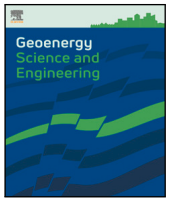
227 (2023) 211929

Contents lists available at ScienceDirect

Geoenergy Science and Engineering

journal homepage: www.sciencedirect.com/journal/geoenergy-science-and-engineering



Quantifying prior model complexity for subsurface reservoir models

Tanteliniaina N. Mioratina [a,b,*], Dean S. Oliver [a]

[a] NORCE Norwegian Research Centre, Norway
[b] University of Bergen, Norway

ARTICLE INFO

Keywords:
Model complexity
Model predictability
Predictive accuracy
Model selection
Prior model
History matching

ABSTRACT

In Bayesian approaches to history matching for subsurface inference, the prior model specifies the uncertain model parameters and the joint probability of those parameters before incorporating production-related data. A good prior model is generally complex enough to capture the future reservoir behavior in the long term, realistic enough to be plausible, consistent with geologic knowledge, and simple enough to allow calibration for data matching. Model complexity is often associated with the number of model parameters, thus the focus on finding the sufficient number of parameters needed for history matching and quantifying future uncertainty.

This work explores model choice based on concepts of complexity and informativeness of models for subsurface reservoir models. It focuses on the effect of the misspecification of prior models for assimilating flow data and their predictive accuracy. The concept of the effective number of parameters is used to investigate the suitability of various types of prior models with different levels of complexity, ranging from a highly simplified polynomial trend model to a more realistic multipoint statistical model(MPS) and a family of isotropic Gaussian models and explore the effect of level of model complexity on the robustness of forecasting. The numerical experiments were performed with different combinations of data type, prior informativeness, forecast type, and model type to compare the effect of different prior models on the robustness of the results. The effective number of parameters was computed for each prior model and their accuracy for predicting future reservoir behavior was analyzed.

The results suggest that effective model dimension is a useful measure of model complexity for history matching problems, although it is not independent of the data used for model calibration and the number of effective model parameters is generally much smaller than the number of model parameters. In a data-rich problem, realism of a model is much less important than the complexity of a model, while for a problem with few data, realism was beneficial for reliable forecasts.

1. Introduction

The use of the Bayesian paradigm in the field of inverse problems is well established (Tarantola, 1987; Stuart, 2010) and it has proven its robustness in various applications, including reservoir history matching (Oliver and Chen, 2011; Oliver et al., 2021; Evensen et al., 2022). Despite its efficiency, Bayesian history matching inherits the subjective nature of Bayesian analysis, especially in the specification of prior uncertainty and prior model selection. The *subjective prior* refers to the prior distribution for model parameters based on experts' prior knowledge and beliefs. As these may be personal beliefs, each expert may choose a different prior model and reach different conclusions. On the other hand, objective prior distributions are designed to be minimally informative and remove subjectivity. The relative benefits of objectivity and subjectivity in the specification of prior distribution remain controversial. Gelman et al. (1995), Gelman and Hennig (2017), Berger (1985), Oakley and O'Hagan (2004) and Simpson et al. (2017) present extended discussions of subjective and objective priors, and their effects on statistical analysis. The subjective nature of Bayesian analysis allows considerable flexibility in the choice of the prior model, but (O'Hagan, 2013) cautions on possible pitfalls in the specification of subjective prior probability, including the possibility of specifying logically inconsistent priors. For history matching, the main issue, however, is probably not the choice of objective versus subjective prior, as the usefulness of incorporating prior geologic knowledge in history matching is almost universally accepted; it is the choice of parameters to include in the uncertainty model. Neglecting essential parameters often results in biased and overconfident predictions (Gelman et al., 2014; Vink et al., 2015; Oliver and Alfonzo, 2018).

In the context of reservoir modeling, the prior describes the confidence in the model parameters based on prior information and general

* Corresponding author at: NORCE Norwegian Research Centre, Norway.
E-mail address: tmio@norceresearch.no (T.N. Mioratina).

https://doi.org/10.1016/j.geoen.2023.211929
Received 12 January 2023; Received in revised form 16 May 2023; Accepted 17 May 2023
Available online 22 May 2023

scientific knowledge. Prior models are usually defined without reference to data before history matching; the choice of model parameters used is subject to practitioners' best knowledge and beliefs. This results in many possible prior scenarios and considerable uncertainty in geological interpretations. Specification of prior models involves, among other things, constraints on model parameters and joint constraints on the spatial distribution of parameter values, reflecting the need for geological realism and consistency. The model parameterizations for history matching are often selected based on their simplicity and parsimony (Dake, 2001; Williams et al., 2004). However, this approach runs the risk of underestimating uncertainty in parameters that could be relevant in understanding future reservoir behavior but might not necessarily be required to match historical data (Hunt et al., 2007). On the other hand, geologists typically value realism in their subsurface characterization. Some estimates of the reservoir parameters might not be acceptable if they do not fit into their notion of what the true subsurface should be like. The concept of realism in geological modeling is reviewed by Linde et al. (2015) and discussed by Heße et al. (2019). Realistic models tend to be computationally expensive to simulate and difficult to fit into consistent probability models. Reality is complex, however, so model complexity appears to be needed to accurately represent all the aspects of reality (Oreskes, 2003). Draper (1995) suggests the model should be "as big as a house"; that is, it should incorporate all the parameters needed so that model will be consistent with observations, will remain consistent with future data, and be able to forecast future behavior.

The complexity of a model can be characterized in many ways, including the interpretability of parameters (Okiria et al., 2022), computational time or cost of model simulation, the ability to be conditioned to observed data, the entropy (or differential entropy) of the model, and the types of physical processes that are included in the forward simulation. In statistical literature, the complexity of a model is sometimes defined by the effective number of parameters or degrees of freedom. In this case, the complexity of the prior model is a function of the data that must be assimilated. Parameters of the prior model that do not influence the value of predicted data are not relevant to this definition of complexity. A more complex model has greater flexibility and generally results in a better fit to data.

Hansen (2021) developed a methodology for efficiently computing the entropy of a prior model using products of sequential simulation. The approach provides a measure of the effective number of parameters that can be used for calibration to point measurements of the parameter field but does not provide an estimate of the effective number of parameters that are available for the calibration of non-local data. Although entropy provides a measure of complexity, entropy may not be useful to quantify the complexity of models for history matching because entropy fails to account for the type of data that must be assimilated. The complexity of a model, when defined in terms of the number of degrees of freedom, might be different for the assimilation of flow data or well-log data. Thus, measuring the model's complexity should consider both the observations and the prior information (Spiegelhalter et al., 2002).

One of the main reasons that the complexity of a model is of interest in history matching and reservoir data assimilation is that it affects predictability. If the complexity of the model is taken to be the number of degrees of freedom available for history matching, one might worry that too many adjustable parameters would result in overfitting and poor predictability of the model. On the other hand, too few adjustable parameters would result in underfitting and, again, poor predictability.

The concept of predictability is central to the investigation of complexity. The goal is not necessarily to provide an objective approach to choosing the best prior model but rather to provide guidance on the types of prior models that will have high predictability for various types and amounts of data. It is expected that the choice of types of models will depend on a number of factors, including the quantity of data that is available and the quantity that is to be predicted. In history matching, the quality of a prior model is often characterized by its ability to assimilate available production data. However, while the quality of the match to available data is important as a predictor of the quality of the match to unseen data, it is almost always optimistic as a measure of predictability.

This paper is organized as follows: Section 2 presents an overview of the metrics used to evaluate the predictability and complexity of a given model. Log pointwise predictive accuracy is used to measure the ability of the model to predict data that were used to history match the model. A probabilistic scoring function is used to compare probabilistic future forecasts. The deviance information criterion (DIC) and the effective number of parameters are applied to quantify model complexity and serve as metrics to compare candidate models.

Section 3 illustrates key concepts of model complexity with two numerical examples. The first example is a 1D Gauss-linear inverse problem, where an extensive comparison of various measures of model complexity and predictability is presented. Undertaking such a comprehensive comparison is feasible with this test case as it is not limited by computational expense. The second numerical example investigates the effect of model complexity on the ability to assimilate production data and provide usable forecasts with a 2-dimensional, 2-phase porous media flow study, where the effect of the misspecification of prior models in a history-matching exercise is examined. This investigation is performed with three different prior models: a simple polynomial model, a family of weakly informative Gaussian models with different correlation lengths, and a more geologically "realistic" multi-point statistics (MPS) model (Guardiano and Srivastava, 1993; Strebelle, 2002; Mariethoz et al., 2010). In this study, an MPS model defined by a training image (TI) is used to portray the high-connectivity channel features that are present in the data-generating model. Furthermore, this example uses a combination of data sets of varying lengths with short and long periods of historical production data to give an insight into the impact of the data availability in forecasting. Each competing model's complexity is assessed numerically, and their predictive accuracy in the immediate future and in long-term forecasts are evaluated.

This paper uses two history-matching techniques to obtain posterior samples. Levenberg–Marquard iterative ensemble smoother (LM-IES) (Chen and Oliver, 2013) is applied to history match the polynomial trend and Gaussian processes models. History matching an MPS model is not feasible with the IES method, a Markov chain Monte Carlo (McMC) extended Metropolis algorithm (Hansen et al., 2013) is used to match the MPS model. The overall findings are presented in Section 5, followed by discussion and concluding remarks.

2. Methodology

2.1. Quantifying model predictability

History matching aims not to estimate the model's parameters but to predict future reservoir behavior or reservoir behavior with a different set of controls. The methodologies for quantifying the quality of a probabilistic forecast are briefly presented in the following section.

2.2. Ability to match unseen data

Let y denote the data used for history matching. For predictability of an unseen dataset $\tilde{y}_i$ after history matching, the posterior predictive distribution $p(\tilde{y}_i|y)$ should be as close to the distribution for the true data-generating process $p_t(\tilde{y}_i)$ as possible. The Kullback–Leibler divergence is a useful measure of the difference between those two distributions

$$\begin{aligned} D_{KL}(p_t\|p) &= \int p_t(\tilde{y}_i)\log\frac{p_t(\tilde{y}_i)}{p(\tilde{y}_i|y)}\,d\tilde{y}_i \\ &= \int p_t(\tilde{y}_i)\log p_t(\tilde{y}_i)\,d\tilde{y}_i - \int p_t(\tilde{y}_i)\log p(\tilde{y}_i|y)\,d\tilde{y}_i. \end{aligned} \tag{1}$$

The KL-divergence will be minimized when $p(\tilde{y}_i|y) = p_t(\tilde{y}_i)$. The first term on the right side of (1) can be ignored as it only involves the true

distribution and is thus the same for all models. Consequently, using the terminology of Vehtari et al. (2017), the *expected log pointwise predictive density* for a new dataset, is

$$\text{elpd} = \sum_{i=1}^{n} \int p_t(\tilde{y}_i) \log p(\tilde{y}_i|y)\, d\tilde{y}_i \tag{2}$$

where $p_t(\tilde{y}_i)$ is the distribution representing the true data-generating process for $\tilde{y}_i$. Unfortunately, while $\log p(\tilde{y}_i|y)$ can be relatively easily evaluated from the posterior ensemble of reservoir models (θ) conditioned to data y, i.e.

$$p(\tilde{y}_i|y) = \int p(\tilde{y}_i|\theta) p(\theta|y)\, d\theta$$

the distribution of the true unseen data $p_t(\tilde{y}_i)$ is unknown.

Something that *can* be computed is the ability of the model to predict the data that were used to history match the model. In this case, the measure of predictability is derived from the posterior distribution of model parameters conditioned to data, $p(\theta|y)$. For a problem with independent observations errors, the *log pointwise predictive density* (Gelman et al., 2014) of the history-matched model to the data that were used for history matching is

$$\begin{aligned} \text{lppd} &= \log \prod_{i=1}^{n} p(y_i|y) \\ &= \sum_{i=1}^{n} \log \int p(y_i|\theta) p(\theta|y)\, d\theta \end{aligned} \tag{3}$$

which can be estimated easily using draws from the posterior distribution, $p(\theta|y)$,

$$\widehat{\text{lppd}} = \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta_s) \right). \tag{4}$$

The log pointwise predictive density (4) does not always provide a reasonable estimate of predictability of data that were not used for history matching (2) because its ability to match data generally becomes better as more degrees of freedom are added. On the other hand, the predictability for unseen data may decrease when the complexity of the model becomes higher than optimal. The difference between the estimate of predictability based on lppd (4) and an estimate based on elpd (2) is sometimes called the *optimism* in the machine learning literature (Hastie et al., 2009).

If lppd (3) is used as a measure of the predictability of out-of-sample data, then it needs to be corrected for the optimism. The goal is to estimate the expected log pointwise predictive density (elpd) for a new dataset. Various methods of approximating this quantity are evaluated in this study and the differences are used to quantify the effective number of parameters in the model. Of course, the model may have many times more uncertain parameters (millions in many reservoir models), but the magnitude of most parameters will generally be determined by the prior model distribution (Spiegelhalter et al., 2002). The approach taken is to identify p as the difference between the computation of lppd for observed data and an estimate of the elppd for unseen data (Gelman et al., 2014). Several potential methods for approximating elppd are investigated here. Note, however, that all of the methods considered are only capable of approximating the predictability of unseen data that is similar to data that has been used for history matching. The most straightforward approach to estimating the predictability of unseen data is through cross-validation, which allows one to evaluate directly the predictability of data that were not used for calibration by using subsets of the data for training and for validation. In the *leave-one-out* (loo) formulation of cross-validation, the data set is repeatedly partitioned into a training set that consists of all data points except the ith, and the held-out data y_i which is then used to evaluate predictability. The sum of the individual log pointwise predictive densities provides an estimate of the elpd (Gelman et al., 2014; Vehtari et al., 2017; Gronau and Wagenmakers, 2019):

$$\text{lppd}_{\text{loo-cv}} = \sum_{i=1}^{n} \log p(y_i|y_{(-i)}), \tag{5}$$

calculated as

$$\sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta^{is}) \right) \tag{6}$$

where θ^{is} denotes the s realization of the posterior, trained without the ith observation.

Burman (1989) points out that the predictive fit from standard loo cross-validation is biased because the lppd (4) is computed for calibration to n data while the lppd for unseen data (5) is computed from calibration to $n-1$ data. The difference is negligible for large n, but a correction might be justified for small n (or when using k-fold cross-validation). Using a first-order bias correction the following estimate is obtained

$$\overline{\text{lppd}}_{-i} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \log p(y_j|y_{(-i)}),$$

which is calculated as

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_j|\theta^{is}) \right). \tag{7}$$

The bias correction is rarely used as it is usually small, but it is included in the computations of numerical results for completeness. To make comparisons to other methods, an estimate of the effective number of parameters based on leave-one-out cross-validation is

$$p_{\text{loo-cv}} = \overline{\text{lppd}}_{-i} - \text{lppd}_{\text{loo-cv}}. \tag{8}$$

where $\overline{\text{lppd}}_{-i}$ is computed from (7) and $\text{lppd}_{\text{loo-cv}}$ is computed using (6). As an estimator of the predictability of unseen data, loo cross-validation is robust but too expensive to use for large-scale, expensive data assimilation problems as it requires the model to be recalibrated N_d times. Consequently, several approximations to loo-cv are evaluated, including p_{DIC} (Spiegelhalter et al., 2002), $p_{\text{psis-loo}}$ (Vehtari et al., 2017) and the effective dimension (Agapiou et al., 2017).

Instead of using the expensive loo-cv approach to estimating the predictability of unseen data, one could use an importance sampling approach to estimating $\text{lppd}_{\text{loo-cv}}$. Gelfand et al. (1992) noted that (5) can be written, without approximation, as

$$\int p(y_i|\theta)\, p(\theta|y_{-i})\, d\theta = \int p(y_i|\theta)\, \frac{p(\theta|y_{-i})}{p(\theta|y)}\, p(\theta|y)\, d\theta$$

and that the importance weight on a sample θ^s from $p(\theta|y)$ is

$$w_i^s = \frac{p(\theta|y_{-i})}{p(\theta|y)} = \frac{1}{p(y_i|\theta)} \tag{9}$$

which is the same as (6) in Vehtari et al. (2017). The importance sampling approximation of the predicted data at the ith location given the data at all other locations using the Monte Carlo approximation is thus

$$p(\tilde{y}_i|y_{-i}) \approx \frac{\sum_{s=1}^{S} w_i^s p(\tilde{y}_i|\theta^s)}{\sum_{s=1}^{S} w_i^s}.$$

If this expression is evaluated at locations of observed data, $\tilde{y}_i = y_i$, then $w_i^s p(y_i|\theta^s) = 1$ can be simplified and consequently an importance sampling estimate is obtained,

$$\text{lppd}_{\text{is-loo}} = \frac{\sum_{s=1}^{S} 1}{\sum_{s=1}^{S} \frac{1}{p(y_i|\theta^s)}} = \frac{1}{\frac{1}{S} \sum_{s=1}^{S} \frac{1}{p(y_i|\theta^s)}}. \tag{10}$$

Vehtari et al. (2017) point out that the importance sampling estimate (10) is often sensitive to the magnitude of weights on a few samples. They suggest that the loo estimate can be improved using Pareto smoothing of the smallest weights. For the computation of Pareto smoothed importance sampling weights, the 'psisloo' routine (Vehtari, 2018) is used, which takes as input the ensemble of log-likelihood values for each of the data. It then fits the Pareto distribution and outputs revised weights. A measure of the effective number of parameters for

the problem can then be obtained as the difference between lppd (3) and $\text{lppd}_{\text{psis-loo}}$:

$$p_{\text{psis-loo}} = \text{lppd} - \text{lppd}_{\text{psis-loo}} \tag{11}$$

The Pareto smoothed importance sampling approach is relatively inexpensive as it only requires the posteriori ensemble from a single calibration to all data. The approach's applicability is limited, however, by the magnitude of the variability of the log-likelihood for a single datum. The method works best with highly informative priors.

2.3. Ability to extrapolate (scoring)

Using a scoring rule is the best way to evaluate the quality of the predictive performance of a model. This paper focuses on the Logarithmic score (Good, 1952), which is defined by (12) to compare the predictability of different models,

$$\text{LogS}(F, y^o) = -\log(f(y^o)) \tag{12}$$

where f is the estimated probability density function pdf of the forecast, which is modeled as a Gaussian distribution, and y^o is the observed forecast.

The steps to compute a score for the quantity of interest x are defined as follows

- Generate the ensemble of forecasted values.
- Fit a Gaussian pdf to the ensemble of predicted values from the mean μ and the standard deviation σ of the QoI, x from the ensemble:

$$p(x|d, H) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

- Compute the logarithmic score for the true QoI, x^{tr}:

$$\text{score}(x^{tr}, d, H) = -\log(\sigma\sqrt{2\pi}) - \frac{1}{2}\frac{(x^{tr}-\mu)^2}{\sigma^2}$$

Given different forecaster models, the models with the higher score over multiple forecasts have higher predictive performance. The score is affected by both the quality of the calibration (the second term) and the sharpness of the prediction (the first term). Hence for two predictions with the same calibration quality, the one with the narrower pdf obtains the higher score.

2.4. Quantifying model complexity

Providing a general definition of model complexity is a challenge since the complexity of a model can be described and assessed in many different ways. This study is aligned with Van der Linde (2012) in defining model complexity as the ability of the parameters to explain the observations and the ability of the observations to determine the parameters. Therefore, measuring model complexity implies measuring the dependence between the observations and parameters.

A measure of the effective number of parameters in a model can be derived based on information theoretic arguments. The measure defined by Spiegelhalter et al. (2002) is particularly useful in data assimilation as it is easy to compute from an ensemble of posterior realizations. Because this measure is used in the DIC to penalize model complexity, it will be referred as p_{DIC}. It is obtained from (2 times) the difference between the log-likelihood evaluated at the mean posterior model and the mean of the log-likelihoods evaluated at the ensemble of draws from the posterior.

$$p_{\text{DIC}} = 2\big(\log p(y|E_{\text{post}}\theta) - E_{\text{post}} \log p(y|\theta)\big), \tag{13}$$

where the expectations are with respect to the posterior distribution of θ. The Monte Carlo computation of p_{DIC} uses simulations $\theta^s, s = 1, \ldots, S$ from the posterior as,

$$p_{\text{DIC}} = 2\left(\log p(y|\hat{\theta}_{\text{post}}) - \frac{1}{S}\sum_{s=1}^{S} \log p(y|\theta^s)\right). \tag{14}$$

Although p_{DIC} is relatively stable, negative estimates can be produced when the posterior mean is far from the mode.

When the posterior distribution is Gaussian, the effective number of parameters can be computed directly from the sensitivity, G, of data to model parameters, the observation error covariance matrix C_d and the posterior covariance for model parameters $C_{m'}$:

$$\begin{aligned} p_{DIC-G} &= \text{Tr}(G^{\text{T}} {C_d}^{-1} G C_{m'}) \\ &= \text{Tr}(C_{m'}^{1/2} G^{\text{T}} {C_d}^{-1} G C_{m'}^{1/2}) \end{aligned} \tag{15}$$

Agapiou et al. (2017) derives an equivalent measure which they interpret as the effective dimension of the Bayesian linear model. In an ensemble Kalman approach to data assimilation, p_{DIC-G} is trivial to approximate from the ensemble of updated predicted data. This quantity will be referred to as the effective dimension (efd), although it is identical to p_{DIC-G},

$$\text{efd} = \text{Tr}({C_d}^{-1} \Delta d (\Delta d)^{\text{T}}/(n_e - 1)). \tag{16}$$

Although the dimension of C_d is often large in geoscience inverse problems, the efd can be efficiently computed from the singular values of $C_d^{-1/2}\Delta d/\sqrt{n_e - 1}$ where Δd is the matrix whose columns are realizations of predicted data after subtraction of the ensemble means.

3. Numerical example: 1D Gauss-linear inverse problem

In this example, a Gaussian random variable is defined on a one-dimensional lattice of length 1 that has been discretized into 150 segments. The data are generated from a Gaussian process with a Gaussian (squared exponential) covariance with a practical range of 0.2. Observations are generated by adding independent zero-mean noise with a standard deviation 0.2 to the true data. Fig. 1 shows the data-generating model (solid curve) and the noisy observations (square dots) for the case in which half of the lattice points have been observed. This example has been chosen because it is possible to analyze results from several methods of quantifying model complexity and model predictability thoroughly, including cross-validation, which is not possible for more expensive models.

For this simple problem, the goal is to evaluate methods of computing model complexity for a variety of prior models. The choice of a prior model is restricted to Gaussian processes with squared exponential covariance type and known variance. In this case, it is possible to investigate model complexity as a function only of the range of the prior model covariance. For each selected correlation length, the log pointwise predictive densities (7) and a measure of the ability to predict unseen data (8) are evaluated. The *expected* log pointwise predictive density (8) is used to evaluate the ability to predict unseen (but similar) data—it avoids the optimism inherent in simply choosing the model that best fits the data. Fig. 2 shows the lppd and the lppd-loo-cv values as functions of the correlation length. As expected, the lppd increases monotonically as the correlation length in the prior covariance becomes shorter, indicating an increased ability to match data as the number of degrees of freedom increases. On the other hand, the predictability of data at locations of held-out data (orange curve) peaks when the correct correlation range is used for data assimilation, i.e., the correlation length from the data-generating model (shown as the vertical red line in Fig. 2). The difference between the blue and orange curves is a measure of model complexity. The models with very short correlation ranges have high model complexity and have over-fit the data.

The complexity of a model does not depend only on the choice of the prior; it also depends on the type and number of observations. In the previous computations, the number of observations was fixed at 75. In Fig. 3, realizations from the posterior predictive distributions are shown for correlation lengths of 0.02, 0.2, and 2.0 and numbers of observations varying from 2 to 75. In general, one can conclude that the prior model with long correlation length and many data (Fig. 3(i)) is insufficiently complex to predict data accurately. Similarly, one could

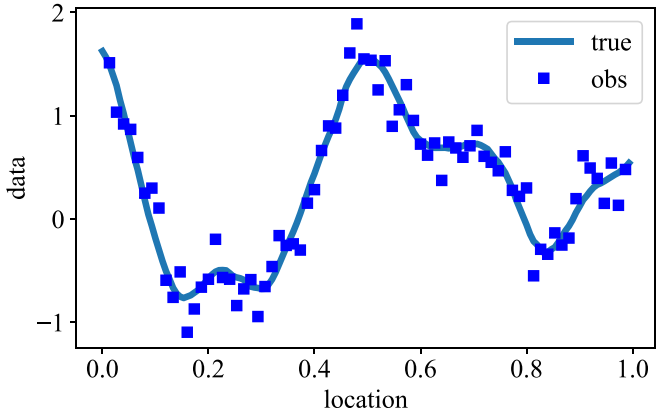


Fig. 1. The data-generating model (solid line) and observations (squares) for the one-dimensional linear-Gaussian inverse problem.

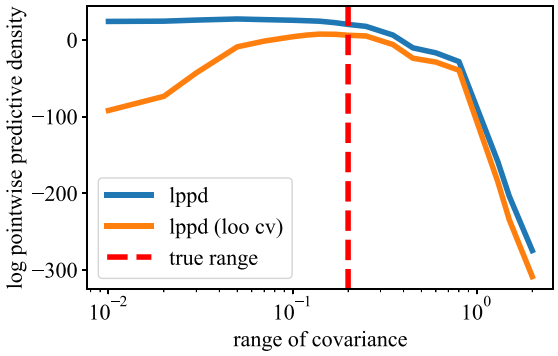


Fig. 2. The log pointwise predictive density (lppd) and the expected log pointwise predictive density for unseen data, computed using leave-one-out cross-validation as a function of correlation range in the trial models.

conclude that the model with a very short range (Fig. 3(a)) is overfitting the data and hence poor at predicting unseen data. However, the "complexity" of each model/data system is not immediately apparent from these figures.

Fig. 4 compares four different methods of estimating model complexity as functions of different amounts of data assimilated and different correlation ranges in the prior covariance. The leave-one-out cross-validation estimate of predictability (upper left) is known to be approximately unbiased. However, it has high variance because the data sets used for model calibration are very similar (Hastie et al., 2009). The p_{DIC} estimate of model complexity (14) and the efd estimate (16) are essentially the same for this example because the posteriori distribution is Gaussian. In this case, the slight differences are due to the details of the computation. Although the psis-loo approach (upper right corner) is intended to approximate the loo-cv approach at a much-reduced cost, all of the examples appear to be outside the range of applicability. In some cases, it might be feasible to augment psis-loo results with a number of loo-cv results at increased cost, but this approach is not pursued as it seems unlikely to be practical for flow problems.

The effect of the number of observations on model complexity is shown in Fig. 5. The information is similar to the information in Fig. 4, but the dependence of the effective number of parameters is shown in Fig. 5 as a function of the number of observations for two fixed correlation ranges. In the case for which the correlation range is very small ($\rho_{\mathrm{trial}} = 0.01$), the effective number of parameters measured by either p_{DIC} or efd (black dots in Fig. 5(c)) is identical to the actual number of observations. When the correlation range for the model covariance is large, the effective number of parameters is limited (red dots in Fig. 5(c)). The other measures of the effective number of parameters, loo-cv (Fig. 5(a)) and psis-loo (Fig. 5(b)) show similar trends but with different magnitudes.

4. Numerical experiment: 2D porous flow problem

A second 2D porous flow test problem, with similarity to realistic history-matching problems, illustrates the effect of prior model complexity on the ability to history-match data and forecast future behavior. Given a true data-generating model, this numerical experiment investigates the complexity of three prior model families with respect to the flow observations from the data-generating model. In this investigation, different possible scenarios in constructing the prior models are considered, emphasizing weakly informative priors with respect to the observations. The aim is to analyze the relationship between the length of historical data, the quality of history matches, and the forecast quality.

The prior predictive performance of each prior model is assessed by comparing predictions with observations before history matching. This provides an indication of the validity of the model. The ability of the model to be calibrated to real observations is subsequently evaluated, as models with insufficient degrees of freedom are incapable of being history matched. Then, the complexity of each prior model is estimated by computing the predictability of held-out data as discussed in Section 2.4. Finally, the predictive performance of each model is evaluated using the probabilistic accuracy of future forecasts.

4.1. The models

The reservoir simulator used for the test case simulates the flow of two immiscible incompressible mobile fluid phases – water and oil – through a porous medium of uniform porosity.[1] Corey (1954) power-law relative permeability curves with exponents of 2 are used to compute the mobilities of both fluid phases. The viscosity of the oil phase is assumed to be the same as the viscosity of the water phase. The data used for history matching are the "water cuts" at producing wells. (Water cut is the fraction of the producing fluid that is water.) Water is injected at fixed, equal rates into four wells and produced at fixed, equal rates at 9 producing wells. The data-generating model is discretized on a 200 × 200 grid, while the data assimilation model is discretized on a 30 × 30 grid.

The true synthetic observations are obtained from a channelized reservoir that was originally created to model a tidal flat environment (Biver et al., 2015). This data-generating model Fig. 6(c) was created using the truncated bi-Gaussian method. A similar bi-Gaussian model was shown previously to be difficult to history match because of the non-monotonic features of the threshold map (Oliver and Chen, 2018). The truncation rule results in the assignment of one of three facies to each cell of the grid. Each facies is then assigned permeability values of 1, 20, and 1000 in dimensionless units, respectively. (Note that because of the fixed-flux boundary conditions, the water cut behavior depends only on the ratio of permeability values for the three facies and the mobility ratios of the two fluid phases.)

In a channelized model, the water is expected to move rapidly from the injectors to the producers through the high-permeability channels, leaving much of the oil behind. This behavior is what is observed in the model (Fig. 7). The presence of channels in a low-permeability background explains the early production of water in Fig. 10 compared to the production from more homogeneous models.

The observations consist of water-cut data with four different history lengths. The first data set has a very short history period with data

[1] The simulator is available from the Github repository of Patrick Raanes: https://github.com/patnr/TPFA-ResSim.

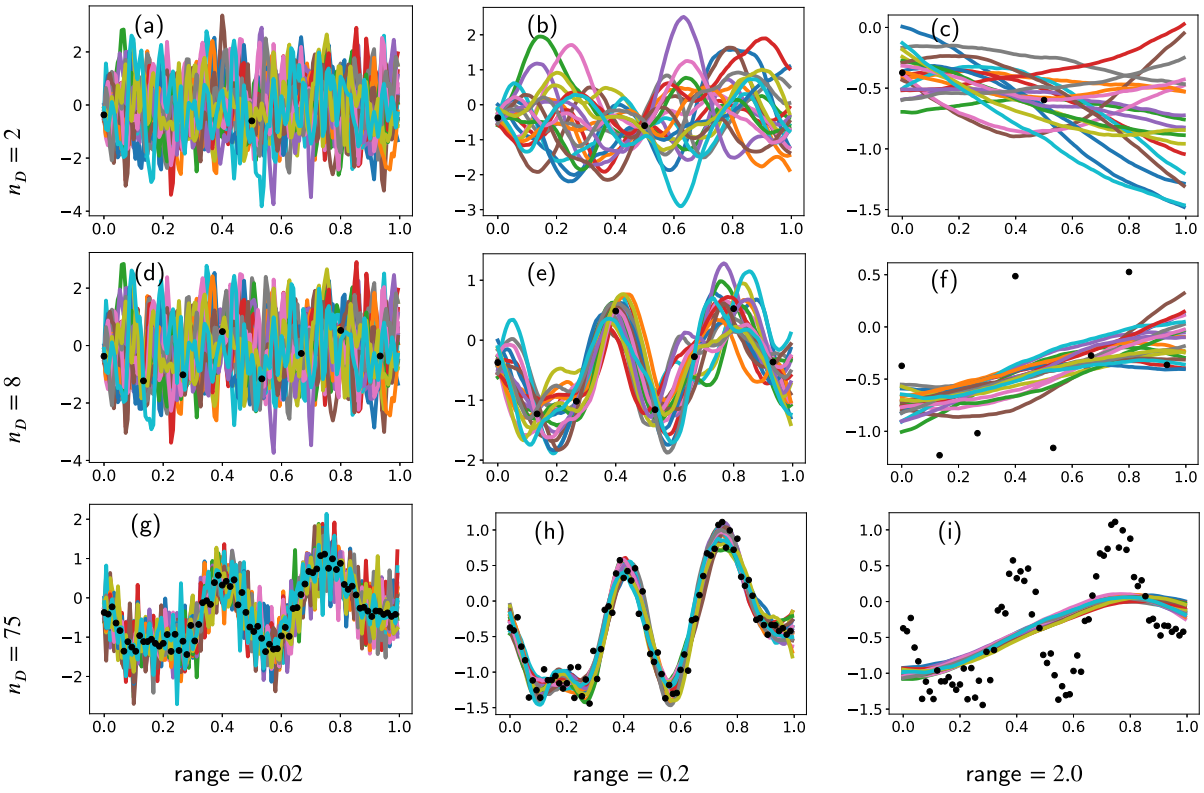


Fig. 3. Realizations from the posterior distribution for three correlation ranges in the prior model and for three different numbers of equally spaced observations.

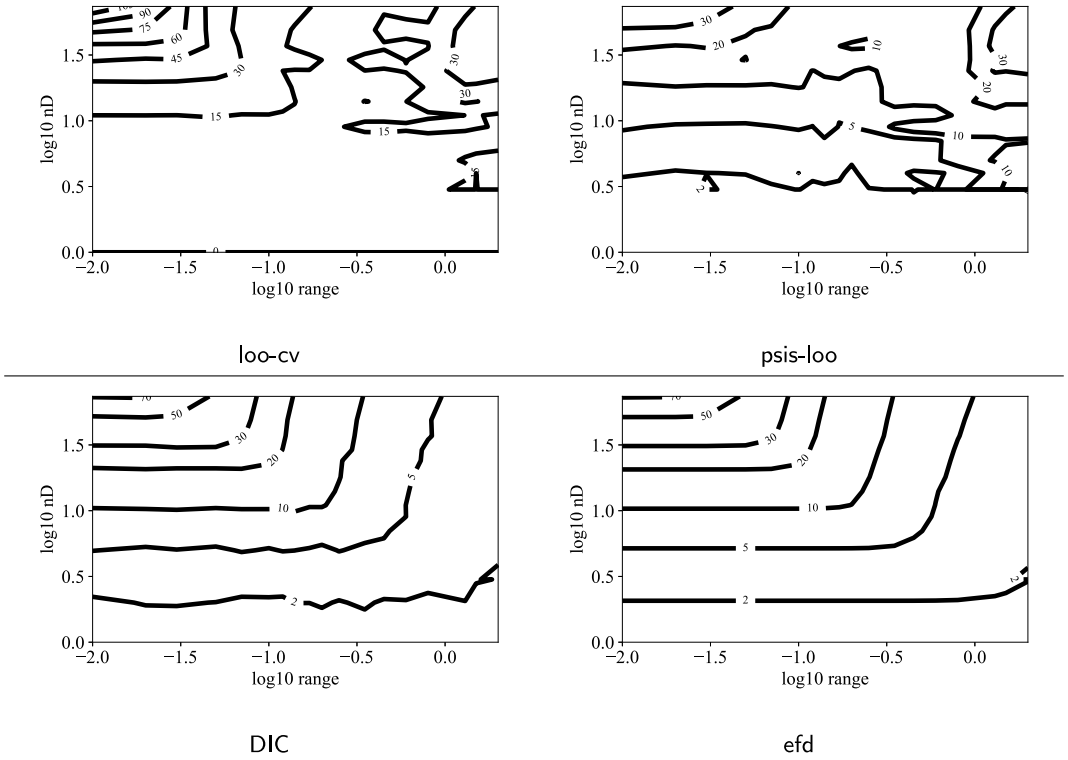
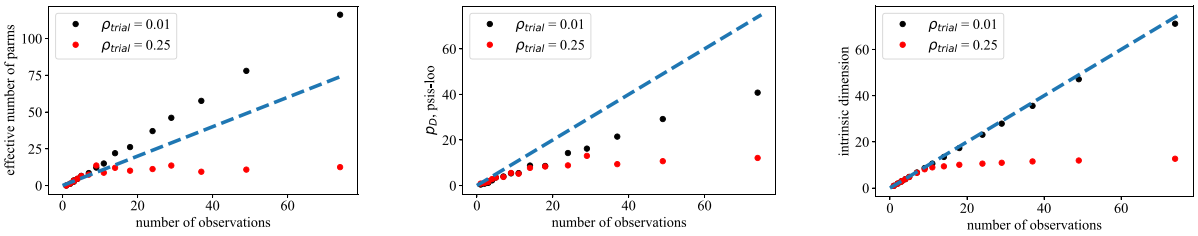


Fig. 4. Contour plots showing the effective number of parameters for the 1D linear inverse problem as a function of the number of actual observations and the correlation range in the prior covariance.



(a) loo-cv (b) psis-loo (c) efd

Fig. 5. The posterior predictive distributions for the one-dimensional linear-Gaussian inverse problem with varying amounts of data and varying correlation lengths.

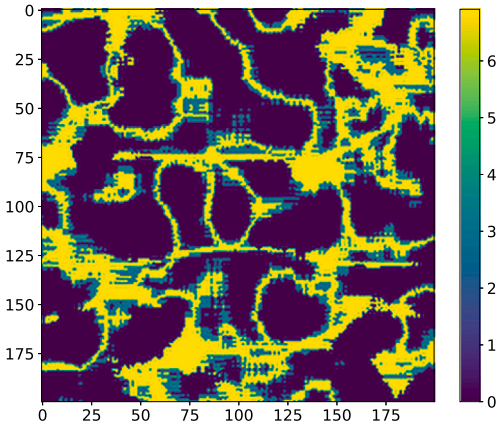


Fig. 6. The true log-permeability field used to generate data.

only to $t = 10$ at which time approximately half of the wells have water production. The second and third data sets include data up to $t = 30$ and $t = 50$, respectively. The fourth data set includes data up to $t = 80$ when all the wells have high water cuts (Fig. 8).

Because models are always only approximations of reality, prior models from three model classes all of which are different from the data-generating model are used for the numerical experiments. The first class of prior models is the family of Gaussian random fields, $\theta \sim \mathcal{N}(\theta^{\text{pr}}, C_\theta)$, with squared exponential isotropic covariance. A wide range of spatial correlation lengths was used to provide a variety of prior models with differing complexity. A realization from a Gaussian prior with a short covariance range is shown in Fig. 9(b). The second class of models is a polynomial trend family in which the log-permeability fields are polynomial functions of spatial location with uncertain coefficients. This model type was selected to include models that might be used for manual history matching for which parsimony is valued. A prior realization of the polynomial prior is shown in Fig. 9(a). The third class of models is the MPS prior based on a single training image. In this investigation, a widely used training image from Strebelle (2002) is utilized, consisting of two facies: a high permeability continuous channel facies and a low permeability background facies. The channels are assigned a dimensionless permeability value of 1000 and a permeability value of 1 to the background, as these are the true values in the data-generating model. Fig. 9(c) shows realization from the MPS model.

For this particular set of production observations, the geoscientist might speculate that the early breakthrough times are a result of the presence of channels in the real system; thus, the MPS model that has channel features similar to the data-generating model might be characterized as "realistic". On the other hand, the Gaussian and the polynomial priors would have been excluded from consideration for history matching based on a comparison of observations with predictions from the models. One would, in fact, conclude that these models are demonstrably "wrong" and the wrongness would become apparent as more data is available for assimilation.

4.2. Predictability before history matching

Fig. 10 shows the prior predictive distributions of production data and the corresponding observations from the first well of each prior model. The inconsistency between the observations and the realizations from Gaussian and the polynomial prior are very apparent in Fig. 10(a) and Fig. 10(b). In practice, based on the inconsistency, these two models would have been rejected as candidates for the prior model. On the other hand, the prior predictive distribution from the MPS model covers the observations (Fig. 10(c)). The log-score (12) was used to quantitatively evaluate how each model performs at predicting the unseen observations from the data-generating model at several time steps of the study. The logarithmic score shows that the MPS model performs better at predicting future behavior for the situation when no data have been used for calibration. In summary, a "realistic" prior may be appropriate for prediction when few data are available.

4.3. History matching – IES

Given model parameters θ following a Gaussian distribution $\theta \sim \mathcal{N}(\theta^{\text{pr}}, C_\theta)$, the *prior* distribution for the model parameters is given by:

$$p(\theta) \propto \left(-\frac{1}{2}(\theta - \theta_{\text{pr}})^T \mathrm{C}_\theta^{-1}(\theta - \theta_{\text{pr}})\right) \tag{17}$$

where the C_θ is the covariance of model parameters and θ_{pr} is the prior mean of model variables. Assuming that the observations errors are zero-mean Gaussian with covariance C_y, the *posterior* distribution for the model parameters is:

$$p(\theta|y) = p(y|\theta)p(\theta) \propto \exp\left(-\frac{1}{2}(\theta - \theta_{\text{pr}})^T \mathrm{C}_\theta^{-1}(\theta - \theta_{\text{pr}}) - \frac{1}{2}(g(\theta) - y)^T \mathrm{C}_Y^{-1}(g(\theta) - y)\right) \tag{18}$$

where $g(\theta)$ is the vector of predicted data values for the variables θ. Approximate sampling from the posterior distribution is obtained by minimizing randomized cost functions (Oliver et al., 2008),

$$O = (y_{\text{obs},i} - g(\theta_i))^T C_Y^{-1}(y_{\text{obs},i} - g(\theta_i)) + (\theta_{\text{pr},i} - (\theta_i)^T) C_\Theta^{-1}(\theta_{\text{pr},i} - \theta_i), \tag{19}$$

where the $y_{\text{obs},i}$ are perturbed observations.

In large inverse problems for which the forward model is a partial differential equation, an iterative ensemble smoother approach is often used to compute minimizers of (19), as the iES does not require the adjoint of the forward model. Here, the approximate form of the Levenberg–Marquardt IES (Chen and Oliver, 2013) is used for sampling from the posterior distribution. Iterations are stopped when either the maximum number of iterations exceeds 50 or the tuning parameter λ increases in 3 successive steps. Local analysis (Chen and Oliver, 2017) is used to remove spurious correlations from the updates and to increase the number of degrees of freedom. The localization radius in these cases is based on the true correlation range and on the distance between

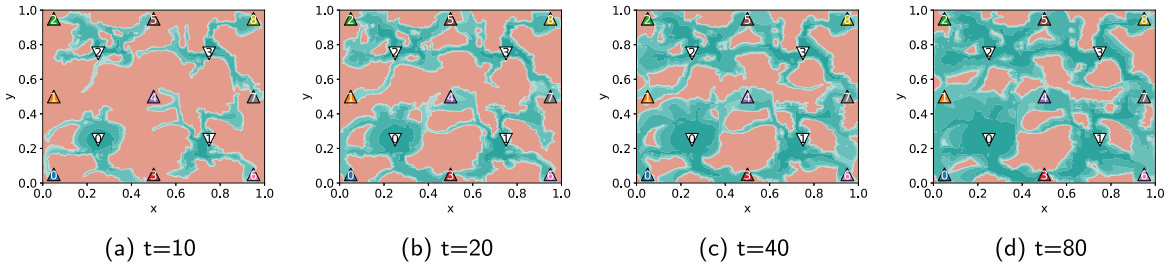


Fig. 7. Evolution of the water saturation in the true field at different time-steps. Positions of 9 producing wells are shown by triangles pointing up. The positions of 4 water injectors are shown by triangles pointing down.

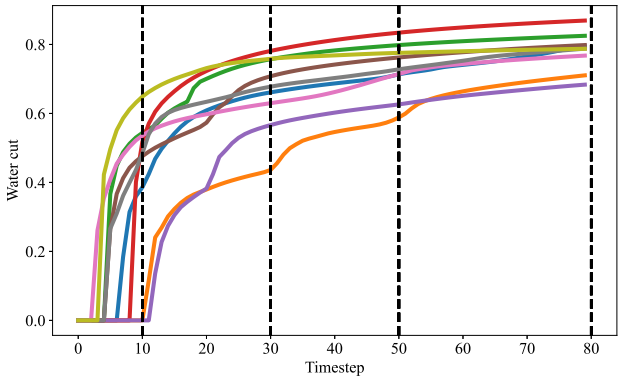


Fig. 8. Evolution of water cuts from the data-generating model from the 9 producing wells at different times. The true observations consist of the water cut data obtained up to t = 10, 30, 50, and 80, as indicated by dashed vertical lines.

wells. The ensemble model update at the lth iteration of the LM-IES algorithm is as follows.

$$\delta\theta_{\ell+1} = -\Delta\theta_\ell \, \Delta Y_\ell^T \left((1+\lambda_\ell)C_Y + \Delta Y_\ell \Delta Y_\ell^T\right)^{-1} \left(g(\theta_\ell) + \epsilon^* - y^{\text{obs}}\right) \tag{20}$$

where y_{obs} is the observed data, Y is the ensemble of predicted data. $\Delta\theta$ is the ensemble of mean-removed vectors of parameter realizations, and ΔY is the ensemble of variations of Y from the ensemble mean. λ_ℓ is the Levenberg–Marquardt tuning parameter that controls the size of the update step.

4.4. History matching – mcmc

The IES methods require that the prior distribution for model variables be Gaussian, so an IES cannot be used to history match the MPS models. Instead, a McMC method is used to generate approximate samples from the posterior distribution, and because it is not feasible to compute the posterior probability that would be required in a standard Metropolis–Hastings test, the extended Metropolis algorithm (Mosegaard and Tarantola, 1995) is implemented, which requires only evaluation of the ratio of the likelihoods of the proposed state and the current state.

The key steps of the extended Metropolis algorithm are summarized as follows:

1. Generate an initial model θ_0 at iteration 0 from the prior distribution for θ.
2. At iteration t generate a proposed model θ_* based on perturbation of the current model, θ_t.
3. Assign $\theta_{t+1} = \theta_*$ with probability $P = \min\left(1, \frac{p(\theta_*|y)}{p(\theta_t|y)}\right)$, else assign $\theta_{t+1} = \theta_t$.
4. $t = t + 1$, return to step 2.

The history matching process for the MPS model makes use of the SIPPI platform (Hansen et al., 2013), which provides tools for the generation of MPS realizations and proposal generation. The proposed model (Step 2 above) is obtained by replacing the permeabilities within a square region of the current model with new permeabilities conditioned to values on the boundaries of the region. The Gibbs sampler is used for generation of the permeabilities in the interior of the square. If the perturbation in the proposal step makes a small change to the current model, it will generally result in slow mixing of the chain but a high probability of acceptance. In contrast, a large change to the current model will lead to a small probability of acceptance. In order to achieve a reasonably high rate of mixing, the size of the square is selected adaptively to try to obtain an acceptance rate of approximately 0.2. Because the Gibbs sampler is used to generate proposed perturbations from the prior distribution, the acceptance test (Step 3 above) requires only the ratio of the likelihoods of the current model and the proposed model.

History matching consists of minimizing the mismatch between the observations and the simulated data from the updated models by adjusting uncertain parameters in the prior realizations. The value of the squared data mismatch part of the objective function is generally used as a criterion for judging the quality of the history match.

$$O_y = \frac{1}{2}\sum_{j=1}^{N}\left(\frac{y_j^{\text{sim}} - y_j^{\text{obs}}}{\sigma_j}\right)^2 \tag{21}$$

where N is the total number of data, y_j^{sim} and y_j^{obs} are the jth simulated and observed data, respectively; and σ_j is the standard deviation of the corresponding data-noise.

4.5. Evaluating model complexity

For the history matching test cases, the only measure of model complexity that was feasible to use was the effective number of parameters (efd). The efd values of each prior model were computed using (16) applied to the ensemble of posterior (history-matched) realizations. For the Gaussian model, with a fixed amount of data, efd is a monotonic function of the correlation length of the prior covariance (Fig. 11a). The effective dimension of the model decreases as the prior covariance range increases. As might be expected, for the polynomial prior, the model complexity (as quantified by the efd) increases monotonically as the number of the uncertain coefficients is increased for a fixed amount of flow data (Fig. 11(b)). The magnitudes of the efd for the polynomial models with relatively small numbers of coefficients are, however, much smaller than the magnitudes of the efd for the Gaussian prior models that were considered.

Because the MPS model is much more costly to history match than the Gaussian or polynomial models, a single MPS training image that does not allow rotation or scaling of the image was used. The MPS model consequently has a single value of efd for each length of the history matching period. The efd for the short data period, which ended at $t = 30$, was 4.4, while the efd for the longest history matching period, which ended at $t = 80$, was 9.8. For comparison, the Gaussian model with the shortest correlation length had an efd of 28.2 for the short data period and an efd of 37.9 for the longest data period, i.e., the Gaussian

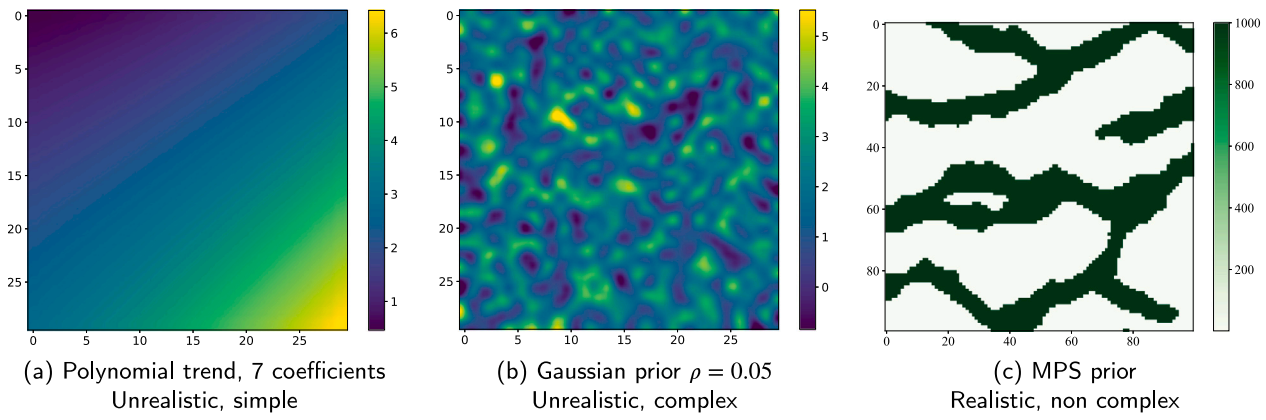


Fig. 9. Prior permeability fields.(a) The permeability field from the polynomial linear trend. (b) Permeability of Gaussian prior with shorter covariance range. (c) Permeability of MPS prior.

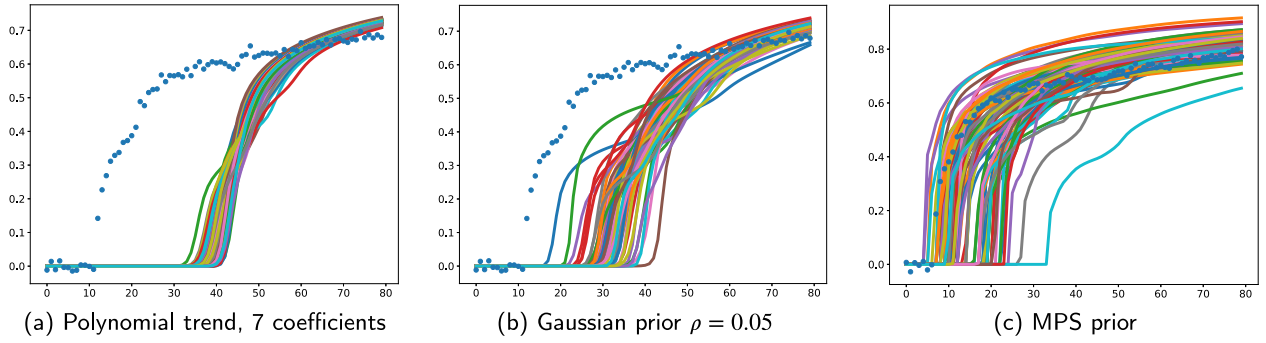


Fig. 10. Prior data before history matching and the true data. The blue dots show the observations, and the multicolored curves show the prior realizations from well 1 of each model.

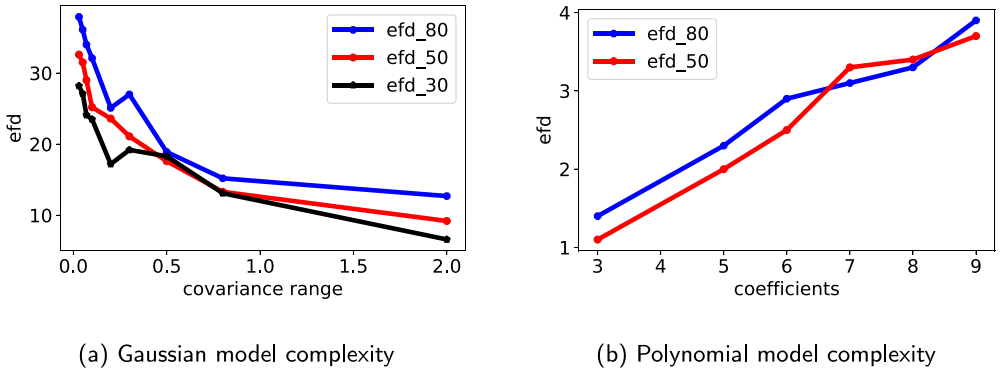


Fig. 11. (a) Effective number of parameters (efd) as a function of covariance range for the Gaussian model. (b) Effective number of parameters (efd) as a function of the number of coefficients of polynomials.

models with short correlation range are much more complex than either the polynomial model or the MPS model with the flow data.

4.6. Ability to match historical data

The quality of the history matching for each of the prior models is perhaps best judged by visual inspection of the comparisons between the actual observations and the posterior data predictions from each of the models. Fig. 12 shows the comparisons at three of the producing wells for one of the Gaussian models with a short correlation length, the polynomial model with 9 uncertain coefficients, and the MPS model. Because the observation error is small ($\sigma_d = 0.02$) the samples from the posterior should be very similar to the observations, and the spread of the predictions should also be small. The Gaussian model with short correlation length appears to capture all important features of the data, including breakthrough time and late time water-cut (Fig. 12(a–c)), while the match of the data for the polynomial model is inaccurate and overconfident (Fig. 12(d–f)). The matches to the data for the MPS model are not as accurate as should be expected if these predictions are actually from the posterior of the data-generating model. In particular, the predictions for Well 2 are biased over most of the history

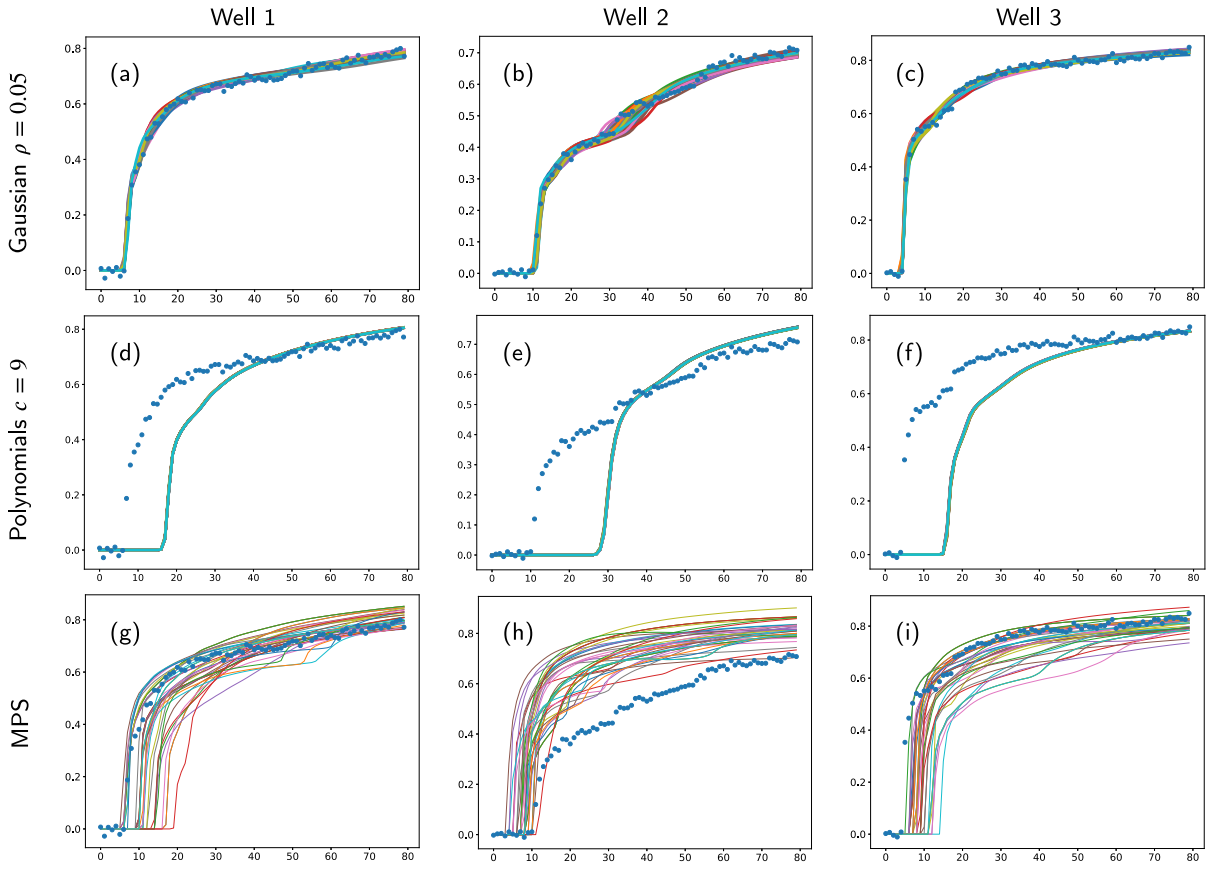


Fig. 12. A comparison of the posterior predictive distribution of each prior models with the historical data of length 80 from the three first wells.

matching interval, but the spread in predictions is too large for all wells (Fig. 12(g–i)).

A quantitative assessment of the quality of the history match is provided by the data objective function, O_y, which is half the normalized squared data mismatch (21). In a perfect model scenario with perfect sampling from the posterior, the expected value of O_y is equal to half the number of observations, and the standard deviation of O_y is $\sqrt{N/2}$. For the long period of history matching, the number of data is $N = 720$ so the expected value of the objective function is $E[O_y] = 360$, and the standard deviation is approximately 19. For comparison, the value of O_y for the Gaussian prior with the shortest correlation range is 337, which is in the expected range. The value of O_y for the polynomial model with 9 coefficients is 49 380, and the value of O_y for the MPS model is 20 600, neither of which is in the acceptable range.

The posterior predictive distributions for the Gaussian model shown in Fig. 12((a–c)) are for history matching up to $t = 80$ with correlation range $\rho = 0.05$. However, when long correlation ranges were used in the Gaussian model (e,g. when $\rho = 2$), the data mismatch was much larger after history matching. The complete tabulation of predictability and complexity measures for Gaussian models is shown in Table 1.

Additionally, the approximation of the ability of each prior model to be used to predict unseen data from the history matching period is measured with the lppd (3). Fig. 13(a) summarizes the results of the lppd computations as functions of model complexity, quantified by the efd. Models with higher efd generally have higher lppd. The polynomial models (orange points) have the lowest lppd and also the lowest efd. The MPS models also have small efd, but it is not apparent from Fig. 13(a) how the predictability (lppd) of the MPS models compares to the Gaussian models. Limiting the range of the lppd to focus on the models with higher predictability, Fig. 13(b) shows the relationship only for the Gaussian and MPS models. It shows, however, not just the dependence of efd on the prior model (e.g. correlation length for the Gaussian models), but also the dependence on the amount of data.

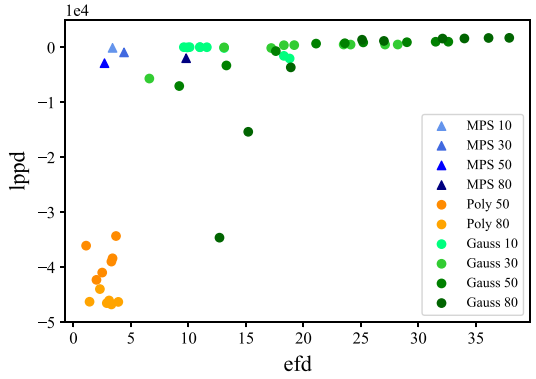
Table 1
Predictability results for the Gaussian prior model with correlation length 0.05 and data to $t = 80$.

ρ_{pr}	lppd	lppd (loo)	p_D (loo)	efd	p_{DIC}	mean obj
0.80	−16 156	−17 856	1700	14.9	65.2	19 369
0.50	−4 667	−5 659	992	18.3	171.9	7 498
0.30	1 207	697	510	27.7	150.7	1 333
0.20	1 453	1 051	402	28.4	187.0	878
0.10	1 672	1 484	188	32.0	150.5	461
0.07	1 742	1 502	240	35.5	212.3	430
0.05	1 717	1 530	187	36.0	12.8	419
0.03	1 761	1 571	191	36.5	63.0	337

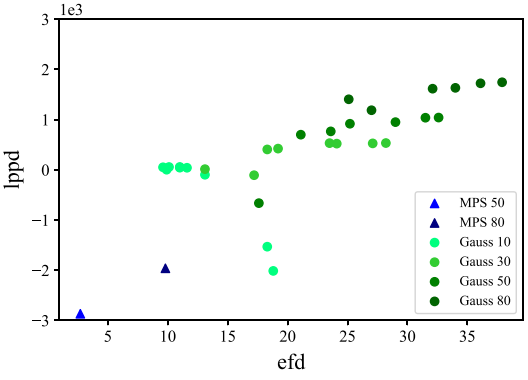
In this range, the relationship between model complexity (efd) and predictability (lppd) is clear—higher model complexity leads to higher predictability of data that were held out from the history matching period. The Gaussian models as a class have the highest predictability, and the Gaussian models with short correlation length have the highest lppd within that class. As discussed in Section 4.7, these conclusions may not be valid for predictions of data governed by different processes.

4.7. Evaluating the predictability

The primary purpose of modeling and history matching is to improve the reliability of predictions of future reservoir behavior. Not all models are equally suitable for history matching and one of the main criteria in the model selection exercise is the model's predictive power at probabilistic forecasting. This numerical experimental study investigates the impact of the length of the observation period on



(a) Relationship between the effective number of parameters, the log-pointwise predictive density, and the length of the historical data.



(b) A zoom of (a) focusing on the most relevant region.

Fig. 13. Predictability of observed data. The green circles correspond to the Gaussian model data points, in orange for the polynomial model, and the blue triangles for the MPS model. For each of these colors, the dark shades are for the longest period of historical observations. Lighter shades are for the shortest data length.

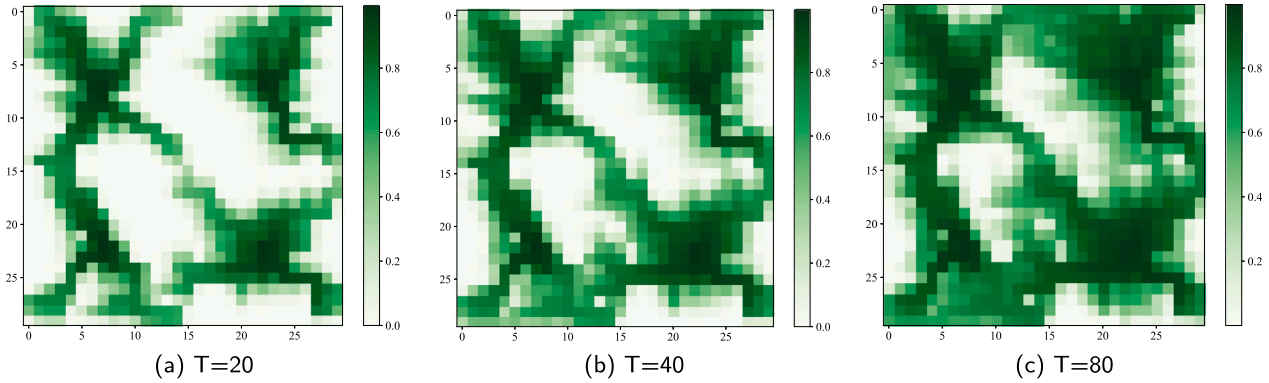


Fig. 14. Posterior mean of water saturation of the Gaussian prior with $\rho = 0.05$ at three different times for data length of 80.

the predictability of future near-term behavior ($T_e + 10$) and future long-term behavior ($T_e + 60$) for each model type. The assessment of the predictability is performed by comparing the logarithmic score which is a function of the approximation of the pdf for the forecast from the posterior ensemble and the actual forecast from the true data-generating model.

Table 2 shows the results of the computed values of the logarithmic score (12) for extrapolations of historical behavior in the three prior model classes and the true predictions from the data-generating model. Here predictability is evaluated at two times: 10 time steps after the end of the history period and 60 time steps after the end of the history period.

Fig. 15 compares the ensemble of forecasts from three of the nine producing wells with the forecasts from the data-generating model when the models have been history matched to data from the first 80 time-steps. Neither the polynomial model nor the MPS model are well-history-matched. The data predictions from the Gaussian model exhibit very little spread in water cut during the historical period, which is what should be seen when the data are accurate. The spread in predictions from the Gaussian model increases somewhat in the forecast period, but the spread is relatively small. The predictability score for the Gaussian model is relatively high because the mean forecast is accurate for all nine wells and the spread is relatively small, yet covers the truth.

The ensemble of predictions from the polynomial model, on the other hand, has small spread in the history matching period and in the forecast period (Fig. 15(d-f)). The mean forecast at Well 1 is quite good, but poor at Well 2. The predictability score is based on forecasts at all nine wells. The polynomial model receives a low predictability score primarily because the forecasts are overly confident (the truth is not within the ensemble).

Like the polynomial model, the MPS model is also incapable of matching the data in the history matching period, but its spread is larger because it was necessary to increase the assumed observation error to achieve any mixing in the McMC ((Fig. 15(g–i)). And, like the polynomial model, the mean forecast at Well 2 ((Fig. 15(h)) is poor, although the log-score for MPS is better than the log-score for the polynomial model because the predicted forecasts are not overly confident.

Predictability for the Gaussian and MPS models are illustrated in Fig. 16 for the situation in which only a small amount of data are available for history matching. In this case, the Gaussian model performs well in a few wells that have already experienced water breakthrough (Fig. 16(a) and (d)), but poorly at the short-term forecast for Well 2 which had not yet seen water production. Interestingly, the ensemble of models show very little spread in the model predictions in the history-matched periods, but reasonable spreads for Wells 1 and 3 in the forecast period. The MPS model receives a higher log-score for the forecasts in this case because the ensemble of predictions contain some with early break through times, even though water production had not been seen in the history period (Fig. 16(e)).

The short term predictability results from numerical experiments with history matching and forecasting using the MPS and Gaussian models are summarized in Fig. 17. One model (MPS) might be considered "realistic" in the sense that it is trained to generate realizations with connected high-permeability channels, but it is extremely costly to

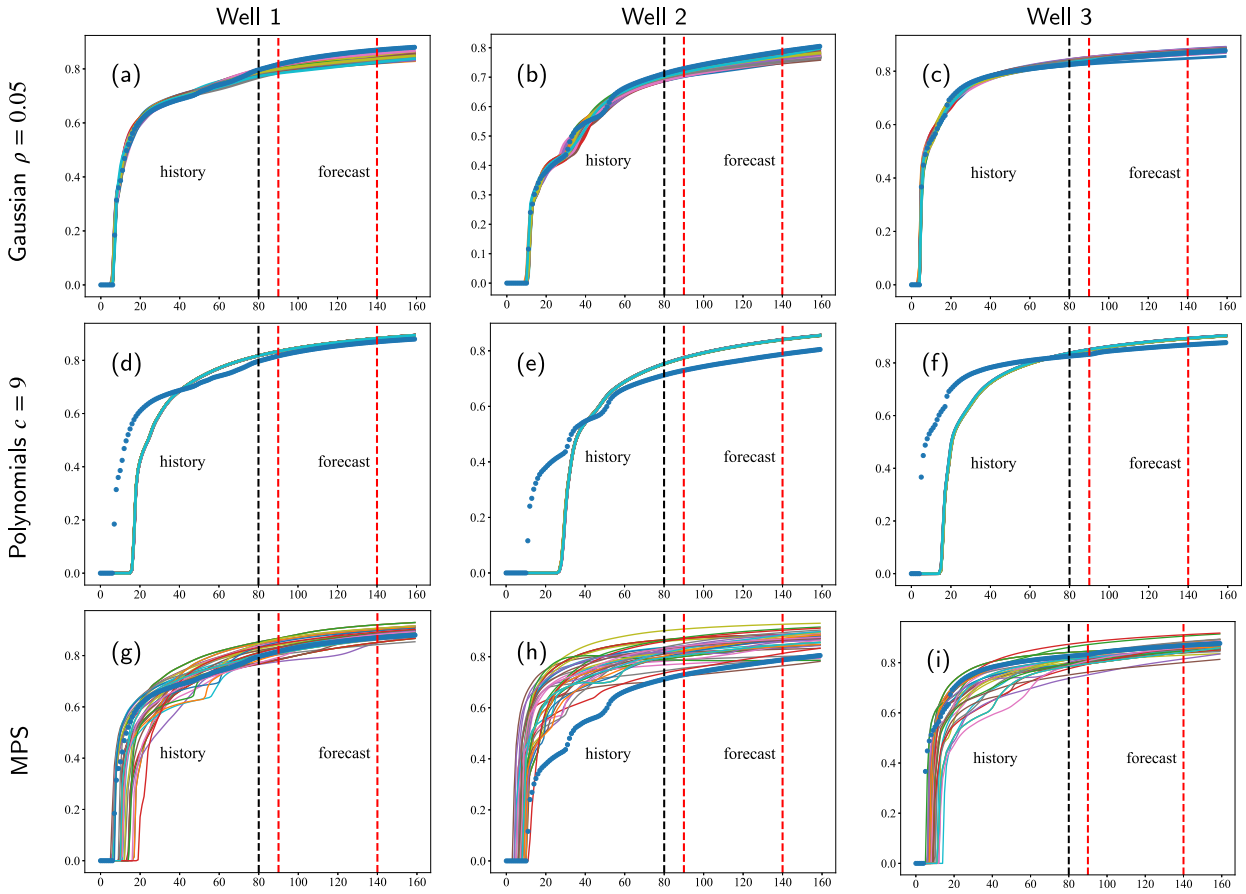


Fig. 15. A comparison of forecast from each prior models with the historical data of length 80 from the three first wells. Data are history matched only using data to $t = 80$, after which the forecasts are compared with actual behavior of the data-generating model.

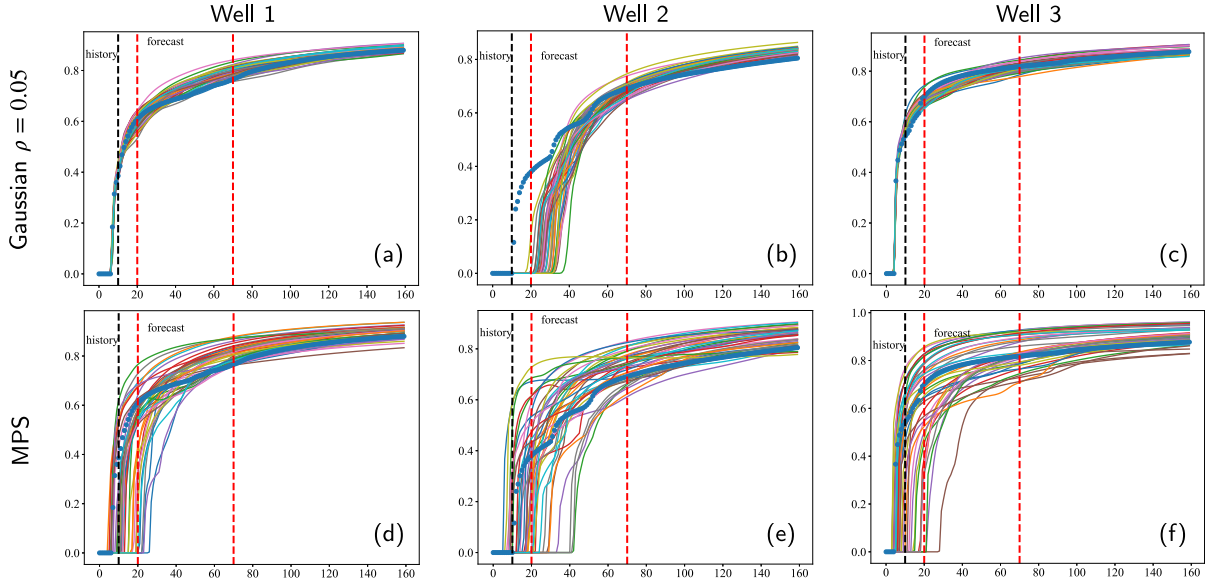


Fig. 16. A comparison of forecast from Gaussian and MPS models with the historical data of length 10 from the three first wells. Data are history matched only using data to $t = 10$, after which the forecasts are compared with actual behavior of the data-generating model.

Table 2
Log-score values between forecast from each prior and the true observed data.

Time	Prior type	Lppd	$T_e + 10$	$T_e + 60$
$T_e = 10$ (before HM)	MPS(*)	–	−10.9	11.8
	Gaussian 0.05	–	−2.47e+181	−26.1
	Polynomial (c = 9)	–	−8.13e+61	−263.5
$T_e = 10$	MPS(*)	−10.12	−1.6	−12.4
	Gaussian 0.05	60.3	−1.92e+08	−16.3
$T_e = 30$	MPS(*)	−57.90	−7.7	−12.3
	Gaussian 0.05	526.7	3.4	−5.4
$T_e = 50$	MPS(*)	−117.82	30.9	36.9
	Gaussian 0.05	1037.8	22.9	9.3
	Polynomial (c = 9)	−34 313.8	−2.52e+04	−5.50e+04
$T_e = 80$	MPS(*)	−176.27	−10.2	−11.4
	Gaussian 0.05	1723.1	7.8	6.4
	Polynomial (c = 9)	−46 301.0	−2.42e+04	−6.28e+04

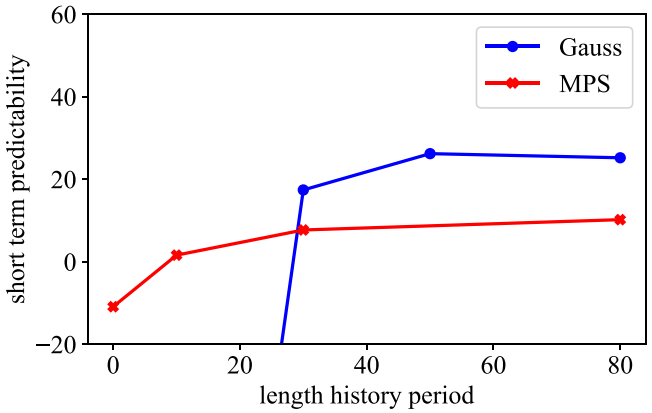


Fig. 17. Short-term predictability for two different model types and various lengths of data.

history match and does not have enough complexity (effective dimension) to assimilate even modest amounts of data. On the other hand, the Gaussian models are decidedly "unrealistic" as models of a channelized reservoir. None of the realizations exhibit the early breakthrough times characteristic of channelized reservoirs. Yet, the Gaussian model has sufficient complexity to match the water production behavior when data, including water breakthrough, are assimilated and to develop channel-like features after assimilating sufficient production data (Fig. 14).

4.8. Discussion of flow example

Although the key purpose of the study was to identify a measure of model complexity beyond computational expense, it is still worth noting that geologically realistic models are generally challenging to history match to a proper misfit level while maintaining realism. For history matching using the IES, an ensemble size of 100 and a maximum number of iterations of 50 were used, so 100 approximately independent model realization were obtained from the posterior at the cost of 50 × 100 reservoir simulations plus a minor additional cost of computing the changes to the model. For the MPS model, it was impossible to obtain convergence with the correct observation error, so the observation error was increased and 5 independent realizations were obtained from 30,000 simulation runs plus the cost of running the Gibbs sampler at each iteration. In summary, it required about 390 minutes of CPU time to generate one independent realization of the MPS model (still without proper data misfit in the history matching period) compared to only 4.3 seconds for a Gaussian model and 1.1 seconds for the polynomial model. In this case, large computational expense did not result in well-calibrated models or good predictability. For reference, the simulations were performed using a i9-9900 CPU @ 3.10 GHz × 16 processor and 64 GB of memory.

Fig. 17 shows that at early times (especially before data is available), it might be advantageous to use a geologically realistic prior to predict future reservoir behavior. Later in the life of a field, however, the models with larger potential complexity and ability to assimilate data (e.g. Gaussian) should be preferred. Incidentally, the short-term predictability of the polynomial models is not shown in Fig. 17 as the predictability is too poor to appear on the plot. The polynomial models lack the realism to mimic channels and the complexity (the effective dimension) to assimilate the flow data.

These simple results suggest that it may be usefull to abandon the goal of geologic realism in brown-field modeling studies if the goal is improved forecasting. A model with sufficiently high complexity (large numbers of effective parameters) should be preferred for history-matching mature fields when large amounts of data must be history matched. Isotropic Gaussian models with short correlation lengths were surprisingly effective at history-matching the data and providing accurate short-term forecasts, despite the data-generating model being characterized by narrow high-permeability channels in a low-permeability background medium. The isotropic Gaussian models would, in fact, have been falsified by analysis of the prior predictive distribution. The posteriori Gaussian realizations also developed channel-like features that allowed regions of the reservoir between channels to remain undrained (Fig. 14). The Gaussian models have the flexibility to match a wide variety of reservoir behaviors, leading to possible concerns over the possibility of overfitting. It was not, however, observed in the flow examples—although the models with the best fit to data were the most complex, they had not done so at the expense of overfitting.

5. Conclusions

Model complexity can be defined in many ways including, for example, realism, the computational time required for simulation, and the difficulty of conditioning to observations. This paper focused on the number of effective parameters as a useful measure of complexity. The rationale for this choice is that the predictability of unseen data suffers when models are either "over parameterized" or "under parameterized". In the first case, it may be possible to match the historical data, but the model may have excessive variability in predictions. In the case of under-parameterization, the model will neither fit historical data nor be useful for predicting unseen data (bias). The effective number of parameters for history matching is a function of both the prior model and the data. Although it is a function of the number of actual parameters, the effective number of parameters is generally much smaller than the actual number. In the flow problems, for example, there were 900 values of permeability that could be adjusted, yet the effective number of parameters was on the order of 36 for cases in which the data were well matched. This paper investigated several methods of estimating predictability and the effective number of parameters, including leave-one-out cross-validation and various approximations based on information criteria.

The methods were tested on a simple linear Gaussian problem with noisy partial observations and a 2D porous flow problem in which the data-generating model had narrow high-permeability channels embedded in a low-permeability background medium. Leave-one-out cross-validation always provided useful results for the 1D linear problem but was too expensive to apply to practical history matching problems. An importance sampling approximation gave useful cross-validation approximations without the expense of cross-validation. However, it failed when applied to the history matching problem, primarily because the estimate was highly sensitive to a few extreme values of the estimated predictability. The most successful methods for estimating the effective number of parameters were based on the DIC. The use of the effective number of parameters as a metric to measure the complexity

of the model can be easily applied for a real field application. It is straightforward to compute since it only requires the posterior covariance for model parameters, an approximation of which is obtained from the ensemble of history-matched realizations.

The results from the numerical experiments performed in this study suggest key elements to formalize the process of prior model selection for history matching for practical fields. From Fig. 17, it is apparent that the prior model type should be chosen based on the modeling objective and data availability. Realistic models and complex models are not necessarily the same, and each may be required for certain types of problems. If the modeling goal is to predict the reservoir behavior in a green field for which only a few data are available, it is beneficial to incorporate realistic characteristics of the geology, such as channels, into the model. In the flow problem example, the water cut from the arguably more realistic MPS model (Fig. 10(c)) provided a better prediction of future behavior before data were available. The MPS model, however, lacked the effective parameters required to match observed flow data and was not useful for short-term forecasting when more observations become available.

The methodology we presented for computing model complexity is extremely efficient when IES methods are used for history matching. They would not, however, always provide useful measures when the posterior distribution of model parameters is multimodal. In practice, it may be difficult to select a prior model with the appropriate level of complexity as the effective dimension of the model can only be evaluated after history matching. The findings here do, however, suggest useful guidelines for prior model selection in both data-rich and data-poor scenarios. These guidelines do not obviate the need for model elicitation as described, for example, by O'Hagan (2013).

CRediT authorship contribution statement

Tanteliniaina N. Mioratina: Investigation, Methodology, Software, Formal analysis, Writing – original draft. **Dean S. Oliver:** Conceptualization, Methodology, Writing – review & editing, Formal analysis, Supervision.

Declaration of competing interest

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

Data availability

No data was used for the research described in the article.

Acknowledgments

The authors acknowledge financial support from the NORCE Norwegian Research Centre cooperative research project "Assimilating 4D Seismic Data: Big Data Into Big Models" which is funded by industry partners, Equinor Energy AS, Lundin Energy Norway AS, Repsol Norge AS, Shell Global Solutions International B.V., TotalEnergies E&P Norge AS and Wintershall Dea Norge AS, as well as the Research Council of Norway through the Petromaks2 program (NFR project number: 295002).

References

Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., Stuart, A.M., 2017. Importance sampling: Intrinsic dimension and computational cost. Statist. Sci. 32 (3), 405–431. http://dx.doi.org/10.2307/26408299.

Berger, J.O., 1985. Prior information and subjective probability. In: Statistical Decision Theory and Bayesian Analysis. Springer, pp. 74–117.

Biver, P.Y.A., Allard, D., Pivot, F., Ruelland, P., 2015. Recent advances for facies modelling in pluri-Gaussian formalism. In: Petroleum Geostatistics, 7–11 September, Biarritz, France. EAGE, http://dx.doi.org/10.3997/2214-4609.201413615.

Burman, P., 1989. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. Biometrika (ISSN: 0006-3444) 76 (3), 503–514. http://dx.doi.org/10.1093/biomet/76.3.503.

Chen, Y., Oliver, D.S., 2013. Levenberg-marquardt forms of the iterative ensemble smoother for efficient history matching and uncertainty quantification. Comput. Geosci. 17 (4), 689–703. http://dx.doi.org/10.1007/s10596-013-9351-5.

Chen, Y., Oliver, D.S., 2017. Localization and regularization for iterative ensemble smoothers. Comput. Geosci. 21 (1), 13–30. http://dx.doi.org/10.1007/s10596-016-9599-7.

Corey, A.T., 1954. The interrelation between gas and oil relative permeabilities. Prod. Mon. 19, 38–41.

Dake, L.P., 2001. The Practice of Reservoir Engineering, Vol. 36. Elsevier.

Draper, D., 1995. Assessment and propagation of model uncertainty. J. R. Stat. Soc. Ser. B Stat. Methodol. 57 (1), 45–70.

Evensen, G., Vossepoel, F.C., van Leeuwen, P.J., 2022. Data Assimilation Fundamentals: A Unified Formulation of the State and Parameter Estimation Problem. Springer Nature.

Gelfand, A.E., Dey, D.K., Chang, H., 1992. Model Determination Using Predictive Distributions with Implementation Via Sampling-Based Methods. Technical report, Stanford University, Dept of Statistics.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995. Bayesian Data Analysis. Chapman and Hall/CRC.

Gelman, A., Hennig, C., 2017. Beyond subjective and objective in statistics. J. R. Stat. Soc. Ser. A (Stat. Soc.) 180 (4), 967–1033.

Gelman, A., Hwang, J., Vehtari, A., 2014. Understanding predictive information criteria for Bayesian models. Stat. Comput. (ISSN: 0960-3174) 24 (6), 997–1016. http://dx.doi.org/10.1007/s11222-013-9416-2.

Good, I.J., 1952. Rational decisions. J. R. Stat. Soc. Ser. B Stat. Methodol. (ISSN: 00359246) 14 (1), 107–114.

Gronau, Q.F., Wagenmakers, E.-J., 2019. Limitations of Bayesian leave-one-out cross-validation for model selection. Comput. Brain Behav. 2 (1), 1–11.

Guardiano, F.B., Srivastava, R.M., 1993. Multivariate geostatistics: beyond bivariate moments. In: Geostatistics Troia'92. Springer, pp. 133–144.

Hansen, T.M., 2021. Entropy and information content of geostatistical models. Math. Geosci. 53 (1), 163–184.

Hansen, T.M., Cordua, K.S., Looms, M.C., Mosegaard, K., 2013. SIPPI: A matlab toolbox for sampling the solution to inverse problems with complex prior information: Part 2 – application to crosshole GPR tomography. Comput. Geosci. 52 (1), 481–492.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, second ed. Springer-Verlag, ISBN: 978-0-387-84857-0.

Heße, F., Comunian, A., Attinger, S., 2019. What we talk about when we talk about uncertainty. Toward a unified, data-driven framework for uncertainty characterization in hydrogeology. Front. Earth Sci. 7, 118. http://dx.doi.org/10.3389/feart.2019.00118.

Hunt, R.J., Doherty, J., Tonkin, M.J., 2007. Are models too simple? Arguments for increased parameterization. Ground Water 45 (3), 254–262.

Linde, N., Renard, P., Mukerji, T., Caers, J., 2015. Geological realism in hydrogeological and geophysical inverse modeling: A review. Adv. Water Resour. (ISSN: 0309-1708) 86, Part A, 86–101.

Mariethoz, G., Renard, P., Straubhaar, J., 2010. The direct sampling method to perform multiple-point geostatistical simulations. Water Resour. Res. 46 (11).

Mosegaard, K., Tarantola, A., 1995. Monte Carlo sampling of solutions to inverse problems. J. Geophys. Res. Solid Earth 100 (B7), 12431–12447.

Oakley, J.E., O'Hagan, A., 2004. Probabilistic sensitivity analysis of complex models: a Bayesian approach. J. R. Stat. Soc. Ser. B Stat. Methodol. (ISSN: 1467-9868) 66 (3), 751–769.

O'Hagan, A., 2013. Bayesian inference with misspecified models: Inference about what? J. Statist. Plann. Inference (ISSN: 0378-3758) 143 (10), 1643–1648. http://dx.doi.org/10.1016/j.jspi.2013.05.016.

Okiria, E., Okazawa, H., Noda, K., Kobayashi, Y., Suzuki, S., Yamazaki, Y., 2022. A comparative evaluation of lumped and semi-distributed conceptual hydrological models: Does model complexity enhance hydrograph prediction? Hydrology (ISSN: 2306-5338) 9 (5), http://dx.doi.org/10.3390/hydrology9050089.

Oliver, D.S., Alfonzo, M., 2018. Calibration of imperfect models to biased observations. Comput. Geosci. 22 (1), 145–161.

Oliver, D.S., Chen, Y., 2011. Recent progress on reservoir history matching: a review. Comput. Geosci. 15 (1), 185–221.

Oliver, D.S., Chen, Y., 2018. Data assimilation in truncated plurigaussian models: impact of the truncation map. Math. Geosci. 50 (8), 867–893.

Oliver, D.S., Fossum, K., Bhakta, T., Sandø, I., Nævdal, G., Lorentzen, R.J., 2021. 4D seismic history matching. J. Pet. Sci. Eng. 207, 109119.

Oliver, D.S., Reynolds, A.C., Liu, N., 2008. Inverse Theory for Petroleum Reservoir Characterization and History Matching. Cambridge University Press, http://dx.doi.org/10.1017/CBO9780511535642.

Oreskes, N., 2003. The role of quantitative models in science. In: Lauenroth, W.K. (Ed.), Models in Ecosystem Science. Princeton University Press, pp. 13–30.

Simpson, D., Rue, H., Riebler, A., Martins, T.G., Sørbye, S.H., 2017. Penalising model component complexity: A principled, practical approach to constructing priors. Statist. Sci. 32 (1), 1–28.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc. Ser. B Stat. Methodol. (ISSN: 1467-9868) 64 (4), 583–639.
Strebelle, S., 2002. Conditional simulation of complex geological structures using multiple-point statistics. Math. Geol. 34 (1), 1–21.
Stuart, A.M., 2010. Inverse problems: A Bayesian perspective. Acta Numer. 19, 451–559.
Tarantola, A., 1987. Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation. Elsevier, Amsterdam.
Van der Linde, A., 2012. A Bayesian view of model complexity. Stat. Neerl. 66 (3), 253–271.
Vehtari, A., 2018. PSIS. https://github.com/avehtari/PSIS.
Vehtari, A., Gelman, A., Gabry, J., 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat. Comput. 27 (5), 1413–1432. http://dx.doi.org/10.1007/s11222-016-9696-4.
Vink, J.C., Gao, G., Chen, C., 2015. Bayesian style history matching: Another way to under-estimate forecast uncertainty? In: SPE Annual Technical Conference and Exhibition. OnePetro.
Williams, G.J.J., Mansfield, M., MacDonald, D.G., Bush, M.D., et al., 2004. Top-down reservoir modelling. In: SPE Annual Technical Conference and Exhibition Held in Houston, Texas, 26–29 September. Society of Petroleum Engineers.