Computational Geosciences (2019) 23:1331–1347
https://doi.org/10.1007/s10596-019-09889-6

ORIGINAL PAPER

Evaluating prior predictions of production and seismic data

Miguel Alfonzo[1,2] · **Dean S. Oliver**[2]

Received: 10 December 2018 / Accepted: 23 August 2019 / Published online: 8 October 2019

Abstract

It is common in ensemble-based methods of history matching to evaluate the adequacy of the initial ensemble of models through visual comparison between actual observations and data predictions prior to data assimilation. If the model is appropriate, then the observed data should look plausible when compared to the distribution of realizations of simulated data. The principle of data coverage alone is, however, not an effective method for model criticism, as coverage can often be obtained by increasing the variability in a single model parameter. In this paper, we propose a methodology for determining the suitability of a model before data assimilation, particularly aimed for real cases with large numbers of model parameters, large amounts of data, and correlated observation errors. This model diagnostic is based on an approximation of the Mahalanobis distance between the observations and the ensemble of predictions in high-dimensional spaces. We applied our methodology to two different examples: a Gaussian example which shows that our shrinkage estimate of the covariance matrix is a better discriminator of outliers than the pseudo-inverse and a diagonal approximation of this matrix; and an example using data from the Norne field. In this second test, we used actual production, repeat formation tester, and inverted seismic data to evaluate the suitability of the initial reservoir simulation model and seismic model. Despite the good data coverage, our model diagnostic suggested that model improvement was necessary. After modifying the model, it was validated against the observations and is now ready for history matching to production and seismic data. This shows that the proposed methodology for the evaluation of the adequacy of the model is suitable for large realistic problems.

Keywords Prior predictive distribution · Model criticism · Model improvement · Mahalanobis distance · Production data · RFT data · Acoustic impedance · Seismic inversion · Correlated observation error · History matching · Norne field

1 Introduction

The process for learning about the subsurface from observations and for making model-based forecasts of future behavior is sometimes separated into two parts. The first can be termed model criticism, while the second can be termed parameter estimation. In practice, we subdivide model criticism into a criticism of the model *before* parameter estimation and a model criticism that occurs *after* parameter estimation. Although there are similarities in the methods of criticism that might be applied in both periods, the criticism that is based on the *prior predictive distribution* is highly sensitive to errors in the prior distribution for model parameters and to errors in the forward model. Criticism based on the *posterior predictive distribution* is highly sensitive to errors in the characterization of measurement error [7, 33]. The purpose of model criticism is not to determine if the model is "wrong," but rather to determine if it might adequately represent reality by comparison of simulated data from the model with observed data [6]. Some methodologies for subsurface model criticism or falsification compute the probability of a model by a type of Bayesian model averaging which implicitly assumes that at least one of the model realizations or scenarios in the study is adequate [21, 35, 40]. The process of falsification is illustrated on an example in which none of the scenarios are adequate but, in the example, the method for falsification is qualitative [21]. We propose a general quantitative methodology that assesses the adequacy of the model by direct comparison of

✉ Miguel Alfonzo
miguelalfonzo74@hotmail.com

[1] University of Bergen, Bergen, Norway

[2] Norwegian Research Centre (NORCE), Bergen, Norway

the perturbed simulated data to the observed data without requiring an assumption on the validity of the model. If the perturbed simulated observations are not consistent with the actual observations, then one should consider modifying the assumptions about the prior distribution of model parameters, add new model parameters, or modify the distribution of observation errors.

It is standard in any ensemble-based data assimilation method to perform a superficial check of the adequacy of the initial ensemble by visual comparison of simulated data with actual observed data. The comparison is typically performed on the basis of "coverage" of the individual observations, i.e., checking to see if each observation is contained within the spread of the ensemble when examined one-by-one. In many cases, this has been sufficient to show that the initial ensemble is inadequate [12, 14]. We note, however, that the converse is not necessarily true, that is even when the coverage is good, it is sometimes difficult to obtain a suitable match to all data [13].

More powerful statistical tests of adequacy can be obtained from subsets of the observations, instead of a one-by-one examination. Using the notation of Box [7], the prior predictive distribution can be written as follows:

$$p(\mathbf{y}|A) = \int p(\mathbf{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|A)\, d\boldsymbol{\theta} \tag{1}$$

where simulated data are denoted by $\mathbf{y}$, model parameters are denoted as $\boldsymbol{\theta}$, and A denotes the totality of assumptions that have been made about the model. Although we cannot evaluate (1) for practical history matching problems with large numbers of data and large numbers of model parameters, it is relatively easy to generate samples from the distribution $p(\mathbf{y}|A)$. To do that, we simply generate samples from $p(\boldsymbol{\theta}|A)$ (the initial ensemble) then, for the ith sample from the initial or prior distribution, we generate simulated data and perturb it according to our model of observation error, i.e., $\mathbf{y}_i = \mathbf{g}(\boldsymbol{\theta}_i) + \boldsymbol{\epsilon}_i$, where typically $\boldsymbol{\epsilon}_i \sim N[0, \mathbf{C}_D]$. Superficially at least, the task then is to compare the actual observed data $\mathbf{y}^{\text{obs}}$ with the samples from the prior predictive distribution with the goal of deciding if the collection of actual data appears to be an outlier when compared to the ensemble of perturbed predictions. Chandola et al. [10] review many methods for determining if the observation appears to be an outlier when compared with an ensemble of predictions. For linear model-data relationships and Gaussian prior probability densities for parameters, the Mahalanobis distance provides a useful measure of the difference between the vector of observations and the mean of the ensemble of predictive observations. For nonlinear problems, it might be reasonable to apply kernel principal component analysis to the problem of outlier detection [22] and in some cases, outliers can be detected visually on suitable plots [16].

The aim of this paper is to present a methodology to be used prior to history matching for evaluation of the initial ensemble of large models with large amounts of data, such as those encountered when assimilating production and seismic data. In high-dimensional spaces, it is not straightforward to make a comparison between an actual observation and a small ensemble of samples that are used to represent $p(\mathbf{y}|A)$. It may be useful to instead compare $F(\mathbf{y}^{\text{obs}})$ to $p[F(\mathbf{y}|A)]$, where $F(\cdot)$ is some appropriate functional of the data [7]. In this paper, we will focus on functionals F that mimic the Mahalanobis distance and show how these can be used in practical field cases where the number of samples used to characterize $p(\mathbf{y}|A)$ is on the order of 100, while the dimension of $\mathbf{y}$ is often on the order of 10^3–10^5. The Mahalanobis distance has been used previously in the petroleum literature to ascertain if $\mathbf{d}_{\text{obs}}$ is likely to be a sample from the same distribution that generated the ensemble of predictions [20]. In that study, however, they did not address the situation in which the number of data is larger than the size of the ensemble. When this situation occurs, as it certainly will with 4D seismic data, a straightforward computation of Mahalanobis distance is not possible.

It is easy to create an example illustrating the difficulties with a naive approach to evaluation of the prior ensemble that looks only at the coverage of the data. The left two subplots in Fig. 1a show 10 "observations" (black dots) with insignificant measurement error plotted with an ensemble of model predictions. The model that generated the realizations is multivariate Gaussian. A visual comparison of actual observations to the ensemble of predictions as in Fig. 1a seems to suggest that there is no inconsistency between the "observed data" and the ensemble of "simulated data." If, however, we were to compute the projections of the observed data and the ensemble members on the first two principal components, we see that the prior predictive ensemble does not actually cover the observations (Fig. 1b). We would need to improve the model before attempting to calibrate it (see also [20]).

In the remainder of this paper, we develop a methodology for evaluation of the initial ensemble based on comparison of the prior distribution of simulated data with observed data using the Mahalanobis distance between the actual data and the mean of the ensemble. In Section 2, we describe a method for computing an approximation of the Mahalanobis distance in high dimensions from small numbers of samples. We also discuss how to evaluate the meaning of the Mahalanobis distance for the cases in which the number of data is larger than the ensemble size.

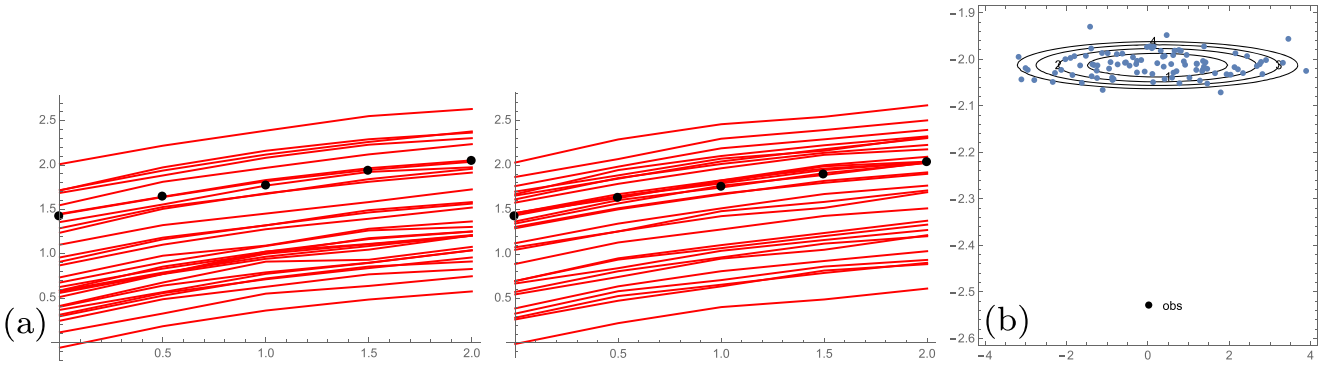

Fig. 1 In **a**, red curves are initial ensembles of predictions at two locations. Black dots show the actual observations, which are "covered" by the ensemble. In **b**, simulated data (blue dots) and the actual observed data (black dot) are projected onto the plane spanned by the first two singular vectors. Contours are distance normalized by variance in the principal directions

2 Approximating the Mahalanobis distance from small samples

The Mahalanobis distance of a vector $\mathbf{x} \in R^m$ from the mean $\boldsymbol{\mu} \in R^m$ of a set of samples with $m \times m$ covariance matrix $\boldsymbol{\Sigma}$ is defined to be as follows:

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}. \tag{2}$$

In general, the mean, $\boldsymbol{\mu}$, and the covariance, $\boldsymbol{\Sigma}$, must be estimated from the set of n_e samples, $\mathbf{X}$. If the samples are possibly contaminated with outliers, then it might be necessary to use robust estimates of the mean and covariance [23]. In our applications to reservoir data assimilation, the samples are from the prior ensemble. We will not require robust estimates, but because our sample size (n_e) is much smaller than the number of data (n_d), the sample covariance is not full rank so its inverse cannot be computed. If we use the pseudo-inverse, the magnitude of the Mahalanobis distance will be much different from the magnitude obtained using the full-rank covariance. Extensive literature exists on improving the estimate of the Mahalanobis distance for outlier detection (e.g., [45]).

We compute an approximation of the Mahalanobis distance [2] using a regularized estimate of the covariance [26, 39] and in particular, we will use Target B of [39] to shrink the sample covariance towards a diagonal matrix with a constant value equal to the average variance, ν. (Note that if the variance is thought to be spatially varying, then it would be appropriate to use a different target.)

Let the estimate of the covariance matrix be as follows:

$$\hat{\boldsymbol{\Sigma}} = \delta\mathbf{T} + (1-\delta)\mathbf{S} \tag{3}$$

where $\mathbf{T}$ is the target covariance matrix, $\mathbf{S}$ is the sample covariance matrix, and δ is the shrinkage parameter, for which we use in line 10 of Algorithm 1 an estimate of the optimal value [28] as follows:

$$\delta = \frac{2}{n_e+2}.$$

For $\mathbf{S}$, we use the maximum likelihood estimate of the sample covariance, i.e.,

$$\mathbf{S} = \frac{1}{n_e-1}\mathbf{X}\mathbf{X}^{\mathrm{T}} \tag{4}$$

where columns of $\mathbf{X}$ are mean removed. For simplicity of notation, we define the following:

$$\hat{\mathbf{X}} = \sqrt{\frac{1-\delta}{n_e-1}}\mathbf{X}$$

and

$$\hat{\mathbf{T}} = \delta\mathbf{T}$$

in which case the regularized estimate of the covariance can be written as follows:

$$\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{T}} + \hat{\mathbf{X}}\hat{\mathbf{X}}^{\mathrm{T}}.$$

Computation of the Mahalanobis distance requires the inverse of the regularized covariance, which can be obtained using the Sherman-Morrison-Woodbury formula as follows:

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}^{-1} &= \left(\hat{\mathbf{T}} + \hat{\mathbf{X}}\hat{\mathbf{X}}^{\mathrm{T}}\right)^{-1} \\ &= \hat{\mathbf{T}}^{-1} - \hat{\mathbf{T}}^{-1}\hat{\mathbf{X}}\left(\mathbf{I} + \hat{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{T}}^{-1}\hat{\mathbf{X}}\right)^{-1}\hat{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{T}}^{-1}.\end{aligned} \tag{5}$$

Since our choice of the target matrix is a scaled identity matrix, i.e., $\mathbf{T} = \nu\mathbf{I}$, the formula for the inverse can be simplified as follows:

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}^{-1} &= \tfrac{1}{\delta\nu}\mathbf{I} - \tfrac{1}{\delta^2\nu^2}\left(\tfrac{1-\delta}{n_e-1}\right)\mathbf{X}\left(\mathbf{I} + \tfrac{1}{\delta\nu}\left(\tfrac{1-\delta}{n_e-1}\right)\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} \\ &= \tfrac{1}{\delta\nu}\mathbf{I} - \tfrac{1}{\delta\nu}\mathbf{X}\left(\delta\nu\left(\tfrac{n_e-1}{1-\delta}\right)\mathbf{I} + \mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}.\end{aligned} \tag{6}$$

To compute the Mahalanobis distance,

$$\begin{aligned} D_M^2(\mathbf{x}) &= (\mathbf{x}-\hat{\boldsymbol{\mu}})^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}). \\ &= \tfrac{1}{\delta\nu}(\mathbf{x}-\hat{\boldsymbol{\mu}})^{\mathrm{T}}(\mathbf{x}-\hat{\boldsymbol{\mu}}) \\ &\quad - \tfrac{1}{\delta\nu}(\mathbf{x}-\hat{\boldsymbol{\mu}})^{\mathrm{T}}\mathbf{X}\left(\delta\nu\left(\tfrac{n_e-1}{1-\delta}\right)\mathbf{I}+\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \\ &\quad \mathbf{X}^{\mathrm{T}}(\mathbf{x}-\hat{\boldsymbol{\mu}}) \\ &= \tfrac{1}{\delta\nu}(\mathbf{x}-\hat{\boldsymbol{\mu}})^{\mathrm{T}}(\mathbf{x}-\hat{\boldsymbol{\mu}}) \\ &\quad - \tfrac{1}{\delta\nu}(\mathbf{x}-\hat{\boldsymbol{\mu}})^{\mathrm{T}}\mathbf{X}\mathbf{L}^{-\mathrm{T}}\mathbf{L}^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{x}-\hat{\boldsymbol{\mu}}) \end{aligned} \tag{7}$$

where we have utilized a Cholesky factorization (also called Cholesky decomposition) of $\left(\delta\nu\left(\frac{n_e-1}{1-\delta}\right)\mathbf{I}+\mathbf{X}^{\mathrm{T}}\mathbf{X}\right) = \mathbf{L}\mathbf{L}^{\mathrm{T}}$. Note that the matrix that required inversion is generally quite small ($n_e \times n_e$), so the cost of factorization is negligible. The product should be performed as follows:

$$\mathbf{y} = \mathbf{L}^{-1}\left(\mathbf{X}^{\mathrm{T}}(\mathbf{x}-\hat{\boldsymbol{\mu}})\right) \tag{8}$$

or by solving $\mathbf{L}\mathbf{y} = \mathbf{X}^{\mathrm{T}}(\mathbf{x}-\hat{\boldsymbol{\mu}})$ in line 14 of Algorithm 1.

Criticism of the initial ensemble is based on the prior predictive distribution, which compares the observation vector $\mathbf{z}$ to the distribution of predictions, represented empirically by the ensemble of predictions $\mathbf{X}$. (In application, we might check to see if the set of actual repeat formation tester (RFT) or production observations is an outlier when compared to an ensemble of perturbed predictions from the initial ensemble.) If the ensemble is large enough, the comparison could be made using the Mahalanobis distance from $\mathbf{z}$ to the mean of the ensemble $\hat{\boldsymbol{\mu}} = \sum \mathbf{X}/n_e$ line 2 of Algorithm 1,

$$D_M^2(\mathbf{z},\hat{\boldsymbol{\mu}}) = (\mathbf{z}-\hat{\boldsymbol{\mu}})^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{z}-\hat{\boldsymbol{\mu}}) \tag{9}$$

where $\hat{\boldsymbol{\Sigma}}$ is an estimate of the population covariance of $\mathbf{X}$.

In most practical history matching cases, the maximum likelihood estimate of the covariance (4) will not be full rank, so that some type of regularization must be applied before inversion. In that case, however, the magnitudes of the Mahalanobis distance will not be distributed as chi-squared with n_d degrees of freedom [18]. One way to decide if the initial ensemble is adequate is to compare the prior distribution of an approximation to the Mahalanobis distance to an equivalent approximation for the observations.

To generate a predictive distribution of values of $D_M^2(\mathbf{x},\hat{\boldsymbol{\mu}})$, we require samples $\mathbf{x}$ from the initial distribution. Our only source for these is the ensemble of prior samples. We take ensemble members one-at-a-time, and compute the distance between the selected sample $\mathbf{x}_i$ and the mean obtained by leaving the ith sample out.

$$D_M^2(\mathbf{x}_i,\hat{\boldsymbol{\mu}}_i) = (\mathbf{x}_i-\hat{\boldsymbol{\mu}}_i)^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_i), \tag{10}$$

where $\hat{\boldsymbol{\Sigma}}_i$ is the covariance estimate obtained by leaving the ith sample out. After looping through all ensemble members in $\mathbf{X}$, we have n_e samples of $D_M^2(\mathbf{x},\hat{\boldsymbol{\mu}})$. These define our prior predictive distribution of Mahalanobis distance.

An evaluation of $D_M^2(\mathbf{z},\hat{\boldsymbol{\mu}})$ is required for comparison with the prior predictive distribution of $D_M^2(\mathbf{x},\hat{\boldsymbol{\mu}})$. A straightforward approach would be to use $\hat{\boldsymbol{\Sigma}}$ from the entire ensemble, but the magnitude can be quite sensitive to the dimension of the ensemble. Thus, we instead compute an ensemble of estimates as follows:

$$D_M^2(\mathbf{z},\hat{\boldsymbol{\mu}}_i) = (\mathbf{z}-\hat{\boldsymbol{\mu}}_i)^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{z}-\hat{\boldsymbol{\mu}}_i). \tag{11}$$

where each covariance estimate is computed by leaving one out. The spread in these estimates is typically quite small when the ensemble size is of order 100. From the reference ensemble of $D_M^2(\mathbf{x}_i,\hat{\boldsymbol{\mu}}_i)$, we create an empirical cumulative distribution function (CDF), and then compare the median of the distribution of $D_M^2(\mathbf{z},\hat{\boldsymbol{\mu}})$ to the CDF.

One limitation with the use of the empirical CDF is that it is unable to discriminate the difference between two values that are outside of the range of values in the reference ensemble. To make that discrimination, we can simply use a measure of distance of the "observed" Mahalanobis distance from the mean or median of the predictive distribution. Here, we choose to use the modified z-score [24], which is a normalized measure of distance from the median.

Thus, in addition to providing an estimate of the probability from the empirical CDF obtained from the prior predictive distribution, we also report the modified z-score in line 23 of Algorithm 1 as follows:

$$z\text{-score} = 0.6745\,\frac{\mathrm{median}\left(D_M^2(\mathbf{z},\hat{\boldsymbol{\mu}})\right)-\mathrm{median}\left(D_M^2(\mathbf{x},\hat{\boldsymbol{\mu}})\right)}{\mathrm{MAD}} \tag{12}$$

where the median absolute deviation (MAD) is computed from the ensemble of $D_M^2(\mathbf{x}_i,\hat{\boldsymbol{\mu}}_i)$. Algorithm 1 shows the steps of this methodology.

3 Applications of methodology

We illustrate the methodology with two examples. In the first example, the reference ensemble is Gaussian and the test observation vectors come from three distributions. In the second example, we use actual production, repeat formation tester (RFT) and seismic data from a segment of the Norne field to assess the adequacy of the reservoir simulation model and seismic model prior to data assimilation.

3.1 1D toy test cases

This toy problem tests the ability to discriminate outliers in a high-dimensional space when ensemble size is modest. The "observations" are linear and the prior is Gaussian so the problem is quite simple. The ensemble size is, however, realistic ($n_e = 100$) and the dimension of the

Algorithm 1 Computation of Mahalanobis distance between median of ensemble of simulated data $\mathbf{X}$ and observed data $\mathbf{z}$.

function COMPUTEMAHALANOBIS($\mathbf{X}$,$\mathbf{z}$,$\mathbf{L}$)
  $D_M \leftarrow \frac{1}{\delta\nu}\left(\mathbf{z}^{\mathrm{T}}\mathbf{z} - (\mathbf{L}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{z})^{\mathrm{T}}(\mathbf{L}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{z})\right)$
  return D_M ▷ approximation to Mahalanobis distance squared
end function

function MODELDIAGNOSTICCV($\mathbf{X}$, $\mathbf{z}$)
input: $\mathbf{X}$ is the ensemble of simulated data , $\mathbf{z}$ is a vector observed data
  $n_x \leftarrow \text{shape}(\mathbf{X})[0]$
  $n_e \leftarrow \text{shape}(\mathbf{X})[1]$
  $n_s \leftarrow n_e - 1$ ▷ Ensemble size after 'leave-one-out'
  $\nu \leftarrow \text{var}(\mathbf{X})$
  $\delta \leftarrow \frac{2}{n_s+2}$ ▷ This is the amount of shrinkage
  for $i \leftarrow 0, n_e - 1$ **do**
    $\mathbf{X}_s \leftarrow \mathbf{X}_{i/}$ ▷ Delete the ith column of $\mathbf{X}$
    $\mathbf{A}_s \leftarrow \delta\nu\frac{(n_s-1)}{(1-\delta)}\mathbf{I} + \mathbf{X}_s^{\mathrm{T}}\mathbf{X}_s$
    $\mathbf{L}_s\mathbf{L}_s^{\mathrm{T}} \leftarrow \mathbf{A}_s$ ▷ Cholesky factorization of $\mathbf{A}_s$
    $\hat{\boldsymbol{\mu}} \leftarrow \frac{1}{n_s}\sum_{j=0}^{n_s-1}\mathbf{x}_{s,j}$
    $\mathbf{x} = \mathbf{X}_i$
    $D_i^x \leftarrow$ COMPUTEMAHALANOBIS($\mathbf{X}_s, \mathbf{x} - \hat{\boldsymbol{\mu}}, \mathbf{L}_s$)
    $D_i^z \leftarrow$ COMPUTEMAHALANOBIS($\mathbf{X}_s, \mathbf{z} - \hat{\boldsymbol{\mu}}, \mathbf{L}_s$)
  end for
  $\sigma_x = 1.4826 * \text{MAD}(\mathbf{D}^x)$ ▷ Robust measure of scale
  $\mu_x \leftarrow \text{med}(\mathbf{D}^x)$
  $\mu_z \leftarrow \text{med}(\mathbf{D}^z)$
  z-score $\leftarrow \frac{\mu_z-\mu_x}{\sigma_x}$ ▷ Empirical estimate of $p(\mu_{DY}|X)$
  ecdf $\leftarrow$ ECDF($\mathbf{D}^x$)
  $p_z \leftarrow \text{ecdf}(\mu_z)$
  return $1 - p_z$, z-score
end function

data vector ($n_d = 1000$) is large enough to evaluate the methodology with different approximations to the inverse of the covariance matrix.

We first define a "reference ensemble," which would correspond to the perturbed predictive ensemble in applications where the objective is to determine if the data vector is consistent with the initial ensemble. In the reference ensemble, the realizations are Gaussian with stationary mean equal to zero, and stationary covariance that is Gaussian with practical correlation range of 25, i.e.,

$$\rho_{\text{ref}}(x) = \exp\left(-3\left(\frac{x}{25}\right)^2\right) \tag{13}$$

where $\rho_{\text{ref}}(x)$ is the covariance of Gaussian random variables whose locations are separated by distance x.

We then create test vectors from three different distributions:

Case 1: Test vector is from the reference distribution (Fig. 2a)

Case 2: Test vector has mean 0, but Gaussian covariance with range 50 (Fig. 2b)

Case 3: Test vector has a non-stationary mean and a smaller variance (Fig. 2c)

We compare three different methods for approximating the Mahalanobis distance in the small ensemble situation. The simplest approach would be to use the diagonal matrix whose entries are the variance of the predictive distribution to approximate the covariance matrix used in the distance measure. A second approach is to approximate the inverse of the sample covariance matrix using the pseudo-inverse based on the singular value decomposition of the sample vectors. A third approach is to use the inverse of the shrinkage estimate of the covariance as described in Section 2. We compare results from these three approaches when applied to cases 1–3. Because the conclusions might be significantly different for different realizations of the test vector, we apply the tests for 100 different test vectors.

Our criterion for comparison will be based on the ability to judge correctly that the test vector is not from the

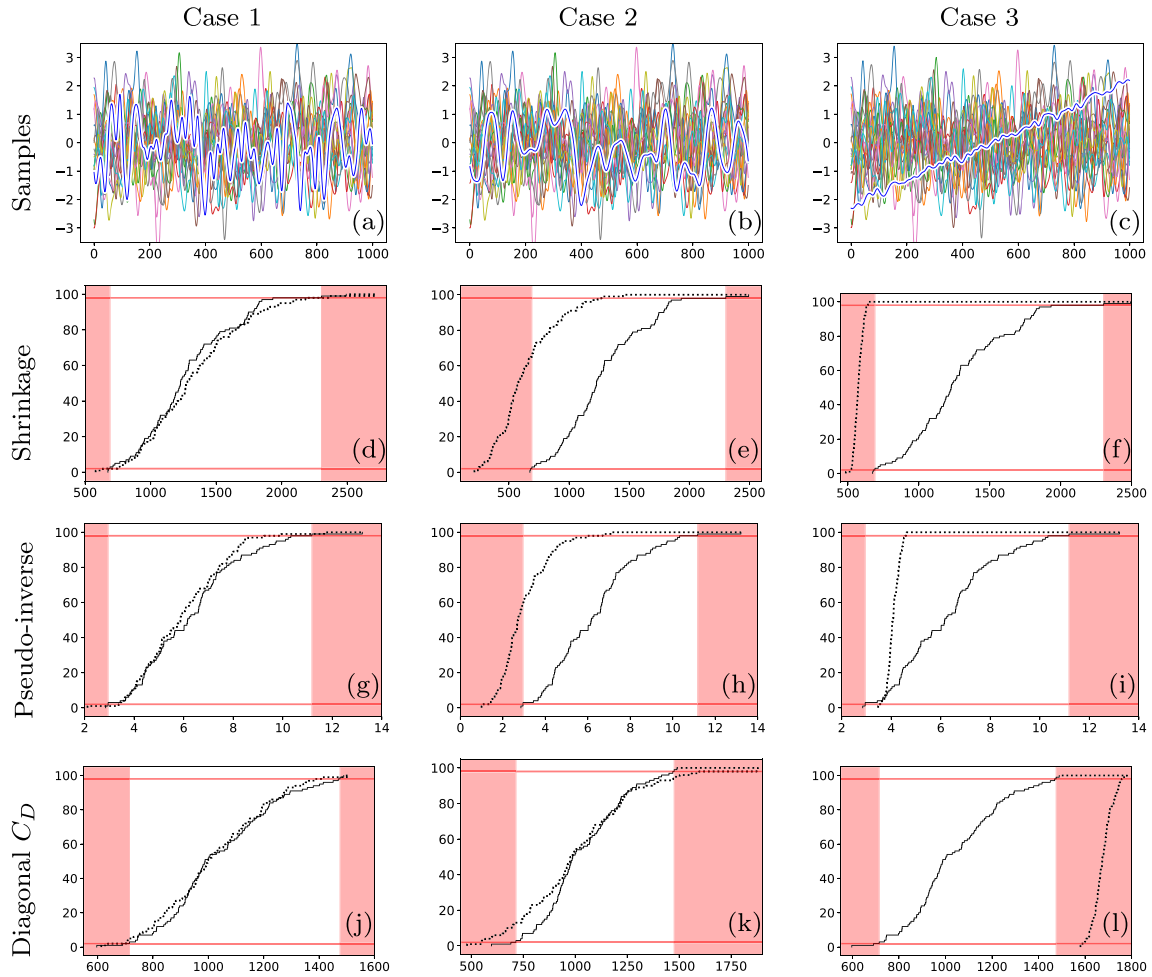

Fig. 2 Summary of discrimination results using distribution of Mahalanobis distance for three distributions of test vectors (blue-white curves in top row) and three different methods of approximating the precision matrix. Black curves in (**d–l**) show the empirical CDFs of Mahalanobis distance for samples from the reference distribution. Dotted curves show the empirical CDF of Mahalanobis distance of test vectors from the three different distributions. Red shaded areas are rejected samples at 2% level

reference distribution. Our decision will be to accept the test vector if its Mahalanobis distance is in the 96% credible intervals based on the empirical CDF (the unshaded regions in Fig. 2d–l). We then examine the frequency that this test correctly identifies violations of the assumptions, i.e., for cases 2 and 3 we identify what fraction of the test vectors would be accepted (falling in the unshaded regions).

Figure 2 a, b, and c shows the ensemble of perturbed predictions from the reference distribution with a single realization of the test vector. In case 1, test data are drawn from the reference distribution. In case 2, test data are drawn from a distribution with longer covariance range. In case 3, test data are drawn from a distribution with a trend in the mean. In each of the three cases, the test vector is "covered" by the reference ensemble when the data are examined one-by-one as in the plots. Although the coverage appears to be good in each case, visual inspection of the entire sequence shows that the test vectors in Fig. 2b and c are unlikely to be samples from the same distribution from which the reference ensemble was drawn. Because the Mahalanobis distance uses the covariance of the reference distribution and looks at the entire data set simultaneously, it offers a more powerful discriminator than simple one-by-one data coverage. In this example, however, it fails to clearly identify the discrepancy for case 2.

When the test vector is a sample from the same distribution as the reference distribution, the predictive distributions for any of the three approximations to the Mahalanobis distance would give similar results—rejecting a small number of correct test samples (Fig. 2d, g, j). Results from cases 2 and 3 are more interesting. In case 2, where the covariance range for the test vector differs from the covariance range of realizations in the reference ensemble, a test based on the Mahalanobis distance using the diagonal covariance matrix would accept 81% of the realizations of the test vector, while the use of

the pseudo-inverse would accept 39% and the shrinkage-based covariance would accept 31%. In case 3, where the trend for the test vector differs from the trend in the reference ensemble, both the diagonal and shrinkage-based methods correctly reject all test vectors, while the method based on the pseudo-inverse fails to reject any. In these tests, only the shrinkage-based approximation of the covariance provided consistently useful discrimination. Table 1 summarizes numerically the results shown graphically in Fig. 2.

3.2 Application to Norne G-segment

In this application, we use production data, RFT data, and seismic data from the G-segment of the Norne field to evaluate the suitability of the reservoir flow model and the seismic model before assimilation of data. We then improve the model by adding uncertainty to parameters that had been fixed in the initial model, and repeat the model criticism step. One reason for use of the Norne G-segment is that the pressure behavior and transport of water in this part of the field appear to be relatively complex.

3.2.1 Initial model creation

The Norne field is an oil field located in the Norwegian Sea, approximately 200 km from the coast of Norway. Production from the field began in 1997 and continues to the present. Data from the period 1997 to 2006, including repeat seismic surveys, production rates, a reservoir simulation model, and geologic reports, were released by Equinor and partners in 2010 in conjunction with an SPE Advanced Technology Workshop. The data have been used repeatedly to test various history matching methods (e.g., [13, 15, 29, 32, 38, 41, 47]). In this paper, we focus on the evaluation of the initial ensemble *before history matching* [33]. Because the purpose of this manuscript is to present methods for evaluation of the initial ensemble, we use the initial ensemble of reservoir properties from Chen and Oliver [13], which was based on properties in the reference model provided by Equinor. Variability was added where it was believed that the values of the model parameters were necessarily uncertain and that the uncertainty would be needed for history matching or for forecast assessment. In all cases, the assessment of uncertainty was subjective, but often based on observed spatial variability. For some properties, such as porosity, the variability can be estimated from logs. For other properties such as relative permeability, no actual data were available, so uncertainty was represented simply by variability in the endpoint relative permeability. Fault transmissibility multipliers were assumed to be uncertain to the extent that 90% of the probability fell in the range bounded below by 0.1 of the base estimate for fault transmissibility multipliers, and bounded above by 10 times the base estimate.

The flux between the various reservoir segments in the Norne reference model was also governed by a transmissibility multiplier MULTREGT. In general, the location of the application of these transmissibility multipliers corresponded to the location of seismic faults or to vertical flow barriers. Values for these multipliers had been set in the base model to approximately match observed pressure behavior. When the exploration well 6608/10-4 was drilled in the G-segment in 1993, RFT data showed an apparently continuous pressure profile above and below the Not shale, and the pressures seemed to be in equilibrium with the rest of the field. On the other hand, when RFT data in the E-4H well were recorded in March 2000, the pressure above the Not shale had increased approximately 5 bars, while the pressure below the Not shale had decreased approximately 25 bars. This was stated as evidence that the G-segment was in communication with the main field (although it has a different oil-water contact).[1] Although the RFT data showed that the G-segment must be in pressure communication with the main part of the field, it appeared that the pressure support was relatively weak because production at E-4 AHT2 (hereafter referred to as E-4 AH for conciseness) declined rapidly and had to be shut in for lack of pressure support. To allow for a small amount of communication between the G-segment and the main C-segment, the base reservoir simulation model created by Equinor assigned MULTREGT values of 0.005 in the Garn formation, 0.01 in the Ile formation, and 0.01 in the Tofte formation at the boundary between segments. These values remained fixed in the initial ensemble.

Table 1 Summary of discrimination results using distribution of Mahalanobis distance for three different methods of approximating the precision matrix

	Approximation to inverse covariance		
	Diagonal	Pseudo-inverse	Shrinkage
Case 1	94	100	98
Case 2	81	39	31
Case 3	0	100	0

The numbers in each cell are the percentage of test vectors that are accepted. For cases 2 and 3 low numbers are better. The two "boxed" values indicate methods that have failed at discrimination

[1] Geological and Petrophysical Report, Norne Field, PL 128, Wells 6608/10-E-4 H, 6608/10-E-4 T2 H, 6608/10-E-4 AH, 6608/10-E-4 A T2 H, Norne PETEK, March 2001

3.2.2 Petro-elastic model (PEM)

A rock-physics or petro-elastic model (PEM) is a set of relationships that aim to convert certain reservoir properties (e.g., porosity and NTG ratio) and reservoir state variables (e.g., saturations and pressures) into elastic properties, such as velocities, densities, and impedances. The Chen and Oliver [13] history match of Norne did not include a PEM because it did not assimilate seismic data. We therefore need to choose a PEM to generate acoustic impedance predictions from the reservoir simulation model and simulation outputs for comparison with actual seismic impedance (I_p) data. Details of our PEM model and the methodology used for simulation of the acoustic impedances can be found in Appendices A and B, respectively.

3.2.3 Data for model criticism

Because we focus on the G-segment, the data is limited to production data from the E-4AH well (Fig. 5c), injection rates from the F-4H well, RFT data from both wells (Fig. 5a), and inverted acoustic impedance data at four survey times. The production and RFT data have been described in previous publications, but the seismic data requires description as it was inverted for this study. The Norne seismic dataset consists of four pre-stack time-migrated volumes acquired in 2001, 2003, 2004, and 2006. In order to improve the repeatability between the seismic surveys, various 4D seismic calibration steps were applied [3]. We subsequently performed a sequential 4D post-stack seismic inversion workflow on the calibrated volumes in order to generate inverted acoustic impedance data for each of the four seismic surveys. This process involved well log QC, horizon QC, wavelet estimation, low-frequency model construction, inversion parameterization, and a linear and deterministic seismic inversion step. The four seismic impedance volumes were then converted from the two-way time (TWT) domain to the depth domain using a calibrated version of the interval velocity model provided in the Norne dataset. Figure 3 (top row) shows the Ip2006-Ip2001 4D difference estimated from the inverted seismic data in the G-segment.

In criticism of the initial ensemble, characterization of the magnitude and correlation of errors in data are of secondary importance, but for RFT data, we assumed uncorrelated measurement errors with standard deviation of 0.01 bar, which appears to be close to the actual measurement error. For production rates, we assumed that errors are unbiased and uncorrelated with standard deviations of 50 m^3/days, which is lower than the value used in [13]. Errors in the inverted seismic acoustic impedance data were estimated using factorial co-Kriging [2]. From this process, we obtained variogram models corresponding to the 3D seismic noise in each seismic survey (Ip2001, Ip2003, Ip2004, and Ip2006) and the 4D seismic noise in each of the 4D impedance volumes: Ip2003-Ip2001, Ip2004-Ip2001, and Ip2006-Ip2001. Since the variogram structures for all the 3D noise models are very similar, we chose to use as 3D seismic noise the model obtained from the independent Ip2001 residuals. Similarly, as all the 4D noise models are very similar to one another, we selected the 4D seismic noise obtained from the Ip2006-Ip2001 seismic differences. In each case, the errors appear to be spatially correlated. For the 3D noise in acoustic impedance, the covariance estimate has a nugget effect with amplitude 983 (units of impedance squared) and a cubic covariance with amplitude 8347 and ranges of 10, 8, and 4 grid cells in the i, j, k directions, respectively. The covariance estimate of the 4D seismic noise, on the other hand, has a nugget effect with amplitude 1,236 (units of impedance squared) and a cubic covariance with amplitude 27,260 and ranges of 10, 9, and 5 grid cells in the i, j, k directions, respectively. These covariance structures were used to generate 100 realizations of 3D seismic noise and 100 realizations of 4D seismic noise. A single realization of 3D seismic noise and a realization of 4D seismic noise based on our characterization are shown in Fig. 4. The realizations of seismic noise were then added to the impedance predictions to generate perturbed impedance predictions for comparison with actual seismic data (Fig. 3, middle and bottom rows).

3.2.4 Initial model criticism

We used several subsets of data from the G-segment of the Norne field to challenge the model. Our subsets included RFT data from the E-4AH and F-4H wells, both separately and jointly. The purpose of examining the data sets jointly is to ensure that the pressure behavior at both wells can be jointly explained by the same set of parameters. We also compared predictions of well water production rate (WWPR) and well gas production rate (WGPR) at well E-4AH to actual observations. Finally, we compared the initial ensemble of perturbed simulated acoustic impedance to the observations of acoustic impedance and to changes in impedance (4D differences).

Figure 5 summarizes visually the most important aspects of the comparison of the perturbed prior predictions to the observations. First, we observe from the joint RFT data that the ensemble has (nearly) enough variability to cover the observations, but the discontinuity in observed pressure for the F-4H well at depth approximately 2640 m is not seen in the perturbed predictions (Fig. 5a). This failure is identified by the Mahalanobis distance which falls slightly outside of the credible interval (Fig. 5b). The observed water production rate at well E-4AH (Fig. 5c) shows early water breakthrough in a substantial number of model realizations,

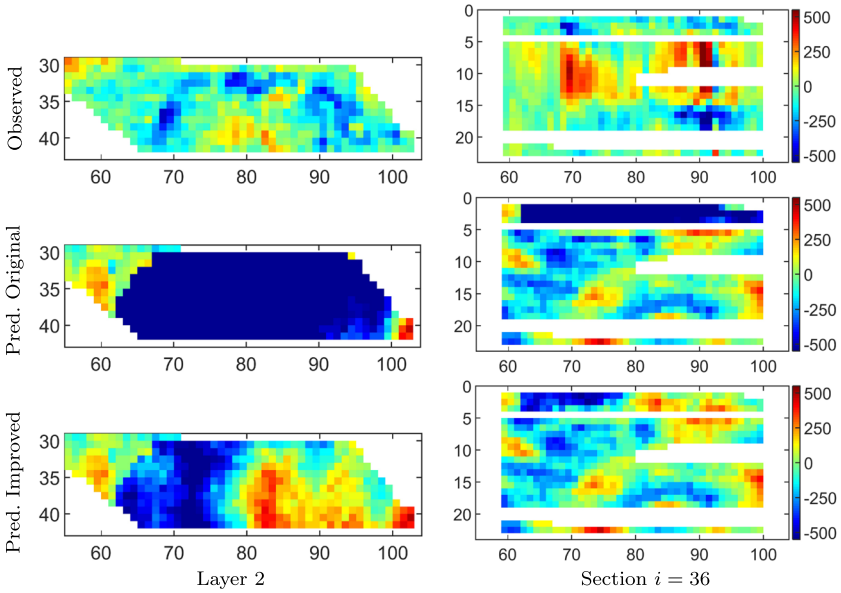

Fig. 3 Observed Ip2006-Ip2001 and realization 1 of perturbed simulated Ip2006-Ip2001 from the original and the improved ensembles. Units of impedances: (m/s).(g/cc)

but is not identified as inconsistent (Fig. 5d) because a few of the models agree with the data by not predicting early water production.

The ensemble of perturbed predictions of gas production at well E-4AH (Fig. 5e) covers the observed gas production and appears, from the Mahalanobis distance, to be consistent with the observations (Fig. 5f). Although the mean simulated pressure from the prior ensemble is much too high, and consequently the mean simulated change in I_p is much too negative in the Garn formation of the G-segment (Fig. 5g), there is sufficient variability in the pressure behavior and in the I_p behavior that the Mahalanobis distance for acoustic impedance change does not invalidate the model (Fig. 5h).

Modified z-scores (12) and cumulative probabilities for evaluation of the credibility are shown in the first column of results in Table 2. In addition to results described above, we note the clear inconsistency of the RFT observations from well F-4H in the Garn formation with the predictions. The failure is much larger when data from the Garn formation is evaluated separately.

3.2.5 Initial model improvement

The G-segment in the Norne field is fairly complex, with several identified faults and barriers to vertical flow. The degree of connectivity with the main part of the field is uncertain. The operators of the field concluded that there must be some connectivity between the G- and C-segments because there was pressure drawdown observed in well E-4H when it was drilled. Yet, the producer E-4AH died fairly quickly due to lack of pressure support. Also, we note that almost all simulation models of the G-segment predict early water breakthrough at well E-4AH, but this behavior was not observed in the field (Fig. 5c). The RFT in well F-4H, which was drilled about 1 year after E-4AH began producing, showed a discontinuity in pressure in the Garn formation. The simulated gas production at well E-4AH,

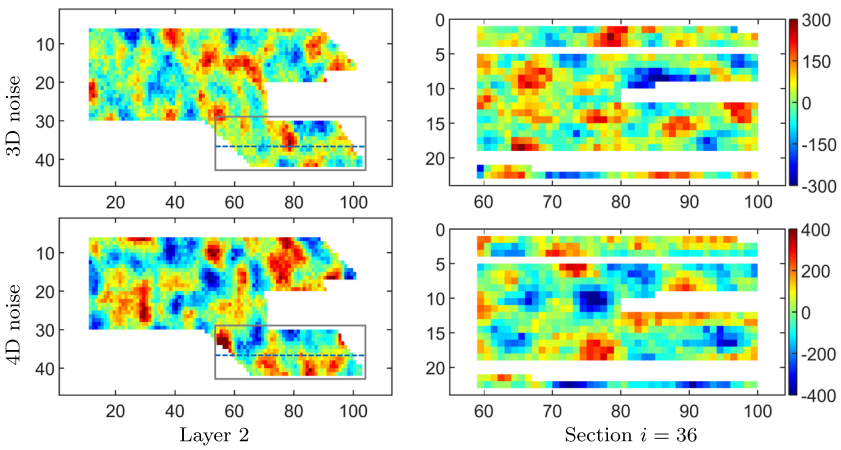

Fig. 4 One realization of 3D seismic noise (top) and one realization of 4D seismic noise (bottom) for the Norne model. The G-segment is inside the rectangular area. Units of impedances: (m/s).(g/cc)

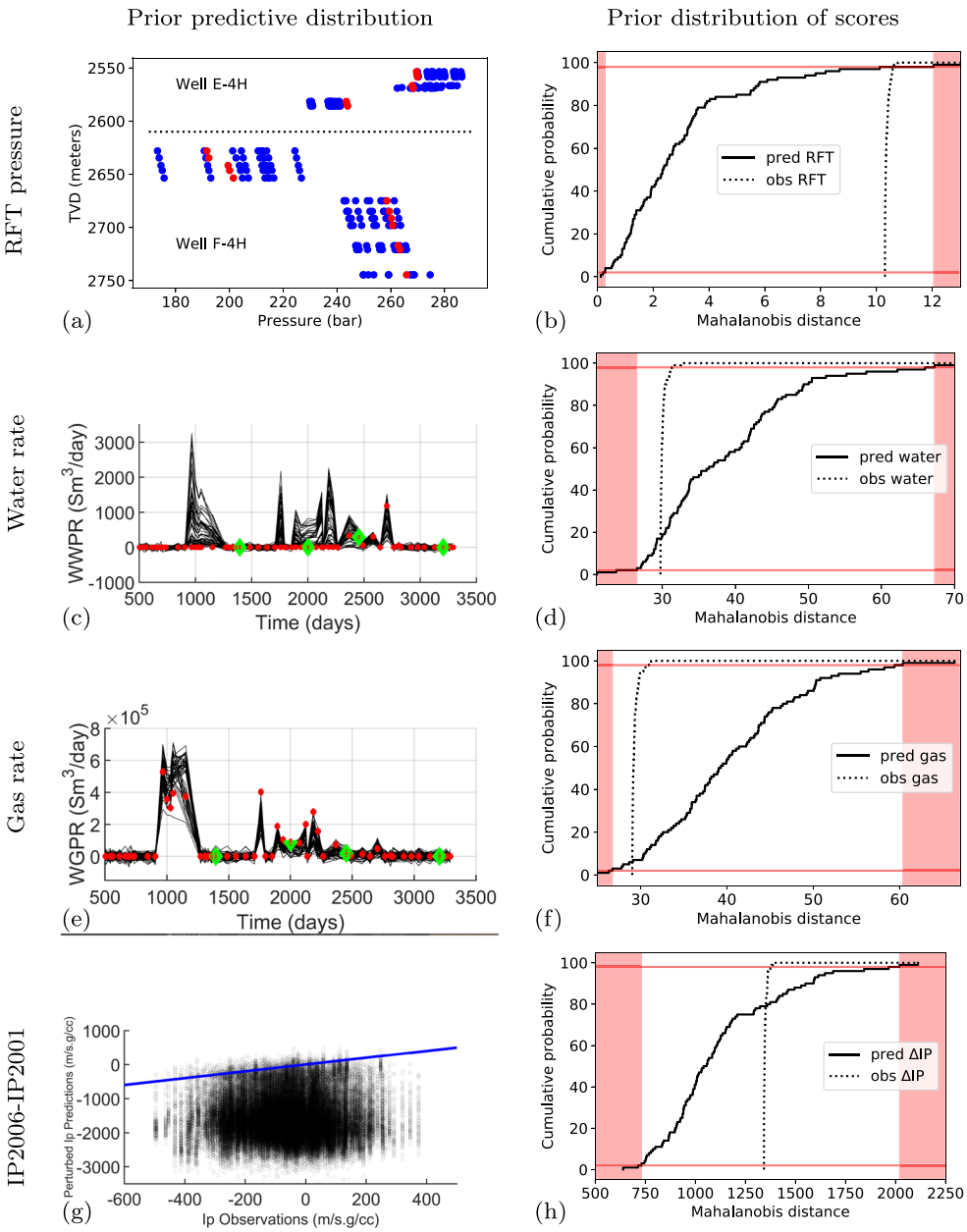

Fig. 5 Prior predictive distribution based on the initial model of uncertainty. Top row shows 20 realizations of perturbed simulated RFT data (blue) in wells E-4H (March 2000) and F-4H (July 2001) compared with observations (red). Same for water production rate in well E-4AH, with the times of the 4D seismic surveys shown by the green diamonds (middle row), and for change in simulated acoustic impedance in the Garn formation of the G-segment (bottom row). Right column compares the simulated distribution of scores (solid curves) with the distribution of estimated scores for the observations (dashed curves)

Table 2 Model diagnostics for various subsets of data from the initial ensemble, and from the ensemble after model improvement

		Initial model		Improved model		Move completion	
Data	Well	Score	95% CI	Score	95% CI	Score	95% CI
	E-4H and F-4H	3.3	0.98	2.2	0.98	− 0.02	0.50
	E-4H	9.0	1.0	5.4	1.0	4.4	0.97
RFT	F-4H	3.7	0.94	4.2	0.96	0.15	0.58
	F-4H (Garn)	755	1.0	192	1.0	6.8	0.92
WWPR		− 0.5	0.17	− 1.0	0.03	0.7	0.83
WGPR	E-4AH	− 0.8	0.04	− 0.6	0.12	− 0.7	0.12
Ip 2001		1.4	0.91	0.8	0.76	0.7	0.75
ΔIp (4D)	Garn only	0.9	0.80	0.5	0.69	− 0.3	0.40

Second column for each case shows the fraction of realizations that fall in the 95% credible interval (CI)

on the other hand, seems consistent with the observed gas production. Finally, the average pressure change simulated by the models was much larger than the observed pressure changes, so the average simulated impedance change was much different from the observation (Figs. 3 and 5g). The different obtained results highlight the importance of criticizing the initial model using subsets of data separately.

The prior predictive distributions for RFT data, production data, and impedance data (Table 2) confirm that some aspects of the uncertainty model for the G-segment are inconsistent with the observed behavior. The model of uncertainty should either be improved, or the model for observation error should be modified to reflect the presence of model error. Unfortunately, it is not straightforward to automate the process of model improvement, so we apply an empirical method in which we first identify the data that was inconsistent with the prior model, and then try to identify possible reasons for the inconsistency. To address the problem of inconsistency in simulated pressure, we added uncertainty in the connection between the G- and C-segments of the Norne field (MULTREGT parameters). We also decreased the mean vertical transmissivity (MULTZ parameter) between layers 1 and 2 to allow simulation of pressure discontinuities, and decreased the mean multiplier for the endpoint relative permeability of water (KRW parameter) in the Garn formation, thinking that it might delay the advance of water. Table 4 in Appendix C summarizes these changes to the simulation model. We also added uncertainty to the PEM (Table 3), not because the model diagnostic indicated inconsistency, but simply because the parameters are uncertain.

The center column of Table 2 shows the scores and probabilities after the initial attempt at model improvement. Visual comparisons of observations and perturbed predictions for three types of data are shown in Fig. 6. Improvements can be noted in the modeling of gas production data and impedance data, but simulated RFT data are still inconsistent with actual observations (Table 2). It is also noted that most model realizations still simulate early breakthrough in water production (Fig. 6c), and that the score for this type of data increases in the new model (Table 2). The ensemble of simulated water production cannot, however, be deemed inconsistent with the WWPR observations (Fig. 6d). Although the ensembles of perturbed simulated impedance data and change in impedance data were both consistent with the observed impedance data in the original model, the model is further improved when uncertainty is added to the parameters of the PEM (Fig. 3, bottom row, Table 2 and cross-plot in Fig. 6g). After the changes to the initial ensemble, however, some actual observations are still inconsistent with the distribution of predictions.

In order to make the prior predictive distribution consistent with RFT observations, several additional changes had to be made to the model. Because of vertical offsets along several faults, layers 1 and 2 are in communication in the G-segment. It does not appear to be possible to simulate a pressure discontinuity between layers 1 and 2 at well F-4H by reducing vertical transmissibility at the base of layer 1 without also reducing the fault transmissibilities. The resulting simulated pressures will then have a pressure discontinuity in the Garn, but it would not be similar to the observed discontinuity. The observed pressure had fallen farther in layer 1 than in layer 2 (Fig. 5a), but the producing well was supposedly in layer 2, which should have resulted in lower pressure in layer 2 (reversal of the observed). In order to simulate the observed behavior, we believe that it is necessary to allow for the possibility that well E-4AH was actually completed in layer 1 (above the cementation barrier), not below it. Figure 7 shows the visual comparisons of observations and perturbed predictions in the new model with changed completion interval for well E-4AH. These changes, coupled with a reduction in flow across faults, allowed the pressure discontinuity at well F-4H to be simulated (see Fig. 7a). Results from the prior predictive distribution for RFT after these changes are greatly improved (right column of Table 2 and Fig. 7b). Although an error in completion location for well E-4AH is almost certainly not the only way to explain the data, it does not appear to be inconsistent with uncertainty in layer depths or location of the well path. Comparison of simulated water production, gas production rates, and seismic impedances were either unchanged or improved by the change (see WWPR, WGPR, and Ip results in Table 2). It is noted that after the change, most models no longer predict early water breakthrough (Fig. 7c), and that the perturbed predictions of WWPR are consistent with the observations (Fig. 7d). Despite the slight increase in the WGPR score in the new model (Table 2), the ensemble of perturbed predictions of gas production remains consistent with the observations (Fig. 7e, f). Results from the prior predictive distribution for impedances and changes in impedances are also improved (Table 2 and Fig. 7h), with a number of model realizations predicting pore pressures which yield impedance values (less negative impedance changes) that are closer to the observed impedance changes (Fig. 7g).

3.2.6 Discussion of data dimension

We evaluate the adequacy of our prior model by asking the question "Do the data appear to be sampled from the initial ensemble of perturbed prediction observations?" If the answer is "yes," then we can consider using the model and the ensemble for data assimilation. If the answer is "no" then we must reevaluate our predictive model and attempt to improve it, since the data have shown that it is inadequate.

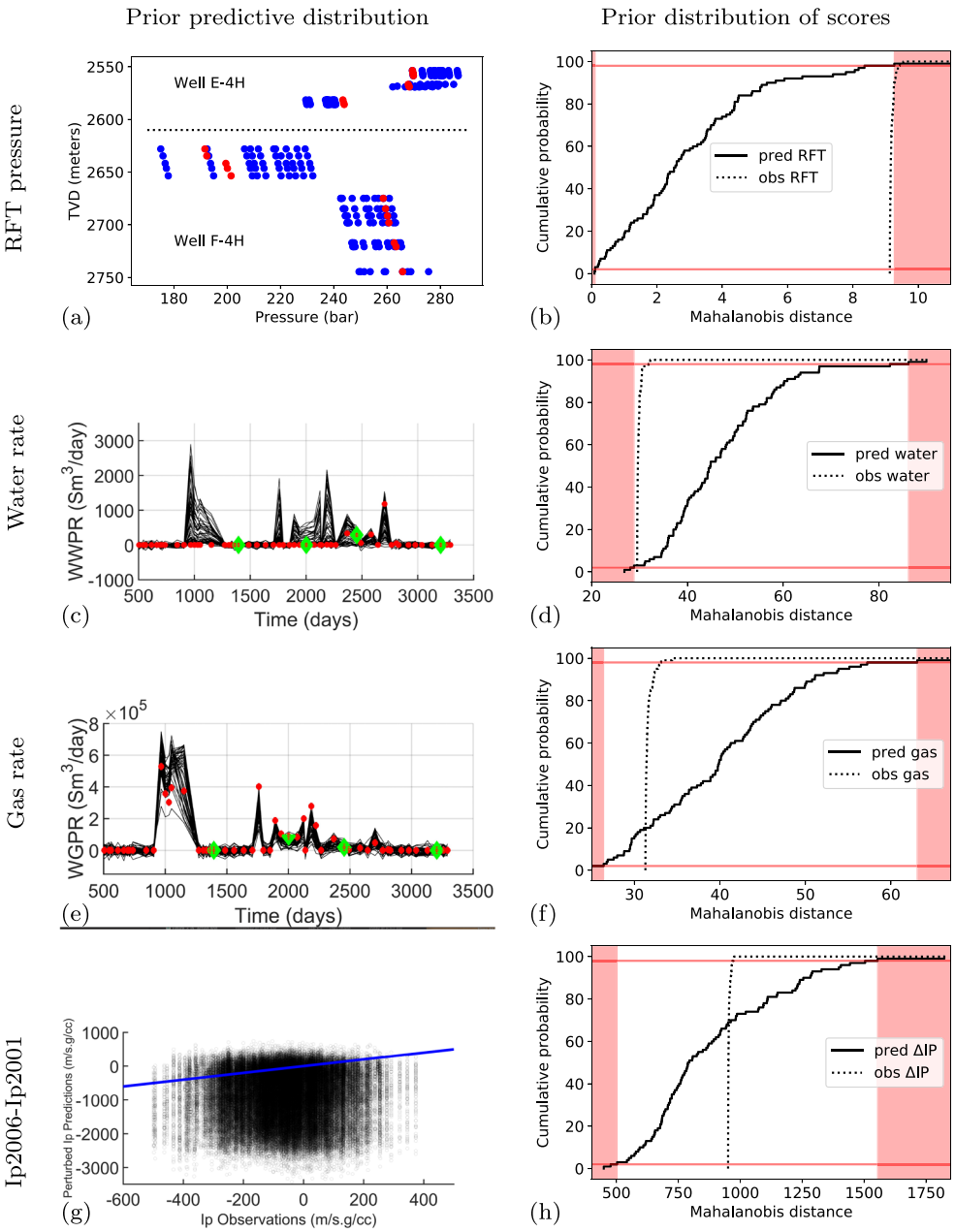

Fig. 6 Prior predictive distribution based on the improved model. Top row shows 20 realizations of perturbed simulated RFT data (blue) in wells E-4H (March 2000) and F-4H (July 2001) compared with observations (red). Same for water production rate in well E-4AH, with the times of the 4D seismic surveys shown by the green diamonds (middle row), and for change in simulated acoustic impedance in the Garn formation of the G-segment (bottom row). Right column compares the predicted distribution of scores (solid curve) with the distribution of estimated scores for the observations (dashed curve)

In fact, the most common tests of adequacy of the initial ensemble are prone to accept poor initial ensembles and poor models. One reason is that it is standard practice to compare individual observations with the ensemble of predictions using the concept of "coverage." Unfortunately, coverage alone is not a powerful method of model checking as coverage of an observation by an ensemble of perturbed observations can often be obtained by simply increasing uncertainty or variability in a single model parameter. The toy problem shown in Fig. 1 illustrated this failure to discriminate. In that case, the model appeared perfectly adequate based on coverage, but was inadequate when data from both series were evaluated simultaneously. On the other hand, if all observations in a large data set are simultaneously compared to model predictions using a measure of proximity, then we suffer from a curse of dimensionality and the concept of proximity or distance may not even be meaningful, depending on the exact metric used [1, 25].

Consequently, it is generally useful to evaluate subsets of the data for which the model errors are expected to be correlated. Hence, when we evaluate the Norne model, we evaluate RFT data separately from seismic and production data. The largest scores are observed when we focus on RFT data in the Garn formation only, but even when all RFT from both wells are evaluated simultaneously, the initial ensemble

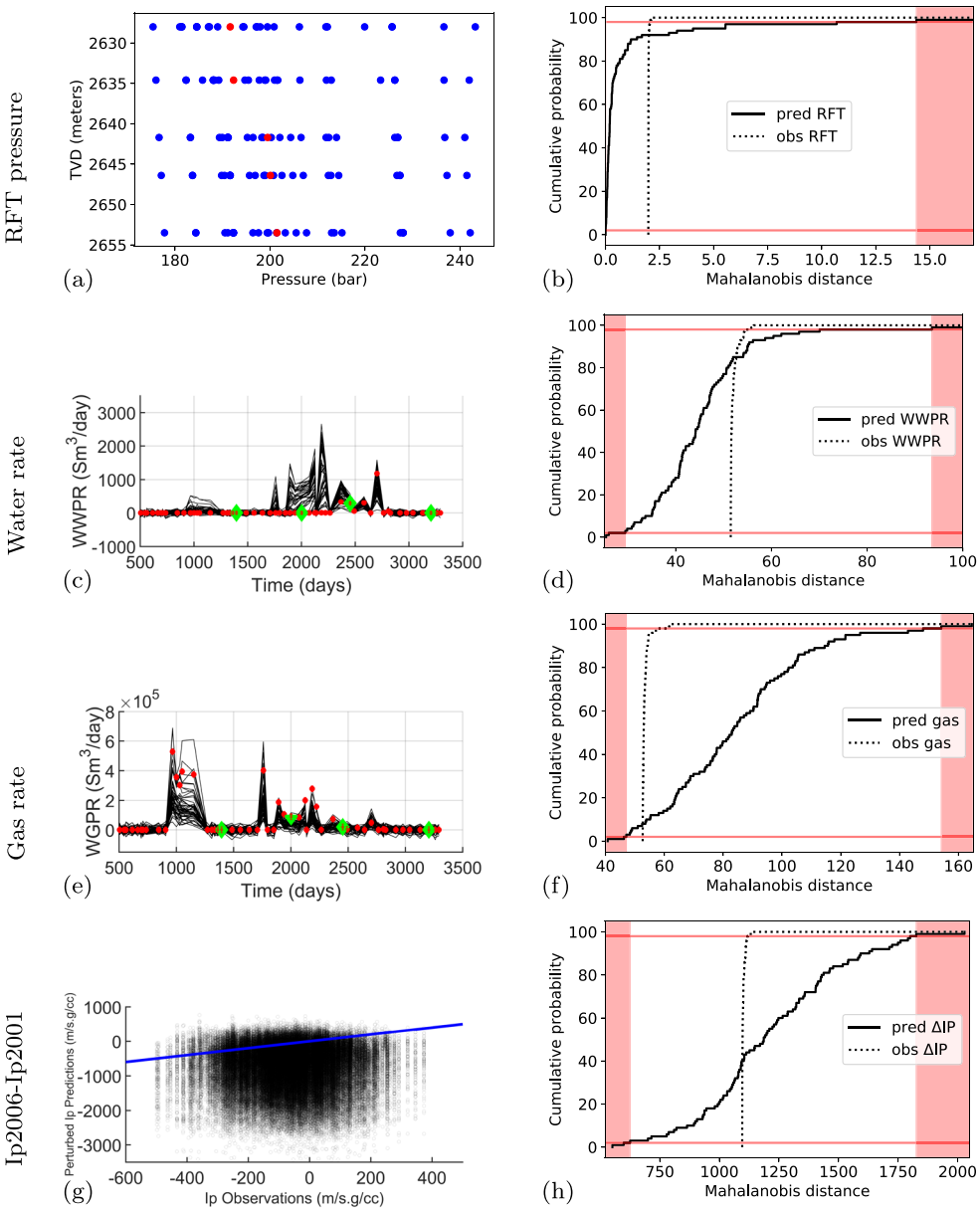

Fig. 7 Prior predictive distribution based on the model after moving completion of well E-4AH. Top row shows 20 realizations of perturbed simulated RFT data (blue) in the Garn formation of well F-4H (July 2001) compared with observations (red). Water and gas production rates in well E-4AH are shown in middle two rows, with the times of the 4D seismic surveys shown by the green diamonds. Bottom row shows change in simulated acoustic impedance in the Garn formation of the G-segment. Right column compares the simulated distribution of scores (solid curve) with the distribution of estimated scores for the observations (dashed curve)

is seen to be inadequate until improvements are made in the connectivity.

4 Conclusions

In this paper, the ability to criticize large models by comparison of large amounts of data to predictions was demonstrated. We showed that simple "coverage" of data by the ensemble of perturbed predictions does not always provide a useful evaluation of the prior ensemble. Poor coverage can be useful for demonstrating inadequacy of the model, but good coverage does not necessarily demonstrate that the initial ensemble is valid.

When coverage is insufficient as a criterion, we propose use of a model diagnostic based on the Mahalanobis distance, which measures the distance of the observations from the ensemble of predictions. Computation of Mahalanobis distance is not straightforward for models with large numbers of data and small numbers of samples as the estimate of the covariance matrix in that case cannot be inverted. We proposed using a shrinkage estimate of the covariance matrix that has full rank, even in high dimensions. We showed that this estimate of the covariance was better at identifying outliers for a toy example with 1000 observations than estimates based on the pseudo-inverse or on a diagonal approximation of the covariance.

We also applied the Mahalanobis distance to the problem of criticism of a model of the G-segment of the Norne field by comparison with production data, RFT data, and inverted seismic data, independently. We showed that the model used by Chen and Oliver [13] was inadequate when compared to RFT data, production data and 4D seismic data, so that model improvement was necessary. After two rounds of model improvement, the observations were consistent with the ensemble of perturbed simulated data. At this point, the model and data are ready for data assimilation and additional model criticism based on the distribution of residuals [2, 33].

We showed that the methodology for model criticism and improvement we proposed in this paper is feasible for realistic problems with large numbers of model parameters, large amounts of data, and correlated observation errors, which is the case in history matching to production and seismic data.

Acknowledgements The authors thank Equinor (operator of the Norne field) and its license partners Eni Norge and Petoro for the release of the Norne data. The authors acknowledge the Center for Integrated Operations at NTNU for cooperation and coordination of the Norne Cases. The view expressed in this paper are the views of the authors and do not necessarily reflect the views of Equinor and the Norne license partners.

We are grateful to Geovariances for providing a license for the use of Isatis for factorial co-kriging, and to Schlumberger for providing Eclipse and Petrel licenses.

Funding information This study is supported by the CIPR/IRIS cooperative research project "4D Seismic History Matching" which is funded by industry partners Eni Norge, Petrobras, and Total, as well as the Research Council of Norway through the Petromaks2 program.

Appendix A: Petro-elastic model

We use the Gassmann model [17, 46] to model the bulk modulus, K_{sat}, and the shear modulus, μ_{sat}, of the saturated rock [31], as follows:

$$K_{\text{sat}} = K_{\text{dry}} + \left(1 - \frac{K_{\text{dry}}}{K_{\text{min}}}\right)^2 \left(\frac{\phi}{K_{\text{fluid}}} + \frac{1-\phi}{K_{\text{min}}} + \frac{K_{\text{dry}}}{{K_{\text{min}}}^2}\right)^{-1} \tag{14}$$

and

$$\mu_{\text{sat}} = \mu_{\text{dry}}, \tag{15}$$

where K_{dry} is the bulk modulus of the dry rock, K_{min} is the bulk modulus of the mixture of minerals in the rock matrix, K_{fluid} is the bulk modulus of the mixture of fluids (water, oil and gas) in the porous medium, ϕ is the porosity of the rock, and μ_{dry} is the shear modulus of the dry rock.

The P-wave velocity of the saturated rock is obtained from the following relationship:

$${V_{\text{sat}}}^p = ((K_{\text{sat}} + (4/3)\mu_{\text{sat}})/\rho_{\text{sat}})^{1/2} \tag{16}$$

The acoustic or P-wave impedance of the saturated rock is then computed from the P-wave velocity and the bulk density of the rock as follows:

$${I_{\text{sat}}}^p = \rho_{\text{sat}} {V_{\text{sat}}}^p. \tag{17}$$

We model the elastic moduli of the minerals in the rock matrix using the Hashin-Shtrikman bounds [19] for a sand-shale mixture, assuming that the rock matrix is composed of sand and shale, and hence, quartz and clay minerals. Our estimates of K_{min} (which is required for computation of K_{sat} (14)) and μ_{min} are obtained by arithmetic averaging of upper and lower Hashin-Shtrikman bounds [31]. The mineral density is the volume-weighted average of the densities of quartz and clay minerals. The relative volumes of each mineral are assumed to be related directly to the net-to-gross ratio of the reservoir model.

Elastic moduli of the dry rock are computed using a slightly modified version [4] of the dependence of the dry-rock moduli on porosity [27, 36], coupled with the dependence on changes in effective stress [30]. This combination is used in the context of 4D seismic data, where two or more seismic times, and hence reservoir conditions, are involved. The dynamic part of the elastic dry moduli (K_{dry} and μ_{dry}), which describes the dependence on stress, are computed as follows:

$$K_{\text{dry}} = K_{\text{dry,stat}} \left(1 + E_K e^{-P_{\text{eff,init}}/P_K}\right) \left(1 + E_K e^{-P_{\text{eff}}/P_K}\right)^{-1} \tag{18}$$

$$\mu_{\text{dry}} = \mu_{\text{dry,stat}} \left(1 + E_\mu e^{-P_{\text{eff,init}}/P_\mu}\right) \left(1 + E_\mu e^{-P_{\text{eff}}/P_\mu}\right)^{-1} \tag{19}$$

In Eqs. 18 and 19, $P_{\text{eff,init}}$ is the effective pressures at initial reservoir conditions, P_{eff} is the effective pressure at a particular reservoir condition (for example, at the time of a certain 4D seismic survey), and E_k, P_k, E_μ, and P_μ are stress sensitivity parameters [30].

The static parts the dry elastic moduli, which describe the dependence of dry-rock elastic properties on porosity are given by the following:

$$K_{\text{dry,stat}} = K_{\text{min}}(1-\phi)/(1+\beta\phi) \tag{20}$$

$$\mu_{\text{dry,stat}} = \mu_{\text{min}}(1-\phi)/(1+\beta\phi) \tag{21}$$

where β is the consolidation factor, which is assumed to be linearly related to the bulk composition [4] as follows:

$$\beta = aV_{\text{sand}} + bV_{\text{shale}} + c\phi \tag{22}$$

where V_{sand} and V_{shale} are the volumes of sand and shale in the whole rock, respectively, ϕ is the porosity, and a, b, and c are consolidation parameters determining the sensitivity of β to V_{sand}, V_{shale}, and ϕ. The elastic moduli of the dry rock (K_{dry} and μ_{dry}) in Eqs. 18 and 19 are required inputs to Gassmann's relations (14 and 15).

The proportions of sand and shale in the bulk rock are assumed to be determined by porosity and net-to-gross ratio as follows:

$$V_{\text{shale}} = (1 - NTG)(1 - \phi) \tag{23}$$

$$V_{\text{sand}} = NTG(1 - \phi) \tag{24}$$

Effective pressure (P_{eff}) at each grid cell of the model is computed from $P_{\text{eff}} = P_{\text{lith}} - P$ where P is fluid pressure and P_{lith} is lithostatic pressure computed from the following formula:[2]

$$P_{\text{lith}}[z] = -49.6 + 0.2027z + 6.127 \times 10^{-6} \times z^2 \quad \text{[bars]} \tag{25}$$

where z is vertical subsea depth in meters.

Elastic moduli of the mixture of fluids at reservoir conditions are modeled using relationships from Batzle and Wang [5]. For the Norne field, the reservoir temperature was assumed to be 98.3 °C, water salinity 15000 ppm, oil gravity 32.7 API, and gas gravity 0.645. The bulk modulus of the mixture of fluids (K_{fluid}) is computed using the Reuss average (Mavko et al., 2009) as follows:

$$K_{\text{fluid}} = (S_w/K_w + S_o/K_o + S_g/K_g)^{-1}. \tag{26}$$

where S_w, S_o, and S_g, are the saturation of water, oil, and gas, respectively, and K_w, K_o, and K_g are the bulk moduli of water, oil, and gas.

The PEM that we selected (Eqs. 14 to 26) requires that some parameters be estimated and calibrated for the Norne field case. We separated the PEM into two vertical regions comprising the Garn formation (layers 1 to 3 of the Norne simulation model) and the formations underlying the Not shale (Ile to Tofte in layers 5 to 22), as the geological information available on the Norne field suggests that the Garn formation and the formations below the Not shale are reasonably different [44].

Values for the consolidation parameters a, b, and c in the original model were obtained from a weighted average at four different wells in the Norne field [8, 9]. Values for elastic moduli of the mineral parameters in the rock matrix were obtained from well log calibration in the Norne field, and the density of sand and shale in the rock matrix were set to 2.689 g/cc and 2.635 g/cc, respectively [8]. Finally, values of parameters for the pressure sensitivity of dry elastic moduli (Eqs. 18 and 19) are obtained from core measurements from the Schiehallion field [30]. All values used in the PEM are shown in Table 3.

[2]Data for P_{lith} provided by Fridtjof Riis of the Norwegian Petroleum Directorate. Compaction is based on [42].

Table 3 Input parameters of the Norne PEM. In the original model, the parameters were set to fixed values

Parameter	Original model	Improved model
Garn formation		
a	5.73	5.73 ± 0.4
b	5.57	5.57 ± 0.5
c	-2.63	-2.63 ± 1.5
Ile, Tofte and Tilje formations		
a	7.93	7.93 ± 1.0
b	9.40	9.40 ± 1.0
c	-2.88	-2.88 ± 0.4
All formations		
$K_{\text{sand,matrix}}$ [GPa]	23	23.0 ± 6.0
$\mu_{\text{sand,matrix}}$ [GPa]	16	16.0 ± 4.5
$K_{\text{shale,matrix}}$ [GPa]	22	22.0 ± 3.0
$\mu_{\text{shale,matrix}}$ [GPa]	12	12.0 ± 2.0
P_k [MPa]	5.62	5.62 ± 1.0
P_μ [MPa]	7.97	7.97 ± 1.0
E_k	1.128	1.128
E_μ	1.083	1.083
I_p [(m/s)(g/cc)]	6500	6500 ± 600

In the improved model, uncertainty was added to some of these parameters

Appendix B: Seismic modeling

We generate realizations of acoustic impedance predictions from the Norne simulation model, using the PEM (Appendix A) and applying a vertical seismic filter to account for differences in resolution between the simulation model and the inverted seismic data.

Due to the gridding in the simulation model, the seismic predictions obtained from application of the PEM to the saturations, pressures, and porosities at the reservoir simulation grid scale may contain high-frequency features that are outside of the observed seismic-frequency spectrum. Although other solutions for dealing with lack of smoothness in synthetic seismic data have been proposed [11, 43], we have applied a vertical seismic filter to the acoustic impedance predictions in the simulation model in order to make them more comparable to actual seismic data [2, 37]. In other words, perturbed impedance predictions and inverted impedance data are compared at the scale of the Norne seismic data.

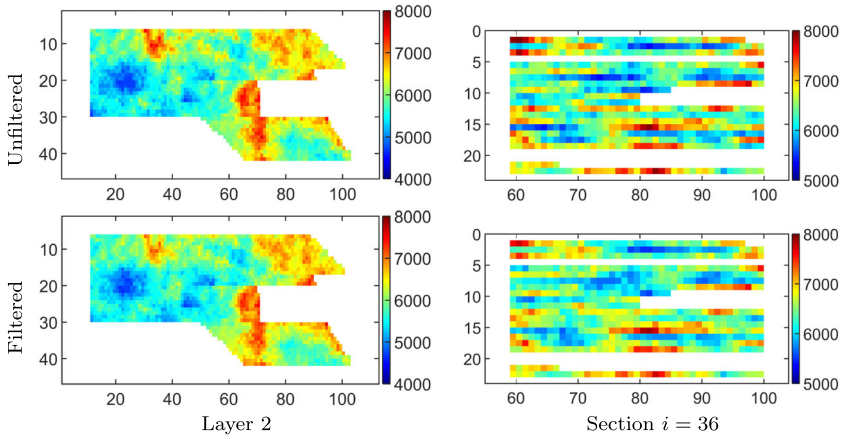

Fig. 8 One realization of acoustic impedance predictions in the G-segment from the PEM (top row) and after filtering (bottom row). Units of impedances: (m/s).(g/cc)

This filter was created based on the frequency spectrum of the actual inverted acoustic impedance data of the four Norne seismic surveys (Ip2001, Ip2003, Ip2004, and Ip2006), all of which have very similar frequency content. We modeled the observed spectra by a low-pass Ormsby filter [34], which is a trapezoidal filter applied in the frequency domain that removes all the frequencies above some user-defined cut frequencies. To approximately match the frequency spectrum of real impedance data, we set the cut frequencies of the Ormsby filter at 0-0-100-120 Hz. Prior to filtering, we populated the inactive cells of the Norne simulation model with a Not shale acoustic impedance value of 6500 (m/s).(g/cc). Figure 8 shows one realization of the impedance predictions at the time of the 2001 seismic survey (baseline) before and after filtering.

Appendix C: Simulation model

Table 4 summarizes the parameters of the simulation model that were modified in the improved and move-completion models (Section 3.2.5). In this table, MULTREGT defines the transmissibility multiplier between flux regions, MULTZ describes the vertical transmissivity multiplier between two layers, and KRW corresponds to the endpoint relative permeability of water.

References

1. Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the surprising behavior of distance metrics in high dimensional space. In: Van den Bussche, J., Vianu, V. (eds.) Database Theory — ICDT 2001, pp. 420–434. Springer, Berlin (2001)
2. Alfonzo, M., Oliver, D.S.: Seismic data assimilation with an imperfect model. Computational Geosciences online 10 July. https://doi.org/10.1007/s10596-019-09849-0 (2019)
3. Alfonzo, M., Oliver, D.S., MacBeth, C.: Analysis and calibration of 4D seismic data prior to 4D seismic inversion and history matching – Norne Field case. In: 79th EAGE Conference and Exhibition, Amsterdam (2017)
4. Amini, H.: A pragmatic approach to simulator-to-seismic modelling for 4D seismic interpretation. Ph.D. thesis, Heriot-Watt University (2014)
5. Batzle, M., Wang, Z.: Seismic properties of pore fluids. Geophysics **57**(11), 1396–1408 (1992)
6. Bayarri, M.J., Berger, J.O., Paulo, R., Sacks, J., Cafeo, J.A., Cavendish, J., Lin, C.H., Tu, J.: A framework for validation of computer models. Technometrics **49**(2), 138–154 (2007)
7. Box, G.E.P.: Sampling and Bayes' inference in scientific modelling and robustness. J. R. Stat. Soc. Series A (General) **143**(4), 383–430 (1980)
8. Briceño, A., MacBeth, C., Mangriotis, M.D.: Towards an effective petroelastic model for simulator to seismic studies. In: 78th EAGE Conference and Exhibition 2016 (2016)
9. Briceño Yañez, A.E.: Calibration and use of the petroelastic model for 4D seismic interpretation. Ph.D. thesis, Heriot-Watt University (2017)
10. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv **41**(3), 15:1–15:58 (2009)

Table 4 Parameters of the Norne simulation model modified in the improved and move-completion models

Parameter		Original model	Improved model
	Garn	-2.3	-2.3 ± 0.5
	Ile	-2.0	-2.0 ± 0.5
MULTREGT	Tofte	-2.0	-2.0 ± 0.5
($\log_{10}$)	Tilje 4/3	-2.0	-2.0 ± 0.5
	Tilje 1/2	-1.0	-1.0 ± 0.5
MULTZ ($\log_{10}$)	Garn (layers 1 and 2)	-2.0 ± 0.5	-3.0 ± 0.5
KRW parameter	Garn	0.8–1.5	0.5–1.2

11. Chen, J., Schuster, G.T.: Resolution limits of migrated images. Geophysics **64**(4), 1046–1053 (1999)
12. Chen, Y., Oliver, D.S.: Ensemble-based closed-loop optimization applied to Brugge field. SPE Reserv. Eval. Eng. **13**(1), 56–71 (2010)
13. Chen, Y., Oliver, D.S.: History matching of the Norne full-field model with an iterative ensemble smoother. SPE Reserv. Eval. Eng. **17**(2), 244–256 (2014). https://doi.org/10.2118/164902-PA
14. Emerick, A.A., Reynolds, A.C.: History matching a field case using the ensemble Kalman filter with covariance localization. SPE Reserv. Eval. Eng. **14**(4), 443–452 (2011)
15. Ferreira, C.J., Davolio, A., Schiozer, D.J.: Use of a probabilistic and multi-objective history matching for uncertainty reduction for the Norne benchmark case. In: SPE Europec featured at 79th EAGE Conference and Exhibition. Society of Petroleum Engineers (2017)
16. Filzmoser, P., Garrett, R.G., Reimann, C.: Multivariate outlier detection in exploration geochemistry. Comput. Geosci. **31**(5), 579–587 (2005)
17. Gassmann, F.: Elastic waves through a packing of spheres. Geophysics **16**, 673–685 (1951)
18. Härdle, W.K., Simar, L.: Applied Multivariate Statistical Analysis, 3rd edn. Springer, Berlin (2007)
19. Hashin, Z., Shtrikman, S.: A variational approach to the theory of the elastic behaviour of multiphase materials. J. Mech. Phys. Solids **11**(2), 127–140 (1963)
20. He, J., Tanaka, S., Wen, X.H., Kamath, J.: Rapid S-curve update using ensemble variance analysis with model validation. In: SPE Western Regional Meeting, Bakersfield, California, 23 April. Soc.of Petrol.Engineers (2017)
21. Hermans, T., Nguyen, F., Caers, J.: Uncertainty in training image-based inversion of hydraulic head data constrained to ERT data: Workflow and case study. Water Resour. Res. **51**(7), 5332–5352 (2015)
22. Hoffmann, H.: Kernel PCA for novelty detection. Pattern Recogn. **40**(3), 863–874 (2007)
23. Huber, P.J.: Robust statistics. In: Lovric, M. (ed.) International Encyclopedia of Statistical Science, pp. 1248–1251. Springer, Berlin (2011)
24. Iglewicz, B., Hoaglin, D.C.: How to Detect and Handle Outliers, vol. 16. ASQ Press (1993)
25. Kriegel, H.P., Schubert, M., Zimek, A.: Angle-based outlier detection in high-dimensional data. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 444–452. ACM (2008)
26. Ledoit, O., Wolf, M.: A well-conditioned estimator for large-dimensional covariance matrices. J. Multivar. Anal. **88**(2), 365–411 (2004)
27. Lee, M.W.: Proposed moduli of dry rock and their application to predicting elastic velocities of sandstones. Tech. Rep. Scientific Investigations Report 2005–5119 US Geological Survey (2005)
28. Leung, P.L., Chan, W.Y.: Estimation of the scale matrix and its eigenvalues in the Wishart and the multivariate F distributions. Ann. Inst. Stat. Math. **50**(3), 523–530 (1998)
29. Lorentzen, R., Bhakta, T., Grana, D., Luo, X., Valestrand, R., Naevdal, G.: History matching of real production and seismic data in the Norne field. In: ECMOR XVI (2018)
30. MacBeth, C.: A classification for the pressure-sensitivity properties of a sandstone rock frame. Geophysics **69**(2), 497–510 (2004)
31. Mavko, G., Mukerji, T., Dvorkin, J.: The Rock Physics Handbook: Tools for Seismic Analysis of Porous Media. Cambridge University Press (2009)
32. Morell, E.: History Matching of the Norne Field. Master's thesis, NTNU. Department of Petroleum Engineering and Applied Geophysics, Trondheim (2010)
33. Oliver, D.S., Alfonzo, M.: Calibration of imperfect models to biased observations. Comput. Geosci. **22**(1), 145–161 (2018). https://doi.org/10.1007/s10596-017-9678-4
34. Ormsby, J.F.A.: Design of numerical filters with applications to missile data processing. J. ACM **8**(3), 440–466 (1961)
35. Park, H., Scheidt, C., Fenwick, D., Boucher, A., Caers, J.: History matching and uncertainty quantification of facies models with multiple geological interpretations. Comput. Geosci. **17**(4), 609–621 (2013). https://doi.org/10.1007/s10596-013-9343-5
36. Pride, S.R.: Relationships between seismic and hydrological properties. In: Rubin, Y., Hubbard, S.S. (eds.) Hydrogeophysics, pp. 253–290. Springer (2005)
37. Roggero, F., Lerat, O., Ding, D.Y., Berthet, P., Bordenave, C., Lefeuvre, F., Perfetti, P.: History matching of production and 4D seismic data: application to the Girassol Field, Offshore Angola. Oil Gas Sci. Technol. — Rev. IFP Energies nouvelles **67**(2), 237–262 (2012)
38. Rwechungura, R.W., Dadashpour, M., Kleppe, J.: Application of particle swarm optimization for parameter estimation integrating production and time lapse seismic data (SPE-146199). In: SPE offshore Europe Oil and Gas Conference and Exhibition, 6-8 September 2011. Aberdeen (2011)
39. Schäfer, J., Strimmer, K.: A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat. Appl. Genet. Mol. Biol. **4**(1, 32), 1–30 (2005)
40. Scheidt, C., Jeong, C., Mukerji, T., Caers, J.: Probabilistic falsification of prior geologic uncertainty with seismic amplitude data: application to a turbidite reservoir case. Geophysics **80**, M89–M100 (2015). https://doi.org/10.1190/geo2015-0084.1
41. Schulze-Riegert, R., Nwakile, M., Skripkin, S., Willen, Y.: Scalability and performance efficiency of history matching workflows using MCMC and adjoint techniques applied to the Norne North Sea reservoir case study. In: 78th EAGE Conference and Exhibition (2016)
42. Sclater, J.G., Christie, P.A.F.: Continental stretching: an explanation of the post-Mid-Cretaceous subsidence of the central North Sea Basin. J. Geophys. Res.: Solid Earth **85**(B7), 3711–3739 (1980)
43. Toxopeus, G., Thorbecke, J., Wapenaar, K., Petersen, S., Slob, E., Fokkerna, J.: Simulating migrated and inverted seismic data by filtering a geologic model. Geophysics **73**(2), T1–T10 (2008)
44. Verlo, S.B., Hetland, M.: Development of a field case with real production and 4D data from the Norne Field as a benchmark case for future reservoir simulation model testing. Master's thesis, NTNU. Trondheim (2008)
45. Wang, F., Sun, J.: Survey on distance metric learning and dimensionality reduction in data mining. Data Min. Knowl. Disc. **29**(2), 534–564 (2015)
46. Wang, Z.Z.: Y2K tutorial: Fundamentals of seismic rock physics. Geophysics **66**(2), 398–412 (2001)
47. Zhang, Y., Leeuwenburgh, O., Carpentier, S., Steeghs, P.: 4D seismic history matching of the Norne field model using ensemble-based methods with distance parameterization. In: IOR 2017–19th European Symposium on Improved Oil Recovery (2017)