Ensemble Kalman filter for multiscale inverse problems

Assyr Abdulle* Giacomo Garegnani* Andrea Zanoni*

Abstract

The abstract goes here.

AMS subject classifications. 62G05, 65N21, 74Q05.

Key words. Inverse problems, Multiscale modelling, Homogenization, Ensemble Kalman filter, Bayesian inference, Modelling error.

1 Introduction

2 Problem setting

Let us consider multiscale inverse problems of the following form

$$\text{find } u^\varepsilon \in X \text{ given observations } y = \mathcal{G}^\varepsilon(u^\varepsilon) + \eta \in Y, \tag{1}$$

where the parameter space X and the observation space Y are Hilbert spaces, $\mathcal{G}^\varepsilon \colon X \to Y$ is the forward response operator mapping the unknown u to the observation space, and $\eta \in Y$ is a source of additive noise, modelled as a mean zero random variable with covariance operator Γ. Usually, in applications, the forward operator $\mathcal{G}^\varepsilon$ can be written as $\mathcal{G}^\varepsilon = \mathcal{O} \circ \mathcal{S}^\varepsilon$, where $\mathcal{O} \colon H_0^1(\Omega) \to Y$ is an observation operator and $\mathcal{S}^\varepsilon$ is the solution operator of a multiscale partial differential equation (PDE). Let Ω be a bounded open domain, $\mathcal{S}^\varepsilon$ maps the unknown u to the solution p^ε of

$$\begin{cases} -\nabla \cdot (A_u^\varepsilon \nabla p^\varepsilon) = 0 & \text{in } \Omega, \\ p^\varepsilon = 0 & \text{on } \partial\Omega, \end{cases} \tag{2}$$

therefore $\mathcal{S}^\varepsilon \colon X \to H_0^1(\Omega)$ and $\mathcal{O} \colon H_0^1(\Omega) \to Y$, for which we assume that the following property holds true.

Assumption 1.
The observation operator $\mathcal{O} \colon H_0^1(\Omega) \to Y$ satisfies for all $p_1, p_2 \in H_0^1(\Omega)$

$$\|\mathcal{O}(p_1) - \mathcal{O}(p_2)\|_2 \leq m \|p_1 - p_2\|_{L^2(\Omega)},$$

where m is a positive constant.

Note that Assumption 1 is a stronger property than being Lipschitz. The tensor A_u^ε belongs to the class of parametrized multiscale tensors which admit explicit scale separation between slow and fast spatial variables

$$A_u^\varepsilon(x) = A\left(u(x), \frac{x}{\varepsilon}\right),$$

where the map $(t, x) \to A\left(t, \frac{x}{\varepsilon}\right)$ is assumed to be known and A is periodic in its second argument. If ε is small, we have to employ a fine discretization to resolve the smallest scale in the evaluation

*Institute of Mathematics, École Polytechnique Fédérale de Lausanne

of $\mathcal{G}^\varepsilon$, which in turn leads to a high computational cost. Considering also that the PDE has to be solved several times in the framework of inverse problems, this procedure can be infeasible.
Therefore, we apply the homogenization theory (see e.g. [4]), which ensures the existence of an homogenized tensor A^0, such that the solution p^0 of the problem

$$\begin{cases} -\nabla \cdot (A_u^0 \nabla p^0) = 0 & \text{in } \Omega, \\ p^0 = 0 & \text{on } \partial\Omega, \end{cases} \tag{3}$$

is the weak limit for $\varepsilon \to 0$ of the functions p^ε, i.e.

$$p^\varepsilon \rightharpoonup p^0 \qquad \text{in } H^1(\Omega).$$

Hence, the function p^0 is a good approximation of p^ε when the multiscale parameter ε is small and, in this case, the multiscale problem (2) can be replaced by its homogenized version (3). In order to solve numerically (3), we adopt a finite element discretization and we denote $\mathcal{G}_h^0 \colon \mathcal{O} \circ \mathcal{S}_h^0$ the forward operator which involves the solution of this problem. Thus, even if, as written in (1), the observations of the inverse problem come from a multiscale model represented by the forward operator $\mathcal{G}^\varepsilon$, inspired by [9], we employ the forward operator given by its discrete homogenized version $\mathcal{G}_h^0$ to solve the inverse problem. Finally, we introduce assumptions on the multiscale and homogenzed tensors A^ε and A^0.

Assumption 2.
The tensors A^ε for the multiscale problem (2) and A^0 for the homogenized problem (3) satisfy for all $u, u_1, u_2 \in X$

$$\begin{aligned} \|A^\varepsilon(u_1) - A^\varepsilon(u_2)\|_{L^\infty(\Omega;\mathbb{R}^{N\times N})} &\leq M \|u_1 - u_2\|_X, \\ \left\|A^0(u_1) - A^0(u_2)\right\|_{L^\infty(\Omega;\mathbb{R}^{N\times N})} &\leq M \|u_1 - u_2\|_X, \end{aligned}$$

and

$$A^\varepsilon(u)\xi \cdot \xi \geq \alpha_0 \|\xi\|_2^2, \quad A^0(u)\xi \cdot \xi \geq \alpha_0 \|\xi\|_2^2, \qquad \text{for all } \xi \in \mathbb{R}^N,$$

where M and α_0 are positive constants.

3 Ensemble Kalman filter for multiscale inverse problems

We solve the inverse problem by applying the ensemble Kalman filter EnKF, for which we present here a small introduction and refer to [7] for further details. Differently from our work, in [7] the authors apply the ensemble Kalman method to one-scale inverse problems.
Inverse problems are often ill-posed and they require regularization, which can be achieved through many techniques, both variational and Bayesian. In this method regularization is obtained by searching for the solution in a finite dimensional and compact subset $\mathcal{A}$ of X, which incorporates prior knowledge of u. In order to approximate the true unknown by means of the Kalman filter theory, we need to introduce artificial dynamics based on state augmentation.
We define the space $Z = X \times Y$ and the mapping $\Xi \colon Z \to Z$ by

$$\Xi(z) = \begin{bmatrix} u \\ \mathcal{G}_h^0(u) \end{bmatrix}, \quad \text{for} \quad z = \begin{bmatrix} u \\ v \end{bmatrix} \in Z,$$

which induces artificial dynamics as

$$z_{n+1} = \Xi(z_n).$$

We assume that data related to the artificial dynamics has the form

$$y_{n+1} = H z_{n+1} + \eta_{n+1},$$

where $H \colon Z \to Y$ is a projection operator defined by $H = \begin{bmatrix} 0 & I \end{bmatrix}$ and $\{\eta_n\}_{n\in\mathbb{N}}$ is an i.i.d. sequence of random variables distributed as $\eta_n \sim \mathcal{N}(0, \Gamma)$. Consequently we get

$$y_{n+1} = H\Xi(z_n) + \eta_{n+1} = \mathcal{G}_h^0(u_n) + \eta_{n+1}.$$

The ensemble Kalman method uses an ensemble of states of dimension J, which is sequentially updated by means of the Kalman formula for N iterations. The last detail missing is the initial ensemble of particles $\{z_0^{(j)}\}_{j=1}^J$. This first guess can be defined by constructing an ensemble $\{\psi^{(j)}\}_{j=1}^J$ in X and taking

$$z_0^{(j)} = \begin{bmatrix} \psi^{(j)} \\ \mathcal{G}_h^0(\psi^{(j)}) \end{bmatrix}.$$

The initial ensemble is related to the definition of the space $\mathcal{A}$, which incorporates prior knowledge of the solution. We assume that the prior knowledge is a probability measure denoted by μ_0, then $\{\psi^{(j)}\}_{j=1}^J$ is constructed drawing J samples from this distribution. Finally, the set $\mathcal{A}$ is given by the subspace generated by the initial ensemble members

$$\mathcal{A} = \text{span } \{\psi^{(j)}\}_{j=1}^J.$$

Each iteration of the ensemble Kalman method is divided in two steps: prediction and analysis. In the first step we map the current ensemble of particles $\{z_n^{(j)}\}_{j=1}^J$ into the data space, introducing information about the forward model $\mathcal{G}_h^0$, which is contained in the definition of Ξ, and obtaining prior estimates of the state variables

$$\hat{z}_{n+1}^{(j)} = \Xi(z_n^{(j)}).$$

In the second step the ensemble of particles $\{z_n^{(j)}\}_{j=1}^J$ is updated comparing the mapped ensemble $\{H\hat{z}_{n+1}^{(j)}\}_{j=1}^J$ with versions of the data perturbed with noise $\{y_{n+1}^{(j)}\}_{j=1}^J$ via the standard Kalman update formula

$$z_{n+1}^{(j)} = \hat{z}_{n+1}^{(j)} + K_{n+1}(y_{n+1}^{(j)} - H\hat{z}_{n+1}^{(j)}) = (I - K_{n+1}H)\hat{z}_{n+1}^{(j)} + K_{n+1}y_{n+1}^{(j)},$$

where $y_{n+1}^{(j)} = y + \eta_{n+1}^{(j)}$ and

$$K_{n+1} = C_{n+1}H^*(HC_{n+1}H^* + \Gamma)^{-1},$$

where H^* is the adjoint operator of H and

$$C_{n+1} = \frac{1}{J}\sum_{j=1}^J (\hat{z}_{n+1}^{(j)} - \bar{z}_{n+1})(\hat{z}_{n+1}^{(j)} - \bar{z}_{n+1})^T = \frac{1}{J}\sum_{j=1}^J \hat{z}_{n+1}^{(j)}(\hat{z}_{n+1}^{(j)})^T - \bar{z}_{n+1}\bar{z}_{n+1}^T.$$

Note that the update equation for the unknown is

$$u_{n+1}^{(j)} = u_n^{(j)} + C_{n+1}^{up}(C_{n+1}^{pp} + \Gamma)^{-1}(y_{n+1}^{(j)} - \mathcal{G}(u_n^{(j)})).$$

Finally, the solution is computed by averaging over the ensemble of particles

$$u_{\text{EnKF}} = \frac{1}{J}\sum_{j=1}^J H^\perp z_N^{(j)} = \frac{1}{J}\sum_{j=1}^J u_N^{(j)},$$

where $H^\perp \colon Z \to Y$ is a projection operator defined by $H^\perp = \begin{bmatrix} I & 0 \end{bmatrix}$.

Remark 1.

The computational cost of the iterative ensemble Kalman algorithm can be measured in terms of the number of evaluations of the forward operator, indeed the operations to compute the Kalman gain and update the ensemble of particles are negligible with respect to the resolution of the homogenized problem, which is the most expensive task. Therefore, the complexity of the algorithm is $\mathcal{O}(JN)$, where J is the dimension of the ensemble and N is the number of iterations. Moreover, note that the ensemble Kalman method is easily parallelizable. Indeed, at each iteration, we apply the forward operator to every particle in the ensemble individually. Hence, if we call n_{CPU} the number of central processing units available and we neglect the cost of exchanging of information between processors, the computational cost of the iterative ensemble Kalman method becomes

$$\mathcal{O}\left(\frac{J}{n_{\text{CPU}}}N\right).$$

Thus, if we have a big number of CPUs available, a parallelized version of the code reduces significantly the computational time for the resolution of the inverse problem.

Following [13], the ensemble Kalman filter can be interpreted also from a Bayesian point of view. Given a prior distribution μ_0 for the unknown, we want to estimate the conditional posterior distribution μ of the unknown given the data, which is distributed according to the measure

$$\mu(du) = \frac{1}{Z} e^{-\Phi(u;y)} \mu_0(du),$$

where Z is a normalization constant and $\Phi(u;y)$ is the least squares functional

$$\Phi(u;y) = \frac{1}{2} \left\| \Gamma^{-1/2}(y - \mathcal{G}(u)) \right\|_2^2.$$

We follow an iterative procedure of N steps and we define the intermediate measures

$$\mu_n(du) = \frac{1}{Z_n} e^{-n\Delta\Phi(u;y)} \mu_0(du),$$

where $\Delta = 1/N$. Note that $\mu_N = \mu$ is the desired final measure. Then we obtain

$$\mu_{n+1}(du) = \frac{Z_n}{Z_{n+1}} e^{-h\Phi(u;y)} \mu_n(du).$$

The posterior distribution μ_n is approximated by a sum of Dirac masses centered in the particles of the ensemble at the n-th iteration

$$\mu_n \simeq \frac{1}{J} \sum_{j=1}^{J} \delta_{u_n^{(j)}}. \tag{4}$$

The mapping of particles at time n into those at time $n+1$ is given by the ensemble Kalman filter update formula, where Γ has been replaced by $\Delta^{-1}\Gamma$

$$u_{n+1}^{(j)} = u_n^{(j)} + C^{up}(u_n)(C^{pp}(u_n) + \Delta^{-1}\Gamma)^{-1}(y_{n+1}^{(j)} - \mathcal{G}(u_n^{(j)})).$$

Finally, let us introduce an assumption on the algorithm, which will be employed in the analysis.

Assumption 3.
The algorithm is stable, in the sense that all the particles in the ensemble at each iteration lie in the ball $B_R(u^*)$ for some $R > 0$ sufficiently big, where u^* is the true unknown.

4 Convergence analysis

In this section we show the convergence of the ensemble of particles generated by the EnKF algorithm using the FEM discretization of the homogenized problem as forward operator to the ensemble of particles generated by the EnKF algorithm using the multiscale problem as forward operator. Moreover, from the Bayesian perspective, we show the convergence of their posterior distributions. Under further assumptions, we also provide a rate of convergence. The whole analysis is done in the finite dimensional case, where the ensemble of particles lie in $\mathbb{R}^M$ and the observations are in $\mathbb{R}^L$, but it can be generalized to the infinite dimensional setting. Therefore, the forward operators under consideration are $\mathcal{G}^\varepsilon \colon \mathbb{R}^M \to \mathbb{R}^L$, $\mathcal{G}^\varepsilon = \mathcal{O} \circ \mathcal{S}^\varepsilon$ and $\mathcal{G}_h^0 \colon \mathbb{R}^M \to \mathbb{R}^L$, $\mathcal{G}_h^0 = \mathcal{O} \circ \mathcal{S}_h^0$. We also introduce the forward operator $\mathcal{G}^0 \colon \mathbb{R}^M \to \mathbb{R}^L$, $\mathcal{G}^0 = \mathcal{O} \circ \mathcal{S}^0$, where $\mathcal{S}^0$ is the continuous solution operator of problem (3), which maps the unknown u to the solution p^0.
Given an ensemble of particles $u = \{u^{(j)}\}_{j=1}^J$, we define the ensemble norm in the following way

$$\|u\| := \frac{1}{J} \sum_{j=1}^{J} \left\| u^{(j)} \right\|_2, \tag{5}$$

and note that it satisfies the properties of a norm

- $\|u\| \geq 0$,

- $\|u\| = 0$ if and only if $\left\|u^{(j)}\right\|_2 = 0$ for all $j = 1, \dots, J$, which is true if and only if $u^{(j)} = 0$ for all $j = 1, \dots, J$, which is equivalent to $u = 0$,
- $\|\lambda u\| = \frac{1}{J}\sum_{j=1}^{J}\left\|\lambda u^{(j)}\right\|_2 = |\lambda|\,\frac{1}{J}\sum_{j=1}^{J}\left\|u^{(j)}\right\|_2 = |\lambda|\,\|u\|$,
- $\|u+v\| := \frac{1}{J}\sum_{j=1}^{J}\left\|u^{(j)}+v^{(j)}\right\|_2 \leq \frac{1}{J}\sum_{j=1}^{J}\left[\left\|u^{(j)}\right\|_2 + \left\|v^{(j)}\right\|_2\right] = \|u\| + \|v\|$.

In the last point, the sum of two ensembles is intended to be the ensemble whose particles are the sum of the single particles in each ensemble. In the next sections we analyse separately the point-wise convergence of the particles and the convergence of the posterior distributions.

4.1 Convergence of the point estimate

The main result is stated in Theorem 1.

Theorem 1.
Let $u^0_{N,h} = \{u^{0^{(j)}}_{N,h}\}_{j=1}^J$, $u^\varepsilon_N = \{u^{\varepsilon^{(j)}}_N\}_{j=1}^J$ be the ensembles of particles at the last iteration of the iterative ensemble Kalman filter for the forward operators $\mathcal{G}^0_h$ and $\mathcal{G}^\varepsilon$ respectively. Then, under Assumption 1, Assumption 2 and Assumption 3, we have

$$\mathbb{E}\left[\left\|u^\varepsilon_N - u^0_{N,h}\right\|\right] \to 0 \qquad \text{as } \varepsilon, h \to 0.$$

In particular, if the exact solution p^0 of the homogenized problem (3) is in $H^{q+1}(\Omega)$ with $q \geq 1$, $A^0 \in C^q(\Omega; \mathbb{R}^{N\times N})$, $f \in H^{q-1}(\Omega)$, $\partial\Omega \in C^{q+1}$, and we use polynomials of degree r for the finite element basis, we have the following rate of convergence

$$\mathbb{E}\left[\left\|u^\varepsilon_N - u^0_{N,h}\right\|\right] \leq C(\varepsilon + h^{s+1}),$$

where $s = \min\{r, q\}$.

We analyse separately the multiscale convergence and the finite element discretization convergence. We first show the convergence of the ensemble of particles generated by the EnKF algorithm using the multiscale problem as forward operator to the ensemble of particles generated by the EnKF algorithm using the homogenized problem as forward operator. Then we show the convergence of the ensemble of particles generated by the EnKF algorithm using a FEM discretization of the homogenized problem as forward operator to the ensemble of particles generated by the EnKF algorithm using the true homogenized problem as forward operator.
Let $e\colon \mathbb{R}\times\mathbb{R}^M \to \mathbb{R}$ be a function modelling the error due to the replacement of the multiscale problem with the homogenized one

$$e(\varepsilon, u) = \frac{1}{J}\sum_{j=1}^{J}\left\|\mathcal{G}^\varepsilon(u^{(j)}) - \mathcal{G}^0(u^{(j)})\right\|_2. \tag{6}$$

and $\tilde{e}\colon \mathbb{R}\times\mathbb{R}^M \to \mathbb{R}$ be the error induced by the use of a FEM discretization for the homogenized problem

$$\tilde{e}(h, u) = \frac{1}{J}\sum_{j=1}^{J}\left\|\mathcal{G}^0_h(u^{(j)}) - \mathcal{G}^0(u^{(j)})\right\|_2. \tag{7}$$

Before proving the main theorem, we need some preliminary results.

Lemma 1.
Let A and B be square invertible matrices, then

$$\left\|A^{-1} - B^{-1}\right\|_2 \leq \left\|A^{-1}\right\|_2\left\|B^{-1}\right\|_2\|A - B\|_2.$$

Proof. The proof can be found in the appendix. □

Lemma 2.
Let A and B be square, symmetric matrices, such that A is positive semidefinite and B is positive definite, then

$$\left\|(A+B)^{-1}\right\|_2 \leq \left\|B^{-1}\right\|_2 .$$

Proof. The proof can be found in the appendix. □

Lemma 3.
Let $\mathcal{G}\colon \mathbb{R}^M \to \mathbb{R}^L$, $\mathcal{G} = \mathcal{O} \circ \mathcal{S}$ be the forward operator defined by the composition of an observation operator $\mathcal{O}\colon H_0^1(\Omega) \to \mathbb{R}^L$ and a solution operator $\mathcal{S}\colon \mathbb{R}^M \to H_0^1(\Omega)$, which assigns to $u \in \mathbb{R}^M$ the solution $p \in H_0^1(\Omega)$ of the elliptic partial differential equation

$$\begin{cases} -\nabla \cdot (A(u)\nabla p) = f & \text{in } \Omega, \\ p = 0 & \text{on } \partial\Omega, \end{cases} \tag{8}$$

where $\Omega \subset \mathbb{R}^N$ is a domain, $A(u) \in L^\infty(\Omega; \mathbb{R}^{N\times N})$ and $f \in L^2(\Omega)$. Let $\mathcal{O}$ be Lipschitz and A such that

$$\|A(u_1) - A(u_2)\|_{L^\infty(\Omega;\mathbb{R}^{N\times N})} \leq M \|u_1 - u_2\|_2 ,$$

and

$$A(u)\xi \cdot \xi \geq \alpha \|\xi\|_2^2 \qquad \text{for all } \xi \in \mathbb{R}^N,$$

where M and α are positive constants. Then $\mathcal{G}$ is Lipschitz.

Proof. Let $u_1, u_2 \in \mathbb{R}^M$, then the weak formulations of problem (8) for these two values are

$$\int_\Omega A(u_1)\nabla p_1 \cdot \nabla v = \int_\Omega fv \qquad \text{and} \qquad \int_\Omega A(u_2)\nabla p_2 \cdot \nabla v = \int_\Omega fv,$$

for all $v \in H_0^1(\Omega)$. Hence we have

$$\int_\Omega A(u_1)\nabla p_1 \cdot \nabla v - \int_\Omega A(u_1)\nabla p_2 \cdot \nabla v + \int_\Omega A(u_1)\nabla p_2 \cdot \nabla v - \int_\Omega A(u_2)\nabla p_2 \cdot \nabla v = 0,$$

which is equivalent to

$$\int_\Omega A(u_1)(\nabla p_1 - \nabla p_2) \cdot \nabla v = -\int_\Omega (A(u_1) - A(u_2))\nabla p_2 \cdot \nabla v.$$

Take $v = p_1 - p_2 \in H_0^1(\Omega)$. Then, using the hypotheses on A and the Hölder inequality, we obtain

$$\begin{aligned} \alpha \|\nabla p_1 - \nabla p_2\|_{L^2(\Omega;\mathbb{R}^N)}^2 &\leq \int_\Omega A(u_1)(\nabla p_1 - \nabla p_2) \cdot (\nabla p_1 - \nabla p_2) \\ &= -\int_\Omega (A(u_1) - A(u_2))\nabla p_2 \cdot (\nabla p_1 - \nabla p_2) \\ &\leq \left| \int_\Omega (A(u_1) - A(u_2))\nabla p_2 \cdot (\nabla p_1 - \nabla p_2) \right| \\ &\leq \|A(u_1) - A(u_2)\|_{L^\infty(\Omega;\mathbb{R}^{N\times N})} \|\nabla p_2\|_{L^2(\Omega;\mathbb{R}^N)} \|\nabla p_1 - \nabla p_2\|_{L^2(\Omega;\mathbb{R}^N)} \\ &\leq M \|u_1 - u_2\|_2 \|\nabla p_2\|_{L^2(\Omega;\mathbb{R}^N)} \|\nabla p_1 - \nabla p_2\|_{L^2(\Omega;\mathbb{R}^N)} , \end{aligned}$$

which implies

$$\|\nabla p_1 - \nabla p_2\|_{L^2(\Omega;\mathbb{R}^N)} \leq \frac{M}{\alpha} \|\nabla p_2\|_{L^2(\Omega;\mathbb{R}^N)} \|u_1 - u_2\|_2 . \tag{9}$$

Now we still have to bound $\|\nabla p_2\|_{L^2(\Omega;\mathbb{R}^N)}$, so we consider the weak formulation of problem (8) for the value u_2 and we take $v = p_2$, then we have

$$\begin{aligned} \alpha \|\nabla p_2\|_{L^2(\Omega;\mathbb{R}^N)}^2 &\leq \int_\Omega A(u_2)\nabla p_2 \cdot \nabla p_2 \\ &= \int_\Omega f p_2 \\ &\leq \|f\|_{L^2(\Omega)} \|p_2\|_{L^2(\Omega)} , \end{aligned}$$

and using Poincaré inequality with constant C_p we obtain

$$\alpha \|\nabla p_2\|^2_{L^2(\Omega;\mathbb{R}^N)} \leq C_p \|f\|_{L^2(\Omega)} \|\nabla p_2\|_{L^2(\Omega;\mathbb{R}^N)}.$$

Thus we derive

$$\|\nabla p_2\|_{L^2(\Omega;\mathbb{R}^N)} \leq \frac{C_p}{\alpha} \|f\|_{L^2(\Omega)},$$

and from (9) we obtain

$$\|\nabla p_1 - \nabla p_2\|_{L^2(\Omega;\mathbb{R}^N)} \leq \frac{MC_p}{\alpha^2} \|f\|_{L^2(\Omega)} \|u_1 - u_2\|_2 = L_{\mathcal{S}} \|u_1 - u_2\|_2,$$

which shows that $\mathcal{S}$ is Lipschitz with constant

$$L_{\mathcal{S}} = \frac{MC_p}{\alpha^2} \|f\|_{L^2(\Omega)}.$$

Finally, $\mathcal{G}$ is the composition of two Lipschitz operators, so it is Lipschitz. Indeed, letting $L_{\mathcal{O}}$ be the Lipschitz constant of the observation operator $\mathcal{O}$, we have

$$\begin{aligned}
\|\mathcal{G}(u_1) - \mathcal{G}(u_2)\|_2 &= \|\mathcal{O}(\mathcal{S}(u_1)) - \mathcal{O}(\mathcal{S}(u_2))\|_2 \\
&= \|\mathcal{O}(p_1) - \mathcal{O}(p_2)\|_2 \\
&\leq L_{\mathcal{O}} \|\nabla p_1 - \nabla p_2\|_{L^2(\Omega;\mathbb{R}^N)} \\
&\leq L_{\mathcal{O}} L_{\mathcal{S}} \|u_1 - u_2\|_2,
\end{aligned}$$

which concludes the proof. □

Lemma 4.
Let Ω be a domain and $\{p^\varepsilon\}$ be a sequence in $H^1_0(\Omega)$ such that

$$p^\varepsilon \rightharpoonup p^0 \qquad \text{in } H^1_0(\Omega),$$

then

$$p^\varepsilon \to p^0 \qquad \text{in } L^2(\Omega).$$

Lemma 5.
Let e be defined as in (6). Under Assumption 1, we have for all $u \in \mathbb{R}^M$

$$e(\varepsilon, u) \to 0 \quad \text{as } \varepsilon \to 0.$$

Moreover, if the solution of the homogenized problem (3) is sufficiently regular, namely $p^0 \in H^2(\Omega)$, then we have a linear rate of convergence, indeed

$$e(\varepsilon, u) \leq K\varepsilon.$$

Proof. The proof can be found in the appendix. □

Lemma 6.
Let $C^{up}(u)$ and $C^{pp}(u)$ be defined as

$$C^{up}(u) = \frac{1}{J} \sum_{j=1}^{J} (u^{(j)} - \bar{u})(\mathcal{G}(u^{(j)}) - \bar{\mathcal{G}})^T \quad \in \mathbb{R}^{M \times L},$$

$$C^{pp}(u) = \frac{1}{J} \sum_{j=1}^{J} (\mathcal{G}(u^{(j)}) - \bar{\mathcal{G}})(\mathcal{G}(u^{(j)}) - \bar{\mathcal{G}})^T \quad \in \mathbb{R}^{L \times L},$$

where

$$\bar{u} = \frac{1}{J} \sum_{j=1}^{J} u^{(j)} \quad \in \mathbb{R}^M,$$

$$\bar{\mathcal{G}} = \frac{1}{J} \sum_{j=1}^{J} \mathcal{G}(u^{(j)}) \quad \in \mathbb{R}^L,$$

and $\mathcal{G}$ is L-Lipschitz. Then there exist four constants C_1, C_2, C_3 and C_4 such that

- $\|C^{up}(u)\|_2 \leq C_1$
- $\|C^{pp}(u)\|_2 \leq C_2$
- $\|C^{up}(u_1) - C^{up}(u_2)\|_2 \leq C_3 \|u_1 - u_2\|$
- $\|C^{pp}(u_1) - C^{pp}(u_2)\|_2 \leq C_4 \|u_1 - u_2\|$

for all the ensembles $u, u_1, u_2 \subset B_R(u^*)$, where $R > 0$ is a constant and $u^* \in \mathbb{R}^M$ is a fixed value with $\|u^*\|_2 = g$ and $\|\mathcal{G}(u^*)\|_2 = G$.

Proof. The proof can be found in the appendix. □

We now present the main result about multiscale convergence, where we show the convergence of the ensemble of particles generated by the EnKF algorithm using the multiscale problem as forward operator to the ensemble of particles generated by the EnKF algorithm using the homogenized problem as forward operator.

Proposition 1.
Let $u_N^0 = \{u_N^{0^{(j)}}\}_{j=1}^J$, $u_N^\varepsilon = \{u_N^{\varepsilon^{(j)}}\}_{j=1}^J$ be the ensembles of particles at the last iteration of the iterative ensemble Kalman filter for the forward operators $\mathcal{G}^0$ and $\mathcal{G}^\varepsilon$ respectively. Then, under Assumption 1, Assumption 2 and Assumption 3, we have

$$\mathbb{E}\left[\left\|u_N^\varepsilon - u_N^0\right\|\right] \to 0 \qquad \text{as } \varepsilon \to 0.$$

Moreover, if the solution of the homogenized problem (3) is sufficiently regular, namely $p^0 \in H^2(\Omega)$, then the error decreases linearly

$$\mathbb{E}\left[\left\|u_N^\varepsilon - u_N^0\right\|\right] \leq K_1 \varepsilon.$$

Proof. First, by the hypothesis on $\mathcal{O}$ and applying Poincaré inequality with constant C_p we have

$$\|\mathcal{O}(p_1) - \mathcal{O}(p_2)\|_2 \leq m \|p_1 - p_2\|_{L^2(\Omega)} \leq mC_p \|\nabla p_1 - \nabla p_2\|_{L^2(\Omega;\mathbb{R}^N)},$$

which shows that $\mathcal{O}$ is Lipschitz with constant mC_p and, applying Lemma 3, we deduce that both $\mathcal{G}^0$ and $\mathcal{G}^\varepsilon$ are Lipschitz with constant $L_\mathcal{G}$ independent of ε.
One step of the iterative ensemble Kalman filter for both the forward operators is

$$u_{n+1}^{\varepsilon^{(j)}} = u_n^{\varepsilon^{(j)}} + C^{up}(u_n^\varepsilon)(C^{pp}(u_n^\varepsilon) + \Gamma)^{-1}(y_{n+1} - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}})), \tag{10}$$

$$u_{n+1}^{0^{(j)}} = u_n^{0^{(j)}} + C^{up}(u_n^0)(C^{pp}(u_n^0) + \Gamma)^{-1}(y_{n+1} - \mathcal{G}^0(u_n^{0^{(j)}})). \tag{11}$$

Let $\{\psi^{(j)}\}_{j=1}^J$ be the initial ensemble, then the first step of the algorithm is

$$u_1^{\varepsilon^{(j)}} = \psi^{(j)} + C^{up}(\psi)(C^{pp}(\psi) + \Gamma)^{-1}(y_1^{(j)} - \mathcal{G}^\varepsilon(\psi^{(j)})), \tag{12}$$

$$u_1^{0^{(j)}} = \psi^{(j)} + C^{up}(\psi)(C^{pp}(\psi) + \Gamma)^{-1}(y_1^{(j)} - \mathcal{G}^0(\psi^{(j)})). \tag{13}$$

Using (12) and (13), we compute the expected error at the first iteration

$$\begin{aligned}
\mathbb{E}\left[\left\|u_1^\varepsilon - u_1^0\right\|\right] &= \frac{1}{J}\sum_{j=1}^J \mathbb{E}\left[\left\|u_1^{\varepsilon^{(j)}} - u_1^{0^{(j)}}\right\|_2\right] \\
&= \frac{1}{J}\sum_{j=1}^J \mathbb{E}\left[\left\|C^{up}(\psi)(C^{pp}(\psi) + \Gamma)^{-1}(\mathcal{G}^0(\psi^{(j)}) - \mathcal{G}^\varepsilon(\psi^j))\right\|_2\right] \\
&\leq \frac{1}{J}\sum_{j=1}^J \mathbb{E}\left[\|C^{up}(\psi)\|_2 \left\|(C^{pp}(\psi) + \Gamma)^{-1}\right\|_2 \left\|\mathcal{G}^0(\psi^{(j)}) - \mathcal{G}^\varepsilon(\psi^j)\right\|_2\right].
\end{aligned}$$

Thanks to Lemma 6 and Lemma 2 we have the following bounds

$$\begin{aligned}&\left\|C^{up}(\psi)\right\|_2 \leq C_1,\\ &\left\|(C^{pp}(\psi)+\Gamma)^{-1}\right\|_2 \leq \left\|\Gamma^{-1}\right\|_2.\end{aligned}$$

We recall the definition of e given in (6)

$$e(\varepsilon,\psi) = \frac{1}{J}\sum_{j=1}^{J}\left\|\mathcal{G}^0(\psi^{(j)}) - \mathcal{G}^\varepsilon(\psi^{(j)})\right\|_2,$$

then, defining $\beta = C_1\left\|\Gamma^{-1}\right\|_2$, we obtain

$$\mathbb{E}\left[\left\|u_1^\varepsilon - u_1^0\right\|\right] \leq C_1\left\|\Gamma^{-1}\right\|_2\mathbb{E}[e(\varepsilon,\psi)] = \beta\mathbb{E}[e(\varepsilon,\psi)]. \tag{14}$$

Subtracting (10) and (11) we have

$$\begin{aligned}\mathbb{E}\left[\left\|u_{n+1}^\varepsilon - u_{n+1}^0\right\|\right] &= \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\left[\left\|u_{n+1}^{\varepsilon^{(j)}} - u_{n+1}^{0^{(j)}}\right\|_2\right]\\ &= \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\Big[\Big\|u_n^{\varepsilon^{(j)}} + C^{up}(u_n^\varepsilon)(C^{pp}(u_n^\varepsilon)+\Gamma)^{-1}(y_{n+1}^{(j)} - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}}))\\ &\qquad - u_n^{0^{(j)}} - C^{up}(u_n^0)(C^{pp}(u_n^0)+\Gamma)^{-1}(y_{n+1}^{(j)} - \mathcal{G}^0(u_n^{0^{(j)}}))\Big\|_2\Big],\end{aligned}$$

and using the triangle inequality we obtain

$$\begin{aligned}&\mathbb{E}\left[\left\|u_{n+1}^\varepsilon - u_{n+1}^0\right\|\right]\\ &\leq \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\left[\left\|u_n^{\varepsilon^{(j)}} - u_n^{0^{(j)}}\right\|_2\right] &(15)\\ &\quad + \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\left[\left\|C^{up}(u_n^\varepsilon) - C^{up}(u_n^0)\right\|_2\left\|(C^{pp}(u_n^\varepsilon)+\Gamma)^{-1}\right\|_2\left\|y_{n+1}^{(j)} - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}})\right\|_2\right] &(16)\\ &\quad + \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\left[\left\|C^{up}(u_n^0)\right\|_2\left\|(C^{pp}(u_n^\varepsilon)+\Gamma)^{-1} - (C^{pp}(u_n^0)+\Gamma)^{-1}\right\|_2\left\|y_{n+1}^{(j)} - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}})\right\|_2\right] &(17)\\ &\quad + \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\left[\left\|C^{up}(u_n^0)\right\|_2\left\|(C^{pp}(u_n^0)+\Gamma)^{-1}\right\|_2\left\|\mathcal{G}^0(u_n^{0^{(j)}}) - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}})\right\|_2\right]. &(18)\end{aligned}$$

Equation (15) can be written as

$$\frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\left[\left\|u_n^{\varepsilon^{(j)}} - u_n^{0^{(j)}}\right\|_2\right] = \mathbb{E}\left[\left\|u_n^\varepsilon - u_n^0\right\|\right]. \tag{19}$$

The first part of (16) can be bounded using Lemma 6

$$\left\|C^{up}(u_n^\varepsilon) - C^{up}(u_n^0)\right\|_2 \leq C_3\left\|u_n^\varepsilon - u_n^0\right\|,$$

and for the second part we use Lemma 2

$$\left\|(C^{pp}(u_n^\varepsilon)+\Gamma)^{-1}\right\|_2 \leq \left\|\Gamma^{-1}\right\|_2.$$

Regarding the last part of (16), using the definition of $y_{n+1}^{(j)}$ we have

$$\begin{aligned}\left\|y_{n+1}^{(j)} - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}})\right\|_2 &= \left\|y + \xi_{n+1}^{(j)} - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}})\right\|_2\\ &= \left\|\mathcal{G}^\varepsilon(u^*) + \xi_{n+1}^{(j)} - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}})\right\|_2\\ &\leq \left\|\mathcal{G}^\varepsilon(u^*) - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}})\right\|_2 + \left\|\xi_{n+1}^{(j)}\right\|_2,\end{aligned}$$

and since $\mathcal{G}^\varepsilon$ is Lipschitz we obtain

$$\begin{aligned}\left\|y_{n+1}^{(j)} - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}})\right\|_2 &\leq L_\mathcal{G}\left\|u^* - u_n^{\varepsilon^{(j)}}\right\|_2 + \left\|\xi_{n+1}^{(j)}\right\|_2 \\ &\leq L_\mathcal{G} R + \left\|\xi_{n+1}^{(j)}\right\|_2.\end{aligned}$$

Thus (16) can be bounded by

$$\frac{1}{J} C_3 \left\|\Gamma^{-1}\right\|_2 \sum_{j=1}^{J} \mathbb{E}\left[\left\|u_n^\varepsilon - u_n^0\right\| \left(L_\mathcal{G} R + \left\|\xi_{n+1}^{(j)}\right\|_2\right)\right]$$

and, since the noise is i.i.d. and independent of the ensembles, we obtain

$$C_3 \left\|\Gamma^{-1}\right\|_2 \left(L_\mathcal{G} R + \mathbb{E}[\|\xi\|_2]\right)\mathbb{E}\left[\left\|u_n^\varepsilon - u_n^0\right\|\right].$$

Moreover, $\xi \sim \mathcal{N}(0, \Gamma)$, therefore we have

$$\mathbb{E}[\|\xi\|_2] \leq \sqrt{\mathbb{E}[\|\xi\|_2^2]} = \sqrt{\mathrm{tr}(\Gamma)},$$

and defining $\alpha_1 = C_3 \left\|\Gamma^{-1}\right\|_2 \left(L_\mathcal{G} R + \sqrt{\mathrm{tr}(\Gamma)}\right)$, the bound for (16) is

$$\alpha_1 \mathbb{E}\left[\left\|u_n^\varepsilon - u_n^0\right\|\right]. \tag{20}$$

The first part of (17) can be bounded using Lemma 6

$$\left\|C^{up}(u_n^0)\right\|_2 \leq C_1,$$

and for the second part we use Lemma 1, Lemma 2 and Lemma 6

$$\begin{aligned}&\left\|(C^{pp}(u_n^\varepsilon) + \Gamma)^{-1} - (C^{pp}(u_n^0) + \Gamma)^{-1}\right\|_2 \\ &\leq \left\|(C^{pp}(u_n^\varepsilon) + \Gamma)^{-1}\right\|_2 \left\|(C^{pp}(u_n^0) + \Gamma)^{-1}\right\|_2 \left\|C^{pp}(u_n^\varepsilon) - C^{pp}(u_n^0)\right\|_2 \\ &\leq C_4 \left\|\Gamma^{-1}\right\|_2^2 \left\|u_n^\varepsilon - u_n^0\right\|.\end{aligned}$$

The third part of (17) is equal to the third part of (16), thus (17) can be bounded by

$$\frac{1}{J} C_1 C_4 \left\|\Gamma^{-1}\right\|_2^2 \sum_{j=1}^{J} \mathbb{E}\left[\left\|u_n^\varepsilon - u_n^0\right\| \left(L_\mathcal{G} R + \left\|\xi_{n+1}^{(j)}\right\|_2\right)\right],$$

and repeating the previous procedure, defining $\alpha_2 = C_1 C_4 \left\|\Gamma^{-1}\right\|_2^2 \left(L_\mathcal{G} R + \sqrt{\mathrm{tr}(\Gamma)}\right)$, we obtain the final bound for (17)

$$\alpha_2 \mathbb{E}\left[\left\|u_n^\varepsilon - u_n^0\right\|\right]. \tag{21}$$

The first part of (18) is equal to the first part of (17) and for the second part we use Lemma 2

$$\left\|(C^{pp}(u_n^0) + \Gamma)^{-1}\right\|_2 \leq \left\|\Gamma^{-1}\right\|_2.$$

Regarding the third part of (18) we use the triangle inequality and the fact that $\mathcal{G}^\varepsilon$ is Lipschitz with constant $L_\mathcal{G}$

$$\begin{aligned}\left\|\mathcal{G}^0(u_n^{0^{(j)}}) - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}})\right\|_2 &\leq \left\|\mathcal{G}^0(u_n^{0^{(j)}}) - \mathcal{G}^\varepsilon(u_n^{0^{(j)}})\right\|_2 + \left\|\mathcal{G}^\varepsilon(u_n^{0^{(j)}}) - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}})\right\|_2 \\ &\leq \left\|\mathcal{G}^0(u_n^{0^{(j)}}) - \mathcal{G}^\varepsilon(u_n^{0^{(j)}})\right\|_2 + L_\mathcal{G}\left\|u_n^{0^{(j)}} - u_n^{\varepsilon^{(j)}}\right\|_2.\end{aligned}$$

Hence a bound for (18) is

$$C_1 \left\|\Gamma^{-1}\right\|_2 \mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}\left\|\mathcal{G}^0(u_n^{0^{(j)}}) - \mathcal{G}^\varepsilon(u_n^{0^{(j)}})\right\|_2 + L_{\mathcal{G}}\frac{1}{J}\sum_{j=1}^{J}\left\|u_n^{0^{(j)}} - u_n^{\varepsilon^{(j)}}\right\|_2\right],$$

which is equal to

$$C_1 \left\|\Gamma^{-1}\right\|_2 \mathbb{E}\left[e(\varepsilon, u_n^0)\right] + C_1 \left\|\Gamma^{-1}\right\|_2 L_{\mathcal{G}}\mathbb{E}\left[\left\|u_n^0 - u_n^\varepsilon\right\|\right],$$

and defining $\alpha_3 = C_1 \left\|\Gamma^{-1}\right\|_2 L_{\mathcal{G}}$ and $\gamma = C_1 \left\|\Gamma^{-1}\right\|_2$ we have the final bound for (18)

$$\alpha_3\mathbb{E}\left[\left\|u_n^0 - u_n^\varepsilon\right\|\right] + \gamma\mathbb{E}\left[e(\varepsilon, u_n^0)\right]. \tag{22}$$

Therefore, using the results (19), (20), (21) and (22), we obtain

$$\mathbb{E}\left[\left\|u_{n+1}^\varepsilon - u_{n+1}^0\right\|\right] \leq (1 + \alpha_1 + \alpha_2 + \alpha_3)\mathbb{E}\left[\left\|u_n^\varepsilon - u_n^0\right\|\right] + \gamma\mathbb{E}\left[e(\varepsilon, u_n^0)\right],$$

and defining $\alpha = 1 + \alpha_1 + \alpha_2 + \alpha_3$ we have

$$\mathbb{E}\left[\left\|u_{n+1}^\varepsilon - u_{n+1}^0\right\|\right] \leq \alpha\mathbb{E}\left[\left\|u_n^\varepsilon - u_n^0\right\|\right] + \gamma\mathbb{E}\left[e(\varepsilon, u_n^0)\right]. \tag{23}$$

Iterating (23) and using (14), after N iterations, at the end of the algorithm we get

$$\begin{aligned}\mathbb{E}\left[\left\|u_N^\varepsilon - u_N^0\right\|\right] &\leq \alpha^{N-1}\mathbb{E}[\left\|u_1^\varepsilon - u_1^0\right\|] + \gamma\sum_{i=0}^{N-2}\alpha^i\mathbb{E}\left[e(\varepsilon, u_{N-1-i}^0)\right] \\ &\leq \alpha^{N-1}\beta\mathbb{E}\left[e(\varepsilon, \psi)\right] + \gamma\sum_{i=0}^{N-2}\alpha^i\mathbb{E}\left[e(\varepsilon, u_{N-1-i}^0)\right],\end{aligned}$$

and since ψ is the initial ensemble u_0^0 we can write

$$\begin{aligned}\mathbb{E}\left[\left\|u_N^\varepsilon - u_N^0\right\|\right] &\leq \alpha^{N-1}\beta\mathbb{E}\left[e(\varepsilon, u_0^0)\right] + \gamma\sum_{i=0}^{N-2}\alpha^i\mathbb{E}\left[e(\varepsilon, u_{N-1-i}^0)\right] \\ &\leq \max\{\beta, \gamma\}\sum_{i=0}^{N-1}\alpha^i\mathbb{E}\left[e(\varepsilon, u_{N-1-i}^0)\right] \\ &= \delta\sum_{i=0}^{N-1}\alpha^{N-1-i}\mathbb{E}\left[e(\varepsilon, u_i^0)\right],\end{aligned}$$

where $\delta = \max\{\beta, \gamma\}$. Finally, applying Lemma 5, we have $e(\varepsilon, u_i^0) \to 0$ for all $i = 0, \ldots, N-1$, hence as $\varepsilon \to 0$

$$\mathbb{E}\left[\left\|u_N^\varepsilon - u_N^0\right\|\right] \to 0.$$

Moreover, if the solution of the homogenized problem p^0 belongs to $H^2(\Omega)$, then, by Lemma 5, we have the estimate

$$e(\varepsilon, u_i^0) \leq K\varepsilon.$$

Therefore we obtain

$$\mathbb{E}\left[\left\|u_N^\varepsilon - u_N^0\right\|\right] \leq \delta\left(\sum_{i=0}^{N-1}\alpha^i\right)K\varepsilon = \delta\frac{\alpha^N - 1}{\alpha - 1}K\varepsilon,$$

and defining $K_1 = \delta(\alpha^N - 1)K/(\alpha - 1)$ we have

$$\mathbb{E}\left[\left\|u_N^\varepsilon - u_N^0\right\|\right] \leq K_1\varepsilon,$$

which is the desired result. □

Regarding the second step, Proposition 2 is the equivalent formulation of Proposition 1 for the finite element convergence, where we show the convergence of the ensemble of particles generated by the EnKF algorithm using a FEM discretization of the homogenized problem as forward operator to the ensemble of particles generated by the EnKF algorithm using the true homogenized problem as forward operator.

Proposition 2.
Let $u_N^0 = \{u_N^{0^{(j)}}\}_{j=1}^J$, $u_{N,h}^0 = \{u_{N,h}^{0^{(j)}}\}_{j=1}^J$ be the ensembles of particles at the last iteration of the iterative ensemble Kalman filter for the forward operators $\mathcal{G}^0$ and $\mathcal{G}_h^0$ respectively. Then, under Assumption 1, Assumption 2, Assumption 3 and if the exact solution p^0 of the homogenized problem (8) is in $H^{q+1}(\Omega)$, $A^0 \in C^q(\Omega; \mathbb{R}^{N\times N})$, $f \in H^{q-1}(\Omega)$, $\partial\Omega \in C^{q+1}$ and we use polynomials of degree r for the finite element basis, we have

$$\mathbb{E}\left[\left\|u_{N,h}^0 - u_N^0\right\|\right] \leq K_2 h^{s+1},$$

where $s = \min\{r, q\}$, and

$$\mathbb{E}\left[\left\|u_{N,h}^0 - u_N^0\right\|\right] \to 0 \qquad \text{as } h \to 0.$$

Proof. The proof of Proposition 2 is identical to the proof of Proposition 1, except that all the ensembles $\{u_n^\varepsilon\}_{n=1}^N$ obtained by the multiscale operator $\mathcal{G}^\varepsilon$ have to be replaced by the ensembles $\{u_{n,h}^0\}_{n=1}^N$ obtained by the finite element discretization of the homogenized operator $\mathcal{G}_h^0$. Moreover Lemma 5 for the error e has to be replaced by Lemma 7 for the error $\tilde{e}$, which follows. □

Lemma 7.
Let $\tilde{e}$ be defined as in (7). Under Assumption 1 and if the exact solution p^0 of the homogenized problem (8) is in $H^{q+1}(\Omega)$, $A^0 \in C^q(\Omega; \mathbb{R}^{N\times N})$, $f \in H^{q-1}(\Omega)$, $\partial\Omega \in C^{q+1}$, and we use polynomials of degree r for the finite element basis, then we have

$$\tilde{e}(h, u) \leq \tilde{K} h^{s+1},$$

where $s = \min\{r, q\}$, which implies

$$\tilde{e}(h, u) \to 0 \text{ as } h \to 0.$$

Proof. By definition of $\tilde{e}$ and using the assumption for $\mathcal{O}$, for all $u \in \mathbb{R}^M$ we have

$$\begin{aligned}
\tilde{e}(h, u) &= \left\|\mathcal{G}_h^0(u) - \mathcal{G}^0(u)\right\|_2 \\
&= \left\|\mathcal{O}(\mathcal{S}_h^0(u)) - \mathcal{O}(\mathcal{S}^0(u))\right\|_2 \\
&= \left\|\mathcal{O}(p_h^0) - \mathcal{O}(p^0)\right\|_2 \\
&\leq m \left\|p_h^0 - p^0\right\|_{L^2(\Omega)}.
\end{aligned}$$

Thanks to the finite element theory (see e.g. [10, Theorem 4.7]) we have

$$\left\|p_h^0 - p^0\right\|_{L^2(\Omega)} \leq C \left|p\right|_{H^{s+1}(\Omega)} h^{s+1},$$

and, using higher order boundary regularity results for elliptic partial differential equations (see e.g. [5, Theorem 6.3.5]), letting $\tilde{C} > 0$, we have

$$\left\|p_h^0 - p^0\right\|_{L^2(\Omega)} \leq C \left\|p\right\|_{H^{s+1}(\Omega)} h^{s+1} \leq C \left\|p\right\|_{H^{q+1}(\Omega)} h^{s+1} \leq C\tilde{C} \left\|f\right\|_{H^{q-1}(\Omega)} h^{s+1},$$

hence we obtain

$$\tilde{e}(h, u) \leq mC\tilde{C} \left\|f\right\|_{H^{q-1}(\Omega)} h^{s+1},$$

then we define $\tilde{K} = mC\tilde{C} \left\|f\right\|_{H^{q-1}(\Omega)}$ and the sequence $\tilde{e}(h, u)$ converges to 0 as h vanishes. □

Applying Proposition 1 and Proposition 2, we finally prove Theorem 1.

Proof of Theorem 1. Using the triangle inequality we have

$$\mathbb{E}[\|u_N^\varepsilon - u_{N,h}^0\|] \leq \mathbb{E}[\|u_N^\varepsilon - u_N^0\|] + \mathbb{E}[\|u_N^0 - u_{N,h}^0\|],$$

and applying Proposition 1 and Proposition 2 we get

$$\begin{aligned}\mathbb{E}[\|u_N^\varepsilon - u_{N,h}^0\|] &\leq K_1\varepsilon + K_2 h^{s+1} \\ &\leq \max\{K_1, K_2\}(\varepsilon + h^{s+1}).\end{aligned}$$

Finally, we define $C = \max\{K_1, K_2\}$ and we obtain

$$\mathbb{E}[\|u_N^\varepsilon - u_{N,h}^0\|] \leq C(\varepsilon + h^{s+1}),$$

which is the desired result. □

Remark 2.
Note that if the exact solution of the homogenized problem in (8) $p^0 \in H^2(\Omega)$ and we use polynomials of degree $r = 1$ for the finite element basis, then we have

$$\mathbb{E}[\|u_N^\varepsilon - u_{N,h}^0\|] \leq C(\varepsilon + h^2).$$

Therefore, in order to balance the two sources of error, the parameter discretization h for the finite element approximation of the homogenized problem has to be taken

$$h = O(\varepsilon^{1/2}), \tag{24}$$

which guarantees linear convergence with respect to ε. With this choice for h, a much smaller computational cost is needed with respect to the solution of the multiscale problem, for which h^ε has to be taken

$$h^\varepsilon \ll \varepsilon,$$

in order to be able to well approximate the smallest scale. For example, suppose $\varepsilon = 10^{-2}$, then to solve the multiscale problem we could choose $h^\varepsilon = 10^{-3}$, but to solve the homogenized problem, by (24), it is enough to choose $h = 10^{-1}$.

4.2 Convergence of the posterior distributions

We now consider the Bayesian interpretation of the ensemble Kalman method and we introduce the family of Wassertein distances in order to show the convergence of the posterior distribution obtained using the multiscale problem as forward operator

$$\mu^\varepsilon = \frac{1}{J}\sum_{j=1}^{J} \delta_{u_n^{\varepsilon(j)}}$$

to the posterior distribution obtained using the finite element discretization of the homogenized problem as forward operator

$$\mu_h^0 = \frac{1}{J}\sum_{j=1}^{J} \delta_{u_{n,h}^{0(j)}}$$

as the multiscale and discretization parameters ε, h vanish.
Let $u^* \in \mathbb{R}^M$ and let $B_R(u^*)$ be the ball of radius R centered in u^* with respect to the norm $\|\cdot\|_s$ with $s \in [1, \infty]$. We define the Wasserstein distances in the metric space $(B_R(u^*), \|\cdot\|_s)$ following [12].

Definition 1.
Let μ and ν be two probability measures on the metric space $(B_R(u^*), \|\cdot\|_s)$. The Wasserstein distance between μ and ν is defined for all $p \in [1, \infty)$ as

$$W_{p,s}(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{B_R(u^*)\times B_R(u^*)} \|u - v\|_s^p \, d\gamma(u, v)\right)^{1/p}, \tag{25}$$

where $\Gamma(\mu, \nu)$ denotes the collection of all joint distributions on $B_R(u^*) \times B_R(u^*)$ with marginals μ and ν on the first and second factors respectively.

Remark 3.
If μ and ν are two discrete distributions on finite state spaces, respectively $\Omega_1 = \{u_1, \dots, u_{K_1}\}$ and $\Omega_2 = \{v_1, \dots, v_{K_2}\}$ included in $B_R(u^*)$, then (25) can be written as

$$W_{p,s}(\mu,\nu) = \left(\inf_{\gamma \in \mathbb{R}^{K_1 \times K_2}} \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \|u_i - v_j\|_s^p \gamma_{ij} \right)^{1/p}, \tag{26}$$

where the matrix γ has to satisfy the following constrains

$$\sum_{j=1}^{K_2} \gamma_{ij} = \mu(u_i) \qquad \text{for all } i = 1, \dots K_1,$$
$$\sum_{i=1}^{K_1} \gamma_{ij} = \nu(v_j) \qquad \text{for all } j = 1, \dots K_2.$$

Remark 4.
The Wasserstein distance with $p = 1$ can be written in an equivalent formulation using its duality representation

$$W_{1,s}(\mu,\nu) = \sup_{\varphi \in \Phi} \left\{ \int_{B_R(u^*)} \varphi d(\mu - \nu) \right\},$$

where Φ is the set of all continuous functions $\varphi \colon B_R(u^*) \to \mathbb{R}$ with minimal Lipschitz constant $L \leq 1$ with respect to the norm $\|\cdot\|_s$.

In Proposition 3 we show that $W_{1,2}$ is bounded by the distance induced by the ensemble norm defined in 5. This result will be crucial later to deduce the convergence of the posterior distribution μ_h^0 to μ^ε from Theorem 1.

Proposition 3.
Let $u_1 = \{u_1^{(j)}\}_{j=1}^J$, $u_2 = \{u_2^{(j)}\}_{j=1}^J$ be two ensembles of particles and let μ_1, μ_2 be the corresponding distributions defined as sum of Dirac masses

$$\mu_1 = \frac{1}{J} \sum_{j=1}^{J} \delta_{u_1^{(j)}}, \qquad \mu_2 = \frac{1}{J} \sum_{j=1}^{J} \delta_{u_2^{(j)}}.$$

Then

$$W_{p,s}(\mu_1,\mu_2) \leq \left(\frac{1}{J} \sum_{j=1}^{J} \left\| u_1^{(j)} - u_2^{(j)} \right\|_s^p \right)^{\frac{1}{p}}$$

and, in particular,

$$W_{1,2}(\mu_1,\mu_2) \leq \|u_1 - u_2\|.$$

Proof. Take γ^* defined as

$$\gamma^*(u_1^{(j)}, u_2^{(i)}) = \begin{cases} \frac{1}{J} & \text{if } i = j \\ 0 & \text{if } i \neq j, \end{cases}$$

which satisfies the constraints

$$\sum_{i=1}^{J} \gamma^*(u_1^{(j)}, u_2^{(i)}) = \mu_1(u_1^{(j)}) = \frac{1}{J} \qquad \text{for all } j = 1, \dots J,$$
$$\sum_{j=1}^{J} \gamma^*(u_1^{(j)}, u_2^{(i)}) = \mu_2(u_2^{(i)}) = \frac{1}{J} \qquad \text{for all } i = 1, \dots J,$$

and note that

$$\sum_{j=1}^{J}\sum_{i=1}^{J}\left\|u_1^{(j)}-u_2^{(i)}\right\|_s^p\gamma^*(u_1^{(j)},u_2^{(i)})=\frac{1}{J}\sum_{j=1}^{J}\left\|u_1^{(j)}-u_2^{(j)}\right\|_s^p.$$

Therefore, by definition of Wasserstein distance for discrete distributions on finite spaces (26), we deduce that

$$W_{p,s}(\mu_1,\mu_2)\leq\left(\frac{1}{J}\sum_{j=1}^{J}\left\|u_1^{(j)}-u_2^{(j)}\right\|_s^p\right)^{\frac{1}{p}},$$

which is the desired result. Finally, taking $p=1$ and $s=2$ and recalling the ensemble norm defined in (5), we obtain

$$W_{1,2}(\mu_1,\mu_2)\leq\frac{1}{J}\sum_{j=1}^{J}\left\|u_1^{(j)}-u_2^{(j)}\right\|_2=\|u_1-u_2\|,$$

which concludes the proof. □

Let us remark that posterior distributions with the form of (4) are random probability measures because of the randomness of the points in which the masses are centered caused by the noises $\xi_n^{(j)}$. Hence, we define the convergence of random probability measures and we analyse its connection with the Wasserstein distances.

Definition 2.
Let $(\Omega,\mathcal{A},P)$ be a probability space. A sequence of random probability measures $\{\mu_n\}_{n\in\mathbb{N}}$ dependent on a random variable ξ on $(\Omega,\mathcal{A},P)$ is said to weakly converge in $L^1(\Omega)$ to a random probability measure μ if for all bounded continuous functions $f\in C_B^0$ we have

$$\mathbb{E}_\xi\left[\left|\int f d\mu_n-\int f d\mu\right|\right]\to 0.$$

In this case we write

$$\mu_n\xrightarrow{L^1}\mu.$$

In Theorem 2 we show that convergence with respect to the expectation of the Wasserstein distances implies weak L^1 convergence of random probability measures. The fact that convergence with respect to the Wasserstein distances implies weak convergence of distribution was proved in [12] for non-random measures, but here we extend the result to random probability measures.

Theorem 2.
Let $(\Omega,\mathcal{A},P)$ be a probability space. Let the sequence $\{\mu_n\}_{n\in\mathbb{N}}$ and μ be random probability measures on the metric space $(B_R(u^*),\|\cdot\|_s)$ dependent on the random variable ξ on $(\Omega,\mathcal{A},P)$. If

$$\mathbb{E}_\xi[W_{1,s}(\mu_n,\mu)]\to 0,$$

then

$$\mu_n\xrightarrow{L^1}\mu.$$

Proof. The proof can be found in the appendix. □

Finally, applying Theorem 1, we show the convergence of the posterior distribution μ^ε to μ_h^0 as the multiscale and discretization parameters ε,h vanish.

Theorem 3.
Let the hypotheses of Theorem 1 be satisfied. Define the posterior random probability measures

$$\mu^\varepsilon=\frac{1}{J}\sum_{j=1}^{J}\delta_{u_N^{\varepsilon(j)}}\qquad\text{and}\qquad\mu_h^0=\frac{1}{J}\sum_{j=1}^{J}\delta_{u_{N,h}^{0(j)}},$$

then as $\varepsilon,h\to 0$

$$\mu^\varepsilon-\mu_h^0\xrightarrow{L^1}0.$$

Proof. By Theorem 1 we know that the average of the ensemble norm of the difference of u_N^ε and $u_{N,h}^0$ vanishes as ε and h go to zero

$$\mathbb{E}[\|u_N^\varepsilon - u_{N,h}^0\|] \to 0,$$

and applying Proposition 3 we deduce that

$$\mathbb{E}[W_{1,2}(\mu^\varepsilon, \mu_h^0)] \to 0.$$

Note that the only difference in the update step is that Γ is replaced by $\Delta^{-1}\Gamma$ where $\Delta = 1/N$. The constants of the proof of Theorem 1 depend on $\left\|\Gamma^{-1}\right\|_2$, which is now replaced by $\left\|(\Delta^{-1}\Gamma)^{-1}\right\|_2$, which can be bounded by $\left\|\Gamma^{-1}\right\|_2$

$$\left\|(\Delta^{-1}\Gamma)^{-1}\right\|_2 = \Delta \left\|\Gamma^{-1}\right\|_2 \leq \left\|\Gamma^{-1}\right\|_2.$$

Finally, by Theorem 2, we obtain

$$\mu^\varepsilon - \mu_h^0 \xrightarrow{L^1} 0,$$

which is the desired result. □

5 Modelling error

We want to predict the exact unknown u^* from observations originated by the model

$$y = \mathcal{G}^\varepsilon(u^*) + \eta, \tag{27}$$

where $\eta \sim \mathcal{N}(0, \Gamma)$ is the noise, but we use the discretization of the homogenized operator $\mathcal{G}_h^0$. Note that (27) can be written as

$$y = \mathcal{G}_h^0(u^*) + \left[\mathcal{G}^\varepsilon(u^*) - \mathcal{G}_h^0(u^*)\right] + \eta,$$

and defining

$$\mathcal{E}(u^*) = \mathcal{G}^\varepsilon(u^*) - \mathcal{G}_h^0(u^*),$$

which represents the modelling error between the multiscale model and the discretization of the homogenized one, we have

$$y = \mathcal{G}_h^0(u^*) + \mathcal{E} + \eta. \tag{28}$$

Equation (28) shows that the observed data y can be seen as data originating by the discrete homogenized model which is affected by two sources of errors, the original noise and the modelling error. This idea was originally presented in [8], and then applied to multiscale inverse problems in [1].
We assume that the modelling error is a Gaussian random variable independent of the noise η, so that $\mathcal{E} \sim \mathcal{N}(m, \Sigma)$ for all u, and we write

$$y = \mathcal{G}_h^0(u^*) + m + \zeta + \eta, \tag{29}$$

where $\zeta \sim \mathcal{N}(0, \Sigma)$. Then we define

$$\tilde{y} = y - m \qquad \text{and} \qquad \tilde{\eta} = \eta + \zeta \sim \mathcal{N}(0, \Gamma + \Sigma)$$

and, from (29), we obtain

$$\tilde{y} = \mathcal{G}_h^0(u^*) + \tilde{\eta}. \tag{30}$$

Therefore, once we know the mean m and the covariance Σ of the modelling error, in order to get a more reliable approximation of the unknown u^*, we can apply the iterative ensemble Kalman method considering $\tilde{y}$ as the vector of observations and $\Gamma + \Sigma$ as the covariance of the noise.
The modelling error distribution, namely its mean and covariance, is approximated offline. We

sample $N_{\mathcal{E}}$ unknowns $\{u_i\}_{i=1}^{N_{\mathcal{E}}}$ from the prior distribution μ_0 and, for all $i = 1, \dots, N_{\mathcal{E}}$, we apply both the forward operators $\mathcal{G}^{\varepsilon}(u_i)$ and $\mathcal{G}_h^0(u_i)$. Then we compute

$$\mathcal{E}_i = \mathcal{G}^{\varepsilon}(u_i) - \mathcal{G}_h^0(u_i),$$

and the mean m and the covariance Σ are given by

$$m = \frac{1}{N_{\mathcal{E}}} \sum_{i=1}^{N_{\mathcal{E}}} \mathcal{E}_i,$$
$$\Sigma = \frac{1}{N_{\mathcal{E}}} \sum_{i=1}^{N_{\mathcal{E}}} (\mathcal{E}_i - m)(\mathcal{E}_i - m)^T.$$

This procedure is computationally very expensive because we have to solve $N_{\mathcal{E}}$ times a multiscale problem, but it has to be done only once, then we can solve the inverse problem for any value of the unknown u^*.
In order to obtain a more reliable approximation of the distribution of the modelling error, inspired by [3], we can follow a dynamic approach, based on the estimation of the mean m and the covariance Σ online, during the resolution of the ensemble Kalman algorithm. We sequentially apply the ensemble Kalman method for $\mathcal{L}$ levels and, in each level, we use a different approximation of the distribution of the modelling error, that we denote $\nu^{\ell} = \mathcal{N}(m^{\ell}, \Sigma^{\ell})$ for any $\ell = 1, \dots, \mathcal{L}$. Moreover, let

$$\mu_n^{\ell} = \frac{1}{J} \sum_{j=1}^{J} \delta_{u_n^{\ell(j)}}$$

be the approximation of the distribution of the particles at iteration n at level ℓ, $\mu_0^{\ell+1} = \mu_{N^{\ell}}^{\ell}$ and $\mu_0^1 = \mu_0$, where N^{ℓ} is the number of iterations at level ℓ. At the beginning of each level ℓ, we approximate the distribution ν^{ℓ} by sampling $N_{\mathcal{E}}^{\ell}$ particles $\{u_i^{\ell}\}_{i=1}^{N_{\mathcal{E}}^{\ell}}$ from the distribution μ_0^{ℓ} and computing the mean m^{ℓ} and the covariance Σ^{ℓ} as

$$m^{\ell} = \frac{1}{N_{\mathcal{E}}^{\ell}} \sum_{i=1}^{N_{\mathcal{E}}^{\ell}} \mathcal{E}_i^{\ell},$$
$$\Sigma^{\ell} = \frac{1}{N_{\mathcal{E}}^{\ell}} \sum_{i=1}^{N_{\mathcal{E}}^{\ell}} (\mathcal{E}_i^{\ell} - m^{\ell})(\mathcal{E}_i^{\ell} - m^{\ell})^T,$$

where

$$\mathcal{E}_i^{\ell} = \mathcal{G}^{\varepsilon}(u_i^{\ell}) - \mathcal{G}_h^0(u_i^{\ell}).$$

This approach gives a better approximation of the modelling error, indeed, instead of taking the samples from the prior distribution, they are drawn from distributions which are closer to the real posterior distribution. On the other hand, this procedure has to be done online and it is computationally expensive because it requires the resolution of $N_{\mathcal{E}} = \sum_{\ell=1}^{\mathcal{L}} N_{\mathcal{E}}^{\ell}$ full multiscale problems.

Remark 5.
Note that the modelling error is negligible when ε is small, but it becomes important when ε is big, thus when the multiscale problem is not computationally very expensive to solve. Therefore we can avoid estimating the modelling error when ε is small and solving a multiscale problem is computationally expensive because we already have a good approximation of the unknown, while it is necessary when ε is big, but in this case solving a multiscale problem is relatively cheap.

Finally, in Proposition 4, we show that the number $N_{\mathcal{E}}$ of full multiscale problems, which have to be solved in order to have a reliable approximation of the true mean m^* of the modelling error, decreases as the multiscale parameter ε and the discretization parameter h become smaller. This is a good result, indeed, as stated in Remark 5, the solution of a full multiscale problem is computationally cheap wean ε is relatively large and the solution of the corresponding homogenized

problem is even cheaper when h is big. Before stating Proposition 4, let us recall the Hoeffding's inequality, which will be used in the proof. Let $\{X_i\}_{i=1}^N$ be independent random variables bounded by the interval $[a,b]$, that is $a \leq X_i \leq b$ for all $i = 1, \dots, N$, let $\bar{X} = \frac{1}{N}\sum_{i=1}^N X_i$ and $t \in \mathbb{R}$, then

$$\mathbb{P}(\left|\bar{X} - \mathbb{E}[X]\right| \geq t) \leq 2\exp\left\{-\frac{2t^2 N}{(b-a)^2}\right\}.$$

Proposition 4.
Let $\alpha \in (0,1)$, $t > 0$ and $C_{\mathcal{E}} = \max\{K, \tilde{K}\}$, where K and $\tilde{K}$ are the constants of Lemma 5 and Lemma 7. Let $\{\mathcal{E}_i\}_{i=1}^{N_\mathcal{E}} \subset \mathbb{R}^L$ given by

$$\mathcal{E}_i = \mathcal{G}^\varepsilon(u_i) - \mathcal{G}_h^0(u_i) \qquad \text{for all } i = 1, \dots, N_\mathcal{E}$$

for a sample of realizations $\{u_i\}_{i=1}^{N_\mathcal{E}}$ from the standard normal distribution $\mathcal{N}(0, I)$ and let m be its sample mean

$$m = \frac{1}{N_\mathcal{E}} \sum_{i=1}^{N_\varepsilon} \mathcal{E}_i,$$

and $m^* = \mathbb{E}[\mathcal{E}_i]$ be the true mean of the modelling error, which is unknown. If

$$N_\mathcal{E} \geq 4C_\mathcal{E}^2 \frac{L}{t^2} \log\left(\frac{2L}{\alpha}\right)\left[\varepsilon^2 + h^{2(s+1)}\right],$$

where s is given by Lemma 7, then

$$\mathbb{P}\left(\|m - m^*\|_2 \leq t\right) \geq 1 - \alpha.$$

Proof. First, note that the modelling error is bounded, indeed by Lemma 5 and Lemma 7, we have for each $i = 1, \dots, N_\mathcal{E}$

$$\|\mathcal{E}_i\|_2 = \left\|\mathcal{G}^\varepsilon(u_i) - \mathcal{G}_h^0(u_i)\right\|_2 \leq \left\|\mathcal{G}^\varepsilon(u_i) - \mathcal{G}^0(u_i)\right\|_2 + \left\|\mathcal{G}^0(u_i) - \mathcal{G}_h^0(u_i)\right\|_2 \leq K\varepsilon + \tilde{K}h^{s+1},$$

so each component $(\mathcal{E}_i)_l$, for $l = 1, \dots, L$, is bounded by the same constant

$$|(\mathcal{E}_i)_l| \leq \|\mathcal{E}_i\|_2 \leq K\varepsilon + \tilde{K}h^{s+1} \leq C_\mathcal{E}(\varepsilon + h^{s+1}). \tag{31}$$

Observe that if

$$|m_l - m_l^*| \leq \frac{t}{\sqrt{L}} \qquad \text{for each } l = 1, \dots, L,$$

then

$$\|m - m^*\|_2 = \left(\sum_{l=1}^L |m_l - m_l^*|^2\right)^{\frac{1}{2}} \leq \left(L\frac{t^2}{L}\right)^{\frac{1}{2}} = t,$$

which implies that

$$\mathbb{P}(\|m - m^*\|_2 \leq t) \geq \mathbb{P}\left(|m_l - m_l^*| \leq \frac{t}{\sqrt{L}} \quad \forall\, l = 1, \dots, L\right). \tag{32}$$

Using (31) and applying Hoeffding's inequality we have

$$\mathbb{P}\left(|m_l - m_l^*| \geq \frac{t}{\sqrt{L}}\right) \leq 2\exp\left\{-\frac{2t^2 N_\mathcal{E}^2}{4LN_\mathcal{E}C_\mathcal{E}^2(\varepsilon + h^{s+1})^2}\right\} \leq 2\exp\left\{-\frac{t^2 N_\mathcal{E}}{4LC_\mathcal{E}^2(\varepsilon^2 + h^{2(s+1)})}\right\}. \tag{33}$$

Define the events $A_l = \left\{|m_l - m_l^*| \leq \frac{t}{\sqrt{L}}\right\}$ for each $l = 1, \dots, L$, then we have

$$\mathbb{P}\left(|m_l - m_l^*| \leq \frac{t}{\sqrt{L}} \quad \forall\, l = 1, \dots, L\right) = \mathbb{P}\left(\bigcap_{l=1}^L A_l\right),$$

and, applying the De Morgan's laws and the union bound, we obtain

$$\mathbb{P}\left(\bigcap_{l=1}^{L} A_l\right) = 1 - \mathbb{P}\left(\left(\bigcap_{l=1}^{L} A_l\right)^C\right) = 1 - \mathbb{P}\left(\bigcup_{l=1}^{L} A_l^C\right) \geq 1 - \sum_{l=1}^{L} \mathbb{P}(A_l^C). \tag{34}$$

Therefore, thanks to (32), (33) and (34), we have

$$\begin{aligned}\mathbb{P}(\|m - m^*\|_2 \leq t) &\geq 1 - L\mathbb{P}\left(|m_l - m_l^*| \geq \frac{t}{\sqrt{L}}\right) \\ &\geq 1 - 2L\exp\left\{-\frac{t^2 N_{\mathcal{E}}}{4LC_{\mathcal{E}}^2(\varepsilon^2 + h^{2(s+1)})}\right\},\end{aligned}$$

and, if $N_{\mathcal{E}}$ satisfies the hypothesis, we obtain

$$\mathbb{P}(\|m - m^*\|_2 \leq t) \geq 1 - 2L\exp\left\{-\log\left(\frac{2L}{\alpha}\right)\right\} = 1 - \alpha,$$

which is the desired result. □

Remark 6.
Note that, in Proposition 4, as expected, the number $N_{\mathcal{E}}$ of full multiscale problems tends to infinity if we require no error between the sample and the true mean ($t \to 0$) or certainty that the error is below a certain value ($\alpha \to 0$).

6 Numerical experiments

In this chapter, using the setting of [1], we present some numerical experiments to illustrate the iterative ensemble Kalman method to solve multiscale inverse problems.
Let Ω be a bounded open domain. We consider a class of parametrized multiscale locally periodic tensors of the type $A^{\varepsilon}_{\sigma^*}(x) = A(\sigma^*(x), x/\varepsilon)$, where $\sigma^*\colon \Omega \to \mathbb{R}$. We assume to know the map $(t, x) \to A(t, x/\varepsilon)$ for all $x \in \Omega$ and $t \in \mathbb{R}$ and we want to estimate the function σ^* given measurements computed from the model

$$\begin{cases} -\nabla \cdot (A^{\varepsilon}_{\sigma^*} \nabla p^{\varepsilon}) = 0 & \text{in } \Omega, \\ p^{\varepsilon} = g & \text{on } \partial\Omega. \end{cases} \tag{35}$$

Remark 7.
Note that the theory has been developed for Dirichlet homogeneous boundary conditions, but it can be applied to the non-homogeneous case by considering an extension of the function at the boundary and slightly modifying the PDE. For more details we refer to [11, Remark 8.10].

For the unknown σ^* we consider the following admissible set

$$\Sigma = \{\sigma \in L^{\infty}(\Omega)\colon \sigma^- \leq \sigma(x) \leq \sigma^+\},$$

where σ^- and σ^+ are two given values.
The measurements, which we take into account, are the integrals of the normal flux multiplied by some functions with compact support in a portion of the boundary of the domain. More precisely, we consider $I \in \mathbb{N}$ disjoint portions of Ω, which we denote by $\Gamma_i \in \partial\Omega$, $i = 1, \ldots, I$, $\Gamma_i \cap \Gamma_j = \emptyset$ for $i \neq j$, and I functions $\varphi_i \in H^{1/2}(\partial\Omega)$ with compact support $\operatorname{supp}(\varphi_i) \subset \Gamma_i$ for all $i = 1, \ldots, I$. Moreover, we solve (35) for $K \in \mathbb{N}$ Dirichlet data, denoted by g_k with $k = 1, \ldots, K$. Then we define the multiscale operator $\mathcal{F}^{\varepsilon}\colon \Sigma \to \mathbb{R}^L$ where $L = IK$ by components

$$\mathcal{F}^{\varepsilon}(\sigma)_{ik} = \mathcal{F}^{\varepsilon}(\sigma)_l = \int_{\Gamma_i} A^{\varepsilon} \nabla p_k^{\varepsilon} \cdot \nu \varphi_i ds, \qquad i = 1, \ldots, I, k = 1, \ldots, K. \tag{36}$$

where p_k^ε is the solution of problem (35) with Dirichlet boundary condition g_k and ν is the exterior unit normal vector to $\partial\Omega$. The final vector of observations y is given by the sum of the operator $\mathcal{F}^\varepsilon$ and a noise

$$y = \mathcal{F}^\varepsilon(\sigma^*) + \eta,$$

where $\eta \sim \mathcal{N}(0, \Gamma)$ and Γ is a given covariance matrix, which, in our experiments, is a multiple of the identity $\Gamma = \gamma^2 I$ and γ is a given value. Observations are computed with a refined Finite Element Method (FEM) with mesh size $h_{\text{obs}} \ll \varepsilon$, while the homogenized version of problem (35) is solved using a macro mesh size $h \gg h_{\text{obs}}$. We call $\mathcal{T}_h$ the macro triangulation and N_h the total number of nodes defining $\mathcal{T}_h$. We assume that the prior distribution for the discretization of the unknown σ^* on the macro triangulation $\mathcal{T}_h$ is given by $\mathcal{N}(\sigma_0, C)$, where σ_0 is a given discretization of a function in Σ and $C \in \mathbb{R}^{N_h \times N_h}$ is defined by

$$C_{ij} = \delta \exp\left(-\frac{\|x_i - x_j\|_2}{\lambda}\right)$$

where $\delta, \lambda \in \mathbb{R}^+$ and $\{x_i\}_{i=1}^{N_h}$ are the nodes of the macro triangulation $\mathcal{T}_h$. The parameter λ is a correlation length that describes how the values at different positions of the functions supported by the prior measure are related, while the parameter δ is an amplitude scaling factor.
In order to reduce the dimensionality of the unknown we use a truncated Karhunen-Loève expansion. Any sample from the prior distribution $\mathcal{N}(\sigma_0, C)$ can be represented as

$$\sigma = \sigma_0 + \sum_{m=1}^{N_h} \sqrt{\lambda_m} u_m \varphi_m, \tag{37}$$

where $\{\varphi_m\}_{m=1}^{N_h}$ is an orthonormal set of eigenvectors of C with corresponding eigenvalues $\{\lambda_m\}_{m=1}^{N_h}$ in decreasing order, and $\{u_m\}_{m=1}^{N_h}$ is an i.i.d sequence with $u_m \sim \mathcal{N}(0, 1)$. Note that the Karhunen-Loève expansion works also in the infinite dimensional setting, where $\sigma_0 \in \Sigma$, C is a covariance operator and $\{\lambda_m, \varphi_m\}_{m=1}^{\infty}$ is an orthonormal set of eigenvalues-eigenfunctions with respect to the scalar product in $L^2(\Omega)$. Then the truncated Karhunen-Loève expansion of the discretization of σ consists of taking the first M components of the series in (37)

$$\sigma \simeq \sigma_0 + \sum_{m=1}^{M} \sqrt{\lambda_m} u_m \varphi_m, \tag{38}$$

and the actual unknown becomes the vector $u \in \mathbb{R}^M$, whose components are the coefficients u_m in (38). Then we define the multiscale forward operator $\mathcal{G}^\varepsilon \colon \mathbb{R}^M \to \mathbb{R}^L$ as the composition of $\mathcal{F}^\varepsilon$ with the truncated Karhunen-Loève expansion

$$\mathcal{G}^\varepsilon(u) = \mathcal{F}^\varepsilon\left(\sigma_0 + \sum_{m=1}^{M} \sqrt{\lambda_m} u_m \varphi_m\right).$$

On the other hand, in the iterative ensemble Kalman method we do not compute the exact solution of problem (35), but we solve its homogenized version numerically using the macro triangulation $\mathcal{T}_h$, therefore we obtain the homogenized discrete solution p_h^0. The problem is solved applying the Finite Element Heterogeneous Multiscale Method (FE-HMM), which is described in [2]. Hence, analogously to the multiscale case, we define the discrete homogenized operator $\mathcal{F}_h^0 \colon \Sigma \to \mathbb{R}^L$ as

$$\mathcal{F}_h^0(\sigma)_l = \mathcal{F}_h^0(\sigma)_{ik} = \int_{\Gamma_i} A^0 \nabla p_{h_k}^0 \cdot \nu \varphi_i ds, \qquad i = 1, \dots, I, k = 1, \dots, K, \tag{39}$$

and the discrete homogenized forward operator $\mathcal{G}_h^0 \colon \mathbb{R}^M \to \mathbb{R}^L$, which is actually used in the algorithm, as

$$\mathcal{G}_h^0(u) = \mathcal{F}_h^0\left(\sigma_0 + \sum_{m=1}^{M} \sqrt{\lambda_m} u_m \varphi_m\right).$$

Finally, we call u_{EnKF} the solution of the iterative ensemble Kalman algorithm and the estimated σ_{EnKF} is obtained from the truncated Karhunen-Loève expansion

$$\sigma_{\text{EnKF}} = \sigma_0 + \sum_{m=1}^{M} \sqrt{\lambda_m} u_{\text{EnKF}_m} \varphi_m.$$

6.1 Data

In the numerical results presented in the following section the computational domain is the unit square

$$\Omega = (0,1)^2 \subset \mathbb{R}^2.$$

For the discretization parameters we set $\varepsilon = 1/64$ and $h_{\text{obs}} = 1/4096$ and for the forward homogenized problem we use a macro mesh size $h = 1/32$, which is much larger than h_{obs} and reduces the computational cost significantly.
We solve the problem for $K = 3$ Dirichlet conditions $\{g_k\}_{k=1}^3$ and $g_k = \sqrt{\mu_k}\psi_k$ where $\{(\mu_k, \psi_k)\}_{k=1}^3$ are couples of eigenvalues and eigenfunctions of the one dimensional discrete Laplacian operator corresponding to the first $K = 3$ smallest eigenvalues. For each g_k we consider its restriction to the boundary $\partial\Omega$ in order to obtain a Dirichlet condition. These functions are orthonormal with respect to the scalar product in $L^2(\Omega)$ and this ensures that each function gives independent information.
To compute the boundary integrals in (36) and (39), we consider $I = 12$ boundary portions, three for each side of the square Ω. In particular, for each side, all Γ_i have length equal to 0.2 and they consist of the intervals $(0.1, 0.3), (0.4, 0.6)$ and $(0.7, 0.9)$. The functions $\{\varphi_i\}_{i=1}^{12}$ are hat functions with $\text{supp}\,(\varphi_i) = \Gamma_i$, which take value one at the midpoint and value 0 at the extremes of Γ_i. Then the parameter of the noise, which perturbs the observations, is $\gamma = 0.01$.
Moreover, regarding the prior distribution for the unknown, we consider $\sigma_0 = 0$ and the parameters of the covariance matrices are $\delta = 0.05$ and $\lambda = 0.5$. In the truncated Karhunen-Loève expansion we take $M = 100$.
Finally, about the ensemble Kalman method, we consider $J = 1000$ particles for each ensemble and 500 iterations.
The exact tensor $A^{\varepsilon}_{\sigma^*}$ is given by

$$\begin{aligned}
a_{11}\left(\sigma^*(x), \frac{x}{\varepsilon}\right) &= e^{\sigma^*(x)}\left(\cos^2\left(\frac{2\pi x_1}{\varepsilon}\right) + 1\right) + \cos^2\left(2\pi\frac{x_2}{\varepsilon}\right), \\
a_{12}\left(\sigma^*(x), \frac{x}{\varepsilon}\right) &= 0, \\
a_{21}\left(\sigma^*(x), \frac{x}{\varepsilon}\right) &= 0, \\
a_{22}\left(\sigma^*(x), \frac{x}{\varepsilon}\right) &= e^{\sigma^*(x)}\left(\sin\left(\frac{2\pi x_2}{\varepsilon}\right) + 2\right) + \cos^2\left(2\pi\frac{x_1}{\varepsilon}\right),
\end{aligned}$$

where

$$\sigma^*(x) = \log(1.3 + 0.3\mathbb{1}_{D_1} - 0.4\mathbb{1}_{D_2}),$$

and

$$\begin{aligned}
D_1 &= \left\{x = (x_1, x_2)\colon \left(x_1 - \frac{5}{16}\right)^2 + \left(x_2 - \frac{11}{16}\right)^2 \leq 0.025\right\}, \\
D_2 &= \left\{x = (x_1, x_2)\colon \left(x_1 - \frac{11}{16}\right)^2 + \left(x_2 - \frac{5}{16}\right)^2 \leq 0.025\right\}.
\end{aligned}$$

Figure 1 shows the exact unknown σ^*. Note that σ^* is a non-continuous function, but, in order to approximate it, we are using a truncated Karhunen-Loève expansion, where the eigenfunctions are smooth.

6.2 Results

In Figure 2 we plot the estimation σ_{EnKF} after $10, 50, 250$ and 500 iterations of the ensemble Kalman algorithm. We clearly see that the approximation gets better as the number of iterations increases and that convergence has been reached, indeed we do not note a significant difference between the last two plots. We point out that we obtain a quite good approximation of the real unknown σ^* ,

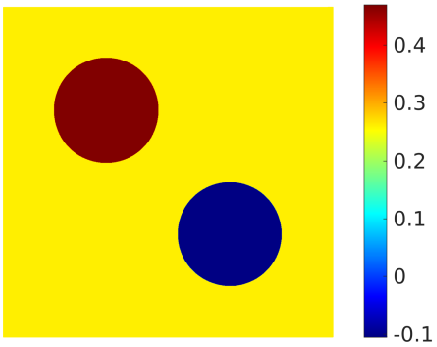

Figure 1: Plot of the exact unknown σ^*

indeed we are trying to recover a non-continuous function in the whole domain given only some observations at the boundary.

We perform a sensitivity analysis with respect to the dimension of the ensemble and the multiscale parameter ε. In Figure 3 we vary the number of particles J and we compare the results obtained at the end of the algorithm after 500 iterations. As expected, the approximation becomes better when the ensemble contains more particles. In particular, note that if the number of particles is too small, e.g. $J = 10$, then the approximation is completely different from the true unknown.

In Figure 4 we compare the results obtained for different values of the multiscale parameter ε, in particular we take $\varepsilon = 1/4, 1/8, 1/32, 1/64$. We notice that the approximation becomes worse when ε is bigger, indeed the homogenized problem becomes too different with respect to the multiscale one and, if ε is too big, the solution does not approximate the true unknown.

Moreover, in order to obtain good results even in case ε is not close to the asymptotic limit $\varepsilon \to 0$, in Figure 5 we apply offline modelling error estimation with $N_{\mathcal{E}} = 20$ and we plot the solution of the inverse problem (30) for different values of the multiscale parameter ε. Comparing these plots with the ones in Figure 4, in particular for $\varepsilon = 1/4$, we observe that the modelling error estimation significantly improves the results.

Finally, in Figure 6 we show the results obtained by applying the ensemble Kalman method with dynamic updating of the modelling error distribution with $\mathcal{L} = 5$ levels, $N_{\mathcal{E}}^{\ell} = 4$ samples and $N^{\ell} = 100$ iterations at each level $\ell = 1, \dots, \mathcal{L}$. The number of resolutions of the full multiscale problem is 20 and the total number of iterations is 500, which are equal to the previous approach, where the distribution of the modelling error was approximated offline. Comparing these plots with the ones in Figure 5, we note that updating the distribution of the modelling error dynamically still improves the results.

Conclusion

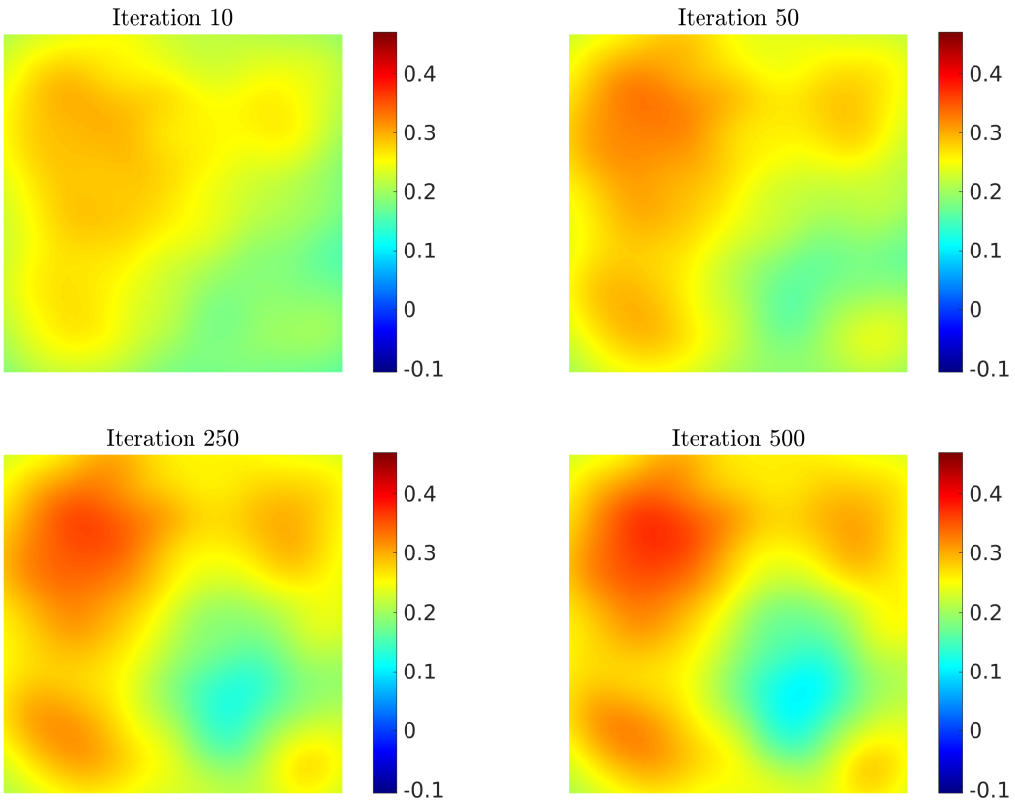

Figure 2: Plot of the unknown σ_{EnKF} estimated by the iterative ensemble Kalman method after $10, 50, 250$ and 500 iterations.

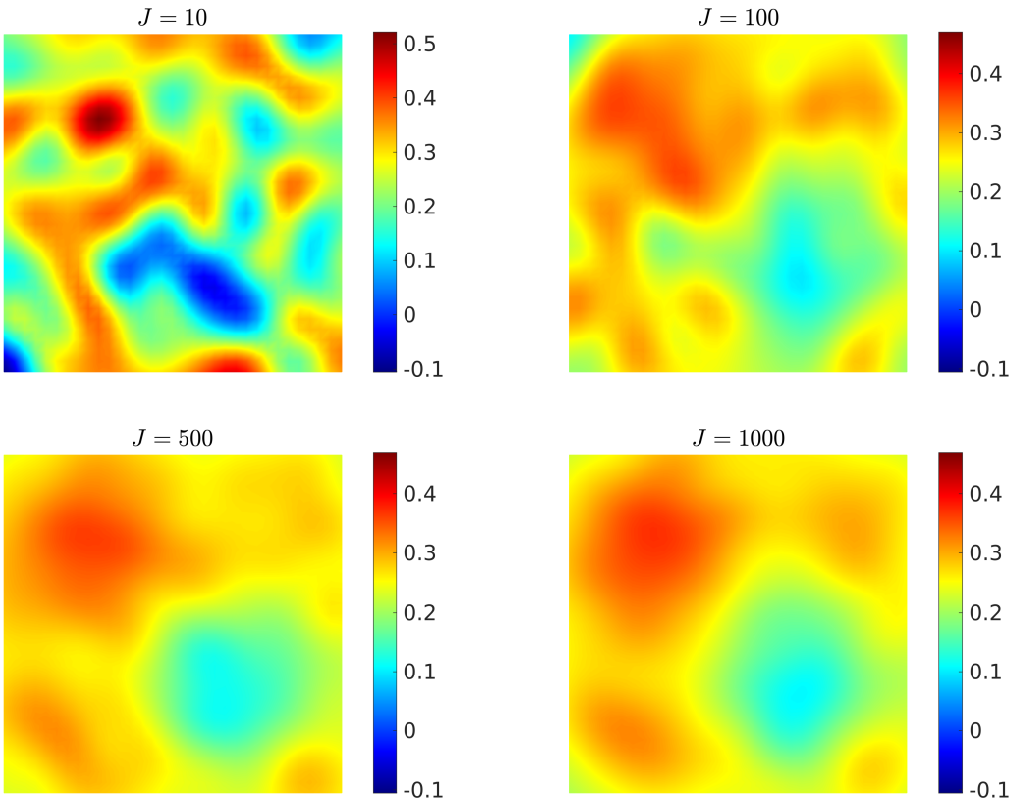

Figure 3: Plot of the unknown σ_{EnKF} estimated by the iterative ensemble Kalman method after 500 iterations for different numbers of particles per ensemble $J = 10, 100, 500, 1000$.

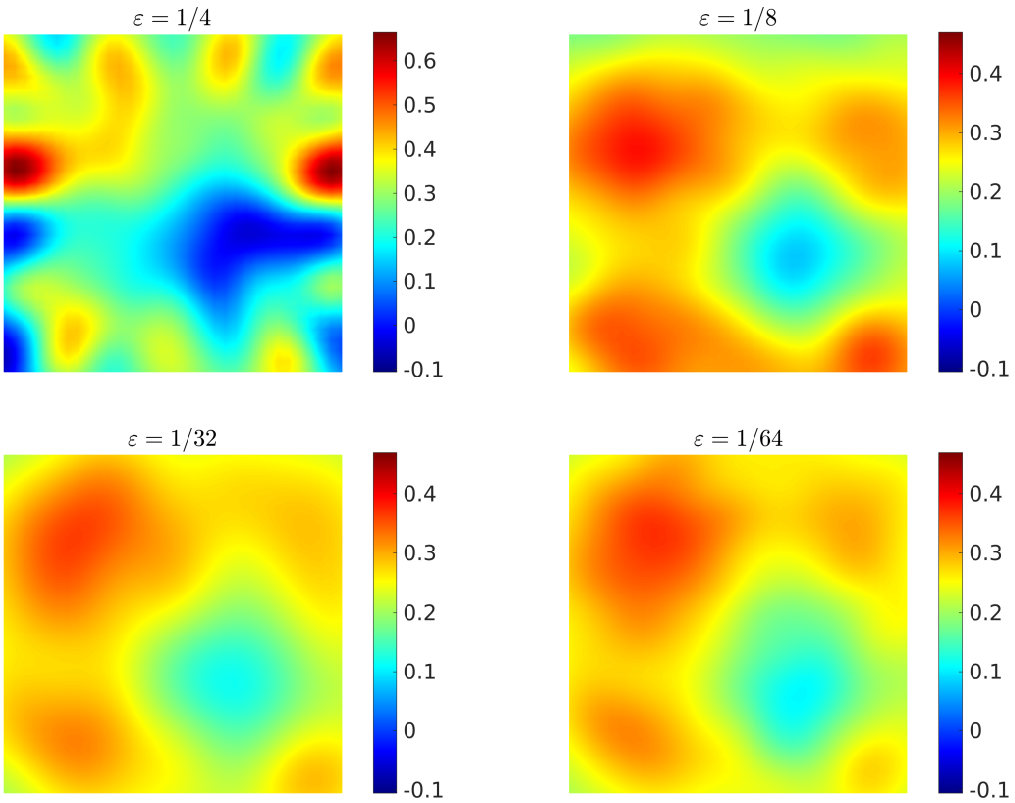

Figure 4: Plot of the unknown σ_{EnKF} estimated by the iterative ensemble Kalman method after 500 iterations for different values of the multiscale parameter $\varepsilon = 1/4, 1/8, 1/32, 1/64$.

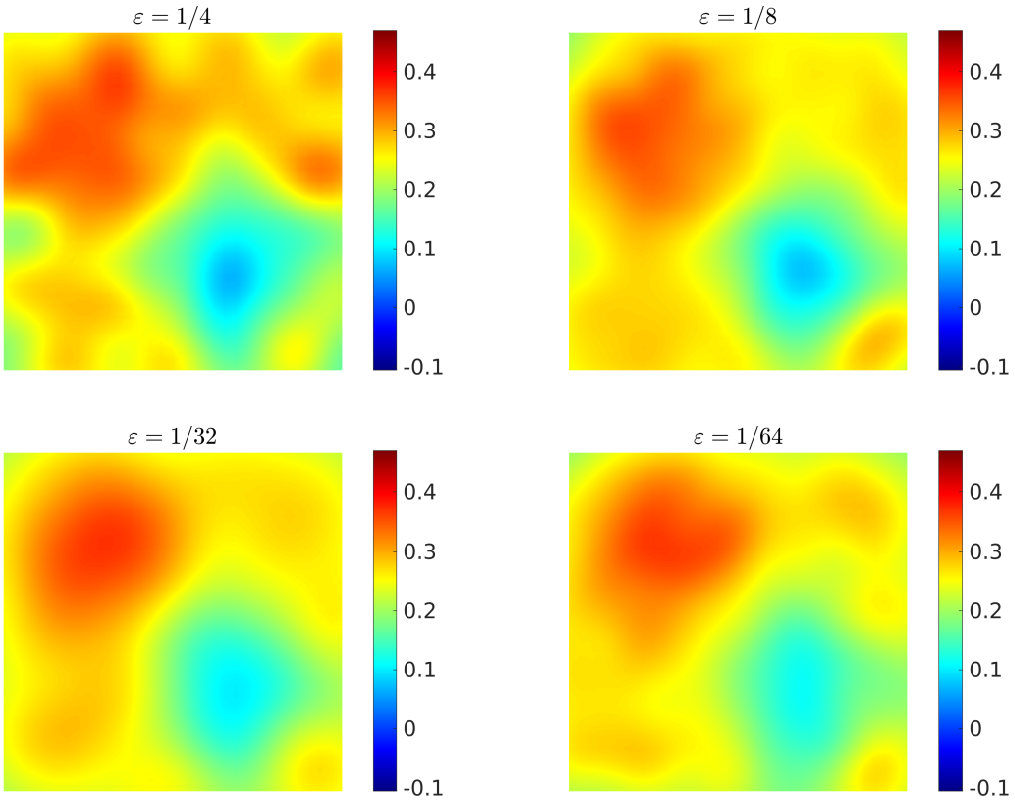

Figure 5: Plot of the unknown σ_{EnKF} estimated by the iterative ensemble Kalman method with model error estimation after 500 iterations for different values of the multiscale parameter $\varepsilon = 1/4, 1/8, 1/32, 1/64$.

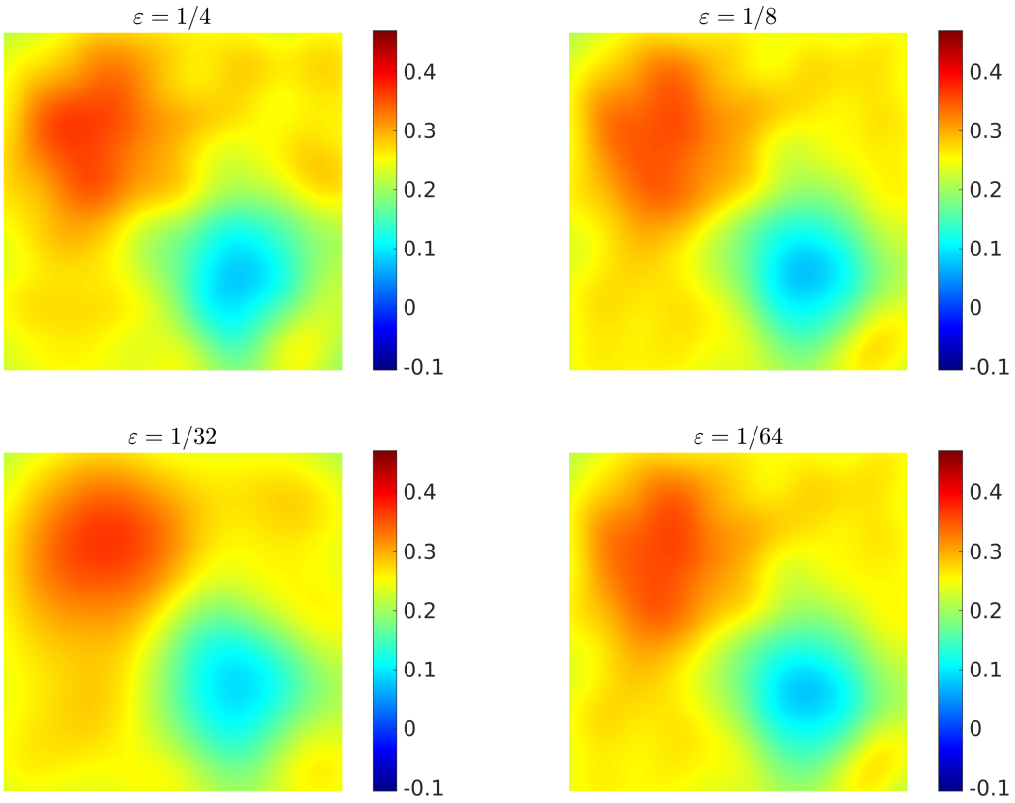

Figure 6: Plot of the unknown σ_{EnKF} estimated by the iterative ensemble Kalman method with dynamic updating of the model error estimation after 500 iterations for different values of the multiscale parameter $\varepsilon = 1/4, 1/8, 1/32, 1/64$.

Appendix

Proof of Lemma 1

Note that

$$A^{-1} - B^{-1} = A^{-1}(I - AB^{-1}) = A^{-1}(B - A)B^{-1},$$

therefore we have

$$\left\|A^{-1} - B^{-1}\right\|_2 \leq \left\|A^{-1}\right\|_2 \|B - A\|_2 \left\|B^{-1}\right\|_2,$$

which is the desired result.

□

Proof of Lemma 2

Let n be the dimension of the matrices, since A is symmetric positive semidefinite and B is symmetric positive definite, then $A + B$ is symmetric positive definite, and the eigenvalues of $A + B$ and B are real and positive, thus they can be written

$$0 < \lambda_1(\cdot) \leq \lambda_2(\cdot) \leq \cdots \leq \lambda_n(\cdot),$$

counted with their multiplicity. First, notice that, using the Rayleigh quotient and the fact that $x^T Ax \geq 0$ for all x, we have

$$\lambda_1(A + B) = \min_{x \neq 0} \frac{x^T(A + B)x}{x^T x} = \min_{x \neq 0} \frac{x^T Ax + x^T Bx}{x^T x} \geq \min_{x \neq 0} \frac{x^T Bx}{x^T x} = \lambda_1(B),$$

which implies

$$\left\|(A + B)^{-1}\right\|_2 = \frac{1}{\lambda_1(A + B)} \leq \frac{1}{\lambda_1(B)} = \left\|B^{-1}\right\|_2,$$

which is the desired result.

□

Proof of Lemma 5

By definition of e and using the assumption on $\mathcal{O}$, for all $u \in \mathbb{R}^M$ we have

$$\begin{aligned} e(\varepsilon, u) &= \left\|\mathcal{G}^\varepsilon(u) - \mathcal{G}^0(u)\right\|_2 \\ &= \left\|\mathcal{O}(\mathcal{S}^\varepsilon(u)) - \mathcal{O}(\mathcal{S}^0(u))\right\|_2 \\ &= \left\|\mathcal{O}(p^\varepsilon) - \mathcal{O}(p^0)\right\|_2 \\ &\leq m \left\|p^\varepsilon - p^0\right\|_{L^2(\Omega)}. \end{aligned}$$

By homogenization theory, we know that $p^\varepsilon \rightharpoonup p^0$ in $H_0^1(\Omega)$, and, by Lemma 4, we obtain $p^\varepsilon \to p^0$ in $L^2(\Omega)$, which implies

$$e(\varepsilon, u) \to 0.$$

Moreover, if the solution of the homogenized problem p^0 is sufficiently regular, namely $p^0 \in H^2(\Omega)$, letting $C > 0$ be a constant, we have the following estimate, which can be found in [6]

$$\left\|p^\varepsilon - p^0\right\|_{L^2(\Omega)} \leq C\varepsilon,$$

hence we obtain

$$e(\varepsilon, u) \leq mC\varepsilon,$$

and we finally define $K = mC$.

□

Proof of Lemma 6

Let L be the Lipschitz constant of $\mathcal{G}$. For all $x \in B_R(u^*)$ we have

$$\|\mathcal{G}(x)\|_2 \leq \|\mathcal{G}(x) - \mathcal{G}(u^*)\|_2 + \|\mathcal{G}(u^*)\|_2 \leq L\|x - u^*\|_2 + \|\mathcal{G}(u^*)\|_2 \leq LR + G,$$
$$\|x\|_2 \leq \|x - u^*\|_2 + \|u^*\|_2 \leq R + g,$$

and we define the bounds $M = LR + G$ and $m = R + g$. The same bounds can be deduced for the mean values

$$\|\bar{u}\|_2 \leq \frac{1}{J}\sum_{j=1}^{J}\left\|u^{(j)}\right\|_2 \leq \frac{1}{J}Jm = m,$$
$$\left\|\bar{\mathcal{G}}\right\|_2 \leq \frac{1}{J}\sum_{j=1}^{J}\left\|\mathcal{G}(u^{(j)})\right\|_2 \leq \frac{1}{J}JM = M.$$

By definition of 2-norm of a matrix we have

$$\begin{aligned}
\|C^{up}(u)\|_2 &= \sup_{x\in\mathbb{R}^L:\ \|x\|_2=1}\left\|\frac{1}{J}\sum_{j=1}^{J}(u^{(j)} - \bar{u})(\mathcal{G}(u^{(j)}) - \bar{\mathcal{G}})^T x\right\|_2 \\
&\leq \sup_{x\in\mathbb{R}^L:\ \|x\|_2=1}\frac{1}{J}\sum_{j=1}^{J}\left|(\mathcal{G}(u^{(j)}) - \bar{\mathcal{G}})^T x\right|\left\|u^{(j)} - \bar{u}\right\|_2 \\
&\leq \sup_{x\in\mathbb{R}^L:\ \|x\|_2=1}\frac{1}{J}\sum_{j=1}^{J}\left\|\mathcal{G}(u^{(j)}) - \bar{\mathcal{G}}\right\|_2\|x\|_2\left\|u^{(j)} - \bar{u}\right\|_2,
\end{aligned}$$

and using (6.2) and (6.2) and the fact that $\|x\|_2 = 1$ we obtain

$$\begin{aligned}
\|C^{up}(u)\|_2 &\leq \frac{1}{J}\sum_{j=1}^{J}\left(\left\|\mathcal{G}(u^{(j)})\right\|_2 + \left\|\bar{\mathcal{G}}\right\|_2\right)\left(\left\|u^{(j)}\right\|_2 + \|\bar{u}\|_2\right) \\
&\leq \frac{1}{J}J(M+M)(m+m) = 4Mm,
\end{aligned}$$

and we define $C_1 = 4Mm$. The procedure is similar for the matrix $C^{pp}(u)$, where we have

$$\begin{aligned}
\|C^{pp}(u)\|_2 &= \sup_{x\in\mathbb{R}^L:\ \|x\|_2=1}\left\|\frac{1}{J}\sum_{j=1}^{J}(\mathcal{G}(u^{(j)}) - \bar{\mathcal{G}})(\mathcal{G}(u^{(j)}) - \bar{\mathcal{G}})^T x\right\|_2 \\
&\leq \sup_{x\in\mathbb{R}^L:\ \|x\|_2=1}\frac{1}{J}\sum_{j=1}^{J}\left|(\mathcal{G}(u^{(j)}) - \bar{\mathcal{G}})^T x\right|\left\|\mathcal{G}(u^{(j)}) - \bar{\mathcal{G}}\right\|_2 \\
&\leq \sup_{x\in\mathbb{R}^L:\ \|x\|_2=1}\frac{1}{J}\sum_{j=1}^{J}\left\|\mathcal{G}(u^{(j)}) - \bar{\mathcal{G}}\right\|_2^2\|x\|_2,
\end{aligned}$$

and using bound (6.2) and the fact that $\|x\|_2 = 1$ we obtain

$$\begin{aligned}
\|C^{up}(u)\|_2 &\leq \frac{1}{J}\sum_{j=1}^{J}\left(\left\|\mathcal{G}(u^{(j)})\right\|_2 + \left\|\bar{\mathcal{G}}\right\|_2\right)^2 \\
&\leq \frac{1}{J}J(M+M)^2 = 4M^2,
\end{aligned}$$

and we define $C_2 = 4M^2$.
Before proving the last two results of the lemma we need the following estimates for the ensemble

of particles u_1 and u_2

$$\begin{aligned}
\|\bar{u}_1 - \bar{u}_2\|_2 &= \left\| \frac{1}{J}\sum_{j=1}^{J}(u_1^{(j)} - u_2^{(j)}) \right\|_2 \leq \frac{1}{J}\sum_{j=1}^{J}\left\|u_1^{(j)} - u_2^{(j)}\right\|_2 = \|u_1 - u_2\|, \\
\left\|\bar{\mathcal{G}}_1 - \bar{\mathcal{G}}_2\right\|_2 &= \left\| \frac{1}{J}\sum_{j=1}^{J}(\mathcal{G}(u_1^{(j)}) - \mathcal{G}(u_2^{(j)})) \right\|_2 \leq \frac{1}{J}\sum_{j=1}^{J}\left\|\mathcal{G}(u_1^{(j)}) - \mathcal{G}(u_2^{(j)})\right\|_2 \\
&\leq L\frac{1}{J}\sum_{j=1}^{J}\left\|u_1^{(j)} - u_2^{(j)}\right\|_2 \\
&= L\|u_1 - u_2\|.
\end{aligned}$$

By definition of 2 norm of a matrix and using the triangle inequality we have

$$\begin{aligned}
&\|C^{up}(u_1) - C^{up}(u_2)\|_2 \\
&= \sup_{x\in\mathbb{R}^L:\ \|x\|_2=1} \left\| \frac{1}{J}\sum_{j=1}^{J}\left[(u_1^{(j)} - \bar{u}_1)(\mathcal{G}(u_1^{(j)}) - \bar{\mathcal{G}}_1)^T x - (u_2^{(j)} - \bar{u}_2)(\mathcal{G}(u_2^{(j)}) - \bar{\mathcal{G}}_2)^T x\right] \right\|_2 \\
&\leq \sup_{x\in\mathbb{R}^L:\ \|x\|_2=1} \frac{1}{J}\sum_{j=1}^{J}\left\|(u_1^{(j)} - \bar{u}_1)(\mathcal{G}(u_1^{(j)}) - \bar{\mathcal{G}}_1)^T x - (u_1^{(j)} - \bar{u}_1)(\mathcal{G}(u_2^{(j)}) - \bar{\mathcal{G}}_2)^T x\right\|_2 \\
&\quad + \sup_{x\in\mathbb{R}^L:\ \|x\|_2=1} \frac{1}{J}\sum_{j=1}^{J}\left\|(u_1^{(j)} - \bar{u}_1)(\mathcal{G}(u_2^{(j)}) - \bar{\mathcal{G}}_2)^T x - (u_2^{(j)} - \bar{u}_2)(\mathcal{G}(u_2^{(j)}) - \bar{\mathcal{G}}_2)^T x\right\|_2,
\end{aligned}$$

which implies

$$\begin{aligned}
&\|C^{up}(u_1) - C^{up}(u_2)\|_2 \\
&\leq \sup_{x\in\mathbb{R}^L:\ \|x\|_2=1} \frac{1}{J}\sum_{j=1}^{J}\left\|(u_1^{(j)} - \bar{u}_1)[(\mathcal{G}(u_1^{(j)}) - \mathcal{G}(u_2^{(j)})) + (\bar{\mathcal{G}}_2 - \bar{\mathcal{G}}_1)]^T x\right\|_2 \\
&\quad + \sup_{x\in\mathbb{R}^L:\ \|x\|_2=1} \frac{1}{J}\sum_{j=1}^{J}\left\|[(u_1^{(j)} - u_2^{(j)}) + (\bar{u}_2 - \bar{u}_1)](\mathcal{G}(u_2^{(j)}) - \bar{\mathcal{G}}_2)^T x\right\|_2 \\
&\leq \sup_{x\in\mathbb{R}^L:\ \|x\|_2=1} \frac{1}{J}\sum_{j=1}^{J}\left\|u_1^{(j)} - \bar{u}_1\right\|_2 [\left\|\mathcal{G}(u_1^{(j)}) - \mathcal{G}(u_2^{(j)})\right\|_2 + \left\|\bar{\mathcal{G}}_2 - \bar{\mathcal{G}}_1\right\|_2] \|x\|_2 \\
&\quad + \sup_{x\in\mathbb{R}^L:\ \|x\|_2=1} \frac{1}{J}\sum_{j=1}^{J}[\left\|u_1^{(j)} - u_2^{(j)}\right\|_2 + \|\bar{u}_2 - \bar{u}_1\|_2] \left\|\mathcal{G}(u_2^{(j)}) - \bar{\mathcal{G}}_2\right\| \|x\|_2.
\end{aligned}$$

Using bounds (6.2) and (6.2) and the fact that $\mathcal{G}$ is Lipschitz with constant L, we obtain

$$\begin{aligned}
&\|C^{up}(u_1) - C^{up}(u_2)\|_2 \\
&\leq \frac{1}{J}\sum_{j=1}^{J}\left\{(\left\|u_1^{(j)}\right\|_2 + \|\bar{u}_1\|_2)(L\left\|u_1^{(j)} - u_2^{(j)}\right\|_2 + L\|u_1 - u_2\|)\right\} \\
&\quad + \frac{1}{J}\sum_{j=1}^{J}\left\{(\left\|u_1^{(j)} - u_2^{(j)}\right\|_2 + \|\bar{u}_2 - \bar{u}_1\|_2)(\left\|\mathcal{G}(u_2^{(j)})\right\|_2 + \left\|\bar{\mathcal{G}}_2\right\|_2)\right\} \\
&\leq \frac{1}{J}\sum_{j=1}^{J}\{2m(LJ\|u_1 - u_2\| + L\|u_1 - u_2\|) + (J\|u_1 - u_2\| + \|u_1 - u_2\|)2M\} \\
&\leq 2(J+1)\max\{mL, M\}\|u_1 - u_2\|,
\end{aligned}$$

and we define $C_3 = 2(J+1)\max\{mL, M\}$. The computation is similar for the last point of the statement, for which we have

$$
\begin{aligned}
&\|C^{pp}(u_1) - C^{pp}(u_2)\|_2 \\
&\quad = \sup_{x\in\mathbb{R}^L:\ \|x\|_2=1} \left\| \frac{1}{J}\sum_{j=1}^{J} \left[(\mathcal{G}(u_1^{(j)}) - \bar{\mathcal{G}}_1)(\mathcal{G}(u_1^{(j)}) - \bar{\mathcal{G}}_1)^T x - (\mathcal{G}(u_2^{(j)}) - \bar{\mathcal{G}}_2)(\mathcal{G}(u_2^{(j)}) - \bar{\mathcal{G}}_2)^T x \right] \right\|_2 \\
&\quad \leq \sup_{x\in\mathbb{R}^L:\ \|x\|_2=1} \frac{1}{J}\sum_{j=1}^{J} \left\| (\mathcal{G}(u_1^{(j)}) - \bar{\mathcal{G}}_1)(\mathcal{G}(u_1^{(j)}) - \bar{\mathcal{G}}_1)^T x - (\mathcal{G}(u_1^{(j)}) - \bar{\mathcal{G}}_1)(\mathcal{G}(u_2^{(j)}) - \bar{\mathcal{G}}_2)^T x \right\|_2 \\
&\qquad + \sup_{x\in\mathbb{R}^L:\ \|x\|_2=1} \frac{1}{J}\sum_{j=1}^{J} \left\| (\mathcal{G}(u_1^{(j)}) - \bar{\mathcal{G}}_1)(\mathcal{G}(u_2^{(j)}) - \bar{\mathcal{G}}_2)^T x - (\mathcal{G}(u_2^{(j)}) - \bar{\mathcal{G}}_2)(\mathcal{G}(u_2^{(j)}) - \bar{\mathcal{G}}_2)^T x \right\|_2,
\end{aligned}
$$

which implies

$$
\begin{aligned}
&\|C^{pp}(u_1) - C^{pp}(u_2)\|_2 \\
&\quad \leq \sup_{x\in\mathbb{R}^L:\ \|x\|_2=1} \frac{1}{J}\sum_{j=1}^{J} \left\| (\mathcal{G}(u_1^{(j)}) - \bar{\mathcal{G}}_1)[(\mathcal{G}(u_1^{(j)}) - \mathcal{G}(u_2^{(j)})) + (\bar{\mathcal{G}}_2 - \bar{\mathcal{G}}_1)]^T x] \right\|_2 \\
&\qquad + \sup_{x\in\mathbb{R}^L:\ \|x\|_2=1} \frac{1}{J}\sum_{j=1}^{J} \left\| [(\mathcal{G}(u_1^{(j)}) - \mathcal{G}(u_2^{(j)})) + (\bar{\mathcal{G}}_2 - \bar{\mathcal{G}}_1)](\mathcal{G}(u_2^{(j)}) - \bar{\mathcal{G}}_2)^T x \right\|_2, \\
&\quad \leq \sup_{x\in\mathbb{R}^L:\ \|x\|_2=1} \frac{1}{J}\sum_{j=1}^{J} \left\| \mathcal{G}(u_1^{(j)}) - \bar{\mathcal{G}}_1 \right\|_2 [\left\| \mathcal{G}(u_1^{(j)}) - \mathcal{G}(u_2^{(j)}) \right\|_2 + \left\| \bar{\mathcal{G}}_2 - \bar{\mathcal{G}}_1 \right\|_2] \|x\|_2 \\
&\qquad + \sup_{x\in\mathbb{R}^L:\ \|x\|_2=1} \frac{1}{J}\sum_{j=1}^{J} [\left\| \mathcal{G}(u_1^{(j)}) - \mathcal{G}(u_2^{(j)}) \right\|_2 + \left\| \bar{\mathcal{G}}_2 - \bar{\mathcal{G}}_1 \right\|_2] \left\| \mathcal{G}(u_2^{(j)}) - \bar{\mathcal{G}}_2 \right\| \|x\|_2 .
\end{aligned}
$$

Using bounds (6.2) and (6.2) and the fact that $\mathcal{G}$ is Lipschitz with constant L, we obtain

$$
\begin{aligned}
&\|C^{pp}(u_1) - C^{pp}(u_2)\|_2 \\
&\quad \leq \frac{1}{J}\sum_{j=1}^{J} \{2M(LJ\|u_1 - u_2\| + L\|u_1 - u_2\|) + (LJ\|u_1 - u_2\| + L\|u_1 - u_2\|)2M\} \\
&\quad = 4ML(J+1)\|u_1 - u_2\|,
\end{aligned}
$$

and we define $C_4 = 4ML(J+1)$.

□

Proof of Theorem 2

We recall the duality formula (4) for the Wasserstein distance $W_{1,s}$

$$
W_{1,s}(\mu_n, \mu) = \sup_{\varphi\in\Phi} \left\{ \int_{B_R(u^*)} \varphi d(\mu_n - \mu) \right\},
$$

where Φ is the set of all globally Lipschitz continuous functions $\varphi\colon B_R(u^*) \to \mathbb{R}$ with Lipschitz constant $L \leq 1$. Note that if $\varphi \in \Phi$, then also $-\varphi \in \Phi$. Therefore we deduce that

$$
W_{1,s}(\mu_n, \mu) = \sup_{\varphi\in\Phi} \left\{ \int_{B_R(u^*)} \varphi d(\mu_n - \mu) \right\} = \sup_{\varphi\in\Phi} \left\{ \left| \int_{B_R(u^*)} \varphi d(\mu_n - \mu) \right| \right\}.
$$

Indeed we have

$$\int_{B_R(u^*)} \varphi d(\mu_n - \mu) \leq \left| \int_{B_R(u^*)} \varphi d(\mu_n - \mu) \right|,$$

which implies the first inequality

$$\sup_{\varphi \in \Phi} \left\{ \int_{B_R(u^*)} \varphi d(\mu_n - \mu) \right\} \leq \sup_{\varphi \in \Phi} \left\{ \left| \int_{B_R(u^*)} \varphi d(\mu_n - \mu) \right| \right\}.$$

On the other hand, we also have

$$A = \left\{ \int_{B_R(u^*)} \varphi d(\mu_n - \mu) \colon \varphi \in \Phi \right\} \supseteq \left\{ \left| \int_{B_R(u^*)} \varphi d(\mu_n - \mu) \right| : \varphi \in \Phi \right\} = A',$$

because if $c \in A'$, which means that there exists $\varphi \in \Phi$ such that

$$c = \left| \int_{B_R(u^*)} \varphi d(\mu_n - \mu) \right|,$$

then we can take $\tilde{\varphi} \in \Phi$ defined as

$$\tilde{\varphi} = \begin{cases} \varphi & \text{if } \int_{B_R(u^*)} \varphi d(\mu_n - \mu) > 0 \\ -\varphi & \text{if } \int_{B_R(u^*)} \varphi d(\mu_n - \mu) < 0, \end{cases}$$

and note that that

$$c = \int_{B_R(u^*)} \tilde{\varphi} d(\mu_n - \mu),$$

which implies that $c \in A$. Therefore, by (6.2), we deduce the opposite inequality

$$\sup_{\varphi \in \Phi} \left\{ \int_{B_R(u^*)} \varphi d(\mu_n - \mu) \right\} \geq \sup_{\varphi \in \Phi} \left\{ \left| \int_{B_R(u^*)} \varphi d(\mu_n - \mu) \right| \right\}.$$

Then, thanks to (6.2), we have

$$\begin{aligned} \sup_{\varphi \in \Phi} \mathbb{E}_\xi \left[\left| \int_{B_R(u^*)} \varphi d(\mu_n - \mu) \right| \right] &\leq \mathbb{E}_\xi \left[\sup_{\varphi \in \Phi} \left\{ \left| \int_{B_R(u^*)} \varphi d(\mu_n - \mu) \right| \right\} \right] \\ &= \mathbb{E}_\xi [W_{1,s}(\mu_n, \mu)], \end{aligned}$$

and the right hand side vanishes by hypothesis, so we obtain

$$\sup_{\varphi \in \Phi} \mathbb{E}_\xi \left[\left| \int_{B_R(u^*)} \varphi d(\mu_n - \mu) \right| \right] \to 0.$$

Hence

$$\mathbb{E}_\xi \left[\left| \int_{B_R(u^*)} \varphi d\mu_n - \int_{B_R(u^*)} \varphi d\mu \right| \right] \to 0 \qquad \text{for all } \varphi \in \Phi.$$

It remains to show that (6.2) holds true for all functions $f \in C^0(B_R(u^*))$. First, we consider any Lipschitz function ψ with Lipschitz constant L. We define $\varphi = \psi / L$, then $\varphi \in \Phi$, indeed

$$|\varphi(x) - \varphi(y)| = \left| \frac{1}{L}\psi(x) - \frac{1}{L}\psi(y) \right| = \frac{1}{L} |\psi(x) - \psi(y)| \leq \frac{1}{L} L \|x - y\|_s = \|x - y\|_s.$$

Therefore we have

$$\mathbb{E}_\xi \left[\left| \int_{B_R(u^*)} \psi d\mu_n - \int_{B_R(u^*)} \psi d\mu \right| \right] = L \mathbb{E}_\xi \left[\left| \int_{B_R(u^*)} \varphi d\mu_n - \int_{B_R(u^*)} \varphi d\mu \right| \right] \to 0.$$

By density, any continuous bounded function $f \in C^0(B_R(u^*))$ can be approximated by a sequence of Lipschitz functions $\{\psi_k\}_{k\in\mathbb{N}}$ such that $\|\psi_k\|_{L^\infty(B_R(u^*))} \leq C$ for all $k \in \mathbb{N}$ where C is a constant dependent on f and $\|\psi_k - f\|_{L^\infty(B_R(u^*))} \to 0$ as $k \to \infty$. Thanks to (6.2) we have

$$\mathbb{E}_\xi \left[\left| \int_{B_R(u^*)} \psi_k d\mu_n - \int_{B_R(u^*)} \psi_k d\mu \right| \right] \to 0,$$

and, applying Lebesgue dominated convergence theorem, we can pass to the limit as $k \to \infty$. We can exchange the limit with the expectation and the integral because the integrand functions are bounded by C and the measures μ_n and μ are finite, since they are probability measures. Thus we obtain

$$\mathbb{E}_\xi \left[\left| \int_{B_R(u^*)} f d\mu_n - \int_{B_R(u^*)} f d\mu \right| \right] \to 0,$$

for all bounded continuous functions $f \in C^0(B_R(u^*))$ which means

$$\mu_n \xrightarrow{L^1} \mu,$$

which is the desired result.

□

References

[1] A. Abdulle and A. Di Blasio, *A Bayesian numerical homogenization method for elliptic multiscale inverse problems*. Submitted to SIAM UQ, 2018.

[2] A. Abdulle, W. E, B. Engquist, and E. Vanden-Eijnden, *The heterogeneous multiscale method*, Acta Numer., 21 (2012), pp. 1–87.

[3] D. Calvetti, M. Dunlop, E. Somersalo, and A. Stuart, *Iterative updating of model error for Bayesian inversion*, Inverse Problems, 34 (2018), pp. 025008, 38.

[4] D. Cioranescu and P. Donato, *An introduction to homogenization*, vol. 17 of Oxford Lecture Series in Mathematics and its Applications, Oxford University Press, New York, 1999.

[5] L. C. Evans, *Partial differential equations*, vol. 19 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, second ed., 2010.

[6] G. Griso, *Interior error estimate for periodic homogenization*, C. R. Math. Acad. Sci. Paris, 340 (2005), pp. 251–254.

[7] M. A. Iglesias, K. J. H. Law, and A. M. Stuart, *Ensemble Kalman methods for inverse problems*, Inverse Problems, 29 (2013), pp. 045001, 20.

[8] J. Kaipio and E. Somersalo, *Statistical and computational inverse problems*, vol. 160 of Applied Mathematical Sciences, Springer-Verlag, New York, 2005.

[9] J. Nolen, G. A. Pavliotis, and A. M. Stuart, *Multiscale modeling and inverse problems*, in Numerical analysis of multiscale problems, vol. 83 of Lect. Notes Comput. Sci. Eng., Springer, Heidelberg, 2012, pp. 1–34.

[10] A. Quarteroni, *Numerical Models for Differential Problems*, vol. 2 of Modeling, Simulation & Applications, Springer, 2009.

[11] S. Salsa, *Partial differential equations in action*, vol. 99 of Unitext, Springer, [Cham], third ed., 2016. From modelling to theory, La Matematica per il 3+2.

[12] F. Santambrogio, *Optimal transport for applied mathematicians*, vol. 87 of Progress in Nonlinear Differential Equations and their Applications, Birkhäuser/Springer, Cham, 2015. Calculus of variations, PDEs, and modeling.

[13] C. Schillings and A. M. Stuart, *Analysis of the ensemble Kalman filter for inverse problems*, SIAM J. Numer. Anal., 55 (2017), pp. 1264–1290.