TU Delft

Delft University of Technology

Clustering Techniques for Value-of-information Assessment in Closed-loop Reservoir Management

Goncalves Dias De Barros, Eduardo; Yap, F.K.; Insuasty, E; Van den Hof, PMJ; Jansen, Jan Dirk

DOI
10.3997/2214-4609.201601858

Publication date
2016

Document Version
Final published version

Published in
Proceedings of the 15th European Conference on the Mathematics of Oil Recovery

Citation (APA)
Goncalves Dias De Barros, E., Yap, F. K., Insuasty, E., Van den Hof, PMJ., & Jansen, J. D. (2016). Clustering Techniques for Value-of-information Assessment in Closed-loop Reservoir Management. In *Proceedings of the 15th European Conference on the Mathematics of Oil Recovery: Amsterdam, Netherlands* (pp. 1-19). EAGE. https://doi.org/10.3997/2214-4609.201601858

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' – Taverne project

https://www.openaccess.nl/en/you-share-we-take-care

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.

EAGE

ECMOR XV

Th Gra 06

Clustering Techniques for Value-of-information Assessment in Closed-loop Reservoir Management

E.G.D. Barros* (Delft University of Technology), F.K. Yap (Delft University of Technology), E. Insuasty (Eindhoven University of Technology), P.M.J. van den Hof (Eindhoven University of Technology) & J.D. Jansen (Delft University of Technology)

SUMMARY

Closed-loop reservoir management (CLRM) is a combination of life-cycle optimization and computer-assisted history matching. The application of the CLRM framework to real field cases can be computationally demanding. An even higher computational load results from procedures to assess the value of information (VOI) in CLRM. Such procedures, which are performed prior to field operation, i.e. during the field development planning (FDP) phase, require extreme amounts of simulations. Therefore, we look for alternatives to reduce this computational cost. In particular we compare various clustering techniques to select a limited number of representative members from an ensemble of reservoir models. Using K-means clustering, multi-dimensional scaling and tensor decomposition techniques, we test the effectiveness of different dissimilarity measures such as distance in parameter space, distance in terms of flow patterns and distance in optimal sets of controls. As a first step towards large-scale application we apply several of these measures to a VOI-CLRM exercise using a simple 2D reservoir model which results in a reduction of the necessary number of forward reservoir simulations from millions to thousands.

Introduction

Modern reservoir management workflows include uncertainty quantification (UQ) based on reservoir simulation models. The most rigorous UQ practice in the reservoir engineering community consists of using ensembles of reservoir model realizations to account for the geological uncertainties (i.e., quasi-Monte Carlo method), which contributes to increasing the computational costs of these workflows. Closed-loop reservoir management (CLRM) is a combination of life-cycle optimization and computer-assisted history matching, both accounting for uncertainties and demanding a significant amount of simulations. For this reason, the application of the CLRM framework in combination with UQ can be extremely computationally expensive. Workflows to assess the value of information (VOI) in CLRM during the field development planning (FDP) phase require even more simulations, which makes real-field applications unfeasible (Barros *et al*., 2016). Therefore, we look for alternatives to reduce this computational cost.

The development of more practical ways of a-priori assessing the value of future measurements has been a topic of several studies recently. Some of these have focused on the use of proxy models to reduce the number of high-fidelity reservoir simulations required for the VOI analysis (He *et al*., 2016). Others have proposed a more approximate definition of VOI which simplifies their procedure (Le and Reynolds, 2014a and 2014b). Eidsvik *et al*. (2015) have envisaged more sophisticated design of experiments to be a promising alternative to alleviate the computational costs of VOI assessment workflows. This paper explores the use of clustering techniques to select a subset of representative model realizations to achieve what Eidsvik *et al*. (2015) suggest within the workflow for VOI assessment proposed by Barros *et al*. (2016).

In the *Background* section we briefly recap our previously proposed methodology for VOI assessment in CLRM and review some previous work on cluster analysis. Next, in the *Methodology* section, we identify opportunities to apply clustering within the original procedure and we describe our approach to reduce the computational costs in different steps of the VOI assessment to come up with a more practical workflow. Thereafter, in the *Examples* section, we illustrate the application of the proposed measures to accelerate VOI calculations and we compare the results with those obtained with the original procedure. Finally, in the *Conclusion* section, we discuss the benefits and weaknesses of the use of clustering, and we comment on points for future work.

Background

Closed-loop reservoir management CLRM

CLRM is a combination of frequent life-cycle production optimization and data assimilation (also known as computer-assisted history matching). Life-cycle optimization aims at maximizing a financial measure, typically net present value (NPV), over the producing life of the reservoir by optimizing the production. This may involve well location optimization, or, in a more restricted setting, optimization of well rates and pressures for a given configuration of wells, on the basis of one or more numerical reservoir models. Data assimilation involves modifying the parameters of one or more reservoir models, or the underlying geological models, with the aim to improve their predictive capacity, using measured data from a potentially wide variety of sources such as production data or time-lapse seismics. For further information on CLRM see, e.g., Jansen *et al*. (2005, 2008, 2009), Naevdal *et al*. (2006), Sarma *et al*. (2008); Chen *et al*. (2009), Wang *et al*. (2009) or Hou *et al*. (2015).

VOI assessment in CLRM

Recently, we have proposed a new methodology to assess the VOI of future measurements by making use of the CLRM framework (Barros *et al*., 2016). Our approach presented there consists of 'closing the loop' in the design phase to simulate how information obtained during the producing life-time of the reservoir comes into play in the context of optimal reservoir management. By considering both data assimilation and optimization in the procedure, we are able "to not only quantify how information changes knowledge, but also how it influences the results of decision making". This is possible because a new production strategy is obtained every time the models are updated with new information, and the strategies with and without additional information can be compared in terms of

the value of the optimization objective function (typically NPV) obtained when applying these strategies to the virtual asset (synthetic truth); see flowchart in Figure 1.

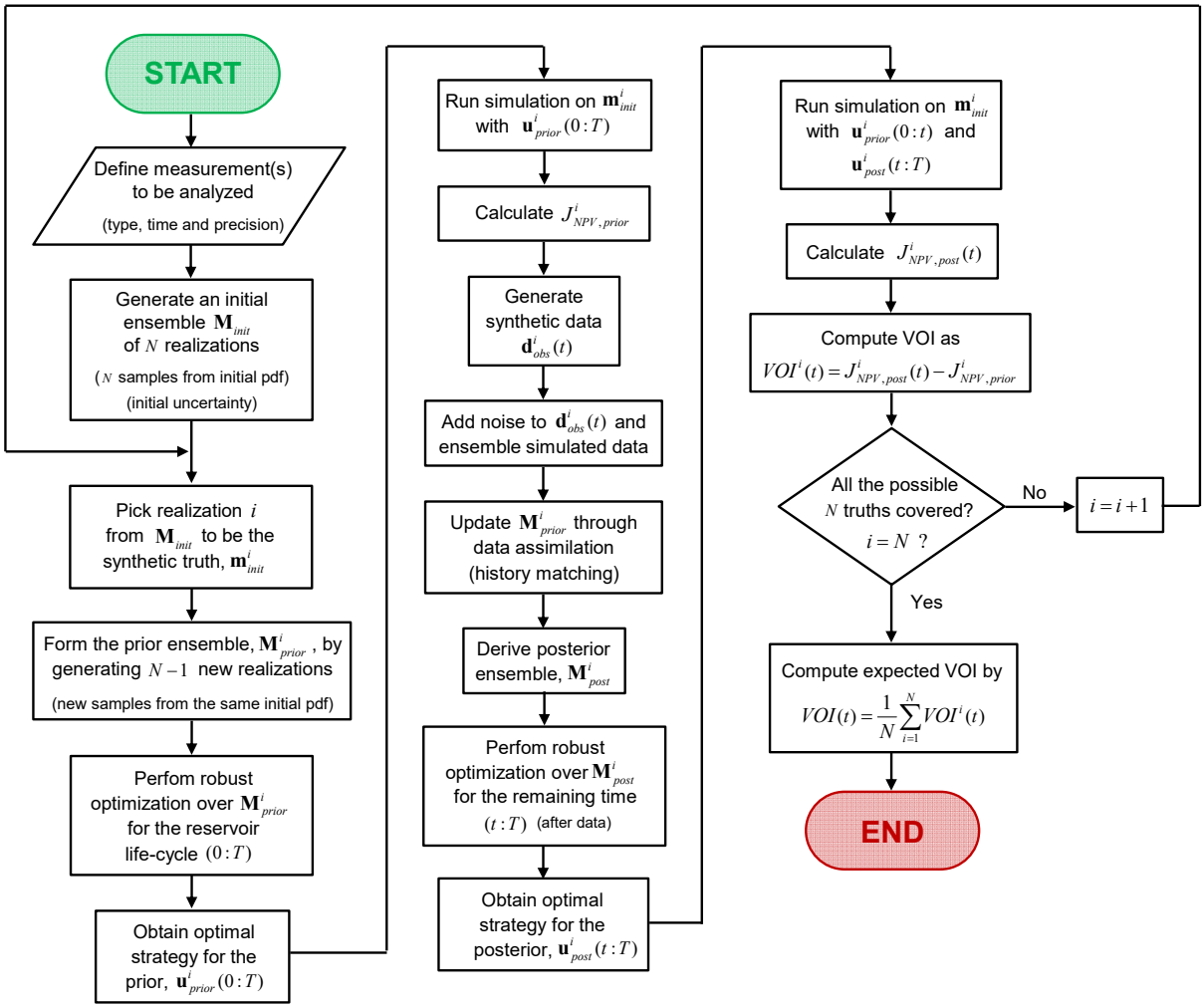

Figure 1 *Complete workflow to compute the expected VOI (from Barros et al., 2016).*

One of the key aspects of this methodology is the idea of using an ensemble of N plausible truths to account for the fact that in reality we do not know the true reservoir nor the outcome of the future measurements whose value we would like to assess. Note that Barros *et al*. (2016) propose also an accelerated procedure where the number of prior robust optimizations is reduced from N to 1. This is not depicted in Figure 1, but it is considered in the next sections of this paper.

Robust optimization

Robust life-cycle optimization uses one or more ensembles of geological realizations (reservoir models) to account for uncertainties and to determine the production strategy that maximizes a given objective function over the ensemble $\mathbf{M} = \{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_N\}$; see, e.g., Yeten *et al*. (2003) or Van Essen *et al* (2009). Typically, the objective function optimized is the net present value (NPV). J_{NPV} is defined as

$$J_{NPV} = \mu_{NPV} = \frac{1}{N}\sum_{i=1}^{N} J_i \quad , \tag{1}$$

where μ_{NPV} is the ensemble mean of the objective function values J_i of the individual realizations. The objective function J_i for a single realization i is defined as

$$J_i = \int_{t=0}^{T} \frac{q_o(t,\mathbf{m}_i)\, r_o - q_{wp}(t,\mathbf{m}_i)\, r_{wp} - q_{wi}(t,\mathbf{m}_i)\, r_{wi}}{(1+b)^{t/\tau}} dt \ , \tag{2}$$

where t is time, T is the producing life-time of the reservoir, q_o is the oil production rate, q_{wp} is the water production rate, q_{wi} is the water injection rate, r_o is the price of oil produced, r_{wp} is the cost of water produced, r_{wi} is the cost of water injected, b is the discount rate expressed as a fraction per year, and τ is the reference time for discounting (typically one year). The outcome of the optimization procedure is a vector $\mathbf{u}$ containing the settings of the control variables over the producing life of the reservoir. Note that, although the optimization is based on N models, only a single strategy $\mathbf{u}$ is obtained. Typical elements of $\mathbf{u}$ are monthly or quarterly settings of well head pressures, water injection rates, valve openings etc.

Model selection

We use multiple ensembles of realizations to account for geological uncertainties. Typical ensembles are formed by tens or hundreds of realizations, turning the procedures involved into computationally intensive applications. The cost, in terms of the amount of simulations required, of robust optimization and history matching algorithms tends to scale linearly with the size of the ensemble (i.e., O(N)), while the VOI assessment workflow described above scales with the square of the number of realizations (i.e., O(N^2)). Thus, a decrease in the number of realizations considered in the analysis may lead to significant reduction in the computational cost and make the VOI assessment problem more tractable. The challenge is on how to cleverly select a subset of realizations which can represent the full ensemble concerning the ability of quantifying the uncertainty. Others have worked on this problem; see, e.g., Armstrong *et al*. (2013) and Sarma *et al*. (2013). This work focusses on the use of clustering techniques to automate the selection of representative model realizations.

Clustering

Cluster analysis aims to group a set of N objects

$$\mathbf{\Theta} = \begin{bmatrix} \boldsymbol{\theta}_1 & \boldsymbol{\theta}_2 & \cdots & \boldsymbol{\theta}_N \end{bmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1N} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{M1} & \theta_{M2} & \cdots & \theta_{MN} \end{bmatrix} , \tag{3}$$

into N_{repr} clusters according to the similarity between objects given; see, e.g., Baker (2015). Note that here the objects have been chosen as vectors $\boldsymbol{\theta}_i$, i = 1,2,…N, in an M-dimensional space (e.g., N realizations of M grid block permeability values) but they could also be scalars, matrices or higher-order objects (tensors). Clustering has been widely used in pattern recognition, machine learning and also in statistics (Arabie and Hubert, 1996) and is broadly classified into partitional and hierarchical categories. As the name suggests, partitional clustering separates the objects into exclusive clusters such that the objects within a cluster are more similar than the objects in another cluster. On the other hand, hierarchical clustering, also known as connectivity-based clustering, connects objects to form clusters based on their distance. The connected objects in clusters can then be represented using a dendrogram.

K-means clustering is one of the most used partitional clustering methods due to its simplicity (Caers, 2011). The user predefines the number N_{repr} of sets C_j, j = 1,2,…N_{repr}, which contain the N_j indices of the objects belonging to each cluster, where the clusters are not necessarily of equal size. The algorithm then attempts to iteratively improve the partitioning to achieve the lowest intra-cluster distance according to

$$C_{opt} = \arg\min_{C} \sum_{j=1}^{N_{repr}} \sum_{k \in C_j} d_{jk} \left(\boldsymbol{\theta}_k, \overline{\boldsymbol{\theta}}_j \right)^2 \quad , \tag{4}$$

where C is the set of N_{repr} clusters, i.e. a set of sets of indices, and $d_{jk}(\boldsymbol{\theta}_k, \overline{\boldsymbol{\theta}}_j)$ is the distance between one of the N_j data points within each cluster and the cluster centroid $\overline{\boldsymbol{\theta}}_j$ computed as

$$\overline{\boldsymbol{\theta}}_j = \frac{1}{N_j} \sum_{k \in C_j} \boldsymbol{\theta}_k \quad . \tag{5}$$

The first step to use cluster analysis consists of choosing a feature operator F to compare the models, whereby the goal is to distinguish or correlate the given samples (i.e., the full ensemble $\mathbf{M}$ of model realizations). The feature operator could just select a number of parameters (e.g. grid block permeability values) of the vector of model parameters $\mathbf{m}$, or it could represent a more complex operation like a full simulation to compute the NPV or a sequence of saturation snapshots. Using this operator, the set $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N]$ is formed, where $\boldsymbol{\theta}_i = F(\mathbf{m}_i)$. The clustering algorithm can then generate the distances required to determine C_{opt}. It has been shown that the choice of the appropriate feature is extremely case-dependent; thus, there is no one-fits-all solution.

Note that the sets $\boldsymbol{\theta}_i$ are elements of an M-dimensional space, where M can be very large. Unfortunately, most clustering algorithms do not work efficiently in higher dimensional spaces because of the inherent sparsity of the data, and as M grows, distance measures become increasingly meaningless (Keim *et al*., 1997; Parsons *et al*., 2004). A solution to this problem is to eliminate some of the dimensions of the feature space. However, if done wrongly, this may cause information loss and introduce wrong correlations to the samples. Aggarwal *et al*. (1999) have shown that the projection of high-dimensional data spaces on subspaces leads to improved clustering results.

Projection methods

There is more than one method to project datasets into a reduced-order space. (Note that this is sometimes referred to as reducing the dimensionality of the "feature space".) One of them involves the use of tensor decomposition techniques. Tensor decomposition is strongly related to principal component analysis (PCA) or singular value decomposition (SVD). It enables the transformation of data into a compact representation while honoring their structure (e.g., spatial/temporal correlations) which is usually degraded with the vectorization step in SVD approaches. These techniques can be used to compress large datasets stored as tensors by constructing low-rank approximations with minimal approximation or reconstruction error. For instance we may form a dataset $\boldsymbol{\Theta}$ in the form of a tensor representation of the data, $\boldsymbol{\Theta} = F(\mathbf{m}_1, \mathbf{m}_{2,\ldots}\ \mathbf{m}_N)$, which better preserves their structure than using the vector representation described above. E.g., we could construct a 3D tensor $\boldsymbol{\Theta}$ by stacking up the two-dimensional permeability fields (matrices) of an ensemble of 2D model realizations. We then perform the following minimization (Insuasty *et al*., 2015)

$$\min_{\varphi_{1:I}, \psi_{1:J}, \chi_{1:K}} \left\| \boldsymbol{\Theta} - \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \sigma_{ijk} \left(\boldsymbol{\varphi}_i \otimes \boldsymbol{\psi}_j \otimes \boldsymbol{\chi}_k \right) \right\|_F \tag{6}$$

$$\text{subject to } \boldsymbol{\varphi}_i^T \boldsymbol{\varphi}_{i''} = \delta_{i'i''}, \boldsymbol{\psi}_{j'}^T \boldsymbol{\psi}_{j''} = \delta_{j'j''}, \boldsymbol{\chi}_{k'}^T \boldsymbol{\chi}_{k''} = \delta_{k'k''},$$

where $\boldsymbol{\varphi}_i$, $\boldsymbol{\psi}_j$, $\boldsymbol{\chi}_k$ are orthonormal basis functions, $\|.\|_F$ represents the Frobenius norm, and the symbol $\otimes$ denotes the tensor product over a vector space. This can be extended for tensors with more dimensions. For more information, we refer to Insuasty *et al*. (2015), who also show that the solution for tensor decomposition in (6) can be approximated by performing high-order SVD (HOSVD). In this case, the tensor is flattened in a planar matrix structure where we can operate similarly to classical SVD. This allows us to determine the basis functions and the coefficients associated with them. Like in classical SVD, a truncation can then be applied to retain only the basis functions which explain the most dominant patterns in the data, thus resulting in the lower-dimensional representation we were aiming for. One of the dimensions of $\boldsymbol{\Theta}$ in our applications typically refers to the model uncertainty, characterized by the N model realizations. We can use the HOSVD coefficients associated with this dimension to perform clustering. Insuasty *et al*. (2015) show that this approach allows us to compare model realizations based on very rich datasets, such as the temporal evolution of the spatial

distribution of pressures and saturations inside the reservoir. They are able to select a subset of realizations representative in terms of dynamic flow patterns and form reduced ensembles to perform robust production optimization more efficiently.

Another tool to represent samples in a lower dimension is multidimensional scaling (MDS). It refers to techniques that use distance measures to produce a $\tilde{\boldsymbol{\theta}}_i$ representation of data points $\boldsymbol{\theta}_i$ in a reduced N_{MDS}-dimensional space with $N_{MDS} << M$. MDS was first introduced for the analysis of proximity data. In recent years, the machine learning community has applied MDS for nonlinear dimension reduction. Kruskal (1976) argues that MDS can be complementary to clustering techniques. Scheidt and Caers (2009) introduced MDS in the reservoir simulation community, and since then many reservoir applications have been documented; see, e.g., Caers (2011).

When applying MDS, a measure of fit, referred to as 'stress' in the MDS literature, can be calculated to quantify the conformance of the representation $\tilde{\boldsymbol{\Theta}}$ to the original data $\boldsymbol{\Theta}$, which can be a criterion to define the number N_{MDS} of dimensions of the reduced space; see Kruskal (1964). Low values of stress (i.e., below 5%) indicate an excellent fit between dissimilarities and distances, thus a good representation of the original samples in the reduced dimension space. We can then use this stress to determine the appropriate number of dimensions N_{MDS} to proceed with: we start with a small number of dimensions and we increase this number until it drops below an acceptable level (e.g., 5%). For further information on dimensionality reduction, including MDS and PCA approaches we refer to Cunningham *et al*. (2015).

Methodology

As discussed in the previous section, the VOI assessment workflow proposed by Barros *et al*. (2016) and depicted in Figure 1 requires an excessive amount of simulations to be applied in practice. This is mainly due to the extensive use of robust optimization and history matching and to the fact that multiple plausible truths are considered. The most demanding steps constitute opportunities for considerable acceleration of the workflow. The focus of this work is on the use of model selection to achieve this goal. Thus, it is about looking for approximated results by compromising the rigor in UQ for the sake of computation speed-up.

In this section, we describe first how to select representative models to speed-up robust optimizations and how to assess the quality of the results with the accelerated procedure. Afterwards, we explain how we can accelerate the VOI analysis by picking representative plausible truths. We also discuss the choice of the most appropriate feature to distinguish model realizations in the different parts of the workflow. Finally, we combine both measures to come up with a new and faster VOI assessment workflow.

Speeding-up robust optimization

The whole idea behind accelerating robust optimization is to reduce the number of reservoir simulations required. This is done by reducing the number of model realizations in the ensemble used in the optimization.

We start with a full ensemble $\mathbf{M}_{full}$ of N model realizations. The first step is deciding the number N_{repr} of representative realizations to form the reduced ensemble $\mathbf{M}_{repr}$. This number should reflect the speed-up factor we would like to achieve or the maximum ensemble size we can afford to use with the available computational resources.

The second step is choosing a feature F relevant to the problem to be optimized and building our dataset to apply the clustering algorithms. In our case, we are using reservoir simulation to assist the optimization of the water flooding process. Therefore, it seems to be important to distinguish the model realizations regarding their simulated dynamical behavior. One option is to rely on the fact that model parameters (e.g., permeabilities and porosities) tend to correlate with the reservoir flow characteristics and use them as the feature to distinguish the realizations $\mathbf{m}_i$. The alternative is to generate simulated data and work with features associated with the dynamics of the system. In this case we can consider relying on model states or flow patterns (e.g., pressure/saturation snapshots and streamlines). Or we can look at the model outputs (e.g., well production data and NPV evolution over time), which are closely related to the objective function to be optimized.

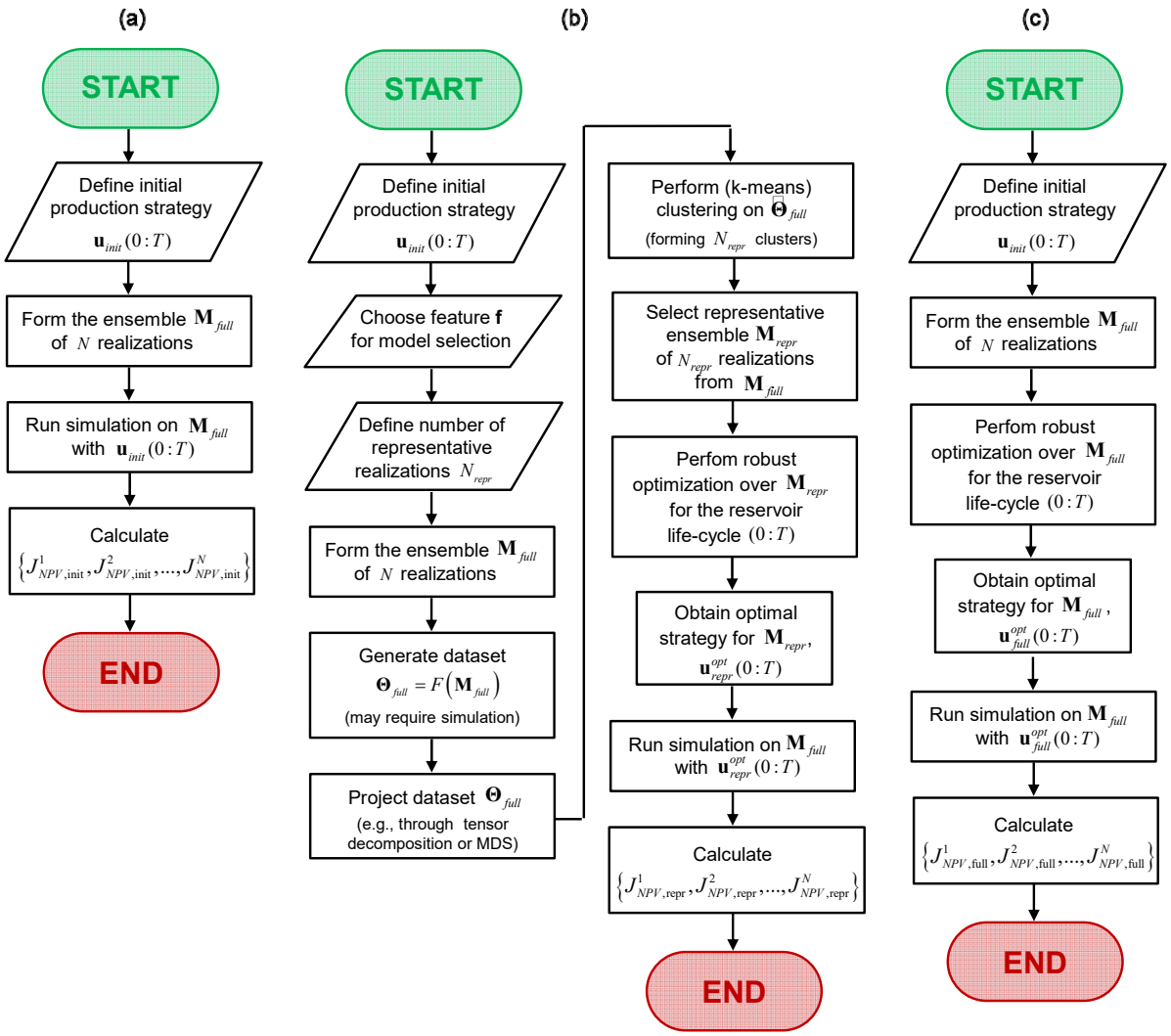

Figure 2 *Workflow to evaluate the use of representative realizations for efficient robust optimization. (a) Computation of objective function values for unoptimized production, (b) robust optimization using representative realizations and (c) robust optimization using the full ensemble (reference).*

Representative plausible truths

The selection of fewer plausible truths for the VOI analysis can help reducing the computational cost of the workflow at a different level than the acceleration of robust optimization. The plausible truths are model realizations which we pick to play the role of truth in the CLRM framework. Thus, the goal remains the same: to select representative model realizations.

The challenge is to find relevant features to distinguish these realizations considering their role in the workflow. Although we are still interested in the reservoir management problem (e.g., water flooding process in our case), the plausible truths are not directly involved in the optimization procedure; we perform the optimizations on the realizations of the prior and posterior ensembles. Due to this difference in roles, the features which are relevant to select representative realizations for the robust optimization are most likely not appropriate to distinguish plausible truths. As we mentioned before, literature suggests there is no one-fits-all solution for choosing the selection features and, therefore, we look for fit-for-purpose solutions.

The plausible truths are related to the optimal reservoir management that we seek with the application of the CLRM framework. Moreover, we use them to generate synthetic data for the future measurements analyzed. Thus, the plausible truths represent the behavior we intend to track by doing history matching in our ensembles of models and to control by doing optimization. This suggests that we should look for features able to distinguish the optimal configurations related to each of the

plausible truths if we are to select a representative subset of them. We refer to these as *optimization features*, opposed to the *model features* considered in the section *Speeding-up robust optimization* (i.e., model parameters, model states and model outputs).

The best way of identifying characteristics of the optimal configurations related to each of the plausible truths is to perform separate nominal optimizations on them. If we consider an initial ensemble of N plausible truths, this means we need to perform N optimizations. As a result, we obtain a set of N optimal production strategies and N optimal objective function values. The optimal production strategies tend to be very different because the plausible truths are different model realizations. However, the optimal objective function values tend to be close to each other because the different optimal production strategies compensate for the differences in the model realizations.

Since the goal is to distinguish the plausible truths, the optimal production strategies seem to be appropriate to support the clustering. The main potential problem of this approach is the risk of non-uniqueness of optimal solutions for the production optimization problem, due to the possible presence of redundant degrees of freedom in the high-dimensional space of control variables (van Essen *et al*., 2009). This may result in multiple production strategies that are equally optimal, which could put the validity of this approach at stake. A possible way to avoid the redundancy is to perform an a-priori tensor- or SVD-based decomposition of a large set of possible controls and perform the optimization in a reduced control space. Alternatively, one could impose temporal and/or spatial correlations on the controls which also reduces the degrees of freedom in the control space. A downside of such a-priori measures is that they may lead to lower NPV values. As an alternative, we therefore apply an a-posteriori tensor decomposition of the set of optimal production strategies and retain a fraction of the basis functions by truncation based on their energy. By doing so, we intend to capture only the main trends of the data and reduce the effects of the non-uniqueness of the optimal production strategies, although we note that this a-posteriori decomposition of the controls does not guarantee an improvement of the situation.

The optimal objective function values by themselves are less suitable to help in the selection of representative plausible truths. However, in combination with the objective function values obtained with a robust strategy (i.e., optimized to maximize the mean objective function of the initial ensemble), these data reveal how much we may benefit if we learn or observe the truth for each one of the plausible truths. We can then distinguish plausible truths according to the gains associated with their optimal configurations, and this can be useful for our purposes. One of the advantages of using these data as features is avoiding the problem of non-unique optimal solutions discussed in the previous paragraph.

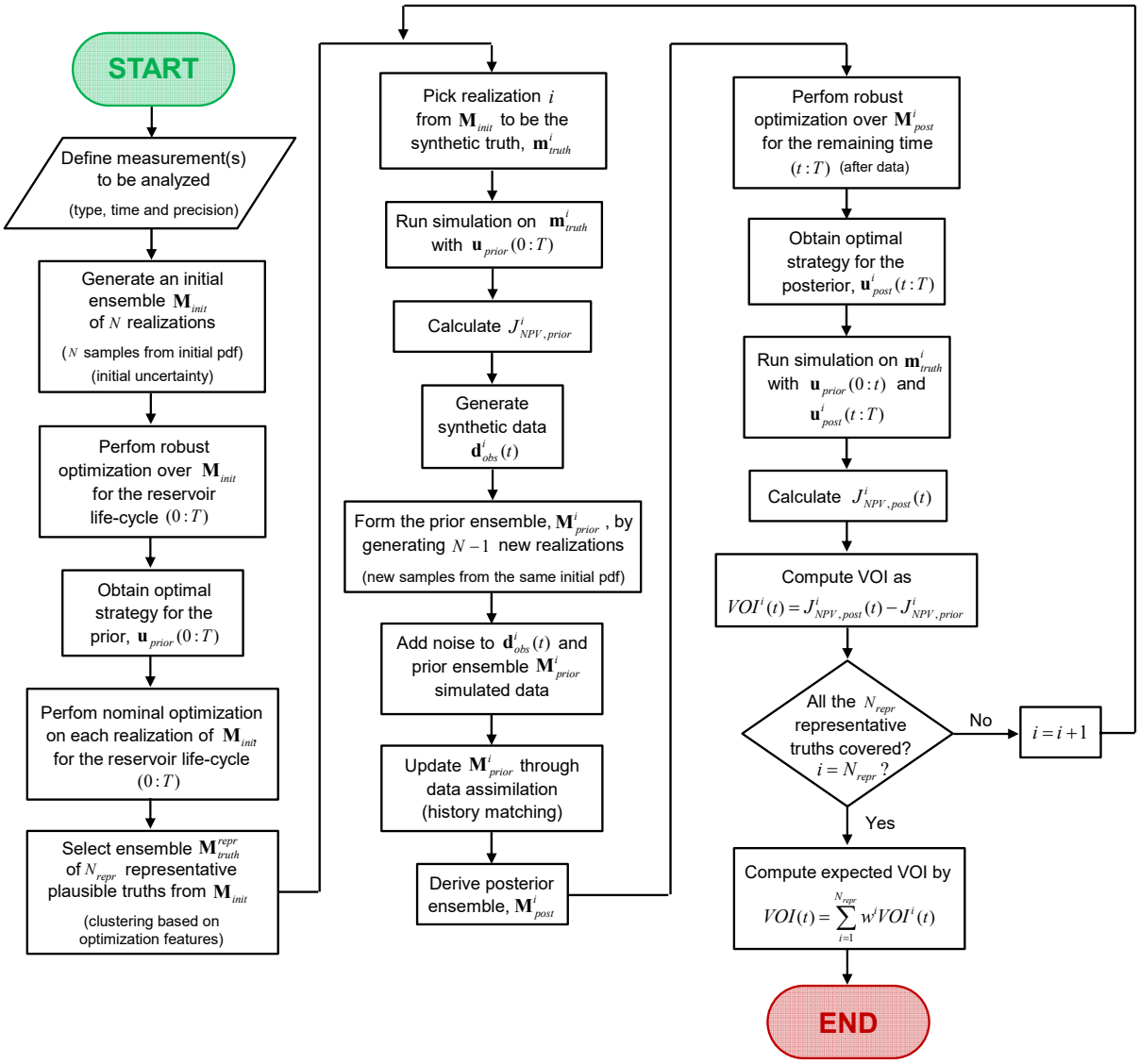

Figure 3 *Workflow to compute the expected VOI using representative plausible truths.*

We introduce the selection of representative plausible truths to the original VOI assessment workflow and we obtain the procedure depicted in Figure 3. The main difference is that, before entering the loop where each one of the N realizations of the initial ensemble $\mathbf{M}_{init}$ is picked to be the truth $\mathbf{m}^i_{truth}$, we have a few more pre-processing steps. First, a step where we optimize each one of the realizations and then the clustering to select N_{repr} representative plausible truths based on the optimization features as explained above. Another minor change in the workflow refers to the computation of the statistics of VOI: before, the plausible truths were considered (for simplicity) to be equiprobable, but, now, the selected plausible truths in $\mathbf{M}^{repr}_{init}$ may have different weights w_i assigned by our selection procedure. Note that there is a computational cost associated with the additional N nominal optimizations required, but that this extra cost is minor when compared to the cost of the full workflow. Another point to realize is that these N nominal optimizations would be performed anyway if we carry out a value of clairvoyance (VOC; see Barros *et al.*, 2016) to determine the upper bound for the VOI analysis.

Accelerated VOI assessment

We combine the ideas of the two previous sections to the original VOI workflow of Figure 1. The result is a new workflow for accelerated VOI assessment, shown in Figure 4.

We note that, like in Barros *et al.* (2016), in this flowchart we consider the size of the initial ensemble and the prior ensembles to be N and N-1. But, in fact, they do not have to be related. The same holds

for the number of representative realizations for robust optimization and the number of representative plausible truths, which here are considered both to be N_{repr}.

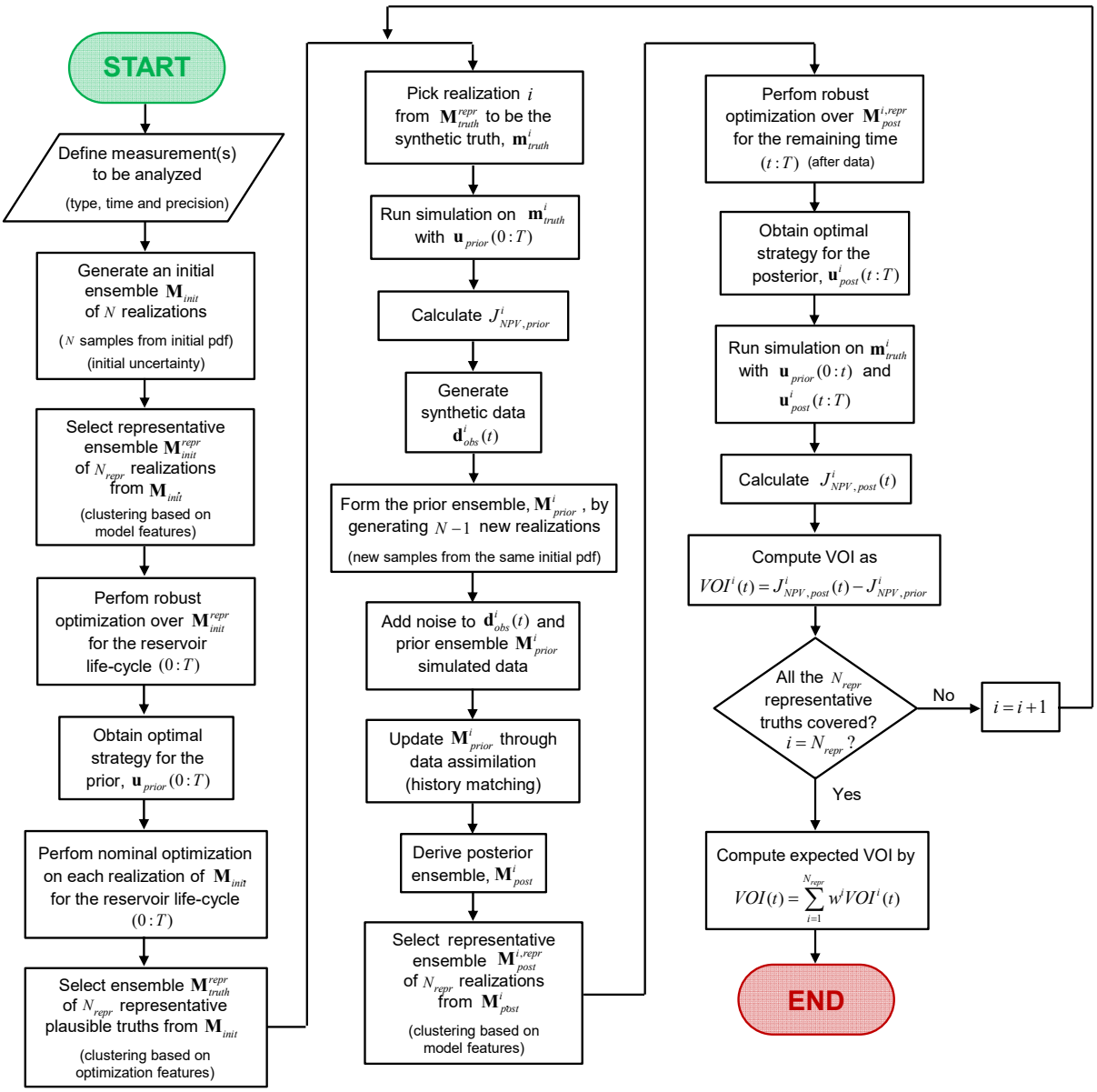

Figure 4 *Workflow to compute the expected VOI using representative plausible truths and reduced ensembles of representative model realizations in the robust optimizations.*

Another point to highlight is that the history matching is carried out on the full prior ensembles. Obviously, representative model selection can also be applied in this step. However, our current implementation of the workflow relies on ensemble based history matching techniques, which require a certain minimal amount of samples to function.

Given these considerations, we can expect a speed-up factor of the order of $O((N/N_{repr})^2)$ by using this accelerated procedure. This means that, if we are able to select reduced ensembles with 10 times fewer realizations, we can reduce the number of required reservoir simulations by a factor of 100.

Examples

Efficient robust optimization

2D models

As a first step to illustrate our approach, we used two different variations of a simple two-dimensional reservoir simulation model with an inverted five-spot well pattern. The first variation follows the exact same configuration with heterogeneous permeability and porosity fields as the example in our previous paper. See Barros *et al*. (2016) for a more detailed description. We refer to it as the 2D five-spot model. The second example has the same settings, but with different permeability and porosity fields. For this case, we used the SNESIM algorithm (Strebelle, 2002) available in the geostatistical modeling software SGeMS to generate ensembles of channelized realizations, based on the training image used for modeling layer 3 of the Stanford VI (Castro *et al*., 2015) reservoir model. The realizations are constrained to hard well data. We refer to this second example as the 2D channelized model.

In the two examples, we have original ensembles of $N = 50$ realizations to test the performance of robust optimization over reduced ensembles. Figure 5 shows a few realizations of the permeability fields of both examples. The optimization was run for a 1,500-day time horizon with well controls updated every 150 days, i.e. $M = 10$, and, with five wells, $\mathbf{u}$ had 50 elements. We applied bound constraints to the optimization variables (200 bar $\leq p_{prod} \leq$ 300 bar and 300 bar $\leq p_{inj} \leq$ 500 bar). The initial control values were chosen as mid in-between the upper and lower bounds. The whole exercise was performed in the open-source reservoir simulator MRST (Lie *et al*., 2012), by modifying the adjoint-based optimization module to allow for robust optimization. Further details concerning the optimization parameters can be found in Barros *et al*. (2016).

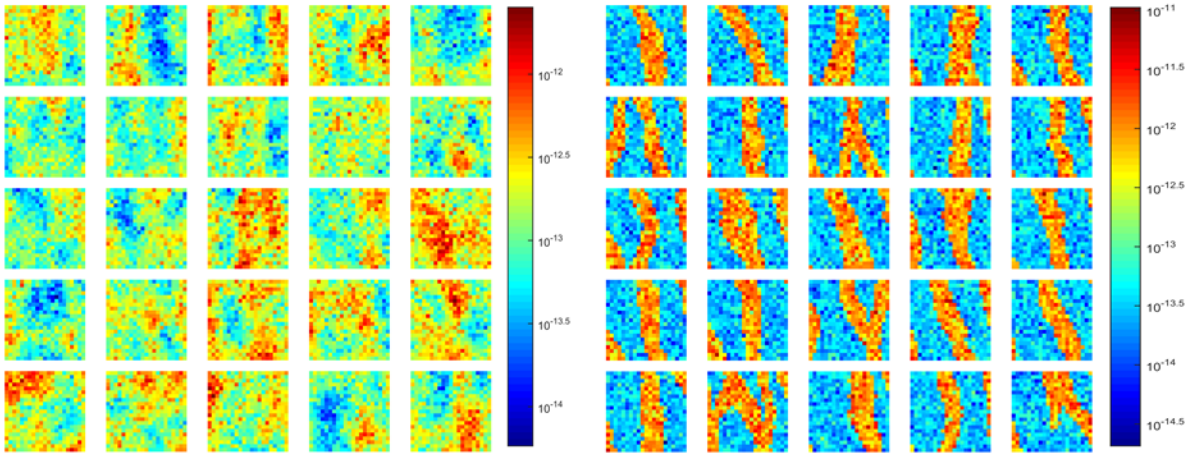

Figure 5 *25 randomly chosen realizations of the permeability field of the 2D five-spot model (left); 25 realizations of the 2D channelized model (right).*

We considered different numbers of representative realizations N_{repr}: 3, 5 and 10, representing approximately 5%, 10% and 20% of the number of realizations in the full ensembles ($N = 50$). We also evaluated the performance using different features for clustering: permeability field, oil saturation snapshots (every 150 days) and NPV time series. The effect of projection methods was evaluated by using both MDS and tensor decomposition to perform the projection of the feature space before clustering. For MDS we used the implementation available in Matlab, and for the tensor decomposition we used the implementation of the HOSVD described in Insuasty *et al*. (2015).

By applying the steps as in Figure 2, we obtained results in terms of NPV. Figures 6 and 7 depict the results for the two examples in cumulative distribution function (CDF) plots. The NPV values for the unoptimized production (Figure 2a) are shown in grey and the reference results (Figure 2c) in black. The results obtained using the reduced ensembles (Figure 2b) are represented by the colored lines according to the number of representative realizations. For both examples, we repeated the procedure with several ensembles $\mathbf{M}_{full}$ and we obtained similar results as the ones shown here.

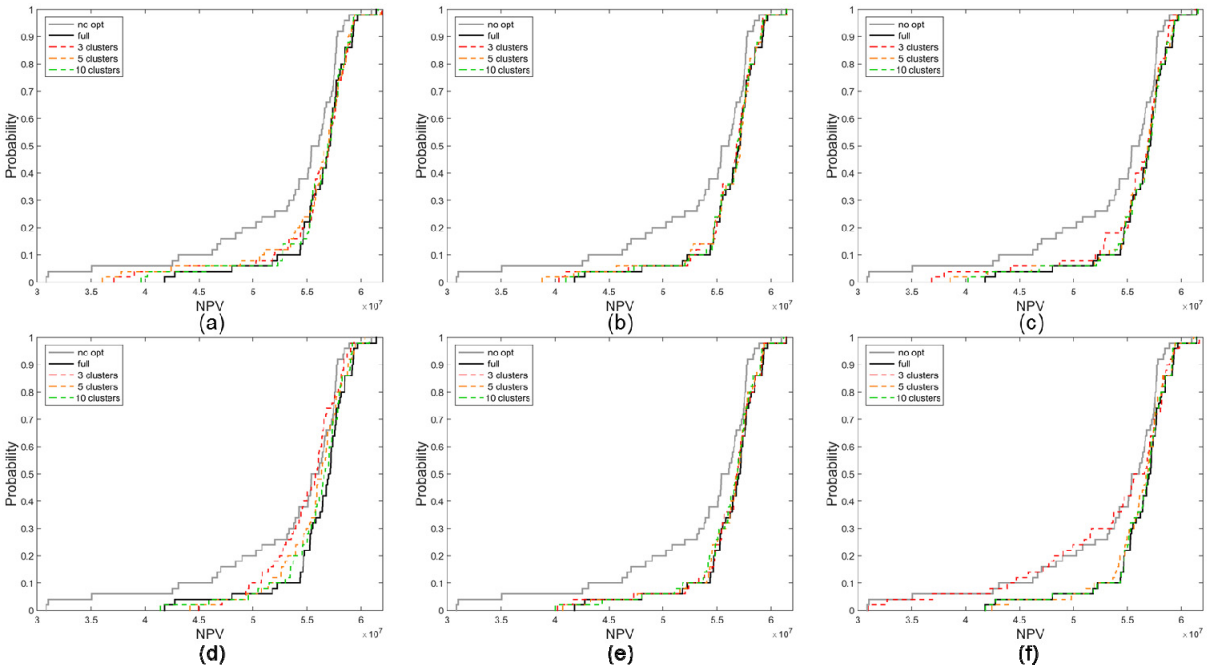

Figure 6 *NPV CDF plot of an ensemble from the 2D five-spot model. (a) Permeability as feature and MDS as projection method, (b) NPV time series as feature and MDS as projection method, (c) oil saturation snapshots as feature and MDS as projection method, (d) Permeability as feature and tensor decomposition as projection method, (e) NPV time series as feature and tensor decomposition as projection method, and (f) oil saturation snapshots as feature and tensor decomposition as projection method.*

Generally, all the production strategies $\mathbf{u}_{repr}^{opt}$ optimized with the representative ensemble performed very well when compared with the strategy $\mathbf{u}_{full}^{opt}$ for the full ensemble. We notice that the selection with three representative realizations (5%) from the ensemble performs poorer in most cases, and fails badly in the case depicted in Figure 6f. This shows that, for this example, taking only 5% representative realizations from the total ensemble is insufficient whereas taking 10% and 20% give good results in all cases. Also, we notice that the MDS transformation helps the selection of representative realizations better than the tensor decomposition, especially when using the permeability field as feature (Figures 6a and d). Overall, oil saturation snapshots and NPV time series seem to be the most suitable selection features based on the results from 2D five-spot model.

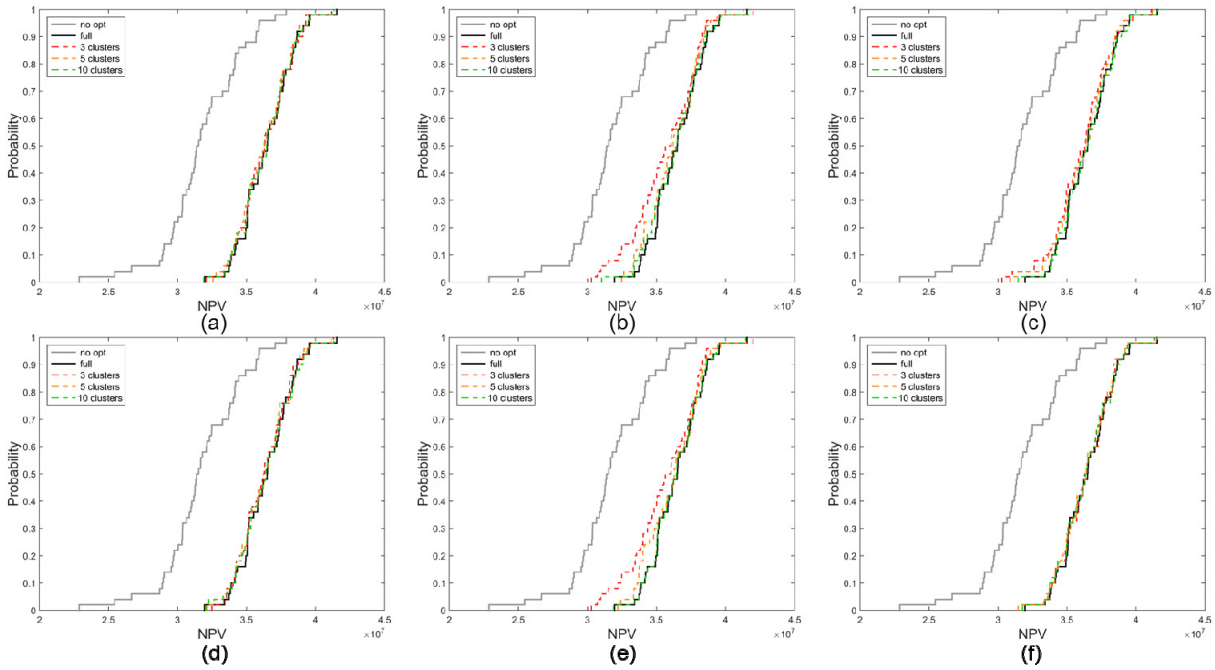

***Figure** 7 Final NPV CDF plot of an ensemble from the 2D channelized model. (a) Permeability as feature and MDS as projection method, (b) NPV time series as feature and MDS as projection method, (c) oil saturation snapshots as feature and MDS as projection method, (d) Permeability as feature and tensor decomposition as projection method, (e) NPV time series as feature and tensor decomposition as projection method, and (f) oil saturation snapshots as feature and tensor decomposition and projection method.*

For the 2D channelized model, the realizations selected with both MDS and tensor decomposition projections perform really well and are rather indistinguishable in terms of quality of approximation. However, from the features used, we observe that NPV time series has a poorer performance compared to the others used as seen in Figures 7b and e. Once again, we see that cases with three representative realizations perform noticeably worse than the rest.

Accelerating VOI assessment

2D five-spot model

To illustrate the selection of representative plausible truths, we applied the workflow in Figure 3 to the 2D five-spot model. The workflow was repeated for different observation times, t_{data} = {150, 300, … , 1350} days. The history matching step was performed with the EnKF module of MRST. We assessed the VOI of the production data (total flow rates and water-cuts) with absolute measurement errors (ε_{flux} = 5 m³/day and ε_{wct} = 0.1). The VOC and the VOI were computed for each of the nine observation times, and we compared the results against those obtained using the original workflow (Figure 1) with the full ensemble $\mathbf{M}_{init}$ of plausible truths (N = 50), which serve as reference.

First, we checked our hypothesis that the *model features* are not suitable for selecting plausible truths. For that, we selected N_{repr} = 5 plausible truths through clustering based on: permeability field, NPV evolution profile and flow patterns (i.e., pressure and saturation snapshots). No projection methods (e.g., MDS) were used. We also tested a random selection. The results are shown in Figure 8, including the reference results. The different lines represent percentiles and mean of the VOC and VOI distribution as a function of the time when the additional information (or clairvoyance) becomes available. We do not interpret or explain the VOI and VOC results here; we refer to Barros *et al.* (2016) for this purpose. Here we assess the ability to obtain similar results with fewer plausible truths.

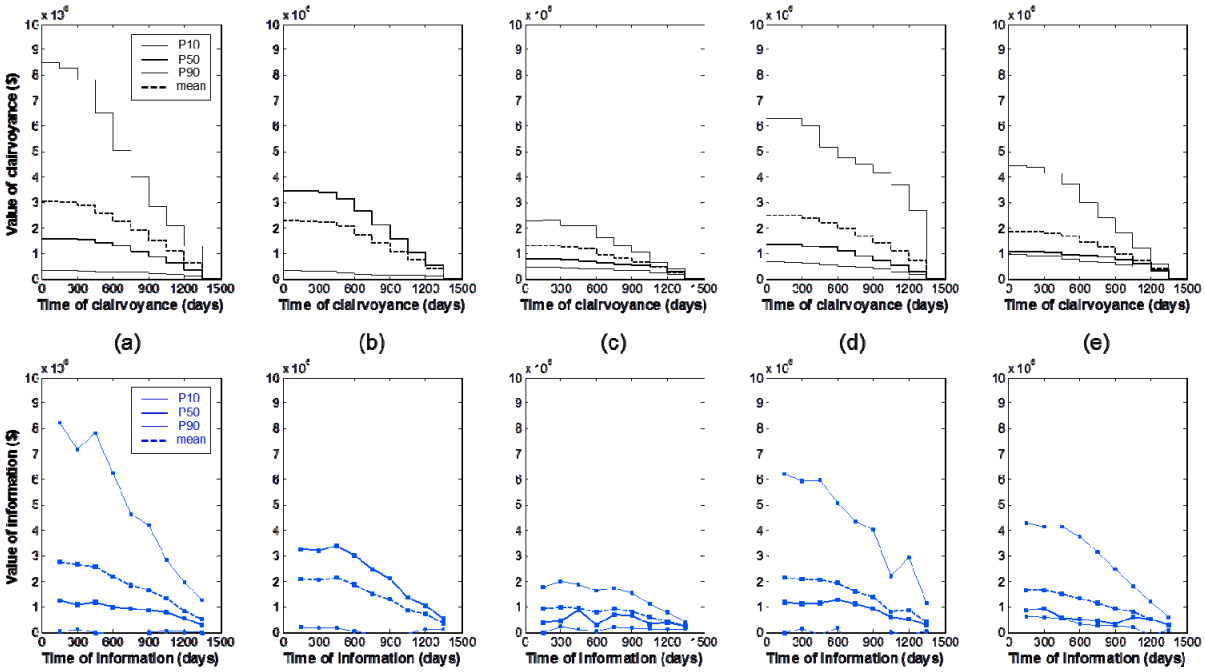

Figure 8 *Results of the VOI assessment for the 2D five-spot model using selection of 5 plausible truths based on model features. (a) Reference obtained using the original workflow, (b) random selection, (c) selection based on permeability field, (d) selection based on NPV time series and (e) selection based on flow patterns.*

Overall, none of the selections in Figure 8 is able to satisfactorily reproduce the reference results, shown in Figure 8a. Although a selection based on some feature (Figures 8c, d and e) is clearly better than a random selection (Figure 8b), our hypothesis that the *model features* are not the most appropriate means to select representative plausible truths seems to be correct.

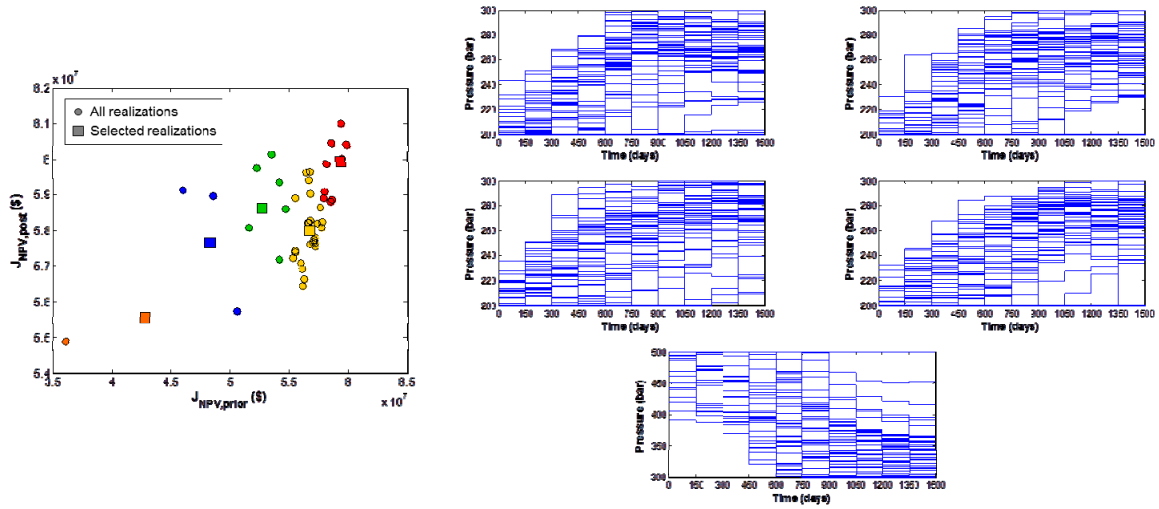

Figure 9 *Proposed optimization features for the selection of representative plausible truths. Objective function values, before and after nominal optimizations (left). Optimal production strategies (bottom-hole pressures at all the wells for every control interval) (right).*

Next, we repeated the same procedure with our proposed *optimization features*. This required a nominal optimization on each of the $N = 50$ plausible truths initially considered. Figure 9 shows the data we obtain from these optimizations, which can be used for clustering and model selection. It becomes clear that these features create a space in which we can distinguish the samples and select those $\mathbf{M}_{init}^{repr}$ that can better represent the entire population $\mathbf{M}_{init}$. Figure 9 (left) displays each of the plausible truths plotted in a two-dimensional space: the first dimension corresponds to the objective

function values (i.e., NPV) before the nominal optimizations and the second to the objective function values after the optimizations. The clusters are shown in different colors and the selected plausible truths are marked as squares. Figure 9 (right) exhibits the optimal production strategy for each plausible truth: the plots show the BHP controls at each of the five wells of the 2D five-spot model.

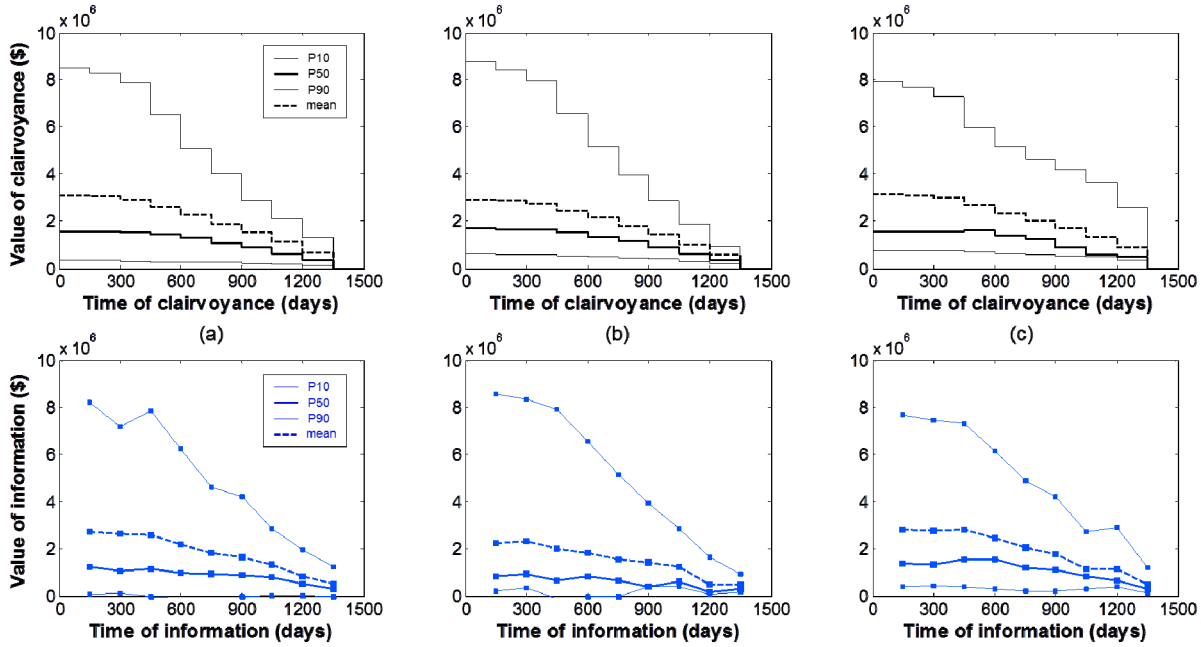

Figure 10 *Results of the VOI assessment for the 2D five-spot model using selection of 5 plausible truths based on optimization features. (a) Reference obtained using the original workflow, (b) selection based on objective function values and (c) selection based on optimal production strategies.*

Figure 10 depicts the results obtained by picking N_{repr} = 5 representative plausible truths according to these new features. This time we observe better selections, which are able to repeat the reference results (Figure 10a) almost perfectly. We note that the selection based on the objective function values (Figure 10b) succeeds in reproducing the reference results for VOC, but less for VOI. In contrast, the selection based on optimal production strategies (Figure 10c) performs fairly well for both VOC and VOI. Therefore, the optimal production strategy is the feature we chose to select representative plausible truths.

After that, we investigated the impact of the non-uniqueness of optimal production strategies in the model selection. For that, we carried out nominal optimizations on each of the plausible truths starting from three different initial solutions. Figure 11 shows the results: Figures 11b, c and d correspond to the selection obtained with the three different starting points and Figure 11e to the selection based on the data of the three optimizations altogether. We observe that the results are not the same, which confirms that the optimal production strategies may be not unique and, thus, not suitable for model selection purposes.

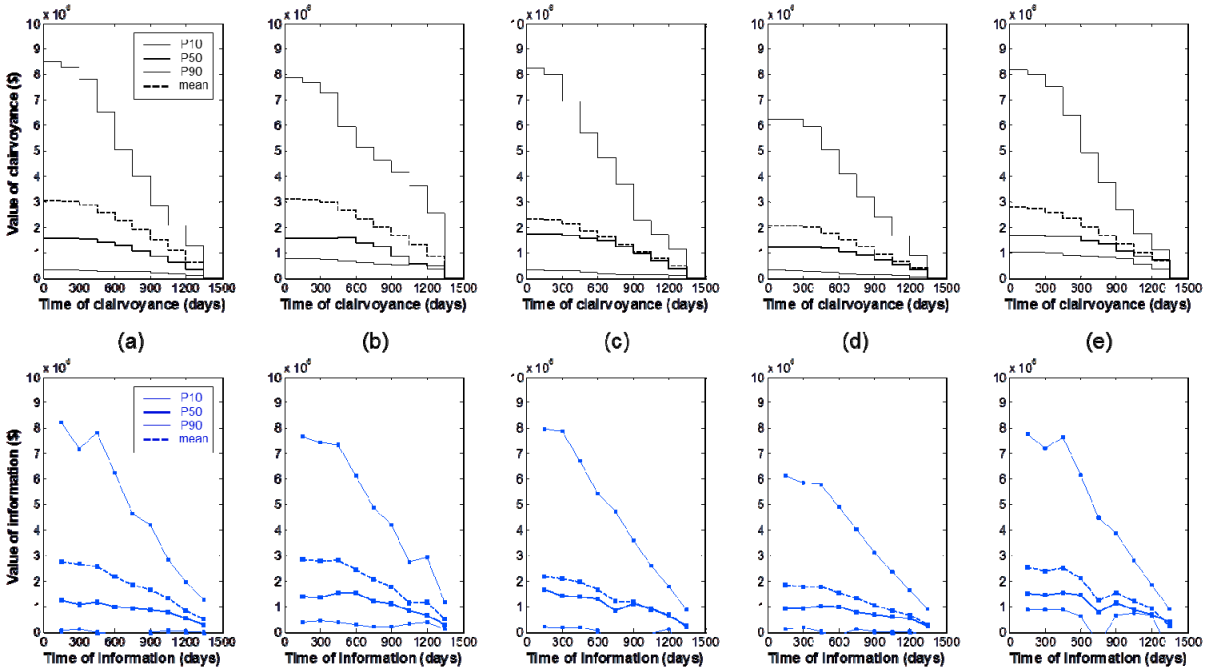

Figure 11 *Results of the VOI assessment for the 2D five-spot model selecting representative plausible truths based on optimal production strategies obtained with different starting solutions. (a) Reference results, (b) starting from robust optimal solution, (c) starting from greedy controls (maximum injection and maximum drawdown in the producers), (d) starting from mid in-between bounds and (e) selection based on data from all the three optimizations.*

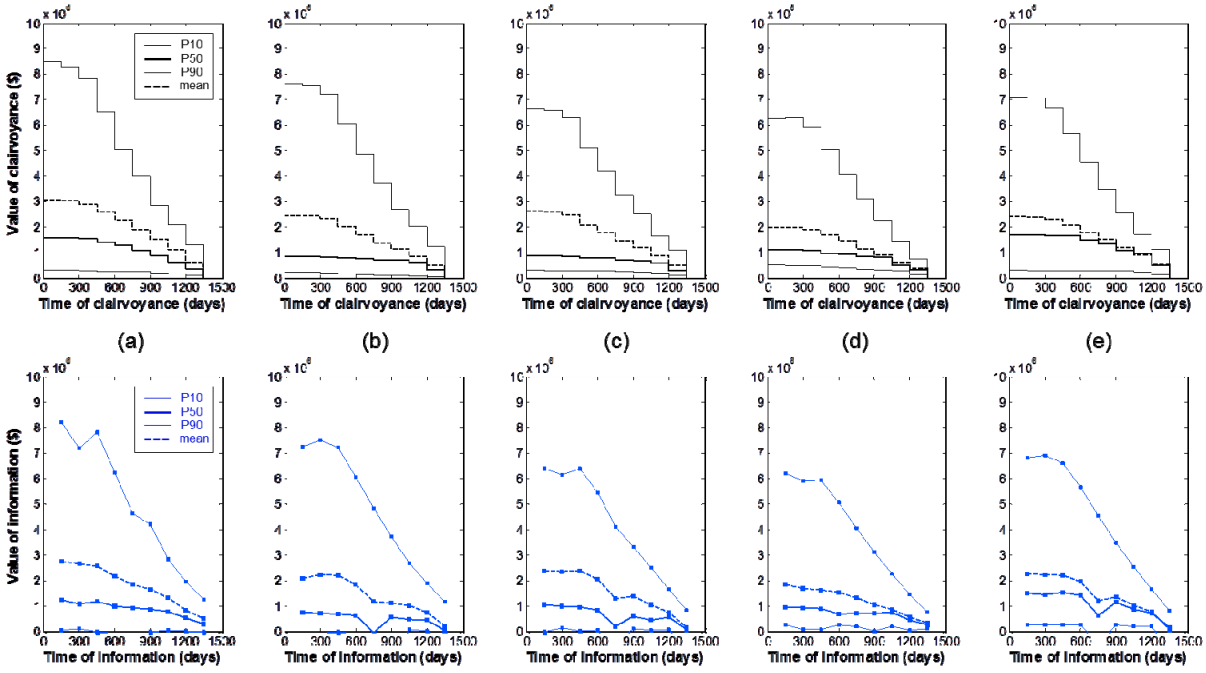

Figure 12 *Results of the VOI assessment for the 2D five-spot model selecting representative plausible truths based on optimal production strategies projected by tensor decomposition. (a) Reference results, (b) starting from robust optimal solution, (c) starting from greedy controls (maximum injection and maximum drawdown in the producers), (d) starting from mid in-between bounds and (e) selection based on data from all the three optimizations.*

As a measure to remediate this problem, we applied a projection based on truncated tensor decomposition to the optimal production strategies dataset. By retaining the fraction of the basis functions which preserves 90% of the energy of the singular values, we hope to capture only the main

trends and reduce the effect of non-uniqueness of the optimal solutions. Figure 12 shows the results obtained with such a projection for the same optimal production strategies. The difference between the results for the three different optimization starting points is smaller than in Figure 12.

Finally, we applied all the measures we discussed so far following the accelerated procedure for VOI assessment depicted in Figure 4. Figure 13 presents the results. Again, Figure 13a exhibits the reference results with N = 50 plausible truths and robust optimization over ensembles of N = 50 realizations. Figure 13b corresponds to the results obtained by using N_{repr} = 5 representative plausible truths and full ensembles for the robust optimizations. Figure 13c shows the results when considering all the plausible truths and reduced ensembles for robust optimizations. And Figure 13d displays the results obtained with N_{repr} = 5 plausible truths and ensembles of N_{repr} = 5 realizations for the optimizations.

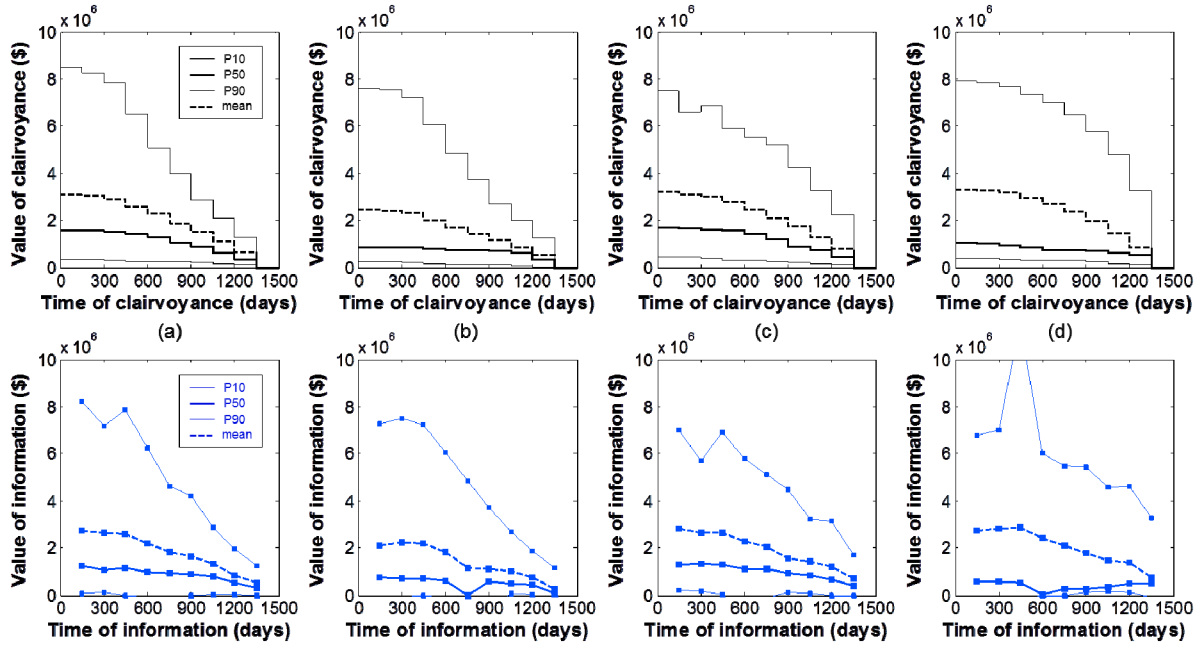

Figure 13 *Results of the new accelerated VOI assessment for the 2D five-spot model. (a) Reference results, (b) only selecting representative plausible truths, (c) only reducing the ensembles for robust optimization, and (d) combining both measures.*

We see that the acceleration measures allow us to obtain similar results by considering only 10% of the original number of realizations. We also observe that even the combination of the two acceleration measures described in this paper is still able to correctly approximate the main trend of the reference results. Note that the lines plotted in Figures 13b and d represent percentiles and mean values of VOI based on N_{repr} = 5 samples, while in Figures 13a and c these values are computed with N = 50 samples, and this should be taken into account when interpreting the quality of the results.

In terms of computational cost, the results in Figure 13a require approximately 1.5 million simulations. This number includes forward and backward simulations. The results shown in Figure 13b and c require 150,000 and 170,000 simulations, respectively. And the results in Figure 13d need 17,000 simulations to be computed. Thus, by applying all measures to reduce the number of model realizations considered in the assessment, we were able to alleviate the computational cost of the workflow by a factor of 88, which is quite significant.

Conclusion

We applied two different measures to decrease the amount of simulations required in VOI assessment workflows. First, we showed how to make the robust optimizations more efficient by reducing the size of the ensembles considered. We used several model features to select representative realizations through clustering and compared them against reference results. For the 2D examples shown in this paper, we concluded that oil saturation snapshots data transformed by MDS provide a good basis for the selection of reduced ensembles for robust optimization. Second, we introduced two new features

to support the selection of representative plausible truths as another alternative to speed-up the VOI assessment procedure. We confirmed that optimal production strategies are most suitable for this purpose and that the choice of the selection feature is case-dependent even within the same workflow. A disadvantage of this approach is the challenge in deriving meaningful statistics of the VOI given the reduced number of plausible truths. Finally, we combined both aforementioned acceleration measures to design a new procedure for faster VOI assessment. For the 2D example used in this paper, we were able to reduce the amount of required reservoir simulations from millions to tens of thousands. This significant reduction in computational costs represents an important step for the use of the VOI assessment in larger examples.

However, there is still scope for future research to further accelerate these workflows. A first step would be to extend the selection of reduced ensembles to the data assimilation problem. Representative models for production optimization might not be as relevant for history matching. Moreover, a challenge may arise if we are to work with reduced sets of models in ensemble-based frameworks. Furthermore, the combination of the ideas presented in this paper with the use of surrogate (e.g., proxies) or multiscale models may be useful in the future development of practical tools for VOI assessment to be applied in real-field applications.

Acknowledgements

This research was carried out within the context of the ISAPP Knowledge Centre. ISAPP (Integrated Systems Approach to Petroleum Production) is a joint project of TNO, Delft University of Technology, ENI, Statoil and Petrobras. The EnKF module for MRST was developed by Olwijn Leeuwenburgh (TNO) and can be obtained from http://www.isapp2.com/data-sharepoint/enkf-module-for-mrst. The authors thank Siep Weiland of Eindhoven University of Technology for useful discussions on tensor decomposition.

References

Aggarwal, C.C., Wolf, J.L., Yu, P.S., Procopiuc, C., Park, J.S. [1999] Fast algorithms for projected clustering. *SIGMOD Record*, **28**(2), 61-72. DOI: 10.1145/304181.304188.

Arabie, P. and Hubert, L.J. [1996] An overview of combinatorial data analysis. *Clustering and Classification*, 5-63. World Scientific Pub., New Jersey.

Armstrong, M., Ndiaye, A., Razanatsimba, R. and Galli, A. [2013] Scenario reduction applied to geostatistical simulations. *Mathematical Geosciences* **45**, 165-182. DOI: 10.1007/s11004-012-9420-7.

Baker, M. [2015] Use of cluster analysis to improve representative model selection: a case study. Paper SPE 176408-MS presented at SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition, Nusa Dua, Bali, Indonesia, 20-22 October. DOI: 10.2118/176408-MS

Barros, E.G.D., Van den Hof, P.M.J. and Jansen, J.D. [2016] Value of information in closed-loop reservoir management. *Computational Geosciences* **20**(3), 737-749. DOI: 10.1007/s10596-015-9509-4.

Caers, J. [2011] *Modeling uncertainty in the earth sciences*. John Wiley & Sons.

Castro, S.A., Caers, J. and Mukerji, T. [2005] The Stanford VI reservoir. 18th Annual Report. Stanford Center for Reservoir Forecasting. Stanford University, May 2005.

Chen, Y., Oliver, D.S. and Zhang, D. [2009] Efficient ensemble-based closed-loop production optimization. *SPE Journal* **14**(4) 634-645.

Cunningham, J.P. and Ghahramani, Z. [2015] Linear dimensionality reduction: survey, insights, and generalizations. *Journal of Machine Learning Research* **16** 2859-2900.

Eidsvik, J., Mukerji, T. and Bhattacharjya, D. [2015] *Value of information in the earth sciences – Integrating spatial modeling and decision analysis*. 1st ed, Cambridge University Press.

Hou, J., Zhou, K., Zhang, X.S. Kang, X.D. and Xie, H. [2015] A review of closed-loop reservoir management. *Petroleum Science* **12** 114-128. DOI: 10.1007/s12182-014-0005-6.

He, J., Xie, J., Sarma, P., Wen, X.H., Chen, W.H. and Kamath, J. [2016] Proxy-based work flow for a priori evaluation of data-acquisition programs. *SPE Journal*. DOI: 10.2118/173229-PA.

Insuasty, E., Van den Hof, P.M.J., Weiland, S., & Jansen, J. D. [2015] Spatial-temporal tensor decompositions for characterizing control-relevant flow profiles in reservoir models. SPE 173238-MS, SPE Reservoir Simulation Symposium, Houston, USA, 23-25 February. DOI: 10.2118/173238-MS.

Jansen, J.D., Brouwer, D.R., Nævdal, G. and van Kruijsdijk, C.P.J.W. [2005] Closed-loop reservoir management. *First Break*, **23**(1), 43-48.

Jansen, J.D., Bosgra, O.H. and van den Hof, P.M.J. [2008] Model-based control of multiphase flow in subsurface oil reservoirs. *Journal of Process Control* **18**, 846-855.

Jansen, J.D., Douma, S.G., Brouwer, D.R., Van den Hof, P.M.J., Bosgra, O.H. and Heemink, A.W. [2009] Closed-loop reservoir management. SPE 119098, SPE Reservoir Simulation Symposium, The Woodlands, USA, 2-4 February.

Keim, D., Berchtold, S., Böhm, C., Kriegel., H. P. [1997] A cost model for nearest neighbor search in high dimensional data space. *Proc. of the 16th Symposium on Principles of Database Systems*, 78-86. DOI: 10.1145/263661.263671.

Kruskal, J.B. [1964] Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, **29**(1), 1-27.

Kruskal, J.B. [1976] The relationship between multidimensional scaling and clustering. *Proc. of Advanced Seminar Conducted by the Mathematics Research Center*, the University of Wisconsin-Madison, 3-5 May, 17-44. DOI: 10.1016/B978-0-12-714250-0.50006-1

Le, D.H. and Reynolds, A.C. [2014a] Optimal choice of a surveillance operation using information theory. *Computational Geosciences*, **18** 505-518. DOI: 10.1007/s10596-014-9401-7.

Le, D.H. and Reynolds, A.C. [2014b] Estimation of mutual information and conditional entropy for surveillance optimization. *SPE Journal*, **19**(4) 648-661. DOI: 10.2118/163638-PA.

Lie, K.-A., Krogstad, S., Ligaarden, I.S., Natvig, J.R., Nilsen, H.M. and Skalestad, B. [2012] Open source MATLAB implementation of consistent discretisations on complex grids. *Computational Geosciences*, **16**(2), 297-322.

Naevdal, G., Brouwer, D.R. and Jansen, J.D. [2006] Waterflooding using closed-loop control. *Computational Geosciences*, **10**(1) 37-60.

Parsons, L., Haque, E. and Liu, H. [2004] Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter*, **6**(1) 90-105.

Sarma, P., Durlofsky, L.J. and Aziz, K. [2008] Computational techniques for closed-loop reservoir modeling with application to a realistic reservoir. *Petroleum Science and Technology*, **26**(10&11) 1120-1140.

Sarma, P., Chen, W.H. and Xie, J. [2013] Selecting representative models from a large set of models. Paper SPE 163671 presented at the SPE Reservoir Simulation Symposium, The Woodlands, USA, 18-20 February.

Scheidt, C. and Caers, J. [2009] Representing spatial uncertainty using distances and kernels. *Mathematical Geosciences*, **41**(4) 397-419.

Strebelle, S. [2002] Conditional simulation of complex geological structures using multiple-point statistics. *Mathematical Geology*, **34**(1) 1-22. DOI: 10.1023/A:1014009426274.

Van Essen, G.M., Zandvliet, M.J., Van den Hof, P.M.J., Bosgra, O.H. and Jansen, J.D. [2009] Robust waterflooding optimization of multiple geological scenarios. *SPE Journal,* **14**(1), 202-210.

Yeten, B., Durlofsky, L.J. and Aziz, K. [2003] Optimization of nonconventional well type, location and trajectory. *SPE Journal*, **8**(3) 200-210.

Wang, C., Li, G. and Reynolds, A.C. [2009] Production optimization in closed-loop reservoir management. *SPE Journal*, **14**(3) 506-523.