energies

Article

Efficient Dimensionality Reduction Methods in Reservoir History Matching

Amine Tadjer [1,*], Reider B. Bratvold [1] and Remus G. Hanea [1,2]

[1] Department of Energy Resources, University of Stavanger, 4021 Stavanger, Norway; reidar.bratvold@uis.no (R.B.B.); remus.hanea@uis.no (R.G.H.)
[2] Equinor ASA, Forusbeen 50, 4035 Stavanger, Norway
* Correspondence: amine.tadjer@uis.no

Abstract: Production forecasting is the basis for decision making in the oil and gas industry, and can be quite challenging, especially in terms of complex geological modeling of the subsurface. To help solve this problem, assisted history matching built on ensemble-based analysis such as the ensemble smoother and ensemble Kalman filter is useful in estimating models that preserve geological realism and have predictive capabilities. These methods tend, however, to be computationally demanding, as they require a large ensemble size for stable convergence. In this paper, we propose a novel method of uncertainty quantification and reservoir model calibration with much-reduced computation time. This approach is based on a sequential combination of nonlinear dimensionality reduction techniques: t-distributed stochastic neighbor embedding or the Gaussian process latent variable model and clustering K-means, along with the data assimilation method ensemble smoother with multiple data assimilation. The cluster analysis with t-distributed stochastic neighbor embedding and Gaussian process latent variable model is used to reduce the number of initial geostatistical realizations and select a set of optimal reservoir models that have similar production performance to the reference model. We then apply ensemble smoother with multiple data assimilation for providing reliable assimilation results. Experimental results based on the Brugge field case data verify the efficiency of the proposed approach.

Keywords: uncertainty quantification; history matching; reservoir simulation; data assimilation; dimensionality reduction

Citation: Tadjer, A.; Bratvold, R.B.; Hanea, R.G. Efficient Dimensionality Reduction Methods in Reservoir History Matching. *Energies* **2021**, *14*, 3137. https://doi.org/10.3390/en14113137

Academic Editors: Vasily Demyanov; Leonardo Azevedo

Received: 25 February 2021
Accepted: 24 May 2021
Published: 27 May 2021

1. Introduction

Research scientists have worked for many years to develop viable methods to calibrate complex reservoir models. However, the uncertainty associated with reservoir models is highly significant, introducing considerable errors in the modeling process. There are several ways to quantify uncertainty in reservoirs. One is the conditioning of reservoir parameters to observed production data, a process referred to as inverse problem or history matching (HM). The first step of HM is parameterization, namely to independently define and vary the model variables in a numerical reservoir simulation model: porosity, permeability, the density and permeability of fractures, the initial depths of oil-water and gas-oil contacts, relative permeability curves, capillary pressure curves, fluid composition, aquifer strength, and the size and fault transmissibility [1]. It is not realistic to do so, however, because of the large area of possible adjustment caused by the large number of grid blocks and variables; the number of varying parameters should therefore be as small as possible. To do this, a reparameterization method based on the pilot point method, the spline function method, the wavelet function method, Karhunen–Loeve reparameterization, and discrete cosine transform was used [2]. The second step is to select the production data, which must be sensitive to the parameters needed to be history matched. The sensitivity becomes more complex, however, in cases using reservoirs with multiphase flow. In these cases, the cross-covariance of production data to model variables is used instead, its main advantage being that it is generally smoother and can show a more global relationship between data and variables, since it is a product of sensitivities and covariances.

The algorithms for HM are diverse. Evolutionary algorithms are often considered the standard approach, since, by generating a new model combining two Gaussian reservoir models, the gradual deformation algorithm reduces the HM problem to a one-dimensional minimization problem [2]. Sambridge (1999) [3] introduced a neighborhood algorithm in which a resampling of the parameters is led by using information in an available ensemble. In addition, several other methods have been introduced to optimize reservoir models via particle swarm optimization [4], simulated annealing [5], and simultaneous perturbation stochastic approximation [6]. When solving the history matching problem, a key issue must be considered: uncertainty quantification. Uncertainty quantification requires strong knowledge of the reservoir characteristics, and uncertainty should be represented by a set of reservoir models (or realizations) instead of a single history-matched model. The Markov chain Monte Carlo method (McMC) [7], the randomized maximum likelihood method [8], the EnKF method [8–10], the ensemble smoother (ES) [11], and the ensemble smoother with multiple data assimilation (ES-MDA) [12,13] are useful methods to quantify uncertainty. For all of these techniques, accuracy and speed are two main factors due to the non-unique solutions and the ill-posed inverse problems.

Many parametrization methods used in DR have already been introduced. For instance, Vo and Durlofsky [14] used principal component analysis (PCA) to reparametrize high dimension data into low dimensional space, then regenerated new realizations based on principal parameters from PCA for data assimilation, while others have used singular value decomposition [15] and Kernel PCA (KPCA) [16]. Muzammil. H et al. [17] applied PCA to account for the model-error component during model calibration. Kang et al. (2017) and Kang et al. (2019) [18,19] also introduced PCA to select suitable models for EnKF. Tolstukhin et al. [20] demonstrated how data analytics can improve efficiency of ensemble history matching by analyzing the statistics that link the static model ensemble and the dynamic model ensemble update. Satija et al. [21] proposed a method known as direct forecasting (DF) based on projecting the prior predictions into a low-dimensional canonical space to maximize the projected oil data and estimate the joint distribution of historical and forecasted data through linear Gaussian regression; they concluded that this method provided uncertainty estimates regarding production forecast that reasonably agreed with rejection sampling. Park et al. [22] proposed an extended approach based on direct forecasting, where both of the geological model parameters and dynamic data are simultaneously used. Our approach in the current paper is different from the previous work in Kang et al. (2019) [19]. Dimensionality reduction techniques such as PCA and SVD, however, are linear approaches that may not accurately represent the relationship between high dimensional parameters and latent variables in reduced space, which likely lead to poor performance of model assimilation and prediction. In addition, the use of EnKF tends to be computationally prohibitive in certain circumstances and also generates spurious correlations leading to loss of geological realism and underestimation of uncertainties (ensemble collapse). In this work, we propose a novel scheme to reduce the number of ensemble members while preserving the prediction quality by combining ES-MDA with machine learning DR techniques and cluster analysis. In this paper, we demonstrate the efficiency of using the non-linear DR techniques t-distributed stochastic neighbor embedding (t-SNE) [11] and Gaussian process latent variable model (GPLVM) [23,24] along with clustering K-means to select effective reservoir models and save computational time without simulating and assimilating the entire initial ensemble. This study uses the Brugge field reservoir case to demonstrate that the new implementation can make computation faster and more robust than the standard procedure proposed in [12,13] and can provide appropriate posterior uncertainty quantification.

The paper is structured as follows. In the next section, we present the complete methodology, by which we tested the proposed workflow in the well-known Brugge field reservoir model. In addition, several cases, involving different reference models, are considered. Finally, some concluding remarks and possible future work directions are provided.

2. Materials and Methods

The procedure applied in this study has four main stages:

1. The first stage includes generating ensemble reservoir models and analyzing whether the observed (reference) prior data can predict posterior distribution that appertains to the prior range.
2. The second stage involves reducing the ensemble dimension and constructing a 2D space by using t-SNE and GPLVM.
3. The third stage uses clustering K-means to extract a set of reservoir models with the least production error compared to the reference model.
4. After extracting the models and selecting the most informative ones, we began the HM process using ES-MDA, and finally we compared the performance of history matching analysis of the proposed workflow with the standard ES-MDA without using dimensionality reduction techniques.

The general steps of the approach are shown in Figure 1 and algorithm solutions employed are described in more detail later.

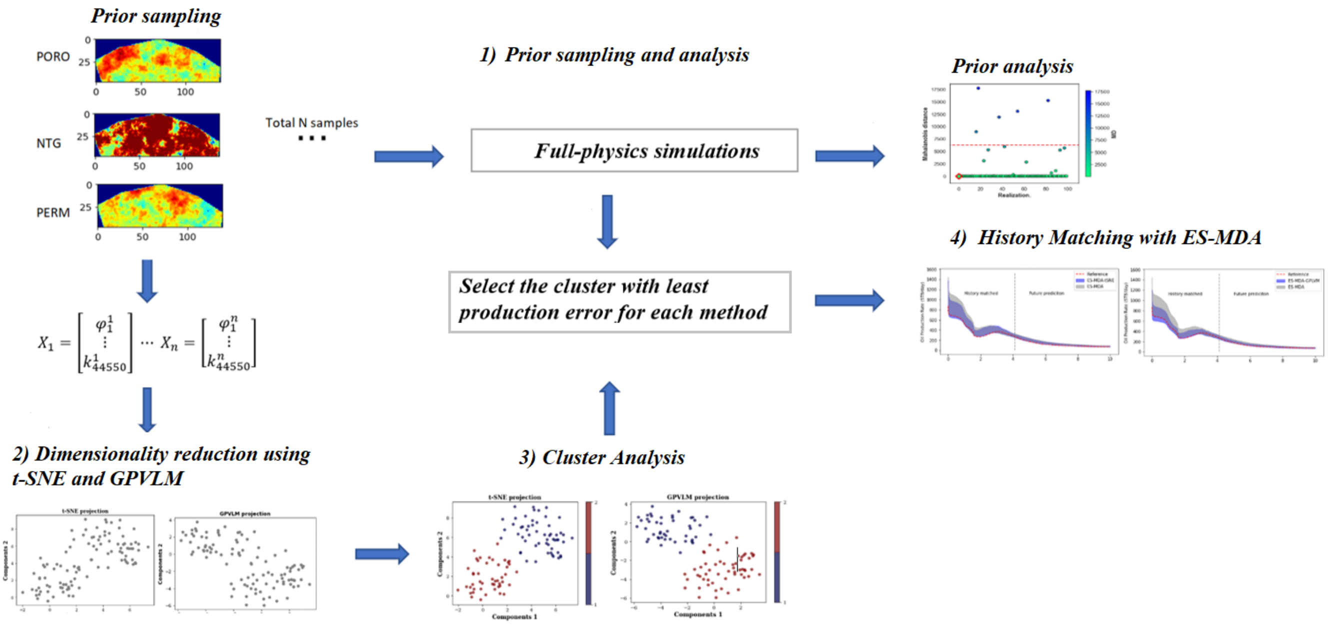

Prior sampling
PORO
NTG
PERM
Total N samples
1) Prior sampling and analysis
Full-physics simulations
Prior analysis
4) History Matching with ES-MDA
Select the cluster with least production error for each method
$X_1 = \begin{bmatrix} \varphi_1^1 \\ \vdots \\ k_{44550}^1 \end{bmatrix} \cdots X_n = \begin{bmatrix} \varphi_1^n \\ \vdots \\ k_{44550}^n \end{bmatrix}$
2) Dimensionality reduction using t-SNE and GPVLM
3) Cluster Analysis
t-SNE projection
GPVLM projection
Components 1
Components 2

Figure 1. Flow chart for the history matching with dimensionality reduction framework.

2.1. Prior Sampling and Analysis

Due to the high dimensionality and nonlinearity of subsurface systems physics-based models, Monte Carlo simulations were used to sample and identify the possible prior range of model parameterization and probability distribution for each geological parameter (e.g., the structural model, rock types, the petrophysical model, and subsurface fluid distribution). Let $m \in \mathcal{R}^N$ denote the vector of uncertain static parameters of a reservoir model with a dynamic data variable (e.g., oil production and water cuts) as vector d. The nonlinear function data forward model is defined as

$$d = G_d(m) \tag{1}$$

The function $\mathcal{G}_d$ is generated through a reservoir simulator and by applying it to prior geological model realizations, $m = \{m^1, m^2, m^3, \ldots\ldots\ldots m^N\}$. We obtained a set of N samples of dynamic data variables, $d = \{d^1, d^2, d^3, \ldots\ldots\ldots d^N\}$. We refer to the vector of observation data as d_{obs}. Once the prior samples are generated, it is important to check that the observed data can be predicted by the prior model, in order for the posterior distribution to appertain in the prior range. Otherwise, there is a risk that the prediction will be erroneous. If the prior model is falsified, which indicates inconsistency with the data, we must revise the prior data distribution herein to evaluate the quality of the prior model and its ability to predict the data. We proposed a statistical procedure based on Robust Mahalanobis distance (RMD) [25,26], which handles high dimensional and different types of measurements of the data, the main objective being to detect outliers and determine if the prior model is falsified or not. The RMD for each data variable realization d or d^{obs} was computed as follows:

$$RMD\ (d^{(n)}) = \sqrt{(d^{(n)} - \rho)\beta^{-1}(d^{(n)} - \rho)}, \quad for\ \ n = 1, 2, 3 \ldots\ldots\ldots, N \tag{2}$$

where ρ and β are the mean and covariance of the data d. Assuming the distribution of the data is multivariate Gaussian, the distribution of $[RMD(dn)]^2$ would be chi squared x_d^2. We set the 95th percentiles of x_d^2 as the tolerance threshold for multivariate dimensional point d^n. If $RMD\ (d_{obs})$ fell outside of the tolerance threshold $(RMD\ (d_{obs}) > RMD\ (d^n)$, the d_{obs} would be regarded as outliers, and the prior model would be falsified, as it has a very small probability. It should also be noted that this method requires data distribution to be Gaussian; if it is not, other outlier detection techniques such as isolation forest [27], local outliers detection [28], and one-class support vector machines [29] are highly recommended.

2.2. Dimensional Reduction

A single reservoir model is represented by numerous grid blocks, each with unique reservoir properties, such as permeability, porosity, and net-to-gross. Accordingly, we construct a vector X containing the reservoir properties of all grid blocks. We also use multiple ensembles of realizations to account for geological uncertainties $X \in R^{N,m}$. Furthermore, typical ensembles are formed by hundreds of realizations, in that we are faced with a high-dimensional problem. Geological realization with similar geological parameters trends will have similar production histories. As we aim to analyze the main geological distribution of the data, reducing the data dimensions is reasonable. Therefore, we utilize two different DR methods: t-SNE [11] and GPLVM [23,24], to characterize reservoir parameters efficiently by projecting the parameters into a 2D plane. However, t-SNE is a non-linear DR algorithm developed for exploring high-dimensional data. It maps multi-dimensional data to a two- or three-dimensional dataset that can be visualized in a scatter plot. Additionally, t-SNE learns joint probabilities defined by two points on a two-dimensional space to be as close as possible to conditional probabilities, defined by two points on high-dimensional space. For more details about t-SNE, one can refer to [11] and Appendix A. GPLVM differs from t-SNE, primarily because it is a Bayesian non-parametric DR method that uses Gaussian process to learn a low-dimensional representation of high-dimensional data. The main advantage of the GPLVM is that it allows the use of nonlinear covariance functions, i.e., that it can represent non-linear functions from the latent space to the data space. The probabilistic nature of the GPLVM also gives it advantages in dealing with missing data values. For more details about GPLVM, one can refer to [23,24] and Appendix A.

2.3. Clustering K-Means:

K-means clustering is an unsupervised learning method that is widely used because of its efficiency and simplicity. K-means is used to find the cluster configuration that minimizes the square error over all K clusters [30]:

$$J = \sum_{k=1}^{K} \sum_{x^{(l)} \in c_k}^{M} \left\| x^{(l)} - \mu^k \right\|^2, u^k = \frac{\sum^{x^{(l)} \in c_k} x^{(l)}}{|S_k|} \tag{3}$$

with u^k as the centroid of the cluster, and c_k refers to the mean of a point within the cluster, $|S_k|$ is the number of samples in the cluster c_k.

K-means clusters provide an optimal solution by minimizing of the distances between data and their centroids. The centroid is computed by the average of the data in each cluster. Several methods exist that allow the selection of the cluster sizes, including the gap statistics, elbow-method as well as the silhouette-method. In this study, we use the silhouette-method to determine the optimal number of clusters. The silhouette index varies between -1 and 1, where a value close to 1 means that the data is appropriate within its cluster. For all of the data-points, the silhouette value $s(i)$ can be determined with the following equation:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{4}$$

where $a(i)$ represents average distances within a specific cluster and $b(i)$ is the minimum average distance from data in other separate clusters. Specifically, $a(i)$ shows how the i-th data is grouped within its cluster, and $b(i)$ indicates the closest distance of adjacent clusters. Therefore, if $a(i)$ is small, that means that the data are well grouped; however, if the silhouette value is close to 1, the $b(i)$ is large.

2.4. ES-MDA and the Localization Technique

Opposite to the production forecast where the "unknown" reservoir behavior is predicted by using the "known" reservoir model variables, history matching inverses the process and estimates the "unknown" reservoir model variables with the "known" observed reservoir behavior. The general objective function for history matching is

$$J(m) = \frac{1}{2}\|g(m) - d_{obs}\|^2 \tag{5}$$

where $g(m)$ is the simulated data with model variables m composed of reservoir variables (e.g., permeability, facies, porosity, and net to gross), and d_{obs} is the observed data. The goal of history matching is to minimize the objective function (m) by finding acceptable model variables m and $minJ(m)$. ES-MDA is an ensemble-based method introduced by Emerick and Reynolds [12]. In its simplest form, the method employs a standard smoother analysis equation a pre-defined number of times, with the covariance matrix of the measured data error multiplied by coefficient a. The coefficients must be selected such that the following condition is satisfied:

$$\sum_{k=1}^{N_a} \frac{1}{a_k} = 1 \tag{6}$$

where N_a is the number of times the analysis step is repeated. The ES-MDA analysis applied to a vector of model parameters, m, can be written as

$$m_i^a = m_i^b + K(d_{obs} - d_{sim,i}), \quad for\ i = 1 \ldots\ldots N \tag{7}$$

Here, i is defined as the ith ensemble members; m_i^a is defined as an updated uncertainty vector, m_i^b, the initial or previous uncertainty vector; K, the Kalman gain matrix, which is used to compute by regularizing with SVD using 99.9 % of all the energy; $d_{sim,i}$ refers to simulation data obtained from previous models. Ensemble-based HM updates N reservoir models simultaneously. In addition, the Kalman gain matrix can be determined as follows:

$$K = C_{md}\left(C_{dd} + a_p C_D\right)^{-1} \tag{8}$$

C_{md} is the cross-covariance matrix between the vector of model parameters m and predicted data d; C_{dd} is the auto covariance matrix of predicted data d; a_p is the coefficient to inflate C_D, which refers the covariance matrix of the observed data measurement error.

However, there are still conceptual and computational challenges associated with ES-MDA, one issue is that of ensemble collapse, which may result in unrealistic uncertainty and difficulty to cover the target distribution. To avoid this, a localization technique is implemented in the equation by introducing a correlation matrix R via an element-by-element multiplication, also known as Schur product ($\circ$). There are different ways of computing R. One of the most common approaches is the distance-dependent localization [31], in which all data points (oil rate, water rate) and model variables (permeability, porosity) are presumed to have certain physical locations connection. The ES-MDA equation is updated to:

$$m_i^a = m_i^b + R \circ K(d_{obs} - d_{sim,i}), \quad for\ i = 1 \ldots\ldots N \tag{9}$$

The parameter R is assumed to be from 0 to 1 depending on the distances for well locations [32]:

$$R(h, L) = \left\{ \begin{matrix} -\frac{1}{4}(\frac{h}{L})^5 + \frac{1}{2}(\frac{h}{L})^4 + \frac{5}{8}(\frac{h}{L})^3 + \frac{5}{8}(\frac{h}{L})^2 + 1 & , 0 \leqslant h < L \\ \frac{1}{12}(\frac{h}{L})^5 - \frac{1}{2}(\frac{h}{L})^4 + \frac{5}{8}(\frac{h}{L})^3 + -\frac{5}{3}(\frac{h}{L})^2 - 5(\frac{h}{L}) + 4 - \frac{2}{3}(\frac{h}{L})^{-1} + 1 & , L < h \leqslant 2L \\ 0 & , h > 2L \end{matrix} \right\} \tag{10}$$

where h is the Euclidean distance between a specific grid cell and well location, and L refers to the critical length, corresponding to influential regions for every well data. Therefore, a high value of R means that grid blocks are close to the wells.

2.5. General Setup

We tested the performance of the proposed methodology in the Brugge field case study. The Brugge field is a complex oilfield constructed by TNO [33]. The model consists of nine layers, and each layer has 139 × 48 gridblocks. The total number of gridblocks is 60,048, with 44,550 active cells. There are 20 producers and 10 injectors in the reservoir models.The reservoir is being depleted by voidage replacement. The producers and injectors are "smart wells", i.e., with vertical flow control, with three perforation intervals per well. For each producer well, the fluid rate is set to max value of 3000 bbl/day, and flowing bottom hole pressure superior to 50 Bar. For each injector well, the fluid rate is set to a max value of 4000 bbl/day, and the corresponding flowing bottom hole pressure less than 180 Bar. We used 104 initial geostatistical realizations provided by TNO and assumed one of 104 as a reference model. In the HM analysis, oil production rates (OPR), water cuts (WCT), and the bottom hole pressure (BHP) were considered, and the model variables to be updated included permeability (PERMX, PERMY, and PERMZ), porosity, and NTG in all active cells. Figure 2 shows the log permeability in the first layer for six random realizations. For more information about the Brugge benchmark, see [33].

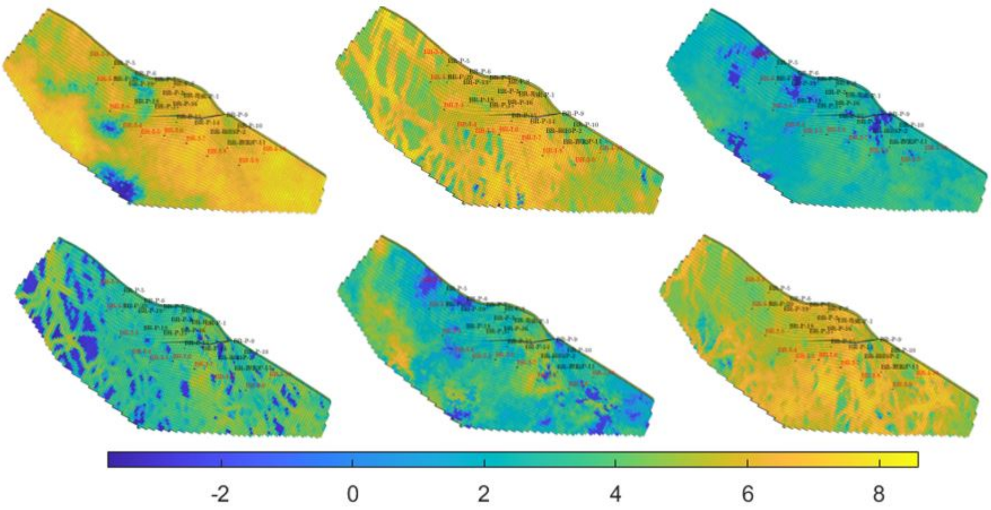


Figure 2. Example log of permeability (K) distribution for six of 103 different geological realizations of the Brugge field.

3. Results

3.1. ES-MDA with DR

The data assimilations were conducted using ES-MDA with localization, with Na = 5. Besides the ES-MDA-t-SNE and ES-MDA-GPLVM, we performed the HM on assimilation observation data for the first 4 years and the rest (6 years) for forecasting.

To assess the quality of prior models, a field oil production rate of 103 prior models was used with d_{obs} by applying the RMD outlier detection. The RMD of d_{obs} was found to be 8.802, which is below the 95th percentile threshold. This means that the prior is not incorrect. Figure 3 shows a comparison between RMD with d_{obs} and RMD with 103 prior models.

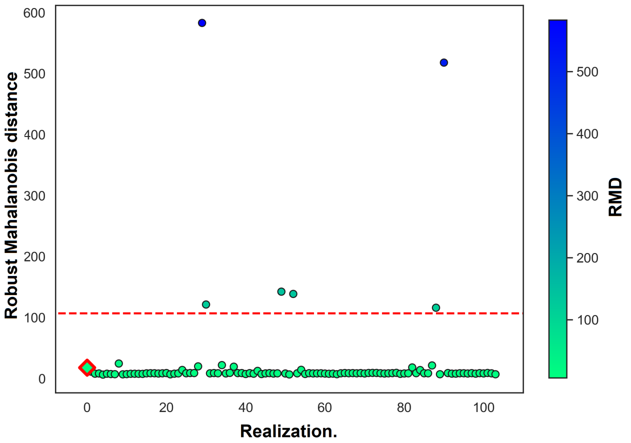


Figure 3. Prior falsification using RMD. The red diamond is the RMD for d_{obs}. Circle dots refer to the RMD results of 103 data variable samples, and the red dashed line is the 95th percentile of the chi-squared distributed RMD.

We applied t-SNE and GPLVM to reservoir models and reduced the dimension into 2D space. Additionally, the silhouette method is used to find optimal cluster numbers (ranges of 2 to 7 clusters) for both GPLVM and t-SNE 2D space. As displayed in Figure 4, the dashed line is used to denote the average silhouette, the silhouette plot with 2 clusters showed the highest value. Therefore, each model was divided into 2 clusters.

Figures 5 and 6 show a scatter plot of the 103 models on a 2D plane, with each dot indicating individual models. When selecting the cluster with the least production error and comparing the forecast accuracy of different forecasting methods among several data sets, there are many performance measures from which to select. In this study, we chose to evaluate, for our forecasting results, a probabilistic metric called the mean continuous ranked probability score (CRPS). The mean CRPS quantifies both accuracy and precision [34], and higher values of the CRPS indicate less accurate results. The mathematical formulations of the mean CRPS are listed in Appendix A. We compared the field oil production rate (FOPR) errors between each cluster and reference model and selected the cluster with the least production error for the data assimilation process, as displayed in Table 1. Only 46 models were selected using t-SNE and 44 models using GPLVM. In Figure 7, we compare the average permeability values between initial 103 models and the selected 46 and 44 models using t-SNE and GPLVM, respectively. Both selected models with t-SNE and GPLVM have a quite similar distribution and quite similar selected reservoir models, and differences were found only in two models.

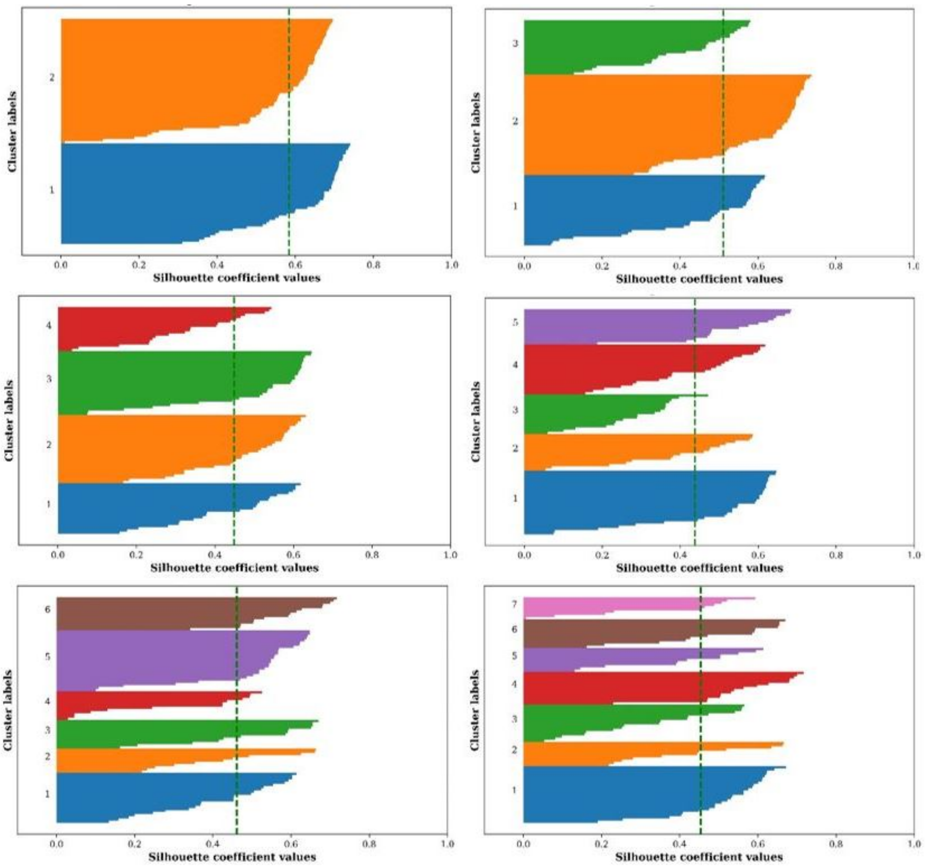


Figure 4. Silhouette plots with different cluster numbers—t-SNE 2D space.

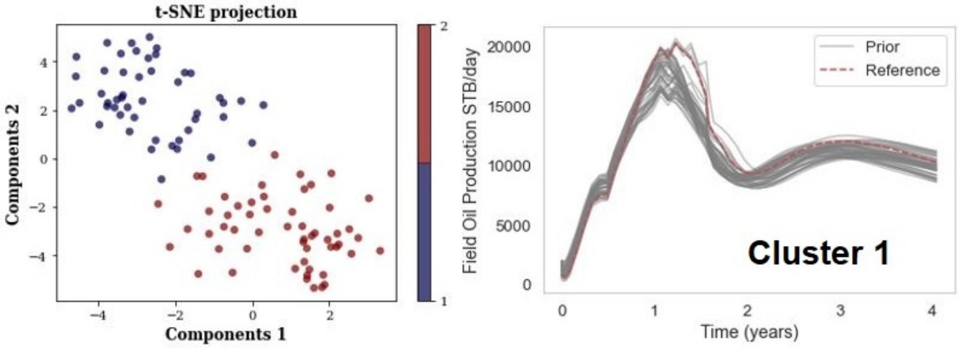


Figure 5. Model selection using t-SNE.

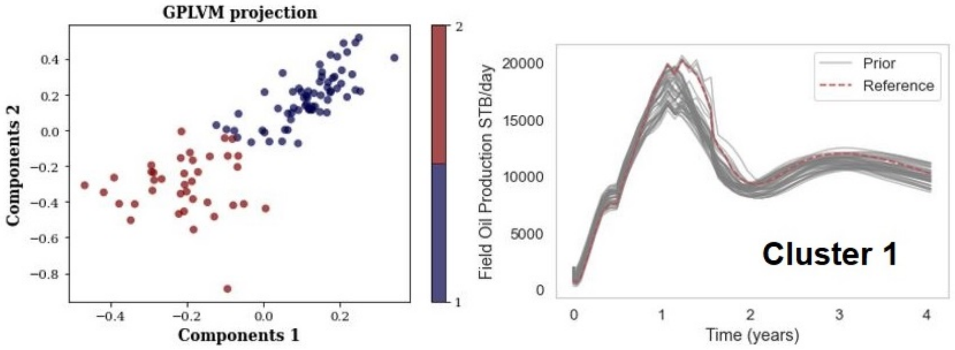
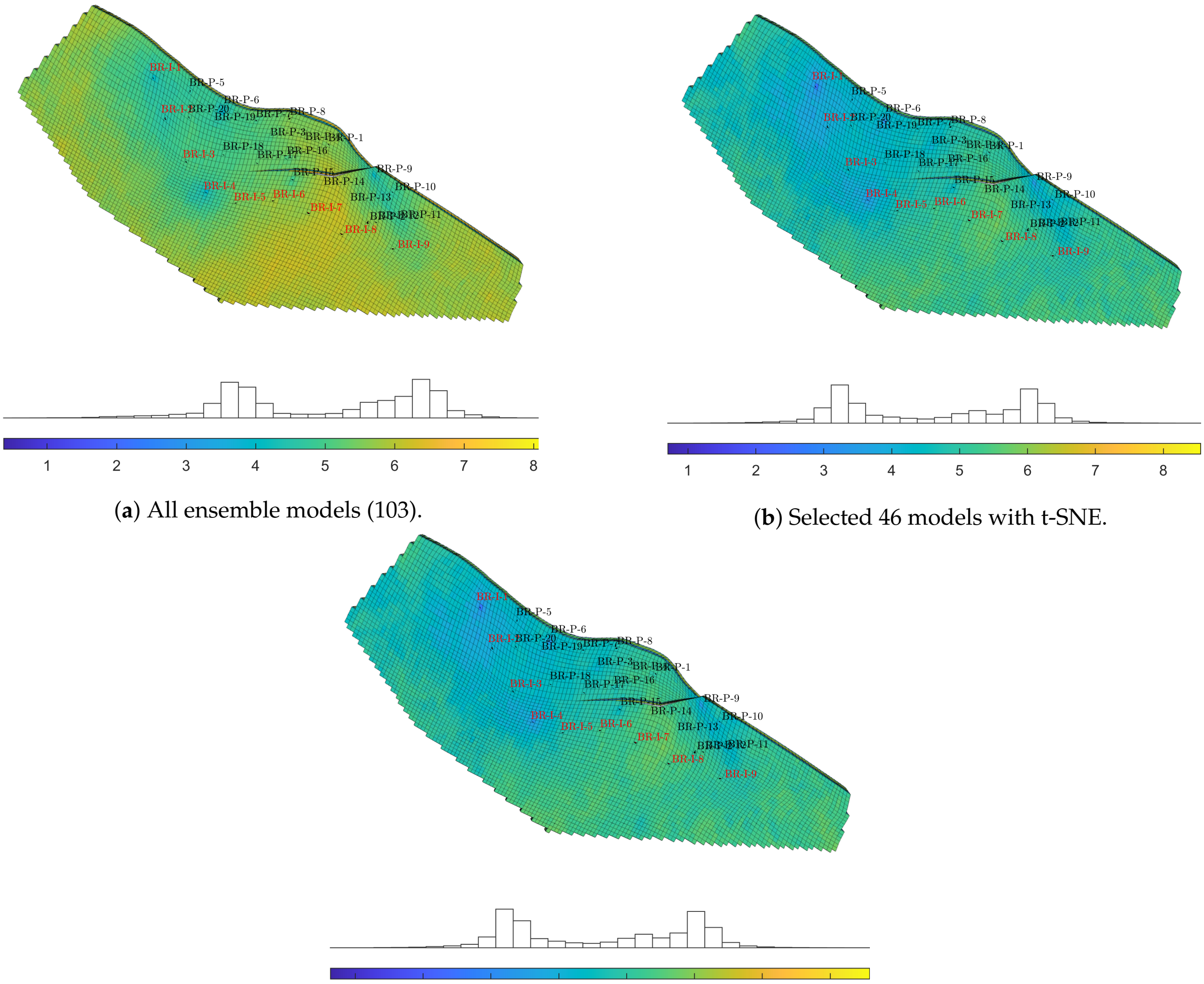


Figure 6. Model selection using GPLVM.

Table 1. Measurement error between the reference model and each cluster.

Methods	**t-SNE**		**GPLVM**	
	CRPS	**Realization**	**CRPS**	**Realization**
Cluster 1	89.78	46	96.77	44
Cluster 2	130.67	57	128.66	59



(**a**) All ensemble models (103).

(**b**) Selected 46 models with t-SNE.

(**c**) Selected 44 model with GPLVM.

Figure 7. Mean of permeability Darcy values in logarithmic scale.

The total simulation for each method is listed in Table 2. We can see that ES-MDA uses around 220 min for the entire process, while ES-MDA-t-SNE and ES-MDA-GPLVM use around 120 and 101.5, respectively. By employing reduction techniques, more than 45% of the total simulation time was saved.

Table 2. CPU time for the whole process.

Methods	**CPU Time (Minutes)**	**Time Reduction**
ES-MDA	220	0
ES-MDA-t-SNE	120	45.5%
ES-MDA-GPLVM	101.5	53.86%

Figures 8 and 9 compare the ensemble means distribution of the reconstructed updated posterior log-perm and porosity on Layer 1 using the standard ES-MDA, ES-MDA-t-SNE, and ES-MDA-GPLVM to their counterparts in the prior models. The results suggest a slight change on the posterior model in areas where the wells are located. Moreover, the uncertainty reduction is achieved, as the posterior samples are conditioned to the dynamic data variables of the well that are contained within the prediction domain. The results show quite similar posterior permeability and porosity distribution for both ES-MDA-t-SNE and ES-MDA-GPLVM, which is expected, as they differ only in two models.

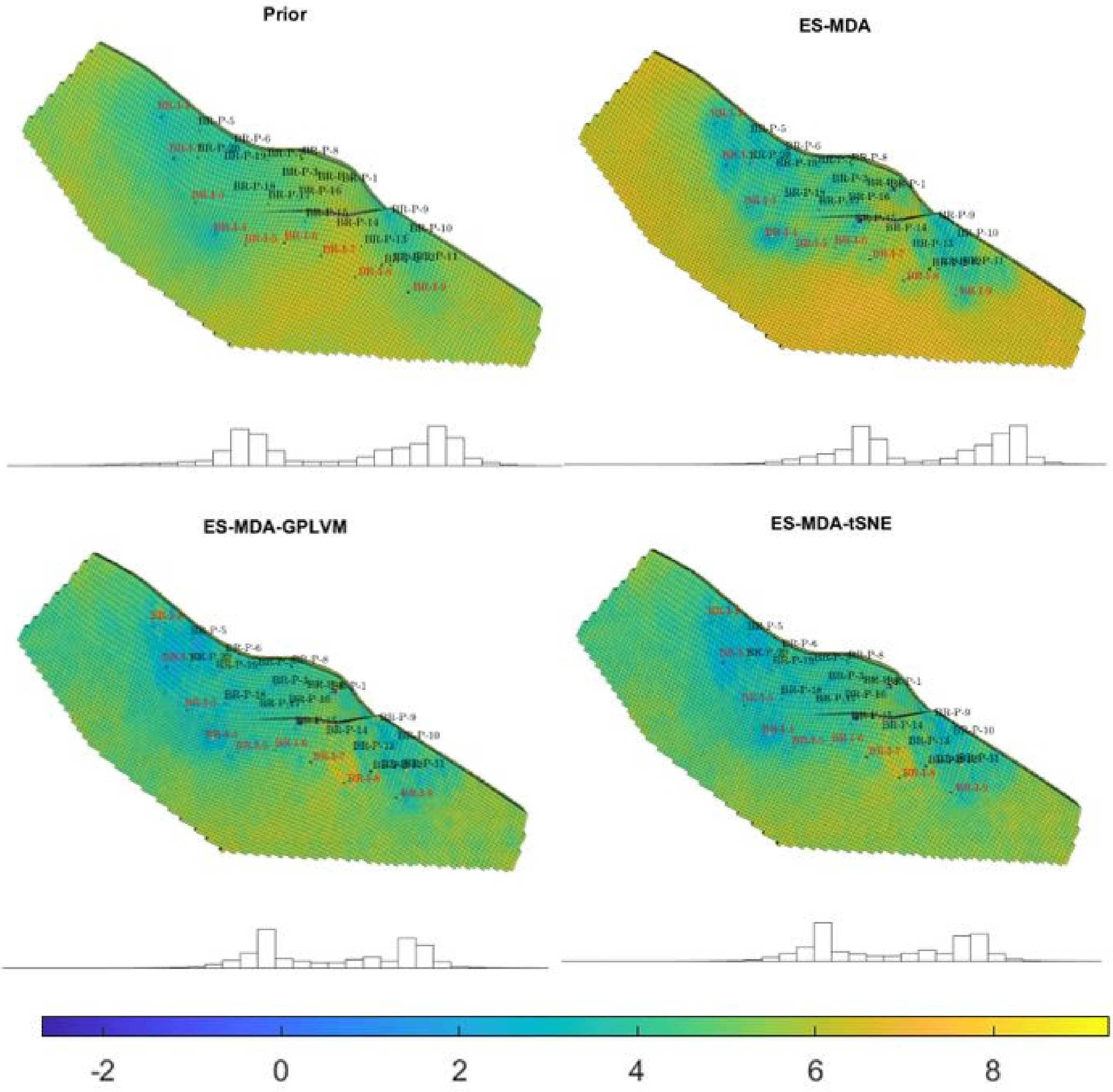


Figure 8. Average log–permeability distribution on Layer 1 from an initial ensemble, the corresponding updated model by ES-MDA, ES-MDA-t-SNE, and ES-MDA-GPLVM.

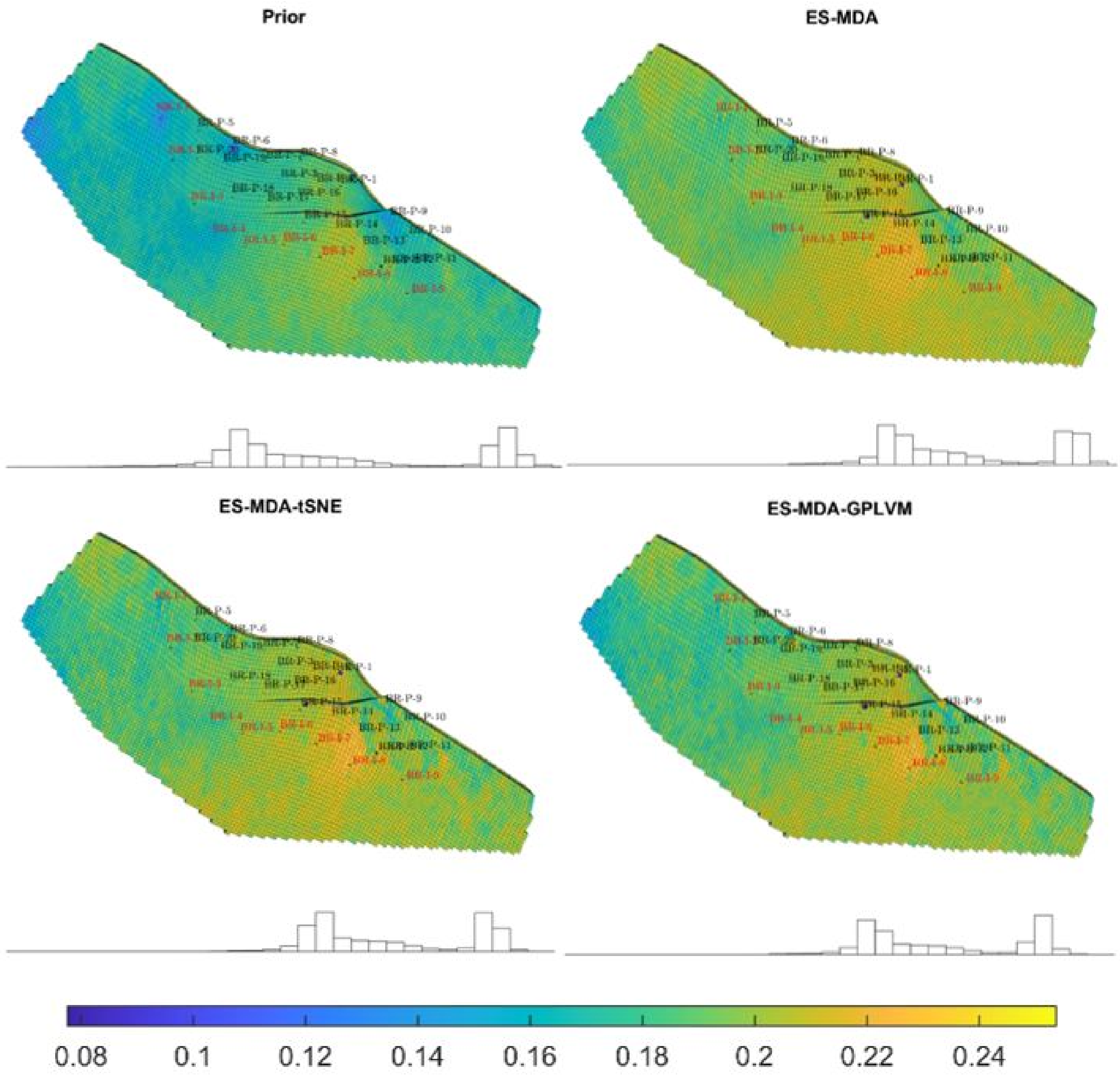


Figure 9. Average porosity distribution on Layer 1 from an initial ensemble, the corresponding updated model by ES-MDA, ES-MDA-t-SNE, and ES-MDA-GPLVM.

Figures 10 and 11 depict the HM profiles for both oil and water cuts of two methods (ES-MDA-t-SNE and ES-MDA-GPLVM) at producers BR-P5, BR-P6, and BR-P19, with respect to the standard ES-MDA and reference model. The vertical dashed line represents the last time of the HM process. The production forecast seems to be reasonable and reliable in the two methods compared to the standard ES-MDA, although only 45.54% and 53.86% of the simulation time is required for ES-MDA-t-SNE and ES-MDA-GPLVM, respectively. The ES-MDA-t-SNE, however, predicts the WOPR data at BR-P6 better than the ES-MDA-GPLVM does, which is likely related to the fact that the WWCT data of BR-P6 are better when using ES-MDA-t-SNE. The matching and forecast ranges with ES-MDA-GPLVM, however, deviate from the reference, especially in BR-P5 and BR-P6.

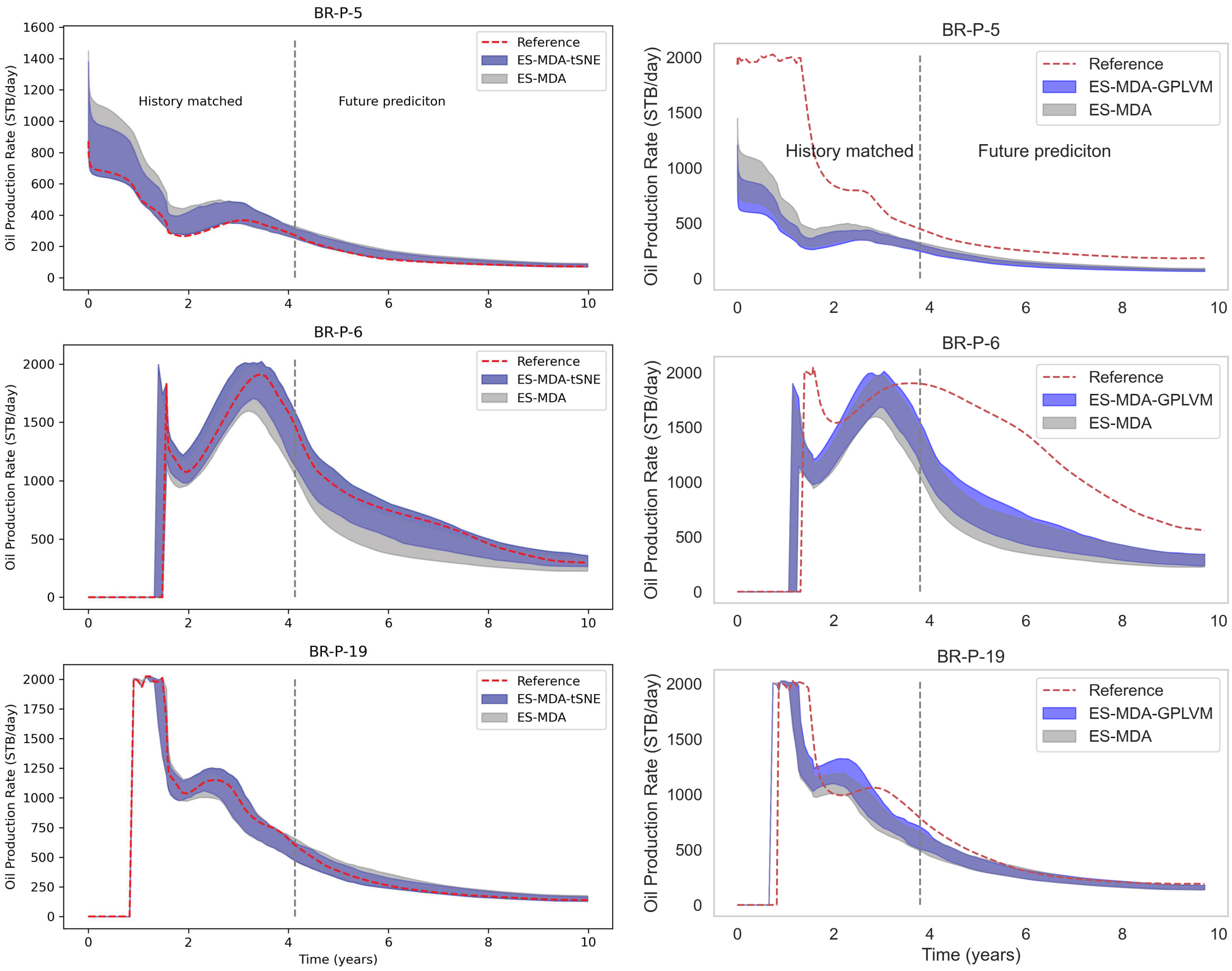


Figure 10. Oil production rate STB/day for three wells with ESMDA, ES-MDA-tSNE and ES-MDA-GPLVM. The blue dashed line is utilized as an indication of the end of historical data and the start of the prediction period. The red dashed line represents the observed data points. The grey region refers to the forecast within P10 and P90 obtained with the ES-MDA. The light blue region represents P10-P90 obtained from ES-MDA-tSNE or ES-MDA-GPLVM.

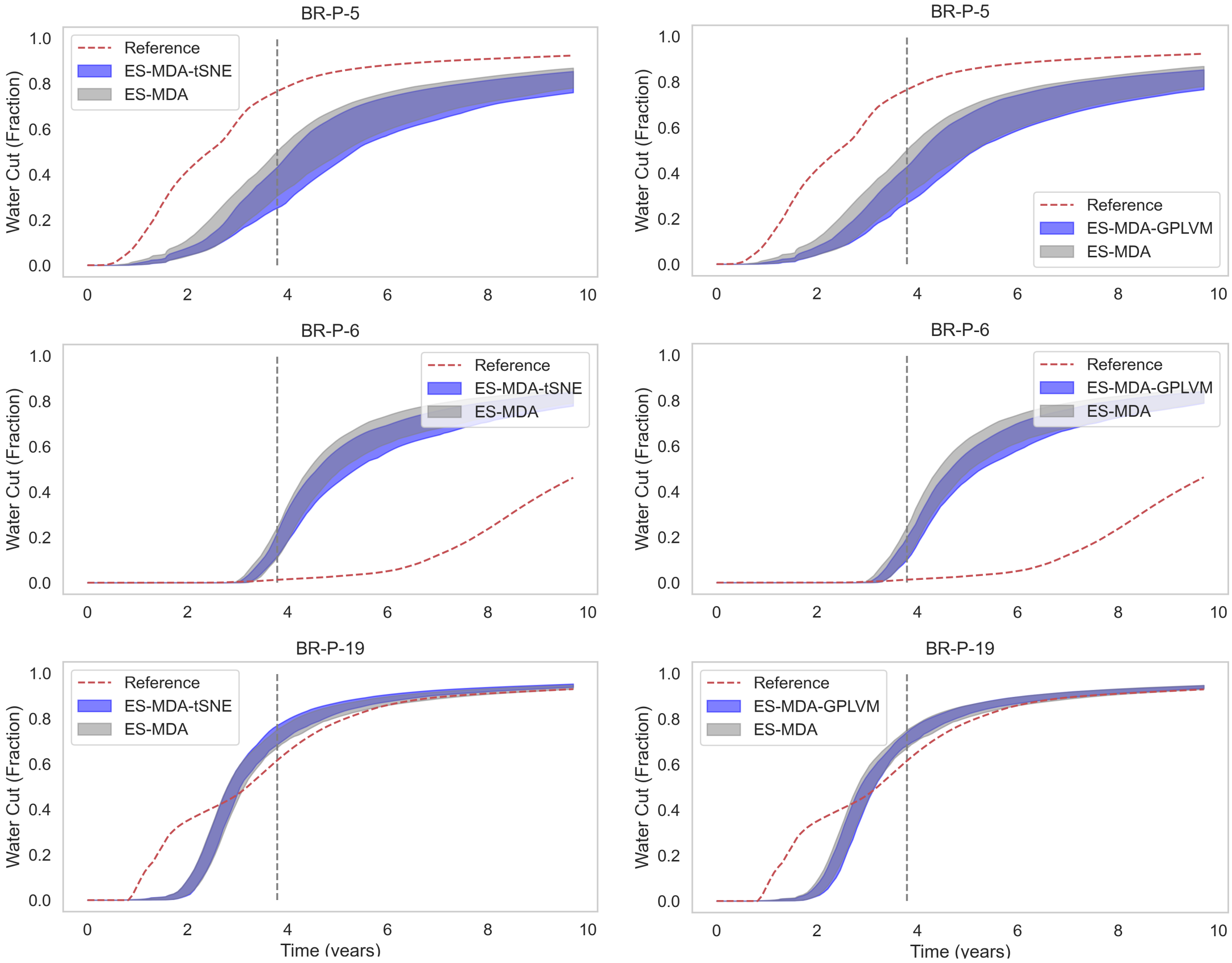


Figure 11. Water cuts for three wells with ESMDA, ES-MDA-tSNE and ES-MDA-GPLVM. The blue dashed line is utilized as an indication of the end of historical data and the start of the prediction period. The red dashed line represents the observed data points. The grey region refers to the forecast within P10 and P90 obtained with the ES-MDA. The light blue region represents P10-P90 obtained from ES-MDA-tSNE or ES-MDA-GPLVM.

For a quantitative comparison, we applied the mean CRPS metric to further evaluate the methods used at all simulated well data from the history-matched ensembles over the historical and prediction period, as displayed in Figure 12. The ES-MDA-t-SNE and ES-MDA-GPLVM provide interesting results, with the lowest CRPS average compared to the prior model, and although we used few ensembles models and saved around 45–53% of the simulation time, the results seem to be comparable to the standard ES-MDA.

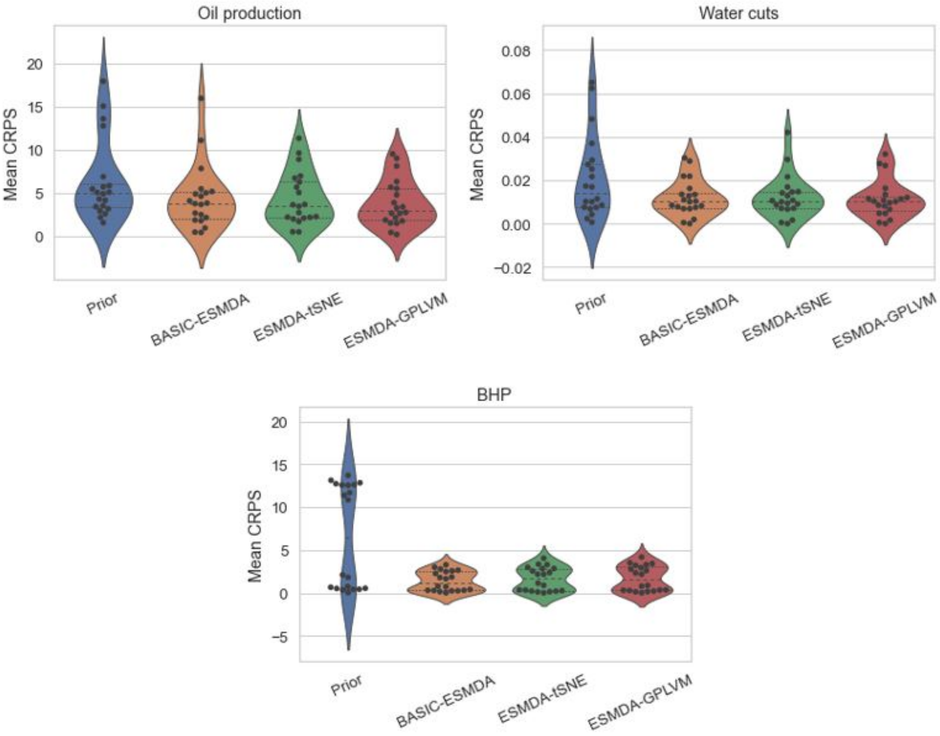


Figure 12. Violin plot of the mean CRPS of the historical and prediction data (OPR, water cuts, and BHP).

Since the uncertainty ranges are quantified in all wells at the three cases, we conducted another comparison by boxplots of the normalized field cumulative oil and water production predicted by each method, as displayed in Figure 13. It should be noted that the values are normalized to the cumulative production from the reference case, which is also added for comparison. Both ES-MDA-t-SNE and ES-MDA-GPLVM cover the true values for both oil and water production in the box ranges.

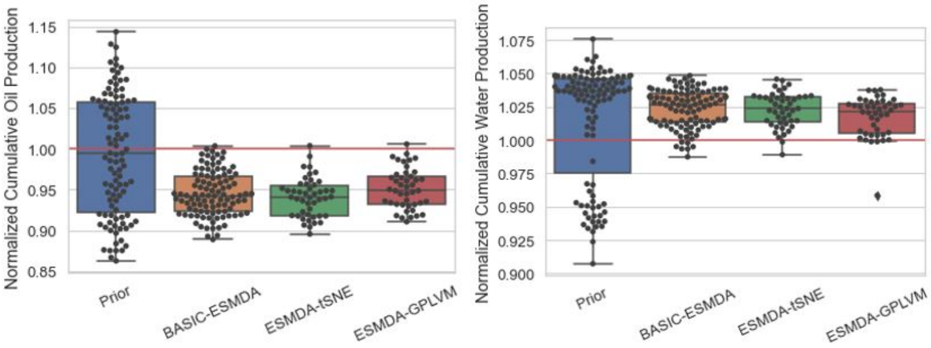


Figure 13. Boxplot of the normalized cumulative production after 10 years.

3.2. Effect of Different "Reference" Models

The previous section evaluated the ES-MDA-DR procedure on a single 'reference' model. We now evaluate the methodology on five additional 'referred' models: Test Case 1, Test Case 2, Test Case 3, Test Case 4, and Test Case 5 (we reiterate that neither reference model was included in the set of N = 103 prior models). Similarly, the data assimilations were conducted using ES-MDA with localization. Apart from the ES-MDA-t-SNE and ES-MDA-GPLVM, we performed the HM on assimilation observation data for the first 4 years and the rest (6 years) for forecasting. Note that the reference value varies considerably between the test cases, as shown in Figure 14. The cumulative distribution function (CDF) for each test case is quite different, except the test case-1 and case-3, which show some similarities.

Table 3 demonstrates the posterior normalized field cumulative oil and water production predicted by each method in terms of P10 and P90 statistics at 10 years. For four cases, the ES-MDA, ES-MDA-t-SNE, and ES-MDA-GPLVM, predictions surround the reference data within the the P10 to P90 range and showed a narrower range of uncertain prediction results. However, in Test Case 2, the cumulative oil production is biased in comparison with the reference production, which is likely explained by whether the problems with the prior ensemble or the selected reference model. The ensemble means distribution of the reconstructed updated permeability posterior for the top-layer for all five cases, as shown in Figure 15. The results suggest a slight change in the posterior model where the wells are located. Additionally, one can expect uncertainty reductions due to the conditioning of the posterior samples to dynamic data variables of wells that are contained within the forecast domain. The results exhibit quite similar posterior permeability distribution for both ES-MDA-t-SNE and ES-MDA-GPLVM. In sum, the results for the five different test cases imply that the DR procedure can indeed provide updated geological models and predictions with different reference models at a significantly reduced computation time.

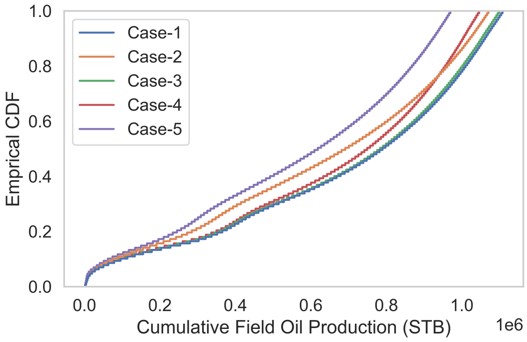
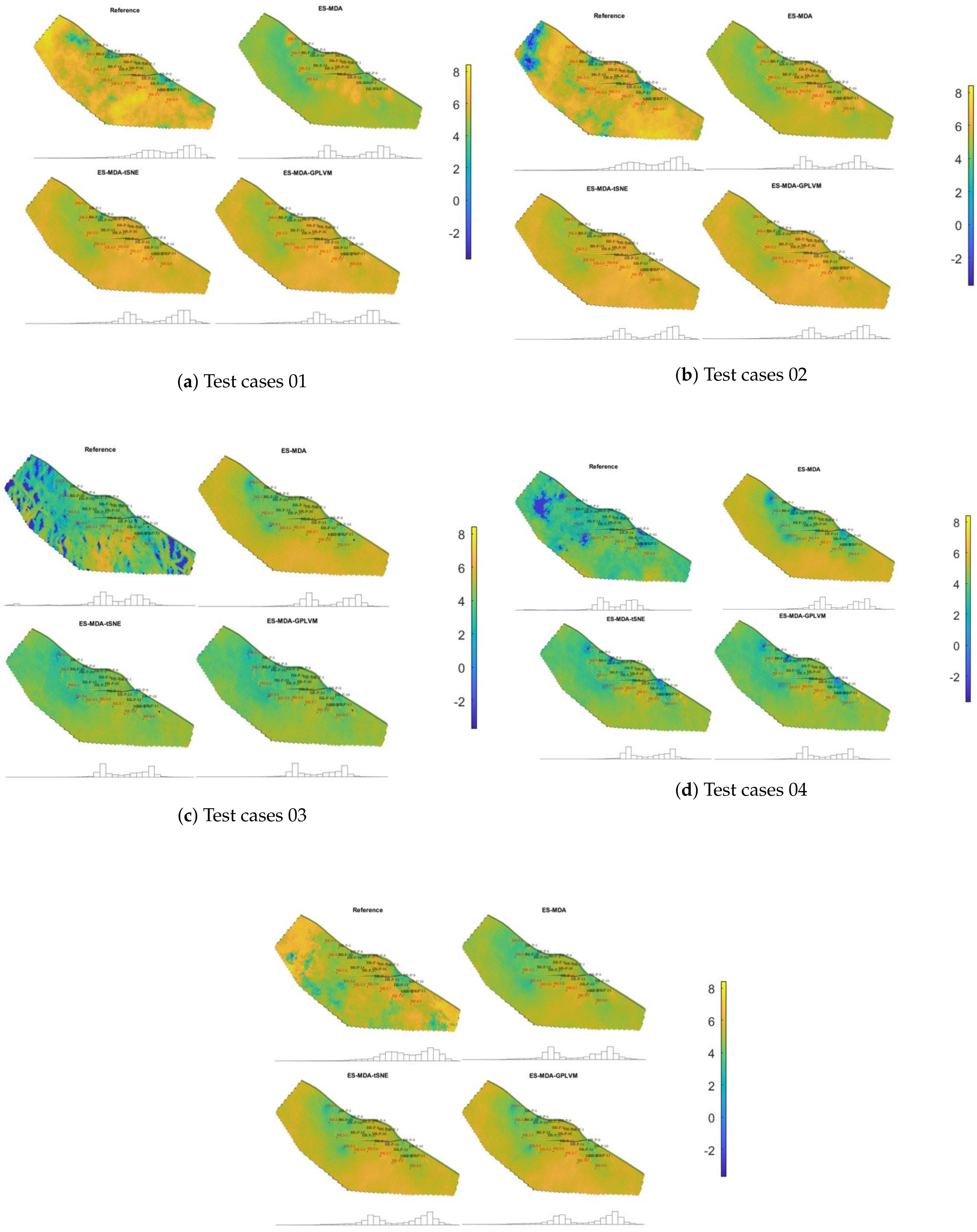


Figure 14. Empirical CDF computer from field oil production total for different test cases.

Table 3. Posterior prediction for different test cases. The P10 and P90 statistics are computed using the forecast results for field cumulative oil at 10 years.

		Reference	P10	P90
	Prior	-	9.099×10^5	1.114×10^6
Test Case 1	ES-MDA	1.108×10^6	1.045×10^6	1.112×10^6
	ES-MDA-t-SNE		1.049×10^6	1.114×10^6
	ES-MDA-GPLVM		1.053×10^6	1.124×10^6
Test Case 2	ES-MDA	1.071×10^6	9.752×10^5	1.042×10^6
	ES-MDA-t-SNE		9.764×10^5	1.013×10^6
	ES-MDA-GPLVM		9.609×10^5	1.016×10^6
Test Case 3	ES-MDA	1.099×10^6	1.037×10^6	1.116×10^6
	ES-MDA-t-SNE		1.052×10^6	1.121×10^6
	ES-MDA-GPLVM		1.050×10^6	1.111×10^6
Test Case 4	ES-MDA	1.046×10^6	1.008×10^6	1.088×10^6
	ES-MDA-t-SNE		1.012×10^6	1.081×10^6
	ES-MDA-GPLVM		1.018×10^6	1.078×10^6
Test Case 5	ES-MDA	9.698×10^5	9.361×10^5	9.993×10^5
	ES-MDA-t-SNE		9.428×10^5	9.986×10^5
	ES-MDA-GPLVM		9.340×10^5	9.867×10^5



(**a**) Test cases 01

(**b**) Test cases 02

(**c**) Test cases 03

(**d**) Test cases 04

(**e**) Test cases 05

Figure 15. Average log permeability distribution on Layer 1 using different reference models.

3.3. Effect of Reference Model Parameters Outside Prior Distribution

In the previous section, we presented the results of ES-MDA, ES-MDA-t-SNE, and ES-MDA-GPLVM for tests where every 'reference' model was within the prior distributions. This indicates that there is consistency between the prior realizations used in the three methods with the underlying 'reference model. Here, we aim to evaluate the performance of the three methods for cases that involve a reference model, which is not consistent with the prior realizations. More precisely, the reference model is characterized by parameters that fall outside the prior ranges. Figure 16 displays the permeability cumulative distribution function (CDF) between the generated reference model along with P10 and P90 prior ensemble. We observe that the 'reference' model permeability parameters for this example lie outside the range of the prior distributions. RMD outlier detection was used as displayed in Figure 17 to verify the prior uncertainty variables (field oil production) on the reference variable. The results show that the RMD of d_{obs} falls above the 95th percentile threshold, in that the prior model is falsified.

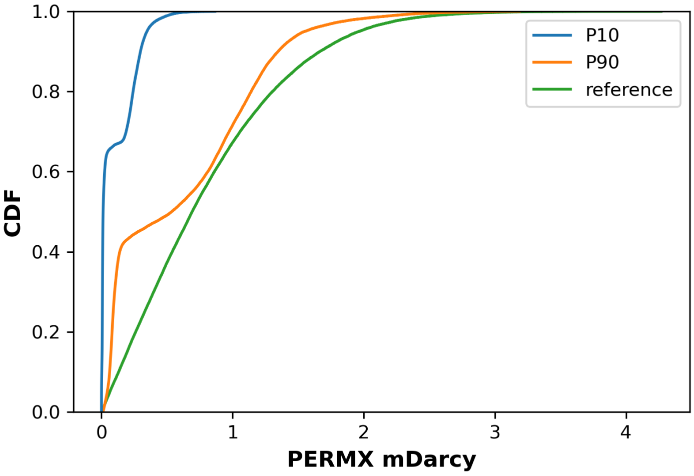


Figure 16. Empirical CDF computed from prior permeability (P10 and P90) and the reference model.

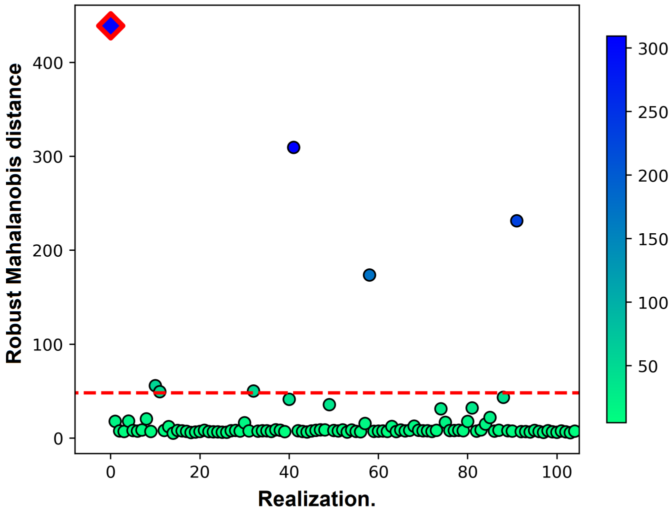


Figure 17. Prior falsification using Robust Mahalanobis distance (RMD). The red diamond is the RMD for d_{obs}. Circle dots refer to the RMD results of 104 data variable samples, and the red dashed line is the 95th percentile of the chi-squared distributed RMD.

Figures 18 and 19 depict the HM profiles for both oil and water cuts of three methods at producers BR-P5, BR-P6, and BR-P19, with respect to the prior ensemble and reference model. The results indicate that the standard ES-MDA and ES-MDA with DR failed to match the reference data, and the predictions from the referenced models do not lie within the predicted P10 to P90 percentile. Figure 20 displays a boxplot of field cumulative water and oil production obtained by each method and the prior ensemble. Overall, the results do not cover the reference values, which is clearly explained by the lack of representativeness of the prior realizations. The previous results evidently demonstrate the success of our procedure based on a degree of the quality in terms of the prior parameter ranges. We emphasize that it is crucial that the prior simulation results contain the observations. Otherwise, we would not expect ES-MDA with DR to provide reasonable posterior predictions, and in practice, we should adapt the prior ensemble before using it for model conditioning.

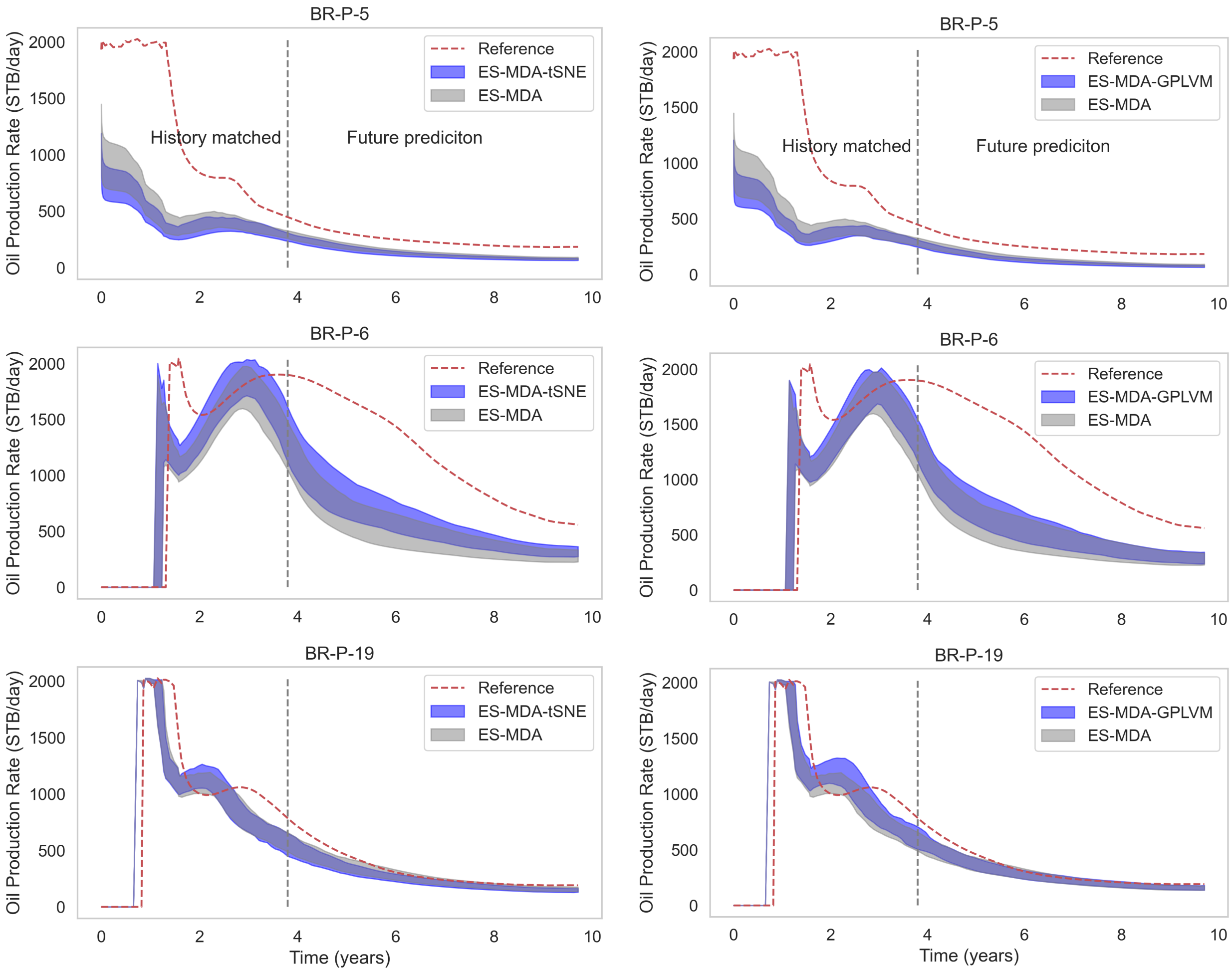


Figure 18. Oil production rate STB/day for three wells with ESMDA, ES-MDA-tSNE and ES-MDA-GPLVM. The blue dashed line is utilized as an indication of the end of historical data and start of prediction period. The red dashed line represents the observed data points. The grey region refers to the forecast within P10 and P90 obtained with the ES-MDA. The light blue section represents P10-P90 obtained from ES-MDA-tSNE or ES-MDA-GPLVM.

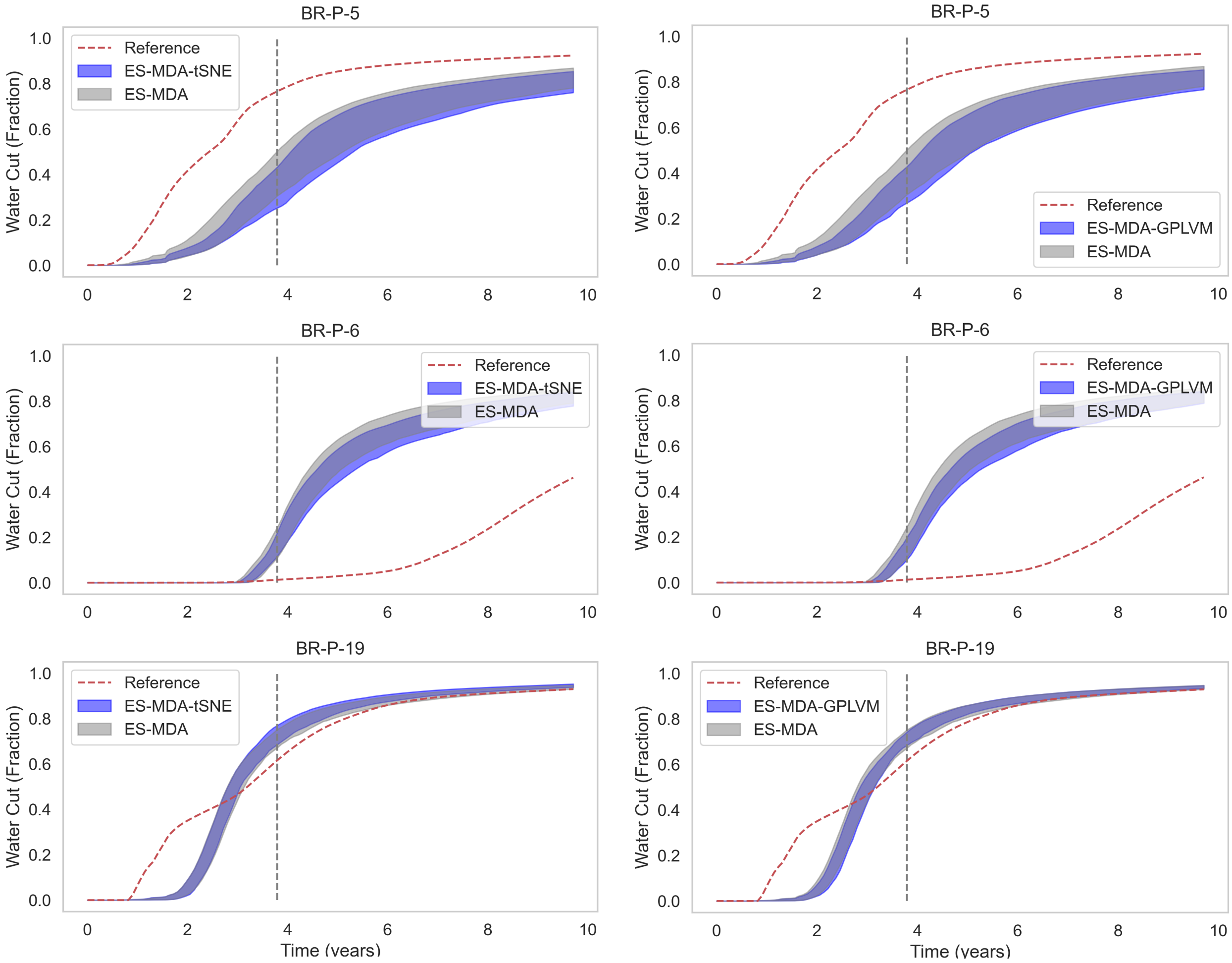


Figure 19. Water cuts for three wells with ESMDA, ES-MDA-tSNE and ES-MDA-GPLVM. The blue dashed line is utilized as an indication of the end of historical data and the start of the prediction period. The red dashed line represents the observed data points. The grey region refers to the forecast within P10 and P90 obtained with the ES-MDA. The light blue section represents P10-P90 obtained from ES-MDA-tSNE or ES-MDA-GPLVM.

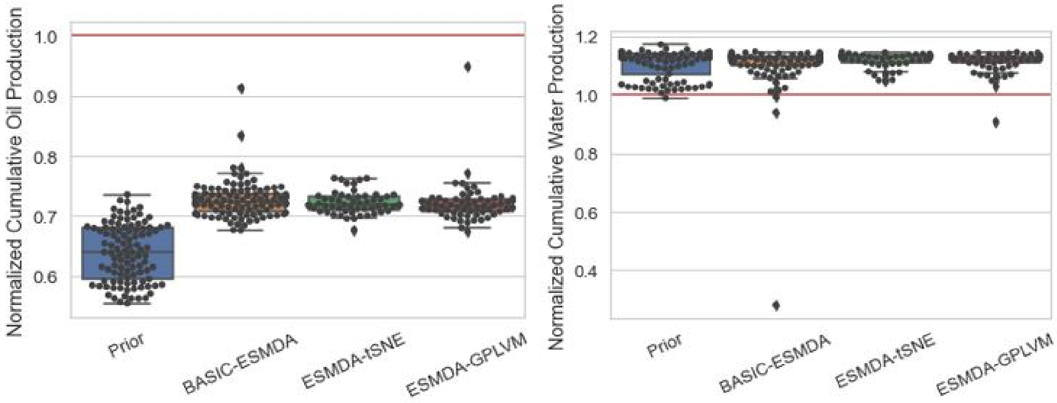


Figure 20. Boxplot of the normalized cumulative production after 10 years. The red line indicates the cumulative production of the reference case.

4. Concluding Remarks

In this study, we presented a novel history matching framework with DR while preserving realistic geology and matching the production data, which was achieved by explicitly integrating t-SNE, GPLVM, and clustering K-means with ESMDA to reduce the simulation time and quantify the uncertainty on reservoir models. The proposed procedure yielded reliable results by selecting a set of good prior ensemble reservoir models, with similar production performance to the reference model, before applying the data assimilation process. Accordingly, we compared the new implementation with the standard ES-MDA in a field reservoir problem with a large number of wells and a long production history. Based on the obtained results, the proposed ES-MDA with DR is concluded to be computationally faster than the original one, and it is very simple to implement and integrate with different types of data and models. We also evaluated our procedure with five different 'reference' models, where we observed that the ES-MDA with DR posterior predictions displayed considerably less uncertainty, and was indeed able to provide improved geological models and predictions at a significantly reduced computation time. Moreover, we also considered a test case where the reference model lay outside the prior distributions, but the results were clearly inconsistent and biased. In conclusion, the accuracy of both methods is highly relied on the ability and quality of the prior realizations to provide appropriate estimates of the prior uncertainty.

We recommend that further studies apply our procedures to more complex geological models such as bimodal channelized systems. This approach can be applied to examine and overcome the challenges in 4D seismic history matching as capturing the value of 4D seismic data can lead to better reservoir management decisions. It will also be interesting to introduce into the framework more non-linear DR techniques, such as a deep autoencoder, a stacked autoencoder, and a generative adversarial network. Additionally, combining the data-space inversion (DSI) method with ES-MDA may more accurately predict oil production with computationally faster simulation.

Author Contributions: A.T. wrote the paper and contributed to tuning the model and analyzing the results R.B.B. and R.G.H. supervised the work and providing continuous feedback. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by Petromaks-2 project DIGIRES (RCN no. 280473).

Acknowledgments: The authors acknowledge financial support from the Research Council of Norway through the Petromaks-2 project DIGIRES (RCN no. 280473) and the industrial partners AkerBP, Wintershall DEA, ENI, Petrobras, Equinor, Lundin, and Neptune Energy. The authors would like also to thank TNO for providing the Brugge reservoir model dataset.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1. t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE is a non-linear dimensionality reduction algorithm developed for exploring high-dimensional data [14]. It maps multi-dimensional data to a two- or three-dimensional dataset $Y = \{y_1, y_2, ..., y_n\}$ that can be visualized in a scatter plot. The t-SNE algorithm begins by computing a joint probability distribution pij over pairs of points x_i,$x_j (i \neq j)$:

$$p_{j|i} = \frac{exp(-\left\|x_i - x_j\right\|^2/2\tau_i^2)}{\sum_{l,s\in[n],l\neq s}(1+\|y_l - y_s\|^2)^{-1}} \quad , p_{ij} = \frac{p_{i|j}+p_{j|i}}{2n} \tag{A1}$$

where τ_i is a tunable parameter that controls the bandwidth of the Gaussian kernel around point x_i. In two-dimensional map $Y = \{y_1, y_2, ..., y_n\} \subset R^2$, the affinity q_{ij} between points y_i and y_j $(i \neq j)$ is defined as:

$$q_{ij} = \frac{(1 + \left\|y_i - y_j\right\|^2)^{-1}}{\sum_{l,s\in[n],l\neq s}(1 + \|y_l - y_s\|^2)^{-1}} \tag{A2}$$

t-SNE then attempts to find points y_i in R^2 that minimize the KL-divergence between p and q:

$$f(y_1, y_2, \ldots, y_n) := KL(p \parallel q) = \sum_{i,j \in [n], i \neq j} p_{ij} log \frac{p_{ij}}{q_{ij}} \tag{A3}$$

The objective function f is minimized using the following gradient descent:

$$\frac{\partial f}{\partial y_i} = 4 \sum_{j \in [n] \setminus \{i\}} (p_{ij} - q_{ij}) q_{ij} Z(y_i - y_j) \quad , i \in [n] \tag{A4}$$

Appendix A.2. Gaussian Process Latent Variable Model (GPLVM)

GPLVM, introduced by Lawrence (2005) [23], is a Bayesian non-parametric dimensionality reduction method that uses a Gaussian process to learn a low-dimensional Q representation of high-dimensional data D. In Gaussian process regression (GP) settings, where we are given inputs X and outputs Y, we choose a kernel and learn hyperparameters that best describe the mapping from X to Y. The GP likelihood function is written as:

$$p(Y \mid X) = \prod_{d=1}^{D} p(y_d \mid X) \tag{A5}$$

Here, y_d represents the d^{th} columns of Y and:

$$p(Y \mid X) = \mathcal{N}(y_d \mid 0, K_{NN} + \beta^{-1} I_N) \tag{A6}$$

$\mathcal{N}$ is the number of the observation, and K_{NN} is the covariance matrix defined by the covariance or kernel function $K(x, \acute{x})$. The kernel function was modified to a squared exponential form to fit the automatic model selection of the dimensionality of latent space:

$$K(x, \acute{x}) = \sigma_f^2 exp\left(\frac{1}{2} \sum_{q=1}^{Q} \alpha_q (x_q - \acute{x}_q)^2\right) \tag{A7}$$

In the GPLVM, we do not have X; we are only given Y. We need to learn X along with the kernel hyperparameters. We do not perform maximum likelihood inference on X. Instead, we set a Gaussian prior for X and learn the mean and variance of the approximate (Gaussian) posterior $p(Y \mid X)$.

$$p(X) = \prod_{n=1}^{N} \mathcal{N}(x_n \mid 0, I_Q) \tag{A8}$$

With each x_n the n^{th} row of X. The joint probability model for the GPLVM model is:

$$p(Y, X) = p(Y \mid X) p(X) \tag{A9}$$

The hyper parameters of the model are the kernel parameters $\theta == (\sigma_f^2, \alpha_1, \alpha_2, \ldots, \alpha_Q)$ and the inverse variance parameter β.

Appendix A.3. Mean Continuous Ranked Probability Score (CRPS)

Mean CRPS is used to quantify both the accuracy and precision of a probabilistic forecast [34]. A higher value of mean CRPS indicates less accurate results. CRPS can be defined as:

$$CRPS = \int_{-\infty}^{\infty} [p(x) - H(x - x_{obs})]^2 dx \tag{A10}$$

Here, $p(x) = \int_{-\infty}^{x} p(y) d_y$ is the cumulative distribution of a quantity of interest, and $H(x - x_{obs})$ is the step function,

$$H(x) = \left\{\begin{matrix} 0 & if < 0 \\ 1 & if \geqslant 0 \end{matrix}\right\} \tag{A11}$$

For N samples, the CRPS can be evaluated as follows:

$$CRPS = \sum_{i=0}^{N} c_i c_i = \alpha_i p_i^2 + \beta_i (1 - p_i)^2 \tag{A12}$$

where $p_i = P(x) = i/N, for\ x_i < x < x_{i+1}$ (piecewise constant function)

$$\alpha_i = \left\{\begin{matrix} 0 & if & x_{obs} < x_i \\ x_{obs} - x_i & if & x_i < x_{obs} < x_{i+1} \\ x_{i+1} - x_i & if & x_{obs} > x_{i+1} \end{matrix}\right\} \tag{A13}$$

$$\beta_i = \left\{\begin{matrix} x_{i+1} - x_i & if & x_{obs} < x_i \\ x_{i+1} - x_{obs} & if & x_i < x_{obs} < x_{i+1} \\ 0 & if & x_{obs} > x_{i+1} \end{matrix}\right\} \tag{A14}$$

References

1. Oliver, D.S.; Chen, Y. Recent progress on reservoir history matching: A review. *Comput. Geosci.* **2011**, *15*, 185–221. [CrossRef]
2. Jafarpour, B.; McLaughlin, D.B. History matching with an ensemble Kalman filter and discrete cosine parameterization. *Comput. Geosci.* **2008**, *12*, 227–244. [CrossRef]
3. Sambridge, M. Geophysical inversion with a neighborhood algorithm—II. Appraising the ensemble. *Geophys. J. Int.* **1999**, *138*, 727–746. [CrossRef]
4. Jin, L.; Alpak, F.O.; van den Hoek, P.; Pirmez, C.; Fehintola, T.; Tendo, F.; Olaniyan, E. A comparison of stochastic data-integration algorithms for the joint history matching of production and time-lapse-seismic data. *SPE Reserv. Eval. Eng.* **2012**, *15*, 498–512. [CrossRef]
5. Jeong, J.; Park, E. Theoretical development of the history matching method for subsurface characterizations based on simulated annealing algorithm. Journal of Petroleum science and engineering. *J. Pet. Sci. Eng.* **2019**, *180*, 545–558. [CrossRef]
6. Gao, G.; Li, G.; Reynolds, A. A Stochastic optimization algorithm for automatic history matching. *SPE J.* **2007**, *12*, 196–208. [CrossRef]
7. Barker, W.J.; Cuypers, M.; Holden, L. Quantifying uncertainty in production forecasts: Another look at the PUNQ-S3 problem. *SPE J.* **2000**, *6*, 433–441. [CrossRef]
8. Gao, G.; Zafari, M.; Reynolds, A.C. Quantifying uncertainty for the PUNQ-S3 problem in a Bayesian setting with RML and EnKF. *SPE J.* **2006**, *11*, 506–515. [CrossRef]
9. Gu, Y.; Oliver, D.S. The ensemble Kalman filter for continuous updating of reservoir simulation models. *J. Energy Resour. Technol.* **2006**, *128*, 79–87. [CrossRef]
10. Evensen, G. Sampling strategies and square root analysis schemes or the EnKF. *Ocean. Dyn.* **2004**, *54*, 539–60. [CrossRef]
11. Van der Maaten, L.; Hinton, G. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
12. Emerick, A.; Reynolds, A. Ensemble smoother with multiple data assimilation. *Comput. Geosci.* **2013**, *55*, 3–15. [CrossRef]
13. Emerick, A. Analysis of the performance of ensemble-based assimilation of production and seismic data. *J. Pet. Sci. Eng.* **2016**, *139*, 219–239. [CrossRef]
14. Vo, H.; Durlofsky, L. A new differentiable parameterization based on principal component analysis for the low-dimensional representation of complex geological models. *Math. Geosci.* **2014**, *46*, 775–308. [CrossRef]
15. Rezaie, J.; Saetrom, J.; Smorgrav, E. Reducing the Dimensionality of Geophysical Data in Conjunction with Seismic History Matching. In *74th EAGE Conference and Exhibition incorporating EUROPEC*; European Association of Geoscientists & Engineers: Houten, The Netherlands, 2012.
16. Sarma, P.; Durlofsky, L.J.; Aziz, K. Kernel principal component analysis for efficient, differentiable parameterization of multi point geostatistics. *Math. Geosci.* **2008**, *40*, 3–32. [CrossRef]
17. Muzammil, R.; Ahmed, H.E.; Yan, C. Identifiability of Model Discrepancy Parameters in History Matching. In Proceedings of the SPE Reservoir Simulation Conference, Galveston, TX, USA, 10–11 April 2019.
18. Kang, B.; Choe, J. Regeneration of initial ensembles with facies analysis for efficient history matching. *J. Energy Resour. Technol.* **2017**, *139*, 042903. [CrossRef]
19. Kang, B.; Kim, S.; Jung, H.; Choe, J.; Lee, K. Characterization of three-dimensional channel reservoirs using ensemble Kalman filter assisted by principal component analysis. *Petroleum Sci.* **2019**, *17*, 182–195. [CrossRef]
20. Tolstukhin, E.; Barrela, E.; Khrulenko, A.; Halotel, J.; Demyanov, V. Ensemble History Matching Enhanced with Data Analytics-A Brown Field Study. *Eur. Assoc. Geosci. Eng.* **2019**, *2019*, 1–5.

21. Satija, A.; Scheidt, C.; Li, L.; Caers,J. Direct Forecasting of Reservoir Performance Using Production Data without History Matching. *Comput. Geosci.* **2017**, *21*, 315–333. [CrossRef]
22. Park, J.; Caers, J. Direct forecasting of global and spatial model parameters from dynamic data. *J. Comput. Geosci.* **2020**, *143*, 104567. [CrossRef]
23. Lawrence, N.D. Learning for larger datasets with the Gaussian process latent variable model. In Proceedings of the Eleventh International Workshop on Artificial Intelligence and Statistics, San Juan, Puerto Rico, 21–24 March 2007.
24. Lawrence, N.D. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *J. Mach. Learn. Res.* **2005**, *6*, 1783–1816.
25. De Maesschalck, R.; Jouan-Rimbaud, D.; Massart, D. The Mahalanobis distance. *Cemometrics Intell. Lab. Syst.* **2000**, *50*, 1–18. [CrossRef]
26. Yin, Z.; Strebelle, S; Caers, J. Automated Monte Carlo-based Quantification and Updating of Geological Uncertainty with Borehole Data (AutoBEL v1.0). *Geosci. Model. Dev.* **2019**, *13*, 651–672. [CrossRef]
27. Liu, F.T.; Ting, K.M.; Hua, Z. Isolation forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008.
28. Breunig, M.M.; Kriegel, H.P.; Ng, R.T.; Sander, J. LOF: Identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, 16–18 May 2000.
29. Schölkopf, B.; Williamson, R.C.; Smola, A.J.; Shawe-Taylor, J.; Platt, J.C. Support vector method for novelty detection. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 29 November–4 December 1999.
30. Jain, A.K. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **2010**, *31*, 651–666. [CrossRef]
31. Watanabe, S.; Datta-Gupta, A. Use of phase streamlines of covariance localization in Ensemble Kalman Filter for three-phase history matching. In Proceedings of the SPE Western North American Region Meeting, Anchorage, AK, USA, 7–11 May 2011.
32. Gaspari, G.; Cohn, S. Construction of correlation functions in two and three dimensions. *Q. J. R. Meteorol. Soc.* **1999**, 723–757. [CrossRef]
33. Peters, E.; Arsts, R.; Brouwer, G.; Geel, C.R.; Cullick, S.; Lorentzen, R.J.; Chen, Y.; Dunlop, N.B.; Vossepoel, C.; Xu, R.; et al. Brugge Brenchmard study for flooding Optimiztion and History Matching. *SPE Reserv. Eval. Eng.* **2010**, *13*, 391–405. [CrossRef]
34. Hersbach, H. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather. Foracsting* **2000**, *15*, 559–570. [CrossRef]