actual strategy used to generate new solutions. While Cavalcante et al. (2017) followed an expectation–maximization approach, in this work, we propose the use of the path-relinking approach (Glover and Laguna 1993) to connect previously available solutions, with the hope of finding better history-matched solutions.

Compared to Cavalcante et al. (2019), there are two major differences introduced in this work:

- (i) The first difference is related to the strategy used to delimit reservoir regions changed during the

history-matching process. The Continuous Learning Algorithm from Cavalcante et al. (2019) uses the hard assignment clustering approach proposed by Cavalcante et al. (2017), and in this work, we introduce a soft clustering strategy to delimit the reservoir regions. With the proposed soft clustering algorithm, the same reservoir grid block can belong to regions of more than one well, which results in regions potentially bigger (with more blocks) than the corresponding regions obtained with the hard clustering algorithm from Cavalcante et al. (2017, 2019). This new way of thinking potentializes the improvements on the matching quality while performing local changes during the history-matching process.

- (ii) The second major difference from Cavalcante et al. (2019) involves the actual strategy used to generate new potential solutions. The approach from Cavalcante et al. (2019) directly uses the patterns of attributes present in the best solution for each well and in the k-best solutions for each variable. The path-relinking approach proposed in this work, although also using the best solution for each well and for each variable, goes far beyond than just directly using the patterns of attributes from these solutions. Our proposal in this work is to use patterns of attributes present in these elite solutions to guide the exploration of solutions that exist in the path of solutions connecting the best global solution to these elite solutions.

Regarding the aforementioned groups of history-matching methods, our learning-from-data approach with soft clustering and path-relinking has the following characteristics: it is a deterministic approach (it does not take any random decision while generating new solutions) that depends only on the data of an initial set of solutions and, hence, has no dependence on the derivatives of a global objective function; it uses a flow simulator to evaluate the performance of each model, so it does not belong to the “reduced physics” group; it is not explicitly designed to be geologically consistent (it does not manipulate geostatistical properties nor does it depend on geostatistical modeling software), but it indeed generates geologically consistent solutions as it will be described in the sections that detail our approach.

Focusing on the novelties introduced in this work, to the best of our knowledge, it is the first time a soft clustering approach is proposed to delimit reservoir regions changed during a history-matching process, and a path-relinking strategy is specially designed to generate solutions for the history-matching problem.

We organize the remainder of this paper into three sections. In Sect. 2, we introduce the learning-from-data

good-quality (or *elite*) solutions to guide the generation of new solutions following a path relinking strategy. The first stage generates new solutions manipulating petrophysical attributes, and the second stage generates new solutions changing global attributes. This division is inspired in the dynamic decision-making framework for history-matching proposed by Cavalcante et al. (2017), which uses a similar learning-from-data approach and also combines two processing stages, one for petrophysical attributes and the other for global attributes.

Figure 1 shows the high-level workflow of the proposed algorithm. It starts with an initial set of solutions (Sect. 2.1), initializes the *Data Manager* (Sect. 2.3) component to manage the access to solutions data, and sequentially executes *Stage 1* (Sect. 2.5) and *Stage 2* (Sect. 2.6) to generate new solutions with the path relinking strategy and changes in the petrophysical and global attributes, respectively. During *Stage 1*, the *Soft Region Delimiter* component (Sect. 2.5.1) is responsible for dynamically delimiting the regions where the changes in the values of petrophysical attributes occur. In both stages, the *Data Analyzer* component (Sect. 2.4) is responsible for the continuous evaluation of the data from available solutions. The algorithm keeps all solutions generated during its execution so that, at the end, one can filter and analyze the quality of the best solutions found. The next

We propose a two-staged algorithm, whereby each stage continuously (i) analyzes the data from previous available solutions to identify the variables which are still mismatched in the global best solution, and to select solutions where these variables are better matched; (ii) uses these

subsections detail the main aspects and components of our algorithm.

Initial solutions

The initial set of solutions for our algorithm can be of any size. However, since the algorithm implements a learning-from-data approach, it is fundamental that the initial solutions cover values of the uncertain attributes, and ideally are representative of good history-matching for each well and variable involved in the history-matching process. Our algorithm is designed to learn from the data, and having examples of good history-matches in the input data certainly improves the algorithm's effectiveness towards finding good-quality solutions. In this work, the initial set of solutions used in the computational experiments (Sect. 3) were all created with the *Discretized Latin Hypercube with Geostatistics* (DLHG) sampling technique from Schiozer et al. (2016).

Quality indicators

Our approach evaluates the quality of each solution from three perspectives: variable perspective, well perspective and global perspective. In all these perspectives, we follow the history-matching methodology introduced by Cavalcante et al. (2017), which defines history-matching quality levels to group solutions with equivalent matching quality. The quality levels are defined based on the values of the normalized quadratic deviation (NQD) indicator, calculated, for each variable involved in the history-matching process, through the following equations:

$$NQD_x = \frac{QD_x}{AQD_x} \quad (1)$$

$$QD_x = \sum_{t=1}^{N_{\text{times}}} (\text{Sim}_{xt} - \text{Hist}_{xt})^2 \quad (2)$$

$$AQD_x = \sum_{t=1}^{N_{\text{times}}} (\text{ToI}_x * \text{Hist}_{xt} + C_x)^2 \quad (3)$$

where

NQD_x is the normalized quadratic deviation of variable x and represents the variable misfit in the solution;

QD_x is the quadratic deviation between history data and simulated data of variable x ;

AQD_x is the acceptable quadratic deviation of variable x , used to normalize QD_x ;

N_{times} is the total number of times in the history data (for each measured temporal point in the history data, the corresponding simulated value is read from the simulation output);

ToI_x is the acceptable tolerance, in the range [0,1], and reflects the reliability on the historical data of variable x (the more reliable the historical data, the lower the acceptable tolerance can be);

Hist_{xt} is the historical data of variable x in time t ;

C_x is a constant used to prevent division by zero for variables which have history data equal to 0 for a long-time interval (e.g., producer water rates);

Sim_{xt} is the simulated data of variable x in time t .

For analysis purposes, sometimes it is useful to know whether the simulated data is predominantly above or below the historical data. In this case, a signed version of the NQD indicator, the normalized quadratic deviation with sign (NQDS), is defined according to the following equations:

$$NQDS_x = \frac{LD_x}{|LD_x|} NQD_x \quad (4)$$

$$LD_x = \sum_{t=1}^{N_{\text{times}}} (\text{Sim}_{xt} - \text{Hist}_{xt}) \quad (5)$$

where

$NQDS_x$ is the normalized quadratic deviation with sign of variable x

LD_x is the linear deviation between history data and simulated data of variable x ;

We adopt the six quality levels defined in Table 1. In this table, the “J” symbol follows the international standard ISO 80,000–2 and denotes an interval endpoint that is not included in the set. Note that the “Excellent” quality level represents matchings where the difference between the simulated data and the history data, for a particular variable x , is

Table 1 Quality levels for NQD_x . The “J” symbol follows the international standard ISO 80,000–2 and denotes an interval endpoint that is not included in the set

Quality level	Excellent	Very good	Good	Regular	Bad	Unsatisfactory
NQD_x	in [0,1]	in]1,2]	in]2,5]	in]5,10]	in]10, 20]	not in [– 20,20]

of the same magnitude as its acceptable deviation defined by Eq. (3).

The variable perspective to evaluate the quality of a given solution corresponds exactly to the NQD_x indicator. The well perspective, in turn, involves the calculation, for each well w , of the well misfit M_w according to Eq. (6):

$$M_w = \sqrt{\sum_{x \in V_w} (NQD_x)^2} \quad (6)$$

where V_w is the set of all variables of well w . A solution is considered “Excellent” for well w if the NQD_x values for all variables x in V_w are in the “Excellent” level, as defined in Table 1. Similarly, a solution is considered “Very Good” for well w , if the NQD_x values for all variables x in V_w are at least in the “Very Good” level. Analogous rules exist for the other quality levels of Table 1.

Finally, the global perspective to measure the quality of a given solution is the global misfit M calculated through the following equation:

$$M = \sqrt{\sum_{x=1}^{N_{var}} (NQD_x)^2} \quad (7)$$

where N_{var} is the total number of variables considered in the history-matching process.

The ultimate purpose of the global misfit M , the misfit M_w , and the NQD_x indicators, Eqs. (7), (6) and (1), respectively, is to allow the identification, by the *Data Analyzer* component (Sect. 2.4), of the best global solution and the best solutions for each well and variable involved in the history-matching process. These solutions, dynamically calculated during the execution of our algorithm, are the basis of our learning-from-data approach with soft clustering and path-relinking approach to the history-matching problem, as will be explored in the following sections.

Data manager

The *Data Manager* component encapsulates all operations related to access and persistence of the data, associated with the solutions manipulated by our algorithm. For each solution, it stores the corresponding values of the uncertain attributes and the quality indicators (Sect. 2.2) associated with each variable involved in the history-matching process.

Data analyzer

The *Data Analyzer* component is the learning core of our algorithm. It is responsible for the continuous analysis of available solutions, and for the calculation of quality indicators—Eqs. (1), (6) and (7)—that are used to measure the

quality of solutions according to the global, well or variable perspectives defined in Sect. 2.2.

This component is used in both stages (Sect. 2.5 and Sect. 2.6) of the proposed algorithm and, once it has calculated the quality indicators for all available solutions, it identifies the following types of good-quality solutions whose patterns of input attributes are used for generating new solutions:

Best global solution ($Best_G$): the solution with the lowest value of the global misfit indicator M , Eq. (7);

Best solution for each well w ($Best_w$): the solution with the lowest value of the well misfit M_w , Eq. (6), for each well;

Best solution for each variable x ($Best_x$): the solution with the lowest value of NQD_x , Eq. (1), for each variable.

Stage 1

The focus of Stage 1 in our learning-from-data approach is to generate new solutions through changes in the values of the petrophysical uncertain attributes. These attributes (e.g., permeability and porosity) are often highly correlated, reason why reservoir simulation models use to define their values in the so-called petrophysical realization (a set of data containing the values of all petrophysical attributes for each reservoir grid block). During a history-matching process, changing one petrophysical attribute without properly changing the others (correlated) attributes in the same block can result in geologically inconsistent models. We designed Stage 1 of our algorithm to be specialist in the manipulation of petrophysical realizations. This stage iteratively identifies variables that still have matching-quality out of the “Excellent” level (Table 1) in the best global solution (Sect. 2.4), and performs regional changes in the petrophysical attributes of this solution with the intention of finding better solutions.

Algorithm 1 depicts a high-level description for Stage 1. It starts with the execution of the *Data Analyzer* component (Sect. 2.4), and, while the stop criteria are not reached, it identifies the variables currently not matched in the best global solution, generates new solutions with petrophysical properties changed locally in the region associated with the mismatched variables, simulates new solutions and re-executes the *Data Analyzer*. Three steps of the execution of Stage 1, however, deserve special attention as they are fundamental pieces of our learning-from-data approach. The first important step is the execution of the *Data Analyzer* component (lines 1 and 13), which allows the identification of what needs to be changed (mismatched variables). The second step to be highlighted is the execution of the *Soft Region Delimiter* (line 4), which dynamically allows the identification of where, from a petrophysical perspective, the change needs to take place (in the reservoir region associated

with the well of the mismatched variable). The third fundamental step of Stage 1 (line 11) is the execution of the *Path Relinking on Petrophysical Attributes* strategy to generate new solutions, which is responsible for how the changes in the petrophysical attributes occur during the generation of new solutions. In Stage 1, the changes are made through progressive incorporation into the best global solution of the regional patterns of petrophysical attributes present in the best available solutions of the mismatched variable, or its well.

judgment about the expected interactions between producer and injector wells, as in the works of Bertolini et al. (2015) and Maschio and Schiozer (2018), or using streamlines, as adopted by Milliken et al. (2001) and Hoffman and Caers (2005).

In a recent work, Cavalcante et al. (2017) proposed a hard clustering (k -means based) strategy to define reservoir regions using spatial coordinates and values of petrophysical properties in each grid block. Each region obtained with this approach contains exactly one well and intends to group

Algorithm 1 Stage 1

Input: *stepSize*, needed by *Path Relinking on Petrophysical Attributes* (Section 3.5.2)

```

1: Execute Data Analyzer component (Section 3.4);
2: while stop criteria not reached
3:   Let  $Best_G$  be the best global solution (Section 3.4) found so far
4:   Execute Soft Region Delimiter (Section 3.5.1) using the petrophysical realization of  $Best_G$ 
5:   Let  $X$  be the set of variables  $x$  involved in the history-matching process, which are not in the
      the Excellent quality level (Table 1) in  $Best_G$ , and sorted in descending order of  $NQD_x$  (Section 3.2)
6:   for each variable  $x$  in  $X$ 
7:     Let  $w$  be the well associated with variable  $x$ ;
8:     Let  $SR_w$  represent the soft regions associated with well  $w$  (Section 3.5.1)
9:     Let  $Best_w$  be the best solution for well  $w$  (Section 3.4)
10:    Let  $Best_x$  be the best solution for variable  $x$  (Section 3.4)
11:    Let  $NS$  be the set of new solutions generated executing Path Relinking on Petrophysical Attributes
        (Section 3.5.2) with  $Best_G$ ,  $Best_w$ ,  $Best_x$ ,  $SR_w$  and stepSize
12:    Simulate the solutions from  $NS$  set
13:    Execute Data Analyzer component (Section 3.4);
14:    if new best global solution (Section 3.4) found
15:      break;
```

The next subsections detail the *Soft Region Delimiter* component and the *Path Relinking on Petrophysical Attributes* strategy.

Soft region delimiter

Partitioning the reservoir grid into regions is one major issue of the history-matching process, especially when the process involves changes in petrophysical attributes (Gervais-Couplet et al. 2007).

The geometrical Voronoi approach (Aurenhammer 1991) is widely used because it is simple and straightforward to implement (Caeiro et al. 2015; Oliveira et al. 2017): It assigns one region to each well (or group of wells) and assigns each reservoir grid block to the region containing its closest well.

Another common strategy to partition the reservoir grid attempts to capture well drainage areas: Each region is assigned to one or more wells, and is defined by the reservoir grid blocks in the drainage area associated with its wells. The core of this partitioning approach is the identification of drainage areas, which can simply be done by the engineer's

all grid blocks spatially close to the well, but also closely related in terms of their petrophysical properties.

The aforementioned strategies of partitioning the reservoir grid into regions shares a common aspect: They use a hard assignment policy with which each reservoir grid block is assigned to one (and only one) region, and the resulting regions are always disjoint. While this characteristic is effective in many cases, the binary assignment of grid blocks to regions may be an excessively strict way of delimiting the influence of a grid block in the performance of one or more wells. Given the heterogeneous nature of petroleum reservoirs, especially for grid blocks located in the frontiers of regions delimited with a hard assignment policy, it is not always obvious that the blocks impact only one of the neighboring regions.

In this work, we propose the use of a soft assignment policy to delimit reservoir regions. Our *Soft Region Delimiter* approach is based on soft k -means clustering algorithm (Appendix A) and, instead of assigning each grid block to a single region, it calculates the degree of membership of each grid block to each region, allowing the same grid block to belong to more than one region, simultaneously.

Algorithm 2 Soft Region Delimiter**Input:** P , petrophysical realization1: Build the R^n space S as the set of data points $v = (v_1, v_2, \dots, v_n)$ such that:2: (i) each data point v is associated with a reservoir grid block from P ;3: (ii) v_1, v_2, v_3 are the x, y, z coordinates of the grid block4: (iii) $v_4, \dots, v_n$ are the normalized values of the petrophysical properties of the grid block5: Let $W = \{w_1, \dots, w_k\}$ be the set of data points v associated with the center block of each well k 6: Initialize the centroids set $C = \{c_1, \dots, c_k\}$ with W , and let R_i denotes the region with centroid c_i 7: **repeat**8: **for** each data point v in S **do**Calculate the membership degree of v to each region R_i

$$z_{vi} = \frac{e^{-\|v - c_i\|^2}}{\sum_{l=1}^k e^{-\|v - c_l\|^2}}$$

9: Update regions centroids

$$c_i = \frac{\sum_v z_{vi} v}{\sum_v z_{vi}}$$

10: **until** centroids stabilize.11: Let $\mathcal{R}_E$ be the set of *empty* regions: $\mathcal{R}_E = \{R_i \mid z_{wi} < \varepsilon, \forall w \in W\}$ 12: Let $\mathcal{R}_{NE}$ be the set of *non-empty* regions: $\mathcal{R}_{NE} = \{R_j \mid z_{wj} > \varepsilon, \text{ for at least one } w \text{ in } W\}$ 13: **for** each region R_i in $\mathcal{R}_E$ **do**14: **for** all v in S , such that $z_{vi} > \varepsilon$, **do**15: Let R_j be the region in $\mathcal{R}_{NE}$ whose centroid c_j is the closest to v 16: Update z_{vj} and z_{vi} : $z_{vj} += z_{vi}$; $z_{vi} = 0$;17: Remove c_i from the set C of centroids.

Algorithm 2 summarizes our *Soft Region Delimiter* algorithm. It works on the same n -dimensional space and starts with the same initial centroids proposed by Cavalcante et al. (2017) (lines 1 to 6). Lines 7 to 10 correspond to the soft k -means clustering algorithm (Appendix A). In line 8, we chose to set to 1 the *stiffness* parameter (which appears in line 3 of Algorithm A.1), following the pragmatic recommendation of Bauckhage (2015). The stop criteria in line 10 is the same from Algorithm A.1, discussed in Appendix A. The final steps of Algorithm 2 (lines 11 to 17) are a post-processing to ensure that each of the final regions has at least one of the wells with nonzero (within an epsilon tolerance) membership degree to it. It is also noteworthy that, although the algorithm starts with one centroid (and one region) for

each well, following the post-processing steps, the number of final regions may be smaller than the number of wells, and some of these regions may have multiple wells with nonzero membership degree to them.

Following an execution of the *Soft Region Delimiter* (Algorithm 2), soft regions associated with a particular well are the regions to which the data point associated with the center block of the well has nonzero membership degree. Each of these regions contains the grid blocks associated to the data points with non-zero membership degree to it, and our final claim is that, for each well, the grid blocks in the soft regions associated with it are the ones that, from space and petrophysical perspectives, are more likely to impact its production.

Path relinking on petrophysical attributes

The *Path Relinking on Petrophysical Attributes* is the component that effectively generates new solutions during the execution of Stage 1 (Sect. 2.5). This component is called in line 11 of Algorithm 1. It is based on the path relinking approach (Appendix A), and performs regional changes in the values of petrophysical attributes with the purpose of improving the history-matching quality of mismatched variables. As explained in Appendix A, path relinking approach explores trajectories connecting elite solutions with the reasoning that these paths can potentially contain better solutions. In this work, we propose the use of path relinking to explore the paths between the best global solution and the best solutions corresponding to each mismatched well and variable involved in the history-matching process. Our motivation here is: In the best global solution, some wells and variables are still mismatched, but among the available solutions, eventually there are solutions (such as well best solutions and variable best solutions, Sect. 2.4) where these wells and variables are better matched. So, exploring the paths that connect the best global solution and the well, and variable best solutions, can result in solutions with improved history-matching quality.

Algorithm 3 shows a high-level description for the *Path Relinking on Petrophysical Attributes* component. The algorithm receives, as input parameters, three elite solutions—global best solution ($Best_G$), best solution for a well w ($Best_w$) and best solution for a variable x ($Best_x$) (Sect. 2.4)—which implicitly compose the reference set for our path relinking approach. The reservoir's soft regions where the petrophysical changes take place are also defined as input data (parameter SR_w). The final input parameter is *stepSize*, a real number in $]0,1[$, which controls the proportion of regional data from guiding solution that is gradually incorporated into the initial solution, during the generation of the intermediate solutions in a path. The lower the *stepSize* value, the more slowly the algorithm moves from the initial solution towards a guiding solution, and the higher

will be the density of intermediate solutions in the paths explored by our algorithm. Note at this point, that we do not use multipliers to tune spatial distributed parameters. Instead, the *Path Relinking on Petrophysical Attributes* component tunes the petrophysical attributes through linear combinations of the petrophysical values presented in the best global solution and in the best solutions for each well and (mismatched) variable involved in the history-matching process.

The proposed *Path Relinking on Petrophysical Attributes* component explores three different paths, all of them starting from the global best solution ($Best_G$ input parameter). The first and second paths (generated through lines 3–10 in Algorithm 3) connect, respectively, the best global solution to the best solution for the well and to best solution to the variable ($Best_w$ and $Best_x$ input parameters). The third path (constructed through lines 11–20) explores solutions that progressively incorporate, into the global best solution, values of attributes present in the combination of the regional petrophysical patterns of the well and variable best solutions. Our reasoning to explore these three different types of paths is to try to incorporate into the best global solution, where some variables are still mismatched, regional petrophysical patterns present in solutions where the variables themselves, or their corresponding wells, are better matched. Note that, although the preservation of geological consistency is not a primary design concern of this component, assuming that elite solutions are themselves geologically consistent, the path-relinking approach will only combine geologically consistent patterns, what will certainly keep the geological consistency in the generated solutions. Algorithm 3 returns the set of newly generated solutions.

One final and important remark is that the number of new solutions generated in each execution of Algorithm 3 is bounded to “ $3 * stepSize$ ”, and all these solutions are independent and can be simulated in parallel, promoting performance gains in the total execution time of our learning-from-data approach.

Algorithm 3 Path Relinking on Petrophysical Attributes

Input: $Best_G$, best global solution (Section 3.4);
 $Best_w$, best solution for well w (Section 3.4);
 $Best_x$, best solution for variable x (Section 3.4);
 SR_w , soft regions for well x (Section 3.5.1)
 $stepSize$, real number in $]0,1[$

- 1: Let NS be the set that will contain the new solutions generated by this algorithm: $NS \leftarrow \emptyset$
- 2: Set $s = Best_G$ and let $\mathcal{P}_s$ be its petrophysical pattern in regions SR_w
- 3: **for** g in $\{Best_w, Best_x\}$ **do**
- 4: Let $\mathcal{P}_g$ be petrophysical pattern of g in regions SR_w
- 5: Set $\alpha = 1$; $\beta = 0$;
- 6: **do**
- 7: Generate $\bar{s}$ from s , setting the petrophysical pattern in soft regions SR_w to
 $(\alpha - stepSize) \mathcal{P}_s + (\beta + stepSize) \mathcal{P}_g$
- 8: Add $\bar{s}$ to NS : $NS \leftarrow NS \cup \{\bar{s}\}$
- 9: Set $\alpha -= stepSize$; $\beta += stepSize$;
- 10: **while** $\alpha > 0$
- 11: Set $g_1 = Best_w$, and let $\mathcal{P}_{g_1}$ be its petrophysical pattern in regions SR_w
- 12: Set $g_2 = Best_x$, and let $\mathcal{P}_{g_2}$ be its petrophysical pattern in regions SR_w
- 13: Set $\alpha = 1$; $\beta = 0$;
- 14: **do**
- 15: Generate $\bar{s}$ from s , setting the petrophysical pattern in soft regions SR_w to
 $(\alpha - stepSize) \mathcal{P}_{g_1} + (\beta + stepSize) \mathcal{P}_{g_2}$
- 16: Add $\bar{s}$ to NS : $NS \leftarrow NS \cup \{\bar{s}\}$
- 17: Set $\alpha -= stepSize$; $\beta += stepSize$;
- 18: **while** $\alpha > stepSize$
- 19: Generate $\bar{s}$ from s , setting the petrophysical pattern in soft regions SR_w to $(\mathcal{P}_s + \mathcal{P}_{g_1} + \mathcal{P}_{g_2})/3$
- 20: Add $\bar{s}$ to NS : $NS \leftarrow NS \cup \{\bar{s}\}$
- 21: **return** NS

Figure 2 illustrates the generation of new solutions in one execution of the *Path Relinking on Petrophysical Attributes* component. In this example, the target mismatched variable is the oil rate of well w_1 ($Q_o[w_1]$), and we set the value of the $stepSize$ input parameter to 0.25. In each solution, the small inner square represents the reservoir region (SR_w input parameter of Algorithm 3) associated with well w_1 where the changes occur. The colors indicate petrophysical patterns and, for the sake of simplicity, we use two colors in each solution (one for the affected region, and another for non-affected regions). The figure shows the three different paths generated by the *Path Relinking on Petrophysical Attributes* component, all of them starting from the best global solution ($Best_G$). In each path, the petrophysical pattern of new solutions in the affected region progressively moves towards patterns of the guiding solutions. Considering that the variable $Q_o[w_1]$ was originally mismatched in the initial $Best_G$ solution and that it is better matched in the guiding solutions ($Best_{w_1}$ and $Best_{Q_o[w_1]}$), our algorithm indeed seeks to explore paths that seem promising to improve the history-matching quality of mismatched variables.

Stage 2

Stage 2 of our learning-from-data approach (Fig. 1) is the component responsible for generating new solutions through changes in the values of the global uncertain attributes. Similar to Stage 1 (Sect. 2.5), Stage 2 is also based on the path relinking approach (Appendix A), and also iteratively identifies the variables that are still mismatched in the best global solution, and generates new solutions with the purpose of improving the matching quality of these variables. The difference here is that, while Stage 1 manipulates petrophysical attributes, Stage 2 seeks to find solutions with improved matching quality by manipulating global attributes. Since Stage 2 is executed immediately after Stage 1, the first best global solution considered in Stage 2 is the last best global solution found by Stage 1.

Algorithm 4 shows a description for this component. From a high-level perspective, it is very similar to Algorithm 1, from Stage 1, presented in Sect. 2.5: starting with the execution of the *Data Analyzer* component (line 1), and while the stop criteria are not reached (line 2), Stage 2 continuously

Fig. 2 Illustration of generation of new solutions by the *Path Relinking on Petrophysical Attributes* component. The three explored paths connect, from the perspective of petrophysical attributes, the best

global solution ($Best_G$) to the best solutions for the well and variable ($Best_{w_1}$ and $Best_{Q_o[w_1]}$)

identifies variables not currently matched in the best global solution (line 4) and, for each of these variables, generates new solutions seeking to improve their matching quality (line 9), simulates the new solutions (line 10), and re-executes the *Data Analyzer* component (line 11). The difference here is that, instead of generating new solutions with regional changes in the values of petrophysical attributes, as was the case of Stage 1, Stage 2 executes the *Path Relinking on Global Attributes* (line 9), which generates new solutions with changes in the values of global attributes. Note that,

similar to Stage 1, the execution of Stage 2 handles the learning aspects of “what needs to be changed” and “how it can be changed”, as it continuously identifies the mismatched variables and executes the *Path Relinking on Global Attributes* to try to improve their matching quality. The concern about “where the change shall take place”, present during the execution of Stage 1, does not make sense during Stage 2, since Stage 2 is focused on the manipulation of global attributes, and changes conducted by this stage have a global impact on the performance of the reservoir model.

Algorithm 4 Stage 2

```

1: Execute Data Analyzer component (Section 3.4);
2: while stop criteria not reached
3:   Let  $Best_G$  be the best global solution (Section 3.4) found so far
4:   Let  $X$  be the set of variables  $x$  involved in the history-matching process, which are not in the
     the Excellent quality level (Table 1) in  $Best_G$  and sorted in descending order of  $NQD_x$  (Section 3.2)
5:   for each variable  $X$  in  $x$ 
6:     Let  $w$  be the well associated with variable  $x$ ;
7:     Let  $Best_w$  be the best solution for well  $w$  (Section 3.4)
8:     Let  $Best_x$  be the best solution for variable  $x$  (Section 3.4)
9:     Let  $NS$  be the set of new solutions generated executing Path Relinking on Global Attributes
       (Section 3.6.1) with  $Best_G$ ,  $Best_w$  and  $Best_x$ 
10:    Simulate the newly generated solutions
11:    Execute Data Analyzer component (Section 3.4);
12:    if new best global solution (Section 3.4) found
13:      break;

```

The next subsection details our *Path Relinking on Global Attributes* strategy.

Path relinking on global attributes

The *Path Relinking on Global Attributes* is the component of our learning-from-data approach that is specialist in the manipulation of global attributes. It is the actual responsible for the generation of new solutions during the execution of Stage 2 (Sect. 2.6). As was the case with the *Path Relinking on Petrophysical Attributes* (Sect. 2.5.2), our motivation here is to explore the paths connecting the best global solution

(where some wells and variables are stills mismatched) to the best solutions for a particular mismatched well and variable involved in the history matching process. The path relinking approach (Appendix A) suggests that paths connecting elite solutions can contain solutions with better quality, and our hope with the *Path Relinking on Global Attributes* component is to find solutions with improved history-matching quality in the paths—constructed with changes in global attributes—that connect the global best solution and the elite solutions for a particular mismatched well or variable.

Algorithm 5 Path Relinking on Global Attributes

```

Input:  $Best_G$ , best global solution (Section 3.4);
        $Best_w$ , best solution for well  $w$  (Section 3.4);
        $Best_x$ , best solution for variable  $x$  (Section 3.4);
1: Let  $NS$  be the set that will contain the new solutions generated by this algorithm:  $NS \leftarrow \emptyset$ 
2: for  $g$  in  $\{Best_w, Best_x\}$ 
3:   Set  $s = Best_G$ , and let  $G_s = \{a_1, a_2, \dots, a_n\}$  be the set of global attribute values in  $s$ 
4:   Let  $G_g = \{a_1', a_2', \dots, a_n'\}$  be the set of global attribute values in  $g$ 
5:   Set  $i = 1$ ;
6:   while  $i \leq n$ 
7:     if  $a_i \neq a_i'$ 
8:       Generate  $\bar{s}$  from  $s$ , replacing  $a_i$  by  $a_i'$ 
9:       Add  $\bar{s}$  to  $NS$ :  $NS \leftarrow NS \cup \{\bar{s}\}$ 
10:      Set  $s = \bar{s}$ ;
11: return  $NS$ 

```

Fig. 3 Illustration of generation of new solutions by the *Path Relinking on Global Attributes* component. The two explored paths connect, from the perspective of global attributes, the best global solution ($Best_G$) to the well best and variable best solutions ($Best_{w_1}$ and $Best_{Q_o[w_1]}$)

Algorithm 5 depicts a high-level description for the *Path Relinking on Global Attributes*. Similar to Algorithm 3 for the *Path Relinking on Petrophysical Attributes* component (Sect. 2.5.2), this algorithm also receives, as input parameters, three elite solutions—global best solution ($Best_G$), well best solution ($Best_w$) and variable best solution ($Best_x$)—which implicitly compose the reference set for our path relinking approach. The difference here is that, to limit the number of new solutions generated, we designed the *Path Relinking on Global Attributes* component to explore only two paths, both starting from $Best_G$ solution, and each of which having $Best_w$ or $Best_x$ as guiding solutions. In these two paths, since the focus is the change of global attributes, the solutions are generated progressively incorporating the value of a global attribute, originally not present in $Best_G$ solution, and present in the corresponding guiding solution (lines 7–10).

The total number of solutions generated in each execution of Algorithm 5 is bounded to “ $2 * nG$ ”, where nG is the number of global attributes considered in the history-matching process. All the new generated solutions are independent and can be simulated in parallel, to profit from the benefits of parallel computing and reduce the total execution time of the proposed approach.

In Fig. 3, we illustrate the generation of new solutions in one execution of the *Path Relinking on Global Attributes* component. In this example, there are three global attributes (A, B, C), each of which have three uncertain values, and the target mismatched variable is the oil rate of well w_1 ($Q_o[w_1]$).

The term *Petro* is shown for each solution just for the completeness of the example, representing the solution’s petrophysical attributes. The figure shows the two paths generated by the *Path Relinking on Global Attributes* component, both starting from the best global solution ($Best_G$). In each path, the *Petro* attributes remain unchanged and the values of the global attributes progressively move towards the patterns of the guiding solutions. Considering that the variable $Q_o[w_1]$ is potentially better matched in the guiding solutions ($Best_{w_1}$ and $Best_{Q_o[w_1]}$) than it is in the initial $Best_G$ solution, the proposed strategy hopes to find, among the new solutions, some solutions with improved history-matching quality.

Experiments and results

We organize the experiments to validate our algorithm in three rounds.

In Round #1 (Sect. 3.3), we evaluate the performance of the proposed approach on a complex benchmark case and show its potential towards finding solutions with history-matching quality much better than the initial solutions, geologically consistent appearance, and also acceptable performance when used to forecast field production.

In Round #2 (Sect. 3.4), we confront results of the proposed soft clustering strategy with the ones obtained with a hard assignment approach recently published in the literature. The comparison indicates that our proposal of using a soft clustering strategy to delimit reservoir regions during

a history-matching process is promising and may result in solutions with better history-matching quality than the ones obtained with a hard assignment policy.

Finally, in Round #3 (Sect. 3.5), we compare results of our learning-from-data approach with the ones from related work for the same benchmark. In terms of matching quality, our approach found solutions with better quality than the best final solutions reported by three of the five compared approaches. Moreover, it required at least 80% less simulations than four of the five approaches considered in the comparison.

Our algorithm was implemented in JAVA (1.7.0_40), and uses libraries from MERO (version 6.0) developed by UNISIM group (<http://www.unisim.cepetro.unicamp.br>). MERO provides a set of tools that automate common tasks present in research lines around reservoir simulation and management. Our algorithm uses MERO tools to launch

simulations and to calculate QD_x indicators according to Eq. (2). Although MERO supports different simulators, in our experiments, we configured it to use the IMEX simulator (2012.10) from Computer Modelling Group (<http://www.cmgl.ca/>).

In the following subsections, we describe the UNISIM-I-H benchmark (Sect. 3.1) and the evaluation metrics (Sect. 3.2) used in the experiments, and detail the three rounds of experiments (Sects. 3.3, 3.4 and 3.5) executed with our learning-from-data approach for the history-matching problem.

UNISIM-I-H benchmark

The UNISIM-I-H benchmark is a complex synthetic case for history matching, proposed by Avansi and Schiozer (2015b). The model is a black-oil simulation model with a medium-resolution grid, and was derived from the UNISIM-I-R model (Avansi and Schiozer 2015a), which is a very-high resolution geological model (3.5 million active blocks), built with data from the Namorado Field (Campos Basin, Brazil). Despite being a synthetic case, UNISIM-I-H poses challenges similar to a real reservoir, and provides the model response for the forecast period. This enables a complete assessment of both the history-matching quality and the forecast capacity of solutions found by different approaches and justifies our choice of using this benchmark during the experiments with our learning-from-data approach for history matching (Table 2).

Table 2 Uncertain attributes of UNISIM-I-H model, and considered in our learning-from-data approach to the history-matching problem

Attribute	Description	Levels
<i>Petro</i>	Petrophysical properties (porosity, horizontal and vertical permeabilities, and net to gross ratio)	500
K_{rw}	Water relative permeability	5
PVT_{R_2}	PVT data in the <i>east block</i>	3
WOC_{R_2}	Oil–water contact in the <i>east block</i>	5
CPOR	Rock compressibility	5
Kz_c	Vertical permeability multiplier	5

Fig. 4 UNISIM-I-H model: grid top and well locations

Table 3 Probability density functions of continuous uncertain attributes of UNISIM-I-H model

Attribute	Unity	PDF
WOC_{R_2}	m	$0, x < 3169 \frac{x-3169}{25},$ $3169 \leq x < 3174 \frac{3179-x}{25},$ $3174 \leq x \leq 3179, 0, x > 3179$
$CPOR \times 10^{-6}$	cm^2/kgf	$0, x < 10$ $\frac{x-10}{9849}, 10 \leq x < 53$ $\frac{53-10}{1849}, 53 \leq x \leq 96$ $0, x > 96$
Kz_c (-)	–	$0, x < 0 \frac{2x}{4.5},$ $0 \leq x < 1.5 \frac{6-2x}{4.5},$ $1.5 \leq x \leq 3$ $0, x > 3$

The UNISIM-I-H grid has $81 \times 58 \times 20$ blocks (of which 36,739 are active), and each block measures in average $100 \times 100 \times 8$ m. The reservoir is drained by 14 producer wells supported by 11 injector wells, located as Fig. 4 indicates. A sealing fault (red line in Fig. 4) divides the reservoir in two regions: Region 1 (biggest reservoir area, located on the west of the sealing fault), and Region 2 (smallest reservoir portion located on the east of the sealing fault, also referenced as *east block*). A detailed description of UNISIM-I-H case and the files of its original characterization are available in <https://www.unisim.cepetro.unicamp.br/benchmarks/en/unisim-i/unisim-i-h>.

In this work, we considered the uncertain attributes summarized in Table 1. For each attribute, the column “Levels” indicates the discretization adopted in our experiments, which is the same discretization used in the work of Cavalcante et al. (2017). The discrete levels of attributes $Petro$, K_{rw} , and PVT_{R_2} are equiprobable and are exactly the ones which correspond to the files provided in the UNISIM-I-H model characterization. For the other three uncertain attributes (WOC_{R_2} , $CPOR$, and Kz_c), which are continuous properties, the levels were created through sampling from their triangular probability density functions (PDF), defined in the UNISIM-I-H model’s characterization and reproduced in Table 3. Stage 1 (Sect. 2.5) of our learning-from-data approach generates new solutions changing the values of the $Petro$ uncertain attributes. These attributes are high-correlated and changing one of them without changing the others could result in geologically inconsistent models. Stage 2 (Sect. 2.6), in turn, focuses on changes in the values of the five remaining uncertain attributes (K_{rw} , PVT_{R_2} , WOC_{R_2} , $CPOR$, and Kz_c).

The history data and forecast response for the UNISIM-I-H benchmark case come from the simulation of the UNISIM-I-R model, which covers a period of 10957 days (30 years), divided into two periods: history period (the first 4018 days or 11 years) and

Table 4 Tolerance and constant values used in NQD_x calculation, Eqs. (1) to (3)

Variable (x)	Tol _x	C _x
Q_o [producer well]	0.01	0
Q_w [producer well]	0.01	20 m ³ /d
Q_g [producer well]	0.01	0
Q_{wi} [injector well]	0.05	0
BHP [producer/injector well]	0.01	0

forecast period (the remaining 6939 days or 19 years). Within the history period, the following data is available, on a monthly basis: oil, water, and gas production rates (Q_o , Q_w , and Q_g) for producer wells; water injection rate (Q_{wi}) for injector wells; and bottom-hole pressure (BHP) for all wells.

In the UNISIM-I-H model, the constrained phases in the simulation are liquid rate for producer wells and water injected rate for injector wells. The output variables to be matched during the history-matching process are: oil, water, and gas production rates (Q_o , Q_w , and Q_g) for producer wells; water injection rate (Q_{wi}) for injector wells; and bottom-hole pressure (BHP) for all wells. Since UNISIM-I-H model has 14 producers and 11 injectors, this results in 78 variables to be matched during the history-matching process of this model and, in 78 variables considered by our learning-from-data approach while evaluating the global misfit, Equation (7), of each solution generated during the experiments with this benchmark.

UNISIM-I-H-A dataset

The UNISIM-I-H-A dataset was publicly released by Cavalcante et al. (2017) to facilitate the comparison of results obtained with different approaches to the history matching of UNISIM-I-H benchmark. The dataset is available on <https://figshare.com/articles/PONE-D-17-12470/5024837> and comprises five scenarios (S1, S2, S3, S4, and S5) based on the UNISIM-I-H benchmark case. For details about these scenarios, the reader is referred to Cavalcante et al. (2017). In summary, each of these scenarios has 100 models, generated with the *Discretized Latin Hypercube with Geostatistics* (DLHG) sampling technique from Schiozer et al. (2016) to combine the levels (Table 2) of the uncertain attributes of the UNISIM-I-H benchmark. In our experiments, we use the models in the scenarios of the UNISIM-I-H-A dataset as initial solutions for the executions with our learning-from-data approach.

Evaluation metrics

In our experiments, we used the following metrics to evaluate the performance of the proposed learning-from-data approach, and to compare it with related work on the same UNISIM-I-H benchmark:

- Global misfit value, Eq. (7), of the best global solution;
- Number of models by max NQD_x , Eq. (1);
- Number of models by number of variables with quality range at least good (Table 1);
- Number of simulations.

For the first metric, we compare the values at the beginning and end of the algorithm's execution. For the second and third metrics, we consider the 100 initial models and the final 100 best models obtained by the algorithm. The number of simulations corresponds to the total number of simulations, and it is ultimately equal to the number of solutions in the initial set of solutions, plus the number of new solutions generated by Stage 1 and Stage 2 of our algorithm (lines 11 and 9, from Algorithm 1 and Algorithm 4, respectively).

In all experiments, we set the values of the acceptable tolerance (Tol_x) and constant (C_x) used in the calculation of the NQD_x indicator, Eqs. (1) to (3), according to Table 4.

Finally, the *stepSize* input parameter used in the *Path Relinking on Petrophysical Attributes* (Sect. 2.5.2) was set to 0.25 in all the executions of our learning-from-data approach. We believe this value provides a reasonable trade-off between the number of new solutions in each path generated by our algorithm and the corresponding number of simulations.

Fig. 5 Misfit of global best solution, initially and, substantially smaller, after the execution of our learning-from-data approach, for each scenario of UNISIM-I-H-A dataset

Table 5 Improvement on the misfit of global best solution with the execution of our learning-from-data approach, for each scenario of UNISIM-I-H-A dataset

Scenario	Improvement (%)
S1	74
S2	70
S3	70
S4	66
S5	77

Experiments-round #1

The purpose of this first round of experiments is to assess the performance of our learning-from-data approach using the five scenarios from the UNISIM-I-H-A dataset (Sect. 3.1.1).

Figure 5 shows, for each scenario from the UNISIM-I-H-A dataset, the misfit, Eq. (7), of the global best solution in the initial set of solutions ('Initial' blue bar) and, in the final set of solutions ('Final' green bar) obtained with the execution of our learning-from-data approach. As Table 5 summarizes, for all scenarios, the improvements on the misfit of the global best solution ranged from 66 to 77%. This result shows how the proposed approach is indeed capable of finding solutions with improved history-matching quality.

Table 6 shows the total number of simulations required in the execution of our algorithm for each scenario of the UNISIM-I-H-A dataset. In all scenarios, the total number of simulations was never higher than 400, which is a low number of simulations for a complex history-matching case, as is the case of the UNISIM-I-H benchmark. The reduced number of simulations is indeed one of the advantages of our learning-from-data approach, and this aspect will be clearer during the discussions of Round #3 experiments in Sect. 3.5, where we compare our approach with previous works for the same UNISIM-I-H benchmark.

Figures 6 and 7 depict two types of plots that enable an evaluation of the overall improvements achieved with the executions of our learning-from-data approach in each scenario of the UNISIM-I-H-A dataset. All plots in these figures compare the initial models with the final 100 best models obtained with the execution of our approach in each scenario. The plots on the left of Figs. 6 and 7 show the number

Table 6 Total number of simulations during the execution of our learning-from-data approach, for each scenario of the UNISIM-I-H-A dataset

Scenario	Total Number of Simulations
S1	344
S2	393
S3	344
S4	264
S5	311

Fig. 6 “Number of models by max NQD” and “Number of models by number of variables with $NQD \leq 5$ ”, considering the initial solutions and the 100 final best solutions obtained with the execution of our learning-from-data approach with scenarios S1 (A), S2 (B), and S3 (C) from the UNISIM-I-H-A dataset. The final (orange) curves

indicate that the 100 final best solutions obtained by our learning-from-data approach are considerably better (with smaller max NQD and higher number of variables with history-matching quality at least “good”, $NQD \leq 5$) than the corresponding initial solutions (blue curves)

Fig. 7 “Number of models by max NQD” and “Number of models by number of variables with $NQD \leq 5$ ”, considering the initial solutions and the 100 final best solutions obtained with the execution of our learning-from-data approach with scenarios S4 (A) and S5 (B) from the UNISIM-I-H-A dataset. The final (orange) curves indicate

that the 100 final best solutions obtained by our learning-from-data approach are considerably better (with smaller max NQD and higher number of variables with history-matching quality at least “good”, $NQD \leq 5$) than the corresponding initial solutions (blue curves)

of models versus the maximum value of NQD indicator (for the 78 variables simultaneously) and, the more vertical and the more the curve tends to the left, the better the result. The plots on the right of Figs. 6 and 7 show the number of models by number of variables with history-matching quality of at least good ($NQD \leq 5$, according to Table 1). In this case, the more the curve tends to the right, the better the result. From both perspectives, the plots in Figs. 6 and 7 show that the 100 final best solutions have superior history-matching quality than the corresponding initial solutions in each scenario. While in the initial sets the max NQD values were in the range [60, 600], in the final sets of 100 best solutions, the max NQD have values in [12, 50]. This represents improvements from 80% to 91.6% in the values of the max

NQD indicator. Also, while initially there was no solution with max $NQD < 50$, in the final 100 best solutions, in all scenarios, at least 28% of solutions have max $NQD \leq 20$, and in three of the five scenarios (S1, S2 and S3), there are more the 90% of the 100 best final solutions with $NQD \leq 25$.

Regarding the number of variables with history-matching quality of at least “good”, the plots of Figs. 6 and 7 also show that solutions generated by our learning-from-data approach are considerably better than the initial solutions in all scenarios of the UNISIM-I-H-A dataset. It is important to note that the only scenario for which the initial set has solutions with more than 60 variables with history-matching quality at least good ($NQD \leq 5$) is scenario S4. However, considering the final 100 best solutions, the plots in Figs. 6

Fig. 8 NQDS plots for producer variables (Q_o , Q_w , Q_g and BHP) and injector variables (Q_{wi} , BHP), in the initial solutions (red columns) and in the 100 final best solutions (blue columns) obtained with the execution of our learning-from-data approach on scenario S1 from

the UNISIM-I-H-A dataset. The final best solutions obtained with our approach are much closer to (or even contained in) the ‘Good’ or ‘Excellent’ quality levels, indicated, respectively, by the horizontal dashed and black lines. The x-axis in the plots indicate well names

and 7 show that, in all scenarios, our learning-from-data approach ended up with solutions with more than 70 variables (89% of the variables involved in the process) with “good” history-matching quality. Moreover, in four out of five scenarios, at least 80% of the 100 final best solutions have more than 65 variables (83% of the total variables) with history-matching quality at least “good”.

In the remainder of this section, we continue the analysis of the performance of our learning-from-data approach, presenting results obtained with scenario S1 from the UNISIM-I-H-A dataset. The results with the other four scenarios from this dataset are similar and, due to space constraints, are not presented here.

Figure 8 shows, in terms of the NQDS indicator, Eq. (4), the history-matching quality of all producer and injector variables considered in the history-matching process of scenario S1 from the UNISIM-I-H-A dataset. In the plots of this figure, the red and blue columns indicate, respectively, the NQDS values for the initial and the final 100 best solutions obtained with the execution of our learning-from-data approach. The horizontal black lines and the horizontal dashed lines in each plot indicate, respectively, the ranges corresponding to the ‘Excellent’ and ‘Good’ quality levels (Table 1). For all the variables, the plots show a significant improvement on the history-matching quality, with the blue columns much closer to (or even contained in) the ‘Good’ or ‘Excellent’ quality thresholds. There are still some (very few) variables that exhibit a not-so-good matching quality. However, even these not perfectly matched variables are better history-matched than they were in the initial solutions. Moreover, a perfect history matching, where all variables are perfectly matched, is hardly possible to achieve, especially for a complex reservoir such as the UNISIM-I-H model. Our method looks at the problem as a whole and seeks to find solutions that have the maximum number of variables with good matching quality as possible. Our primary purpose in this work was to validate the suitability of a learning-from-data approach with soft clustering and path relinking to the history-matching problem, understanding whether it is capable of improving the history-matching quality of an initial set of solutions. The results from Fig. 8 clearly indicate the potential of our approach toward finding models with significantly better history-matching quality.

Figure 9 shows the oil rate curves for all producers in scenario S1 from the UNISIM-I-H-A dataset. Comparing the initial, grey, curves (corresponding to the initial solutions) with the final, green, ones (corresponding to the 100 best solutions obtained with our approach in this scenario), we can observe that the final curves are much less dispersed, and much closer to the ‘History’ curve (red dots) or to the ‘Excellent’ quality level, whose boundaries are indicated by the blue lines.

Figures 10, 11 and 12 illustrate the petrophysical patterns of the solutions generated by our learning-from-data approach. Each of these figures shows the pattern of one petrophysical property in the initial global best solution, and the corresponding pattern in the final global best solution generated by our approach using scenario S1 from the UNISIM-I-H-A dataset. The initial petrophysical patterns are geologically consistent since all initial solutions in UNISIM-I-H-A scenarios use petrophysical images available in the UNISIM-I-H characterization (Sect. 3.1) which, in turn, were generated with a commercial software for geostatistical modeling. At this point, we highlight how the final petrophysical patterns obtained with our learning-from-data approach are visually realistic and consistent from a geological point of view: They are very similar to the initial (and geologically consistent) patterns, and do not have regions of abrupt discontinuity. Moreover, comparing Figs. 10 and 11, we observe that the final petrophysical patterns also maintain the trend between permeability and porosity properties. The preservation of geological consistency, indeed, is an important characteristic of our approach. Although this aspect has not been taken as a primary design concern, our learning-from-data approach implicitly manages geological consistency, generating solutions using combinations of petrophysical patterns present in (or derived from) geologically consistent solutions present in the initial set of solutions.

Our last analysis in this first round of experiments aims to verify how the 100 final best solutions generated by our learning-from-data approach behave when the simulation period is extended to include the forecast period. With this purpose, the simulation time of the selected models was extended to cover a total of 10,957 days (30 years), thus divided: The first 4018 days (11 years) correspond to the history period, and the remaining 6939 days (19 years) are the forecast period. Figure 13 shows the oil and water field production curves in the historical and forecast periods, for the initial solutions and for the final 100 best solutions obtained with the execution of our algorithm with scenario S1 from the UNISIM-I-H-A dataset. It is noteworthy that, during the history period, both oil and water curves are relatively well matched. During the forecast period, although the final (blue) solutions obtained with our algorithm do not completely encompass the forecast reference (red) curve, they indeed capture and follow the essence of the reference curve much better than the initial grey curves. Finally, it is important to highlight that our algorithm is designed to learn from the data, so it can only capture patterns that are actually present in the initial data. The initial (grey) curves in Fig. 13, indicate models that, during the history period, have oil production and water production that are higher and lower than the reference curve, respectively. This explains the fact that the final (blue) solutions obtained with our algorithm, although very close to the reference curve, still present oil

Fig. 9 Oil rate production curves in the initial solutions (grey curves), and in the 100 final best solutions (green curves) obtained with the execution of our learning-from-data approach on scenario S1 from the UNISIM-I-H-A dataset. For all producers, the final green curves are much closer to the ‘History’ red curve, or to the bounds of the ‘Excellent’ quality level indicated by the blue lines

production that is slightly above, and water production that is, on average, below the reference curve.

Concluding this first round of experiments, we can note that the results obtained with our learning-from-data approach indicate its potential towards finding solutions with history-matching quality much better than the initial solutions, and also with acceptable performance when used to forecast field production.

Experiments–round #2

Dividing the reservoir into regions is a common task during history-matching processes (Gervais-Couplet et al. 2007), and is used with the purpose of performing local changes on the petrophysical values of reservoir model grid blocks. In our learning-from-data approach, we use the regions during Stage 1 (Sect. 2.5), that generates new solutions changing the values of petrophysical attributes.

In this second round of experiments, our goal is to compare the performance of our soft clustering strategy to partition the reservoir grid into regions (Sect. 2.5.1) with a hard assignment strategy. With this purpose, we implemented a “hard assignment” version of our learning-from-data approach, replacing the *Soft Region Delimiter* component in Stage 1 (Fig. 1) with the “*Region Definition*” component

Fig. 10 Porosity map (layer 1) on initial best solution (A), generated with commercial software for geostatistical modeling, and on final best solution (B), obtained with the execution of our learning-from-

data approach on scenario UNISIM-I-H-A S1. Pattern (B), generated by our algorithm, is visually realistic and consistent from a geological point of view

Fig. 11 Horizontal Permeability I map (layer 1) on initial best solution (A), generated with commercial software for geostatistical modeling, and on final best solution (B), obtained with the execution of our

learning-from-data approach on scenario UNISIM-I-H-A S1. Pattern (B), generated by our algorithm, is visually realistic and consistent from a geological point of view

Fig. 12 Net to Gross Ratio map (layer 1) on initial best solution (A), generated with commercial software for geostatistical modeling, and on final best solution (B), obtained with the execution of our learn-

ing-from-data approach on scenario UNISIM-I-H-A S1. Pattern (B), generated by our algorithm, is visually realistic and consistent from a geological point of view

Fig. 13 Oil and water field production curves in the historical and forecast periods, for the initial solutions and the final 100 best solutions obtained with the execution of our algorithm with scenario S1 from the UNISIM-I-H-A dataset

from Cavalcante et al. (2017). The “*Region Definition*” component implements a hard clustering (*k*-means based) strategy to define the reservoir regions, and works on the same space of spatial coordinates and petrophysical properties of each grid block, adopted by our *Soft Region Delimiter* (Sect. 2.5.1). In both cases, once the regions are delimited, the values of the petrophysical attributes on the grid blocks of the regions are changed using “*Path Relinking on Petrophysical Attributes*” (Sect. 2.5.2), that progressively incorporates into the current best global solution, the petrophysical patterns present in the best solutions for a well and for a particular mismatched variable.

Figure 14 shows, for the scenarios from UNISIM-I-H-A dataset, the misfit of the global best solution obtained with

the original version of our learning-from-data approach, and the corresponding value obtained with the hard assignment version of our approach. The values of Fig. 14 show that, in four out of the five scenarios, the misfit value obtained with the soft version was better than the value obtained with the version using the hard assignment strategy. It is also noteworthy that, for scenarios S1, S2, S3, and S4, both hard and soft assignment strategies behaved similarly, while, for scenario S5, the soft assignment strategy was clearly superior. The point that needs to be clear is that both hard and soft algorithms to delimit reservoir regions dynamically use the spatial coordinates and the petrophysical values of each grid block in the current global best solution. Therefore, depending on the petrophysical values of the current

Fig. 14 Misfit of global best solution after the execution of our learning-from-data approach, using the *Soft Region Delimiter* and the (hard) *Region Definition* from Cavalcante et al. (2017), with scenarios from the UNISIM-I-H-A dataset. In four (out of five) of the tested scenarios, the use of a soft clustering strategy to delimit reservoir regions resulted in a smaller misfit of the global best solution than the one obtained with a hard clustering strategy

best global solution of each scenario, the region associated with a particular well may have a higher or a smaller number of grid blocks. The interesting aspect of the proposed soft clustering algorithm to delimit the reservoir regions is that, once depending on the petrophysical values of the best global solution, the same grid block can belong to regions of more than just one well. For this reason, regions obtained by the soft clustering algorithm are potentially bigger (with more grid blocks) than the corresponding regions obtained by the hard clustering algorithm, and hence the impact of changing the soft region is potentially higher (given that

more grid blocks are affected) than the impact of changing the corresponding hard region. For scenarios S1 to S4, the petrophysical values of the global best solution are such that regions delimited with the hard and soft clustering algorithms have a similar number of grid blocks. However, the petrophysical values of the global best solution in scenario S5 result in soft regions with up to $9 \times$ more blocks than the corresponding hard regions, which justifies the observed different behavior of the soft and hard clustering algorithms in Scenario S5 and highlights the representation power of the proposed solution.

Figure 15 shows, for scenario S1 from the UNISIM-I-H-A dataset, the plots “Number of models versus Maximum value of NQD indicator” and “Number of models by Number of variables with history-matching quality at least good”. In both plots, the green and orange curves correspond, respectively, to the final 100 best models obtained with the original (soft) version, and with the modified (hard) version of our learning-from-data approach. The plot on the left of Fig. 14 indicates that, for scenario S1, the use of our soft clustering strategy to partition the reservoir resulted in more models with smaller maximum values of the NQD indicator. Note, for example, that while the ‘Final-soft’ curve shows more than 60 models with maximum NQD 15, the ‘Final-hard’ curve indicates only 40 models respecting this maximum value of *NQD*.

The plot on the right of Fig. 15 shows that, in terms of the number of variables that are at least good, both the soft and hard versions ended up with roughly 20 models with 71 (out of 78) variables with at least good history-matching quality. Note, however, that, for 90% of the 100 final best models obtained with both strategies, the curve corresponding to our original (soft clustering) approach is more vertical and

Fig. 15 “Number of models by max NQD” and “Number of models by number of variables with $NQD \leq 5$ ”, after the execution of our learning-from-data approach, using the *Soft Region Delimiter* and the (hard) *Region Definition* from Cavalcante et al. (2017), with scenario

S1 from UNISIM-I-H-A dataset. The green curves corresponding to solutions obtained with the soft clustering approach indicate more solutions with smaller max NQD and more models with more variables with $NQD \leq 5$

Table 7 Quality of final models and total number of simulations of our learning-from-data approach compared with values reported by five previous studies on the UNISIM-I-H benchmark

	% Final Models with max $NQD \leq 5$	% Final Models with max $NQD \leq 10$	% Final Models with max $NQD \leq 20$	Total Number of Simulations
Our learning-from-data approach	0	0	95	344
Mesquita et al. (2015)	0	0	3	2100
Cavalcante et al. (2017)	0	0	33	185
Oliveira et al. (2017)	0	0	0	3800
Maschio and Schiozer (2018)	0	100	100	10,350
Soares et al. (2018)	0	83.2	≥ 83.2	2000

more displaced to the right (indicating more models with more variables with at least good history-matching quality) than the curve corresponding to the hard assignment version. Therefore, our proposal of using a soft clustering strategy to delimit reservoir regions during a history-matching process seems to be promising and may result in solutions with better history-matching quality than the ones obtained with a hard assignment policy.

Experiments–round #3

Since the UNISIM-I-H benchmark was publicly released by Avansi and Schiozer (2015b), it has been used to validate different history-matching approaches: Mesquita et al. (2015), Bertolini et al. (2015), Oliveira et al. (2017), Maschio and Schiozer (2016, 2018), Cavalcante et al. (2017), and Soares et al. (2018). Although each of these approaches has its specificities and eventually uses a parameterization that is different from the one adopted in this work (Sect. 3.1), most of them use the same NQD indicator (Sect. 2.2) to assess the quality of solutions generated during the history-matching process. In this third round of experiments, our goal is to compare the results obtained with our learning-from-data approach with the best results reported by five of the previous methodologies applied to the UNISIM-I-H benchmark, considering two aspects: (i) the total number of simulations used by each approach; (ii) the quality of solutions found, measured by the percentage of final models having a maximum NQD value of no more than 5, 10 or 20 (values corresponding to the upper limits of quality levels ‘Good’, ‘Regular’, and ‘Bad’ from Table 1). For a summary of these five methodologies, the reader is referenced to Appendix B.

Table 7 shows, for the two mentioned comparison aspects, the best values obtained with our learning-from-data approach, and the corresponding best values reported by these approaches when applied to the UNISIM-I-H benchmark. None of the approaches were able to find models with $\max NQD \leq 5$. This means that, in all the compared approaches, there is no final model where all variables involved in the history-matching process ended up with

quality level ‘Good’ or better (Table 1). This reinforces how challenging the UNISIM-I-H benchmark is.

Regarding the quality of final models, Table 6 shows that the approach from Maschio and Schiozer (2018) found more models with $\max NQD \leq 10$. However, the total number of simulations required by this approach is the highest: it takes more than $30 \times$ the number of simulations when compared to our approach, and it is almost three times higher than the second highest number of simulations among the compared approaches.

At this point, and before continuing the discussion about the values from Table 7, we need to highlight that the simulation time of a reservoir model does not depend on the history-matching approach that launches the simulation. The simulation time depends only on the machine where the simulation is executed. For this reason, the simulation time of the UNISIM-I-H model, regardless of the history-matching approach, if executed in the same machine, will take exactly the same time. The number of simulations is, indeed, the primary indicator about the efficiency of history-matching approaches that are based on numeric reservoir simulation, which is the case of our approach and all the ones used in the comparison of Table 7. Another important performance indicator is the degree of parallelism of each approach, especially with regard to the parallel execution of the simulations. Time indicators such as CPU time are rarely reported in the literature of numeric simulation-based approaches given the diversity of computational resources from one research team to another. In fact, to properly compare the execution time of different history-matching approaches that use numeric reservoir simulations, it would be necessary to execute all of them in the same machine, with the fundamental condition that the machine would allow the optimal (maximal) parallel performance of each of the compared approaches. Unfortunately, analysis about the degree of parallelism of history-matching approaches, although very important, is rarely found in the literature of history-matching. In the specific comparison between the performance of our approach with the one reported by Maschio and Schiozer (2018) for the UNISIM-I-H benchmark, we can say that:

- (i) Both approaches support the parallel execution of simulations;
- (ii) In the approach from Maschio et Schiozer (2018), the degree of parallelism is given by $\max(N_c, nSA)$, where N_c is the number of combinations per iteration of the IDLHC method, and nSA is the number of Simulated Annealing processes. Given that, in the experiments with UNISIM-I-H benchmark, they configured the IDLHCSA algorithm with $N_c = 450$ and $nSA = 30$, the optimal performance of their approach occurs when it is executed in a machine that supports 450 simulations in parallel;
- (iii) In our approach, the degree of parallelism is given by $\max(nS1, nS2)$, where $nS1$ is the number of solutions generated during each execution of the *Path Relinking on Petrophysical Attributes* component, and it is bounded to “3stepSize” (Sect. 2.5.2); and $nS2$ is the number of solutions generated during each execution of the *Path Relinking on Global Attributes* component, and it is bounded to “2 * number of global uncertain attributes” (Sect. 2.6.1). Given that, in our experiments with the UNISIM-I-H benchmark, *stepSize* was set to 0.25 (Sect. 3.2), and we considered five global uncertain attributes (Sect. 3.1), the optimal performance of our approach with the UNISIM-I-H benchmark would occur when it is executed in a machine supporting 12 simulations in parallel.
- (iv) Considering (ii) and (iii), it is noteworthy that the machine requirements for the optimal performance of the approach from Maschio and Schiozer (2018) with the UNISIM-I-H benchmark is much higher than the machine requirements for the optimal performance of our approach with same benchmark: 450 simulations in parallel versus 12 simulations in parallel.
- (v) Hypothetically considering that (a) we could execute both approaches in the ideal machine for the approach from Maschio and Schiozer (2018), and (b) the simulation time of UNISIM-I-H model in such machine takes T minutes, we can say that the optimal (minimal) execution time of the approach from Maschio and Schiozer (2018) for the UNISIM-I-H benchmark in this ideal machine is $(8T + 225T) = 233T$ minutes, since in the experiments the authors configured the IDLHCSA method with 8 IDLHC iterations and 225 SA iterations. Also, in this same ideal machine, the optimal execution time of our approach with scenario S1 from the UNISIM-I-H-A dataset is $(16T + 7T) = 23T$ minutes, since for this scenario the *Path Relinking on Petrophysical Attributes* component was executed 16 times, and the *Path Relinking on Global Attributes* component was executed 7 times. Comparing the values of the optimal performance of both approaches in this “ideal” machine, we can say that our approach is almost

10 times faster than the approach from Maschio and Schiozer (2018). And this superior performance of our approach will be even more evident the further away from the ideal machine configuration is the machine on which the approach from Maschio and Schiozer (2018) is actually run. To clarify a bit more, we executed all the experiments of our manuscript in a Windows 7 machine with Intel® Core™ 2 Duo CPU E 8600 3.33 GHz and 8 GB RAM, where the simulations could only be executed sequentially. In this machine, the simulation time of UNISIM-I-H model was 3 to 5 min. As Table 8 indicates, our approach required 344 simulations, which results in an execution time between 1,032 to 1,720 min (17hs < time < 29hs). If the approach from Maschio and Schiozer (2018), that requires 10,350 simulations, had been executed in the same machine as ours, it would have resulted in an execution time between 31,050 and 51,750 min (about 21 days, in the best case).

Back to the discussion of the values presented in Table 7, the approach from Soares et al. (2018) seems to be the one with the best tradeoff between the quality of the final models and the total number of simulations. Our learning-from-data approach appears as the one with the lowest number of simulations that always found models with $\max NQD \leq 20$ (models for which no variable involved in the history-matching process has matching quality ‘Unsatisfactory’, according to Table 1).

From the compared approaches, the only ones that use the same parameterization we adopted in this work for the UNISIM-I-H benchmark (Sect. 3.1) are the approaches from Mesquita et al. (2015) and Cavalcante et al. (2017). Compared with Mesquita et al (2015), Table 7 shows that our approach is clearly superior, as it is able to find more models with $\max NQD \leq 20$, using a much smaller number of simulations. Compared with Cavalcante et al. (2017), our learning-from-data approach uses more simulations but it is able to find more approach with $\max NQD \leq 20$. Table 8 and Table 9 show, respectively, the total number of simulations

Table 8 Total number of simulations used by our learning-from-data approach, and reported by Cavalcante et al. (2017) for the five scenarios on the UNISIM-I-H-A dataset

	Total Number of Simulations				
	S1	S2	S3	S4	S5
Our learning-from-data approach	344	393	344	264	311
Cavalcante et al. (2017)	190	185	175	185	165

Table 9 Quality of final models found by our learning-from-data approach, and reported by Cavalcante et al. (2017) for the five scenarios on the UNISIM-I-H-A dataset

	% Final Models with max NQD ≤ 20				
	S1	S2	S3	S4	S5
Our learning-from-data approach	95	42	59	38	28
Cavalcante et al. (2017)	25	0	0	33	0

Table 10 Performance of our learning-from-data approach compared with five previous studies on the UNISIM-I-H benchmark. An up-arrow ($\uparrow$) indicates a better result from our approach, and a down-arrow ($\downarrow$) indicates a better result from the corresponding literature approach

	Matching Quality of Final Models	Total Number of Simulations
Mesquita et al. (2015)	$\uparrow$	$\uparrow$
Cavalcante et al. (2017)	$\uparrow$	$\downarrow$
Oliveira et al. (2017)	$\uparrow$	$\uparrow$
Maschio and Schiozer (2018)	$\downarrow$	$\uparrow$
Soares et al. (2018)	$\downarrow$	$\uparrow$

and the percentage of final models with $NQD \leq 20$ obtained by our approach and reported by Cavalcante et al. (2017) for each scenario from the UNISIM-I-H-A dataset. Note that, even if the total number of simulations of our approach is higher than the one reported by Cavalcante et al. (2017), it is still substantially lower than the values reported by the other four approaches considered in the comparison of this section (Table 7). In terms of quality of the final models, the values in Table 9 shows that our approach is clearly superior than the one from Cavalcante et al. (2017) for all scenarios from the UNISIM-I-H-A dataset.

Concluding the analysis of this third round of experiments, Table 10 summarizes the performance of our learning-from-data approach versus the other five previous approaches applied to UNISIM-I-H. In this table, an up-arrow ($\uparrow$) indicates a better result from our approach, and a down-arrow ($\downarrow$) indicates a better result from the corresponding literature approach. From Table 10, we can note that, regarding the number of simulations, our learning-from-data approach performs better than four (out of five) previous approaches applied to the UNISIM-I-H benchmark. In terms of the matching quality of the final solutions, our approach was better than three (out of five) approaches considered in the comparison. Note, however, that the worse matching quality of the final best solutions obtained with our approach, when compared to the approaches from Maschio and Schiozer (2018) and Soares et al. (2018), must

be considered with caution since these two studies use different parameterizations of the UNISIM-I-H benchmark. Both of them consider uncertain attributes which are not present in the parameterization adopted in this work (Sect. 3.1), and even for the uncertain attributes that are also considered in this work, they either discretize them in more levels than we do, or treat them as continuous, non-discretized, attributes. So, the fact that our approach could not find solutions with the same quality found by these two approaches does not invalidate our proposal, especially when we consider that our approach has always found solutions with superior quality when compared to previous approaches that use the same parameterization (Mesquita et al. 2015; and Cavalcante et al. 2017). The results presented here only suggest that the parameterization adopted by Maschio and Schiozer (2018), or by Soares et al. (2018) may be worth trying, which will definitely serve as a guiding factor in our future work.

Conclusions and future work

In this work, we proposed a learning-from-data approach with soft clustering and path relinking for the history-matching problem. We applied our approach to the complex UNISIM-I-H benchmark, and the results show the potential the strategy has towards finding solutions with good matching quality, geologically consistent appearance, and acceptable forecast capability while using a small number of simulations.

From the perspective of assisted techniques to the history-matching problem, we conclude that the proposed and innovative combination of a soft clustering strategy to delimit reservoir regions, with a path-relinking approach to progressively change values of uncertain attributes in the global best solution is clearly efficient and effective in improving the history-matching quality of an initial set of solutions. We also learn that in comparison with a hard assignment strategy, recently published in the literature to partition the reservoir grid into regions, our new soft clustering approach was able to find solutions with better matching quality in four (out of five) scenarios used in the experiments.

Compared with five previous history-matching methodologies, applied to the same UNISIM-I-H benchmark, we conclude from the results obtained with our approach that its learning-from-data nature makes it competitive both in terms of quality and efficiency: the solutions found during our experiments have better quality than the best final solutions found by three of the five compared approaches, and our approach required at least 80% fewer simulations than four of the five compared approaches.

Finally, we conclude that the combination of machine learning with optimization techniques is very suitable to guide a history-matching process towards solutions with

improved matching quality, a reliable forecast of field performance, and all of this with a small number of simulations, what certainly makes strategies such as our learning-from-data approach promising and worth considering while solving history-matching problems.

Future directions of our work include exploring learning strategies that generate new solutions considering not only patterns of input attributes present in the best solutions for each variable (or well), but also in any solution which has a minimum level of quality. Another research possibility involves the introduction of randomness in the learning-from-data approach and the design of a component for changing values of continuous global uncertain attributes, without the need of creating discrete levels for them.

Appendix A

Theoretical background

In this Appendix, we briefly describe the two main techniques that form the underpinnings of our learning-from-data approach to the history-matching problem. For a deeper explanation of these techniques, we point the readers to the key references given in the corresponding subsections.

Soft K-means clustering

In the context of machine learning, clustering is the task of organizing objects (data points) into groups (clusters) based on some similarity criteria, in such a way that data points within the same cluster are more similar to each other than they are to points belonging to a different cluster (Jain et al. 1999).

While several clustering techniques exist (Grabmeier and Rudolph 2002), they are normally divided into two different groups, depending on the policy with which data points are assigned to clusters. In hard clustering strategies, each data point is assigned to one (and only one) cluster. In soft clustering approaches, by contrast, clusters can overlap and each data point can simultaneously belong to multiple clusters.

The soft k-means algorithm (MacKay 2003; Bauckhage 2015) is a soft clustering strategy, which calculates membership degrees to which data points belong to clusters. Algorithm A.1 shows a high-level description for soft k-means clustering, adapted from the one presented in Bauckhage (2015). The algorithm is based on the traditional (hard) k-means procedure (Bishop 2006), differing only in the assignment step. While the hard k-means algorithm implements a binary 0–1 assignment of data points to clusters, depending on distances between data points to centroids, the soft k-means uses these same distances to calculate degrees of cluster membership (line 3). The β parameter in line 3 is called the *stiffness* parameter, and it is such that the larger the value, the smaller the degree of membership to cluster i of a data point x_j far from centroid c_i (Bauckhage 2015). In fact, as the *stiffness* parameter goes to ∞ , the soft k-means algorithm becomes identical to the traditional (hard) k-means procedure, where each data point is assigned to exactly one cluster (MacKay 2003). Although there are a number of general ideas regarding the choice of the β parameter (e.g. setting β as a multiple of the average distance between data points), Bauckhage (2015) recommends setting β equal to 1 as a *pragmatic* choice when one is in doubt about which value to use. The stop criteria of the soft k-means clustering algorithm (line 5) is the same from the traditional (hard) k-means procedure (Bishop 2006): the algorithm stops when the new centroids (calculated in line 4) do not change from one iteration to another, or if the changes in the centroids are below some small *epsilon* value.

Algorithm A.1 Soft K-Means Clustering

Input: $X = \{x_1, \dots, x_n\}$, set of data points; k , number of clusters

1: Randomly initialize the centroids set $C \subset X$:

$$C = \{c_1, \dots, c_k \mid \forall i \neq j, c_i \neq c_j\}$$

2: Repeat

3: For each data point $x_j \in X$, calculate the membership degree of x_j to cluster i

$$z_{ij} = \frac{e^{-\beta \|x_j - c_i\|^2}}{\sum_{l=1}^k e^{-\beta \|x_j - c_l\|^2}}$$

4: Update centroids

$$c_i = \frac{\sum_j z_{ij} x_j}{\sum_j z_{ij}}$$

5: until centroids stabilize.

Table 11 Summary of previous work on the UNISIM-I-H benchmark considered in the comparison with our learning-from-data approach

Authors	Description
Mesquita et al. (2015)	Propose a probabilistic and multi-objective approach that performs history-matching iteratively following two steps. The first step quantifies the matching quality of each model for each objective function (OF) and selects the highest mismatched OF to be matched. The second step constructs a bi-dimensional influence matrix that supports the identification of attributes that are source of the highest deviations. The method defines guidelines of actions to update and reduce uncertainties of the attributes identified in step two and, at each iteration, generates new models combining the updated model uncertainties through a robust sampling technique
Cavalcante et al. (2017)	Propose a deterministic optimization framework that follows a learning-from-data paradigm. Starting from an initial set of solutions, the approach has two optimization stages. The first stage uses unsupervised learning to uncover patterns of petrophysical attributes present in the best solutions for each well. The second stage performs statistical analysis and follows an expectation maximization approach to evaluate the levels of global attributes and select promising ones. At each optimization stage, new models are generated, simulated and their data is dynamically considered to improve the learning process
Oliveira et al. (2017)	Propose a geostatistical history-matching methodology to reduce uncertainties in the characterization of reservoir petrophysical properties. The method does not consider the impact of global attributes, and its main purpose is to generate multiple models that preserve well data, production data and geological knowledge. The method iteratively performs a sensitive analysis of petrophysical properties to identify reservoir regions and attributes that affect the history-matching objective functions, and uses this information as input to perturbation methods that generate new geostatistical realizations (Probability Perturbation Method for facies modelling, and image perturbation with Co-simulation for porosity and permeability modelling)
Maschio and Schiozer (2018)	Propose a stochastic method that combines a sampling phase (based on Iterative Discrete Latin Hypercube, IDLHC) with an intensification phase from multiple start-points (based on Simulated Annealing). During the sampling phase with IDLHC, new models are iteratively generated using the Discrete Latin Hypercube, and a percentage of the best models is selected to guide the update of the probability distribution functions of each uncertain attribute. The second phase starts from the good solutions found in the first phase, and its main purpose is to refine the search, finding local minimum solutions around the best solutions found in the sampling phase. In both phases, the benefits of parallel computing are efficiently explored by the parallel nature of IDLHC and (parallel) Simulated Annealing
Soares et al. (2018)	Propose the use of the Ensemble Smoother with Multiple Data Assimilation (ES-MDA) method (Emerick and Reynolds 2013), in conjunction with two distance-dependent covariance localization approaches. The purpose of the methodology is to reduce spurious correlations and geological inconsistencies while assimilating the data. The application of distance-dependent localization techniques consists in defining elliptic regions of influence for each well. In the first approach, the regions of influence are delimited using the influence area of each producer-injector pair. In the second localization approach, the influence regions are defined using the streamlines on the model constructed with the mean of all initial models in the ensemble

Path relinking

The path relinking approach was introduced by Glover and Laguna (1993) as a strategy to pursue the goals of intensification and diversification while exploring solution spaces. Since then, the method has been proven effective for solving difficult discrete and nonlinear optimization problems, from both classical and real-world settings (Glover et al. 2000; Resende et al. 2010). The method proposes systematic rules to explore trajectories connecting elite solutions (which are, by definition, good-quality solutions) with the reasoning that these solutions may share a significant subset of attributes, and the paths between them can possibly contain better solutions (Glover 1999).

The implementation of a path-relinking approach, regardless of the problem being tackled, must deal with three main

issues (Ho and Gendreau 2006): (i) rules for building the *reference set*, containing elite solutions that will be the extreme points of the generated paths; (ii) rules for choosing the *initial* and *guiding solutions*, the start and ending points of each desired path; and (iii) the *neighborhood structure*, which defines the way the algorithm will move along the paths.

Algorithm A.2 shows a high-level description with the main steps of path relinking. Note that, once the *initial* and *guiding* solutions are chosen (line 3), the algorithm explores solutions obtained through the progressive incorporation into the current solution of attributes contributed by the *guiding* solution (line 7). At each iteration of lines 5 to 11, the solutions explored are farther from the *initial* solution and closer to the *guiding* solution.

Algorithm A.2 Path Relinking

```

1: Build the reference set  $\mathcal{R}$  with elite solutions
2: do
3:   Select initial solution  $s_i$  and guiding solution  $s_g$  from  $\mathcal{R}$ 
4:   Set  $x = s_i$ 
5:   do
6:     Let  $\Delta_{sg}^{s_i}$  be the set of attributes contained in  $s_g$  and not contained in  $x$ :
            $\Delta_{sg}^{s_i} = \{\delta_1, \dots, \delta_n\}$ 
7:     Let  $\mathcal{N}(x)$  be the set of neighbor solutions of  $x$ , each of them obtained through
           the incorporation in  $x$  of one attribute  $\delta_j \in \Delta_{sg}^{s_i}$ 
8:     Select  $\bar{x}$  from  $\mathcal{N}(x)$ 
9:     Evaluate  $\bar{x}$ 
10:    Set  $x = \bar{x}$ 
11:  while  $x \neq s_g$ 
12: while stop criteria not reached

```

Appendix B

Summary of previous work on the UNISIM-I-H benchmark considered in the comparison with our learning-from-data approach

This Appendix summarizes in Table 11 the five previous approaches we considered in the comparison with our learning-from-data approach. Among the previous approaches applied to the UNISIM-I-H, these five studies were the ones publicly available by the time we ran the experiments and wrote this manuscript, and which reported results using the NQD indicator (Sect. 2.2).

Acknowledgements This study was financed in part by the “Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001”. Apart from CAPES, we also thank CNPq (grants 306454/2018-1 and 304497/2018-5), UNISIM-CEPETRO-UNICAMP, Petrobras (SIGER Network), Energi and ANP for their support in the development of this work, and also CMG for the IMEX software licenses.

Declarations

Conflict of interest On behalf of all the co-authors, the corresponding author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are

included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aanonsen SI, Nævdal G, Oliver DS, Reynolds AC, Vallès B (2009) Ensemble Kalman filter in reservoir engineering—a review. SPE J 14(3):393–412
- Abdollahzadeh A, 2014. Adaptive Algorithms for History Matching and Uncertainty Quantification. PhD Thesis, Institute of Petroleum Engineering, Heriot-Watt University, UK.
- Abdollahzadeh A, Reynolds A, Christie MA, Corne D, Williams G, Davies B (2013) Estimation of Distribution Algorithms Applied to History Matching. SPE J 18(3):508–517
- Abu-Mostafa YS, Magdon-Ismael M, Lin HT (2012) Learning from Data. New York, AMLBook
- Anterion, F., Eymard, R., Karcher, B., 1989. Use of parameter gradients for reservoir history matching. SPE-18433, In SPE Symposium on Reservoir Simulation, Society of Petroleum Engineers, pp 6–8 February, Houston, Texas.
- Aurenhammer F (1991) Voronoi Diagrams—A Survey of a Fundamental Geometric Data Structure. ACM Comput Surv 23(3):345–405
- Avansi GD, Schiozer DJ (2015a) UNISIM-I: synthetic model for reservoir development and 1114 management applications. Int J Model Simulat Petrol Ind 9(1):21–30
- Avansi, G.D., Schiozer, D.J., 2015b. A New Approach to History Matching Using Reservoir Characterization and Reservoir Simulation Integrated Studies. Offshore Technology Conference. May, Houston. Paper OTC-26038-MS.
- Bauchhage, C. (2015). Lecture notes on data science: Soft k-means clustering. Technical Report, Univ. Bonn, DOI: <https://doi.org/10.13140/RG.2.1.3582.6643>
- Bertolini AC, Maschio C, Schiozer DJ (2015) A methodology to evaluate and reduce reservoir uncertainties using multivariate distribution. J Petrol Sci Eng 128:1–14

- Bishop C (2006) Pattern Recognition and Machine Learning. Springer
- Caeiro MH, Demyanov V, Soares A (2015) Optimized history matching with direct sequential image transforming for non-stationary reservoirs. *Math Geosci* 47:975–994
- Cavalcante CCB, Maschio C, Santos AA, Schiozer D, Rocha A (2017) History matching through dynamic decision-making. *PLoS ONE* 12(6):e0178507
- Cavalcante CCB, Maschio C, Santos AA, Schiozer D, Rocha A (2019) A continuous learning algorithm for history matching. *Eng Appl Artif Intell* 85:543–568
- Cullick, A.S., Johnson, D., Shi, G., 2006. Improved and more-rapid history matching with a nonlinear proxy and global optimization. SPE 101933-MS in SPE Annual Technical Conference and Exhibition, pp 24–27 September, San Antonio, Texas, USA.
- Dickstein F, Goldfeld P, Pfeiffer GT, Pinto RV (2018) Truncated conjugate gradient and improved LBFGS and TSVD for history matching. *Comput Geosci* 22:309
- Elsheikh AH, Jackson MD, Laforce TC (2012) Bayesian reservoir history matching considering model and parameter uncertainties. *Math Geosci* 44:515–543
- Emerick AA (2016) Analysis of the performance of ensemble-based assimilation of production and seismic data. *J Petrol Sci Eng* 139:219–239
- Emerick, A., 2017. History Matching and Forecasting. *Journal of Petroleum Technology*, 69(4).
- Emerick AA (2018) Deterministic ensemble smoother with multiple data assimilation as an alternative for history-matching seismic data. *Comput Geosci* 22:1175
- Emerick AA, Reynolds AC (2013) Ensemble smoother with multiple data assimilation. *Comput Geosci* 55:3–15
- Ertekin T, Abou-Kassem JH, King GR (2001) Basic Applied Reservoir Simulation. Richardson, Texas, SPE Text Book series
- Evensen, G., Hove, J., Meisingset, H., Reiso, E., Seim, K. S., Espelid, O., 2007. Using the EnKF for Assisted History Matching of a North Sea Reservoir Model. SPE 106184, In SPE Reservoir Simulation Symposium, pp 26–28 February, Houston, Texas, USA.
- Gervais-Couplet, V., Gautier, Y., Le Ravalec-Dupin, M., Roggero, F., 2007. History Matching Using Local Gradual Deformation. SPE 107173-MS, In EUROPEC/EAGE Conference and Exhibition, pp 11–14 June, London, UK. doi:<https://doi.org/10.2118/107173-MS>
- Glover F (1999) Scatter Search and Path Relinking. In: Corne D, Dorigo M, Glover F (eds) *New ideas in optimization*. McGraw Hill, New York, pp 297–316
- Glover F, Laguna M (1993) Tabu search. In: Reeves CR (ed) *Modern heuristic techniques for combinatorial problems*. Blackwell Scientific Publishing, Oxford, pp 70–150
- Glover F, Laguna M, Marti R (2000) Fundamentals of scatter search and path relinking. *Control Cybern* 39:653–684
- Grabmeier J, Rudolph A (2002) Techniques of cluster algorithms in data mining. *Data Min Knowl Disc* 6:303–360
- Hajizadeh Y, Christie M, Demyanov V (2011) Ant colony optimization for history matching and uncertainty quantification of reservoir models. *J Petrol Sci Eng* 77(1):78–92
- Ho SC, Gendreau M (2006) Path relinking for the vehicle routing problem. *J Heuristics* 12:55–72
- Ho YZ, Pepyne DL (2002) Simple explanation of the no-free-lunch theorem and its implications. *J Optim Theory Appl* 115(3):549–570
- Hoffman, B., Caers, J., 2003. Geostatistical History Matching Using a Regional Probability Perturbation Method, SPE 84409-MS, In SPE Annual Technical Conference and Exhibition, pp 5–8 October, Denver, Colorado, USA.
- Hoffman BT, Caers J (2005) Regional probability perturbations for history matching. *J Petrol Sci Eng* 46:53
- Hou J, Zhou K, Zhang X, Kang X, Xie H (2015) A review of closed-loop reservoir management. *Pet Sci* 12(1):114–128
- Jacquard P, Jain C (1965) Permeability distribution from field pressure data. *Soc Pet Eng J* 5(4):281–294
- Jain AK, Murty MN, Flynn PJ (1999) Data Clustering: A Review. *ACM Comput Surv* 31(3):264–323
- Jansen J-D, Brouwer DR, van Kruijsdijk CPJW, Nævdal G (2005) Closed-loop reservoir management. *First Break* 23(1):43–48
- Jansen, J.-D., Brouwer, D.R., Douma, S.G., 2009. Closed-loop reservoir management. In SPE Reservoir Simulation Symposium, The Woodlands: Texas.
- Le DH, Emerick AA, Reynolds AC (2016) An adaptive ensemble smoother with multiple data assimilation for assisted history matching. *Soc Petrol Eng*. <https://doi.org/10.2118/173214-PA>
- Li B, Bhark EW, Gross SJ, Billiter TC, Dehghani K (2018) Best practices of assisted history matching using design of experiments. *Soc Petrol Eng*. <https://doi.org/10.2118/191699-MS>
- MacKay DJC (2003) Information theory, inference, and learning algorithms. Cambridge University Press
- Maschio C, Davolio A, Correia MG, Schiozer DJ (2015) A new framework for geostatistics-based history matching using genetic algorithm with adaptive bounds. *J Petrol Sci Eng* 127:387–397
- Maschio C, Schiozer DJ (2014) Bayesian History matching using artificial neural network and Markov Chain Monte Carlo. *J Petrol Sci Eng* 123:62–71
- Maschio C, Schiozer DJ (2016) Probabilistic history matching using discrete Latin Hypercube sampling and nonparametric density estimation. *J Petrol Sci Eng* 147:98–115
- Maschio C, Schiozer DJ (2018) A new methodology for history matching combining iterative discrete Latin Hypercube with multi-start simulated annealing. *J Petrol Sci Eng* 169:560–577
- Mesquita, F.B., Davolio, A., Schiozer, D.J., 2015. A Systematic Approach to Uncertainty Reduction with a Probabilistic and Multi-Objective History Matching. SPE-174359-MS, SPE EUROPEC 2015, 1–4 June, Madrid, Spain.
- Milliken WJ, Emanuel AS, Chakravarty A (2001) Applications of 3D streamline simulation to assist history matching. *SPE J RE* 4(06):502–508
- Mohamed L, Christie M, Demyanov V (2010) Comparison of stochastic sampling algorithms for uncertainty quantification. *SPE J* 15(1):31–38
- Oliveira GS, Schiozer DJ, Maschio C (2017) History matching by integrating regional multi-property image perturbation methods with a multivariate sensitivity analysis. *J Petrol Sci Eng* 153:111–122
- Oliver DS, Chen Y (2011) Recent progress on reservoir history matching: a review. *Comput Geosci* 15(1):185–221
- Oliver DS, Reynolds AC, Liu N (2008) *Inverse Theory for Petroleum Reservoir Characterization and History Matching*, 1st edn. Cambridge University Press, Cambridge
- Petrovska, I., 2009. Estimation of distribution algorithms for reservoir history matching optimisation. PhD Thesis, Imperial College, London, UK.
- Ranazzi, P.H., Pinto, M.A.S., 2017. ASSISTED HISTORY MATCHING USING COMBINED OPTIMIZATION METHODS. In CILAMCE 2017. Proceedings of the XXXVIII Iberian Latin-American Congress on Computational Methods in Engineering P.O. Faria, R.H. Lopez, L.F.F. Miguel, W.J.S. Gomes, M. Noronha (Editors), ABMEC, Florianópolis, SC, Brazil, November 5–8.
- Resende, M., Ribeiro, C., Glover, F., Marti, R., 2010. Scatter Search and Path-Relinking: Fundamentals, Advances, and Applications. *Handbook of Metaheuristics*. pp 87–107.
- Santhosh EC, Sangwai JS (2016) A hybrid differential evolution algorithm approach towards assisted history matching and uncertainty quantification for reservoir models. *J Petrol Sci Eng* 142:21–35
- Schiozer DJ, Avansi GD, Santos AAS (2016) Risk quantification combining geostatistical realizations and discretized Latin Hypercube. *J Braz Soc Mech Sci Eng* 39(2):575–587

- Schiozer D, Santos A, Santos S, von Hohendorff Filho J (2019) Model-based decision analysis applied to petroleum field development and management. *Oil Gas Sci Technol* 74(46):2019
- Shirangi MG, Durlofsky LJ (2016) A general method to select representative models for decision making and optimization under uncertainty. *Comput Geosci* 96:109–123
- Schulze-Riegert, R., Ghedan, S., 2007. Modern techniques for history matching. In: *Proceedings of the 9th International Forum on Reservoir Simulation*, 9–13 December, Abu Dhabi, United Arab Emirates.
- Skjervheim, J.-A., Evensen, G., 2011. An ensemble smoother for assisted history matching. SPE 141929, in *SPE Reservoir Simulation Symposium*, 21–23, February, The Woodlands, Texas, USA
- Soares RV, Maschio C, Schiozer DJ (2018) Applying a localization technique to Kalman Gain and assessing the influence on the variability of models in history matching. *J Petrol Sci Eng* 169:110–125
- Zachariassen, E., Skjervheim, J.A., Vabø, J.G., Lunt, I., Hove, J., Evensen, G., 2011. Integrated work flow for model update using geophysical monitoring data. In *Proceedings of the 73rd EAGE Conference & Exhibition*, Vienna, Austria.
- Zhang, F., Reynolds, A.C., 2002. Optimization algorithms for automatic history matching of production data. In: *Proceedings of 8th European Conference on the Mathematics of Oil Recovery*, 3–6 September, Freiberg, Germany.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.