

# Ensemble Kalman Filtering with Shrinkage Regression Techniques

Jon Sætrum & Henning Omre, Norwegian University of Science and Technology;

## Summary

The classical Ensemble Kalman Filter (EnKF) is known to underestimate the prediction uncertainty resulting from model overfitting and estimation error. This can potentially lead to low forecast precision and an ensemble collapsing into a single realisation. In this paper we present alternative EnKF updating schemes based on shrinkage methods known from multivariate linear regression. These methods reduce the effects caused by collinear ensemble members, and have the same computational properties as the fastest EnKF algorithms previously suggested. In addition, the importance of model selection and validation for prediction purposes is investigated, and a model selection scheme based on Cross-Validation is introduced. The classical EnKF scheme is compared with the suggested procedures on two toy examples and one synthetic reservoir case study. Significant improvements are seen, both in terms of forecast precision and prediction uncertainty estimates.

## Introduction

The Ensemble Kalman Filter (EnKF) is a Bayesian data assimilation method that in recent years has become popular when considering data assimilation for nonlinear spatio-temporal models (Evensen, 2007; Aanonsen et al., 2009). The EnKF is based on the classical Kalman Filter (KF) (Kalman, 1960). Assuming a Gaussian initial prior, with a linear and Gaussian forward and likelihood model, termed the Gauss-linear model, the KF provides an analytical solution for the posterior probability distribution.

The main assumption of the EnKF is that the unconditional distribution at each timestep approximately follows a Gaussian distribution with unknown mean and covariance. In the EnKF the idea is therefore to use an ensemble of realisations to estimate these two unknown statistics. Under the Gauss-linear model, the EnKF is then consistent with the KF as the ensemble size approaches infinity (Mardia et al., 1979).

From multivariate statistics (Anderson, 2003), we know that the classical EnKF updating scheme (Evensen, 1994) can be formulated as a multivariate regression problem. The least squares solution for the matrix of regression coefficients is then given as the estimated Kalman gain, expressed in terms of the estimated unknown covariance matrices. However, it is not guaranteed that this estimator is good for prediction purposes. This is especially true for regression models where the predictor variables are collinear, or when the regression model is computed based on dependent realisations (Farrer and Glauber, 1967). As shown in Myrseth et al. (2009), the traditional EnKF updating scheme under the Gaussian approximation violates the assumption of independent realisations, because they are coupled through the estimated Kalman gain matrix. Hence, we should expect that improved estimators of the regression coefficient matrix can be found in these cases.

In multivariate linear regression, there exist several different approaches to avoid overfitting a regression model with collinear predictor variables, known as shrinkage methods. Some introduce regularisation terms, while others aim to estimate the regression coefficients based on data in a reduced order space. In this paper we will consider the following three methods: Ridge Regression (RR) (Hoerl and Kennard, 1970), Principal Component Regression (PCR) (Hotelling, 1933; Jolliffe, 2002) and Partial Least Squares Regression (PLSR) (Wold, 1975; Rosipal and Krämer, 2006).

Shrinkage estimators in multivariate regression require that a prior hyperparameter is selected. For methods based on PCR, this means to select the dimension of a reduced order subspace. The most common method used to accomplish this task is to look at the total variance explained by the selected Principal Components (PC) (Jolliffe, 2002). However, such an approach does not take into account the predictive capabilities of the estimated regression model, and can lead to situations where we either overfit, or underfit the model to the data (Seber and Lee, 2003; Cook, 2007). The method usually applied in shrinkage regression, is therefore to base the selection of the hyperparameter on Cross-Validation (CV) (Efron, 2004) in order to avoid this potential problem.

Shrinkage estimators which are based on dimension reduction techniques are for computational reasons the natural approach to consider when assimilating high dimensional data, such as time-lapse seismic. Shrinkage type estimators for the Kalman gain have therefore already been used in an EnKF setting (Skjervheim et al., 2005). It has also been noted (Evensen, 2007, Chapter 14) that the results obtained when using shrinkage estimators are indeed dependent on the selection of the prior hyperparameter used. However, the importance of the number of components retained in the reduced order space to

avoid overfitting, has seemingly been overlooked in most of the EnKF literature. Moreover, because PCR is based on an unsupervised dimension reduction technique, all components that are important for predictive purposes can potentially be discarded, unless model validation is performed (Jolliffe, 1982, 2002; Hadi and Ling, 1998; Cook, 2007).

In this paper we have formulated the EnKF using shrinkage-based regression techniques. The suggested procedures have the same computational complexity and memory requirements as the fastest implementations of the EnKF (Evensen, 2003). We have further demonstrated the approach on the following case studies: A Gauss-linear model, a non-linear forward model with a linear Gaussian likelihood, and a synthetic reservoir example.

### Notation and Model Formulation

Throughout this paper the notation  $\mathbf{x} \in \mathbb{R}^{n_x \times 1}$  will be used to denote that  $\mathbf{x}$  is an  $n_x$ -dimensional column vector in the real space and  $\mathbf{x}^T$  will denote its transpose. Similarly, we will write  $\mathbf{A} \in \mathbb{R}^{a \times b}$  to denote that  $\mathbf{A}$  is a matrix in the real space containing  $a$  rows and  $b$  columns. Note that the same notation will be used for both scalars and random variables.

Probability density functions (pdfs) will be denoted by  $f(\mathbf{x})$ , and the notation  $\mathbf{x} \sim f(\mathbf{x})$ , implies that the random vector  $\mathbf{x}$  follows the pdf  $f(\mathbf{x})$ . As a special case we will let a random vector  $\mathbf{x}$ , following the Gaussian distribution with mean vector  $\boldsymbol{\mu}_x$  and covariance matrix  $\boldsymbol{\Sigma}_x$ , be denoted by  $\mathbf{x} \sim \text{Gauss}_{n_x}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ . Conditional pdfs of  $\mathbf{x}$  given  $\mathbf{y}$  will further be denoted by  $f(\mathbf{x}|\mathbf{y})$ .

Consider the stochastic Directed Acyclic Graph (DAG) (Ripley, 1996) outlined in **Fig. 1**. Here,  $\mathbf{x}_{t_k} \in \mathbb{R}^{n_x \times 1}$  denotes the state of the unknown random vector of interest at timestep  $k$  and time  $t_k$ , and similarly  $\mathbf{d}_{t_k} \in \mathbb{R}^{n_d \times 1}$  denotes the vector of observed data. For notational convenience we will from now on drop the subscript  $t_k$ , and simply write  $\mathbf{x}_k, \mathbf{d}_k$ . Also note that we will for simplicity refer to  $\mathbf{x}$  and  $\mathbf{d}$  as the state and observation vector respectively.

Through the Markov property of a stochastic DAG, we have conditional independence between  $\mathbf{x}_{k+1}$  and  $\mathbf{x}_l, l = 0, \dots, k-1$  given  $\mathbf{x}_k$ , meaning:

$$f(\mathbf{x}_{k+1}|\mathbf{x}_k, \mathbf{x}_{k-1}, \dots, \mathbf{x}_0) = f(\mathbf{x}_{k+1}|\mathbf{x}_k), k = 0, \dots, K.$$

In general assume that

$$[\mathbf{x}_{k+1}|\mathbf{x}_k] = \boldsymbol{\omega}(\mathbf{x}_k, \boldsymbol{\epsilon}_{\mathbf{x}_k}), k = 0, \dots, K, \dots \dots \dots (1)$$

where  $\boldsymbol{\omega} : (\mathbb{R}^{n_x} \times \mathbb{R}^{n_x}) \rightarrow \mathbb{R}^{n_x}$  is a known, possibly non-linear forward function. Here  $\boldsymbol{\epsilon}_{\mathbf{x}_k}$  represents random model errors or numerical errors in the forward model, assumed to follow a known probability distribution. Further assume that observed data  $\mathbf{d}_k$  is connected to  $\mathbf{x}_k$  by:

$$[\mathbf{d}_k|\mathbf{x}_k] = \boldsymbol{\zeta}(\mathbf{x}_k, \boldsymbol{\epsilon}_{\mathbf{d}_k}), k = 0, \dots, K, \dots \dots \dots (2)$$

where  $\boldsymbol{\zeta} : (\mathbb{R}^{n_x} \times \mathbb{R}^{n_d}) \rightarrow \mathbb{R}^{n_d}$ , is a known, possibly non-linear, likelihood function and  $\boldsymbol{\epsilon}_{\mathbf{d}_k}$  represents the observation error, again assumed to follow a known pdf.

The aim in this model setting is to solve the spatio-temporal forecast problem of finding the unknown state vectors:

$$\mathbf{x}_k^c = [\mathbf{x}_k|\mathbf{d}_0, \dots, \mathbf{d}_k] \quad \text{and} \quad \mathbf{x}_{k+1}^u = [\mathbf{x}_{k+1}|\mathbf{d}_0, \dots, \mathbf{d}_k],$$

for  $k = 1, \dots, K$ . Bayesian inversion provides a sequential solution to this problem. By defining a prior probability model for the state vector at the initial timestep,  $f(\mathbf{x}_0)$ , the state of the unknown vectors  $\mathbf{x}_k^c$  and  $\mathbf{x}_{k+1}^u$ , can be assessed by sampling from the respective posterior distributions. Through the use of Bayes rule and the Markov properties of a DAG, these pdfs are given as:

$$f(\mathbf{x}_k^c) \propto f(\mathbf{x}_k^u) f(\mathbf{d}_k | \mathbf{x}_k)$$

$$f(\mathbf{x}_{k+1}^u) = \int f(\mathbf{x}_{k+1} | \mathbf{x}_k) f(\mathbf{x}_k^c) d\mathbf{x}_k. \dots \dots \dots (3)$$

Note that the conditional pdfs  $f(\mathbf{x}_{k+1} | \mathbf{x}_k)$  and  $f(\mathbf{d}_k | \mathbf{x}_k)$  corresponds to the forward and likelihood models  $\omega(\mathbf{x}_k, \epsilon_{\mathbf{x}_k})$  and  $\zeta(\mathbf{x}_k, \epsilon_{\mathbf{d}_k})$  defined in Eqs. (1) and (2) respectively.

Generally we only know the conditional distributions defined in Eq. (3) up to an unknown normalising constant. Computationally demanding techniques such as Markov chain Monte Carlo (MCMC) or Rejection Sampling can therefore be used to assess the correct posterior distribution (Doucet et al., 2000). For applications such as petroleum reservoir evaluation, using these techniques are, however, computationally prohibitive as even a single evaluation of  $f(\mathbf{x}_{k+1} | \mathbf{x}_k) = \omega(\mathbf{x}_k, \epsilon_{\mathbf{x}_k})$ , known as fluid flow simulation, can take several hours, or even days.

An approximative approach can be defined by assuming that  $\mathbf{x}_k^u$  and  $\mathbf{d}_k$  are from a family of probability density functions that ensures analytical tractability of  $f(\mathbf{x}_k^c)$ . As an example, assume that  $\mathbf{x}_k^u$  and  $\mathbf{d}_k$  are jointly Gaussian. Then the posterior distribution at timestep  $k$  will also be Gaussian. These model assumptions are equivalent to those made in the classical EnKF, which we will consider next.

### Classical Ensemble Kalman Filter

In general assume that the output of the forward- and likelihood model are distributed as:

$$\begin{bmatrix} \mathbf{x}_k^u \\ \mathbf{d}_k \end{bmatrix} \sim \text{Gauss}_{n_y} \left( \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{x}_k} \\ \boldsymbol{\mu}_{\mathbf{d}_k} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{x}_k} & \boldsymbol{\Sigma}_{\mathbf{x}, \mathbf{d}_k} \\ \boldsymbol{\Sigma}_{\mathbf{d}, \mathbf{x}_k} & \boldsymbol{\Sigma}_{\mathbf{d}_k} \end{bmatrix} \right), \dots \dots \dots (4)$$

where  $n_y = n_x + n_d$ . For notational convenience, we will from now on omit the subscript  $k$ , because the main focus is a single timestep.

Under the Gaussian assumption above, the posterior pdf  $f(\mathbf{x}_k^c)$  will be Gaussian (Mardia et al., 1979) with analytically obtainable mean:

$$\boldsymbol{\mu}_{\mathbf{x} | \mathbf{d}} = \boldsymbol{\mu}_{\mathbf{x}} + \boldsymbol{\Sigma}_{\mathbf{x}, \mathbf{d}} \boldsymbol{\Sigma}_{\mathbf{d}}^{-1} (\mathbf{d} - \boldsymbol{\mu}_{\mathbf{d}}), \dots \dots \dots (5)$$

and covariance matrix:

$$\boldsymbol{\Sigma}_{\mathbf{x} | \mathbf{d}} = \boldsymbol{\Sigma}_{\mathbf{x}} - \boldsymbol{\Sigma}_{\mathbf{x}, \mathbf{d}} \boldsymbol{\Sigma}_{\mathbf{d}}^{-1} \boldsymbol{\Sigma}_{\mathbf{x}, \mathbf{d}}^T. \dots \dots \dots (6)$$

Let  $\mathbf{y}^{(i)} = [\mathbf{x}^{u(i)T}, \mathbf{d}^{(i)T}]^T \in \mathbb{R}^{n_y \times 1}$  be a realisation from the Gaussian distribution defined in Eq. (4). Then,

$$\mathbf{x}^{c(i)} = \mathbf{x}^{u(i)} + \boldsymbol{\Sigma}_{\mathbf{x}, \mathbf{d}} \boldsymbol{\Sigma}_{\mathbf{d}}^{-1} (\mathbf{d} - \mathbf{d}^{(i)}), \dots \dots \dots (7)$$

is a realisation from the Gaussian posterior distribution with mean and covariance given in Eqs. (5) and (6). This can easily be shown by recognising that the Gaussian distribution is closed under linear operations and by computing the mean and

covariance of  $\mathbf{x}^{c(i)}$ , because the Gaussian distribution is completely specified through the first and second moments (Mardia et al., 1979).

Eq. (7) involves two model parameters, namely  $\Sigma_{\mathbf{x},d}$  and  $\Sigma_d$  forming the Kalman gain matrix,

$$\mathbf{K} = \Sigma_{\mathbf{x},d} \Sigma_d^{-1} \in \mathbb{R}^{n_x \times n_d}.$$

In a general setting such as the one considered here, these model parameters are unknown. The EnKF solution to this problem is therefore to use an ensemble of realisations to obtain empirical estimates of the unknown covariance matrices. Hence, let  $\mathbf{X}$  and  $\mathbf{D}$  be the centred state- and observation vector ensembles respectively. That is,  $\mathbf{X} = [\mathbf{x}^{(1)} - \hat{\boldsymbol{\mu}}_{\mathbf{x}}, \dots, \mathbf{x}^{(n_e)} - \hat{\boldsymbol{\mu}}_{\mathbf{x}}] \in \mathbb{R}^{n_x \times n_e}$ , where  $\hat{\boldsymbol{\mu}}_{\mathbf{x}}$  denotes the classical estimator of the mean value with a corresponding expression for  $\mathbf{D} \in \mathbb{R}^{n_d \times n_e}$ . Consistent estimators for the unknown covariance matrices are then given as

$$\hat{\Sigma}_d = \frac{1}{n_e} \mathbf{D} \mathbf{D}^T$$

and

$$\hat{\Sigma}_{\mathbf{x},d} = \frac{1}{n_e} \mathbf{X} \mathbf{D}^T.$$

Replacing the unknown covariance matrices in Eq. (7) with consistent empirical estimates, then ensures that  $\mathbf{x}^{c(i)}$  is a sample from the Gaussian posterior distribution above as  $n_e \rightarrow \infty$  (Mardia et al., 1979). Note that throughout this paper we will for notational convenience let all ensemble matrices be centred, unless otherwise stated.

Central in the EnKF updating scheme is the estimated Kalman gain matrix:

$$\hat{\mathbf{K}} = \hat{\Sigma}_{\mathbf{x},d} \hat{\Sigma}_d^{-1}. \dots\dots\dots (8)$$

From multivariate statistical theory (Mardia et al., 1979), we know that this is equivalent to the least squares estimate of the matrix of regression coefficients in a multivariate linear regression setting (Seber and Lee, 2003):

$$\hat{\mathbf{K}} = \arg \min_{\mathbf{K}} \text{tr} \{ (\mathbf{X} - \mathbf{K} \mathbf{D}) (\mathbf{X} - \mathbf{K} \mathbf{D})^T \},$$

where  $\text{tr}(\bullet)$  denotes the trace operator. The analytical rank of the Kalman gain is known to be equal to  $\min\{n_x, n_d, n_e - 1\}$ . Hence,  $\hat{\mathbf{K}}$  will be rank deficient if the number of ensemble members is smaller or equal to  $\min\{n_x, n_d\}$ . This will have direct consequences for the computation of  $\hat{\Sigma}_d^{-1}$ , because this matrix will be positive semi-definite.

Although,  $\hat{\Sigma}_d$  can be ensured to have full rank either through regularisation or Monte Carlo simulation (Myrseth et al., 2009),  $\hat{\Sigma}_{\mathbf{x},d}$  may still suffer from rank deficiency. Moreover, both matrices are likely to suffer from estimation uncertainty (Houtekamer and Mitchell, 1998; van Leeuwen, 1999; Houtekamer and Mitchell, 1999; Furrer and Bengtsson, 2007; Sacher and Bartello, 2008) resulting from a limited number of ensemble members. Several different approaches have been suggested in the EnKF literature in order to handle these problems (see Myrseth and Omre (2009) or Aanonsen et al. (2009) and references therein). Most of these methods, however, focus on improving the estimates of the unknown covariance matrices,

and not the Kalman gain itself. One explanation for this might be that in situations with a linear likelihood model, with additive Gaussian noise:  $\mathbf{d} = \mathbf{H}\mathbf{x} + \epsilon_d$ , the Kalman gain matrix can be written as:

$$\mathbf{K} = \Sigma_x \mathbf{H}^T (\mathbf{H} \Sigma_x \mathbf{H}^T + \Sigma_{\epsilon_d})^{-1}, \dots\dots\dots (9)$$

where  $\Sigma_{\epsilon_d} \in \mathbb{R}^{n_d \times n_d}$  is the covariance matrix of the observations errors, assumed to be known. Hence, the covariance matrix of the state vector is the only unknown parameter that requires estimation. Note that this corresponds to the classical estimator used for linear Gaussian likelihood models in an EnKF setting (Evensen, 1994). Further note that for non-linear likelihood models with an additive noise term, the state vector can always be augmented by  $\mathbf{y} = [\mathbf{x}^T, \mathbf{d}^T]^T$ , so that the EnKF updating scheme can be written in a similar manner as described in Eq. (9). Here  $\mathbf{d}$  corresponds to the deterministic part of the likelihood model with additive noise.

In multivariate linear regression there exists several different techniques that improve the estimated matrix of regression coefficients in the presence of collinear data. These are referred to as shrinkage regression methods (Hastie et al., 2009), and we will use these approaches in an EnKF setting to obtain alternative estimates of  $\mathbf{K}$ .

**Shrinkage Regression Methods**

Consider the linear regression problem:

$$[\mathbf{x}|\mathbf{d}] = \mathbf{K}\mathbf{d} + \epsilon_{\mathbf{x}|\mathbf{d}},$$

where  $\epsilon_{\mathbf{x}|\mathbf{d}}$  corresponds to the model error in the regression model. Here, the goal is to estimate the unknown matrix of multiple linear regression coefficients  $\mathbf{K}$ , based on a set of observed data  $\mathbf{X}$  and  $\mathbf{D}$ . An optimal estimate should then be selected such that a regression model with good prediction capabilities is provided.

Consider the well known problem of estimating a function based on the observed data shown in **Fig. 2a**. Fitting the model perfectly to the available data as shown in **Fig. 2b**, increases the model complexity and lowers the prediction bias with respect to the training data. However, when the model is used in a predictive setting when new data is available, this can lead to poor results with a high variability. On the other hand, underfitting the model by selecting a too simplistic model, as shown in **Fig. 2c**, will lower the variability, but increase the bias. Hence, the optimal model shown in **Fig. 2d**, should be selected such that both bias and variability are kept on a tolerable level as illustrated in **Fig. 3**. As we can see from this figure, the prediction error for a test data set will tend to increase in situations where the model is overfitted to the training data.

Shrinkage regression methods focus on the model fitting problem by sacrificing the unbiasedness of the classical least squares estimator. In this paper we will consider the following three methods: Ridge Regression (RR), which is a special case of Tikhonov Regularisation (Tikhonov and Arsenin, 1977), Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR).

**Ridge Regression.** RR can be viewed as a regularisation method, where the RR estimate of the Kalman gain,  $\mathbf{K}_{RR}$ , is

selected by minimising the mean squared error, with additional constraints which gives:

$$\hat{\mathbf{K}}_{\text{RR}} = \arg \min_{\mathbf{K}} \{ \text{tr}\{(\mathbf{X} - \mathbf{K}\mathbf{D})(\mathbf{X} - \mathbf{K}\mathbf{D})^T\} + \text{tr}\{\mathbf{K}\xi\mathbf{K}^T\} \}$$

The solution to this problem is given as (Seber and Lee, 2003):

$$\hat{\mathbf{K}}_{\text{RR}} = \mathbf{X}\mathbf{D}^T (\mathbf{D}\mathbf{D}^T + \xi\mathbf{I})^{-1}, \dots\dots\dots (10)$$

where  $\mathbf{I}$  is the identity matrix of proper dimensions and  $\xi$  is a regularisation parameter that has to be selected.

**Principal Component Regression.** PCR is based on principal component analysis, which aims at explaining the structure of the data ensemble through a small number of vectors, termed Principal Components (PC):

$$\mathbf{z}_1 = (\mathbf{u}_1^T \mathbf{D})^T, \dots, \mathbf{z}_p = (\mathbf{u}_p^T \mathbf{D})^T \in \mathbb{R}^{n_e \times 1}.$$

The sample PC are selected based on the following criteria:

$$\mathbf{z}_i = (\mathbf{u}_i \mathbf{D})^T - \begin{cases} \max_{\mathbf{u}_i} \{ \mathbf{u}_i^T \hat{\Sigma}_d \mathbf{u}_i \} \\ \|\mathbf{u}_i\|_2 = 1 \\ \mathbf{z}_i^T \mathbf{z}_j = 0, \text{ for all } j < i, i = 1, \dots, p, \end{cases}$$

where  $\|\bullet\|_2$  denotes the Euclidean norm. It can be shown (Anderson, 2003) that the  $i$ th sample PC direction  $\mathbf{u}_i$  is given as the  $i$ th eigenvector of the covariance matrix  $\hat{\Sigma}_d$ . Further, it can be shown that the variance explained by the  $i$ th PC is given by the  $i$ th eigenvalue  $\lambda_i$  of  $\hat{\Sigma}_d$ , where  $\lambda_1 > \lambda_2 > \dots > \lambda_p$ . Under the assumption that the  $p$  sample PC,  $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_p]^T \in \mathbb{R}^{p \times n_e}$ , sufficiently represents  $\mathbf{D}$ , the matrix of regression coefficients can then be estimated based on  $\mathbf{Z}$ , which gives:

$$\hat{\mathbf{K}}_{\text{PCR}} = \mathbf{X}\mathbf{Z}^T (\mathbf{Z}\mathbf{Z}^T)^{-1}.$$

As shown in Appendix A, PCR can be efficiently implemented using Singular Value Decomposition (SVD).

Because the variance of each sample PC is given as the corresponding eigenvalue of the sample covariance matrix, the criterion often used for selecting  $p$  in a PCR setting is to look at the total variance explained by the first  $p$  components given as:

$$\sum_{i=1}^p \lambda_i / \sum_{i=1}^{n_e-1} \lambda_i.$$

However, as noted above, this criterion does not take into account the predictive capabilities of the various PC, and we can end up either overfitting or underfitting the regression model.

**Partial Least Squares Regression.** Consider the DAG shown in **Fig. 4**, which gives a graphical representation of the PLSR model assumptions. Similar to PCR, PLSR aims to represent  $\mathbf{D}$  in a reduced order space, before fitting the regression model. The underlying assumption is that  $\mathbf{T} = [\mathbf{t}_1, \dots, \mathbf{t}_p] \in \mathbb{R}^{n_e \times p}$  and  $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_p] \in \mathbb{R}^{n_e \times p}$  represent the information in

$\mathbf{X}$  and  $\mathbf{D}$  respectively. It is further assumed that  $(\mathbf{w}_1, \mathbf{t}_1)$  captures more information than  $(\mathbf{w}_2, \mathbf{t}_2)$ , and so forth, and that the pairs  $(\mathbf{w}_i, \mathbf{t}_i)$ ,  $i = 1, \dots, p$  also explain the correlation between  $\mathbf{X}$  and  $\mathbf{D}$  (Zeng et al., 2007). Hence, while PCR only does a dimension reduction on  $\mathbf{D}$  independent of the values in  $\mathbf{X}$ , the classical PLSR algorithm selects the latent variables  $\mathbf{t}_i = \mathbf{D}^T \boldsymbol{\psi}_i \in \mathbb{R}^{n_e \times 1}$  with the largest dependency on  $\mathbf{w}_i = \mathbf{X}^T \mathbf{v}_i \in \mathbb{R}^{n_e \times 1}$ , that is:

$$\begin{bmatrix} \mathbf{t}_i = \mathbf{D}^T \boldsymbol{\psi}_i \\ \mathbf{w}_i = \mathbf{X}^T \mathbf{v}_i \end{bmatrix} = \begin{cases} \max_{\boldsymbol{\psi}_i, \mathbf{v}_i} \{ \mathbf{v}_i^T \hat{\boldsymbol{\Sigma}}_{\mathbf{x}, \mathbf{d}} \boldsymbol{\psi}_i \} \\ \|\boldsymbol{\psi}_i\|_2 = 1, \|\mathbf{v}_i\|_2 = 1 \\ \mathbf{t}_i^T \mathbf{t}_j = 0, \text{ for all } j < i, i = 1, \dots, p. \end{cases}$$

This problem can be solved sequentially using the Non-linear Iterative Partial Least Squares (NiPALS) procedure (Rosipal and Krämer, 2006), or simultaneously by computing the SVD of the estimated covariance matrix  $\hat{\boldsymbol{\Sigma}}_{\mathbf{x}, \mathbf{d}}$  (Barker and Rayens, 2003). The matrices of latent variables  $\mathbf{T}$  and  $\mathbf{W}$  can then be obtained using the matrices  $\boldsymbol{\Psi} \in \mathbb{R}^{n_d \times p}$  and  $\boldsymbol{\Upsilon} \in \mathbb{R}^{n_x \times p}$ , corresponding to the first  $p$  left and right singular vectors of  $\hat{\boldsymbol{\Sigma}}_{\mathbf{x}, \mathbf{d}}$ .

Similarly to PCR, we assume that the latent variables  $\mathbf{T}$  are good predictors for  $\mathbf{D}$ . In addition we assume that  $\mathbf{W} = \boldsymbol{\beta} \mathbf{T} + \boldsymbol{\epsilon}_{\mathbf{T}}$  where  $\boldsymbol{\beta}$  is a diagonal matrix and  $\boldsymbol{\epsilon}_{\mathbf{T}}$  is a residual term. That is, there is a linear relationship between each element of  $\mathbf{w}_i$  and  $\mathbf{t}_i$ ,  $i = 1, \dots, p$  (Kaspar and Ray, 1993). As shown in Rosipal and Krämer (2006), the PLSR estimate of the Kalman gain  $\hat{\mathbf{K}}_{\text{PLSR}}$  is then given as:

$$\hat{\mathbf{K}}_{\text{PLSR}} = \mathbf{X} \mathbf{T} (\mathbf{W}^T \mathbf{D}^T \mathbf{D} \mathbf{T})^{-1} \mathbf{W}^T \mathbf{D}^T. \dots\dots\dots (11)$$

**Comments.** The RR, PCR and PLSR methods have the same computational complexity and memory requirements as the fastest implementation of the EnKF as explained in Appendix B. Note also that the PCR and PLSR methods are in general not scale invariant. This implies that the data ensemble matrix should be standardised before the dimension reduction, when the data is collected on different scales (Mardia et al., 1979). Finally it is interesting to note that the shrinkage estimators of the Kalman gain, previously applied in the EnKF literature (Evensen, 2007), can be thought of as hybrid methods combining Tikhonov regularisation and PCR. The regularisation term in these approaches is given by the covariance matrix of the observation errors, or a corresponding low rank representation.

**Cross-Validation.** The estimated matrix of regression coefficients will for all three methods be dependent on the choice of some hyperparameter  $\theta$ , denoted  $\mathbf{K}_\bullet(\theta)$ . For RR, this is the size of the regularisation parameter  $\xi$ , and for PCR and PLSR, the dimension of the reduced order space  $p$ . As discussed above, there is a tradeoff between how well a model is fitted to the training data, and how well it is suited for prediction purposes. To determine this tradeoff, CV is often used. The idea used in CV is to randomly split the ensembles into one set used for model fitting:  $\mathbf{Y}_{\text{Train}} = [\mathbf{X}_{\text{Train}}^T, \mathbf{D}_{\text{Train}}^T]^T \in \mathbb{R}^{n_y \times n_{\text{Train}}}$ , and one set used for testing the prediction capabilities:  $\mathbf{Y}_{\text{Test}} = [\mathbf{X}_{\text{Test}}^T, \mathbf{D}_{\text{Test}}^T]^T \in \mathbb{R}^{n_y \times n_{\text{Test}}}$ . Here  $n_{\text{Train}}$  and  $n_{\text{Test}}$  is used to denote the number of members in the training and test ensembles respectively. If  $\mathbf{Y}_{\text{Train}}$  consists of all realisations except  $\mathbf{y}^{(i)}$ , this is referred to as leave-one-out CV, and a brute implementation will increase the computational time with  $O(n_e^2)$  to the regression method used. Splitting the data into  $m$  sized portions randomly, and sequentially using  $m$  data points for testing

and the remaining  $m - 1$  part for model fitting purposes, is referred to as  $m$ -fold CV. This will increase the computational effort with  $O(n_e)$ . Typical values used for  $m$  in  $m$ -fold CV are 5 or 10 (Hastie et al., 2009). The optimal regression model can then be selected by minimising the Predictive Error Sum of Squares (PRESS) statistic, defined as:

$$\text{PRESS}(\theta) = \sum_{i=1}^{n_{\text{Test}}} \|\mathbf{x}_{\text{Test}}^{(i)} - \hat{\mathbf{K}}_{\bullet}(\theta) \mathbf{d}_{\text{Test}}^{(i)}\|_2^2. \dots\dots\dots (12)$$

To avoid overfitting of the regression model to the data when using the PCR and PLSR technique, a penalised version of the original PRESS statistic is given as:

$$\text{PRESS}_{\text{pen}}(p) = \sum_{i=1}^{n_{\text{Test}}} \frac{\|\mathbf{x}_{\text{Test}}^{(i)} - \hat{\mathbf{K}}_{\bullet}(p) \mathbf{d}_{\text{Test}}^{(i)}\|_2^2}{(\min\{n_e, n_d + 1\} - p)^2}.$$

Note that  $\text{PRESS}_{\text{pen}}$  is similar to the generalised PRESS statistic, which is an estimate of the leave one out CV prediction error. This follows because the effective degrees of freedom, defined as  $\text{tr}(\hat{\mathbf{K}}(p)\mathbf{D})$  (Hastie et al., 2009), is equal to  $p$  for both PCR and PLSR.

Finally it should be noted that applying the CV scheme does not necessarily lead to an increase in the computational demands. This follows because CV can be equally performed in the reduced order space (Hastie and Tibshirani, 2004). A more thorough discussion regarding the computational properties of the CV scheme can be found in Appendix B.

### Empirical Study

We consider two test cases similar to the ones used in Myrseth and Omre (2009), where the unknown vector of interest,  $\mathbf{x}_k$ ,  $k = 0, \dots, 10$ , with  $n_x = 100$ . Here  $x_{j,k}$  denotes the variable of interest at timestep  $k$  and location  $j \in \mathcal{L}_x$ , where  $\mathcal{L}_x$  is a regular grid. Observations are assumed to be made at timesteps  $0, \dots, 9$ , and the objective of this study is to assimilate the observed data, and predict at timestep 10.

The first test case, referred to as the linear case, is defined as a Gaussian prior at timestep zero, a linear forward function, and a linear Gaussian likelihood:

$$\begin{aligned} \mathbf{x}_0 &\sim \text{Gauss}_{n_x}(\mathbf{0}, \Sigma_{\mathbf{x}_0}) \\ \mathbf{x}_k &= \mathbf{A}_k \mathbf{x}_{k-1} \\ \mathbf{d}_k &= \mathbf{H} \mathbf{x}_k + \epsilon_{\mathbf{d}_k}, \end{aligned}$$

where  $\mathbf{0}$  is the null-vector of proper dimensions. Here  $\Sigma_{\mathbf{x}_0}$  is constructed based on an exponential covariance function,

$$\text{Cov}(x_{i0}, x_{j0}) = 20 \exp\left\{-\frac{3}{20}|i-j|\right\}, \dots\dots\dots (13)$$

the forward model is defined by the sparse matrix  $\mathbf{A}_k$ , where the elements  $\mathbf{A}_{k_l, m}$  for  $\{5(k-1) < l, m \leq 5(k+1)\}$ , are displayed in **Fig. 5** with  $\mathbf{A}_{k_l, m} = \delta_{l, m}$  otherwise, and  $\epsilon_{\mathbf{d}_k} \sim \text{Gauss}_{n_d}(\mathbf{0}, \mathbf{I})$ .  $\mathbf{H}$  is a sparse matrix with elements equal to one at the grid locations displayed in **Fig. 6**. Here we also see that observation  $d_i = \sum_{l=-1}^1 x_{j+l}$ ,  $i = 1, \dots, n_d$ , at 13 different grid locations  $j$ .

The second test case, referred to as the non-linear case, considers the same prior and likelihood as in the linear case defined

above. Here, however, the forward model is defined as the non-linear function:

$$\mathbf{x}_k = c\mathbf{A}_k(\mathbf{x}_{k-1} + \arctan(\mathbf{x}_{k-1})),$$

where  $c = 0.8$  is a scaling factor which ensures alignment of the variances for the non-linear and linear case, and the functional  $\arctan(\bullet)$  acts on the argument element-by-element.

**Fig. 7**, displays the reference  $\mathbf{x}_9$  and  $\mathbf{x}_{10}$  together with the observed data at timestep 9 for the two test cases. As we can see from this figure, the linear case has a more smooth behaviour in the left part of the grid nodes owing to the construction of  $\mathbf{A}_k$ . We also see a more spiky behaviour for the non-linear case. Also note that the state vector  $\mathbf{x}_k$  contains both dynamic and static variables, similar to what we have in a reservoir setting.

## Results with Discussion

We have considered the following EnKF updating schemes:

- Classical EnKF: Estimated Kalman gain matrix computed based on Eq. (9) using the correct  $\Sigma_{\epsilon_d}$ .
- PCR-0.99-EnKF:  $p$  selected based on the estimated proportion of explained variance (99 percent)
- PCR-CV-EnKF:  $p$  selected based on 10-fold CV and  $\text{PRESS}_{\text{pen}}$ .
- PLSR-CV-EnKF:  $p$  selected based on 10-fold CV and  $\text{PRESS}_{\text{pen}}$ .

The same initial ensemble and random numbers were used for all four updating schemes, with two different ensemble sizes:  $n_e = 100$  and  $n_e = 20$ . We also applied the CV criterion suggested by Hastie et al. (2009), described in Appendix B, to further avoid the problem of overfitting.

**Linear Case.** For the linear case, the prediction mean,  $E[x_{10}^u]$ , and 95% prediction interval are analytically obtainable using the KF recursions, and the results are displayed in **Fig. 8a**. The results obtained when applying the four different schemes outlined above, are displayed in **Figs. 8b** through **i**.

As we can see from Fig. 8b, the result obtained using the classical EnKF updating scheme is relatively reliable for  $n_e = 100$ . The classical EnKF solution matches the KF solution fairly well, although we do not see the same smooth behaviour in the estimated ensemble mean as seen in the KF solution. Moreover, there is a tendency of underestimating the prediction uncertainty. For  $n_e = 20$ , however, both the estimated mean and the prediction interval deviates dramatically from the KF solution. This is particularly true for grid nodes 1 to 35 and 65 to 100, where we have less observed data.

The estimated posterior mean obtained using the PCR-0.99-EnKF updating scheme matches the KF solution fairly well for  $n_e = 100$ . However, the estimated prediction interval is severely underestimated, and we are not able to capture the reference solution within the prediction interval. For the smallest ensemble size,  $n_e = 20$  the estimated posterior mean is highly variable and there is no uncertainty in the predictions as the updated ensemble has collapsed completely.

The obtained posterior mean using the PCR-CV-EnKF updating scheme appears to be a reliable estimate of the KF posterior mean for  $n_e = 100$ . Similarly we see that the estimated prediction interval matches the KF solution reasonably well

between grid nodes 35 and 65. In the area where the data is observed less frequently, however, the PCR-CV-EnKF updating scheme is not able to reduce the uncertainty at the observation sites. For  $n_e = 20$ , the posterior mean deviates more from the KF solution and is less smooth than for  $n_e = 100$ . The reference solution is reasonably well captured within the prediction interval, although the prediction interval is slightly underestimated relative to the KF solution. However, the results still appear to be relatively reliable.

The EnKF-CV-PLSR scheme is able to get a good representation of the KF solution for  $n_e = 100$ , both in terms of the estimated posterior mean and prediction interval. By decreasing the ensemble size to  $n_e = 20$ , the scheme is still able to obtain reasonable results, even though the prediction uncertainty is underestimated.

High variability in the estimated posterior mean and underestimation of the prediction interval are problems occurring in all four schemes when the ensemble size is only 20. The match with the KF solution is, however, reasonably good when using the PCR-CV-EnKF and PLSR-CV-EnKF schemes. For the two other schemes, the estimated posterior mean has a high variability and we see a dramatic underestimation of the prediction uncertainty.

Increasing the ensemble size to  $n_e = 100$ , improves the estimates of the posterior mean and prediction intervals for all four updating schemes. However, the classical EnKF and PCR-0.99-EnKF schemes tend to underestimate the prediction uncertainty. The best overall match with the KF solution is obtained using the PLSR-CV-EnKF updating scheme.

To further quantify the performance of the four updating schemes, the algorithms are rerun 100 times using different initial ensembles. Here we consider the Root Mean Squared Error of the estimated posterior mean (ARMSE) to the KF solution, and the percentage of the reference solution covered by the estimated prediction intervals. The results are shown in **Table 1**, where the estimated ARMSE and coverage of the initial ensemble are also included. As we can see from this table, the ARMSE of the posterior mean decrease significantly for all four EnKF schemes when  $n_e = 100$ . Compared to the ARMSE of the initial ensemble, the PLSR-CV-EnKF scheme shows the largest improvement with a 66 percent decrease. For  $n_e = 20$ , however, the classical EnKF updating scheme is not able to improve the ARMSE, while the PCR-0.99-EnKF scheme leads to an increase in ARMSE compared to the initial ensemble. Again PLSR-CV-EnKF has the smallest ARMSE with a reduction of 34 percent from the initial ensemble.

The coverage of the respective estimated 95% prediction intervals, is seen to be significantly underestimated for both the classical EnKF, and PCR-0.99-EnKF schemes. This is especially true for  $n_e = 20$ , where the prediction interval based on the PCR-0.99-EnKF solution only covers one percent of the reference solution. The PCR-CV-EnKF and PLSR-CV-EnKF updating schemes have similar and more reliable estimates of the prediction intervals, with the latter being slightly better than the former.

**Non-Linear Case.** For the non-linear case, analytical tractability is lost, we therefore use the results obtained with the classical EnKF with  $n_e = 100000$ , displayed in **Fig. 9a**, for comparison. The four different EnKF updating schemes outlined above, provide the results shown in **Figs. 9b** through **i**. Similar to the linear case, we see that the estimated posterior mean for the classical EnKF scheme is highly fluctuating with a severely underestimated prediction interval when  $n_e = 20$ . The most severe problems are in the PCR-0.99-EnKF solution, with a completely collapsed ensemble for  $n_e = 20$ . For both the PCR-CV-EnKF and PLSR-CV-EnKF schemes we observe relatively reasonable results when  $n_e = 20$  and 100. The

PCR-CV-EnKF scheme, however, tends to overestimate the prediction uncertainty at data locations for  $n_e = 100$ .

To quantify the performance the four schemes are rerun using 100 different initial ensembles. The results are shown in **Table 2**. Similar to the linear case, the PLSR-CV-EnKF scheme show the best performance in terms of estimated ARMSE and coverage for both ensemble sizes. Both the classical EnKF, and the PCR-0.99-EnKF schemes fail to cover the reference solution within their respective 95% prediction intervals for  $n_e = 20$ .

**Ensemble Size.** The number of ensemble members needed to achieve at least a 92% coverage of the reference solution in the respective estimated 95% prediction intervals are shown in **Table 3**. As we can see from this table, the classical EnKF updating scheme requires 5 times as many ensemble members as the PLSR-CV-EnKF scheme for the linear case, and 11 times as many ensemble members for the non-linear case.

Note also that the classical EnKF requires three times as many ensemble members in the non-linear case, compared to the linear case. This effect is believed to be caused by the non-linear forward model, because the Gaussian assumption made in Eq. 4 in this case is violated.

**Summary.** For both the linear and non-linear case the PLSR-CV-EnKF updating scheme gave the best representation of the reference posterior mean and prediction intervals. The PCR-CV-EnKF scheme tended to overestimate the prediction uncertainty at grid locations where data was sparsely observed for  $n_e = 100$ . The reason for this behaviour is that the penalised PRESS statistic ensured that only one component was selected at each updating step for both the PCR-CV-EnKF and PLSR-CV-EnKF schemes. Hence, the PCR-CV-EnKF scheme underfitted the model because some of the components important for prediction purposes were discarded. This is apparently not the case for the supervised PLSR-CV-EnKF updating scheme.

The classical EnKF suffers from both estimation uncertainty and model overfitting unless the ensemble size tends to infinity. For a small ensemble size this will lead to an ensemble collapsing, as seen in both the linear and non-linear case.

Amongst the four schemes discussed above, the PCR-0.99-EnKF updating scheme gave the least favourable representation of the prediction uncertainty for both ensemble sizes. This behaviour is expected because  $p$  was selected based on the estimated proportion of explained variance. As noted in Ledoit and Wolf (2004), the estimated eigenvalues of the empirical covariance matrix is known to be severely biased unless  $n_d/n_e \rightarrow 0$ , and the realisations in the data ensemble are independent identically distributed. In addition, no validation of the regression model is performed to evaluate the predictive capabilities.

### Reservoir Example

We consider a small synthetic reservoir model to further evaluate the performance of the PLSR-CV-EnKF and the classical EnKF scheme. The example is similar to the reservoir model used in Hegstad and Omre (2001), although the prior models for the porosity fields,  $\phi$ , and ln-permeability,  $\kappa$ , are different.

**Reservoir Description.** The reservoir grid domain is of size (10 000 x 10 000 x 100) ft, discretised into  $n = (10 \times 10 \times 15)$  regular grid blocks, with the top of the reservoir at depth 8 325 ft. The reference porosity and ln-permeability fields are generated by initially sampling from the Gaussian distribution described in Appendix C. The 100 traces of the reference ln-permeability and porosity are shown in **Figs. 11** and **12**. Initially the reservoir is fully saturated with oil, with pressure

5 800 psi at the equilibrium depth of 8 400 ft.

There are two horizontal production wells (P1, P2), and one vertical gas injection well (I1) at the locations displayed in **Fig. 10**, where the rate of the injection is assumed to be 65 000 Mscf/D. Production data from the reference model is simulated for 4 000 days using the commercial fluid flow simulator ECLIPSE™. Observations are made of the Gas/Oil Ratio (GOR) and Oil Production Rate (OPR) in the two production wells and the Bottomhole Pressure (BHP) in both the injection and production wells. **Figs. 13a** through **g** display the observed reference production data. As we can see from these figures, the production wells switch to BHP control when the pressure reaches 4 100 psi, which happens approximately after 1 200 days of production.

We assume that the model has already been updated using the data from the first 910 days of production. The task is therefore to continue updating the reservoir model based on new data collected every 30 days for the next 540 days of production, and with one final update after 1 640 days of production. Hence, there are seven data observations at 19 different timesteps  $k$ . Measurement errors are assumed to be additive Gaussian with a standard deviation of one percent of the observed value for the OPR and BHP, while for the GOR it is assumed to be 20 percent. In this study the reservoir state vector  $\mathbf{x}_k$ , contains  $\kappa$ ,  $\phi$ , logit-saturation,  $s_k$  and pressure,  $p_k$ .

**Results with Discussion.** The classical EnKF with four different ensemble sizes:  $n_e = 20, 100, 1\,000$  and  $1\,500$ , and the PLSR-CV-EnKF scheme with  $n_e = 20$  are evaluated. Initially we generate  $1\,500$  porosity and ln-permeability fields using the prior model described in Appendix C. To make the results comparable, the initial ensemble members for the smaller ensemble sizes are then selected as the first 20, 100 and 1 000 members of the largest initial ensemble respectively. Note that the initial saturation and pressure are assumed to be known throughout the reservoir.

The forecasted production obtained when we restart the simulator from timestep zero and predict for 4 000 days, based on the updated  $\kappa$  and  $\phi$  values are shown in **Fig 14**. As we can see from this figure, the initial ensemble fully spans the reference solution, and there is a relatively high uncertainty regarding the time of the gas breakthrough. Looking at results based on the classical EnKF with  $n_e = 20$ , we see that the average of the production forecast based on the updated ensemble members are missing the reference production. Moreover, reduced uncertainty in the production forecast causes the ensemble members not to span the reference production. For both the classical EnKF scheme with  $n_e = 1\,500$  and the PLSR-CV-EnKF scheme with  $n_e = 20$ , however, the forecasts are correctly centred at the reference production and the uncertainty is considerably reduced.

To further quantify the results obtained using the classical EnKF and PLSR-CV-EnKF updating schemes, we investigate how well the updated ensemble members span the reference  $\phi$  and  $\kappa$ . The estimated posterior mean and 95 % prediction interval for  $\kappa$ , are displayed **Fig. 15**. Note that the results for porosity look similar, and are therefore not presented.

The updated ensemble based on the classical EnKF scheme is not able to span the true ln-permeability for  $n_e = 20$ , and the ensemble has almost collapsed into a single realisation. Increasing the ensemble size to  $n_e = 100$ , does improve the uncertainty estimates. However, the ensemble mean appears to be highly variable, and we also see a tendency of overestimating  $\kappa$  and  $\phi$  in many of the grid blocks. This leads to a bias in the production forecasts, not shown here, however. For the two largest ensemble sizes both the posterior mean, and prediction intervals appear to be similar. However, when rerunning the

reservoir simulator from timestep zero using the updated  $\kappa$  and  $\phi$  as input for  $n_e = 1\,000$ , we obtain production forecasts that deviates from the reference production curves, again not shown here. This is not the case for  $n_e = 1\,500$ , as shown in Fig. 14. The PLSR-CV-EnKF updating scheme with  $n_e = 20$ , on the other hand, appears to provide a much better representation of the prediction uncertainty, although we are not able to fully cover 95% of the reference solution.

Similar to the empirical study above, the classical EnKF and PLSR-CV-EnKF updating schemes were rerun 100 times using different initial ensembles of size  $n_e = 20$ . The results are summarised in **Table 4**, where we estimate the RMSE between the ensemble members and the reference ln-permeability field (EnRMSE), and the percentage of the reference solution the estimated prediction intervals cover. As we see from these results, the classical EnKF algorithm clearly underestimates the prediction uncertainty for the smallest ensemble size,  $n_e = 20$ .

**Fig. 16** contains two realisations of ln-permeability from the initial ensemble and the corresponding realisations after the final updating step using the classical EnKF updating scheme with four different ensemble sizes and the PLSR-CV-EnKF scheme with  $n_e = 20$ . We also display the estimated posterior means based on the first 20 ensemble members, which are equal at the initial timestep for all four ensemble sizes considered. The two realisations are different at the initial timestep, although the strong spatial correlation is present in both cases. From the average of the initial ensemble we see that there appears to be no particular trend in the initial model.

For the classical EnKF with  $n_e = 20$ , the realisations has collapsed and there appears to be a very high variability between neighbouring grid blocks. Moreover, the ln-permeability is well outside the range of the prior model at many of the grid locations. The realisations for  $n_e = 100$  appear to give a much better representation of the reference ln-permeability. However, we see that the ensemble members fail to capture the low permeable layers around horizontal cross-section 10 in the reference solution. For the two largest ensemble sizes,  $n_e = 1\,000$  and  $1\,500$ , the spatial structure in the reference solution appears to be much better preserved in the updated realisations. We also observe the middle layer of low permeability present in the reference solution for both the updated realisations and ensemble mean. Again we see that the PLSR-CV-EnKF updating scheme with  $n_e = 20$  provides updated realisation which appears to have many of the same features present in the prior model and reference solution.

Note that when the classical EnKF updating scheme with  $n_e = 20$  was rerun several times, we observed that the EnRMSE often became larger at the final updating step than for the initial ensemble. This is caused by spurious correlations in the estimated covariance matrix  $\hat{\Sigma}_x$  for small ensemble sizes (Evensen, 2007). Note, however, that this problem is also present in the classical EnKF updating scheme for  $n_e = 100$  and  $n_e = 1\,000$ . These results suggest that spurious correlations are not only introduced by estimation uncertainty, but also by overfitting the regression model resulting from collinear ensemble members.

In **Figs 17a** and **b**, the EnRMSE and the percentage of the reference ln-permeability located within the estimated 95% prediction interval are displayed as a function of timestep  $k$ . The EnRMSE starts to increase after  $k = 10, 12$  and  $15$ , for  $n_e = 20, 100$  and  $1\,000$  respectively, with the most dramatic effect for the smallest ensemble size. At the same time, we see that the coverage decrease as  $k$  increases, again with the largest effect for  $n_e = 20$ . For this particular reservoir model we have to increase  $n_e$  to  $1\,500$  to have a decreasing trend in the EnRMSE, while preserving the coverage applying the classical

EnKF updating scheme.

The PLSR-CV-EnKF scheme with  $n_e = 20$  is able to preserve the decreasing trend in the EnRMSE, with only a small decrease in the coverage as  $k$  increases. This appears to be caused by the reduced coupling of the updated ensemble members when using PLSR. Note, however, that both the classical EnKF and PLSR-CV-EnKF updating scheme with  $n_e = 1500$  and 20 respectively, will eventually see the same increase in EnRMSE if data assimilation is continued into a distant future.

The increased coupling of the updated ensemble members can be quantified by the estimated rank of the updated In-permeability ensemble, displayed in **Fig. 18**. Here we have computed the loss in relative numerical rank for the updated ensemble at three different timesteps  $k$ . As seen from this figure the relative loss in rank for the updated ensemble members based on the classical EnKF updating scheme increases with timestep  $k$ . This is caused by overfitting of the regression model since the ensemble members become more collinear. We also see that this effect is more prominent for smaller ensemble sizes. Note that by construction the PLSR-CV-EnKF scheme has a rank loss of one for all timesteps, as explained in Appendix D, which for this case makes the approach less vulnerable to overfitting.

## Conclusion

We have formulated an EnKF updating scheme based on shrinkage regression techniques known from multivariate linear regression. The purpose of these methods is to replace the unbiased classical estimator of the Kalman gain matrix with bias alternatives, having improved predictive capabilities. Two of the techniques were considered on small linear and non-linear toy examples, namely Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR). Common to both these methods is that the predictor variables are transformed into a reduced order subspace, before the regression model is computed. The dimension of this subspace is therefore a prior hyperparameter that needs to be selected. When the subspace dimension was selected based on Cross-Validation (CV), both the PCR and PLSR updating schemes performed significantly better than the classical EnKF for small ensemble sizes, with the supervised PLSR scheme providing slightly better results. However, when the subspace dimension used in PCR was selected based on the commonly applied theoretical criterion, the scheme suffered from similar problems as the classical EnKF. That is, low forecast precision and severe underestimation of the prediction uncertainty for small ensemble sizes.

We further compared the classical EnKF and PLSR updating schemes on a synthetic reservoir case study. Using the PLSR scheme, reasonable estimates of the prediction uncertainty for the porosity and permeability fields were obtained using only 20 ensemble members. The classical EnKF, on the other hand, required 75 times as many ensemble members to obtain similar results. This was caused by severe problems of model overfitting, because the ensemble members became increasingly collinear when the reservoir production data was assimilated.

## Nomenclature

- $Cov(\bullet)$  = covariance  
 $\mathbf{d}_k$  = data vector  
 $\mathbf{D}$  = centred data ensemble  
 $\delta_{i,j}$  = Kronecker delta  
 $\epsilon_\bullet$  = model/observation error  
 $f(\mathbf{x})$  = probability density function (pdf)  
 $f(\mathbf{x}|\mathbf{y})$  = conditional pdf of  $\mathbf{x}$  given  $\mathbf{y}$   
 $\zeta(\bullet)$  = likelihood model  
 $\theta$  = shrinkage regression hyperparameters  
Gauss.  $(\bullet, \bullet)$  = Gaussian distribution  
 $\mathbf{G}_\bullet$  = Gram matrix  
 $\mathbf{I}$  = identity matrix  
 $\mathbf{K}$  = Kalman gain matrix  
 $\kappa$  = ln-permeability [md]  
 $\mu_\bullet$  = expectation vector  
 $\hat{\mu}_\bullet$  = estimated mean vector  
 $\|\mathbf{a}\|_2$  = Euclidean vector norm  
 $n_d$  = dimension data vector  
 $n_x$  = dimension state vector  
 $n_e$  = number of ensemble members  
 $n_y = n_x + n_d$   
 $O(n_x^2)$  = upper bound, computational complexity  
 $\omega(\bullet)$  = forward model  
 $p$  = dimension reduced order space, PCR, PLSR  
PRESS( $\theta$ ) = Predictive Error Sum of Squares (PRESS)  
PRESS( $\theta$ )<sub>pen</sub> = Penalised PRESS  
 $p$  = pressure [psi]  
 $\phi$  = porosity  
 $\xi$  = scaling parameter RR  
 $s$  = logit gas saturation  
 $\Sigma_\bullet$  = covariance matrix  
 $\Sigma_{\epsilon_d}$  = covariance of the observation errors  
 $\hat{\Sigma}_\bullet$  = estimated covariance matrix  
 $\text{tr}(\mathbf{A})$  = trace operator  
 $\text{Var}(\bullet)$  = variance  
 $\mathbf{x}_k$  = state vector  
 $\mathbf{X}$  = centred state vector ensemble  
 $\mathbf{x}_k^u$  = uncondition state vector:  $[\mathbf{x}_k | \mathbf{d}_0, \dots, \mathbf{d}_{k-1}]$   
 $\mathbf{x}_k^c$  = conditioned state vector:  $[\mathbf{x}_k | \mathbf{d}_0, \dots, \mathbf{d}_k]$   
 $\mathbf{y} = [\mathbf{x}^T, \mathbf{d}^T]$   
 $\mathbf{Y} = [\mathbf{X}^T, \mathbf{D}^T]^T$   
 $\mathbf{0}$  = vector with entries equal to zero

## Subscript

- $i_x$  = coordinate in  $x$ -direction  
 $i_y$  = coordinate in  $y$ -direction  
 $i_z$  = coordinate in  $z$ -direction  
 $k$  = timestep  
 $t_k$  = time at timestep  $k$   
Test = test data  
Train = training data

## Superscript

- $c$  = conditioned, meaning after data assimilation  
 $(i)$  = realisation number  
 $T$  = matrix/vector transposed  
True = reference solution  
 $u$  = unconditioned, meaning the output of the forward model

## Acknowledgements

This work is funded by the Uncertainty in Reservoir Evaluation (URE) initiative at NTNU. We would also like to thank the Department of Petroleum Engineering and Applied Geophysics at NTNU for granting us access to their computational resources.

## References

- Aanonsen, S. I., Nævdal, G., Oliver, D. S., Reynolds, A. C. and Vallès, B. 2009. Ensemble Kalman filter in reservoir engineering - a review. *SPE Journal*, **14**(3): 393–412.
- Adabir, K. M. and Magnus, J. R. 2005. *Matrix Algebra*. Cambridge University Press, New York.
- Anderson, T. W. 2003. *An introduction to multivariate statistical analysis*, 3 edition. Wiley.
- Barker, M. and Rayens, W. 2003. Partial least squares for discrimination. *Journal of Chemometrics*, **17**: 166–173.
- Cook, D. R. 2007. Fisher Lecture: Dimension Reduction in Regression. *Statistical Science*, **22**(1): 1–26.
- Doucet, A., Godsill, S. and Andrieu, C. 2000. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, **10**: 197–208.
- Efron, B. 2004. The Estimation of Prediction Error: Covariance Penalties and Cross-Validation. *Journal of the American Statistical Association*.
- Evensen, G. 1994. Sequential data assimilation with nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*.
- Evensen, G. 2003. The Ensemble Kalman Filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, pages 343–367.
- Evensen, G. 2007. *Data Assimilation. The Ensemble Kalman Filter*. Springer.
- Farrer, D. E. and Glauber, R. R. 1967. Multicollinearity in Regression Analysis: The Problem Revisited. *The Review of Economics and Statistics*, **49**(1): 92–107.
- Furrer, R. and Bengtsson, T. 2007. Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis*, pages 227–255.
- Hadi, A. S. and Ling, R. F. 1998. Some cautionary notes on the use of principal components regression. *The American Statistician*, **52**(1): 15–19.
- Hastie, T. and Tibshirani, R. 2004. Efficient quadratic regularization for expression arrays. *Biostatistics*, **5**(3): 329–340.
- Hastie, T., Tibshirani, R. and Friedman, J. 2009. *The Elements of Statistical Learning; Data Mining, Inference, and Prediction*, 2 edition. Springer, New York.
- Hegstad, B. K. and Omre, H. 2001. Uncertainty in production forecasts based on well observations, seismic data and production history. *Society of Petroleum Engineers Journal*, pages 409–425.
- Helland, I. S. 2001. Some theoretical aspects of partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, **58**(2): 97–107.
- Hoerl, A. E. and Kennard, R. W. 1970. Ridge Regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**(3): 55–67.
- Hotelling, H. 1933. Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, **24**(6).
- Houtekamer, P. L. and Mitchell, H. L. 1998. Data Assimilation Using an Ensemble Kalman Filter Technique. *Monthly Weather Review*, **126**: 796–811.
- Houtekamer, P. L. and Mitchell, H. L. 1999. Reply. *Monthly Weather Review*, **127**: 1378–1379.
- Höskuldsson, A. 1988. PLS Regression Methods. *Journal of Chemometrics*, **2**: 211–228.
- Jolliffe, I. 1982. A Note on the Use of Principal Components in Regression. *Applied Statistics*, **31**(2).
- Jolliffe, I. T. 2002. *Principal Component Analysis*, 2 edition. Springer.

- Kalivas, J. H. 1999. Cyclic subspace regression with analysis of the hat matrix. *Chemometrics and Intelligent Laboratory Systems*, **45**: 215–224.
- Kalman, R. E. 1960. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME - Journal of Basic Engineering*.
- Kaspar, M. H. and Ray, W. H. 1993. Partial Least Squares Modelling as Successive Singular Value Decomposition. *Computers and Chemical Engineering*, **17**: 985–989.
- Ledoit, O. and Wolf, M. 2004. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, pages 365–411.
- Mardia, K. V., Kent, J. T. and Bibby, J. M. 1979. *Multivariate Analysis*. Academic Press, London.
- Myrseth, I. B. and Omre, H. 2009. *Computational Methods for Large-Scale Inversion Problems and Quantification of Uncertainty*, chapter Ensemble Kalman filter and related filters. Wiley.
- Myrseth, I. B., Sætrum, J. and Omre, H. 2009. Resampling the Ensemble Kalman Filter. Paper submitted for publication.
- Ripley, B. D. 1996. *Pattern Recognition and Neural Networks*. Cambridge University Press.
- Rosipal, R. and Krämer, N. 2006. *Subspace, Latent Structure and Feature Selection Techniques*, volume 2940 of *Lecture Notes in Computer Science*, chapter Overview and Recent Advances in Partial Least Squares, pages 34–51. Springer.
- Rännér, S., Lindgren, F., Gelandi, P. and Wold, S. 1994. A PLS Kernel Algorithm for Data Sets with Many Variables and Fewer Objects. Part 1: Theory and Algorithm. *Journal of Chemometrics*, **8**: 111–125.
- Sacher, W. and Bartello, P. 2008. Sampling Errors in Ensemble Kalman Filtering. Part I: Theory. *Monthly Weather Review*, pages 3035–3049.
- Seber, G. A. F. and Lee, A. J. 2003. *Linear Regression Analysis*. Wiley.
- Skjervheim, J. A., Evensen, G., Aanonsen, S., Ruud, B. and Johansen, T. 2005. Incorporating 4D Seismic Data in Reservoir Simulation Models Using Ensemble Kalman Filter. *SPE*.
- Strang, G. 1988. *Linear Algebra And Its Applications*. Thomson Learning.
- Tikhonov, A. N. and Arsenin, V. A. 1977. *Solution of Ill-posed Problems*. Winston & Sons, Washington.
- van Leeuwen, P. J. 1999. Comments on "Data Assimilation Using an Ensemble Kalman Filter Technique". *Monthly Weather Review*, **127**: 1374–1377.
- Wold, H. 1975. *Quantitative Sociology: International perspectives on mathematical and statistical model building*, chapter Path models with latent variables: The NiPALS approach, pages 307–357. Academic Press.
- Zeng, X.-Q., Wang, M.-W. and Nie, J.-Y. 2007. Text Classification Based on Partial Least Square Analysis. In *The 22nd Annual ACM Symposium on Applied Computing, Special Track on Information Access and Retrieval*.

## Appendix A, Use of Singular Value Decomposition

The eigenvectors of the empirically estimated covariance matrix of the data vector,  $\hat{\Sigma}_d$ , is given as the column vectors  $U \in \mathbb{R}^{n_d \times n_d}$  obtained when performing Singular Value Decomposition (SVD) on  $\frac{1}{\sqrt{n_e}} D = U S V^T$  (Strang, 1988). Moreover, the eigenvalues are given as  $\lambda_i = s_{ii}^2$ , where  $s_{ii}$  is the  $i$ th singular value, given as the  $i$ th diagonal element of the matrix  $S \in \mathbb{R}^{n_d \times r}$  where  $r$  is the rank of  $D$ . Hence, the ensemble matrix  $D$  can be approximated by a truncated version:

$$D_p = U_p S_p V_p^T, \dots \dots \dots \quad (\text{A-1})$$

where  $U_p = [u_1, \dots, u_p] \in \mathbb{R}^{n_d \times p}$ ,  $S_p = \text{diag}_p(s_p) \in \mathbb{R}^{p \times p}$ , and  $V_p = [v_1, \dots, v_p] \in \mathbb{R}^{n_e \times p}$ , with  $v_i$  given as the  $i$ th eigenvector of the matrix  $G_d = D^T D \in \mathbb{R}^{n_e \times n_e}$ . Here the notation  $S_p = \text{diag}_p(s_p)$  is used to denote that the  $p$  dimensional matrix  $S_p$  is a diagonal matrix with the vector  $s_p$  on the main diagonal.

By use of the truncated SVD, the PCR estimate for the matrix of regression coefficient is given as:

$$\hat{\mathbf{K}}_{\text{PCR}} = \mathbf{X} \mathbf{V}_p \mathbf{S}_p^{-1} \mathbf{U}_p^T. \dots\dots\dots (\text{A-2})$$

Thus, the computation of  $\hat{\mathbf{K}}_{\text{PCR}}$  is efficient, both in terms of speed and memory use, compared to working with full dimensional covariance matrices. This is especially true when  $n_d$  is larger than  $n_e$ . It should be noted that setting  $p = n_e - 1$ ,  $\hat{\mathbf{K}}_{\text{PCR}}$  minimises the mean squared error in the rank deficient case. Also note that a similar trick can be utilised for the RR estimate, because  $\mathbf{D}\mathbf{D}^T + \xi\mathbf{I}$  share the same eigenvectors as  $\mathbf{D}\mathbf{D}^T$ , and with eigenvalues given as  $\tilde{\lambda}_i = \lambda_i + \xi$ , where  $\lambda_i$  are the eigenvalues of  $\mathbf{D}\mathbf{D}^T$ ,  $i = 1, \dots, \min\{n_e - 1, n_d\}$ .

### Appendix B, Computational Properties

The classical EnKF can be modified by simply replacing the estimated Kalman gain with either the RR, PCR or PLSR estimates. For all three methods the computational complexity is  $O(\max\{n_d, n_x\} \cdot n_e^2)$ . For the RR and PCR estimate this corresponds to the cost of performing a SVD on  $\mathbf{D}$  and the matrix-matrix multiplication  $\mathbf{X}\mathbf{V}^T$ , while for the PLSR estimate this is caused by the computation of the two matrices  $\mathbf{G}_x = \mathbf{X}^T \mathbf{X} \in \mathbb{R}^{n_e \times n_e}$  and  $\mathbf{G}_d = \mathbf{D}^T \mathbf{D} \in \mathbb{R}^{n_e \times n_e}$ .

When computing the PLSR estimate of the Kalman gain, we see from Eq. (11) that this requires the inversion of a  $p \times p$  dimensional matrix, which in the general case requires  $O(p^3)$  floating point operations (flops). However, it can be shown (Höskuldsson, 1988; Ränner et al., 1994), that this matrix is lower triangular, reducing the number of flops required to  $O(p^2)$ . If both  $n_x$  and  $n_d$  are large, the classical NiPALS algorithm and computing the SVD of  $\hat{\Sigma}_{x,d}$ , will be computationally demanding. Note, however, that this problem can be avoided by the algorithm outlined in Ränner et al. (1994).

When  $m$ -fold CV is applied to select the optimal number of components for PCR and PLSR, the computational complexity is  $O(\max\{n_e \cdot n_x, (m - 1) \cdot n_d\} n_e^2)$ , when the PRESS statistic is computed for all possible values for  $p$  in the truncation. Note that the computational complexity can be further reduced when  $n_d > n_e - 1$ . As explained in Hastie and Tibshirani (2004), the same results will be obtained if we perform the  $m$ -fold CV based on the  $n_e - 1$  PC, instead of the data ensemble. Hence, only a single SVD will be necessary to perform. Further note that if we replace the Euclidean norm used in the PRESS statistic in Eq. (12), with the (pseudo) norm:

$$\|\mathbf{a}\|_{\hat{\Sigma}_x}^2 = n_e \mathbf{a}^T \hat{\Sigma}_x \mathbf{a} = \mathbf{a}^T \mathbf{X} \mathbf{X}^T \mathbf{a},$$

we can further reduce the computational complexity to  $O(\max\{n_x, n_e^2, n_d\} n_e^2)$ . For large dimensional state vectors, we therefore see that the additional computational demands caused by  $m$ -fold CV can be negligible as long as  $n_e^2 < \max\{n_x, n_d\}$ .

When overfitting is a severe problem for PCR or PLSR, the CV rule described in Hastie et al. (2009) can be applied. Rather than selecting  $p$  at the global minimum of the PRESS statistic, the optimal number of components,  $p^*$ , should be given as the smallest number of components such that  $\text{PRESS}(p^*) \leq \text{PRESS}(p^G) + \hat{\sigma}_{\text{PRESS}(p^G)}$ . Here  $\text{PRESS}(p^G)$  and  $\hat{\sigma}_{\text{PRESS}(p^G)}$  corresponds to the estimated value and standard deviation of the PRESS statistic at the global minimum  $p^G$  respectively.

Although CV can lead to higher computational demands if an exhaustive search for the optimal number of components is carried out, it should be noted that testing for all possible combinations of  $p$  is often not required. This is especially true when using the PLSR approach because this tend to require a smaller number of components than PCR (Kalivas, 1999; Helland,

2001).

### Appendix C, Prior Distribution

For the reservoir example the porosity and ln-permeability fields are described by the following prior distribution: Initially, a realisation  $\mathbf{z}^{(i)} \in \mathbb{R}^{n \times 1}$  is generated from a Gaussian distribution with mean 0.25 and covariance matrix  $\Sigma_{\mathbf{z}} \in \mathbb{R}^{n \times n}$ , with  $\Sigma_{\mathbf{z}_\Delta}$  defined by an exponential covariance function:

$$c(\Delta) = \sigma_{\mathbf{z}}^2 \exp \left\{ -\sqrt{\|\Delta\|_l^2} \right\}.$$

Here  $\sigma_{\mathbf{z}} = 0.02$  and

$$\|\Delta\|_l^2 = \left( \frac{\Delta_x}{l_x} \right)^2 + \left( \frac{\Delta_y}{l_y} \right)^2 + \left( \frac{\Delta_z}{l_z} \right)^2,$$

with  $l_x = l_y = 10000$  and  $l_z = 100$ . We further assume that

$$[\mathbf{w}|\mathbf{z}] = 25\mathbf{z} + \epsilon_{\mathbf{w}|\mathbf{z}},$$

where  $\epsilon_{\mathbf{w}|\mathbf{z}} \sim \text{Gauss}_n(\mathbf{0}, \mathbf{I})$ . A convolution between the realisations and a three dimensional standard Gaussian kernel over a (5 x 5 x 5) neighbourhood is then applied to generate porosity and ln-permeability fields. This entails that  $\phi^{(i)} = \mathbf{C}\mathbf{z}^{(i)}$  and  $\kappa^{(i)} = \mathbf{C}\mathbf{w}^{(i)}$ , where  $\mathbf{C} \in \mathbb{R}^{n \times n}$  is the convolution matrix. Note that the purpose of this convolution is simply to increase the spatial continuity in the realisations. Further note that in practice the MATLAB function `smooth3(•)` is used.

### Appendix D, Rank of the Updated Ensemble

Let  $\mathbf{X}^u \in \mathbb{R}^{n_x \times n_e}$  and  $\mathbf{D} \in \mathbb{R}^{n_d \times n_e}$  be two non-centred ensemble matrices. Further let  $\mathbf{C} = \mathbf{I} + \frac{1}{n_e} \mathbf{1}\mathbf{1}^T$  denote the centring matrix, where  $\mathbf{1}$  is a  $n_e$ -dimensional column vector with all entries equal to one. The EnKF updating scheme based on PLSR can then be written as:

$$\mathbf{X}^c = \mathbf{X}^u [\mathbf{I} + \mathbf{T}\mathbf{A}^{-1}\mathbf{W}^T\mathbf{D}^T(\mathbf{d}\mathbf{1}^T - \mathbf{D})], \dots \dots \dots \quad (\text{D-1})$$

where

$$\mathbf{A} = (\mathbf{W}\mathbf{D}^T\mathbf{D}\mathbf{T}).$$

This follows from the definition of  $\hat{\mathbf{K}}_{\text{PLSR}}$  and the identities:

$$\mathbf{T} = \mathbf{C}\mathbf{D}^T\mathbf{\Psi} = \mathbf{C}\mathbf{C}\mathbf{D}^T\mathbf{\Psi} = \mathbf{C}\mathbf{T} \dots \dots \dots \quad (\text{D-2})$$

and  $\mathbf{W} = \mathbf{C}\mathbf{W}$ , where we have used that  $\mathbf{C}$  is an idempotent matrix.

For the EnKF updating scheme defined in Eq. (D-1) using  $p$  components, assuming  $\text{rank}(\mathbf{X}^u) = n_e$ , the following result then holds:

**Result 1.** *Assuming  $\text{rank}(\mathbf{X}^u) = n_e$ , the rank of the updated state vector ensemble for the EnKF updating scheme based on PLSR and PCR is equal to  $n_e - p$ .*

*Proof.* Consider the matrix:

$$\mathbf{H} = \mathbf{I} + \mathbf{T}\mathbf{A}^{-1}\mathbf{W}^T\mathbf{D}^T(\mathbf{d}\mathbf{1}^T - \mathbf{D}) \in \mathbb{R}^{n_e \times n_e}.$$

By use of Eq. (D-2) and the property that  $\mathbf{1}^T\mathbf{C} = 0$ , we obtain the identity  $\mathbf{H} = \mathbf{H}^2$ . Let  $(\lambda_i, \mathbf{e}_i)$ ,  $i = 1, \dots, n_e$  be an eigenvalue, -vector pair of  $\mathbf{H}$ . From the identity

$$\lambda_i \mathbf{e}_i = \mathbf{H}\mathbf{e}_i = \mathbf{H}\mathbf{H}\mathbf{e}_i = \lambda_i^2 \mathbf{e}_i,$$

we then see that all eigenvalues of the idempotent matrix  $\mathbf{H}$  are equal to zero or one. Hence, the result follows from the identity:

$$\text{rank}(\mathbf{X}^c) = \text{rank}(\mathbf{X}^u\mathbf{H}) = \text{rank}(\mathbf{H}) = \text{tr}(\mathbf{H}),$$

where second equality holds because  $\text{rank}(\mathbf{X}^u) = n_e$  Adahir and Magnus (2005) and the third equality holds because  $\mathbf{H}$  is idempotent. We leave the proof for the EnKF updating scheme based on PCR to the reader.  $\square$

**Jon Sæstrom** is a PhD student at the Norwegian University of Science and Technology, Trondheim, Norway. **Henning Omre** is Professor at the Norwegian University of Science and Technology, Trondheim, Norway.

$n_e$	Scheme	ARMSE	Coverage %
20	No Updating	9.96	89
20	Classical EnKF	9.96	24
20	PCR-0.99-EnKF	11.4	1
20	PCR-CV-EnKF	7.26	58
20	PLSR-CV-EnKF	6.53	59
100	No Updating	4.39	96
100	Classical EnKF	1.75	84
100	PCR-0.99-EnKF	2.42	53
100	PCR-CV-EnKF	1.74	92
100	PLSR-CV-EnKF	1.49	96

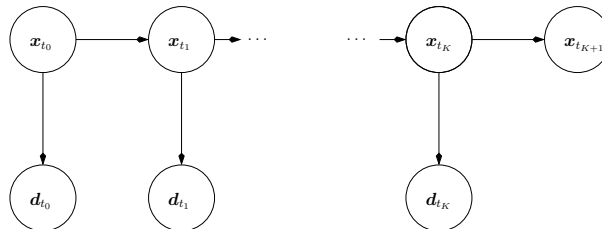
**Table 1—Estimated Root Mean Squared Error of the posterior mean (ARMSE) and coverage of the reference solution in the estimated 95% prediction intervals for the linear case based on 100 different initial ensembles.**

$n_e$	Scheme	ARMSE	Coverage %
20	No Updating	10.2	89
20	Classical EnKF	8.07	21
20	PCR-0.99-EnKF	8.95	1
20	PCR-CV-EnKF	6.63	65
20	PLSR-CV-EnKF	5.88	65
100	No Updating	4.48	96
100	Classical EnKF	1.52	79
100	PCR-0.99-EnKF	2.19	56
100	PCR-CV-EnKF	1.76	92
100	PLSR-CV-EnKF	1.25	93

**Table 2—Estimated Root Mean Squared Error of the posterior mean (ARMSE) and coverage of the reference solution in the estimated 95% prediction intervals for the non-linear case based on 100 different initial ensembles.**

Scheme	Linear Case	Non-Linear Case
Classical EnKF	300	1 000
PCR-0.99-EnKF	550	1 250
PCR-CV-EnKF	100	100
PLSR-CV-EnKF	60	90

**Table 3—Ensemble size  $n_e$  required to achieve at least a 92 percent coverage of the reference state vector within the estimated 95 percent prediction interval for the four EnKF updating schemes considered for the empirical case study. Estimates are based on 100 reruns using different initial ensembles.**



**Fig. 1—Stochastic Directed Acyclic Graph (DAG) of the model considered.**

Scheme		EnRMSE	Coverage %
No Update	$\kappa$	24.1	95
Classical EnKF	$\kappa$	22.0	21
PLSR-CV-EnKF	$\kappa$	19.1	61
No Update	$\phi$	0.86	96
Classical EnKF	$\phi$	0.74	21
PLSR-CV-EnKF	$\phi$	0.64	62

Table 4—Estimated Root Mean Squared Error of the difference between the updated ensemble members  $x_{19}^{(i)c}$  and the reference solution (EnRMSE), and coverage of the reference solution in the estimated 95% prediction intervals for the two static variables in the reservoir example. Estimates computed based on 100 different initial ensembles with  $n_e = 20$ .

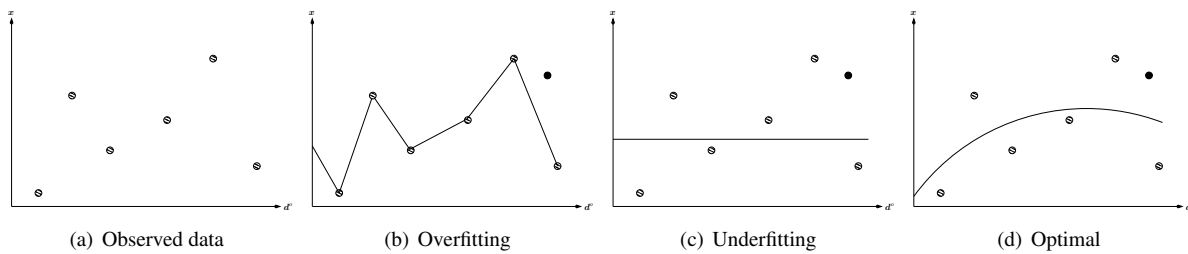


Fig. 2—The figure illustrates the problem of over-/underfitting of a model to the available training data. The training data is shown as dashed bullets, the test data is shown as a solid black bullet, while the fitted model is shown as a solid line.

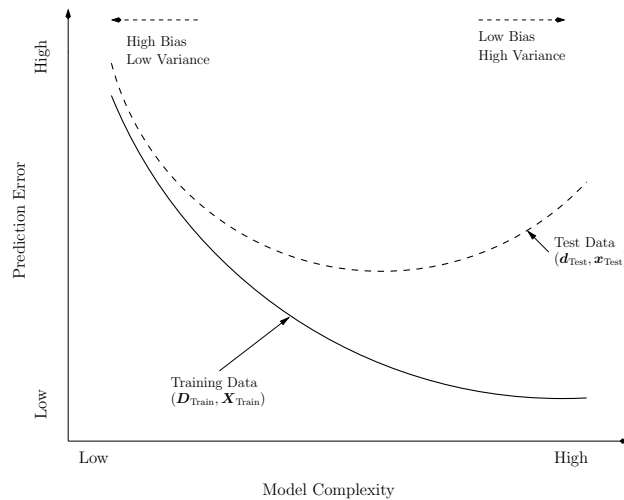


Fig. 3—Tradeoff between Bias and Variance: Overfitting the data by increasing the model complexity tend to increase the variance in model predictions. Based on a similar figure found in Hastie et al. (2009).

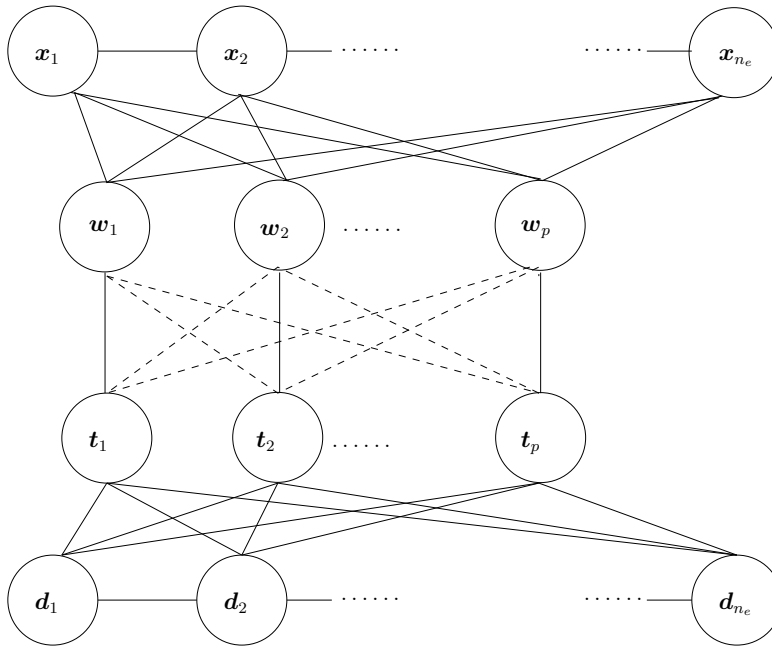


Fig. 4—Graphical presentation of the model assumptions made in PLSR, inspired by a similar figure found in Zeng et al. (2007). Connected lines implies that there is a direct connection between the variables, while dashed lines implies an implicit connection.

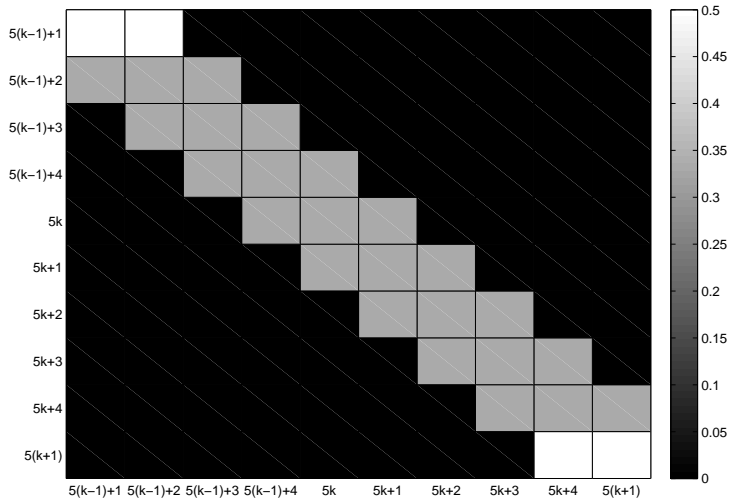


Fig. 5—Graphical presentation of the elements of the matrix  $A_{k,l,m}$ , for  $5(k-1) < l, m \leq 5(k+1)$ , for the forward model in the empirical case study. At all other grid locations  $A_{k,l,m} = \delta_{l,m}$

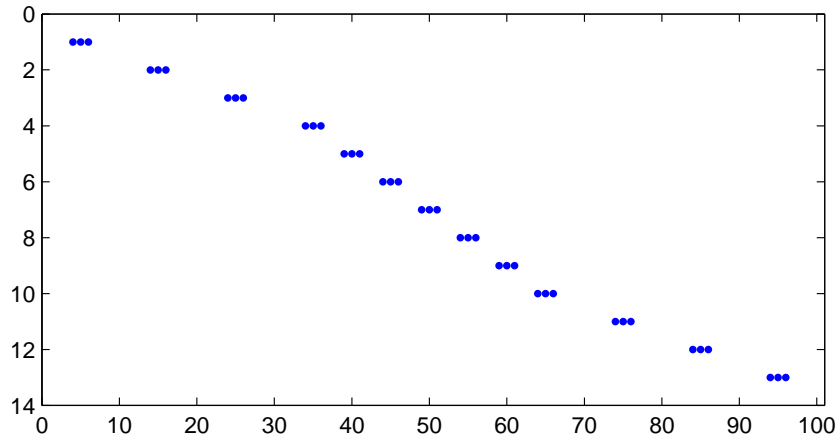


Fig. 6—Graphical presentation of the non-zero elements of the matrix  $H$ , for the likelihood function in the empirical case study.

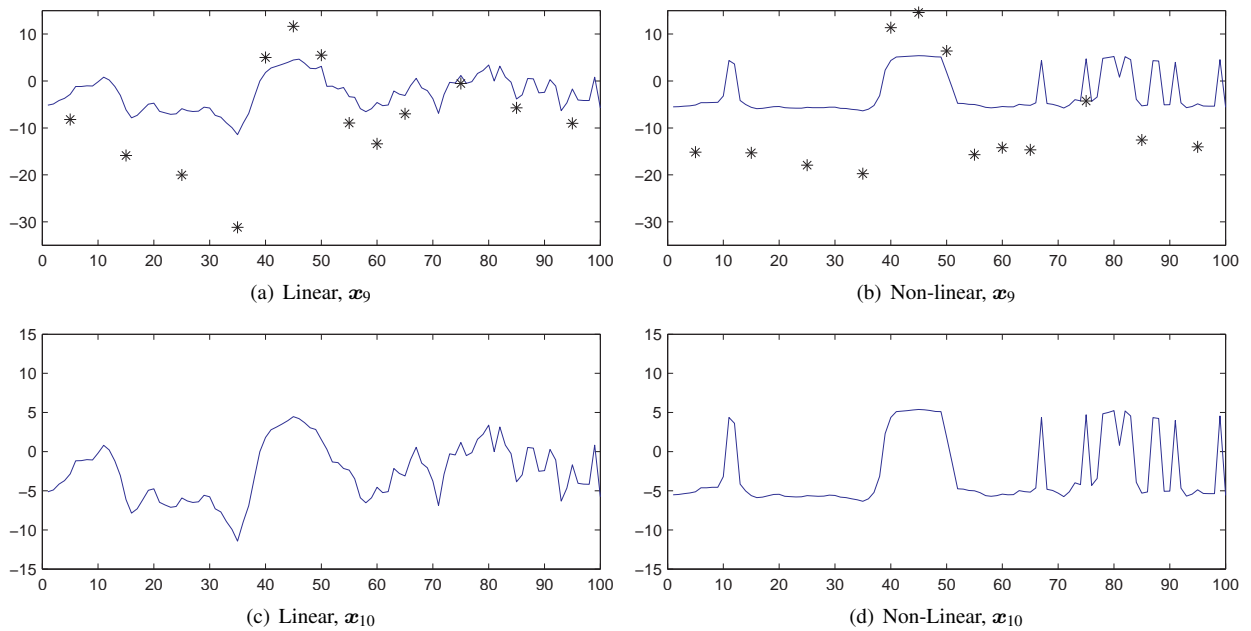
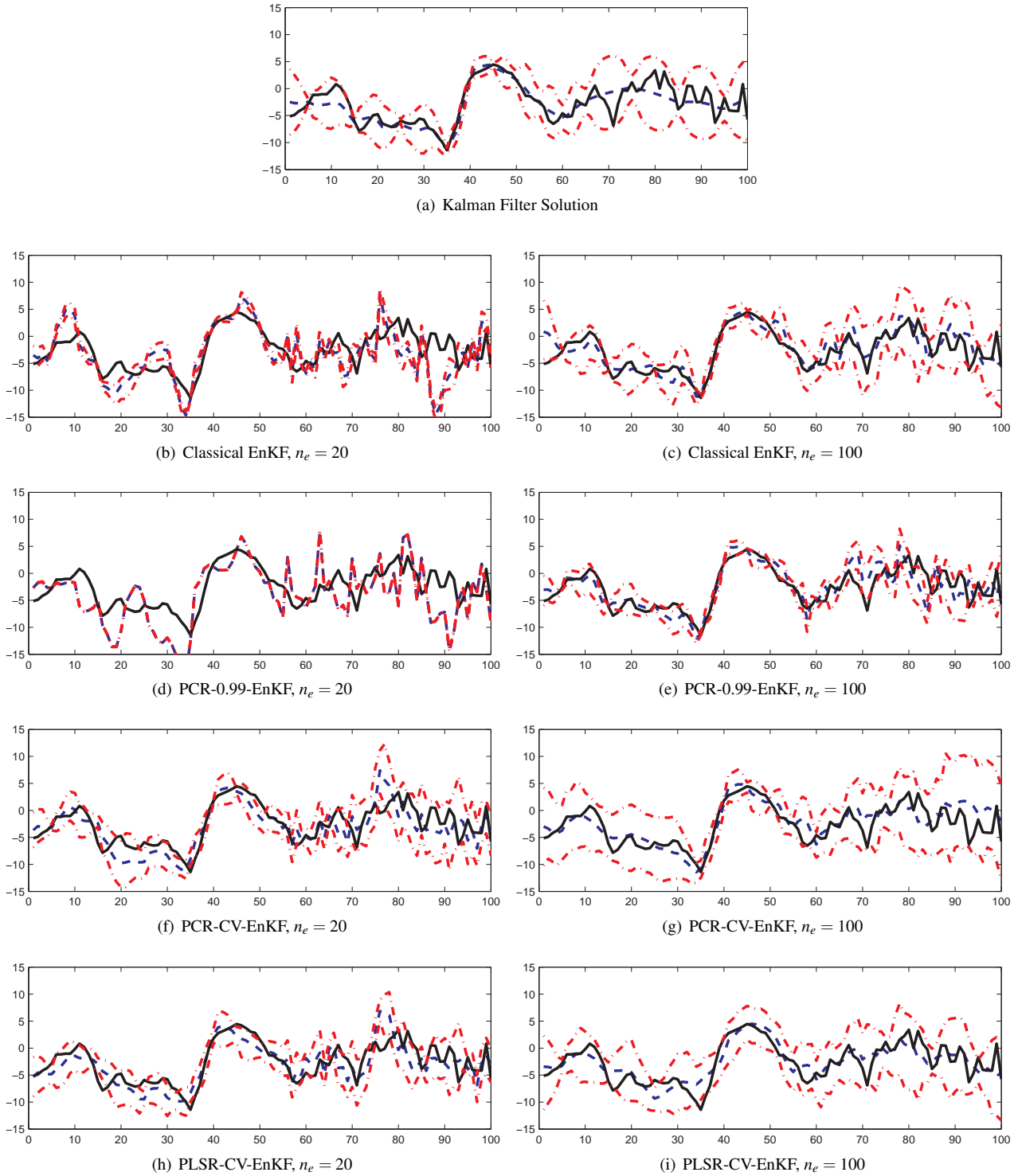
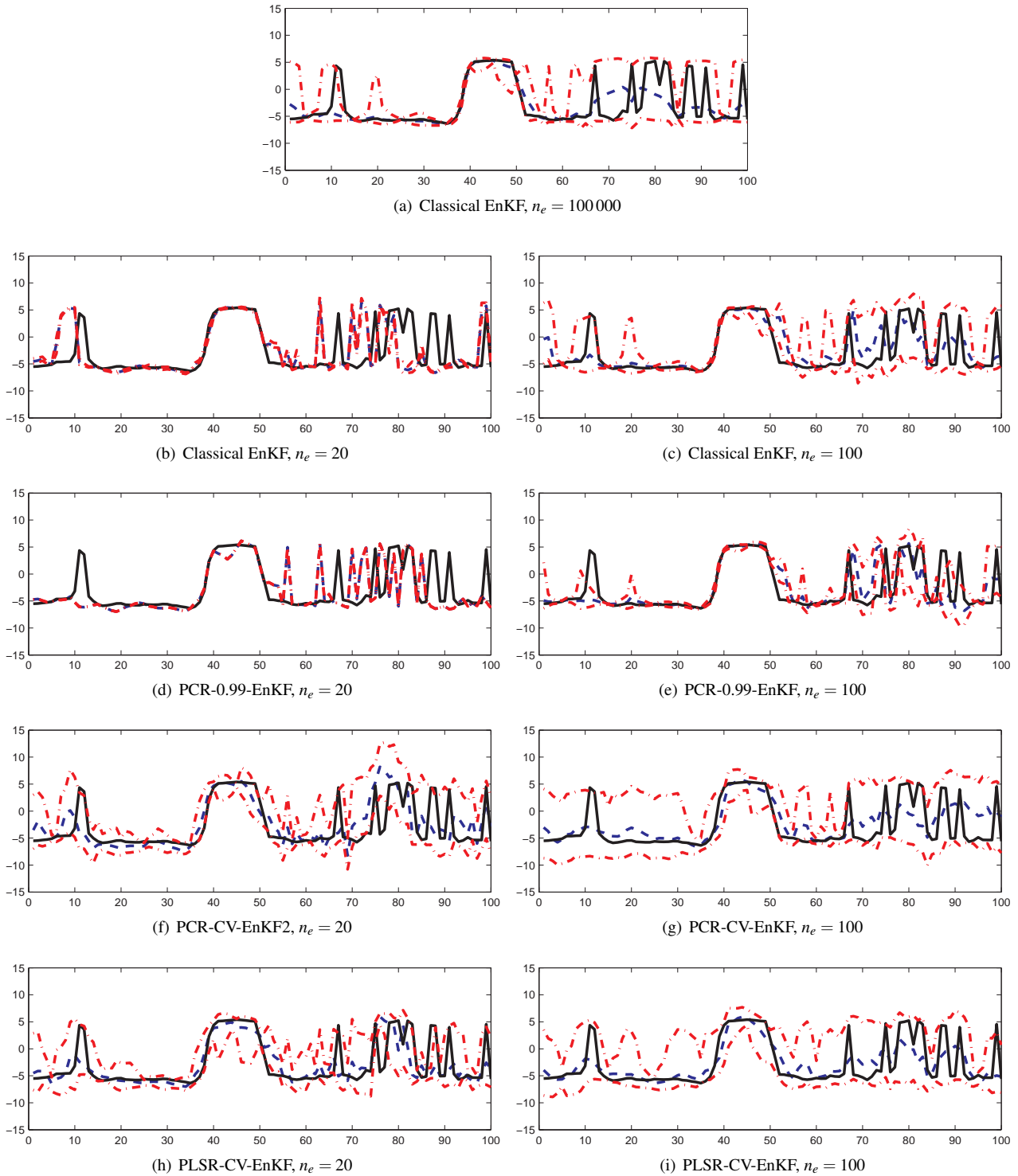


Fig. 7—Reference  $x_9^{\text{True}}$  and  $x_{10}^{\text{True}}$  (line) together with the observed data  $d_9$  (stars) for the linear and non-linear models considered.



**Fig. 8**—Results obtained when running four different EnKF updating schemes on the linear case with two different ensemble sizes. The figure displays the reference  $x_{10}^{\text{true}}$  (solid), the ensemble mean (dashed, blue) and the estimated 95% confidence bounds of the prediction interval (dashed-dotted, red).



**Fig. 9**—Results obtained when running four different EnKF updating schemes on the non-linear case with two different ensemble sizes. The figure displays the reference  $x_{10}^{\text{True}}$  (solid), the ensemble mean (dashed, blue) and the estimated 95% confidence bounds of the prediction interval (dashed-dotted, red).

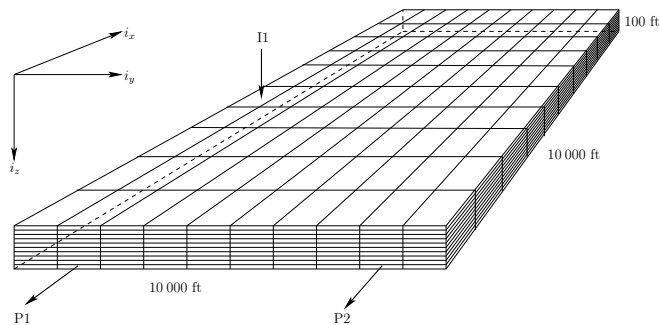


Fig. 10—Description of the synthetic reservoir model discretised into (10 x 10 x 15) grid blocks. The inward pointing arrow indicate the location of the injection well, while the outward pointing arrows indicate the locations of the producer wells.

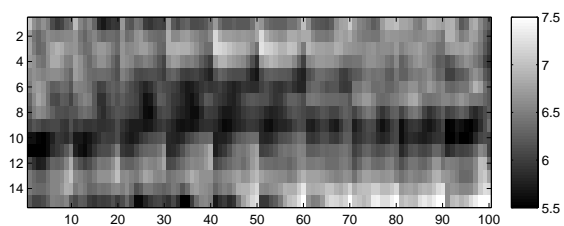


Fig. 11—Image plot of the 100 vertical traces for the reference In-permeability used in the synthetic reservoir example.

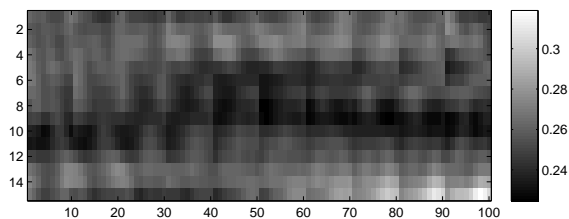


Fig. 12—Image plot of the 100 vertical traces for the reference porosity used in the synthetic reservoir example.

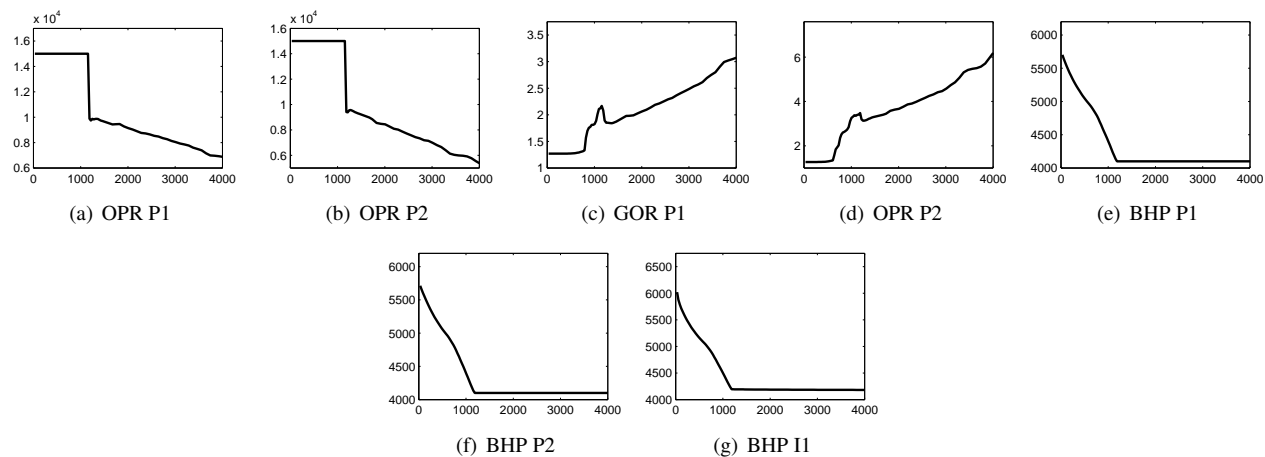
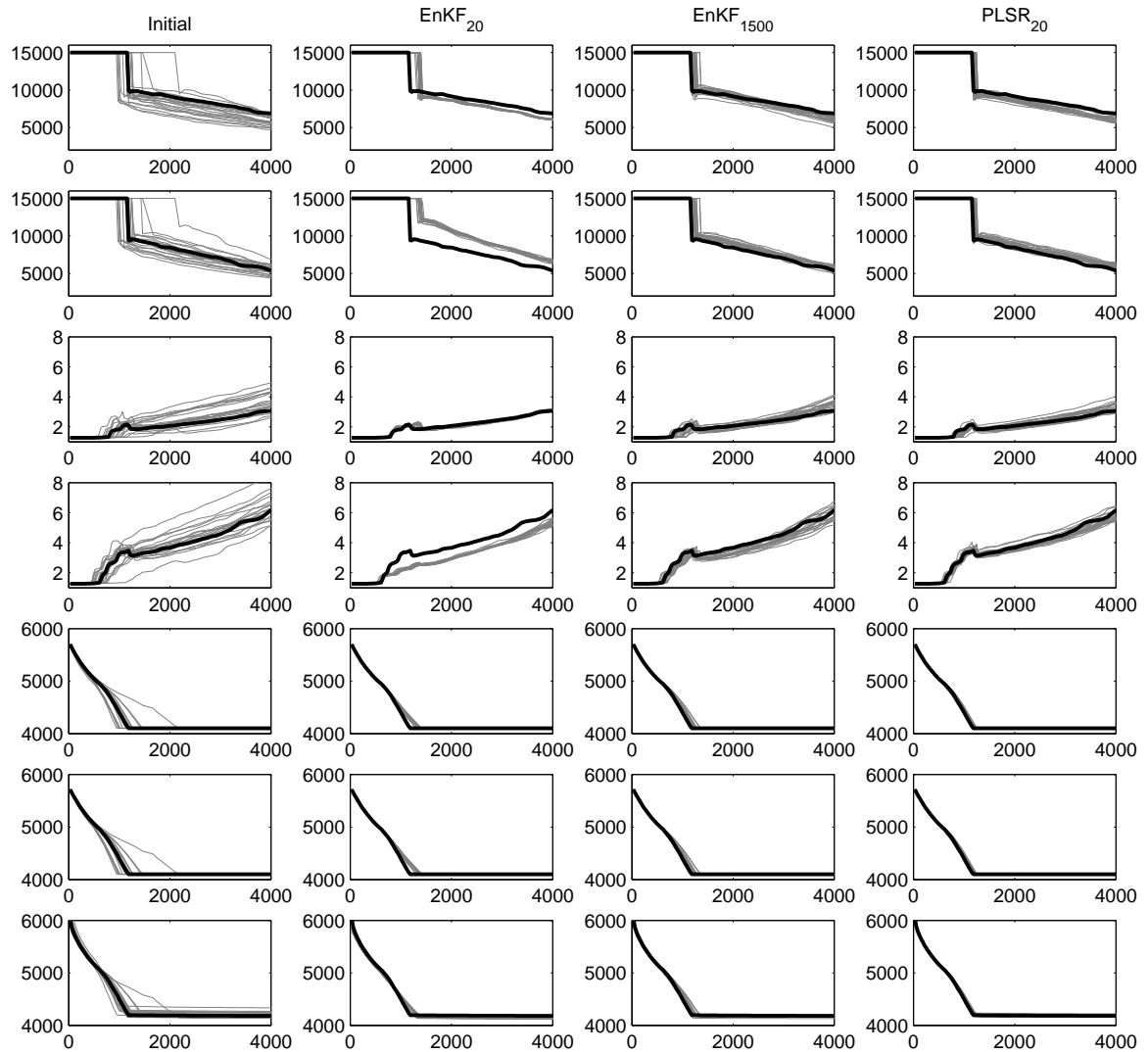


Fig. 13—Reference production data.



**Fig. 14—Forecasted production data based on reruns of the  $n_e = 20$  first ensemble members from timestep zero compared with the reference production data (thick line). The figure shows from left to right the prediction based on the initial ensemble  $x_0^c$ , the updated ensemble  $x_{19}^c$  based on the classical EnKF with  $n_e = 20$  (EnKF<sub>20</sub>),  $n_e = 1\,500$  (EnKF<sub>1500</sub>), and the updated ensemble  $x_{19}^c$  based on the PLSR-CV-EnKF updating scheme with  $n_e = 20$  (PLSR<sub>20</sub>). The production properties considered are from top to bottom OPR P1, OPR P2, GOR P1, GOR P2, BHP P1, BHP P2 and BHP I1.**

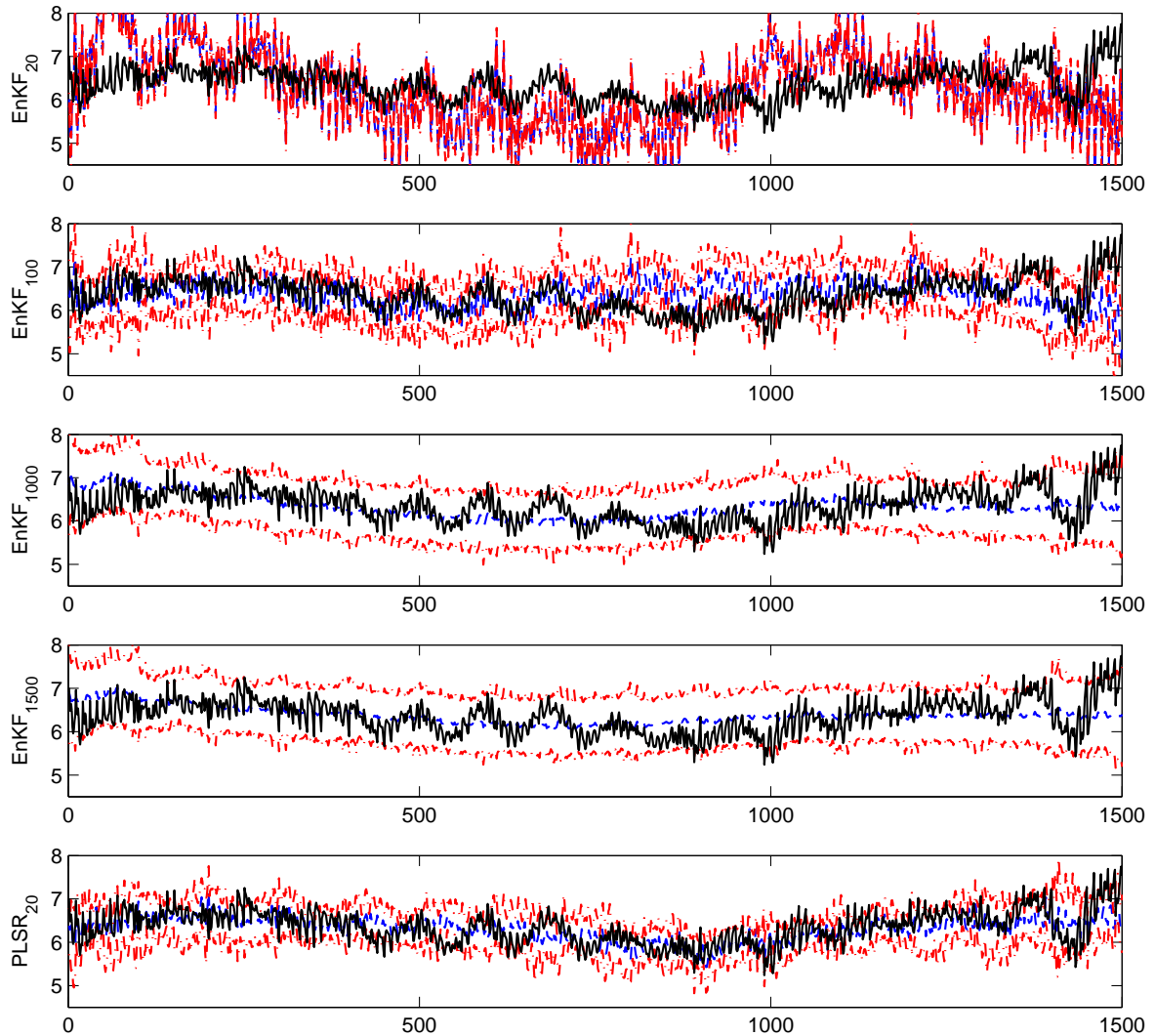


Fig. 15—Updated In-permeability values based on the classical EnKF updating scheme (EnKF) using four different ensemble sizes and the PLSR-CV-EnKF updating scheme (PLSR). The figure displays the reference In-permeability (solid, black), the ensemble mean (dashed, blue) and the estimated 95% confidence bounds of the prediction interval (dashed-dotted, red) obtained based on the updated ensemble members at timestep 19. The subscript denotes the ensemble size used.

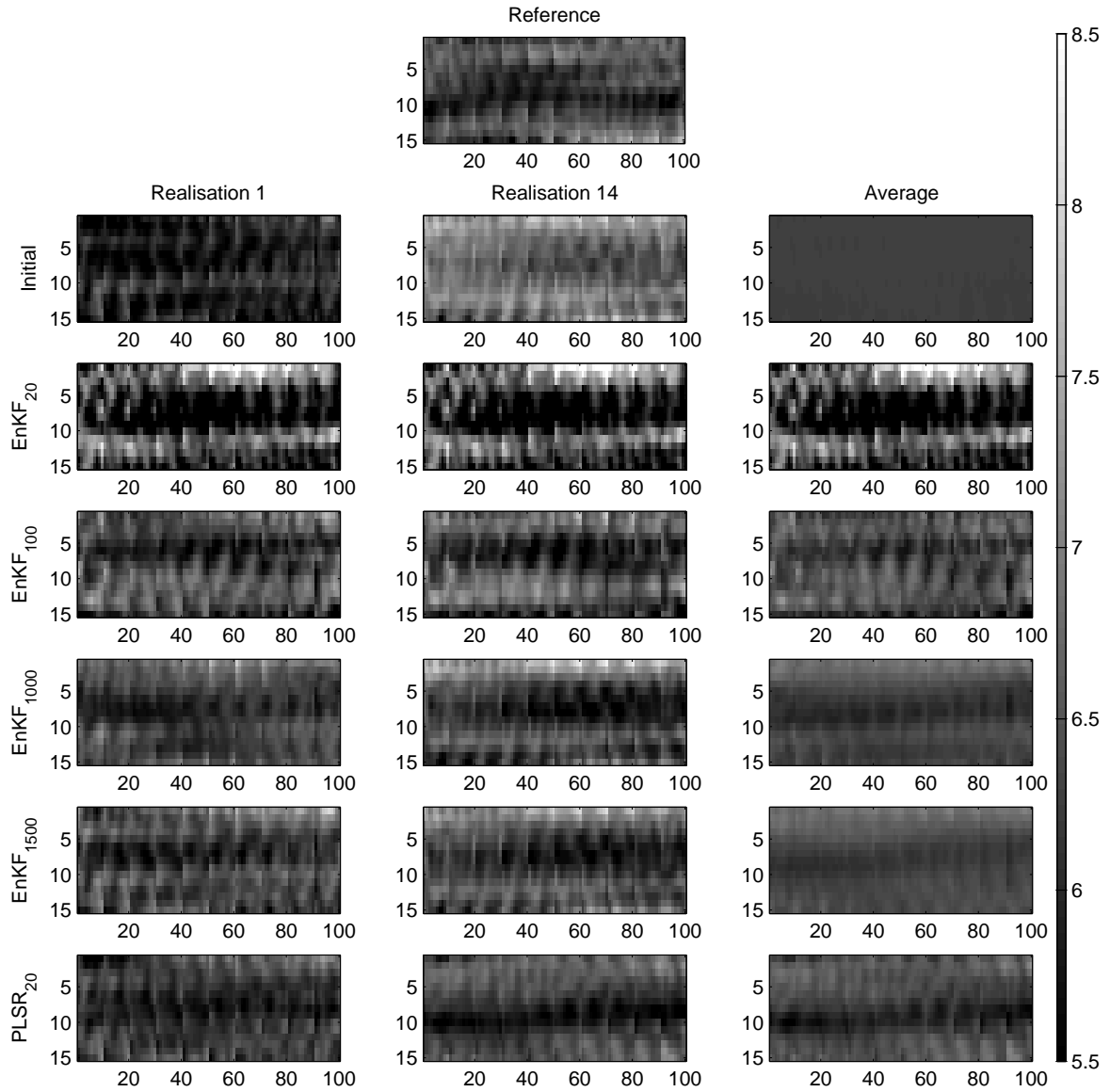


Fig. 16—Two realisations and the estimated ensemble mean based on the first 20 ensemble members for the initial ensemble, for different ensemble sizes using the classical EnKF updating scheme (EnKF) and the PLSR-CV-EnKF scheme (PLSR). The subscript denotes the ensemble size. The reference In-permeability is shown at the top.

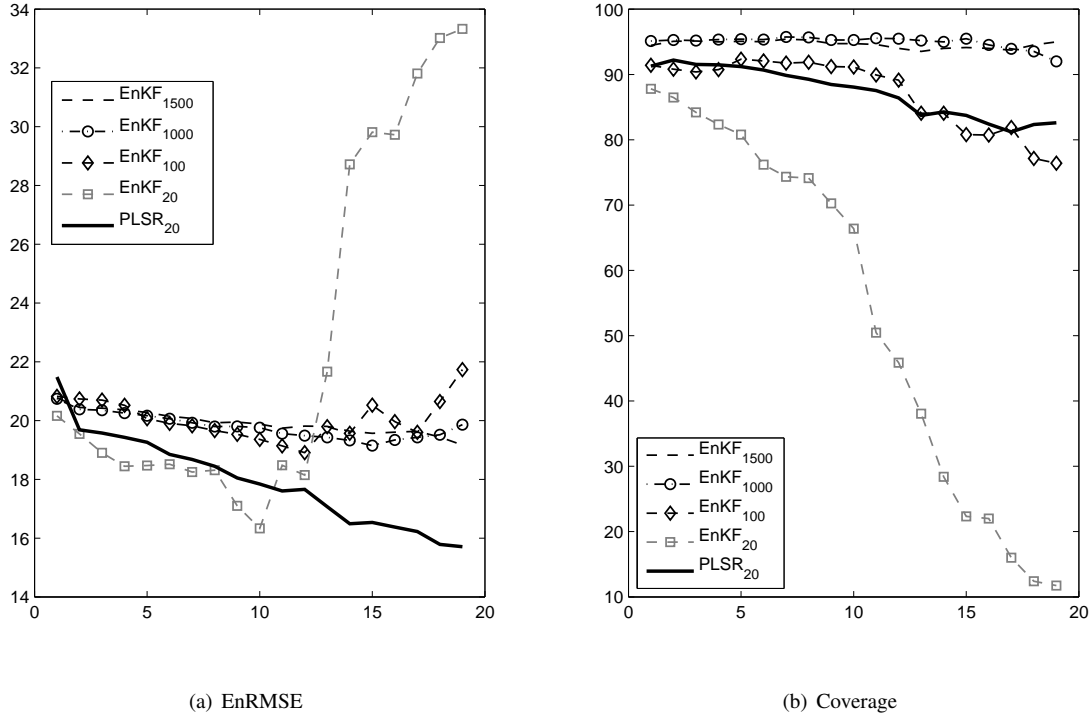


Fig. 17—Root Mean Squared Error of the forecasted In-permeability ensemble members and the reference (EnRMSE), and coverage of the reference solution in the estimated prediction intervals as a function of timesteps  $k$ . Here EnKF corresponds to the classical EnKF updating scheme, while PLSR corresponds to the PLSR-CV-EnKF scheme. The subscript denotes the ensemble size used.

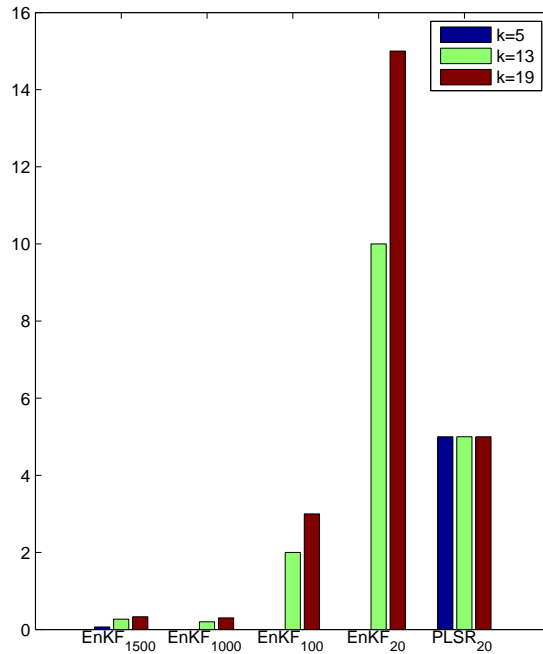


Fig. 18—Estimated relative loss in rank (%) for the updated In-permeability ensemble at three different timesteps. The notation EnKF corresponds to the classical EnKF updating scheme, while PLSR corresponds to the PLSR-CV-EnKF scheme. The subscript denotes the ensemble size used.