

Article

Data Assimilation in Spatio-Temporal Models with Non-Gaussian Initial States—The Selection Ensemble Kalman Model

Maxime Conjard ^{*,†}  and Henning Omre [†]Department of Mathematical Sciences, Norwegian University of Science and Technology,
7491 Trondheim, Norway; henning.omre@ntnu.no

* Correspondence: maxime.conjard@ntnu.no

† These authors contributed equally to this work.

Received: 6 July 2020; Accepted: 17 August 2020; Published: 19 August 2020



Featured Application: Porosity/permeability inversion in petroleum engineering; Log-conductivity inversion with heads and tracer data in hydrogeology; Source contribution identification in air pollution monitoring.

Abstract: Assimilation of spatio-temporal data poses a challenge when allowing non-Gaussian features in the prior distribution. It becomes even more complex with nonlinear forward and likelihood models. The ensemble Kalman model and its many variants have proven resilient when handling nonlinearity. However, owing to the linearized updates, conserving the non-Gaussian features in the posterior distribution remains an issue. When the prior model is chosen in the class of selection-Gaussian distributions, the selection Ensemble Kalman model provides an approach that conserves non-Gaussianity in the posterior distribution. The synthetic case study features the prediction of a parameter field and the inversion of an initial state for the diffusion equation. By using the selection Kalman model, it is possible to represent multimodality in the posterior model while offering a 20 to 30% reduction in root mean square error relative to the traditional ensemble Kalman model.

Keywords: data assimilation; EnKF; multimodality

1. Introduction

Data assimilation of spatio-temporal models is a challenge in many fields of study, including, but not limited to, air pollution mapping, weather forecast, petroleum engineering, and ground water flow assessment. Over the years, methods have been developed to handle increasingly complex problems. It started with the Kalman filter as presented in the seminal publication [1]. The Kalman filter is based on a Gaussian initial model and Gauss-linear forward and observation models. It defined the foundation for data assimilation and is still used in many assimilation studies. The extended Kalman filter (EKF) [2] appeared as a natural methodological extension that allowed for nonlinearity in the Kalman filter framework by linearization. The ensemble Kalman filter (EnKF) [3,4] defined a Monte Carlo approach to the filter and it became popular as it allowed for nonlinearity in the forward and observation models without having to evaluate analytical gradients. The EnKF and its variants have proven to be efficient in solving high-dimensional and nonlinear problems, see [5,6]. In the EnKF, the initial ensemble members represent the initial state which may not have an analytical expression. The forward model then propagates the ensemble members forward in time. Pseudo observations are generated using the observation model. The conditioning of each ensemble member is made with the Kalman weights estimated from the ensemble to give the best linear update. In cases where the initial

model is non-Gaussian, the distribution of the variable of interest conditioned on the data will tend toward Gaussianity as observations are assimilated due to the linear assimilation rule.

Non-Gaussian initial distributions may be conserved by using a univariate transform into Gaussian marginals while assuming multi-Gaussianity in the transformed space. A univariate back transform is then used to return to the original space. This approach has a long history in traditional statistics, geostatistics, and more recently in ensemble methods for data assimilation, which is referred to as copulas [7], normal score transform [8], and Gaussian anamorphosis [9], respectively. The latter has shown to improve the performance of the EnKF in many applications [10,11]. There are however some unresolved issues since Gaussian anamorphosis transforms the marginal distributions rather than the full distribution, and the effect on the resulting variables interdependence is uncertain.

The Ensemble Randomized Maximum Likelihood Filter (EnRML) [12] and its close relative the Iterative EnKF (IEnKF) [6] are primarily used to handle nonlinearities in the forward and observation models, but they will also retain certain non-Gaussian features in the filtering distribution. These filters require gradient evaluations to execute the update which can be complicated even if the adjoint state method is used. One alternative is to evaluate the gradient using the ensemble itself [13], but this approach introduces an approximation with unclear consequences, particularly in models with multimodal marginals.

Multimodality in the prior model can be represented using categorical auxiliary variables to construct Gaussian mixture prior models [14–16]. In a spatial setting, these models appear as a combination of Gaussian random fields whose parameters depend on the value taken by the categorical variable, but in order to retain spatial dependence, the categorical variable must also have a spatial dependence. This indicator spatial variable can be modeled as a Markov [17] or truncated pluri-Gaussian [18] random field. For both of these models, there are challenges related to temporal data assimilation, although some encouraging examples have been developed [19].

We define and study an alternative prior model, the selection-Gaussian random field [20,21], which may represent multimodality, skewness, and peakedness. This random field model is conjugate with respect to Gauss-linear forward and observation models, similarly to the Gaussian random field model. The posterior distribution is therefore analytically tractable under these assumptions [22]. For general forward and observation models, ensemble based algorithms along the lines of the EnKF can be designed. Such selection ensemble Kalman algorithms are the focus of this study, and they are evaluated on a couple of examples.

In Section 2, we introduce the selection ensemble Kalman model. It provides a framework for the use of the selection-Gaussian distribution as a prior in data assimilation. This framework is then used for ensemble filtering and smoothing through the selection EnKF (SEnKF) and the selection EnKS (SEnKS) algorithms. In Section 3, a synthetic case study of the diffusion equation, with two distinct test cases, showcases the ability of the proposed approaches to assess a parameter field and the initial state of a dynamic field. Results from the SEnKF and the SEnKS are compared to that of the traditional EnKF and the EnKS, respectively. In Section 4, potential shortcomings are discussed and the results are put into perspective with respect to applicability in more realistic applications. In Section 5, conclusions are presented.

In this paper, $f(\mathbf{y})$ denotes the probability density function (pdf) of a random variable \mathbf{y} , $\varphi_n(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the pdf of the Gaussian n -vector \mathbf{y} with expectation n -vector $\boldsymbol{\mu}$ and covariance $(n \times n)$ -matrix $\boldsymbol{\Sigma}$. Furthermore, $\Phi_n(A; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the probability of the aforementioned Gaussian n -vector \mathbf{y} to be in $A \subset \mathbb{R}^n$. We also use \mathbf{i}_n to denote the all-ones n -vector, I_n to denote the identity $(n \times n)$ -matrix and $\mathbf{1}(S)$ to denote the indicator function that equals 1 when S is true and 0 otherwise. We consider log-diffusivity to be an adimensional quantity and it will therefore not be given a unit.

2. Materials and Methods

Consider the unknown temporal n -vector \mathbf{r}_t for $t \in \mathcal{T}_r : \{0, 1, \dots, T, T + 1\}$. Let $\mathbf{r} = \{\mathbf{r}_0, \mathbf{r}_1, \dots, \mathbf{r}_T, \mathbf{r}_{T+1}\}$ denote the variable of interest and let $\mathbf{r}_{i:j}$ denote $\{\mathbf{r}_i, \mathbf{r}_{i+1}, \dots, \mathbf{r}_j\}, \forall (i, j) \in \mathcal{T}_r^2, i \leq j$.

Assume that the temporal m -vectors of observations \mathbf{d}_t for $t \in \mathcal{T}_d : \{0, 1, \dots, T\}$ are available, and define $\mathbf{d} = \{\mathbf{d}_0, \mathbf{d}_1, \dots, \mathbf{d}_T\}$ and $\mathbf{d}_{i:j} = (\mathbf{d}_i, \dots, \mathbf{d}_j)$ accordingly. The model specified hereafter defines a hidden Markov (HM) model [23] as displayed in Figure 1.

Prior model: The prior model on \mathbf{r} consists of an initial and a forward model,

$$f(\mathbf{r}) = f(\mathbf{r}_0)f(\mathbf{r}_{1:T+1}|\mathbf{r}_0), \tag{1}$$

where $f(\mathbf{r}_0)$ is the pdf of the initial state and $f(\mathbf{r}_{1:T+1}|\mathbf{r}_0)$ defines the forward model.

(a) *Initial distribution:* The distribution for the initial state $f(\mathbf{r}_0)$ is assumed to be in the class of selection-Gaussian distributions [20,21]. Consider a Gaussian $(n + n)$ -vector $[\tilde{\mathbf{r}}, \mathbf{v}]$,

$$\begin{bmatrix} \tilde{\mathbf{r}} \\ \mathbf{v} \end{bmatrix} \sim \varphi_{2n} \left(\begin{bmatrix} \tilde{\mathbf{r}} \\ \mathbf{v} \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_{\tilde{\mathbf{r}}} \\ \boldsymbol{\mu}_{\mathbf{v}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\tilde{\mathbf{r}}} & \boldsymbol{\Sigma}_{\tilde{\mathbf{r}}}\boldsymbol{\Gamma}_{\mathbf{v}|\tilde{\mathbf{r}}}^T \\ \boldsymbol{\Gamma}_{\mathbf{v}|\tilde{\mathbf{r}}}\boldsymbol{\Sigma}_{\tilde{\mathbf{r}}} & \boldsymbol{\Sigma}_{\mathbf{v}} \end{bmatrix} \right), \tag{2}$$

with n -vectors $\boldsymbol{\mu}_{\tilde{\mathbf{r}}}$ and $\boldsymbol{\mu}_{\mathbf{v}}$, $(n \times n)$ -matrix $\boldsymbol{\Gamma}_{\mathbf{v}|\tilde{\mathbf{r}}}$, and where $\boldsymbol{\Sigma}_{\tilde{\mathbf{r}}}$, $\boldsymbol{\Sigma}_{\mathbf{v}}$, and $\boldsymbol{\Sigma}_{\mathbf{v}|\tilde{\mathbf{r}}}$ are all three covariance $(n \times n)$ -matrices with $\boldsymbol{\Sigma}_{\mathbf{v}} = \boldsymbol{\Gamma}_{\mathbf{v}|\tilde{\mathbf{r}}}\boldsymbol{\Sigma}_{\tilde{\mathbf{r}}}\boldsymbol{\Gamma}_{\mathbf{v}|\tilde{\mathbf{r}}}^T + \boldsymbol{\Sigma}_{\mathbf{v}|\tilde{\mathbf{r}}}$. Define a selection set $A \subset \mathbb{R}^n$ of dimension n and let $\mathbf{r}_0 = [\tilde{\mathbf{r}}|\mathbf{v} \in A]$; then, \mathbf{r}_0 is in the class of selection-Gaussian distribution and its pdf is,

$$\begin{aligned} f(\mathbf{r}_0) &= [\Phi_n(A; \boldsymbol{\mu}_{\mathbf{v}}, \boldsymbol{\Sigma}_{\mathbf{v}})]^{-1} \\ &\times \Phi_n(A; \boldsymbol{\mu}_{\mathbf{v}} + \boldsymbol{\Gamma}_{\mathbf{v}|\tilde{\mathbf{r}}}(\mathbf{r}_0 - \boldsymbol{\mu}_{\tilde{\mathbf{r}}}), \boldsymbol{\Sigma}_{\mathbf{v}|\tilde{\mathbf{r}}}) \times \varphi_n(\mathbf{r}_0; \boldsymbol{\mu}_{\tilde{\mathbf{r}}}, \boldsymbol{\Sigma}_{\tilde{\mathbf{r}}}). \end{aligned} \tag{3}$$

Note that the class of Gaussian distributions constitutes a subset of the class of selection-Gaussian distributions with $\boldsymbol{\Gamma}_{\mathbf{v}|\tilde{\mathbf{r}}} = \mathbf{0} \times \mathbf{I}_n$. The dependence in $[\tilde{\mathbf{r}}, \mathbf{v}]$ represented by $\boldsymbol{\Gamma}_{\mathbf{v}|\tilde{\mathbf{r}}}$ and the selection subset A are crucial user-defined parameters with the latter being temporally constant. The selection-Gaussian model may represent multimodal, skewed, and/or peaked marginal distributions, see [21]. In this study, the initial distribution is defined to be a discretized stationary selection-Gaussian random field with parametrization,

$$\begin{aligned} \boldsymbol{\mu}_{\tilde{\mathbf{r}}} &= \mu_{\tilde{\mathbf{r}}}\mathbf{i}_n \\ \boldsymbol{\mu}_{\mathbf{v}} &= \mu_{\mathbf{v}}\mathbf{i}_n \\ \boldsymbol{\Sigma}_{\tilde{\mathbf{r}}} &= \sigma_{\tilde{\mathbf{r}}}^2\boldsymbol{\Sigma}_{\tilde{\mathbf{r}}}^\rho \\ \boldsymbol{\Sigma}_{\mathbf{v}} &= \gamma^2\boldsymbol{\Sigma}_{\tilde{\mathbf{r}}}^\rho + (1 - \gamma^2)\mathbf{I}_n \\ \boldsymbol{\Gamma}_{\mathbf{v}|\tilde{\mathbf{r}}} &= \gamma\sigma_{\tilde{\mathbf{r}}}^{-1}\mathbf{I}_n. \end{aligned} \tag{4}$$

For a given spatial correlation $(n \times n)$ -matrix $\boldsymbol{\Sigma}_{\tilde{\mathbf{r}}}^\rho$, a stationary selection-Gaussian random field is fully parametrized by $\Theta^{SG} = (\mu_{\tilde{\mathbf{r}}}, \mu_{\mathbf{v}}, \sigma_{\tilde{\mathbf{r}}}, \boldsymbol{\Sigma}_{\tilde{\mathbf{r}}}^\rho, \gamma, A)$. Similarly, a stationary Gaussian random field is parametrized by $\Theta^G = (\mu_r, \sigma_r, \boldsymbol{\Sigma}_r^\rho)$.

(b) *Forward model:* The forward model given the initial state $[\mathbf{r}_{1:T+1}|\mathbf{r}_0]$ is defined as

$$f(\mathbf{r}_{1:T+1}|\mathbf{r}_0) = \prod_{t=0}^T f(\mathbf{r}_{t+1}|\mathbf{r}_t), \tag{5}$$

with

$$[\mathbf{r}_{t+1}|\mathbf{r}_t] = \omega_t(\mathbf{r}_t, \boldsymbol{\epsilon}_t^r) \sim f(\mathbf{r}_{t+1}|\mathbf{r}_t), \tag{6}$$

where $\omega_t(\cdot, \cdot) \in \mathbb{R}^n$ is the forward model with random n -vector $\boldsymbol{\epsilon}_t^r$, independent and identically distributed (iid) for each t . This forward model may be nonlinear, but, since it only involves the variable at the previous time step \mathbf{r}_t , it defines a first-order Markov chain. Note that $f(\mathbf{r}_{t+1}|\mathbf{r}_t)$ cannot generally be written in closed form.

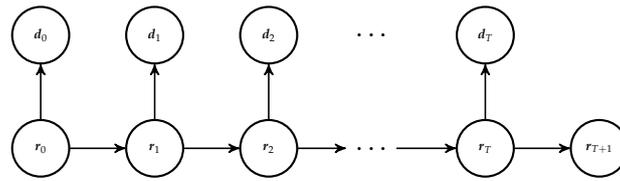


Figure 1. Graph of the hidden Markov model.

Likelihood model: The likelihood model for $[d|r]$ is defined as conditional independent with single-site response,

$$f(d|r) = \prod_{t=0}^T f(d_t|r_t), \tag{7}$$

with

$$[d_t|r_t] = \psi_t(r_t, \epsilon_t^d) \sim f(d_t|r_t), \tag{8}$$

where $\psi_t(\cdot, \cdot) \in \mathbb{R}^m$ is the likelihood function with random m -vector ϵ_t^d , iid for each t . Note that $f(d_t|r_t)$ cannot generally be written in closed form.

Posterior model: The posterior model for the HM model in Figure 1 is given by

$$\begin{aligned} [r|d] \sim f(r|d) &= \text{const} \times f(d|r)f(r) \\ &= \text{const} \times f(d_0|r_0)f(r_0) \prod_{t=1}^T f(d_t|r_t)f(r_t|r_{t-1})f(r_{T+1}|r_T) \\ &= f(r_0|d) \prod_{t=1}^T f(r_t|r_{t-1}, d_{t:T})f(r_{T+1}|r_T), \end{aligned} \tag{9}$$

and is also a Markov chain, see [23,24]. This model is denoted the selection ensemble Kalman model. If the forward and likelihood models are Gauss-linear, the posterior model is also selection-Gaussian and analytically tractable, see [22]. When the forward and/or likelihood models are nonlinear, however, approximate or sampling based assessment of the posterior model must be made. For this purpose, we introduce the selection ensemble Kalman filter (SEnKF) and smoother (SEnKS) in the spirit of the traditional ensemble Kalman model [3].

The traditional EnKF algorithm aims at assessing the forecast pdf $f(r_{T+1}|d_{0:T})$, and it is justified by general HM model recursions, see [23]. The algorithm is initiated by

$$[r_1|d_0] \sim f(r_1|d_0) = \int f(r_1|r_0)[f(d_0)]^{-1}f(d_0|r_0)f(r_0)dr_0, \tag{10}$$

and utilizes the recursion for $t = 1, \dots, T$,

$$[r_{t+1}|d_{0:t}] \sim f(r_{t+1}|d_{0:t}) = \int f(r_{t+1}|r_t)[f(d_t|d_{0:t-1})]^{-1}f(d_t|r_t)f(r_t|d_{0:t-1})dr_t. \tag{11}$$

The expressions are represented by an ensemble of realizations, which in each recursion is conditioned using a linearized approximation with Kalman weights estimated from the ensemble. Thereafter, the ensemble is forwarded to the next time step. The SEnKF introduced in this study relies on the same relations as above, but it operates on the augmented $(n + n)$ -vector $[\tilde{r}, v]$, see Equation (2). Hence, the forward model is defined as

$$\begin{bmatrix} \tilde{r}_{t+1} \\ v_{t+1} \end{bmatrix} \Big| \begin{bmatrix} \tilde{r}_t \\ v_t \end{bmatrix} = \begin{bmatrix} \omega_t(\tilde{r}_t, \epsilon_t^r) \\ v_t \end{bmatrix}. \tag{12}$$

where the auxiliary n -vector \mathbf{v}_t is temporally constant.

The likelihood model is defined as

$$\begin{bmatrix} \mathbf{d}_t \\ \tilde{\mathbf{r}}_t \end{bmatrix} \Big| \mathbf{v}_t = \psi_t(\tilde{\mathbf{r}}_t, \epsilon_t^d). \tag{13}$$

The SEnKF algorithm provides an ensemble representation of

$$\begin{bmatrix} \tilde{\mathbf{r}}_{T+1} \\ \mathbf{v} \end{bmatrix} \Big| \mathbf{d}_{0:T} \sim f(\tilde{\mathbf{r}}_{T+1}, \mathbf{v} | \mathbf{d}_{0:T}), \tag{14}$$

and, based on this ensemble, empirical sampling based inference, see [21], is used to obtain the forecast of interest:

$$[\mathbf{r}_{T+1} | \mathbf{d}_{0:T}] \sim f(\mathbf{r}_{T+1} | \mathbf{d}_{0:T}) = f(\tilde{\mathbf{r}}_{T+1} | \mathbf{d}_{0:T}, \mathbf{v} \in A). \tag{15}$$

The SEnKF algorithm is specified in Algorithm A1 in Appendix A.

The traditional EnKS algorithm aims at evaluating the interpolation pdf $f(\mathbf{r}_{0:T} | \mathbf{d}_{0:T})$ with corresponding HM model recursions, see [23]. The algorithm is initiated by

$$[\mathbf{r}_0 | \mathbf{d}_0] \sim f(\mathbf{r}_0 | \mathbf{d}_0) = [f(\mathbf{d}_0)]^{-1} f(\mathbf{d}_0 | \mathbf{r}_0) f(\mathbf{r}_0), \tag{16}$$

and the recursions for $t = 1, \dots, T$,

$$\begin{aligned} [\mathbf{r}_{0:t} | \mathbf{d}_{0:t}] &\sim f(\mathbf{r}_{0:t} | \mathbf{d}_{0:t}) \\ &= [f(\mathbf{d}_t | \mathbf{d}_{0:t-1})]^{-1} f(\mathbf{d}_t | \mathbf{r}_{0:t}, \mathbf{d}_{0:t-1}) f(\mathbf{r}_t | \mathbf{r}_{0:t-1}, \mathbf{d}_{0:t-1}) f(\mathbf{r}_{0:t-1} | \mathbf{d}_{0:t-1}) \\ &= [f(\mathbf{d}_t | \mathbf{d}_{0:t-1})]^{-1} f(\mathbf{d}_t | \mathbf{r}_t) f(\mathbf{r}_t | \mathbf{r}_{t-1}) f(\mathbf{r}_{0:t-1} | \mathbf{d}_{0:t-1}). \end{aligned} \tag{17}$$

The expressions are represented by an ensemble of realizations. Forwarding is made on the ensemble and the conditioning is empirically linearized. Note that the dimension of the model increases very fast, one may therefore only store the interpolation pdf $f(\mathbf{r}_s | \mathbf{d}_{0:T})$ at the time point s of interest. The SEnKS introduced in this study relies on the relations defined above and uses an extended $(n + n)$ -vector $[\tilde{\mathbf{r}}, \mathbf{v}]$ as defined in Equation (2). The forward and likelihood models are identical to those defined for the filter. The SEnKS algorithm provides an ensemble representation of

$$\begin{bmatrix} \tilde{\mathbf{r}}_{0:T} \\ \mathbf{v} \end{bmatrix} \Big| \mathbf{d}_{0:T} \sim f(\tilde{\mathbf{r}}_{0:T}, \mathbf{v} | \mathbf{d}_{0:T}), \tag{18}$$

and by using empirical sampling based inference, see [21], the interpolation of interest is assessed,

$$[\mathbf{r}_{0:T} | \mathbf{d}_{0:T}] \sim f(\mathbf{r}_{0:T} | \mathbf{d}_{0:T}) = f(\tilde{\mathbf{r}}_{0:T} | \mathbf{d}_{0:T}, \mathbf{v} \in A). \tag{19}$$

The SEnKS algorithm is specified in Algorithm A2 in the Appendix A. Both algorithms, SEnKF and SEnKS, contain empirically linearized conditioning and asymptotic results, when the ensemble size goes to infinity, and are consistent only for Gauss-linear forward and likelihood models. Under these assumptions, the model is analytically tractable; however, see [22]. In spite of this lack of asymptotic consistency for general HM models, the ensemble Kalman scheme has proven surprisingly reliable for high-dimensional, weakly nonlinear models even with very modest ensemble sizes [25].

3. Results

We consider two test cases to illustrate the relevance of the selection ensemble Kalman algorithms presented in Section 2. The model, common to both test cases, is based on the diffusion equation.

The test cases are designed such that it will be opportune to consider bi-modal initial distributions. In the first test case, we compare the SEnKF to the traditional EnKF with a focus on predicting the diffusivity field that contains a high diffusivity channel. In the second test case, we compare the SEnKS to the traditional EnKS with a focus on evaluating the initial temperature field that is divided into two distinct areas where the initial temperature is substantially higher in one than in the other.

3.1. Model

Consider a discretized spatio-temporal random field, $\{r_t(\mathbf{x}), \mathbf{x} \in \mathcal{L}_r \subset \mathbb{R}^2\}$ where $t \in \mathcal{L}_t : \{0, 1, \dots, T\}$ and $r_t(\cdot) \in \mathbb{R}$ that represents temperature ($^{\circ}\text{C}$). Let a discretized spatial random field, $\{\lambda(\mathbf{x}), \mathbf{x} \in \mathcal{L}_r \subset \mathbb{R}^2\}$; with $\lambda(\cdot) \in \mathbb{R}_{\oplus}$ representing diffusivity ($\text{m}^2 \text{s}^{-1}$). Let \mathbf{x} be the spatial reference on the regular spatial grid \mathcal{L}_r on the domain D , while t is the temporal reference on the regular temporal grid \mathcal{L}_t . The number of spatial grid nodes is $n = 21 \times 21$, and they are placed every 10 cm vertically and horizontally. The discretized temperature field at time t may be represented by the n -vector \mathbf{r}_t and the diffusivity field by the n -vector λ . Both are assumed to be unknown. Note that the Kalman models are defined on the joint variable $[\mathbf{r}_t, \lambda]$.

Assume that, given the initial temperature field, the field evolves according to the diffusion equation:

$$\begin{aligned} \frac{\partial r_t(\mathbf{x})}{\partial t} - \nabla \cdot (\lambda(\mathbf{x}) \nabla r_t(\mathbf{x})) &= q \\ \nabla r_t(\mathbf{x}) \cdot \mathbf{n} &= 0, \end{aligned} \tag{20}$$

with \mathbf{n} the outer normal to the domain and q a source term. The expression in Equation (20) is discretized using finite differences and the forward model is defined as

$$[\mathbf{r}_{t+1} | \mathbf{r}_t, \lambda] = \omega^*(\mathbf{r}_t, \lambda), \tag{21}$$

with $\omega^*(\cdot, \cdot) \in \mathbb{R}^n$. Convergence and stability of the numerical method are easily ensured for the finite difference scheme that is used. The initial temperature field \mathbf{r}_0 is considered unknown in the test cases.

The forward model is assumed to be perfect in the sense that there is no model error. The forward model in Equation (6) then takes the form,

$$\omega([\mathbf{r}_t, \lambda], 0\mathbf{i}_n) = \begin{bmatrix} \omega^*(\mathbf{r}_t, \lambda) \\ \lambda \end{bmatrix}. \tag{22}$$

This forward model is nonlinear due to the product of \mathbf{r}_t and λ in Equation (20). Consequently, the assumption of Gauss-linearity required for both the traditional Kalman model [1] and the selection Kalman model [22] is violated and necessitates ensemble based algorithms.

The observations are acquired in a $m = 5$ location pattern on the spatial grid \mathcal{L}_r at each temporal node in \mathcal{L}_t , providing the set of observations m -vectors $\mathbf{d}_t, t \in \mathcal{L}_t$. The corresponding likelihood model is defined as

$$\begin{aligned} [\mathbf{d}_t | \mathbf{r}_t] &= \psi_t(\mathbf{r}_t, \boldsymbol{\epsilon}_t^d) = \mathbf{H}\mathbf{r}_t + \boldsymbol{\epsilon}_t^d \\ f(\mathbf{d}_t | \mathbf{r}_t) &= \varphi_m(\mathbf{d}_t; \mathbf{H}\mathbf{r}_t, \boldsymbol{\Sigma}_{d|r}), \end{aligned} \tag{23}$$

where the observation $(m \times n)$ -matrix \mathbf{H} is a binary selection matrix, while the centered Gaussian m -vector $\boldsymbol{\epsilon}_t^d$ with the covariance $(m \times m)$ -matrix $\boldsymbol{\Sigma}_{d|r} = \sigma_{d|r}^2 \mathbf{I}_m$, and $\sigma_{d|r}^2 = 0.1$ represents independent observation errors. This likelihood model is in Gauss-linear form.

3.2. Test Case 1: Predicting the Parameter Field

The focus of this test case is to predict the unknown diffusivity field λ based on the observations d . Because diffusivity is constant in time, smoothing and filtering give an identical prediction of the field. However, filtering is preferred because it does not require updating the ensemble at all future times in addition to the previous one, see [26]. The posterior model is evaluated using the SENKF, see Appendix A and the results are compared to those from the traditional EnKF algorithm.

The true diffusivity n -vector λ is displayed in Figure 2. The diffusivity λ is always positive. To ensure that ensemble updates do not lead to negative diffusivity values, we work on $\log(\lambda)$. The figure shows a channel in which the diffusivity is higher than in the rest of the field. The diffusivity field is formally defined as

$$\lambda(x) = \lambda_1 \mathbf{1}(x \in D_1) + \lambda_2 \mathbf{1}(x \in D_2), \tag{24}$$

where $D_1 \subset D$ is the low diffusivity area and $D_2 \subset D$ is the high diffusivity channel. The parameter values are $\lambda_1 = e^{-12} \text{ m}^2 \text{ s}^{-1}$ and $\lambda_2 = e^{-5} \text{ m}^2 \text{ s}^{-1}$. The true temperature field is initially at 20°C and the heat source on the lower border of the high diffusivity channel starts pumping in heat at $T = 0$ at a constant volumetric rate $q = 15 \text{ W m}^{-3}$, see Figure 2. The temporal evolution of the temperature field, shown in Figure 3, is obtained by solving the diffusion equation in Equation (20) for the log-diffusivity field in Figure 2 and the initial temperature field defined above. The temperature observations d , see Figure 4, are then collected from the five locations shown in Figure 2 using the likelihood model defined in Equation (23). The measurements are taken every second from $T = 0$ to $T = 100$. As the heat from the source diffuses mostly along the high diffusivity channel, the observed temperature increases substantially at observation locations within the channel.

The unknown initial field for log-diffusivity $\log(\lambda)$ is assigned a stationary selection-Gaussian random field prior model with parameters $\Theta_\lambda^{SG} = (\mu_\lambda^\lambda, \mu_\nu^\lambda, \sigma_\lambda^\lambda, \Sigma_\lambda^\rho, \gamma^\lambda, A)$, see [21] and Equation (2). The parameter values for the prior model are listed in Table 1.

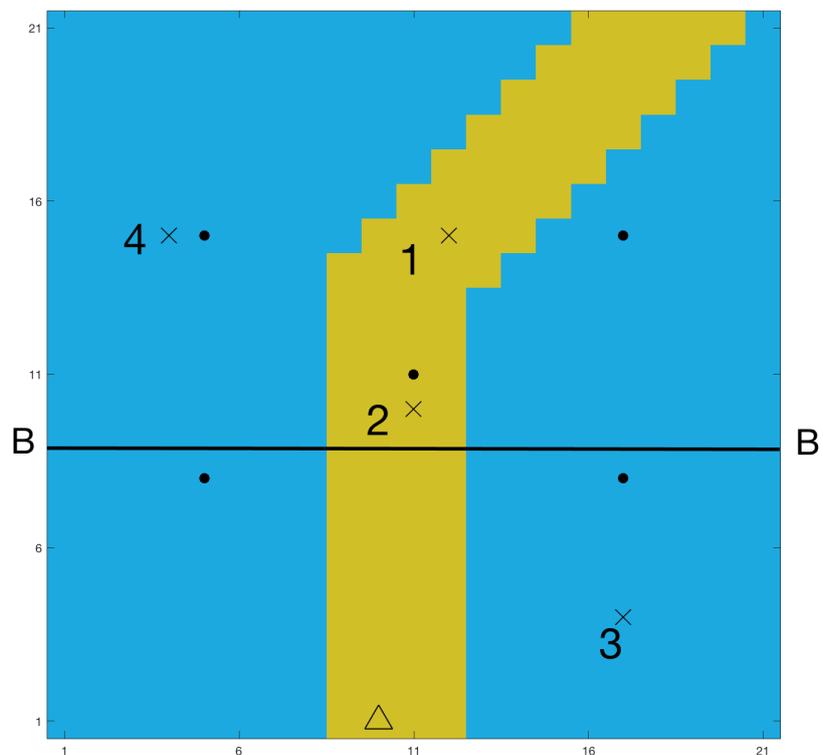


Figure 2. Initial log-diffusivity field with observation locations \cdot , monitoring locations \times , and heat source \triangle .

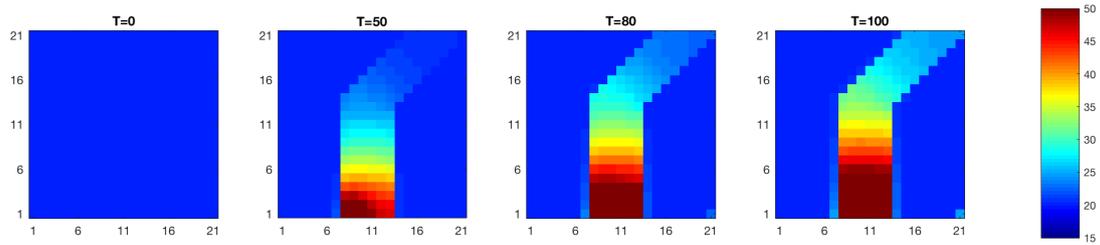


Figure 3. True temperature (°C) field evolution over time.

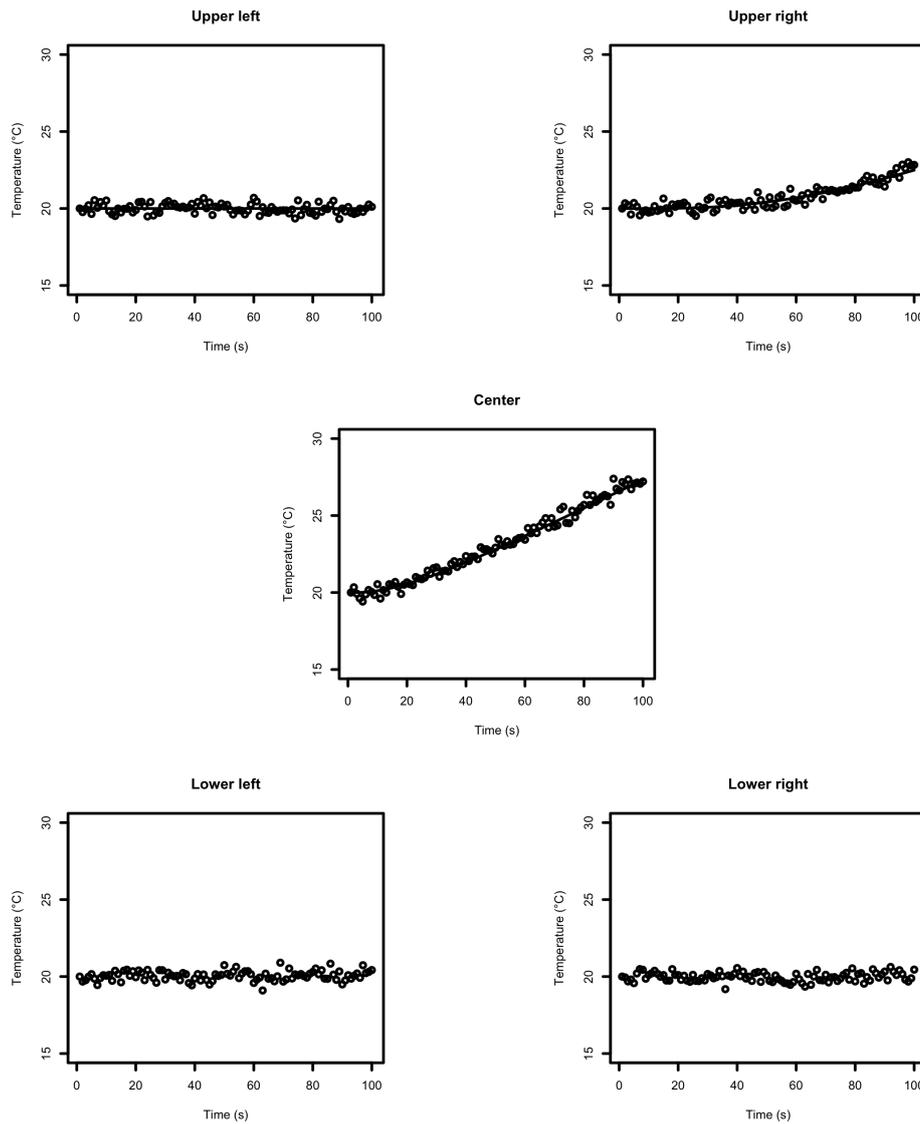


Figure 4. Data collected over time (·) and true temperature evolution (line) at the data collection points.

Table 1. Parameter values for the selection-Gaussian initial distribution for the initial log-diffusivity field.

Parameters	Values
μ_x^λ	-8.5
μ_y^λ	0
σ_x^λ	$\sqrt{1.6}$
γ^λ	0.9
A	$] -\infty, -0.3] \cup [0.5, +\infty[$

The unknown initial temperature field r_0 is assigned a stationary Gaussian random field prior model with parameters $\Theta_r^G = (\mu_r, \sigma_r, \Sigma_r^\rho)$ with expectation and variance levels $\mu_r = 20$ and $\sigma_r^2 = 2$, respectively. The variance level is relatively large as we assume little prior knowledge of the initial temperature field. For both prior models, the spatial correlation ($n \times n$)-matrix Σ_r^ρ is defined by the second order exponential spatial correlation function $\rho(\tau) = \exp(-\tau^2/\delta^2)$; $\delta = 0.15$, with interdistance τ .

Figure 5 contains realizations from the prior model of the log-diffusivity field and their associated spatial histograms. The prior model is specified to be spatially stationary except for boundary effects with bi-modal spatial histograms. The selection set $A \subset \mathbb{R}^n$ for the prior model is chosen to obtain bi-modal marginal distributions with a very dominant mode centered slightly above the value for λ_1 and a very small mode centered slightly below the value for λ_2 . The prior is therefore not centered at the true values. Note that the joint random field $[\log(\lambda), r_0]$ will appear as a bi-variate selection-Gaussian random field, see [21].

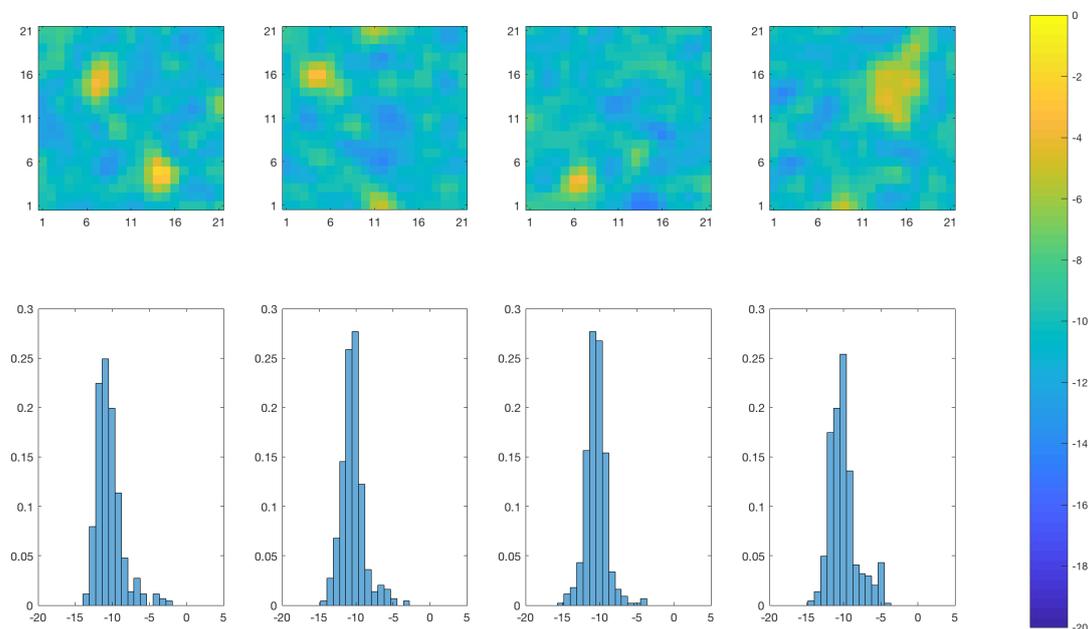


Figure 5. Realizations from the initial selection-Gaussian distribution of the log diffusivity $f(\log(\lambda))$ at time $T = 0$ (upper panels) and associated spatial histogram (lower panels). Lower panels: the horizontal axes represent the log-diffusivity, the vertical axes represent the relative prevalence of each log-diffusivity value for the realization in the panel right above.

The SEnKF operates on the $3n$ -vector $[\log(\tilde{\lambda}), \nu, r_0]$, and therefore we generate an initial ensemble with $n_e = 10,000$ ensemble members that are sampled from the Gaussian $3n$ -vector $[\log(\tilde{\lambda}), \nu, r_0]$ with pdf,

$$\begin{bmatrix} \log(\tilde{\lambda}) \\ \nu \\ r_0 \end{bmatrix} \sim \varphi_{3n} \left(\begin{bmatrix} \log(\tilde{\lambda}) \\ \nu \\ r_0 \end{bmatrix}; \begin{bmatrix} \mu_{\tilde{\lambda}}^\lambda \\ \mu_\nu^\lambda \\ \mu_r \end{bmatrix}, \begin{bmatrix} \sigma_{\tilde{\lambda}}^{\lambda^2} \Sigma_{\tilde{\lambda}}^\rho & \gamma^\lambda \sigma_{\tilde{\lambda}}^\lambda \Sigma_{\tilde{\lambda}}^{\rho T} & 0 \\ \gamma^\lambda \sigma_{\tilde{\lambda}}^\lambda \Sigma_{\tilde{\lambda}}^\rho & \gamma^{\lambda^2} \Sigma_{\tilde{\lambda}}^\rho + (1 - \gamma^{\lambda^2}) \mathbf{I}_n & 0 \\ 0 & 0 & \sigma_r^2 \Sigma_r^\rho \end{bmatrix} \right). \quad (25)$$

The EnKF operates on the $2n$ -vector $[\log(\lambda), r_0]$, and therefore we generate an initial ensemble with $n_e = 10,000$ ensemble members that are sampled from the selection-Gaussian distribution $f(\log(\lambda), r_0)$. The variables $\log(\lambda)$ and r_0 are independent, so we generate them independently: 10,000 samples from the selection-Gaussian n -vector $\log(\lambda)$ with parameters Θ_λ^{SG} and 10,000 samples from the Gaussian n -vector r_0 with parameters Θ_r^G . It is important to understand that both ensemble algorithms are initiated with an ensemble from an identical selection-Gaussian random field prior

model for $[\log(\lambda), r_0]$ at $T = 0$, which reflects the bi-modality of the prior model. Due to the size of the ensemble relative to the dimension of the problem, we are using neither localization nor inflation in the algorithms.

To illustrate the differences between the SEnKF and the EnKF, we present the following results for both algorithms:

1. The marginal posterior distributions $f(\log(\lambda_i)|d_{0:T})$ of the log-diffusivity field at four monitoring locations denoted 1, 2, 3, 4 on Figure 2, at time $T = 0, 50, 80, 100$.
2. The marginal maximum a posteriori (MMAP) prediction of the log-diffusivity field at time at time $T = 0, 50, 80, 100$.
3. Realizations from the posterior distribution $f(\log(\lambda)|d_{0:T})$ at time $T = 100$.
4. The root mean square errors (RMSE) of the MMAP prediction of the log-diffusivity field relative to the true log-diffusivity field at time $T = 100$.

Figures 6 and 7 show the marginal posterior pdfs $f(\log(\lambda_i)|d_{0:T})$ at the four monitoring locations at time $T = 0, 50, 80, 100$ for the SEnKF and EnKF algorithms, respectively. Monitoring locations 1 and 2 are placed within the high diffusivity area while the two other locations are placed far into the low diffusivity area. At $T = 0$, all pdfs are identical, in all locations due to the stationary prior model and for both algorithms due to identical prior models. The SEnKF results appear to preserve bi-modality as observations are assimilated. As more data are made available, the high value mode increases at monitoring locations 1 and 2 that are inside the high diffusivity area. The low value mode remains dominant at monitoring locations within the low diffusivity areas. These results reflect expected behaviors. The traditional EnKF results are significantly different since the bi-modality of the marginal pdfs disappears already at $T = 50$. The marginal pdfs are Gaussian-like and are gently moved toward high and low values depending on which diffusivity area the monitoring locations are in. This regression toward the mean effect of the EnKF is generally recognized as it gives the best prediction in the squared error sense [27].

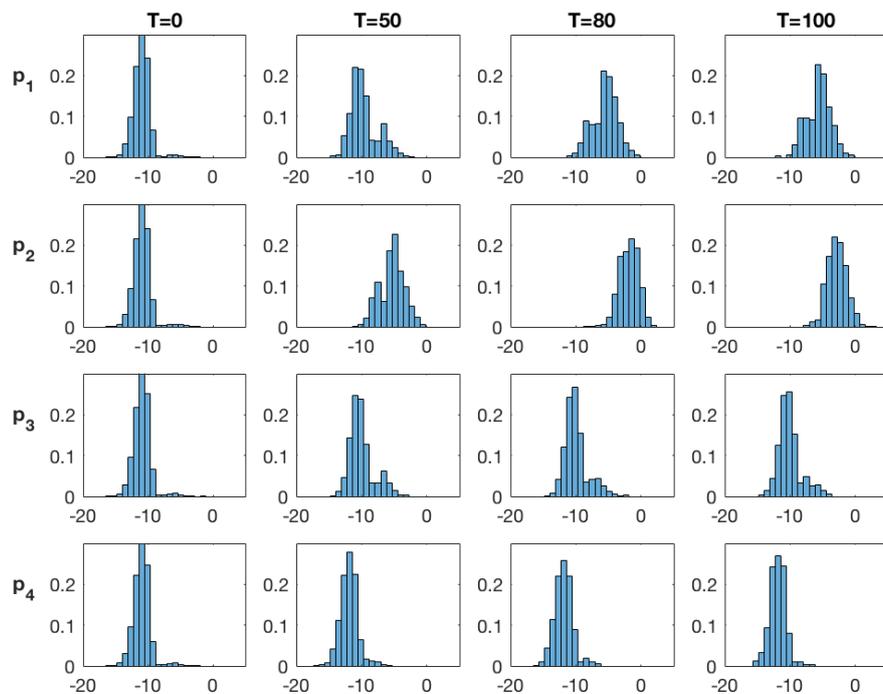


Figure 6. SEnKF approach: Marginal posterior distribution of the log diffusivity $f(\log(\lambda_i)|d_{0:T})$ at time $T = 0, 50, 80, 100$ at the monitoring locations (1, 2, 3, 4) denoted (p_1, p_2, p_3, p_4) .

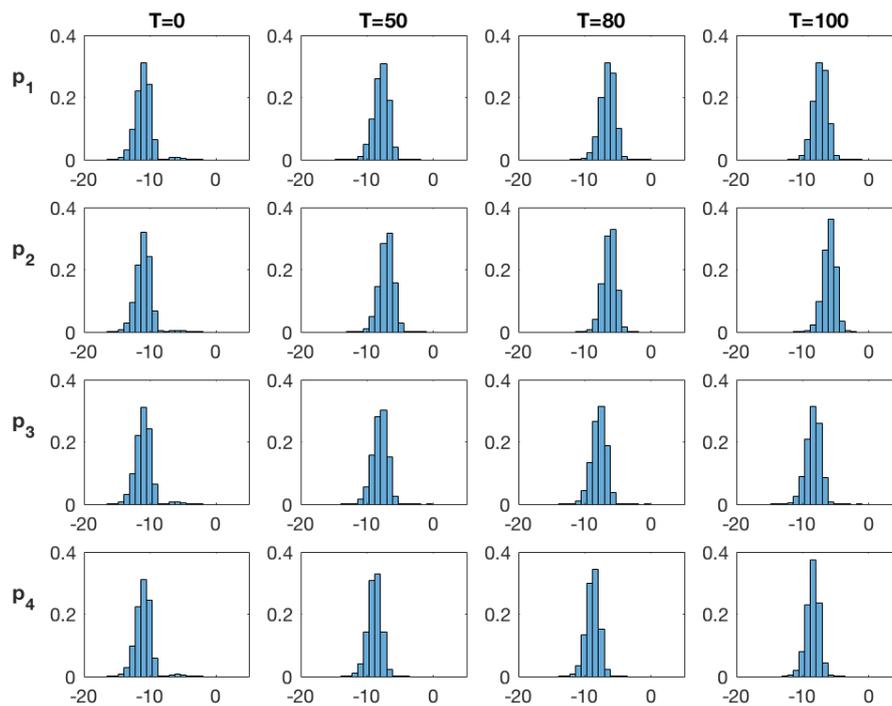


Figure 7. EnKF approach: Marginal posterior distribution of the log diffusivity $f(\log(\lambda_i)|d_{0:T})$ at time $T = 0, 50, 80, 100$ at the monitoring locations (1, 2, 3, 4) denoted (p_1, p_2, p_3, p_4) .

Figure 8 displays the MMAP predictions based on the SENKF and the traditional EnKF at time $T = 0, 50, 80, 100$. At $T = 0$, the predictions from the two algorithms are identical since they use identical prior models. As observations are assimilated, the SENKF predictions reproduce the high diffusivity area relatively well, with clear contrast. The traditional EnKF predictions also indicate the diffusivity areas, but with less contrast. Figure 9 displays the MMAP prediction at $T = 100$ along the section B-B' shown in Figure 2. The high contrast reliable reconstruction of the high diffusivity channel by the SENKF algorithm is confirmed. The traditional EnKF predictions appear less reliable. The 80% highest density interval (HDI) [28] covers the true diffusivity values for the SENKF while these values are far outside the interval for the traditional EnKF results. The results are consistent with the observations made regarding the marginal posterior pdfs in Figures 6 and 7.

Figures 10 and 11 show realizations and spatial histograms from the posterior distribution of the log-diffusivity at time $T = 100$ for the SENKF and traditional EnKF algorithms, respectively. The realizations from the SENKF largely reproduce the channel with clear contrast while the realizations from the EnKF also reproduce the channel, but with much less contrast. The spatial histograms also underline the difference in contrast in that they are clearly bi-modal for the SENKS and much more Gaussian-like for the EnKS.

Table 2 shows that the RMSE of the MMAP prediction relative to the true diffusivity field for the SENKF is approximately 30% lower than for the EnKF .

Table 2. RMSE comparing the MMAP prediction and the true log diffusivity field at time $T = 100$.

	SEnKF	ENKF
$RMSE_{T=100}$	2.72	3.76

This test case clearly illustrates the SENKF’s ability to conserve multimodality in the posterior distribution and it leads to predictions with better contrast and accuracy. We conclude that the reconstruction of the true diffusivity field is done more reliably by the SENKF algorithm than by the EnKF algorithm.

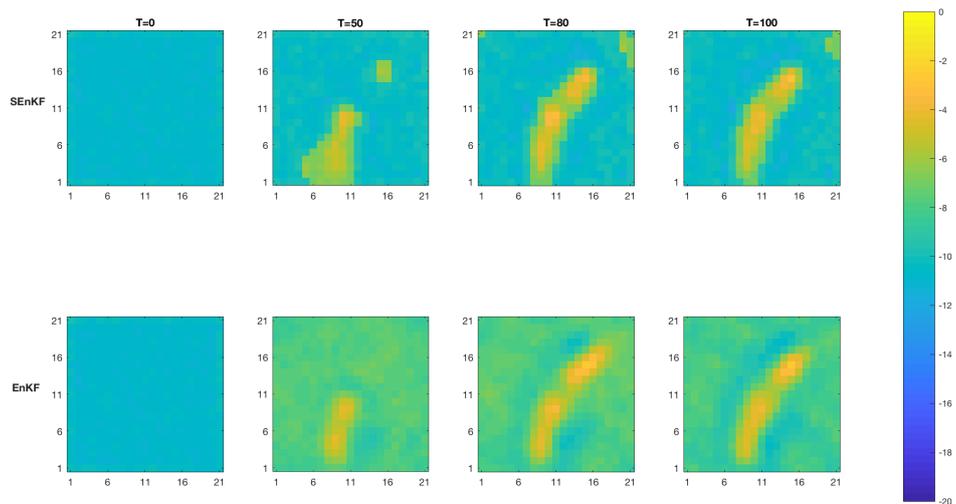


Figure 8. MMAP predictions of the log diffusivity field $(\log(\lambda)|d_{0:T})$ at time $T = 0, 50, 80, 100$ (upper panels—SEnKF approach, lower panels—EnKF approach).

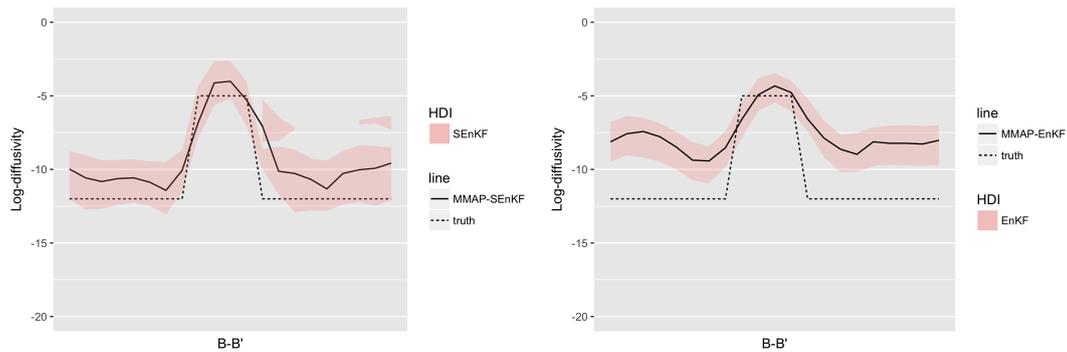


Figure 9. MMAP predictions of the log diffusivity field with 80% HDI in cross section B-B' at time $T = 100$ with SENKF (left) and with EnKF (right).

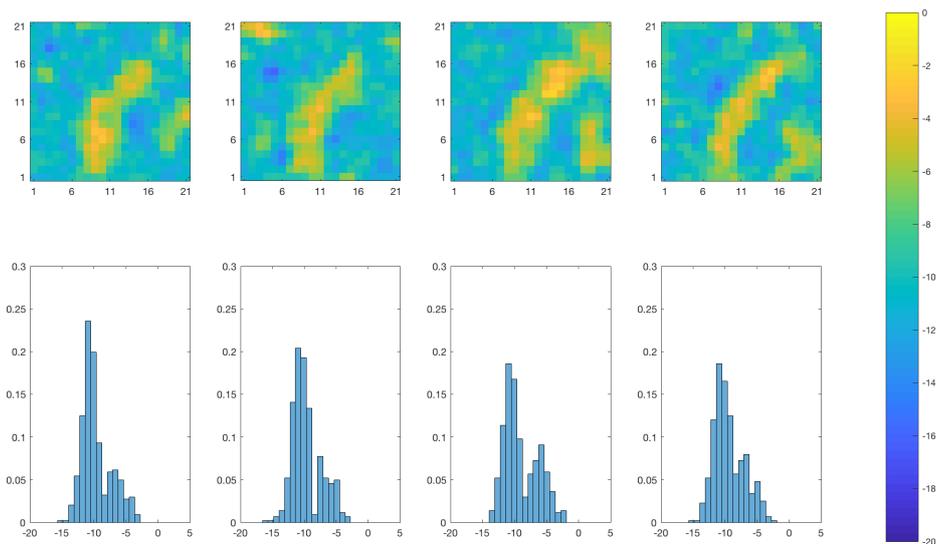


Figure 10. SENKF approach: Realizations of the posterior distribution of the log diffusivity $f(\log(\lambda)|d_{0:T})$ at time $T = 100$ (upper panels) and associated spatial histogram (lower panels). Lower panels: the horizontal axes represent the log-diffusivity, the vertical axes represent the relative prevalence of each log-diffusivity value for the realization in the panel right above.

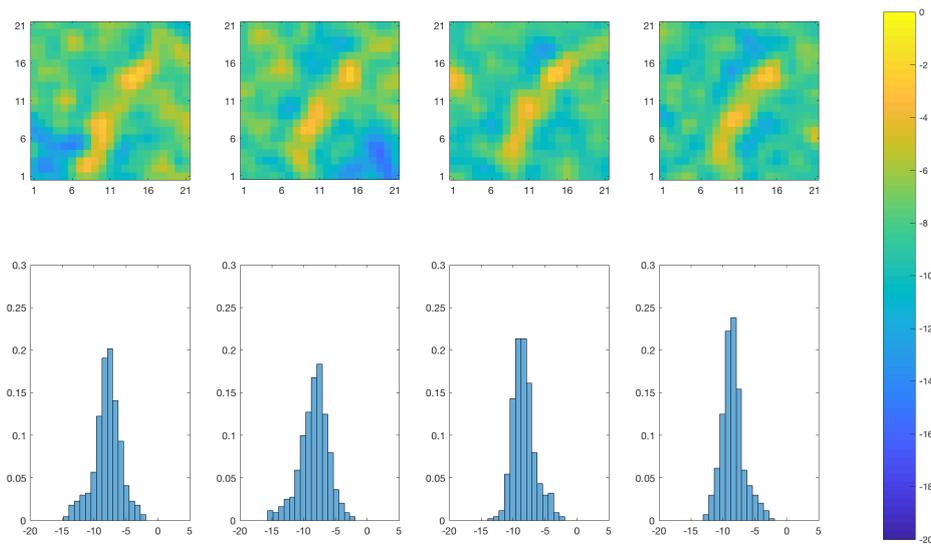


Figure 11. EnKF approach: Realizations of the posterior distribution of the log diffusivity $f(\log(\lambda)|d_{0:T})$ at time $T = 100$ (**upper panels**) and associated spatial histogram (**lower panels**). Lower panels: the horizontal axes represent the log-diffusivity, the vertical axes represent the relative prevalence of each log-diffusivity value for the realization in the panel right above.

3.3. Test Case 2: Reconstructing the Initial Field

The focus of the study is to evaluate the unknown initial state of the temperature field r_0 based on the observations d . The posterior model $f(r_0|d)$ is assessed using the SEnKS, see Appendix A, and the results are compared to those from the traditional EnKS.

The true initial temperature field r_0 is set at 20°C except for a square shaped region with temperature set at 45°C , see Figure 12. The temperature field is formally defined as

$$r_0(x) = \tau_1 \mathbf{1}(x \in D_1) + \tau_2 \mathbf{1}(x \in D_2), \tag{26}$$

where $D_1 \subset D$ is the low temperature area and $D_2 \subset D$ is the high temperature area, and $\tau_1 = 20^\circ\text{C}$ and $\tau_2 = 45^\circ\text{C}$. Figure 12 shows the true log-diffusivity n -vector $\log(\lambda)$. The diffusivity λ is always positive. To ensure that ensemble updates do not lead to negative diffusivity values, we work on $\log(\lambda)$. The heat contained in the high temperature area will diffuse towards the rest of the field according to the diffusion equation in Equation (20), see Figure 13. The temporal observations are collected at five different observation locations according to the likelihood model in Equation (23), see Figure 12. Figure 14 displays the observations d where it is clear that the observed temperature increases substantially only at the observation locations close to the high temperature area. The measurements are taken every second from $T = 0$ to $T = 50$.

The unknown initial temperature field r_0 is assigned a stationary selection-Gaussian random field prior model with parameters $\Theta_r^{SG} = (\mu_{\tilde{r}}, \mu_v, \sigma_{\tilde{r}}, \Sigma_{\tilde{r}}^\rho, \gamma, A)$. The parameter values are listed in Table 3. The unknown log-diffusivity field $\log(\lambda)$ is assigned a stationary Gaussian random field prior model with parameters $\Theta_\lambda^G = (\mu_\lambda, \sigma_\lambda, \Sigma_\lambda^\rho)$ with expectation and variance levels $\mu_\lambda = -8.5$ and $\sigma_\lambda^2 = 2$, respectively. For both prior models, the spatial correlation ($n \times n$)-matrix Σ^ρ is defined by the second order exponential spatial correlation function $\rho(\tau) = \exp(-\tau^2/\delta^2)$; $\delta = 0.15$, with interdistance τ .

Figure 15 contains four realizations from the prior model of the temperature field and their spatial histograms. The marginal initial distributions of the realizations are bi-modal and spatially stationary except for boundary effects. The selection set $A \subset \mathbb{R}^n$ in the prior model is chosen to obtain a bi-modal marginal distribution with one large mode approximately centered about 20°C and a smaller mode centered close to 45°C .

Table 3. Parameter values for the selection-Gaussian initial distribution for the initial temperature prior model.

Parameters	Values
$\mu_{\tilde{r}}$	28.75
μ_v	0
$\sigma_{\tilde{r}}$	$\sqrt{10}$
γ	0.8
A	$] -\infty, -0.2] \cup [0.5, +\infty[$

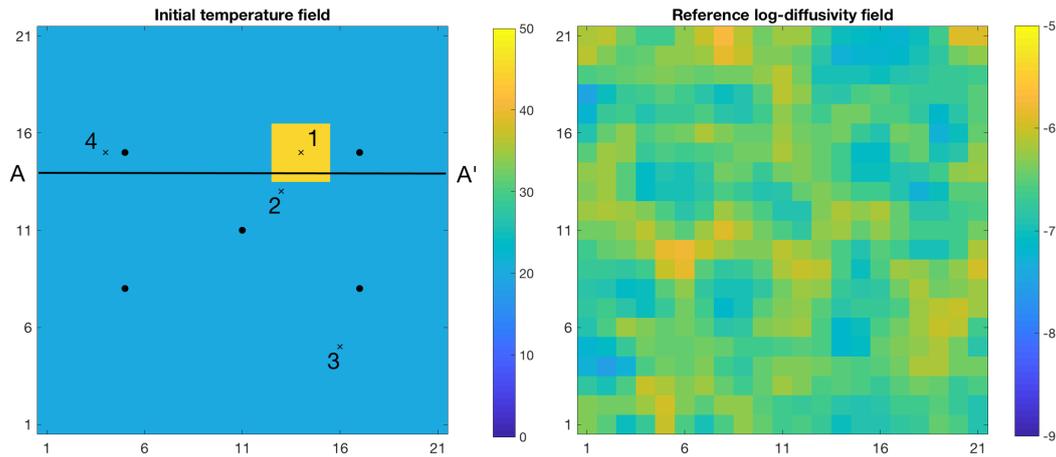


Figure 12. Initial temperature ($^{\circ}\text{C}$) field (left) with data collection points \cdot and monitoring locations \times and reference log-diffusivity field (right).

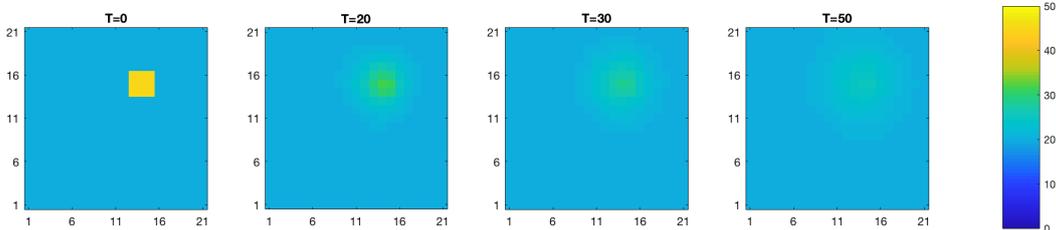


Figure 13. True temperature ($^{\circ}\text{C}$) field evolution over time.

The SEnKS operates on the $3n$ -vector $[\tilde{r}_0, v, \log(\lambda)]$, and therefore we generate an initial ensemble with $n_e = 10,000$ ensemble members that are sampled from the Gaussian $3n$ -vector $[\tilde{r}_0, v, \log(\lambda)]$ with pdf,

$$\begin{bmatrix} \tilde{r}_0 \\ v \\ \log(\lambda) \end{bmatrix} \sim \varphi_{3n} \left(\begin{bmatrix} \tilde{r}_0 \\ v \\ \log(\lambda) \end{bmatrix}; \begin{bmatrix} \mu_{\tilde{r}} \\ \mu_v \\ \mu_{\lambda} \end{bmatrix}, \begin{bmatrix} \sigma_{\tilde{r}}^2 \Sigma_{\tilde{r}}^{\rho} & \gamma \sigma_{\tilde{r}} \Sigma_{\tilde{r}}^{\rho T} & 0 \\ \gamma \sigma_{\tilde{r}} \Sigma_{\tilde{r}}^{\rho} & \gamma^2 \Sigma_{\tilde{r}}^{\rho} + (1 - \gamma^2) I_n & 0 \\ 0 & 0 & \sigma_{\lambda}^2 \Sigma_{\lambda}^{\rho} \end{bmatrix} \right). \quad (27)$$

The EnKS operates on the $2n$ -vector $[r_0, \log(\lambda)]$, and therefore we generate an initial ensemble with $n_e = 10,000$ ensemble members that are sampled from selection-Gaussian distribution $f(r_0, \log(\lambda))$. The variables r_0 and $\log(\lambda)$ are independent, so we generate them independently: 10,000 samples from the selection-Gaussian n -vector r_0 with parameters Θ_r^{SG} and 10,000 samples from the Gaussian n -vector $\log(\lambda)$ with parameters Θ_{λ}^G . Due to the size of the ensemble relative to the dimension of the problem, we used neither localization nor inflation in the algorithms.

To illustrate the differences between the SEnKS and the EnKS we present the following results for both algorithms:

1. The marginal posterior distributions $f(r_{0,i}|d_{0:T})$ of the initial temperature field at four monitoring locations denoted 1, 2, 3, 4 on Figure 12, at time $T = 0, 20, 30, 50$.
2. The marginal maximum a posteriori (MMAP) prediction of the initial temperature field at time at time $T = 0, 20, 30, 50$.
3. Realizations from the posterior distribution $f(r_0|d_{0:T})$ of the initial temperature field at time $T = 50$.
4. The root mean square errors (RMSE) of the MMAP prediction of the initial temperature field relative to the true initial temperature field at time $T = 50$.

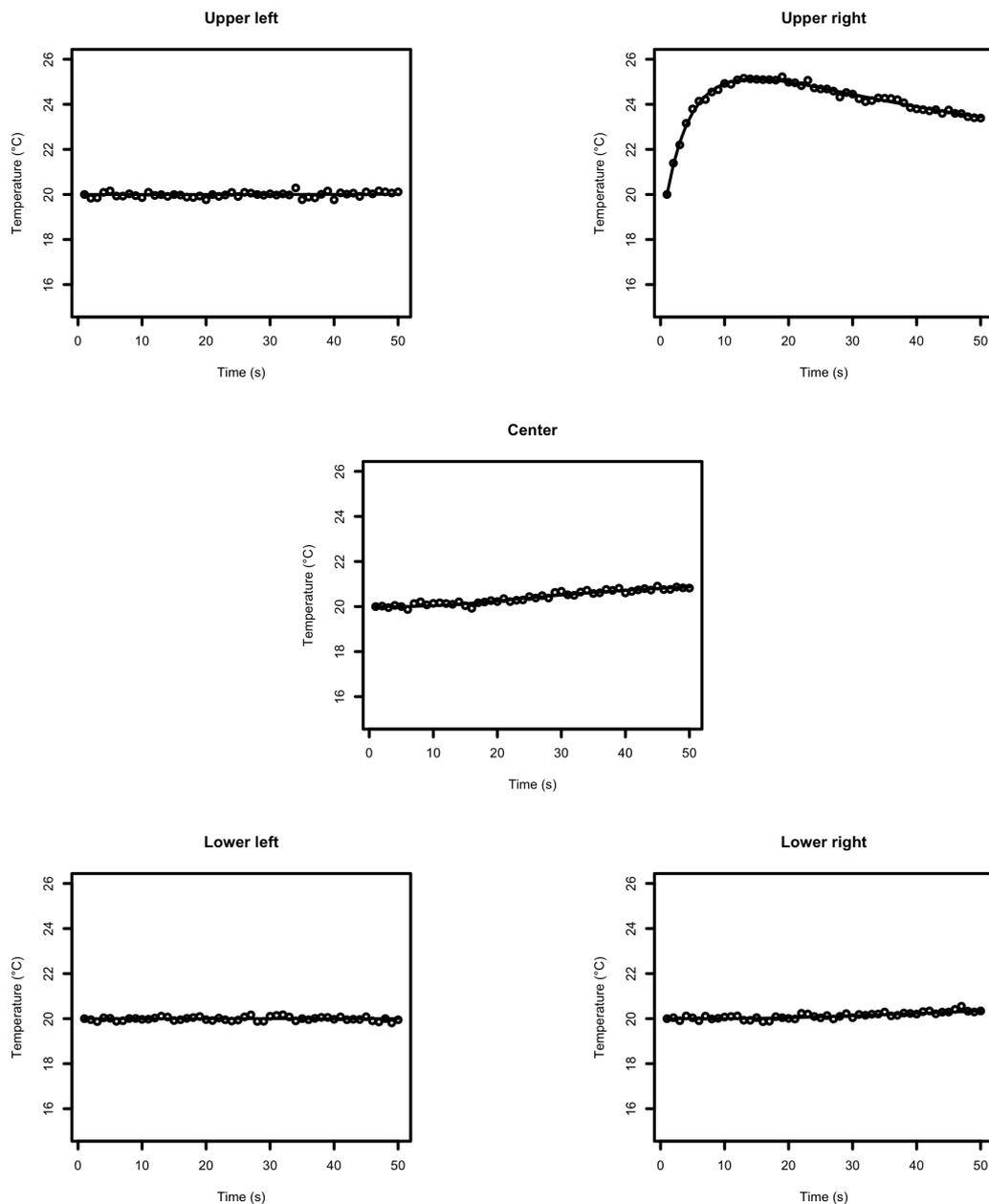


Figure 14. Data collected over time (points) and true temperature ($^{\circ}\text{C}$) evolution at the data collection points (line).

The marginal posterior pdfs $f(r_{0,i}|d_{0:T})$ at the four monitoring locations at time $T = 0, 20, 30, 50$ are displayed in Figures 16 and 17 for the SENKS and EnKS algorithms, respectively. At $T = 0$, the prior models for both algorithms are identical and so are the marginal pdfs. Monitoring location 1 is placed

inside the high temperature area. As observations are assimilated, the marginal pdf from the SENKS remain bi-modal, but the high value mode increases steadily. For the other monitoring locations, all placed outside the high temperature area, the bi-modality is reproduced but with a dominant low value mode. The relative size of the modes reflects the distance to the high temperature area and the observation locations. The marginal pdfs from the EnKS lose their bi-modality after a few assimilation steps and from then on the Gaussian-like marginal pdfs are only slightly shifted by the assimilation of observations.

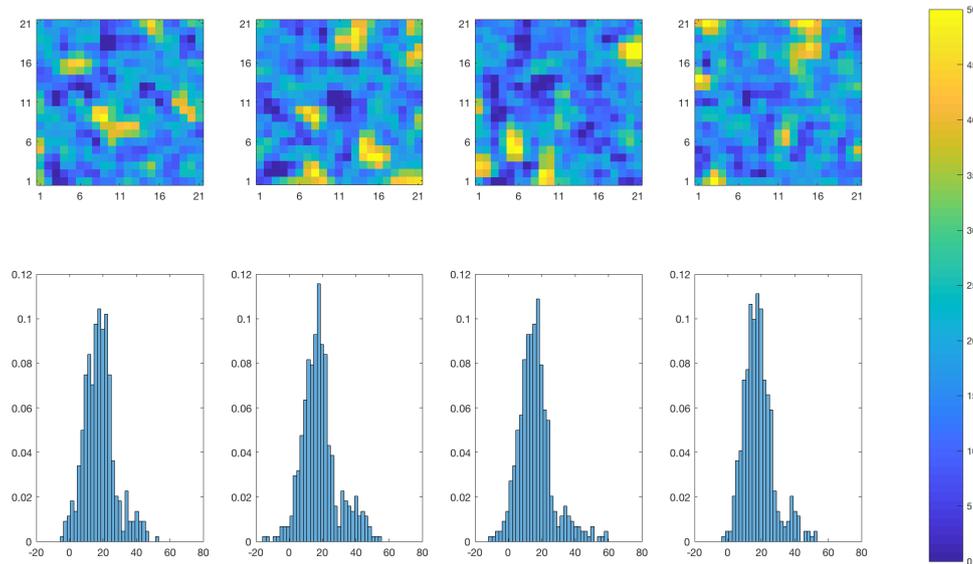


Figure 15. Realizations from the selection-Gaussian initial distribution of the initial temperature field $f(r_0)$ at time $T = 0$ (upper panels) and associated spatial histogram (lower panels). Upper panels: the colorbar gives the temperature in $^{\circ}\text{C}$. Lower panels: the horizontal axes represent the temperature ($^{\circ}\text{C}$), the vertical axes represent the relative prevalence of each temperature value for the realization right above.

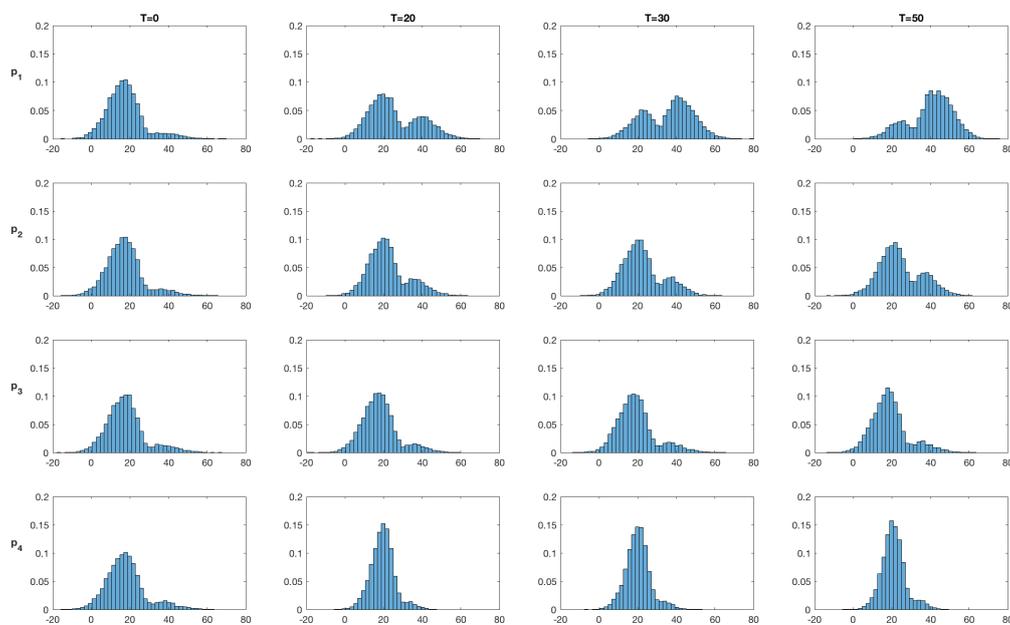


Figure 16. SENKS approach: Marginal posterior distributions of the initial temperature $f(r_{0,i}|d_{0:T})$ at time $T = 0, 20, 30, 50$ at monitoring locations ($i = 1, 2, 3, 4$) denoted (p_1, p_2, p_3, p_4) . The horizontal axes representing temperature are expressed in $^{\circ}\text{C}$.

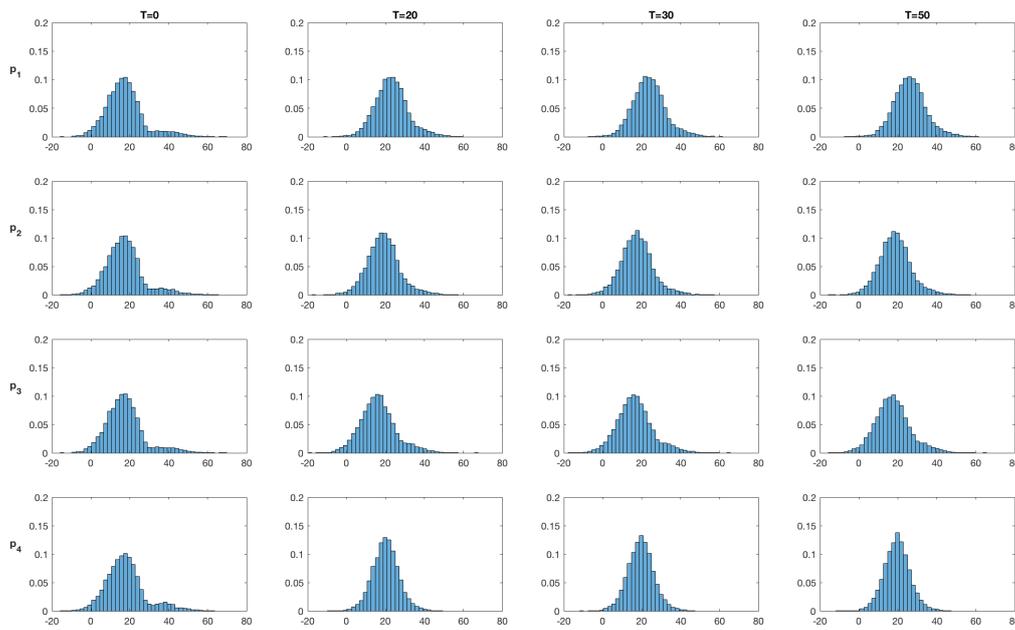


Figure 17. EnKS approach: Marginal posterior distributions of the initial temperature $f(r_{0,i}|d_{0:T})$ at time $T = 0, 20, 30, 50$ at monitoring locations ($i = 1, 2, 3, 4$) denoted (p_1, p_2, p_3, p_4) . The horizontal axes representing temperature are expressed in $^{\circ}\text{C}$.

Figure 18 displays the MMAP predictions of the initial temperature field based on the SENKS and the traditional EnKS at time $T = 0, 20, 30, 50$. For the SENKS, the high temperature area is clearly identifiable with clear contrast from time $T = 30$ while for the EnKS the high temperature area is hardly ever identifiable on the MMAP predictions that show little contrast. Figure 19 displays the MMAP prediction of the initial temperature field at $T = 50$ along the section A-A', see Figure 12, for the SENKS and the traditional EnKS. The SENKS clearly identifies the high temperature area and the 80% HDI covers the truth, while the EnKS clearly fails to identify the high temperature area and the 80% HDI does not even cover it.

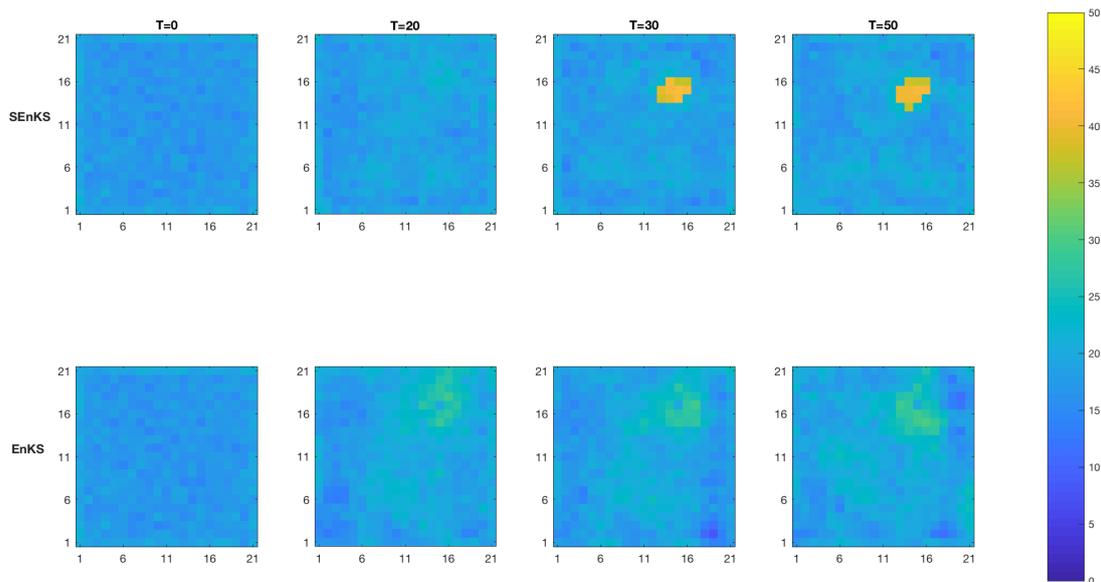


Figure 18. MMAP predictions of the initial temperature ($^{\circ}\text{C}$) field at time $T = 0, 20, 30, 50$ for the SENKS approach (**upper**) and the EnKS approach (**lower**).

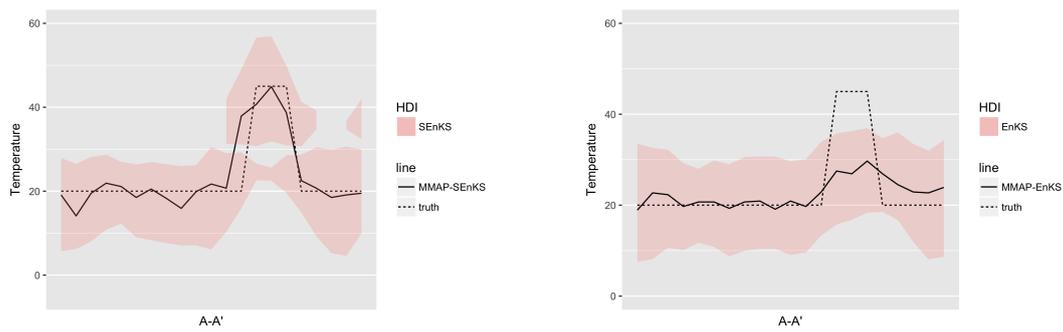


Figure 19. MMAP predictions of the initial temperature ($^{\circ}\text{C}$) field ($r_0|d_{0:T}$) with 80% HDI in cross section A-A' at time $T = 50$ with SENKS (left) and with EnKS (right).

Realizations of the posterior model at $T = 50$ based on the SENKS and the traditional EnKS algorithms are displayed in Figures 20 and 21, respectively. The SENKS produces realizations that appear bi-modal while the ensemble members from the EnKS display more symmetric spatial histograms. Even though the differences between the realizations are quite subtle, they are consistent with previous results.

Table 4 shows that RMSE of the MMAP prediction relative to the true initial temperature field for the SENKS is approximately 20% lower than for the EnKS.

This test case clearly illustrates the ability of the SENKS to conserve multimodality in the posterior distribution, and it leads to predictions with better contrast and accuracy. We conclude that the SENKS algorithm provides a more reliable reconstruction of the initial state of the temperature field than the traditional EnKS algorithm. Note that the posterior model for the unknown diffusivity field $f(\log(\lambda)|d)$ can also be assessed with the two algorithms. When comparing the MMAP predictions relative to the true diffusivity field, see Figure 12, we observe that none of the algorithms provide reliable predictions. We conclude that the small scale variations in the field are not sufficiently distinct to be identified.

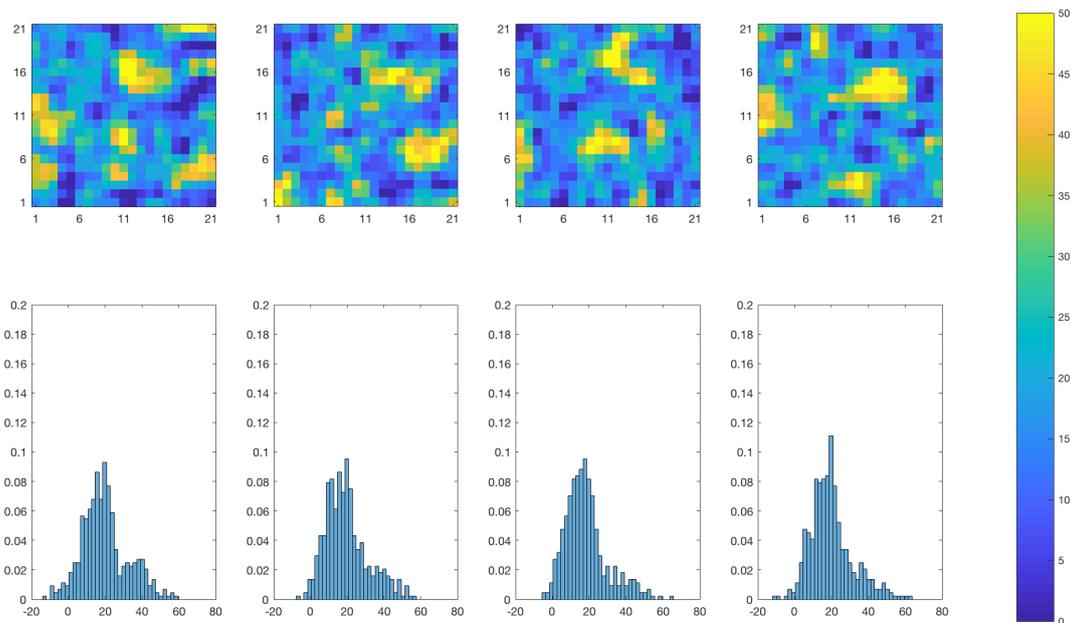


Figure 20. SENKS approach: Realizations from the posterior distribution $f(r_0|d_{0:T})$ of the initial temperature field at time $T = 50$. Upper panels: the colorbar gives the temperature in $^{\circ}\text{C}$. Lower panels: the horizontal axes represent the temperature ($^{\circ}\text{C}$), the vertical axes represent the relative prevalence of each temperature value for the realization right above.

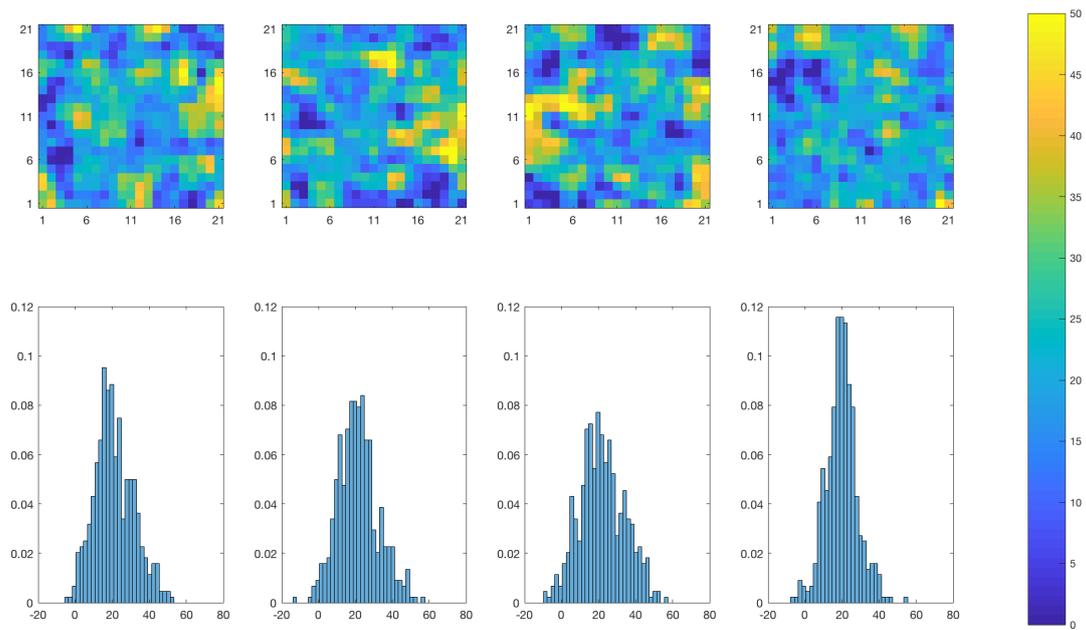


Figure 21. EnKS approach: Realizations from the posterior distribution of the initial temperature field $f(r_0|d_{0:T})$ at time $T = 50$. Upper panels: the colorbar gives the temperature in °C. Lower panels: the horizontal axes represent the temperature (°C), the vertical axes represent the relative prevalence of each temperature value for the realization right above.

Table 4. RMSE comparing the MMAP prediction of the initial temperature field and the initial temperature field at time $T = 50$.

	SEnKS	ENKS
$RMSE_{T=50}$	2.92	3.72

4. Discussion

The traditional EnKF and EnKS algorithms provide an ensemble that directly represents the posterior models $f(r_{T+1}|d_{0:T})$ and $f(r_{0:T}|d_{0:T})$, respectively. Hence, the posterior models can be assessed by displaying statistics based on these ensembles. Reliable assessment of the posterior model in the two test cases can be obtained with approximately 1000 ensemble members. The SEEnKF and SEEnKS algorithms under study provide an ensemble of the augmented posterior models, $f(\tilde{r}_{T+1}, \nu|d_{0:T})$ and $f(\tilde{r}_{0:T}, \nu_{0:T}|d_{0:T})$, respectively. In order to obtain the posterior models of interest, $f(r_{T+1}|d_{0:T})$ and $f(r_{0:T}|d_{0:T})$, the conditioning on $\nu \in A$ must be made by empirical sampling based inference, see [21,22]. This inference requires the estimation of the expectation vector $\mu_{\tilde{r}_\nu}$ and covariance matrix $\Sigma_{\tilde{r}_\nu}$. The two test cases are defined on a (21×21) -grid for both \tilde{r}_t and ν —hence in dimension 882. The expectation and covariance will have 882 and 389,403 unique entries, respectively. Our experience from this study is that approximately 10,000 ensemble members are required to obtain reliable assessment of the posterior models of interest. To reduce the ensemble size, we have tested various localization approaches [29], without notable success, and leave the subject for further research.

5. Conclusions

Data assimilation of spatio-temporal variables with multimodal spatial histograms is challenging. Traditional ensemble Kalman algorithms enforce a regression towards the mean due to the linearized conditioning on observations, hence the multimodality is averaged out. We introduce the selection ensemble Kalman algorithms, termed SEEnKF and SEEnKS. These algorithms are based on recursive expressions similar to the ones justifying the traditional ensemble Kalman algorithms, but they are defined in an augmented space including the selection variable. From the two case studies, we conclude

that multimodality is much better represented by the selection ensemble Kalman algorithms than by the traditional ones. We obtain RMSE reductions in the range of 20 to 30%.

The traditional ensemble Kalman algorithms provide an ensemble representation of the posterior model of interest hence making assessment of the posterior pdf simple. The selection ensemble Kalman algorithms are defined in an augmented space and conditioning on the selection variable must be made a posteriori. For this conditioning to be reliable, the ensemble size needs to be much larger than for the traditional algorithms. Hence, there is a trade-off between improved reproduction of multimodal characteristics of the phenomenon under study and the computational demands. In our case study, the ensemble size needed to be increased by approximately a factor of ten.

We have not fully explored the possibilities of robust estimation of model parameters in the conditioning of the selection variable. This robustification may reduce the ensemble size requirements. Note that parallelization in forwarding of the ensemble is possible and it will reduce the computer demands.

Author Contributions: Conceptualization, M.C. and H.O.; Formal analysis, M.C. and H.O.; Funding acquisition, H.O.; Investigation, M.C. and H.O.; Methodology, M.C. and H.O.; Project administration, H.O.; Resources, H.O.; Software, M.C.; Supervision, H.O.; Validation, M.C. and H.O.; Visualization, H.O.; Writing—original draft, M.C. and H.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research and the APC are funded by the research initiative: ‘Uncertainty in Reservoir Evaluation’ at Department of Mathematical Sciences, NTNU, Trondheim, Norway.

Acknowledgments: The research is a part of the Uncertainty in Reservoir Evaluation (URE) activity at the Norwegian University of Science and Technology (NTNU), Trondheim, Norway.

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

Glossary

$\mathbf{r}_t \in \mathbb{R}^n$	discretized spatial variable at time t .
$[\tilde{\mathbf{r}}_t, \mathbf{v}] \in \mathbb{R}^{2n}$	Gaussian variables; basis and auxiliary variables, at time t .
$A \subset \mathbb{R}^n$	selection set.
$\mathbf{r}_t = [\tilde{\mathbf{r}}_t \mathbf{v} \in A]$	selection Gaussian variable at time t .
$\boldsymbol{\mu}_t \in \mathbb{R}^n$	expectation vector.
$\boldsymbol{\Sigma}_t \in \mathbb{R}^n \times \mathbb{R}^n$	covariance matrix.
$\boldsymbol{\Sigma}_t^o \in \mathbb{R}^n \times \mathbb{R}^n$	correlation matrix.
$\boldsymbol{\Gamma}_{\cdot \cdot} \in \mathbb{R}^n \times \mathbb{R}^n$	matrix cross-correlation
$d_t \in \mathbb{R}^m$	observation variable at time t .
$\omega(\mathbf{r}_t, \boldsymbol{\epsilon}_t^r)$	forward function at time t .
$\psi(\mathbf{r}_t, \boldsymbol{\epsilon}_t^d)$	observation function at time t .
$\rho_{\cdot}(\tau) \in \mathbb{R}_{[-1,1]}$	spatial correlation function.

Appendix A

The algorithms detailed in Algorithms A1 and A2 follow the formalism in [4].

Algorithm A1 description: The SENKF is a two-step algorithm. The first step is a traditional EnKF that evaluates $[\tilde{\mathbf{r}}_{T+1}, \mathbf{v} | \mathbf{d}_{0:T}]$. The second step consists of a sampling step where the target quantity $[\mathbf{r}_{T+1} | \mathbf{d}_0, \dots, \mathbf{d}_T]$ is evaluated using $[\tilde{\mathbf{r}}_{T+1}, \mathbf{v} | \mathbf{d}_{0:T}]$ from the first step.

Algorithm A1 Selection Ensemble Kalman Filter (SEnKF)

A time series of ensembles is defined as $e_t = \{(\tilde{r}_t^{u(i)}, \nu_t^{u(i)}, d_t^i), i = 1, \dots, n_e\}, \forall t = 0, \dots, T$ and the $(2n + m)$ -vector $[\tilde{r}_t, \nu_t, d_t]$ has the following covariance matrix:

$$\Sigma_{\tilde{r}\nu d} = \begin{bmatrix} \Sigma_{\tilde{r}\nu} & \Gamma_{\tilde{r}\nu, d} \\ \Gamma_{d, \tilde{r}\nu} & \Sigma_d \end{bmatrix}$$

1. **Initiate:**
2. $n_e =$ No. of ensemble members
3. Generate $\begin{bmatrix} \tilde{r}_0^{u(i)} \\ \nu_0^{u(i)} \end{bmatrix} \sim f(\tilde{r}_0, \nu), i = 1, \dots, n_e$
4. Generate $\epsilon_0^{d(i)} \sim \mathcal{U}_m[0, 1], i = 1, \dots, n_e$
5. $d_0^i = \psi_0(\tilde{r}_0^{u(i)}, \epsilon_0^{d(i)}), i = 1, \dots, n_e$
6. $e_0 = \{(\tilde{r}_0^{u(i)}, \nu_0^{u(i)}, d_0^i), i = 1, \dots, n_e\}$
7. **Iterate** $t = 0, \dots, T$:
8. **Conditioning:**
9. Estimate $\Sigma_{\tilde{r}\nu d}$ from $e_t \rightarrow \hat{\Sigma}_{\tilde{r}\nu d}$
10. $\begin{bmatrix} \tilde{r}_t^{c(i)} \\ \nu_t^{c(i)} \end{bmatrix} = \begin{bmatrix} \tilde{r}_t^{u(i)} \\ \nu_t^{u(i)} \end{bmatrix} + \hat{\Gamma}_{\tilde{r}\nu, d} \hat{\Sigma}_d^{-1} (d_t - d_t^i), i = 1, \dots, n_e$
11. **Forwarding:**
12. Generate $\epsilon_t^{\tilde{r}(i)} \sim \mathcal{U}_n[0, 1], i = 1, \dots, n_e$
13. $\begin{bmatrix} \tilde{r}_{t+1}^{u(i)} \\ \nu_{t+1}^{u(i)} \end{bmatrix} = \begin{bmatrix} \omega_t(\tilde{r}_t^{c(i)}, \epsilon_t^{\tilde{r}(i)}) \\ \nu_t^{c(i)} \end{bmatrix}, i = 1, \dots, n_e$
14. If $t < T$
15. Generate $\epsilon_{t+1}^{d(i)} \sim \mathcal{U}_m[0, 1], i = 1, \dots, n_e$
16. $d_{t+1}^i = \psi_{t+1}(\tilde{r}_{t+1}^{u(i)}, \epsilon_{t+1}^{d(i)}), i = 1, \dots, n_e$
17. $e_{t+1} = \{(\tilde{r}_{t+1}^{u(i)}, \nu_{t+1}^{u(i)}, d_{t+1}^i), i = 1, \dots, n_e\}$
18. Else
19. $e_{t+1} = \{(\tilde{r}_{t+1}^{u(i)}, \nu_{t+1}^{u(i)}), i = 1, \dots, n_e\}$
20. **End iterate**
21. Estimate $\mu_{\tilde{r}\nu}, \Sigma_{\tilde{r}\nu}$ from $e_{T+1} \rightarrow \hat{\mu}_{\tilde{r}\nu}, \hat{\Sigma}_{\tilde{r}\nu}$
22. Assess
23. $\hat{f}(r_{T+1}|d_0, \dots, d_T) = [\Phi_n(A; \hat{\mu}_\nu, \hat{\Sigma}_\nu)]^{-1} \times \Phi_n(A; \hat{\mu}_\nu + \hat{\Gamma}_{\nu|\tilde{r}}(r - \hat{\mu}_{\tilde{r}}), \hat{\Sigma}_{\nu|\tilde{r}}) \times \varphi_n(r; \hat{\mu}_{\tilde{r}}, \hat{\Sigma}_{\tilde{r}})$
24. End Algorithm

The ensemble e_{T+1} represents $[\tilde{r}_{T+1}, \nu|d_{0:T}]$. To assess $f(r_{T+1}|d_{0:T}) = f(\tilde{r}_{T+1}|d_{0:T}, \nu \in A)$, the sampling algorithm specified in [21] requires $E[\tilde{r}_{T+1}, \nu|d_{0:T}] = \mu_{\tilde{r}\nu}$ and $Cov[\tilde{r}_{T+1}, \nu|d_{0:T}] = \Sigma_{\tilde{r}\nu}$ which are estimated using the ensemble e_{T+1} .

Algorithm A2 description: The SEEnKS is a two-step algorithm. The first step is a traditional EnKS that evaluates $[\tilde{r}_t, \nu|d_{0:T}]$. The second step consists of a sampling step where the target quantity $[r_t|d_0, \dots, d_T]$ is evaluated using $[\tilde{r}_t, \nu|d_{0:T}]$ from the first step.

Algorithm A2 Selection Ensemble Kalman Smoother (SEnKS)

Two time series of ensemble sets are defined as

$$\left. \begin{aligned} \mathbf{e}_t^{\tilde{r}v} &= \{(\tilde{r}_t^i, \mathbf{v}_t^i), i = 1, \dots, n_e\} \\ \mathbf{e}_t^d &= \{d_t^i, i = 1, \dots, n_e\} \end{aligned} \right\} \text{ for } t = 0, \dots, T$$

and the accumulated ensemble set defined as

$$\mathbf{E}_{0:t}^{\tilde{r}v} = \{(\tilde{r}_{0:t}^i, \mathbf{v}_{0:t}^i), i = 1, \dots, n_e\} \text{ for } t = 0, \dots, T$$

The $(2n + m)$ -vector $[\tilde{r}_t, \mathbf{v}_t, \mathbf{d}_t]$ has covariance matrix $\Sigma_{\tilde{r}vd}^t = \begin{bmatrix} \Sigma_{\tilde{r}v}^t & \Gamma_{\tilde{r}v,d}^t \\ \Gamma_{d,\tilde{r}v}^t & \Sigma_d^t \end{bmatrix}$

The $(2n(t + 1) + m)$ -vector $[\tilde{r}_{0:t}, \mathbf{v}_{0:t}, \mathbf{d}_t]$ has covariance matrix $\Sigma_{\tilde{r}vd}^{0:t} = \begin{bmatrix} \Sigma_{\tilde{r}v}^{0:t} & \Gamma_{\tilde{r}v,d}^{0:t} \\ \Gamma_{d,\tilde{r}v}^{0:t} & \Sigma_d^t \end{bmatrix}$

1. **Initiate**
2. $n_e =$ No. of ensemble members
3. Generate $\begin{bmatrix} \tilde{r}_0^{u(i)} \\ \mathbf{v}_0^{u(i)} \end{bmatrix} \sim f(\tilde{r}_0, \mathbf{v}), i = 1, \dots, n_e$
4. $\mathbf{E}_{0:0}^{\tilde{r}v} = \{(\tilde{r}_0^{u(i)}, \mathbf{v}_0^{u(i)}), i = 1, \dots, n_e\}$
5. Generate $\epsilon_0^{d(i)} \sim \mathcal{U}_m[0, 1], i = 1, \dots, n_e$ iid
6. $d_0^i = \psi_0(\tilde{r}_0^{u(i)}, \epsilon_0^{d(i)}), i = 1, \dots, n_e$
7. $\mathbf{e}_0^d = \{d_0^i, i = 1, \dots, n_e\}$
8. Estimate $\Sigma_{\tilde{r}vd}^0$ from $\mathbf{E}_{0:0}^{\tilde{r}v}, \mathbf{e}_0^d \rightarrow \hat{\Sigma}_{\tilde{r}vd}^0$
9. $\begin{bmatrix} \tilde{r}_0^{c(i)} \\ \mathbf{v}_0^{c(i)} \end{bmatrix} = \begin{bmatrix} \tilde{r}_0^{u(i)} \\ \mathbf{v}_0^{u(i)} \end{bmatrix} + \hat{\Gamma}_{\tilde{r}v,d}^0 \hat{\Sigma}_d^{0-1} (d_0 - d_0^i), i = 1, \dots, n_e$
10. **Iterate** $t = 1, \dots, T$:
11. Forwarding
12. Generate $\epsilon_t^{\tilde{r}(i)} \sim \mathcal{U}_n[0, 1], i = 1, \dots, n_e$
13. $\begin{bmatrix} \tilde{r}_t^{u(i)} \\ \mathbf{v}_t^{u(i)} \end{bmatrix} = \begin{bmatrix} \omega_t(\tilde{r}_{t-1}^{c(i)}, \epsilon_t^{\tilde{r}(i)}) \\ \mathbf{v}_{t-1}^{c(i)} \end{bmatrix}, i = 1, \dots, n_e$
14. $\mathbf{E}_{0:t}^{\tilde{r}v} = \{\mathbf{E}_{0:t-1}^{\tilde{r}v}, (\tilde{r}_t^{u(i)}, \mathbf{v}_t^{u(i)}), i = 1, \dots, n_e\}$
15. Generate $\epsilon_t^{d(i)} \sim \mathcal{U}_m[0, 1], i = 1, \dots, n_e$ iid
16. $d_t^i = \psi_0(\tilde{r}_t^{u(i)}, \epsilon_t^{d(i)}), i = 1, \dots, n_e$
17. $\mathbf{e}_t^d = \{d_t^i, i = 1, \dots, n_e\}$
18. Estimate $\Sigma_{\tilde{r}vd}^{0:t}$ from $\mathbf{E}_{0:t}^{\tilde{r}v}, \mathbf{e}_t^d \rightarrow \hat{\Sigma}_{\tilde{r}vd}^{0:t}$
19. $\begin{bmatrix} \tilde{r}_{0:t}^{c(i)} \\ \mathbf{v}_{0:t}^{c(i)} \end{bmatrix} = \begin{bmatrix} \tilde{r}_{0:t}^{u(i)} \\ \mathbf{v}_{0:t}^{u(i)} \end{bmatrix} + \hat{\Gamma}_{\tilde{r}v,d}^{0:t} \hat{\Sigma}_d^{t-1} (d_t - d_t^i), i = 1, \dots, n_e$
20. **End iterate**
21. $\mathbf{E}_{0:T}^S = \{(\tilde{r}_{0:T}^{c(i)}, \mathbf{v}_{0:T}^{c(i)}), i = 1, \dots, n_e\}$
22. Select
23. For arbitrary $t \in [0, T]$, select corresponding ensemble e_t^S from $\mathbf{E}_{0:T}^S$
24. Estimate $\mu_{\tilde{r}v}^t, \Sigma_{\tilde{r}v}^t$ from $e_t^S \rightarrow \hat{\mu}_{\tilde{r}v}^t, \hat{\Sigma}_{\tilde{r}v}^t$
25. Assess
26. $\hat{f}(r_t | d_{0:T}) = [\Phi_n(A; \hat{\mu}_{\tilde{r}v}^t, \hat{\Sigma}_{\tilde{r}v}^t)]^{-1} \times \Phi_n(A; \hat{\mu}_{\tilde{r}v}^t + \hat{\Gamma}_{\tilde{r}v}^t (r - \hat{\mu}_{\tilde{r}v}^t), \hat{\Sigma}_{\tilde{r}v}^t) \times \varphi_n(r; \hat{\mu}_{\tilde{r}v}^t, \hat{\Sigma}_{\tilde{r}v}^t)$
27. End Algorithm

The ensemble $\mathbf{E}_{0:T}^S$ represents $[\tilde{r}_{0:T}, \mathbf{v}_{0:T} | d_{0:T}]$. To assess $f(r_t | d_{0:T}) = f(\tilde{r}_t | d_{0:T}, \mathbf{v} \in A)$, the sampling algorithm in [21] requires $E(\tilde{r}_t, \mathbf{v} | d_{0:T}) = \mu_{\tilde{r}v}^t$ and $\text{Cov}(\tilde{r}_t, \mathbf{v} | d_{0:T}) = \Sigma_{\tilde{r}v}^t$, which are estimated using the sub-ensemble e_t^S of $\mathbf{E}_{0:t}^S$.

References

1. Kalman, R.E. A new approach to linear filtering and prediction problems. *Trans. ASME-J. Basic Eng.* **1960**, *82*, 35–45. [[CrossRef](#)]
2. McElhoe, B.A. An Assessment of the Navigation and Course Corrections for a Manned Flyby of Mars or Venus. *IEEE Trans. Aerosp. Electron. Syst.* **1966**, *AES-2*, 613–623. [[CrossRef](#)]
3. Evensen, G. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.* **1994**, *99*, 10143. [[CrossRef](#)]
4. Myrseth, I.; Omre, H. The Ensemble Kalman Filter and Related Filters. In *Large Scale Inverse Problems and Quantification of Uncertainty*; John Wiley & Sons, Ltd.: London, UK, 2010; chapter 11, pp. 217–246.
5. Houtekamer, P.L.; Mitchell, H.L.; Pellerin, G.; Buehner, M.; Charron, M.; Spacek, L.; Hansen, B. Atmospheric Data Assimilation with an Ensemble Kalman Filter: Results with Real Observations. *Mon. Weather. Rev.* **2005**, *133*, 604–620. [[CrossRef](#)]
6. Sakov, P.; Oliver, D.; Bertino, L. An Iterative EnKF for Strongly Nonlinear Systems. *Mon. Weather. Rev.* **2012**, *140*, 1988–2004. [[CrossRef](#)]
7. Sklar, A. Random variables, joint distribution functions, and copulas. *Kybernetika* **1973**, *9*, 449–460.
8. Isaaks, E.H.; Srivastava, R.M. *Applied Geostatistics*; Oxford University Press: New York, NY, USA, 1989.
9. Bertino, L.; Evensen, G.; Wackernagel, H. Sequential Data Assimilation Techniques in Oceanography. *Int. Stat. Rev.* **2003**, *71*, 223–241. [[CrossRef](#)]
10. Simon, E.; Bertino, L. Application of the Gaussian anamorphosis to assimilation in a 3D coupled physical-ecosystem model of the North Atlantic with the EnKF: a twin experiment. *Ocean. Sci.* **2009**, *5*, 495–510. [[CrossRef](#)]
11. Xu, T.; Gomez-Hernandez, J. Characterization of non-Gaussian conductivities and porosities with hydraulic heads, solute concentrations, and water temperatures. *Water Resour. Res.* **2016**, *52*, 6111–6136. [[CrossRef](#)]
12. Gu, Y.; Oliver, D. An Iterative Ensemble Kalman Filter for Multiphase Fluid Flow Data Assimilation. *SPE J.* **2007**, *12*, 438–446. [[CrossRef](#)]
13. Evensen, G. Analysis of iterative ensemble smoothers for solving inverse problems. *Comput. Geosci.* **2018**, *22*, 885–908. [[CrossRef](#)]
14. Dovera, L.; Della Rossa, E. Multimodal ensemble Kalman filtering using Gaussian mixture models. *Comput. Geosci.* **2010**, *15*, 307–323. [[CrossRef](#)]
15. Rimstad, K.; Omre, H. Approximate posterior distributions for convolutional two-level hidden Markov models. *Comput. Stat. Data Anal.* **2013**, *58*, 187–200. [[CrossRef](#)]
16. Grana, D.; Fjeldstad, T.; Omre, H. Bayesian Gaussian Mixture Linear Inversion for Geophysical Inverse Problems. *Math. Geosci.* **2017**, *49*, 493–515. [[CrossRef](#)]
17. Besag, J. Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Ser.* **1974**, *36*, 192–236. [[CrossRef](#)]
18. Le Loc'h, G.; Beucher, H.; Galli, A.; Doligez, B. Improvement In The Truncated Gaussian Method: Combining Several Gaussian Functions. In *Proceedings of the ECMOR IV—4th European Conference on the Mathematics of Oil Recovery*; European Association of Geoscientists & Engineers: Røros, Norway, 1994. [[CrossRef](#)]
19. Oliver, D.; Chen, Y. Data Assimilation in Truncated Plurigaussian Models: Impact of the Truncation Map. *Math. Geosci.* **2018**, *50*, 867–893. [[CrossRef](#)]
20. Arellano-Valle, R.B.; Branco, M.D.; Genton, M.G. A unified view on skewed distributions arising from selections. *Can. J. Stat.* **2006**, *34*, 581–601. [[CrossRef](#)]
21. Omre, H.; Rimstad, K. Bayesian Spatial Inversion and Conjugate Selection Gaussian Prior Models. *arXiv* **2018**, arXiv:1812.01882.
22. Conjard, M.; Omre, H. Spatio-temporal Inversion using the Selection Kalman Model. *arXiv* **2020**, arXiv:stat.ME/2006.14343
23. Cappé, O.; Moulines, E.; Ryden, T. *Inference in Hidden Markov Models (Springer Series in Statistics)*; Springer-Verlag: Berlin/Heidelberg, Germany, 2005.
24. Moja, S.; Asfaw, Z.; Omre, H. Bayesian Inversion in Hidden Markov Models with Varying Marginal Proportions. *Math. Geosci.* **2018**, *51*, 463–484. [[CrossRef](#)]
25. Evensen, G. *Data Assimilation*; Springer-Verlag: Berlin Heidelberg, Germany, 2006; Volume 307. [[CrossRef](#)]

26. Evensen, G. The Ensemble Kalman filter: Theoretical Formulation and Practical Implementation. *Ocean. Dyn.* **2003**, *53*, 343–367. [[CrossRef](#)]
27. Burgers, G.; Van Leeuwen, P.J.; Evensen, G. On the Analysis Scheme in the Ensemble Kalman Filter. *Mon. Weather. Rev.* **1998**, *126*, 1719–1724. [[CrossRef](#)]
28. Hyndman, R. Computing and Graphing Highest Density Regions. *Am. Stat.* **1996**, *50*, 120–126.
29. Gaspari, G.; Cohn, S. Quarterly Journal of the Royal Meteorological Society. *J. Comput. Graph. Stat.* **1999**, *125*, 723–757.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).