Ensemble Kalman filter for multiscale inverse problems

Assyr Abdulle* Giacomo Garegnani* Andrea Zanoni*

Abstract

We present a novel algorithm based on the ensemble Kalman filter to solve inverse problems involving multiscale elliptic partial differential equations. Our method is based on numerical homogenization and finite element discretization and allows to recover a highly oscillatory tensor from measurements of the multiscale solution in a computationally inexpensive manner. The properties of the approximate solution are analysed with respect to the multiscale and discretization parameters, and a convergence result is shown to hold. A reinterpretation of the solution from a Bayesian perspective is provided, and convergence of the approximate conditional posterior distribution is proved with respect to the Wasserstein distance. A numerical experiment validates our methodology, with a particular emphasis on modelling error and computational cost.

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

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

1 Introduction

In this work we consider the application of techniques derived from the Kalman filter to inverse problems involving multiscale phenomena which can be modelled by means of partial differential equations (PDEs). Inverse problems arise in many fields, such as seismography, meteorology and tomography, all physical domains with a multiscale nature. Our reference mathematical model is given by multiscale elliptic PDEs of the form

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

where $\Omega \subset \mathbb{R}^d$ is the physical domain, A_u^ε is a tensor oscillating with an amplitude described by the parameter ε and u is a possibly infinite-dimensional unknown which parametrizes the tensor A_u^ε. We are then interested in the solution of inverse problems involving the retrieval of the parameter u given noisy observations derived from the solution p^ε.

Multiscale inverse problems of this form have been recently introduced in [12] and analysed extensively in [1, 2]. In particular, in [2] Abdulle and Di Blasio build a coarse-graining approach to solve the inverse problem regularized with a Tikhonov technique. The main idea is replacing the computationally expensive solution of the highly-oscillating multiscale problem with an homogenized surrogate, which eliminates the fast variables and is therefore cheaper. In particular, the theory of homogenization guarantees under certain assumptions, which will be specified throughout this work, that there exists a PDE of the form

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

*Institute of Mathematics, École Polytechnique Fédérale de Lausanne

such that the solution p^0 is the weak limit of the functions p^ε in the vanishing limit for ε, and such that A_u^0 is independent of ε. In [2], the authors showed that employing this homogenized model to the multiscale inverse problem guarantees a good approximation to its solution if a Tikhonov regularization is employed. This framework has been successively enlarged by the same authors to the Bayesian case in [2], where the analysis involves posterior distributions arising from both the multiscale and the homogenized model. In the same work, a technique for estimating the modelling error which was developed in [4, 5] is successfully applied to multiscale inverse problems to account for the homogenization and discretization errors.

The ensemble Kalman filter (EnKF), first introduced in [9], is an algorithm which is widely employed in the engineering community for the estimation of the state of partially-observed dynamical systems whose dynamics are governed by a nonlinear agent. In particular, Kalman filters have long been used successfully in meteorology, oceanography and automation applications. In [11], Iglesias et al. propose the application of the EnKF method to obtain a point-wise solution to inverse problems involving PDEs, and an extension of their analysis giving a Bayesian interpretation of the filtering solution is presented in [15].

In this work, we present a combination of the well-established techniques of homogenization and filtering to build a novel scheme for solving multiscale inverse problems in an efficient and reliable manner. In the same spirit of [1, 2], we prove that it is possible to eliminate the fast scales from the PDE appearing in the inverse problem relying on the theory of homogenization, thus obtaining a solution which is accurate in the vanishing limit for the multiscale parameter ε. In our analysis, we both consider point-wise estimations as in [11] and Bayesian solutions as in [15], thus showing convergence results which are endowed with decay rates under special assumptions on the problem. Inspired by [1, 4, 5], we then consider offline and online techniques for estimating the modelling error and prove a novel result indicating the computational cost which is required for such an estimation for any given multiscale problem.

We identified two main advantages of a filtering approach as the one provided by the EnKF method with respect to other approaches. First, a Bayesian interpretation of the solution to the inverse problem is obtained from the algorithm without any additional cost with respect to a point-wise estimation. A distribution on the unknown provides in fact with a deeper insight and a full uncertainty quantification on the solution, which is therefore in turn more interpretable and robust. Secondly, the EnKF can be simply divided in sequential parallel runs, which we verified in practice to allow a faster computation of the solution of rather complex inverse problems with respect to standard Monte Carlo approaches, such as the Metropolis–Hastings algorithm.

The outline of the work is the following. In Section 2 we introduce the setting of the problem in a rigorous manner, as well as the notation which will be employed throughout this work. Then, in Section 3, we briefly summarize the techniques introduced in [11, 15] to solve inverse problems both in a point-wise and in a Bayesian spirit employing the EnKF method. In Section 4 we present the results of convergence of the EnKF scheme in the multiscale setting, which is the main contribution of this work. Then, Section 5 is dedicated to the estimation of the modelling error, and to a novel theoretical results which strengthens its value in practice. Finally, Section 6 is devoted to numerical experiments which corroborate our analysis.

2 Problem setting

Given a positive parameter ε, let us consider the multiscale inverse problem

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

where the parameter space X and the observation space Y are Hilbert spaces, the multiscale forward operator $\mathcal{G}^\varepsilon \colon X \to Y$ maps the unknown to the observation space, and $\eta \in Y$ is a source of additive noise, which we assume to be distributed as a Gaussian $\mathcal{N}(0, \Gamma)$, where Γ is a positive definite covariance operator. We assume that 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 the observation operator and $\mathcal{S}^\varepsilon \colon X \to H_0^1(\Omega)$ is the solution operator of

a multiscale elliptic partial differential equation (PDE). Letting Ω be a bounded open domain in $\mathbb{R}^d$, the operator $\mathcal{S}^\varepsilon$ maps the unknown u to the solution p^ε of

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

Let us introduce a regularity assumption for the observation operator.

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

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

where m is a positive constant.

Note that since $\mathcal{O}$ is defined on $H_0^1 \subset L^2$, Assumption 1 is stronger than Lipschitz continuity. The tensor A_u^ε belongs to the class of parametrized multiscale tensors which admit explicit scale separation between slow and fast spatial variables, i.e.,

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

where the map $(t, x) \to A(t, x/\varepsilon)$ is assumed to be known and A is periodic in its second argument. If ε is small, a fine discretization is needed to resolve the smallest scale and thus evaluate the forward operator $\mathcal{G}^\varepsilon$, 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.

In order to approach the multiscale problem in a more efficient manner we therefore apply the theory of homogenization (see e.g. [7]), which ensures the existence of an homogenized tensor A^0, such that for $\varepsilon \to 0$ the solution p^ε of (2) tends weakly in $H^1(\Omega)$ to the solution p^0 of the problem

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

Hence, the function p^0 can be assumed to be a good approximation of p^ε when the multiscale parameter ε is small and thus, in this case, the multiscale problem (2) can be replaced by its homogenized version (3). Let us denote by $\mathcal{G}_h^0\colon \mathcal{O} \circ \mathcal{S}_h^0$ the forward operator which maps the unknown parameter into the solution of (3) computed with the finite element method (FEM) with discretization parameter $h > 0$. Inspired by [1, 2, 12], we employ the homogenized operator $\mathcal{G}_h^0$, which is cheaper to evaluate numerically, to solve the inverse problem (**??**) Finally, let us introduce a regularity assumption on the multiscale and homogenzed tensors A^ε and A^0.

Assumption 2. The tensors A^ε and A^0 in (2) and (3) respectively satisfy for all $u, u_1, u_2 \in X$ and $\xi \in \mathbb{R}^d$

$$\begin{aligned} &\|A^\varepsilon(u_1) - A^\varepsilon(u_2)\|_{L^\infty(\Omega;\mathbb{R}^{d\times d})} \leq M \|u_1 - u_2\|_X, \\ &\left\|A^0(u_1) - A^0(u_2)\right\|_{L^\infty(\Omega;\mathbb{R}^{d\times d})} \leq M \|u_1 - u_2\|_X, \\ &A^\varepsilon(u)\xi \cdot \xi \geq \alpha_0 \|\xi\|_2^2, \quad A^0(u)\xi \cdot \xi \geq \alpha_0 \|\xi\|_2^2, \end{aligned}$$

where M and α_0 are positive constants.

3 A Kalman filter solution to inverse problems

In this section, we present a technique to solve an inverse problem of the form (1) based on the ensemble Kalman filter (EnKF). Further details about the EnKF and its applications to inverse problems ought to be found in [11, 15], where the authors develop a framework for inverse problems involving single-scale PDE models. Inverse problems are often ill-posed and hence require

regularization, which can be granted by, for example, variational and Bayesian techniques. The EnKF method achieves regularization by searching for the solution of the inverse problem in a finite dimensional and compact subset $\mathcal{A}$ of X, which incorporates prior knowledge of u. In this section, we consider a general forward map $\mathcal{G}\colon X \to Y$ and the observations to be given by $y = \mathcal{G}(u) + \eta$ for an unknown $u \in X$ and a Gaussian noise $\eta \sim \mathcal{N}(0, \Gamma)$.

The traditional theory of Kalman filters involves the estimation of the state of a dynamical system which is observed in a partial and noisy manner. Therefore, in order to approximate the unknown by means of the Kalman filter theory, we need to introduce artificial dynamics based on state augmentation. Given the space $Z = X \times Y$, let us define the map $\Xi\colon Z \to Z$ as

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

which is employed to construct artificial dynamics as

$$z_{n+1} = \Xi(z_n). \tag{4}$$

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 $\eta_n \sim \mathcal{N}(0, \Gamma)$, i.e., with the same distribution as the noise in (1). Consequently, we get

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

The EnKF method, which we briefly describe here, proceeds by propagating an ensemble $\{z_n^{(j)}\}_{j=1}^J \subset Z$ of J particles for $n = 0, \dots, N$, following the classical Kalman update formulae. Let $\mathcal{A} \subset X$ be such that $\dim(\mathcal{A}) \leq J$, and let $\{\psi^{(j)}\}_{j=1}^J \subset \mathcal{A}$. The initial ensemble $\{z_0^{(j)}\}_{j=1}^J$ is then built as

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

Each iteration of the EnKF method can be split in two parts, the prediction and analysis steps. At the n-th step of the algorithm, the current ensemble of particles $\{z_n^{(j)}\}_{j=1}^J$ is first mapped forward through the augmented dynamics (4), i.e., for all $j = 1, \dots, J$

$$\hat{z}_{n+1}^{(j)} = \Xi(z_n^{(j)}). \tag{5}$$

Let us remark that this step introduces information on the forward model due to how the second component of the map Ξ is defined, which implies that the partially-updated ensemble $\{\hat{z}_{n+1}^{(j)}\}_{j=1}^J$ can be interpreted as a prior estimate. In the analysis step, the ensemble is updated comparing the prior estimate (5) with versions of the data perturbed with noise $\{y_{n+1}^{(j)}\}_{j=1}^J$, where $y_{n+1}^{(j)} = y + \eta_{n+1}^{(j)}$, via the standard Kalman update formula

$$z_{n+1}^{(j)} = \hat{z}_{n+1}^{(j)} + K_{n+1}(y_{n+1}^{(j)} - H\hat{z}_{n+1}^{(j)}), \tag{6}$$

where K_{n+1}, the Kalman gain, is defined as

$$K_{n+1} = C_{n+1}H^*(HC_{n+1}H^* + \Gamma)^{-1}, \tag{7}$$

and is employed to weigh the prior guess provided by Ξ and the information carried by the observations. In the definition (7), the operator H^* is the adjoint of H and the matrix C_{n+1} is the empirical covariance matrix of the partially-updated ensemble $\{\hat{z}_{n+1}^{(j)}\}_{j=1}^J$. At the N-th and final step, the EnKF estimator is obtained by averaging over the ensemble of particles projected on the space X, i.e.,

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

where $H^\perp : Z \to X$ is defined by $H^\perp = \begin{bmatrix} I & 0 \end{bmatrix}$.

The last detail needed to run the EnKF algorithm is the definition of the initial ensemble, which is closely related to the choice of the space $\mathcal{A}$. In particular, the space $\mathcal{A}$ should incorporate all the prior knowledge on the solution of the inverse problem. Let us assume that the prior knowledge can be summarized as a probability measure on X denoted by μ_0. A possible choice for initialization is then to generate the set $\{\psi^{(j)}\}_{j=1}^J$ as J draws from μ_0, and to fix the set $\mathcal{A}$ as

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

Remark 1. The cost of the EnKF method can be measured in terms of the number of evaluations of the forward operator, which is indeed dominating the other operations in terms of computational time. 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. Nonetheless, let us remark that the ensemble Kalman method can be easily parallelized, since at each iteration the forward operator is applied independently to each particle. Hence, if the number of computing threads N_{comp} available is such that $N_{\text{comp}} = \mathcal{O}(J)$, we have that the overall cost is of order $\mathcal{O}(N)$.

In [15], the authors show that the EnKF admits a Bayesian interpretation, which we briefly summarise here. Given a prior distribution μ_0 on X, the posterior distribution μ of the unknown given the data is defined as (see e.g. [16])

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

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

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

The map from the prior distribution μ_0 to the posterior μ can be divided in N sub-steps through 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 and that

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

The distribution μ_n can then be approximated by the discrete probability measure induced by the particles of the EnKF method at the n-th step, i.e.,

$$\mu_n \simeq \frac{1}{J} \sum_{j=1}^J \delta_{u_n^{(j)}}. \tag{8}$$

The particles at time n are still mapped into the the particles at time $n + 1$ with the ensemble Kalman filter update formula described above, with the slight modification that Γ has to be replaced by $\Delta^{-1}\Gamma$ (see [15] for the details).

To conclude this section, we 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 value of the unknown.

4 Convergence analysis

In this section we show the convergence of the ensemble of particles generated by the EnKF algorithm which employs the multiscale forward operator $\mathcal{G}^\varepsilon$ in its update formulae to the ensemble

which is given by the application of the same algorithm with the FEM solution to the homogenized problem, i.e., by the forward operator $\mathcal{G}_h^0$, in the limit $\varepsilon, h \to 0$. Moreover, from the Bayesian perspective, we show the convergence of the associated posterior distributions, thus providing under further assumptions a rate of convergence. The analysis is carried out in the finite dimensional case $X = \mathbb{R}^M$ and $Y = \mathbb{R}^L$, but it can be generalized to the infinite dimensional setting. Summarizing, the forward operators involved 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 solution operator associated to (3), which maps the unknown u to the exact solution p^0. Convergence is shown in the ensemble norm, which we define for an ensemble of particles $u = \{u^{(j)}\}_{j=1}^J$ as

$$\|u\| \coloneqq \frac{1}{J} \sum_{j=1}^{J} \left\| u^{(j)} \right\|_2, \tag{9}$$

and which is indeed a norm. Given a scalar α, let us finally define the linear combination $w = u + \alpha v$ between two ensembles u and v with the same number of particles as $\{w^{(j)} = u^{(j)} + \alpha v^{(j)}\}_{j=1}^J$.

4.1 Convergence of the point estimate

We first consider the convergence of the particle ensembles, which can be summarized by the following theorem.

Theorem 1. *Let $u_{N,h}^0 = \{u_{N,h}^{0(j)}\}_{j=1}^J$, $u_N^\varepsilon = \{u_N^{\varepsilon(j)}\}_{j=1}^J$ be the ensembles after N iterations of the EnKF method for the forward operators $\mathcal{G}_h^0$ and $\mathcal{G}^\varepsilon$ respectively. Then, under Assumption 1, Assumption 2 and Assumption 3, we have*

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

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 employ polynomials of degree r for the finite element basis, then

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

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

It is clear from the statement of Theorem 1 that the effects of homogenization and discretization can be analysed separately. In particular, we first show the convergence of the ensemble generated employing the forward operator $\mathcal{G}^\varepsilon$ to the one generated employing the exact homogenized operator $\mathcal{G}^0$ for $\varepsilon \to 0$. Then, in an analogous fashion, we prove the convergence of the ensemble generated with $\mathcal{G}_h^0$ to the ensemble generated employing $\mathcal{G}^0$. In order to introduce a compact notation, we denote by $\mathcal{U}_{J,M}$ the set of ensembles of dimension J with elements in $\mathbb{R}^M$ and we consider the homogenization error function $e \colon \mathbb{R} \times \mathcal{U}_{J,M} \to \mathbb{R}$, which is defined for a generic ensemble u as

$$e(\varepsilon, u) = \frac{1}{J} \sum_{j=1}^{J} \left\| \mathcal{G}^\varepsilon(u^{(j)}) - \mathcal{G}^0(u^{(j)}) \right\|_2, \tag{10}$$

and a discretization error function $\tilde{e} \colon \mathbb{R} \times \mathcal{U}_{J,M} \to \mathbb{R}$ as

$$\tilde{e}(h, u) = \frac{1}{J} \sum_{j=1}^{J} \left\| \mathcal{G}_h^0(u^{(j)}) - \mathcal{G}^0(u^{(j)}) \right\|_2. \tag{11}$$

Before proving the main theorem, we introduce some preliminary results. In particular, Lemma 1 and Lemma 2 are linear algebra results. In Lemma 3 we prove Lipschitz continuity of forward operators which involve a Lipschitz observation operator and the solution of an elliptic PDE, and in Lemma 4 we show that the homogenization error defined in (10) vanishes in the limit $\varepsilon \to 0$. Finally, in Lemma 5 we consider the particle empirical covariances of ensembles given by the EnKF algorithm, thus proving their boundedness and Lipschitz continuity. The proof of all the results above can be found in the Appendix.

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

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

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

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

Lemma 3. *Let $\mathcal{G}\colon \mathbb{R}^M \to \mathbb{R}^L$, $\mathcal{G} = \mathcal{O}\circ\mathcal{S}$ be a forward operator such that $\mathcal{O}\colon H_0^1(\Omega)\to\mathbb{R}^L$ is Lipschitz and $\mathcal{S}\colon \mathbb{R}^M \to H_0^1(\Omega)$, $\mathcal{S}\colon u\mapsto p$ is defined by the solution of*

$$\begin{cases} -\nabla\cdot(A(u)\nabla p) = f, & \text{in } \Omega, \\ p = 0, & \text{on } \partial\Omega, \end{cases} \tag{12}$$

where $\Omega\subset\mathbb{R}^d$ is an open bounded set, the right hand side $f\in L^2(\Omega)$ and the tensor $A(u)\in L^\infty(\Omega;\mathbb{R}^{d\times d})$ satisfies

$$\|A(u_1)-A(u_2)\|_{L^\infty(\Omega;\mathbb{R}^{d\times d})} \leq M\|u_1-u_2\|_2, \qquad \text{for all } u_1,u_2\in\mathbb{R}^M,$$

and

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

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

Lemma 4. *Let e be defined as in* (10)*. Under Assumption 1, we have for all $u\in\mathcal{U}_{J,M}$*

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

Moreover, if the solution of the homogenized problem (3) *is sufficiently smooth independently of u, namely $p^0\in H^2(\Omega)$, then there exists $K>0$ independent of ε and u such that*

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

Lemma 5. *Let $C^{up}(u)\in\mathbb{R}^{M\times L}$ and $C^{pp}(u)\in\mathbb{R}^{L\times L}$ be defined as*

$$C^{up}(u) = \frac{1}{J}\sum_{j=1}^J \left(u^{(j)}-\bar{u}\right)\left(\mathcal{G}(u^{(j)})-\bar{\mathcal{G}}\right)^T, \qquad C^{pp}(u) = \frac{1}{J}\sum_{j=1}^J \left(\mathcal{G}(u^{(j)})-\bar{\mathcal{G}}\right)\left(\mathcal{G}(u^{(j)})-\bar{\mathcal{G}}\right)^T,$$

where $\bar{u}\in\mathbb{R}^M$ and $\bar{\mathcal{G}}\in\mathbb{R}^L$ are the empirical averages

$$\bar{u} = \frac{1}{J}\sum_{j=1}^J u^{(j)}, \qquad \bar{\mathcal{G}} = \frac{1}{J}\sum_{j=1}^J \mathcal{G}(u^{(j)}),$$

and let $\mathcal{G}\colon\mathbb{R}^M\to\mathbb{R}^L$ be Lipschitz with constant L. Then, there exist four constants $C_i>0$, $i=1,\dots,4$, such that

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

for all ensembles $u,u_1,u_2\in\mathcal{U}_{J,M}$ which are stable in the sense of Assumption 3.

In order to clarify the exposition, we first consider the amplification the error over one step between the EnKF algorithms employing the multiscale and the homogenized forward operators respectively, which is summarized in the following lemma.

Lemma 6. *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, there exist positive constants α and γ such that*

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

where $e(\varepsilon,u)$ is given in (10)*.*

Proof. First, due to Assumption 1 and the Poincaré inequality with constant C_p we have

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

which shows that $\mathcal{O}$ is Lipschitz with constant mC_p. Therefore, applying Lemma 3, we deduce that both $\mathcal{G}^0$ and $\mathcal{G}^\varepsilon$ are Lipschitz with constant $L_\mathcal{G}$ independent of ε. The Kalman update formulae (6) restricted to the u variable read (see [11])

$$u_{n+1}^{\varepsilon^{(j)}} = u_n^{\varepsilon^{(j)}} + C^{up}(u_n^\varepsilon)(C^{pp}(u_n^\varepsilon) + \Gamma)^{-1}(y_{n+1} - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}})), \tag{13}$$

$$u_{n+1}^{0^{(j)}} = u_n^{0^{(j)}} + C^{up}(u_n^0)(C^{pp}(u_n^0) + \Gamma)^{-1}(y_{n+1} - \mathcal{G}^0(u_n^{0^{(j)}})). \tag{14}$$

Combining (13) and (14), we have

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

and using the triangle inequality we obtain

$$\mathbb{E}\left[\left\|u_{n+1}^\varepsilon - u_{n+1}^0\right\|\right] \leq \mathbb{E}\left[\left\|u_n^\varepsilon - u_n^0\right\|\right] + S_1 + S_2 + S_3, \tag{15}$$

where

$$S_1 = \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\left[\left\|C^{up}(u_n^\varepsilon) - C^{up}(u_n^0)\right\|_2 \left\|(C^{pp}(u_n^\varepsilon) + \Gamma)^{-1}\right\|_2 \left\|y_{n+1}^{(j)} - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}})\right\|_2\right], \tag{16}$$

$$S_2 = \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\left[\left\|C^{up}(u_n^0)\right\|_2 \left\|(C^{pp}(u_n^\varepsilon) + \Gamma)^{-1} - (C^{pp}(u_n^0) + \Gamma)^{-1}\right\|_2 \left\|y_{n+1}^{(j)} - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}})\right\|_2\right], \tag{17}$$

$$S_3 = \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\left[\left\|C^{up}(u_n^0)\right\|_2 \left\|(C^{pp}(u_n^0) + \Gamma)^{-1}\right\|_2 \left\|\mathcal{G}^0(u_n^{0^{(j)}}) - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}})\right\|_2\right]. \tag{18}$$

Let us first consider S_1. Due to Lemma 5, we have

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

and due to Lemma 2

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

Moreover, due to the definition of $y_{n+1}^{(j)}$ and since $y = \mathcal{G}^\varepsilon(u^*) + \eta$, where u^* is the true value of the unknown and η is the true realization of the noise, we obtain the bound

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

which, since $\mathcal{G}^\varepsilon$ is Lipschitz, yields

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

Thus (16) can be bounded by

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

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

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

Moreover, the random variable $\zeta \coloneqq \xi + \eta$ is distributed by independence as $\zeta \sim \mathcal{N}(0, 2\Gamma)$, and therefore we get

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

and defining $\alpha_1 \coloneqq C_3 \left\|\Gamma^{-1}\right\|_2 (L_{\mathcal{G}} R + \sqrt{2\mathrm{tr}(\Gamma)})$, the final bound for S_1 reads

$$S_1 \leq \alpha_1 \mathbb{E}\left[\left\|u_n^\varepsilon - u_n^0\right\|\right]. \tag{19}$$

Let us now consider the second term S_2. Due to Lemma 5, we have

$$\left\|C^{up}(u_n^0)\right\|_2 \leq C_1, \tag{20}$$

and an application of Lemma 1, Lemma 2 and Lemma 5 gives

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

The third factor appearing in (17) is equal to the third factor of (16), thus finally S_2 satisfies

$$S_2 \leq \frac{1}{J} C_1 C_4 \left\|\Gamma^{-1}\right\|_2^2 \sum_{j=1}^{J} \mathbb{E}\left[\left\|u_n^\varepsilon - u_n^0\right\| (L_{\mathcal{G}} R + \left\|\xi_{n+1}^{(j)} + \eta\right\|_2)\right],$$

and, as above, defining $\alpha_2 \coloneqq C_1 C_4 \left\|\Gamma^{-1}\right\|_2^2 (L_{\mathcal{G}} R + \sqrt{2\mathrm{tr}(\Gamma)})$, we obtain

$$S_2 \leq \alpha_2 \mathbb{E}\left[\left\|u_n^\varepsilon - u_n^0\right\|\right]. \tag{21}$$

We now consider the last term S_3. The first factor appearing in (18) can be bounded as in (20) and for the second part we use Lemma 2, thus obtaining

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

Regarding the third factor of (18), we apply the triangle inequality and the Lipschitz continuity of the forward operator $\mathcal{G}^\varepsilon$, which yield

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

Hence, a bound for S_3 is given by

$$S_3 \leq C_1 \left\|\Gamma^{-1}\right\|_2 \mathbb{E}\left[\frac{1}{J} \sum_{j=1}^{J} \left\|\mathcal{G}^0(u_n^{0^{(j)}}) - \mathcal{G}^\varepsilon(u_n^{0^{(j)}})\right\|_2 + L_{\mathcal{G}} \frac{1}{J} \sum_{j=1}^{J} \left\|u_n^{0^{(j)}} - u_n^{\varepsilon^{(j)}}\right\|_2\right],$$

which is equivalent to

$$S_3 \leq C_1 \left\|\Gamma^{-1}\right\|_2 \mathbb{E}\left[e(\varepsilon, u_n^0)\right] + C_1 \left\|\Gamma^{-1}\right\|_2 L_{\mathcal{G}} \mathbb{E}\left[\left\|u_n^0 - u_n^\varepsilon\right\|\right],$$

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 S_3, i.e.,

$$S_3 \leq \alpha_3 \mathbb{E}\left[\left\|u_n^0 - u_n^\varepsilon\right\|\right] + \gamma \mathbb{E}\left[e(\varepsilon, u_n^0)\right]. \tag{22}$$

Therefore, using the results (15), (19), (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

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

which is the desired result. □

We now present the main result about global multiscale convergence of the EnKF algorithm.

Proposition 1. *Under the notation and assumptions of Lemma 6, letting $u_0^{\varepsilon} = u_0^0$ be the same initial ensemble, 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 there exists $K_1 > 0$ independent of ε such that*

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

Proof. Iterating Lemma 6 and taking into account that $u_0^{\varepsilon} = u_0^0$, after N iterations, at the end of the EnKF algorithm, we get

$$\mathbb{E}\left[\left\|u_{N}^{\varepsilon}-u_{N}^{0}\right\|\right] \leq \gamma\sum_{i=0}^{N-1}\alpha^{N-1-i}\mathbb{E}\left[e(\varepsilon,u_i^0)\right],$$

Applying Lemma 4, 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 4, we have the estimate

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

Therefore we obtain

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

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

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

which is the desired result. □

We now consider convergence with respect to the FEM discretization of the homogenized problem. First, we introduce a preliminary result, which plays the role of Lemma 4 in the context of numerical convergence and whose proof is given in the Appendix.

Lemma 7. *Let $\tilde{e}$ be defined in* (11)*. Under Assumption 1 and if the exact solution p^0 of the homogenized problem* (12) *is in $H^{q+1}(\Omega)$, $A^0 \in C^q(\Omega; \mathbb{R}^{d\times d})$ independently of u, $f \in H^{q-1}(\Omega)$, $\partial\Omega \in C^{q+1}$, and we employ polynomials of degree r for the finite element basis, then*

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

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

We can now state the main result concerning convergence with respect to the numerical discretization of the homogenized problem.

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 (12) is in $H^{q+1}(\Omega)$, $A^0 \in C^q(\Omega; \mathbb{R}^{d\times d})$, $f \in H^{q-1}(\Omega)$, $\partial\Omega \in C^{q+1}$ 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 K_2 is a positive constant independent of h.

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 4 for the error e has to be replaced by Lemma 7 for the error $\tilde{e}$. □

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

Proof of Theorem 1. An application of the triangle inequality yields

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

The two addends can be bounded applying Proposition 1 and Proposition 2 respectively as

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

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

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

which is the desired result. □

Remark 2. Note that if the exact solution of the homogenized problem in (12) $p^0 \in H^2(\Omega)$ and we use polynomials of degree $r = 1$ for the finite element basis, then we have

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

Therefore, in order to balance the two sources of error, the discretization parameter h for the FEM approximation of the homogenized problem should be chosen as

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

which guarantees linear convergence with respect to ε. With this choice for h, the computational cost is drastically reduced with respect to the solution of the multiscale problem, for which h^ε needs to be chosen as

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

in order to be able to resolve the oscillations of the multiscale solution.

4.2 Convergence of the posterior distributions

We now consider the Bayesian interpretation of the EnKF method. In particular, we consider the multiscale posterior distribution obtained after N steps of the EnKF algorithm, i.e.,

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

and the posterior corresponding to the FEM solution of the homogenized problem, which reads

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

and we study the convergence of μ^ε to μ_h^0 as $\varepsilon, h \to 0$. Due to the discrete nature of the distributions above, we study convergence with respect to the Wasserstein metrics. 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 now report the definition of the Wasserstein distances in the metric space $(B_R(u^*), \|\cdot\|_s)$, which can be found, e.g., in [14].

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

$$W_{p,s}(\mu,\nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{B_R(u^*)\times B_R(u^*)} \|u - v\|_s^p \, d\gamma(u,v) \right)^{1/p}, \tag{23}$$

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 (23) can be written as

$$W_{p,s}(\mu,\nu) = \left(\inf_{\gamma \in \mathbb{R}^{K_1 \times K_2}} \sum_{i=1}^{K_1}\sum_{j=1}^{K_2} \|u_i - v_j\|_s^p \, \gamma_{ij} \right)^{1/p}, \tag{24}$$

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 $L \le 1$ with respect to the norm $\|\cdot\|_s$.

In Lemma 8 we show that $W_{1,2}$ is bounded by the distance induced by the ensemble norm defined in 9. This result will be crucial later to deduce the convergence of the posterior distribution μ_h^0 to μ^ε from Theorem 1.

Lemma 8. *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

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

and, in particular,

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

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 (24), 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 (9), 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 of the form (8) are random probability measures. Indeed, the EnKF algorithm randomizes data at each step, and therefore the resulting ensemble is not deterministic. Therefore, we need to introduce a notion of convergence for random probability measures, and 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 Lemma 9, whose proof is given in the Appendix, 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 [14] for non-random measures, but here we extend the result to random probability measures.

Lemma 9. *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

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

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 2. *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

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

and applying Lemma 8 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 $\left\|\Gamma^{-1}\right\|_2$ as

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

Finally, by Lemma 9, we obtain

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

which is the desired result. □

5 Modelling error

In this section, we consider the effects of model misspecification due to the homogenization and discretization error. All the results presented above deal with the asymptotic case $h, \varepsilon \to 0$, which is unrealistic in applications. Let us recall that the original inverse problem involves predicting the exact unknown u^* from observations originated by the model

$$y = \mathcal{G}^\varepsilon(u^*) + \eta, \tag{25}$$

where $\eta \sim \mathcal{N}(0, \Gamma)$ is the noise. Since evaluating $\mathcal{G}^\varepsilon$ is too expensive and in many applications unfeasible, we wish to employ the cheaper forward operator $\mathcal{G}_h^0$. Hence, we rewrite (25) as

$$y = \mathcal{G}_h^0(u^*) + \mathcal{E}(u^*) + \eta, \tag{26}$$

where

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

The quantity $\mathcal{E}(u^*)$ represents the error introduced by misspecification of the forward model. Equation (26) 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 formulation of modelling error was originally presented in [5], and then applied to multiscale inverse problems in [1]. Following [1,5], 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 write

$$y = \mathcal{G}_h^0(u^*) + m + \zeta + \eta, \tag{27}$$

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 (27), we obtain

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

Therefore, if the mean m and covariance Σ of the modelling error are known, a more reliable approximation of the unknown u^* can be obtained applying the EnKF to (28). The modelling error distribution, by assumption fully determined by its mean and covariance, is approximated offline. We sample $N_{\mathcal{E}}$ unknowns $\{u_i\}_{i=1}^{N_{\mathcal{E}}}$ from 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 then obtained as the empirical mean and covariance of the sample $\{\mathcal{E}_i\}_{i=1}^{N_{\mathcal{E}}}$. This procedure is computationally involved due to the multiple evaluations of $\mathcal{G}^\varepsilon$, but it has to be performed only once and can then be applied to different sets of observations and true values u^*. Moreover, we remarked in practice via numerical experiments that a small number $N_{\mathcal{E}}$ can be employed to obtain a satisfactory approximation of the modelling error. A theoretical insight of this property is provided by Proposition 3.

In order to obtain a more reliable approximation of the distribution of the modelling error, we can follow a dynamic approach based on the estimation of the mean m and the covariance Σ online, i.e., during the run of the EnKF algorithm. This methodology has been developed in [4]. In particular, we sequentially apply the ensemble Kalman method for $\mathcal{L}$ levels and, at each level, we update the distribution of the modelling error. We denote by $\nu^\ell = \mathcal{N}(m^\ell, \Sigma^\ell)$ for any $\ell = 1, \dots, \mathcal{L}$ the approximated distribution at level ℓ. In particular, 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, thus computing the mean m^ℓ and the covariance Σ^ℓ as the empirical mean and covariance of the sample. This approach provides indeed a better approximation of the modelling error as instead of taking the samples from the prior distribution, they are drawn from distributions which are progressively closer to the true posterior. 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. Let us remark that on the one hand, due to the theory of homogenization, the modelling error can be considered negligible when ε is very small, and the expensive estimation of $\mathcal{E}$ may not be necessary. On the other hand, when ε is larger, the homogenized equation does not provide with a good approximation of the multiscale problem, and an estimation of $\mathcal{E}$ is required. One may rightfully argue that in case $\varepsilon = \mathcal{O}(1)$, it is possible to evaluate the forward operator $\mathcal{G}^\varepsilon$ without a large computational effort. Hence, the techniques presented in this section are relevant for mid-range values of ε, for which $\mathcal{E}$ is significant with respect to η.

Finally, we are interested in studying whether the simple offline method for estimating the modelling error provides indeed a good approximation, at least in the mean sense. In this direction, we give in Proposition 3 a criterion on how to choose the number $N_{\mathcal{E}}$ of full multiscale problems which has to be solved in order to have a reliable approximation of the true mean m^* of the modelling error with respect to ε and h. Before stating Proposition 3, let us recall the Hoeffding's inequality, which will be used in the proof. Let $\{X_i\}_{i=1}^N$ be independent random variables with values in $[a, b]$, and let $\bar{X}$ be the sample average of $\{X_i\}_{i=1}^N$. Then, for all $t \in \mathbb{R}$ it holds

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

Proposition 3. *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 4 and Lemma 7. Let $\{\mathcal{E}_i\}_{i=1}^{N_\mathcal{E}} \subset \mathbb{R}^L$ be 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 $\mathcal{N}(0, I)$, let m be the sample mean of $\{\mathcal{E}_i\}_{i=1}^{N_\mathcal{E}}$ and $m^* = \mathbb{E}[\mathcal{E}_i]$. 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 4 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

$$|(\mathcal{E}_i)_l| \leq \|\mathcal{E}_i\|_2 \leq K\varepsilon + \tilde{K}h^{s+1} \leq C_\mathcal{E}(\varepsilon + h^{s+1}). \tag{29}$$

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

$$\mathbb{P}(\|m - m^*\|_2 \leq t) \geq \mathbb{P}\left(|m_l - m_l^*| \leq \frac{t}{\sqrt{L}} \quad \forall\, l = 1, \dots, L\right). \tag{30}$$

Using (29), applying Hoeffding's inequality and Young's inequality we have

$$\begin{aligned} \mathbb{P}\left(|m_l - m_l^*| \geq \frac{t}{\sqrt{L}}\right) &\leq 2\exp\left\{-\frac{2t^2N_\mathcal{E}^2}{4LN_\mathcal{E}C_\mathcal{E}^2(\varepsilon + h^{s+1})^2}\right\} \\ &\leq 2\exp\left\{-\frac{t^2N_\mathcal{E}}{4LC_\mathcal{E}^2(\varepsilon^2 + h^{2(s+1)})}\right\}. \end{aligned} \tag{31}$$

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

$$\mathbb{P}\left(\bigcap_{l=1}^{L} A_l\right) = 1 - \mathbb{P}\left(\left(\bigcap_{l=1}^{L} A_l\right)^C\right) = 1 - \mathbb{P}\left(\bigcup_{l=1}^{L} A_l^C\right) \geq 1 - \sum_{l=1}^{L} \mathbb{P}(A_l^C). \tag{32}$$

Therefore, thanks to (30), (31) and (32), we have

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

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 3, 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 ($\alpha \to 0$). Moreover, note that for any given accuracy the number of samples required $N_{\mathcal{E}}$ is a increasing function of ε and h, so that if the model $\mathcal{G}_h^0$ is a good approximation of $\mathcal{G}$, only a few samples are needed.

6 Numerical experiments

In this section, 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

$$\begin{cases} -\nabla \cdot (A^{\varepsilon}_{\sigma^*} \nabla p^{\varepsilon}) = 0 & \text{in } \Omega, \\ p^{\varepsilon} = g & \text{on } \partial\Omega. \end{cases} \tag{33}$$

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 [13, 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 $\Gamma_i \in \partial\Omega$, $i = 1, \dots, I$, $\Gamma_i \cap \Gamma_j = \emptyset$ for $i \neq j$, and I functions $\varphi_i \in H^{1/2}(\partial\Omega)$ with compact support supp $(\varphi_i) \subset \Gamma_i$ for all $i = 1, \dots, I$. Moreover, we solve (33) for $K \in \mathbb{N}$ Dirichlet data, denoted by g_k with $k = 1, \dots, K$. Then we define the multiscale operator $\mathcal{F}^{\varepsilon}\colon \Sigma \to \mathbb{R}^L$ where $L = IK$ by components

$$\mathcal{F}^{\varepsilon}(\sigma)_{ik} = \mathcal{F}^{\varepsilon}(\sigma)_l = \int_{\Gamma_i} A^{\varepsilon} \nabla p^{\varepsilon}_k \cdot \nu \varphi_i ds, \qquad i = 1, \dots, I, k = 1, \dots, K. \tag{34}$$

where p^{ε}_k is the solution of problem (33) 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 (33) 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

$$\sigma = \sigma_0 + \sum_{m=1}^{N_h} \sqrt{\lambda_m} u_m \varphi_m, \tag{35}$$

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 (35)

$$\sigma \simeq \sigma_0 + \sum_{m=1}^{M} \sqrt{\lambda_m} u_m \varphi_m, \tag{36}$$

and the actual unknown becomes the vector $u \in \mathbb{R}^M$, whose components are the coefficients u_m in (36). 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 (33), 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 [3]. Hence, analogously to the multiscale case, we define the discrete homogenized operator $\mathcal{F}_h^0 \colon \Sigma \to \mathbb{R}^L$ as

$$\mathcal{F}_h^0(\sigma)_l = \mathcal{F}_h^0(\sigma)_{ik} = \int_{\Gamma_i} A^0 \nabla p_{h_k}^0 \cdot \nu \varphi_i ds, \qquad i = 1, \dots, I, k = 1, \dots, K, \tag{37}$$

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

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

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 $h_{\mathrm{obs}} = 1/4096$ 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.

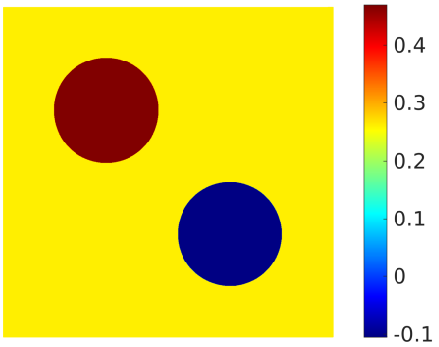

Figure 1: Plot of the exact unknown σ^*

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 (34) and (37), 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 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.

One can verify that the tensor A^ε_σ satisfies Assumption 2. In particular, for $\xi \in \mathbb{R}^2$ we have

$$
A^\varepsilon_\sigma \xi \cdot \xi = a_{1,1}\left(\sigma(x), \frac{x}{\varepsilon}\right)\xi_1^2 + a_{2,2}\left(\sigma(x), \frac{x}{\varepsilon}\right)\xi_2^2 \geq e^{\sigma(x)}(\xi_1^2 + \xi_2^2) = e^{\sigma -} \|\xi\|_2^2 .
$$

Moreover, since the EnKF algorithm estimates the coefficients u of the truncated Karhunen-Loève expansion, we can show that for all $u_1, u_2 \in \mathbb{R}^M$

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

where $A^\varepsilon(u) = A^\varepsilon_{\sigma_u}$ and $\|\cdot\|_2$ stands for the 2-norm of a vector in $\mathbb{R}^M$. We have

$$\begin{aligned}\|A^\varepsilon(u_1) - A^\varepsilon(u_2)\|_{L^\infty(\Omega,\mathbb{R}^{d\times d})} &= \sup_{x\in\Omega} \sup_{\xi\in\mathbb{R}^2, \|\xi\|_2=1} \|(A^\varepsilon(u_1) - A^\varepsilon(u_2))\xi\|_2 \\ &\leq \sup_{x\in\Omega} \sqrt{\left(a_{1,1}\left(\sigma_{u_1}(x), \frac{x}{\varepsilon}\right) - a_{1,1}\left(\sigma_{u_2}(x), \frac{x}{\varepsilon}\right)\right)^2 + \left(a_{2,2}\left(\sigma_{u_1}(x), \frac{x}{\varepsilon}\right) - a_{2,2}\left(\sigma_{u_2}(x), \frac{x}{\varepsilon}\right)\right)^2},\end{aligned}$$

which implies

$$\begin{aligned}\|A^\varepsilon(u_1) - A^\varepsilon(u_2)\|_{L^\infty(\Omega,\mathbb{R}^{d\times d})} &\leq \sup_{x\in\Omega} \sqrt{13\left(e^{\sigma_{u_1}(x)} - e^{\sigma_{u_2}(x)}\right)^2} \\ &\leq \sup_{x\in\Omega} \sqrt{13} e^{\sigma^+} \left|\sigma_{u_1}(x) - \sigma_{u_2}(x)\right| .\end{aligned}$$

Using the truncated Karhunen-Loève expansion we obtain

$$\begin{aligned}\|A^\varepsilon(u_2)\|_{L^\infty(\Omega,\mathbb{R}^{d\times d})} &\leq \sup_{x\in\Omega} \sqrt{13} e^{\sigma^+} \left|\sum_{m=1}^M \sqrt{\lambda_m}\varphi_m(x)(u_{1,m} - u_{2,m})\right| \\ &\leq \sup_{x\in\Omega} \sqrt{13} e^{\sigma^+} \left(\sum_{m=1}^M \lambda_m \varphi_m(x)^2\right)^{1/2} \left(\sum_{m=1}^M (u_{1,m} - u_{2,m})^2\right)^{1/2} \\ &= \sup_{x\in\Omega} \sqrt{13} e^{\sigma^+} \left(\sum_{m=1}^M \lambda_m \varphi_m(x)^2\right)^{1/2} \|u_1 - u_2\|_2 ,\end{aligned}$$

and defining $M = \sup_{x\in\Omega} \sqrt{13} e^{\sigma^+} \left(\sum_{m=1}^M \lambda_m \varphi_m(x)^2\right)^{1/2}$ we get the result.

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^* , 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 (28) 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.

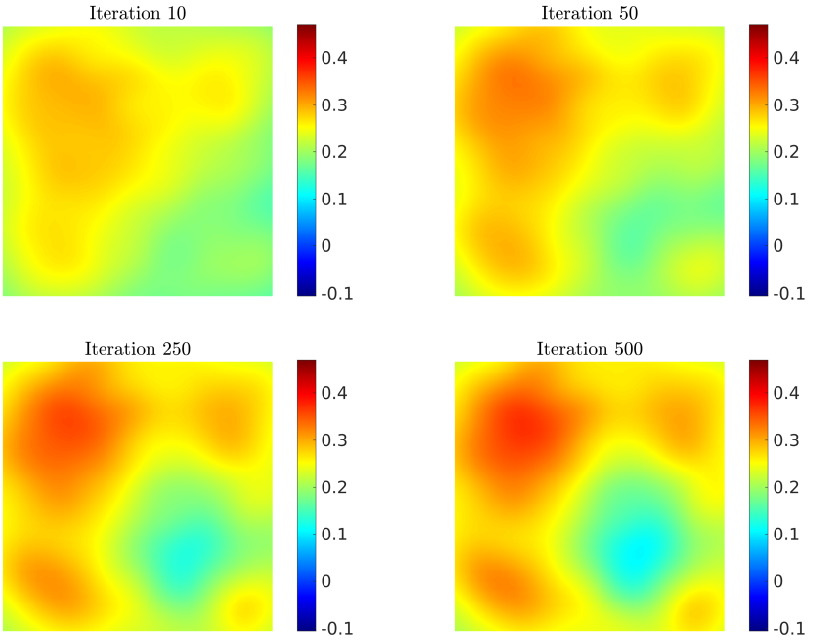

Figure 2: Plot of the unknown σ_{EnKF} estimated by the iterative ensemble Kalman method after $10, 50, 250$ and 500 iterations.

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 $\ell = 1, \ldots, \mathcal{L}$. 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.

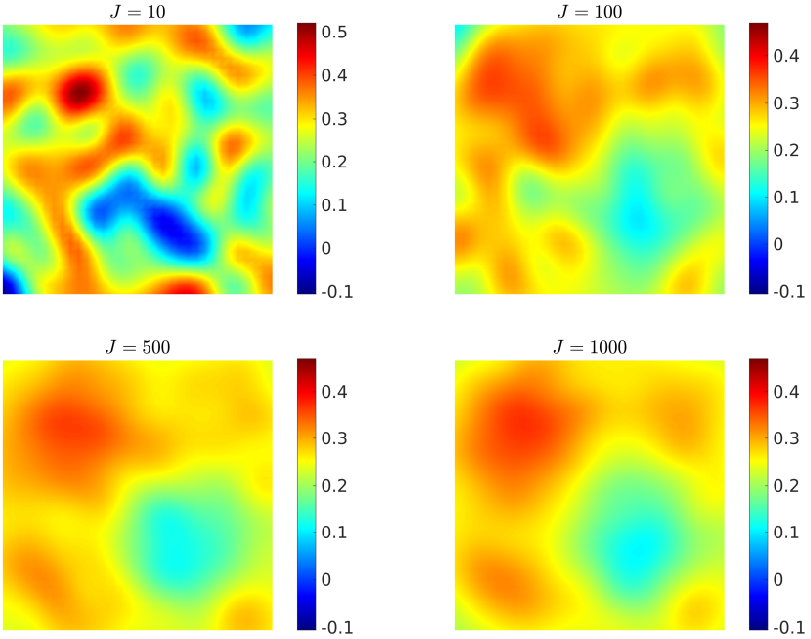

Figure 3: Plot of the unknown σ_{EnKF} estimated by the iterative ensemble Kalman method after 500 iterations for different numbers of particles per ensemble $J = 10, 100, 500, 1000$.

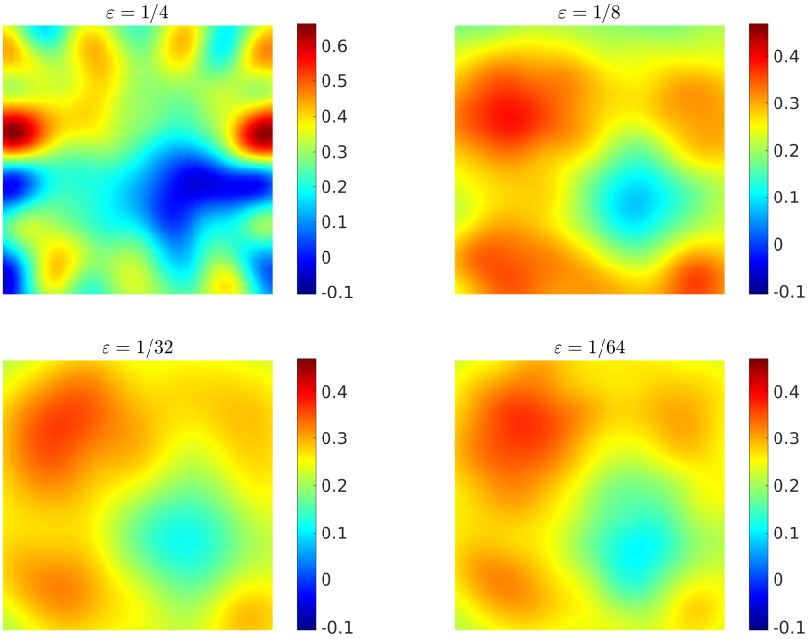

Figure 4: Plot of the unknown σ_{EnKF} estimated by the iterative ensemble Kalman method after 500 iterations for different values of the multiscale parameter $\varepsilon = 1/4, 1/8, 1/32, 1/64$.

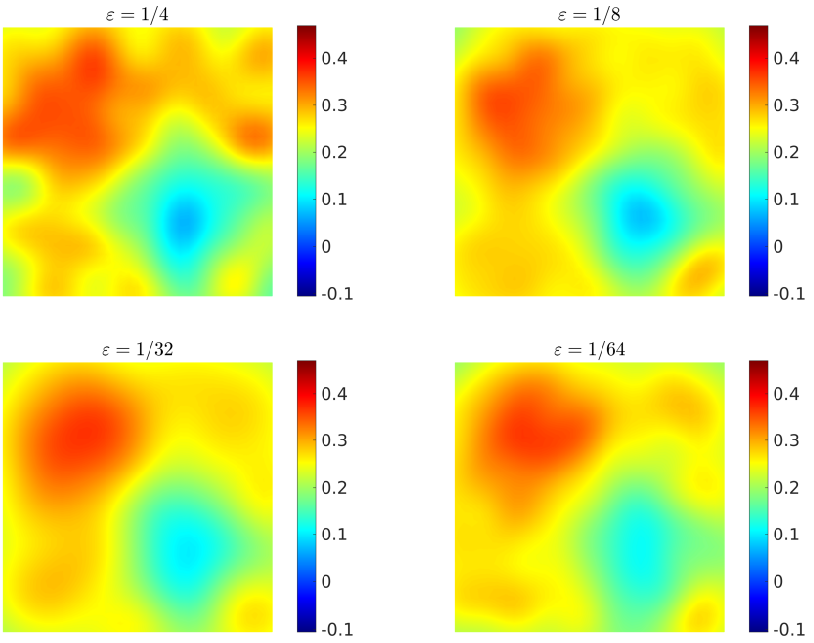

Figure 5: Plot of the unknown σ_{EnKF} estimated by the iterative ensemble Kalman method with model error estimation after 500 iterations for different values of the multiscale parameter $\varepsilon = 1/4, 1/8, 1/32, 1/64$.

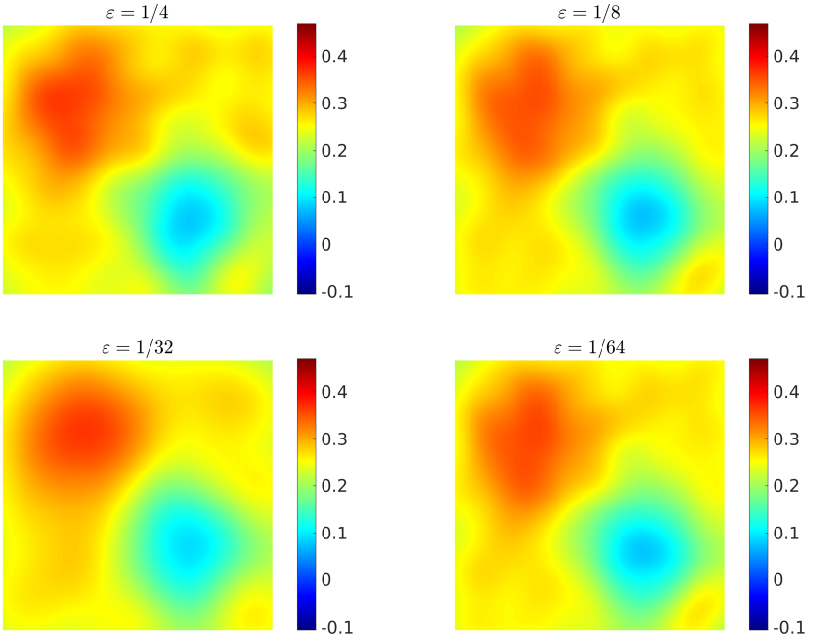

Figure 6: Plot of the unknown σ_{EnKF} estimated by the iterative ensemble Kalman method with dynamic updating of the model error estimation after 500 iterations for different values of the multiscale parameter $\varepsilon = 1/4, 1/8, 1/32, 1/64$.

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 3

Let $u_1, u_2 \in \mathbb{R}^M$, and $p_1 = \mathcal{S}(u_1)$, $p_2 = \mathcal{S}(u_2)$ Then, writing the weak formulations of (12) we get

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

for all $v \in H_0^1(\Omega)$. Adding and subtracting $A(u_1)\nabla p_2$ to the first factor inside the integral and rearranging terms yields

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

Choosing $v = p_1 - p_2$, using the hypotheses on A and the Hölder inequality, we obtain

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

which implies

$$\|\nabla p_1 - \nabla p_2\|_{L^2(\Omega;\mathbb{R}^d)} \leq \frac{M}{\alpha} \|\nabla p_2\|_{L^2(\Omega;\mathbb{R}^d)} \|u_1 - u_2\|_2. \tag{38}$$

It remains to bound $\|\nabla p_2\|_{L^2(\Omega;\mathbb{R}^N)}$, which can be achieved with a standard coercivity argument. In particular, we have

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

where C_p is the Poincaré constant associated to the domain Ω. Replacing in (38), 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 of constant $L_{\mathcal{G}} = L_{\mathcal{O}} L_{\mathcal{S}}$, which concludes the proof. □

Proof of Lemma 4

Let us consider an ensemble $u \in \mathcal{U}_{J,M}$ with particles $u^{(j)} \in \mathbb{R}^M$, for $j = 1, \ldots, J$. For each particle we have

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

where we write explicitly the dependence of p^ε and p^0 on the particle it is generated by. By homogenization theory, we know that $p^\varepsilon(u^{(j)}) \rightharpoonup p^0(u^{(j)})$ in $H_0^1(\Omega)$, and therefore we have $p^\varepsilon(u^{(j)}) \to p^0(u^{(j)})$ in $L^2(\Omega)$, which implies

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

Moreover, if the solution of the homogenized problem p^0 is sufficiently smooth independently of u, namely $p^0 \in H^2(\Omega)$, letting $C > 0$ be a constant, we have for all $j = 1, \ldots, J$ the following estimate, which can be found in [10]

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

hence we finally obtain

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

which proves the result for $K = mC$. □

Proof of Lemma 5

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

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

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

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

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

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

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

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

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

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

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

$$\left\|\bar{u}_1-\bar{u}_2\right\|_2 = \left\|\frac{1}{J}\sum_{j=1}^{J}(u_1^{(j)}-u_2^{(j)})\right\|_2 \le \frac{1}{J}\sum_{j=1}^{J}\left\|u_1^{(j)}-u_2^{(j)}\right\|_2 = \|u_1-u_2\|,$$

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

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

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

which implies

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

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

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

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 Lemma 7

Let us consider an ensemble $u \in \mathcal{U}_{J,M}$ with particles $u^{(j)} \in \mathbb{R}^M$, for $j = 1, \ldots, J$. For each particle we have

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

where we write explicitly the dependence of p^0 and p_h^0 on the particle it is generated by. We now consider the standard a priori error estimates of FEM, (see e.g. [6, Theorem 3.2.5]), which reads

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

Moreover, higher order boundary regularity results for elliptic partial differential equations (see e.g. [8, Theorem 6.3.5]), imply for a constant $\tilde{C} > 0$ and for all $j = 1, \ldots, J$

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

Finally, we obtain

$$
\begin{aligned}
\tilde{e}(h,u) = &\frac{1}{J}\sum_{j=1}^{J} \left\|\mathcal{G}_h^0(u^{(j)}) - \mathcal{G}^0(u^{(j)})\right\|_2 \\
&\leq mC\tilde{C}\|f\|_{H^{q-1}(\Omega)} h^{s+1},
\end{aligned}
$$

which proves the result for $\tilde{K} = mC\tilde{C}\|f\|_{H^{q-1}(\Omega)}$. □

Proof of Lemma 9

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

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

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 $\|\psi_k\|_{L^\infty(B_R(u^*))}\le C$ 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] ——, *Numerical homogenization and model order reduction for multiscale inverse problems.* Accepted in SIAM MMS, 2018.

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

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

[5] D. Calvetti, O. Ernst, and E. Somersalo, *Dynamic updating of numerical model discrepancy using sequential sampling*, Inverse Problems, 30 (2014), pp. 114019, 19.

[6] P. G. Ciarlet, *The finite element method for elliptic problems.*, vol. 40 of Classics Appl. Math., SIAM, Philadelphia, 2002.

[7] 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.

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

[9] G. Evensen, *Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics*, J. Geophys. Res., (1994), pp. 10143–10162.

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

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

[12] 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.

[13] 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.

[14] 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.

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

[16] A. M. Stuart, *Inverse problems: a Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.