Data-Centric Engineering (2025), 6: e19
doi:10.1017/dce.2025.8

RESEARCH ARTICLE

An interacting Wasserstein gradient flow strategy to robust Bayesian inference for application to decision-making in engineering

Felipe Igea[1] and Alice Cicirello[1,2]

[1]Department of Engineering Science, University of Oxford, Oxford, UK
[2]Department of Engineering, University of Cambridge, Cambridge, UK
Corresponding author: Felipe Igea; Email: felipe.igea@hotmail.com

Received: 22 October 2023; **Revised:** 04 December 2024; **Accepted:** 07 January 2025

Keywords: Ambiguity sets; Bayesian model updating; Interacting Wasserstein gradient flows; Robust Bayesian inference; Uncertain prior

Abstract

Bayesian model updating (BMU) is frequently used in Structural Health Monitoring to investigate the structure's dynamic behavior under various operational and environmental loadings for decision-making, e.g., to determine whether maintenance is required. Data collected by sensors are used to update the prior of some physics-based model's latent parameters to yield the posterior. The choice of prior may significantly affect posterior predictions and subsequent decision-making, especially under the typical case in engineering applications of little informative data. Therefore, understanding how the choice of prior affects the posterior prediction is of great interest. In this article, a robust Bayesian inference technique evaluates the optimal and worst-case prior in the vicinity of a chosen nominal prior and their corresponding posteriors. This technique derives an interacting Wasserstein gradient flow that minimizes and maximizes/minimizes the KL divergence between the posterior and the approximation to the posterior, with respect to the approximation to the posterior and the prior. Two numerical case studies are used to showcase the proposed algorithm: a double-banana-posterior and a double-beam structure. Optimal and worst-case priors are modeled by specifying an ambiguity set containing any distribution at a statistical distance to the nominal prior, less than or equal to the radius. The resulting posteriors may be used to yield the lower and upper bounds on subsequent calculations of an engineering metric (e.g., failure probability) used for decision-making. If the metric used for decision-making is not sensitive to the resulting posteriors, it may be assumed that decisions taken are robust to prior uncertainty.

Impact statement

Bayesian model updating may be significantly sensitive to assumptions about the prior distributions chosen for the latent parameters of the physics-based model used to represent the structure's behavior, especially if, due to some restrictions, such as time constraints and cost, the number of observations available is limited. In these cases, the selection of prior distributions may significantly affect the resulting posterior distributions, and as a consequence, the decisions about engineering metrics, such as reliability, useful lifetime, and maintenance of the structure. To address these limitations, a robust Bayesian inference approach based on interacting Wasserstein gradient flows is proposed in this article. It is shown that the proposed approach estimates the optimal and worst cases of prior distributions and calculates their corresponding approximations to the posterior distribution that may be used as lower and upper bounds on subsequent metric calculations used for decision-making. These bounds on the resulting metric can be readily used in decision-making to assess if the decisions taken are robust to prior distribution uncertainty or otherwise.

https://doi.org/10.1017/dce.2025.8 Published online by Cambridge University Press

1. Introduction

Bayesian inference techniques have been frequently used in engineering to estimate the inherent variability of the uncertain latent parameters and/or to identify unknown model parameters of the physical models of real-world structures when measurements on the real-world structure become available (Mottershead and Friswell, 1993; Green and Worden, 2015; Lye et al. 2021; Igea, 2023). The so-called Bayesian model updating (BMU), starting from a physics-based model, a prior distribution on the uncertain parameters, and a suitable description of the discrepancy between the measurements and model predictions (the likelihood), yields the posterior distribution of the uncertain latent parameters (Beck and Katafygiotis 1998; Sedehi et al., 2019; Kennedy and O'Hagan, 2001). This updated probabilistic model is then used to assess the performance of the real-world structure under various conditions, including its reliability (Straub et al., 2015), remaining useful life (Sankararaman, 2015), and to support maintenance decision-making (Verzobio et al., 2018). Bayesian inference is actively used in structural health monitoring (SHM) (Yuen et al., 2006; Farrar and Worden, 2013; Rocchetta et al., 2018). SHM focuses on non-intrusive detection of an abnormal structural condition. SHM can provide early warnings on the health status of engineering structures (Farrar and Worden, 2013). The updated models can be used to locate and assess the damage (Rytter, 1993; Farrar and Worden, 2013; Simoen et al., 2015; Ebrahimian et al., 2017; Verzobio et al., 2018). Once the health state of the structural system is identified, optimal maintenance decisions can be identified (Kamariotis et al., 2020). Real-time engineering decisions (repairment, further observation, etc.) may be mathematically performed using Bayesian decision analysis methods (Kamariotis et al., 2023). In this decision-making process, the target is to define a group of actions (e.g., repair, inspection, or maintenance) that minimize the expected life-cycle costs of the structure. Therefore, the results obtained with BMU are critical in decision-making in engineering. However, those results can be sensitive to: the prior distribution assumption, the likelihood, and the computational strategy employed to evaluate the posterior distribution. Robust Bayesian inference, is a methodology used to investigate the robustness of Bayesian inference results to uncertainty of the prior, and/or likelihood, and/or computational strategy (Berger et al., 1994). When the result does not sensitively depend on the assumptions and calculations made, it is said to be robust. Robustness can be quantified in terms of the range of the result: if the range is small, the result is not sensitive to the different assumptions or calculations explored. This article focuses on the robustness of the BMU results with respect to the prior distribution. A review on recent progress in computational strategy and likelihood approximations used in Bayesian inference is given below.

Bayesian inference techniques obtain approximations to the posterior distributions for given sets of data by using numerical approaches. Approximation techniques are mostly used due to the frequently occurring intractability of the posterior distribution. Two common cases of these approximation methods are Markov Chain Monte Carlo (MCMC) and Variational Inference (VI). The machine learning community has been using optimization methods based on VI to approximate posterior distributions (Blei et al., 2017). Put simply, VI shapes the inference problem into an optimization one. In this optimization problem, the chosen distribution is a member of a family of distributions that shows a smaller Kullback-Leibler (KL) divergence to the posterior (Blei et al., 2017). Methods based on MCMC extract the samples directly from the posterior. These MCMC approaches have as disadvantages their slow convergence and the difficulties that occur when assessing if convergence has been reached (Cheng et al., 2023). The VI methodologies recently developed (Kingma et al., 2016; Acerbi, 2018; Campbell and Li, 2019) if compared to sampling approaches like MCMC benefit from higher numerical scalability, and due to their more advanced optimization features, are better suited to be employed in a more comprehensive range of situations. Recent advances of MCMC strategies in BMU for engineering applications can be found in Lye et al. (2021), while VI methods to BMU to efficiently deal with multimodal posterior distributions can be found in Igea and Cicirello (2023). An ideal combination of the VI and MCMC techniques has been recently developed within the machine learning community: the particle-based VI methods (ParVIs) (Chen et al., 2018; Alvarez-Melis et al., 2021; Fan et al., 2021; Cheng et al., 2023). In these ParVIs approaches, a set of particles is used to represent the distribution to be approximated.

These distributions are updated in an iterative manner through minimization of their KL divergence to the posterior. ParVI methods show greater particle efficiency than MCMC approaches due to particle interactions, and compared to typical VIs, the ParVIs exhibit greater flexibility as a consequence of their non-parametric character (Cheng et al., 2023). The Stein Variational Gradient Descent (SVGD) is the most used VI technique based on particles (Liu and Wang, 2016). In SVGD, the distribution space is chosen in such a manner that inside of it, the gradient flows are tractable (Liu, 2017; Chewi et al., 2020), and the particles are updated by simulation of the KL divergence gradient flows.

Unreliable approximations of the system's posterior distributions are mainly produced by not accounting for all plausible values of the observations that may be obtained from experiments or ignoring spatial and temporal correlation in the measurements (Simoen et al., 2013; Koune et al., 2023). In engineering, this is of particular interest, as the number of experiments that may be run is limited due to the high cost incurred and time constraints. In these cases where the complexities of the likelihood are increased to improve the models' accuracy, and therefore, the reliability of the inferences, techniques such as: mixture models, nonparametric or semiparametric models, and models with heavier tails have been used as likelihood functions (Hooker and Vidyashankar, 2011; Ghosh and Basu, 2016; Chérief-Abdellatif and Alquier, 2019; Matsubara et al., 2021). Nonetheless, the introduction of those methodologies to define the likelihood functions frequently leads to a set of new issues: higher numerical cost, definition of parameters, and harder interpretability. Although these techniques to define complex likelihood functions may improve the model's specification, some amount of inaccuracy is unavoidable. As a result, when a limited number of observations is available, the choice of the prior distribution may substantially affect the posterior distribution obtained. In engineering, this may affect subsequent decisions such as those made to assess the reliability of a structure, its remaining useful lifetime, and whether a structure requires predictive maintenance. Therefore, a method able to quantify the robustness of the posterior prediction when the assumptions of the prior distribution are changed is of great interest. This is the focus of the present article that focuses on the development of a robust Bayesian inference strategy for application to decision-making in engineering.

This work investigates the sensitivity of the posterior distribution to the prior distribution's uncertainty. More specifically, if the prior distributions that either maximize or minimize a certain metric, defined as the worst-case prior and optimal prior distributions, respectively (or vice versa depending on the metric's definition), can be determined, the resulting posterior distribution of each case can be used as lower and upper bounds on subsequent calculations used for decision-making. If the difference between the upper bound and lower bound found using the method is low for the metric used to support a decision, then it may be confirmed that the decision taken is robust to the prior distribution uncertainty. More specifically, it might not be possible to define exactly the prior distribution for the latent variables. This type of situation could arise in the presence of limited prior knowledge on the latent parameters and/or conflicting opinions from experts. For those cases, we would like to explore how the approximation to the posterior might be affected by distributions that are in the neighborhood of an assumed nominal prior distribution, as this might have consequences on subsequent calculations. Therefore, it would be useful to develop a method that could determine the worst or optimal distributions inside that neighborhood of distributions in terms of a particular functional of interest. Differently from a method that uses an a priori defined (typically by experts) informative or non-informative prior distribution for the BMU, the proposed method seeks to obtain the optimal and worst-case prior distributions inside an ambiguity set for a given functional. Therefore, it enables one to quantify how confident one is about the chosen nominal prior distributions by exploring distributions in its neighborhood through the definition of an ambiguity set. In this article, the problem of robustness to prior uncertainty in Bayesian inference is dealt with by developing an interacting Wasserstein gradient flow (WGF) combined with an ambiguity set. An interacting WGF is derived to find: (a) the best approximation to the posterior by minimizing the KL divergence between the posterior and the approximation to the posterior, where the posterior is subject to change (due to the prior distribution also changing); (b) the optimal or worst-case prior distributions—defined as the distribution that either minimizes or maximizes the KL divergence between the posterior and the approximation to the posterior, respectively. The proposed approach calculates the resulting optimal or worst-case prior distributions by constraining the space of distributions to be explored using an ambiguity set. This ambiguity set is defined by a nominal distribution and all the distributions that lie within a specified value of a statistical distance, where both are assumed to be known. The robustness of the method is derived from this distance metric. A useful property of the Wasserstein distance is that distributions that do not share the same support may be investigated inside the ambiguity set (Kuhn et al., 2019). The support of a distribution refers to the values of the random variable that have a probability density bigger than zero. A particle-based interacting WGF Wasserstein-2 space algorithm is developed, and the results from two numerical case studies are presented.

2. Robust Bayesian inference framework

The proposed robust Bayesian inference approach is based on the WGF formulation (Santambrogio, 2016). This method has been developed to deal with situations where the prior distribution is uncertain, but it can be described by an ambiguity set (Bayraksan and Love, 2015). This is useful, as in some Bayesian inference problems, notable difficulties arise to define the prior distributions of the latent parameters to be inferred. For example, when the suggestions of different experts about which prior distributions should be used significantly differ. For the cases where the amount of observed data is limited, significant changes of posterior may be found for different choices of prior distribution, and therefore, decisions to be taken for predictive maintenance may be affected. In these situations, an ambiguity set defined by a nominal prior distribution, a statistical distance, and a radius may be assumed, and the posteriors resulting from identifying the optimal and the worst-case prior distributions can be investigated by limiting the distribution space to priors within a statistical distance ε of the nominal prior distribution. In the next section, the concept of an ambiguity set is defined.

2.1. *Ambiguity set*

An ambiguity set is a set of distributions close to a distribution $p(\boldsymbol{\theta})$ with respect to some statistical distance r (Bayraksan and Love, 2015). An ambiguity set is defined by the nominal distribution $p(\boldsymbol{\theta})$, a statistical distance r, and a radius ε. The ambiguity set is used to restrict the space of distributions that the prior distribution could, in theory, take to solve the optimization of the chosen functional. Figure 1 shows an ambiguity set that is centered at a nominal distribution $p(\boldsymbol{\theta})$ and contains any distribution p^* within a statistical distance r less or equal to ε, this may be expressed as:

$$A(\varepsilon,p)=\{p^*:r(p^*(\theta)\|p(\theta))\leq\varepsilon\} \tag{1}$$

When the ambiguity set is defined, two conditions must be met (Go and Isaac, 2022): $\int_\Omega p^*(\boldsymbol{\theta})\,d\boldsymbol{\theta}=1$ and $r(p^*(\theta)\|p(\theta))\leq\varepsilon$.

Any statistical distance may be used to define the ambiguity set, but care should be taken in choosing this distance, as the distributions that lie inside that ambiguity set are defined by that statistical distance's properties. For example, if a phi divergence is used as the statistical distance in the ambiguity set, all

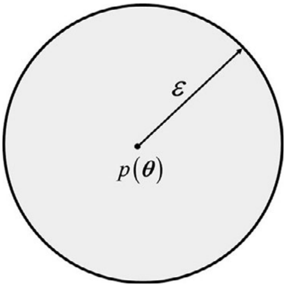


Figure 1. *Ambiguity set centered at* $p(\boldsymbol{\theta})$ *with radius* ε.

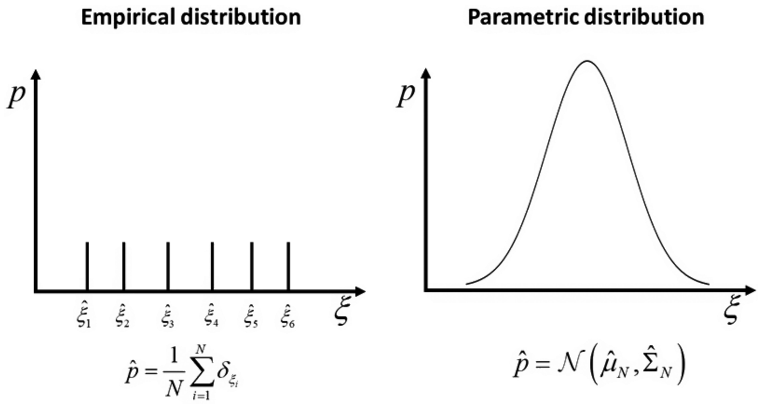

Empirical distribution
Parametric distribution
p
p
$\hat{\xi}_1$ $\hat{\xi}_2$ $\hat{\xi}_3$ $\hat{\xi}_4$ $\hat{\xi}_5$ $\hat{\xi}_6$
ξ
ξ
$\hat{p} = \frac{1}{N}\sum_{i=1}^{N}\delta_{\xi_i}$
$\hat{p} = \mathcal{N}\left(\hat{\mu}_N, \hat{\Sigma}_N\right)$

Figure 2. *Nominal distributions: empirical vs. parametric distribution.*

distributions inside the ambiguity set must be absolutely continuous w.r.t. the nominal distribution (van Parys et al., 2017). However, if the 2-Wasserstein distance is used to define the ambiguity set, then the distributions that lie within the ambiguity set do not need to be absolutely continuous w.r.t. the nominal distribution (Kuhn et al., 2019). The use of the 2-Wasserstein distance also means that distributions that lie within the ambiguity set do not need to share the same support (Kuhn et al., 2019).

Depending on the information that the practitioner has available, the nominal distribution may be given by either an empirical distribution or a parametric distribution (e.g., Gaussian distribution) as shown in Figure 2. In Figure 2, $\widehat{p}$ is a possible nominal distribution, N is the number of data points, δ is the Kronecker delta function, ξ is the parameter of the data, and $\mathcal{N}\left(\widehat{\mu}_N, \widehat{\Sigma}_N\right)$ is a Gaussian distribution with a sample mean $\widehat{\mu}_N$ and sample covariance $\widehat{\Sigma}_N$, obtained from N data points.

The chosen statistical distance for the ambiguity set is the 2-Wasserstein distance. As previously mentioned, this allows us to explore distributions that do not need to share the same support as the nominal distribution.

2.2. Simultaneous optimization of approximated posterior and optimal or worst-case prior distribution

In this article, we explore the problem of the simultaneous optimization of the approximation to the posterior and the optimal or worst-case prior distribution by using an interacting WGF scheme. The proposed approach differs from current Bayesian inference WGF-based approaches, as it formulates a new problem that requires interacting WGFs for the simultaneous optimization of the chosen functional. This interacting WGF simultaneously obtains the best approximation to the posterior and the optimal or worst-case prior distribution that either minimizes or maximizes a certain functional. The ambiguity set is used to restrict the space of distributions that the prior distribution could, in theory, take to solve the optimization of the chosen functional $E(\rho(\theta), p(\theta))$. In this article, the min-max (or min-min) formulation problem that needs to be solved is:

$$\min_{\rho(\theta)\in\mathcal{P}(\Omega)} \; \min \textit{ or } \max_{p*(\theta)\in\mathcal{W}(p(\theta),p*(\theta))\le\varepsilon} E(\rho(\theta), p(\theta))$$

where

$$E(\rho(\boldsymbol{\theta}), p(\boldsymbol{\theta})) := \int \rho(\boldsymbol{\theta}) \log\left(\frac{\rho(\boldsymbol{\theta})}{p(\boldsymbol{\theta})p(\boldsymbol{y}_{obs}|\boldsymbol{\theta})}\right) \tag{2}$$

The distribution $\rho(\boldsymbol{\theta})$ is the approximation to the posterior, the likelihood distribution is $p(\boldsymbol{y}_{obs}|\boldsymbol{\theta})$, the density $p(\boldsymbol{\theta})$ is the prior distribution, and $\mathcal{W}$ is the 2-Wasserstein distance chosen to define the ambiguity set. The chosen functional $E(\rho(\boldsymbol{\theta}), p(\boldsymbol{\theta}))$ is the KL divergence between the unnormalized posterior

$p(\boldsymbol{\theta}, \boldsymbol{y}_{obs})$ and the approximation to the posterior $\rho(\boldsymbol{\theta})$. This functional is chosen as it recently has been used to derive a WGF for Bayesian inference (Gao and Liu, 2020; Wang et al., 2022; Chen et al., 2023).

By using the properties of the logarithm, the functional in equation (2) can be rewritten as:

$$E(\rho(\boldsymbol{\theta}), p(\boldsymbol{\theta})) := \int \rho(\boldsymbol{\theta}) \log(\rho(\boldsymbol{\theta})) d\boldsymbol{\theta} - \int \rho(\boldsymbol{\theta}) \log(p(\boldsymbol{\theta})) d\boldsymbol{\theta} - \int \rho(\boldsymbol{\theta}) \log(p(\boldsymbol{y}|\boldsymbol{\theta})) d\boldsymbol{\theta}$$

The first term of the equation corresponds to the definition of entropy $\mathscr{H}$ w.r.t. the approximation to the posterior, therefore, the functional can be further expressed as:

$$E(\rho(\boldsymbol{\theta}), p(\boldsymbol{\theta})) := -\mathscr{H}(\rho(\boldsymbol{\theta})) - \int \rho(\boldsymbol{\theta}) \log(p(\boldsymbol{\theta})) d\boldsymbol{\theta} - \int \rho(\boldsymbol{\theta}) \log(p(\boldsymbol{y}|\boldsymbol{\theta})) d\boldsymbol{\theta} \tag{3}$$

The purpose of deriving an interacting WGF is to locate the pair of probability distributions (ρ^*, p^*) that balances the simultaneous minimization and maximization (or minimization) of the functional in equation (3). In other words, we are interested in finding simultaneously the distribution $\rho(\boldsymbol{\theta})$ that minimizes the KL divergence between the unnormalized posterior $p(\boldsymbol{\theta}, \boldsymbol{y}_{obs})$ and the approximation to the posterior $\rho(\boldsymbol{\theta})$, and the prior distribution(s) that minimizes/maximizes the KL divergence between the unnormalized posterior $p(\boldsymbol{\theta}, \boldsymbol{y}_{obs})$ and the approximation to the posterior $\rho(\boldsymbol{\theta})$.

In numerous occasions, efforts have been made to prove the convergence of algorithms with interacting WGFs to their global solution (Chizat and Bach, 2018; Mei et al., 2018), but these attempts generally require entropy regularization. The entropy regularization is already included in the formulation of the functional shown in equation (3), where the first term regularizes the partial differential equation of the WGF that minimizes the KL divergence between the posterior and the approximation to the posterior. The second term in equation (3) serves as a regularizer of the WGF that minimizes/maximizes the KL divergence between the posterior and the approximation to the posterior to obtain the optimal or worst-case prior distribution, respectively. In this article, it is assumed that the regularizers allow convergence to the pair of probability distributions that are sought. Proving convergence to this pair of probability distributions is still a problem currently under investigation and not attempted to be solved in the current article; the reader is referred to the Mixed Nash Equilibria literature for more details (Lin et al., 2019; Lu, 2022; Ding et al., 2023).

2.3. Proposed algorithm and workflow

The proposed approach is schematically summarized in Figure 3, and it is composed of three main parts: the inputs, the simultaneous functional optimization, and the outputs. The physics-based model (analytical, numerical, or equivalent surrogate model) of the engineering system of interest, a nominal prior distribution on the unknown latent parameters with a specified radius, a statistical distance to define the ambiguity set, an assumed likelihood, and measurements taken from the engineering system are needed as inputs. The main outputs, as shown in Figure 3, are the optimal or worst-case prior distribution and, consequently, the approximation to the posterior. The optimal or worst-case prior distribution is defined as the distribution that either minimizes or maximizes the KL divergence between the posterior and the approximation to the posterior, respectively.

In Figure 4, the elements of the optimization block shown in Figure 3 of the proposed approach are described. At first, we allow the approximation to the posterior to minimize the KL divergence between the posterior and the approximation to the posterior without changing the prior distribution. This is done by making the step size τ equal to zero on the WGF that results from either maximizing or minimizing the functional in equation (3) with respect to the prior distribution for a prescribed number of iterations N_a. The optimization to find the best approximation to the posterior is performed as follows. For the first iteration, N_0 initial particles are chosen at random (usually drawn from the nominal prior distribution), and the same set of particles is used for both the initial prior distribution $p_0(\boldsymbol{\theta})$ and the initial approximation to the posterior $\rho_0(\boldsymbol{\theta})$. At each iteration $i < N_a$, the physics-based model $PM(\boldsymbol{\theta})$ is run at the corresponding particle positions Θ_i^N of the approximation to the posterior. These numerical simulations at the particle

https://doi.org/10.1017/dce.2025.8 Published online by Cambridge University Press

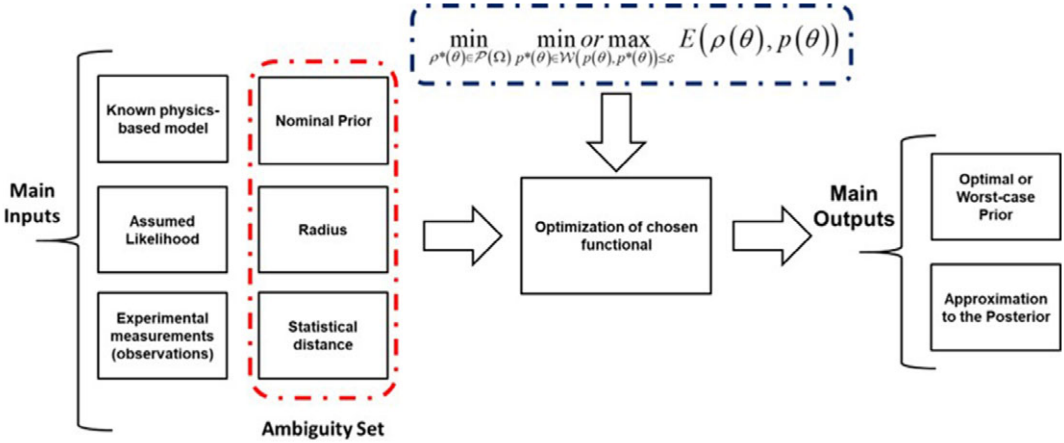

$\min_{\rho^*(\theta)\in\mathcal{P}(\Omega)} \min or \max_{p^*(\theta)\in\mathcal{W}(p(\theta),p^*(\theta))\le\varepsilon} E(\rho(\theta),p(\theta))$
Main Inputs
Known physics-based model
Assumed Likelihood
Experimental measurements (observations)
Nominal Prior
Radius
Statistical distance
Ambiguity Set
Optimization of chosen functional
Main Outputs
Optimal or Worst-case Prior
Approximation to the Posterior

Figure 3. *Main inputs, functional optimization, and main outputs of the proposed approach.*

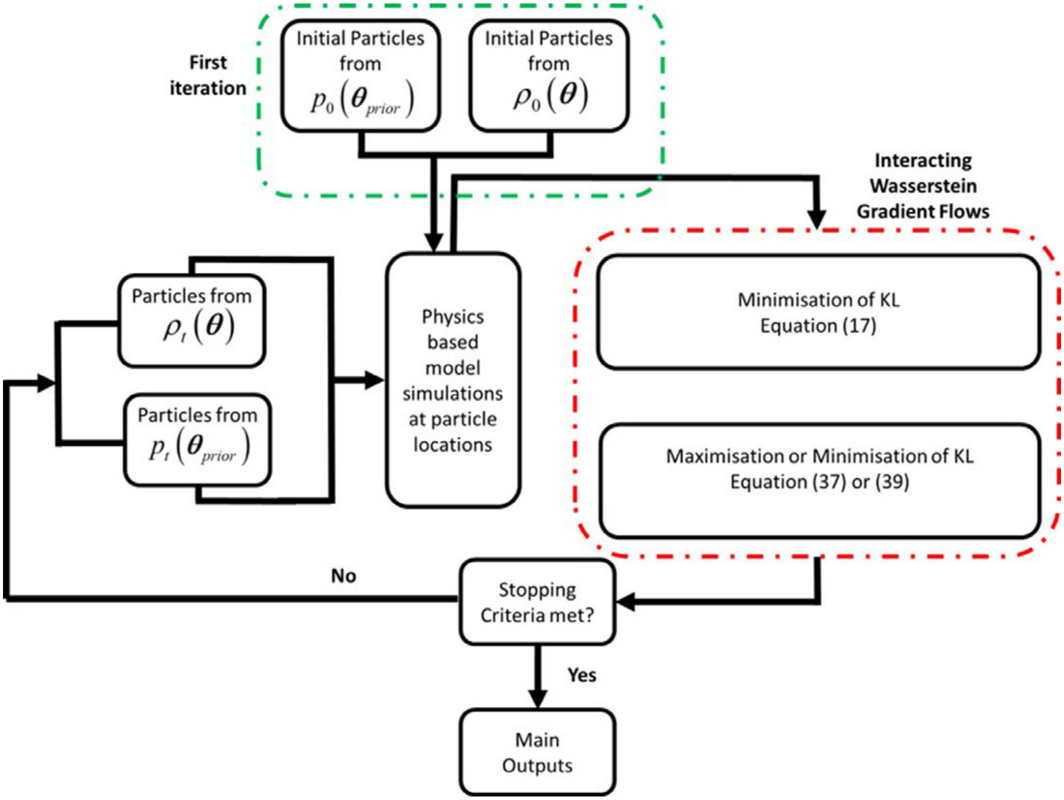

First iteration
Initial Particles from $p_0(\theta_{prior})$
Initial Particles from $\rho_0(\theta)$
Interacting Wasserstein Gradient Flows
Particles from $\rho_t(\theta)$
Particles from $p_t(\theta_{prior})$
Physics based model simulations at particle locations
Minimisation of KL Equation (17)
Maximisation or Minimisation of KL Equation (37) or (39)
No
Stopping Criteria met?
Yes
Main Outputs

Figure 4. *Pictorial description of simultaneous optimization of chosen functional.*

positions Θ_i^N are then used to calculate the gradient of the logarithm of the likelihood $\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{y}_{\text{obs}}|\boldsymbol{\theta})$ at those respective locations. The gradient of the logarithm of the prior distribution $\nabla_{\boldsymbol{\theta}} \log(p(\boldsymbol{\theta}))$ and the gradient of the logarithm to the approximation to the posterior $\nabla \log \rho_t(\boldsymbol{\theta})$, are approximated using a kernel density estimation (KDE) approach, where the bandwidth is chosen using the median approach (Liu and Wang, 2016). Using equation (17), a new set of N particles $\Theta_{i+1}^N \sim \rho_{i+1}(\boldsymbol{\theta})$ is obtained. This process is repeated until the iteration number reaches $i = N_a$, this is done to ensure that the approximation to the

https://doi.org/10.1017/dce.2025.8 Published online by Cambridge University Press

posterior has converged to the true posterior. For the cases studied, it was first assumed and later validated that the parameter N_a was enough to obtain convergence of the approximation to the posterior (by checking the 2-Wasserstein distance between successive iterations).

Once the prescribed number of iterations has been reached, the step size τ is allowed to be non-zero and positive, such that at every iteration $i \geq N_a$, a new set of prior particles $\boldsymbol{\theta}^N_{prior,i+1}$ is obtained using the second equation (40). At this stage, both the approximation to the posterior and prior distributions are updated using equation (40), such that the resulting new set of particles corresponds to independently and identically distributed samples from the distributions $\rho_{i+1}(\boldsymbol{\theta})$ and $p_{i+1}(\boldsymbol{\theta})$ of the next iteration number $i+1$.

Additionally, with the purpose of constraining the distribution to be optimized inside the ambiguity set, the 2-Wasserstein distance from the nominal prior distribution to the prior at iteration i is calculated at each iteration of the proposed method. If the distribution lies outside the ambiguity set, the distribution is discarded, and the size of the step in the particle flow algorithm is reduced until the distribution lies within the ambiguity set. In this way, the step size is controlled to restrict the prior distribution within the radius of the Wasserstein ambiguity set. This is based on the assumption that the distribution that maximizes or minimizes the KL divergence between the actual posterior and the approximation to the posterior lies at the radius of the ambiguity set. Moreover, if a preset number N_b of distributions are discarded when determining whether a distribution belongs in the ambiguity set, the distribution at iteration $i+1$ is reset to the distribution from an earlier iteration $i-N_c$ to avoid the optimization getting trapped at one of the local optima. Convergence of the prior distribution to the optimal or worst-case prior distribution is assumed if the previously mentioned resetting occurs N_{reset} times. In this case, the prior distribution is no longer reset, and in a manner similar to the one defined at the beginning of the algorithm, an additional number of iterations are allowed, so the approximation to the posterior can converge. At this stage, the algorithm checks if the stopping criteria have been fulfilled, if it has not, a new iteration $i+1$ is started. The stopping criteria are set as: (i) the maximum number of allowed iterations $N_{\max}$ is reached; (ii) a maximum number of prior distributions N_{reset} are reset to the distribution from an earlier iteration, and an additional number of iterations N_a are allowed for the approximation to the posterior to converge.

A summary of the steps to be run for the proposed method is given below:

1. Calculate approximation to the posterior for the initial prior distribution (by setting $\tau_t = 0$)
   a. Obtain N_0 initial particles from the prior distribution and approximation to the posterior.
   b. Calculate next set of particles of the approximation to the posterior using equation (17).
   c. Repeat from (1a) until iteration number reaches $i = N_a$.
2. Simultaneous optimization of equation (2) to calculate the approximation of the posterior and optimal or worst-case prior distribution (allow $\tau_t > 0$):
   a. Calculate the next set of particles of the approximation to the posterior and prior distributions using equations (17) and (37) or (39).
   b. Check if the prior distribution lies outside the defined ambiguity set:
      i. if false, continue to (2c).
      ii. if true:
         1. reduce the time step τ_t until it is inside ambiguity set.
         2. check if the number of discarded distributions is less than N_b.
            a. if true, continue.
            b. if false, reset current prior particles to prior particles from iteration $i - N_c$.
         3. check if the number of times prior distributions have been reset is less than N_{reset}.
            a. if true, continue.
            b. if false, skip to step 3.
   c. Repeat from (2a) until iterations reach $N_{\max}$ and stop running the algorithm.
3. Calculate the approximation to the posterior for the final prior distribution (by setting τ_t equal to zero, and allowing an additional number of iterations N_a):
   a. Calculate next set of particles of the approximation to the posterior distribution using equation (17).
   b. Repeat from (3a) until iterations reach $N_{\max}$ or the additional number of iterations is reached.

The system illustrated in Figure 5 is used to show the main results that would be obtained by using the proposed algorithm. A 1D mass-spring system with mass $m = 1$ [kg], stiffness $k = 1$ [N/m], and angular frequency $\omega = \sqrt{\frac{k}{m}}$ [rad/s] is studied. In this example, a Gaussian observational error is assumed when obtaining a numerical observation of the angular frequency $\omega_{obs} = \sqrt{\frac{k}{m}} + \zeta$, where $\zeta \sim \mathcal{N}(0,\sigma)$. It is also assumed that the uncertain parameter is the spring stiffness $k = \theta$ [N/m]. The initial Gaussian prior distribution (which is the same as the nominal prior distribution of the ambiguity set) is assumed to be $p(\theta) = \mathcal{N}(1,0.1)$. Two different runs to obtain the optimal prior distribution and the worst-case prior distribution (and their corresponding approximations to the posterior) w.r.t. the chosen functional in equation (2) are shown in Figure 6. An ambiguity set with a radius $\varepsilon = 0.005$ is chosen. For both the optimal and worst-case prior distributions, the step sizes in the interacting particle flow WGF algorithm cases are and $\tau = 3*10^{-4}$. The values of, N_b, N_c, N_{reset}, and $N_{\max}$ used to run the algorithm are the same as the two numerical examples shown in Section 6, and they can be found in the introduction of that section. Also, in this example, the number of initial particles $N_0 = 100$ is chosen. The obtained probability density functions (pdf) plotted in Figure 6 are calculated using the kde function in *MATLAB* (2022) with the standard options and using 100 samples from their respective distributions.

As expected, it can be seen in Figure 6 that the optimal prior distribution case assigns higher probability density at regions of high posterior density, while the worst-case prior distribution moves prior density away from regions of high posterior density. The optimal prior distribution has its support reduced w.r.t. the initial prior distribution, while for the worst-case prior distribution, its support is increased. A slight multimodality can be seen for the optimal prior distribution, with its main mode at the same location

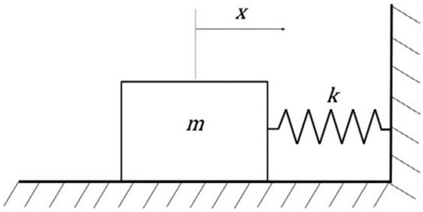
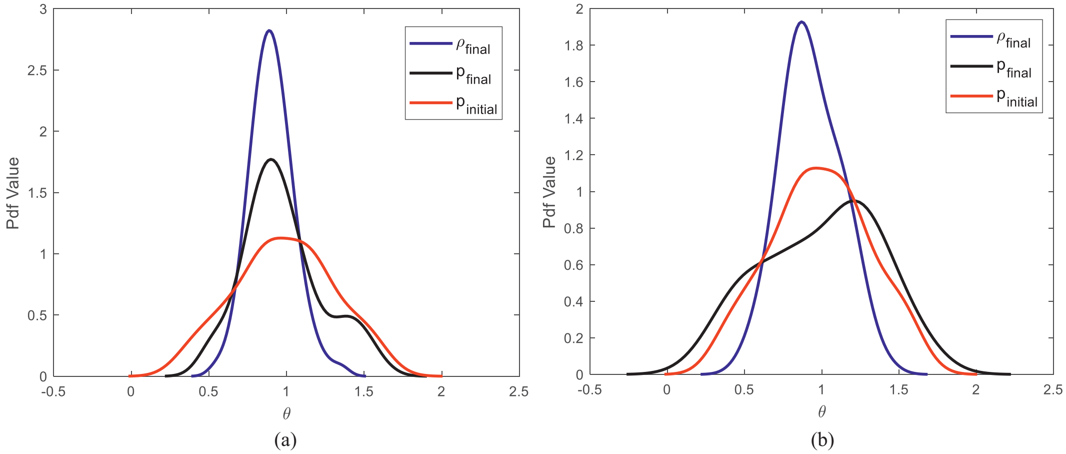


Figure 5. *1-Degree of freedom mass-spring system.*



(a)

(b)

Figure 6. *Kernel density estimates of the distributions for a 1D mass-spring system given an initial/ nominal prior distribution (red—initial prior distribution; blue—final approximation to the posterior distribution; and black—final prior distribution): (a) Optimal prior distribution case and (b) Worst-case prior distribution case.*

https://doi.org/10.1017/dce.2025.8 Published online by Cambridge University Press

as the only mode found for the approximation to the posterior distribution. Both the optimal and worst-case prior distributions are non-Gaussian and non-symmetric, even though the initial prior distribution was Gaussian and therefore symmetric. It should be noted that in this example, the radius and statistical distance used to define the ambiguity set are assumed to be known. The radius should be chosen in such a way that it captures uncertainty on the nominal prior distribution and allows to explore distributions that lie inside the ambiguity set.

The following sections of this article build upon the knowledge needed to understand the main concepts and algorithmic approximations required for the proposed approach.

3. Wasserstein gradient flow

In this article, to be able to consider the optimization of functionals with respect to probability measures, the WGF (Santambrogio, 2015) concept is introduced. The WFG applies on a probability measure space where a 2-Wasserstein metric has been defined.

Let us first consider the functional $E(\rho)$, where $E:\mathcal{P}(\Omega)\to\mathbb{R}$ maps a probability measure to a real value, $\mathcal{P}(\Omega)$ is the space of probability measures on $\Omega\subset\mathbb{R}^D$, and D is the number of dimensions.

To investigate the optimization of the functional $E(\rho)$ as a WGF, the Jordan Kinderleher Otto (JKO) scheme (Ambrosio et al., 2005; Santambrogio, 2016) is used. The JKO scheme solves the variational problem by defining the time discretization of the diffusion process; for this discretization, the approximate probability density, ρ_{i+1_τ} at the $i+1$ timestep is calculated:

$$\rho_{i+1}^{\tau} = \arg\min_{\rho}\left\{E(\rho)+\frac{\mathcal{W}_2^2\left(\rho,\rho_i^{\tau}\right)}{2\tau}\right\} \tag{4}$$

Where $\mathcal{W}_2$ is the 2-Wasserstein distance, $\tau>0$ is the size of the timestep, and as the size of the timestep approaches zero $\tau\to0$, the expression above converges to the exact WGF. The 2-Wasserstein distance (curve length between two distributions) is defined as (Santambrogio, 2015):

$$\mathcal{W}_2^2(\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)}\int_{\Omega\times\Omega}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2\gamma(d\boldsymbol{\theta},d\boldsymbol{\theta}^*) \tag{5}$$

where γ is the deterministic coupling that minimizes equation (5), and γ is inside the set of all possible couplings or joint distributions $\Gamma(\mu,\nu)$ over $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$, where μ and ν are the marginal distributions of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$. In the context of transport optimization, the calculation of the 2-Wasserstein distance can be interpreted as the transformation of elements in the domain μ to the domain ν at a minimum cost. Then, from this perspective, in equation (5) of the 2-Wasserstein distance, $\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2$ is the transportation cost of $\boldsymbol{\theta}$ in μ, to $\boldsymbol{\theta}^*$ in ν (Santambrogio, 2015). By defining the cost function c as $\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2$, equation (5) can be rewritten to:

$$\mathcal{W}_2^2(\mu,\nu) := \inf_{\mathcal{T}}\int_{\Omega}c(\boldsymbol{\theta},\mathcal{T}(\boldsymbol{\theta}))d\mu(\boldsymbol{\theta}) \tag{6}$$

For the cases where there is a unique solution for the problem of minimum transportation cost from $\boldsymbol{\theta}$ in μ, to $\boldsymbol{\theta}^*$ in ν, the unique solution can also be expressed as a mapping $\mathcal{T}:\mathbb{R}^D\to\mathbb{R}^D$, that pushes elements $\boldsymbol{\theta}$ of the domain μ to the domain ν (Chen et al., 2018). The solution is unique when the marginal distribution of probability μ is absolutely continuous w.r.t. the Lebesgue measure (Chen et al., 2018).

If $\{\mu_t\}_{t\in[0,1]}$ is an absolutely continuous curve with finite second-order moments in the probabilistic space $\mathcal{P}(\Omega)$, then the changes of μ_t in that curve will be defined through investigation of $\mathcal{W}_2^2\left(\mu_t,\mu_{t+\tau}\right)$. Studying the changes of μ_t, is related to the original JKO problem (Ambrosio et al., 2005; Santambrogio, 2016) of the minimization of the functional shown in equation (4). These changes can be described using a velocity field given by: $\boldsymbol{v}_t(\boldsymbol{\theta}) := \lim_{\tau\to0}\frac{\mathcal{T}(\boldsymbol{\theta}_t)-\boldsymbol{\theta}_t}{\tau}$. This velocity field $\boldsymbol{v}_t(\boldsymbol{\theta})$ defines in $\mathcal{P}(\Omega)$ the gradient flow (Ambrosio et al., 2005):

$$\partial_t \mu_t + \nabla \cdot (\boldsymbol{v}_t \mu_t) = 0 \tag{7}$$

Solving equation (7) requires finding a velocity field $\boldsymbol{v}(t)$ such that its flow agrees with $\lim_{\tau \to 0}(\boldsymbol{\theta}_\tau(t))$. The WGF can be shown to have a velocity field $\boldsymbol{v}(t)$ that minimizes the functional $E(\rho)$, with the following form $\boldsymbol{v}(t) = -\nabla \frac{\partial E(\rho)}{\partial \rho}$ (Ambrosio et al., 2005), where $\frac{\partial E(\rho)}{\partial \rho}$ is called the first variation of $E(\rho)$ at ρ. Based on this, the WGF may be expressed as:

$$\partial_t \rho_t = -\nabla \cdot (\boldsymbol{v}_t \rho_t) = \nabla \cdot \left(\rho_t \nabla \frac{\partial E(\rho_t)}{\partial \rho_t} \right) \tag{8}$$

Therefore, to derive the WGF for the optimization of the functional $E(\rho)$, the following requirements are introduced:

1. The first variation of the functional $E(\rho)$ with respect to the density $\frac{\partial E(\rho)}{\partial \rho}$ needs to be calculated.
2. A perturbation that follows the formal definition of a derivative in the probability space has to be introduced.
3. The probability ρ is a probability density $\rho \in \mathcal{P}(\Omega)$ that has to be perturbed to $\rho + \varepsilon\chi$, which is also another probability density such that it also lies in the probability space $\mathcal{P}(\Omega)$, in this way, $E(\rho + \varepsilon\chi)$ is well defined.
4. For all small $\varepsilon > 0$, both the perturbed probability density is defined in the probability space $\rho + \varepsilon\chi \in \mathcal{P}(\Omega)$ and $\sigma = \rho + \chi \in \mathcal{P}(\Omega)$.

This can also be rewritten as $\rho + \varepsilon\chi = \rho + \varepsilon(\sigma - \rho) = \rho(1-\varepsilon) + \varepsilon\sigma$, where $\rho(1-\varepsilon) + \varepsilon\sigma \in \mathcal{P}(\Omega)$, as long as ρ and σ are also probability densities.

Now that the requirements have been introduced, the first variation of $E(\rho)$, $\frac{\partial E(\rho)}{\partial \rho}$ can be found, and it is given as (Ambrosio et al., 2005; Santambrogio, 2016):

$$\frac{\partial}{\partial \varepsilon} E(\rho + \varepsilon\chi)\bigg|_{\varepsilon=0} = \int_\Omega \frac{\partial E(\rho)}{\partial \rho} \chi(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{9}$$

for all $\chi = \sigma - \rho$. If a constant C is added, $\int_\Omega \left(\frac{\partial E(\rho)}{\partial \rho} + C \right) \chi(\boldsymbol{\theta}) d\boldsymbol{\theta}$, it can be found that the first variation may be defined uniquely only up to additive constants, as that second integral $\int_\Omega \chi(\boldsymbol{\theta}) d\boldsymbol{\theta}$ includes the difference of 2 probability densities $\chi = \sigma - \rho$.

4. Wasserstein gradient flow for Bayesian inference

Approximations to the posterior can be obtained using many different methods. Recently, methods based on VI have been gaining popularity (Blei et al., 2017). These methods are based on the minimization of the KL divergence between the posterior $p(\boldsymbol{y}_{obs}|\boldsymbol{\theta})$ and a probability density (usually parametric) defined inside a family of distributions $\mathcal{Q}$, to quantify the degree of dissimilarity between two distributions over the same domain:

$$\rho^* = \underset{\rho \in \mathcal{Q}}{\arg\min}\, KL(\rho \| p(\boldsymbol{\theta}|\boldsymbol{y}_{\text{obs}})) \tag{10}$$

where the KL divergence is defined as:

$$KL(\rho | p(\boldsymbol{\theta}|\boldsymbol{y}_{obs})) = \int_\Omega \rho \log\left(\frac{\rho}{p(\boldsymbol{\theta}|\boldsymbol{y}_{obs})} \right) d\boldsymbol{\theta} \tag{11}$$

The approximation to the posterior is obtained by finding the member of the family and its respective hyperparameters that best minimize the KL divergence (Blei et al., 2017).

https://doi.org/10.1017/dce.2025.8 Published online by Cambridge University Press

An alternative to VI would be to use the WGF to define an iterative procedure that uses the set of data $\boldsymbol{y}_{obs}$ to update a chain of $\rho_n(\boldsymbol{\theta})$ with the purpose of approximating $p(\boldsymbol{\theta}|\boldsymbol{y}_{obs})$ given the minimization of a suitable functional $E(\rho)$. In WGF, the optimization of the functional can be solved by using equation (8). To be able to solve this equation, the velocity field $\boldsymbol{v}(t)$ given the chosen functional is required. In a first analysis, it may be thought that as the posterior $p(\boldsymbol{\theta}|\boldsymbol{y}_{obs})$ is not known in advance because of the presence of the normalization constant $p(\boldsymbol{y}_{obs})$), then the functional of equation (11) cannot be used to derive a WGF. But as the first variation of the functional is only uniquely defined up to additive constants, a simpler functional $E(\rho)$ where the posterior $p(\boldsymbol{\theta}|\boldsymbol{y}_{obs})$ is replaced for the unnormalized posterior $p(\boldsymbol{\theta},\boldsymbol{y}_{obs})$ may be used (Gao and Liu, 2020). Therefore, the velocity field that results from replacing the posterior with the unnormalized posterior would be the same as the velocity field as in the functional in equation (11).

By obtaining a WGF of the functional $E(\rho)$, the partial differential equation can be solved to flow the approximation of the posterior $\rho_t(\boldsymbol{\theta})$ to its equilibrium $p(\boldsymbol{\theta}|\boldsymbol{y}_{obs})$ for the observed data. The dynamic system is defined by an initial density $\rho_0(\boldsymbol{\theta})$ that is given by the prior distribution $p(\boldsymbol{\theta})$ and $\rho_\infty(\boldsymbol{\theta})$ tends to the posterior distribution (Gao and Liu, 2020). In a more rigorous manner, in a manifold M in the parameter space, a pushforward density $\rho_t(\boldsymbol{\theta}) = \mathcal{T}_t \# p(\theta) \in M$ is considered, where # is the push forward operator, and the best curve (under certain restrictions) ρ_t, that drives ρ_0 to ρ_∞ has to be found (Gao and Liu, 2020).

The WGF of the chosen functional $E(\rho)$, may be performed by first calculating the first variation (where the bounds are omitted for clarity):

$$\frac{\partial}{\partial \varepsilon} E(\rho+\varepsilon\chi)\bigg|_{\varepsilon=0} = \frac{\partial}{\partial \varepsilon}\left[\int (\rho+\chi\varepsilon)\log((\rho+\chi\varepsilon))d\boldsymbol{\theta} - \int (\rho+\chi\varepsilon)\log(p(\boldsymbol{\theta},\boldsymbol{y}_{obs}))d\boldsymbol{\theta}\right]\bigg|_{\varepsilon=0} =$$

$$= \left[\int \chi\log(\rho+\chi\varepsilon)d\boldsymbol{\theta} + \int (\rho+\chi\varepsilon)\frac{\chi}{(\rho+\chi\varepsilon)}d\boldsymbol{\theta} - \int (\rho+\chi\varepsilon)\log(p(\boldsymbol{\theta},\boldsymbol{y}_{obs}))d\boldsymbol{\theta}\right]\bigg|_{\varepsilon=0} =$$

$$= \int \chi\log(\rho)d\boldsymbol{\theta} + \int \chi d\boldsymbol{\theta} - \int \chi\log(p(\boldsymbol{\theta},\boldsymbol{y}_{obs}))d\boldsymbol{\theta} = \int (\log(\rho) + 1 - \log(p(\boldsymbol{\theta},\boldsymbol{y}_{obs})))\chi d\boldsymbol{\theta} \tag{12}$$

The first variation of the functional $E(\rho)$ with respect to the density ρ is then given by:

$$\frac{\partial E(\rho)}{\partial \rho} = \log(\rho) + 1 - \log(p(\boldsymbol{\theta},\boldsymbol{y}_{obs})) \tag{13}$$

and the velocity field is:

$$\boldsymbol{v}(t) = -\nabla\frac{\partial E(\rho)}{\partial \rho} = \nabla(\log(p(\boldsymbol{\theta},\boldsymbol{y}_{obs})) - \log(\rho) - 1) = \nabla\log(p(\boldsymbol{\theta},\boldsymbol{y}_{obs})) - \nabla\log(\rho) \tag{14}$$

If the first variation of the functional $E(\rho)$ is introduced into the continuity equation, the following equation is obtained (Wang et al., 2022):

$$\partial_t \rho_t = \nabla\cdot(\rho_t(\nabla\log(p(\boldsymbol{\theta},\boldsymbol{y}_{obs})) - \nabla\log(\rho_t))) \tag{15}$$

The KL WGF is an approximation in continuous time of the deterministic mean-field particle system called mean-field Wasserstein dynamics (Wang et al., 2022):

$$d\boldsymbol{\theta}_t = [\nabla\log(p(\boldsymbol{\theta},\boldsymbol{y}_{obs})) - \nabla\log(\rho_t)]dt \tag{16}$$

The mean-field term is derived from the fact that the dynamics' evolution varies with the current density function ρ_t. The deterministic particle descent WGF may be obtained from the mean-field Wasserstein dynamics (Wang et al., 2022):

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha_t(\nabla\log(p(\boldsymbol{\theta},\boldsymbol{y}_{obs})) - \nabla\log(\rho_t)) \tag{17}$$

https://doi.org/10.1017/dce.2025.8 Published online by Cambridge University Press

Equation (17) represents one of the two particle discretization WGF equations needed for the simultaneous optimization of the chosen functional in equation (3) shown in Section 2. In equation (17), an approximation of $\nabla \log(\rho_t)$ is required, as no analytical expression is available. Many different methods may be used to obtain an approximation. In this article, a KDE approach is chosen and explained in Section 5.1. It should be noted that the WGF follows a deterministic rule for the updating, and therefore the initial positions of the system determine the particle interactions and randomness.

5. Approximations in Wasserstein gradient flow for robust Bayesian inference

This section provides a more detailed explanation of some of the mathematical tools required for the application of the algorithm described in Figure 4.

5.1. Approximation to $\nabla_{\boldsymbol{\theta}} \log(\rho)$ *from samples*

When the velocity field $\boldsymbol{v}(t)$ is to be approximated, one of the difficulties that arises is the estimation of $\nabla \log \rho(\boldsymbol{\theta})$ (Liu et al., 2018). Only a finite set of samples $\left\{\theta^{(i)}\right\}_{i=1}^{N}$ of $\rho(\boldsymbol{\theta})$ is known. However, a direct approximation of $\rho(\boldsymbol{\theta})$ using the empirical distribution $\widehat{\rho}(\boldsymbol{\theta}) := \frac{1}{N}\sum_{i=1}^{N}\delta\left(\boldsymbol{\theta}-\boldsymbol{\theta}^{(i)}\right)$, where δ is the Dirac delta function is not possible. The reason why a direct approximation cannot be performed is because the WGF of the KL divergence at $\widehat{\rho}(\boldsymbol{\theta})$ is not defined, a consequence of $\widehat{\rho}(\boldsymbol{\theta})$ not being absolutely continuous. Using the absolutely continuous approximated expression $\tilde{\rho}(\boldsymbol{\theta}) := (\widehat{\rho} * K)(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} K\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(i)}\right)$ ("$*$" symbolizes convolution), the velocity field $\boldsymbol{v}(t)$ can be well-defined by smoothing $\widehat{\rho}(\boldsymbol{\theta})$ through a smooth kernel K on $\boldsymbol{\theta}$.

In this article, the approximation of $\rho(\boldsymbol{\theta})$ is produced using the KDE $\tilde{\rho}(\boldsymbol{\theta})$, where $K\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(i)}\right) : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ is a given positive and differentiable kernel function, and the Gaussian kernel is used:

$$K(\boldsymbol{\theta},\boldsymbol{\theta}^*) = (2\pi h)^{-\frac{N}{2}} \exp\left(-\frac{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2}{2h}\right) \tag{18}$$

where N is the number of samples used to define the kernel function $K(\boldsymbol{\theta},\boldsymbol{\theta}^*)$, h is the bandwidth and it is defined by $h = \text{med}^2/\log(N)$, and med represents the median of distances of the samples (Liu and Wang, 2016).

The reasoning behind this choice of kernel function is based on the fact that $\sum_j k\left(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j\right) \approx n \exp\left(-\frac{1}{h}\text{med}^2\right) = 1$, therefore, for each point $\boldsymbol{\theta}_i$, its own gradient contribution and the effect from the other points balance each other (Liu and Wang, 2016).

When the KDE is used as an approximation of $\rho(\boldsymbol{\theta})$, the following expression may be used to calculate an approximation of $\nabla \log \rho(\boldsymbol{\theta})$ (Wang et al., 2022):

$$\nabla \log \tilde{\rho}(\boldsymbol{\theta}) = \frac{\nabla \tilde{\rho}(\boldsymbol{\theta})}{\tilde{\rho}(\boldsymbol{\theta})} = \frac{\sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} K\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(i)}\right)}{\sum_{i=1}^{N} K\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(i)}\right)} \tag{19}$$

The kernel chosen does not affect the solution of the gradient flow if the size of the ensemble tends to infinity (Lu et al., 2019). Nonetheless, the distribution of particles for a finite number of them may not be unique. An alternative manner to explain this is that for different given kernels, that is, with different particle flows, different results (final positions in the state space of the particles) are obtained. However, for those kernels, as their number of particles increases, the representation of the posterior pdf becomes more accurate.

5.2. Approximation to $\nabla_{\boldsymbol{\theta}} \log(p(\boldsymbol{\theta},\boldsymbol{y}_{obs}))$

In this article, two different ways to estimate the gradient of log-likelihood are considered. The first one uses local estimations of that Jacobian matrix of the model's ensemble, whereas the second one uses Gaussian processes. The first approach is only able to obtain estimates of the gradient of log-likelihood at particle positions where the model has been run previously. However, the Gaussian process approach is able to obtain estimates of the gradient of log-likelihood at particle positions that have not been evaluated by leveraging on the prior distribution assumptions and previous model runs. The choice of approach is usually based on the computational cost of dealing with the physics-based model involved.

In general, $\nabla_{\boldsymbol{\theta}} \log(p(\boldsymbol{\theta}))$ can be calculated analytically, as most of the $\log(p(\boldsymbol{\theta}))$ chosen in Bayesian inference are differentiable. However, if an analytical expression is not available, $\nabla_{\boldsymbol{\theta}} \log(p(\boldsymbol{\theta}))$ may be approximated using equation (19) as long as samples from the prior a distribution are available.

5.2.1. Gradient of log-likelihood using the ensemble method

Assuming a multivariate Gaussian likelihood, $p(\boldsymbol{y}_{\mathrm{obs}}|\boldsymbol{\theta})$ can be written as:

$$p(\boldsymbol{y}_{\mathrm{obs}}|\boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}(\boldsymbol{y}_{\mathrm{obs}} - \boldsymbol{y}_{\mathrm{model}})^T \Sigma^{-1} (\boldsymbol{y}_{\mathrm{obs}} - \boldsymbol{y}_{\mathrm{model}})\right) \tag{20}$$

In the above expression, d refers to the dimensionality of the observation space (the number of observations), $\boldsymbol{y}_{\mathrm{model}}$ and $\boldsymbol{y}_{\mathrm{obs}}$, respectively, are the $n \times 1$ vectors of simulated and observed states, and the inverse of the $n \times n$ error covariance matrix Σ is denoted by Σ^{-1}.

By taking the gradient of the logarithm of the multivariate Gaussian likelihood, the below expression is obtained:

$$\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{y}_{\mathrm{obs}}|\boldsymbol{\theta}) = \frac{1}{2} \nabla_{\boldsymbol{\theta}} \boldsymbol{y}_{\mathrm{model}}{}^T \Sigma^{-1} (\boldsymbol{y}_{\mathrm{obs}} - \boldsymbol{y}_{\mathrm{model}}) \tag{21}$$

In equation (21), $\nabla_{\boldsymbol{\theta}} \boldsymbol{y}_{\mathrm{model}}$ is a matrix of dimensions $n \times D$. The number of model parameters is denoted by D. The elements of the $\nabla_{\boldsymbol{\theta}} \boldsymbol{y}_{\mathrm{model}}$ matrix are the partial derivatives of each simulated state (associated to rows 1, …, n) w.r.t. each parameter (associated to columns 1, …, D). The states are simulated by introducing input parameters $\boldsymbol{\theta}$ into a computational model:

$$\boldsymbol{y}_{\mathrm{model}} = PM(\boldsymbol{x}, \boldsymbol{\theta}) \tag{22}$$

The expression above assumes that the observed states are directly simulated by the model. If the Jacobian is defined as $J(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} PM(\boldsymbol{x}, \boldsymbol{\theta})^T$, the matrix of dimensions $(D \times n)$, equation (21) can be rewritten as:

$$\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{y}_{\mathrm{obs}}|\boldsymbol{\theta}) = \frac{1}{2} J(\boldsymbol{\theta})^T \Sigma^{-1} (\boldsymbol{y}_{\mathrm{obs}} - \boldsymbol{y}_{\mathrm{model}}) \tag{23}$$

As a result, using equation (23), the log-likelihood gradient may be evaluated using local estimations of that Jacobian matrix. Computational difficulties arise during the evaluation of the Jacobian matrix $J(\boldsymbol{\theta})$ of dimensions $(n \times D)$, as the closed form of this matrix is frequently unavailable.

To solve the mentioned difficulty, an approach that consists of obtaining nonintrusive estimations of the Jacobian $J(\boldsymbol{\theta})$ may be taken (Ramgraber et al., 2021). The vector $\boldsymbol{\theta}$, that has the parameters as elements, is perturbed in a small increment in each of its D dimensions, and the Jacobian matrix $J(\boldsymbol{\theta})$ is approximated using the obtained two- or three-point finite difference derivatives. This computational differentiation may produce very accurate results, but it becomes unpractical if the model has a high number of parameters D. If the ensemble size or number of particles is denoted N, and a set of local Jacobians is to be required, the model has to be run $(D+1)N$ times if two-points finite difference derivatives are used, or even more times $(2D+1)N$, if three-points finite difference derivatives are chosen (Ramgraber et al., 2021). Those numbers are well above the number of evaluations of the model that practitioners may consider affordable.

A technique that requires only N model evaluations $PM(\boldsymbol{x},\boldsymbol{\theta})$, and it is able to produce the estimation of the Jacobian matrix $J(\boldsymbol{\theta})$, directly from the ensemble, may be found in Ramgraber et al. (2021). This methodology makes use of the relative differences between particles:

$$\tilde{J}(\boldsymbol{\theta}_r) = \frac{P}{N}\sum_{r=1}^{N}\frac{PM(\boldsymbol{\theta}_r)-PM(\boldsymbol{\theta}_s)}{\|PM(\boldsymbol{\theta}_r)-PM(\boldsymbol{\theta}_s)\|}\cdot\frac{\|PM(\boldsymbol{\theta}_r)-PM(\boldsymbol{\theta}_s)\|}{\|\boldsymbol{\theta}_r-\boldsymbol{\theta}_s\|}\cdot\frac{\boldsymbol{\theta}_r{}^T-\boldsymbol{\theta}_s{}^T}{\|\boldsymbol{\theta}_r-\boldsymbol{\theta}_s\|} \tag{24}$$

In equation (24), P is the rank expected for the Jacobian matrix $J(\boldsymbol{\theta})$, this expected rank is the smallest value between $N-1$ and D. Inside the summation symbol, three fractions are found in correlative order: the vector from particle $\boldsymbol{\theta}_r$ to the particle $\boldsymbol{\theta}_s$ (normalized), the scalar gradient between the observation and the parameter space, and the normalized vector in parameter space. Equation (24) may be simplified as follows (Ramgraber et al., 2021):

$$\tilde{J}(\boldsymbol{\theta}_r) = \frac{P}{N}\sum_{r=1}^{N}\frac{(PM(\boldsymbol{\theta}_r)-PM(\boldsymbol{\theta}_s))\left(\boldsymbol{\theta}_r{}^T-\boldsymbol{\theta}_s{}^T\right)}{\|\boldsymbol{\theta}_r-\boldsymbol{\theta}_s\|^2} \tag{25}$$

The factor $\frac{P}{N}$ external to the sum is made up of a correction factor P to consider that the maximum possible contribution of each vector to the rank of the Jacobian is one, and a factor $\frac{1}{N}$ to account for an arithmetical average. For $N\to\infty$ and an isotropic arrangement of particles, the Jacobian in equation (25) should converge against the correct one (Ramgraber et al., 2021).

5.2.2. Gradient of log-likelihood using Gaussian process

For cases where the physics-based model is expensive to evaluate, an approximation of the gradient of the log-likelihood may be produced using a Gaussian process. This methodology allows the estimation of the gradient at particle positions where the physics-based model has not been evaluated.

Assuming that the likelihood function is given by a multivariate Gaussian with zero error mean and covariance Σ, the log-likelihood function is:

$$\log p(\boldsymbol{y}_{\text{obs}}|\boldsymbol{\theta}) = -\frac{d}{2}\log(2\pi)-\frac{1}{2}\log(\det(\Sigma))-\frac{1}{2}(\boldsymbol{y}_{\text{obs}}-\boldsymbol{y}_{\text{model}})^T\Sigma^{-1}(\boldsymbol{y}_{\text{obs}}-\boldsymbol{y}_{\text{model}}) \tag{26}$$

Focus is placed on the last term of equation (26), as the gradient of the log-likelihood function w.r.t. the parameter $\boldsymbol{\theta}$ only depends on that term. Consequently, the partially observed potential is modeled as (Dunbar et al., 2022):

$$V_L(\theta) = \frac{1}{2}(\boldsymbol{y}_{\text{obs}}-\boldsymbol{y}_{\text{model}})^T\Sigma^{-1}(\boldsymbol{y}_{\text{obs}}-\boldsymbol{y}_{\text{model}}) \tag{27}$$

Where $V_L(\theta)$ is a Gaussian process $f\sim GP(0,k)$ and κ denotes a positive definite kernel on $\mathbb{R}^D$ that has been chosen according to the explanations below.

In this article, κ is a Gaussian radial basis function kernel that has the form:

$$\kappa(\boldsymbol{\theta},\boldsymbol{\theta}^*;\lambda,l) = \lambda\exp\left(-\frac{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|^2}{2l^2}\right)$$

In this expression, $l>0$ denotes the kernel bandwidth and $\lambda>0$ is the amplitude of the kernel. A function f is sought so that for some $\sigma>0$, and for some noisy evaluations of the potential at the ensemble of points $\Theta_t = \left(\Theta_t^1,\dots,\Theta_t^N\right)\in\mathbb{R}^{N\times D}$, then (Dunbar et al., 2022):

$$V_L\left(\Theta_t^i\right) = f\left(\Theta_t^i\right)+\sigma\xi^i,\ \xi^i = \left(\xi^1,\dots,\xi^N\right)\sim\mathcal{N}(0,I) \tag{28}$$

The mean function of the associated Gaussian process posterior for f is (Rasmussen, 2003):

$$\mu(\theta^*) = \kappa(\theta^*,\Theta)K(\Theta,\Theta)^{-1}V_L(\Theta) \tag{29}$$

and the expression for the variance function is (Rasmussen, 2003):

https://doi.org/10.1017/dce.2025.8 Published online by Cambridge University Press

$$\sigma^2(\theta^*) = \kappa(\theta^*,\theta^*) - \kappa(\theta^*,\Theta)K(\Theta,\Theta)^{-1}\kappa(\Theta,\theta^*) \tag{30}$$

Where $K(\Theta,\Theta) = diag(\sigma^2) + \kappa(\Theta,\Theta)$. Equations (31) and (32) express the well-defined gradient of the posterior mean (Rasmussen, 2003):

$$\mathbb{E}\left[\frac{\partial V_L(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}_d^*}\right] = \frac{\partial \mathbb{E}[V_L(\boldsymbol{\theta}^*)]}{\partial \boldsymbol{\theta}_d^*} = \frac{\partial \kappa(\theta^*,\Theta)}{\partial \boldsymbol{\theta}_d^*} K(\Theta,\Theta)^{-1} V_L(\Theta) \tag{31}$$

$$\nabla_{\boldsymbol{\theta}} V_L(\Theta) = \left.\frac{\partial \kappa(\boldsymbol{\theta}^*,\Theta)}{\partial \boldsymbol{\theta}_d^*}\right|_{\theta^*=\Theta} K(\Theta,\Theta)^{-1} V_L(\Theta) \tag{32}$$

Both the energy term $V_L(\theta)$ and the hyperparameters (σ,λ,l) are updated at each iteration and are calculated considering the new incoming data (Rasmussen, 2003).

5.3. Derivations of Wasserstein gradient flow equations for optimal or worst-case prior distribution

In Section 4, the WGF for the case when the approximation to the posterior is made to vary to minimize the KL divergence between the posterior and the approximation to the posterior has been derived. Now, the WGF that either maximizes or minimizes the KL divergence between the posterior and the approximation to the posterior with respect to the prior distribution needs to be calculated. Currently, the interacting WGF has the following form:

$$\begin{cases} \partial_t \rho_t = \nabla \cdot (\rho_t (\nabla \log(p(\boldsymbol{\theta},\boldsymbol{y}_{obs})) - \nabla \log(\rho_t))) \\ \partial_t p_t(\boldsymbol{\theta}) = \eta\left(-\nabla \cdot \left(p_t(\boldsymbol{\theta}) \nabla \left(\frac{\partial E}{\partial p(\boldsymbol{\theta})}(\rho, p(\boldsymbol{\theta}))\right)\right)\right) \end{cases} \tag{33}$$

The first step is to calculate the first variation of the functional $E(\rho(\boldsymbol{\theta}),p(\boldsymbol{\theta}))$ with respect to the prior distribution $p(\boldsymbol{\theta})$. When the optimal prior distribution is of interest, this results in the minimization of equation (3), to obtain an expression of the first variation, we first need to calculate the following:

$$\left.\frac{\partial}{\partial \varepsilon} E(p(\boldsymbol{\theta}) + \varepsilon\chi)\right|_{\varepsilon=0} = \left.\frac{\partial}{\partial \varepsilon}\left[\int \rho \log \rho d\boldsymbol{\theta} - \int \rho \log(p(\boldsymbol{\theta}) + \varepsilon\chi) d\boldsymbol{\theta} - \int \rho \log(p(\boldsymbol{\theta}|\boldsymbol{y}_{obs})) d\boldsymbol{\theta}\right]\right|_{\varepsilon=0} =$$

$$= \left.\left[-\int \rho \frac{\chi}{(p(\boldsymbol{\theta}) + \chi\varepsilon)} d\boldsymbol{\theta}\right]\right|_{\varepsilon=0} = -\int \frac{\rho}{p(\boldsymbol{\theta})} \chi d\boldsymbol{\theta} \tag{34}$$

Now an expression of the first variation of the functional to be optimized can be obtained, and it is given by:

$$\frac{\partial E}{\partial p(\boldsymbol{\theta})}(p(\boldsymbol{\theta})) = -\frac{\rho}{p(\boldsymbol{\theta})} \tag{35}$$

and the velocity field is:

$$\boldsymbol{v}(t) = -\nabla \frac{\partial E}{\partial \rho}(\rho) = \nabla\left(\frac{\rho}{p(\boldsymbol{\theta})}\right) = \frac{p(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} \nabla\left(\frac{\rho}{p(\boldsymbol{\theta})}\right) = \frac{\rho}{p(\boldsymbol{\theta})}(\nabla \log \rho - \nabla \log p(\boldsymbol{\theta})) \tag{36}$$

The resulting particle-based WGF, using an Euler discretization, is given as:

$$\boldsymbol{\theta}^N_{prior,t+1} = \boldsymbol{\theta}^N_{prior,t} + \tau_t \left(\frac{\rho_t(\boldsymbol{\theta}_{prior})}{p_t(\boldsymbol{\theta}_{prior})}(\nabla \log p_t(\boldsymbol{\theta}_{prior}) - \nabla \log \rho_t(\boldsymbol{\theta}_{prior}))\right) \tag{37}$$

https://doi.org/10.1017/dce.2025.8 Published online by Cambridge University Press

If the maximization of the KL divergence is sought instead, this requires the calculation of the worst-case prior distribution, and the resulting velocity field is given as the negative of the previously calculated velocity field:

$$\boldsymbol{v}(t) = \frac{\rho}{p(\boldsymbol{\theta})}(\nabla \log p(\boldsymbol{\theta}) - \nabla \log \rho) \tag{38}$$

Therefore, the resulting particle-based WGF using an Euler discretization is given as:

$$\boldsymbol{\theta}^N_{prior,t+1} = \boldsymbol{\theta}^N_{prior,t} - \tau_t \left(\frac{\rho_t(\boldsymbol{\theta}_{prior})}{p_t(\boldsymbol{\theta}_{prior})} (\nabla \log p_t(\boldsymbol{\theta}_{prior}) - \nabla \log \rho_t(\boldsymbol{\theta}_{prior})) \right) \tag{39}$$

Now that the particle-based WGF for the minimization or maximization of the functional with respect to the prior distribution has been derived, an interacting particle-based WGF can be defined as follows:

$$\begin{cases} \boldsymbol{\theta}^N_{t+1} = \boldsymbol{\theta}^N_t + \alpha_t(\nabla \log(p_t(\boldsymbol{\theta}, \boldsymbol{y}_{obs})) - \nabla \log(\rho_t)) \\ \boldsymbol{\theta}^N_{prior,t+1} = \boldsymbol{\theta}^N_{prior,t} \pm \tau_t \left(\frac{\rho_t(\boldsymbol{\theta}_{prior})}{p_t(\boldsymbol{\theta}_{prior})} (\nabla \log p_t(\boldsymbol{\theta}_{prior}) - \nabla \log \rho_t(\boldsymbol{\theta}_{prior})) \right) \end{cases} \tag{40}$$

The resulting simultaneous equations (40) are composed of: (a) the top equation, which is the particle discretization of the WGF that results from the minimization of the KL divergence between the posterior and the approximation to the posterior w.r.t. the approximation to the posterior; (b) the bottom equation that results from either minimizing or maximizing the KL divergence between the posterior and the approximation to the posterior w.r.t. the prior distribution. These simultaneous equations may be used to obtain the prior distribution that either maximizes or minimizes the functional and their resulting approximations to the posterior. For the case where the step size τ_t of the bottom equation in the simultaneous equation (40) is zero, the original particle-based WGF for Bayesian inference would be recovered, as this would mean the prior distribution is static (it does not change with time).

5.4. Density ratio estimation from samples

Equations (37) and (39) require the calculation of the pdf of the $\rho_t(\boldsymbol{\theta}_{prior})$ and the pdf of $p_t(\boldsymbol{\theta}_{prior})$. This may be done, for example, using KDE. In this article, rather than doing the direct estimation of the densities, the density ratio is calculated directly:

$$g(\boldsymbol{\theta}_{prior}) = \frac{\rho_t(\boldsymbol{\theta}_{prior})}{p_t(\boldsymbol{\theta}_{prior})} \tag{41}$$

Numerous methods have been developed for the calculation of the density ratio in equation (41); the method chosen in this article is the one called Relative unconstrained Least-Squares Importance Fitting (RuLSIF), and the interested reader can find it in (Yamada et al., 2011).

6. Data and numerical models

In this section, the proposed method is validated using two numerical examples. These two case studies have been selected to showcase the applicability of the proposed approach to deal with problems of different complexities, and an engineering case study is included. In the first example, the 2D double-banana posterior problem (Detommaso et al., 2018) is used to show the resulting particles obtained from the optimal and worst-case prior distributions and also the resulting particles from the approximation to the posterior. In the second example, a double-beam system is used to show the differences between the ensemble method and the Gaussian process to numerically estimate the gradient of the logarithm of the likelihood at the particle positions.

In both case studies, the number of initial samples is $N_0 = 100$ for the approximation to the posterior and also the prior distribution, and those initial samples are picked from identically and independently distributed draws from the nominal prior distribution. Each iteration of the algorithm uses the same number of particles ($N = 100$), and it corresponds to evaluations of the physics-based model at the positions of those particles. The choice of $N_0 = N = 100$ samples was validated both in terms of convenience and computational cost constraints. It was chosen in terms of convenience, as using a lower number of particles was found easier to explain some of the key results, and it also allowed to reduce the computational cost that would have been incurred by having a higher number of samples, as in each iteration of the algorithm the physics-based model would have had to be run at the particle locations. For more complex, higher-dimensionality problems, or where higher accuracy is required, a larger number of samples would typically be necessary. As described in Section 2, in the beginning of the method τ is set to zero until the number of iterations reaches $N_a = 50$.

The Gaussian kernel in equation (18), is used to produce the estimations of $\nabla \log \rho$ and $\nabla \log p(\boldsymbol{\theta})$, and the bandwidth is chosen using the median methodology.

If the distribution that is being optimized lies outside the ambiguity set, the size of the step in the particle flow algorithm is reduced to half until the distribution lies within the ambiguity set. Also, the distribution at iteration $i+1$ may be reset to a distribution from an earlier iteration $i - N_c$, where $N_c = 10$, if a preset number of distributions ($N_b = 5$) are discarded when determining whether a distribution belongs to the ambiguity set. The maximum number of prior distributions that are allowed to be reset is $N_{reset} = 2$, once this number is reached, an additional number of iterations N_a are allowed. The total maximum allowed number of iterations $N_{\max} = 400$.

6.1. Double-banana posterior example

This first example is based on the paper (Detommaso et al., 2018) that results in a 2D double-banana-shaped posterior distribution. The equation that defines the model used is given by the logarithmic Rosenbrock function used in Detommaso et al. (2018):

$$PM(\boldsymbol{\theta}) = \log\left((1-\theta_1)^2 + 100\left(\theta_2 - \theta_1^2\right)^2\right) \tag{42}$$

The initial prior distribution chosen is a standard multivariate Gaussian, $\mathcal{N}(0, I)$. The numerical observation used to update the prior knowledge is obtained by $y_{obs} = PM(\boldsymbol{\theta}_{true}) + \zeta$, where $\boldsymbol{\theta}_{true}$ is a random variable drawn from the assumed prior distribution, the standard deviation of the observational error is $\sigma = 0.3$, and $\zeta \sim \mathcal{N}(0, \sigma^2 I)$.

For the ambiguity set, the nominal prior distribution is chosen to be the same as the initial prior distribution. The statistical distance used is the 2-Wasserstein distance, and a radius $\varepsilon = 0.05$ has been chosen.

Using the algorithm inputs described above, the interacting WGFs are used to find the resulting distributions for two different cases: (a) the optimal prior distribution and its resulting approximation to the posterior; (b) the worst-case prior distribution and its resulting approximation to the posterior. In this example, the ensemble method described in Section 5.2.1 is used to calculate an approximation to the gradient of the log-likelihood at the particle positions to be evaluated.

The step sizes in the interacting particle flow WGF algorithm for the optimal prior distribution case are $\alpha = 3*10^{-3}$ and $\tau = 1.5*10^{-3}$. For the worst-case prior distribution case, the step sizes in the interacting particle flow WGF algorithm are $\alpha = 3*10^{-3}$ and $\tau = 3*10^{-4}$.

In this numerical case, two different subcases are run; Figures 7 to Figure 13 correspond to the situations when the optimal prior distribution and its approximation to the posterior are calculated. Figures 14 to 16 correspond to the situations when the worst-case prior distribution and its approximation to the posterior distribution are calculated.

In Figures 7 and 8, respectively, is shown, for each iteration, the positions of the particles from the optimal prior and from the approximation to the posterior distribution. In both plots of Figure 7, it may be

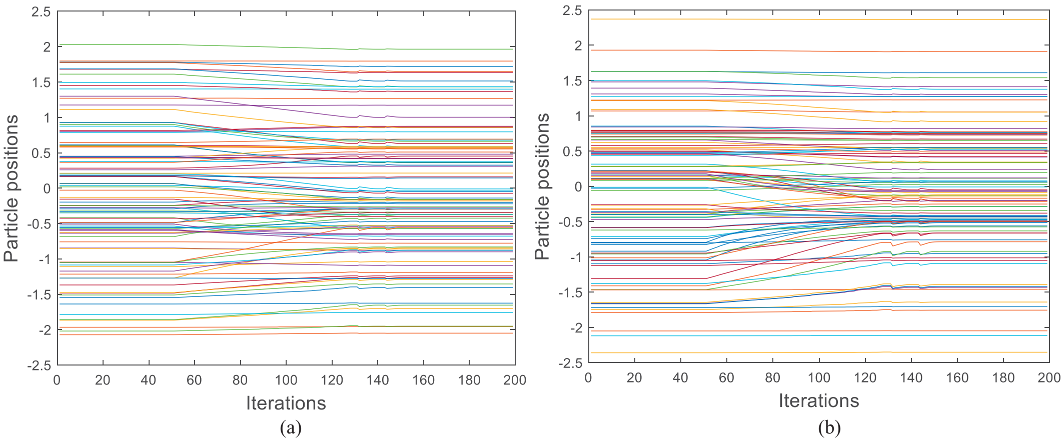


Figure 7. *Optimal prior particle positions at different iterations: (a) particle positions at θ_1; (b) particle positions at θ_2.*

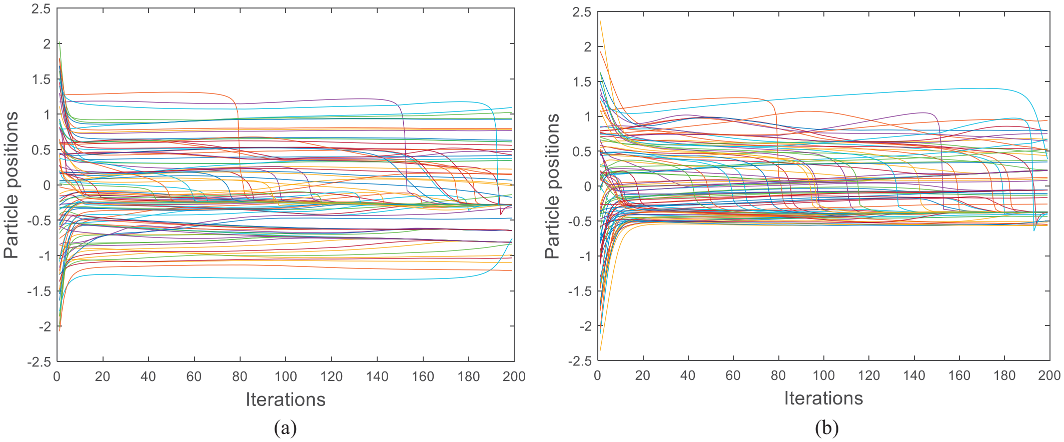
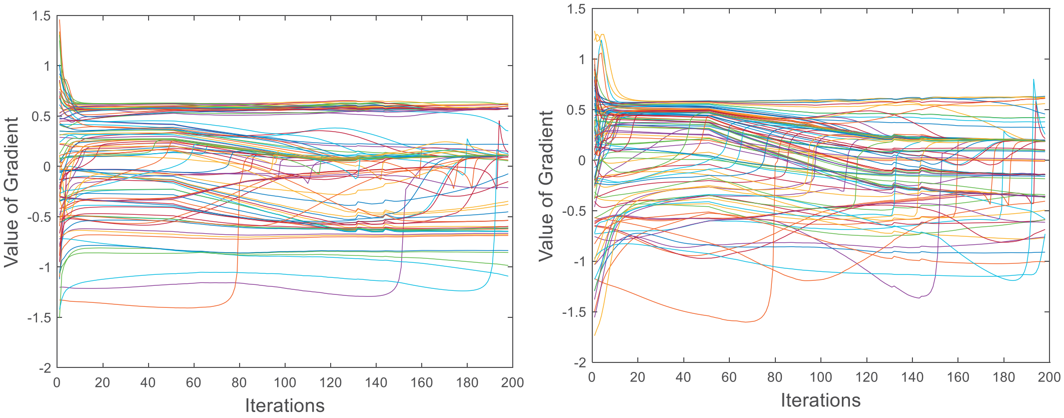


Figure 8. *Approximation to posterior particle positions at different iterations: (a) particle positions at θ_1; (b) particle positions at θ_2.*

seen that after iteration $i = N_a$, the inner particles of the prior distribution initially tend inwards, i.e., to the direction of smaller absolute values of parameters θ_1 and θ_2, this is due to the fact that the prior distribution tries to move the closest particles to the positions of the particles of the approximation to the posterior. It can also be seen that the prior particles after a number of approximately 150 iterations do not change much of position; this occurs because of the step size decrease performed with the purpose of constraining the prior distribution inside the defined ambiguity set. Figure 8 illustrates how the particle positions of the approximation to the posterior start moving into the regions of higher probability density. After iteration $i = N_a$, the particles' positions of the approximation to the posterior concentrate even more into regions of high probability density due to the prior distribution having a greater effect on the positions of the particles.

Figures 9 and 10 show, respectively, for each iteration, the values of the gradient of the logarithm of the prior distribution and the gradient of the logarithm of the likelihood at the particles Θ_i^N positions. Figure 9 shows how the values of the gradient of the logarithm of the prior distribution at the particle positions of the approximation to the posterior start to decrease as the prior particles start concentrating around the particles of the approximation to the posterior. As the iterations progress, the values of the gradient of the

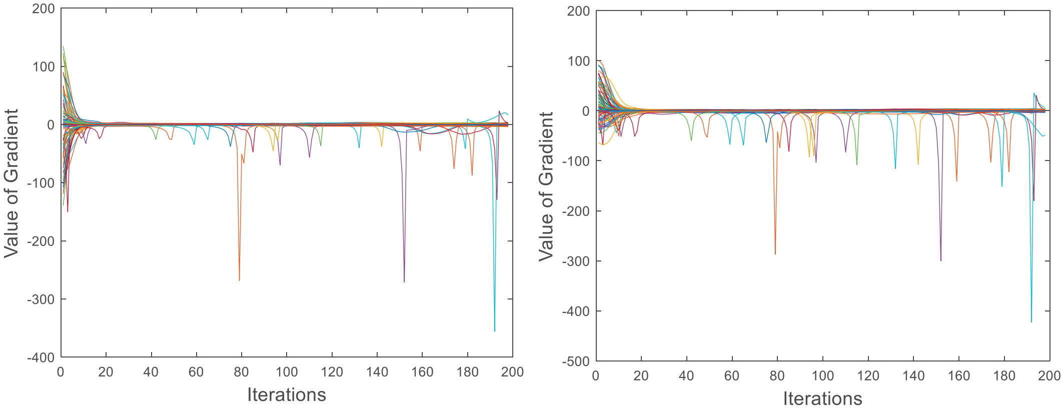
https://doi.org/10.1017/dce.2025.8 Published online by Cambridge University Press



(a)

(b)

Figure 9. *Gradient of log prior at different iterations and at particle positions Θ_i^N w.r.t.: (a) latent parameter θ_1; (b) latent parameter θ_2.*



(a)

(b)

Figure 10. *Gradient of log-likelihood at several iterations and at particle positions Θ_i^N w.r.t.: (a) latent parameter θ_1; (b) latent parameter θ_2.*

logarithm of the prior distribution at the approximation of the posterior particle positions start decreasing; this is because the particles of the prior become closer to the particles of the approximation to the posterior. This also means that the particles of the approximation to the posterior are becoming closer to regions of high prior density as the iterations progress. Figure 10 shows that during the initial iterations, high absolute values of the gradient of the logarithm of the likelihood may be found. This happens because during the initial stages of the algorithm, there are particles that are still distant from the regions of high likelihood density. After around 20 to 30 iterations, the values concentrate in a more defined region, even though some occasional extreme values can still be found.

In Figure 11, the initial particle positions (where the prior distribution and approximation to the posterior particles are the same, shown in red), the final particle positions of the prior (black), and the approximation to the posterior (blue) can be seen. As expected, the final positions of the particles from the approximation to the posterior are shown to resemble the double-banana posterior in Detommaso et al. (2018). It can also be observed that most of the final positions of the particles from the prior distribution (optimal prior distribution) are near the particles of the approximation to the posterior, and a smaller number of particles lie close to the initial prior particles. This means that the optimal prior assigns a high

https://doi.org/10.1017/dce.2025.8 Published online by Cambridge University Press

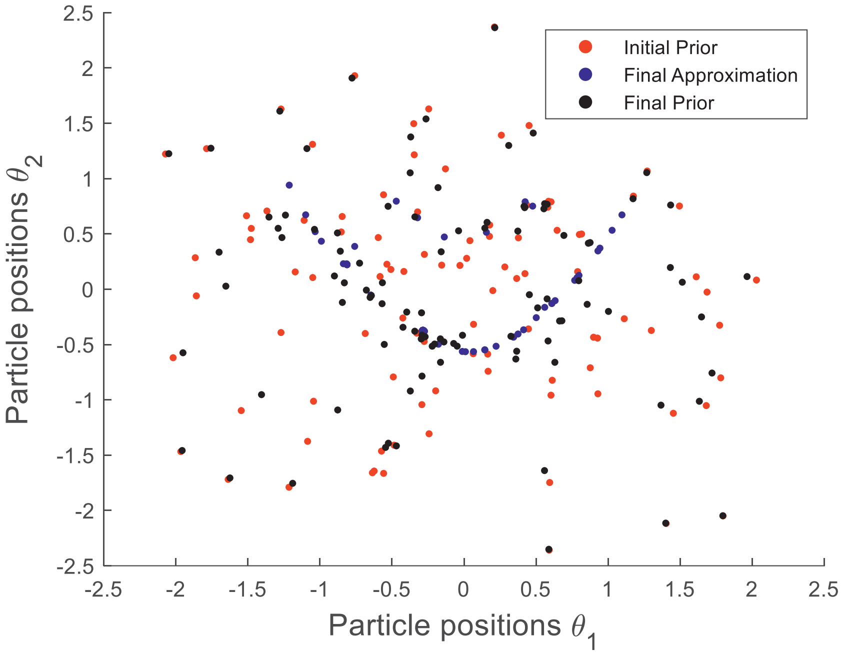


Figure 11. *Initial prior, final approximation to the posterior and final prior particle positions.*

probability to the region close to the approximation to the posterior and a lower probability to the outer particles far from the approximation to the posterior density.

Figure 12 shows a quiver plot, also known as a vector plot, that is produced by the generic function quiver in *MATLAB* (2022). The scaling of the quiver function's default setting is chosen to prevent arrow

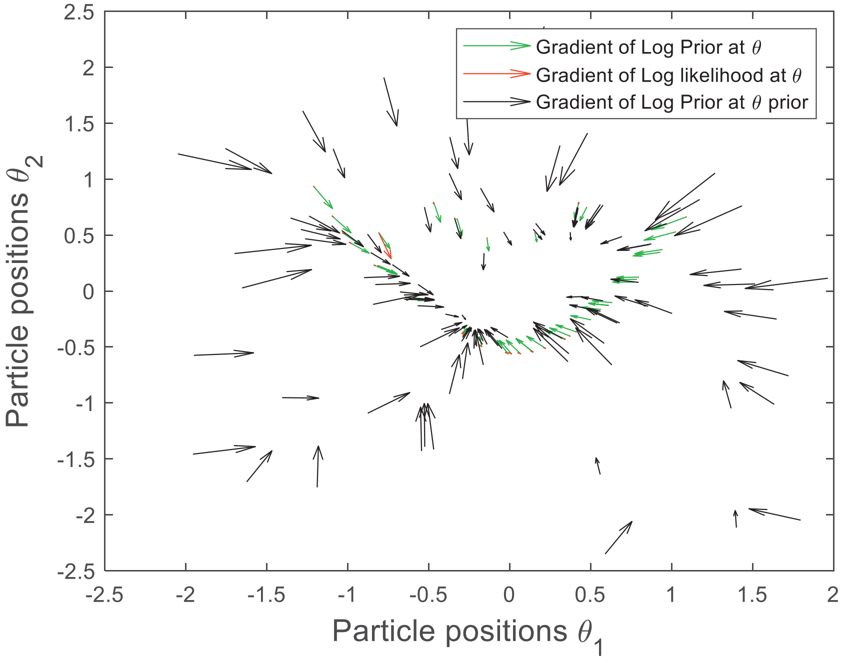


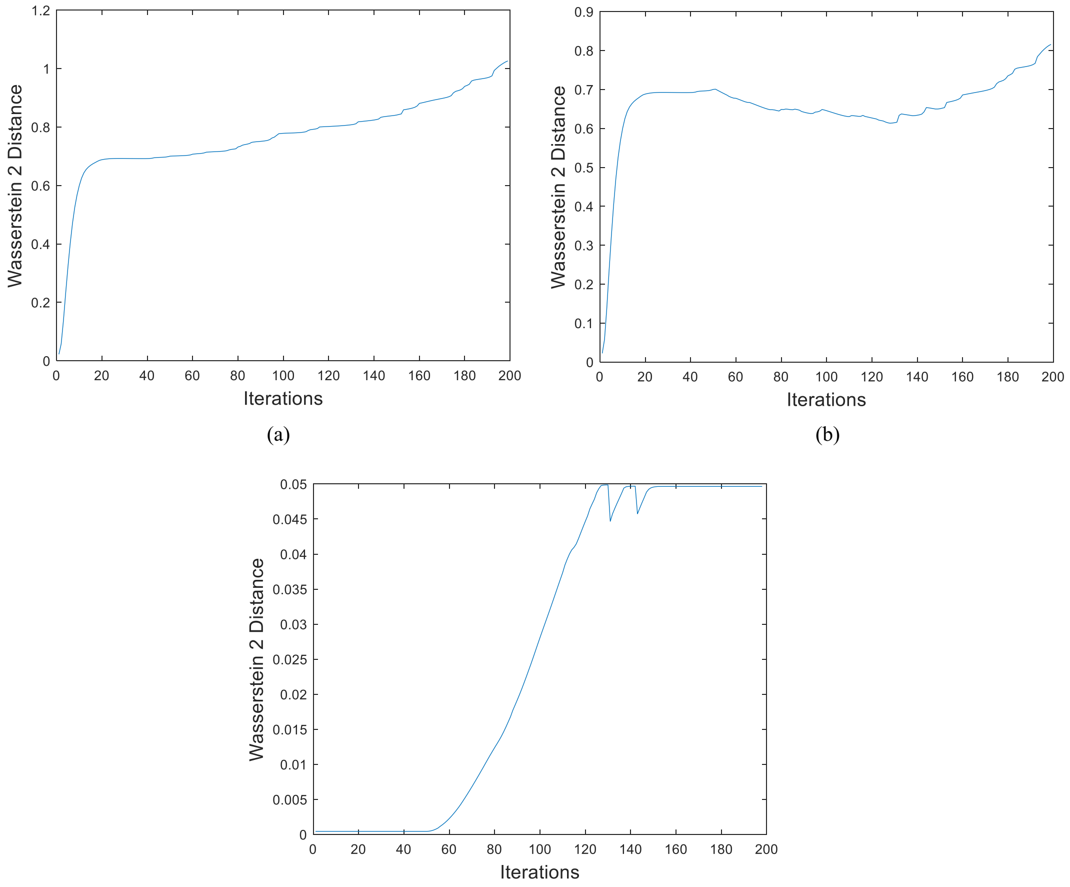
Figure 12. *Gradient/Quiver Plot of log prior and log-likelihood.*

https://doi.org/10.1017/dce.2025.8 Published online by Cambridge University Press

length overlap. In this plot, the gradients of the logarithm of the prior distribution and the logarithm of the likelihood at the particle's positions from the approximation to the posterior in the final iteration are plotted. The gradients of the logarithm of the prior distribution at the final prior particle positions are also shown.

Figure 13 illustrates the 2-Wasserstein distances at each iteration. Three plots can be found. The following distances at each iteration are plotted: the first is from the initial prior distribution to the approximation of the posterior distribution; the second is from the approximation to the posterior distribution and the prior distribution; and the third is from the initial prior to the prior distributions.

In Figures 14 and 15, respectively, the positions of the particles from the worst-case prior distribution and from the approximation to the posterior are shown for each iteration. In Figure 14 (a and b), it may be seen that the inner particles of the worst-case prior distribution tend to move outwards to the direction of higher absolute values of parameters θ_1 and θ_2 as more iterations occur. Figure 14 also illustrates that after approximately 200 iterations, the prior particles do not change much in their positions. This is due to the decreasing size of the time step that is introduced with the purpose of limiting the prior distribution inside the ambiguity set. In a similar manner to what occurs for the optimal prior distribution case, in Figure 15, it can be observed that the particle positions of the approximation to the posterior also move to areas of higher probability density as iterations advance.



(a)

(b)

(c)

Figure 13. *2-Wasserstein distance at different iterations i between: (a) initial prior distribution and approximation to posterior distribution; (b) approximation to posterior and prior distributions; and (c) initial prior and prior distributions.*

https://doi.org/10.1017/dce.2025.8 Published online by Cambridge University Press

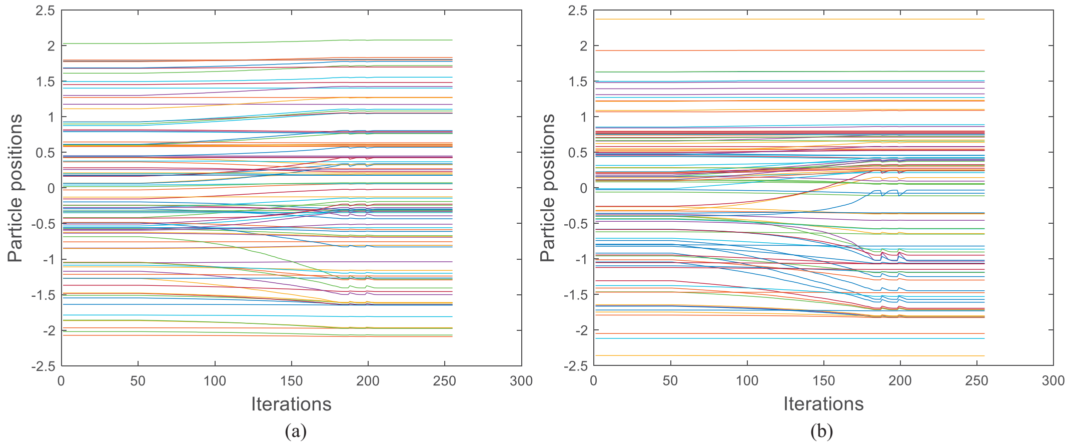


Figure 14. *Worst-case prior particle positions at different iterations: (a) particle positions at θ_1; (b) particle positions at θ_2.*

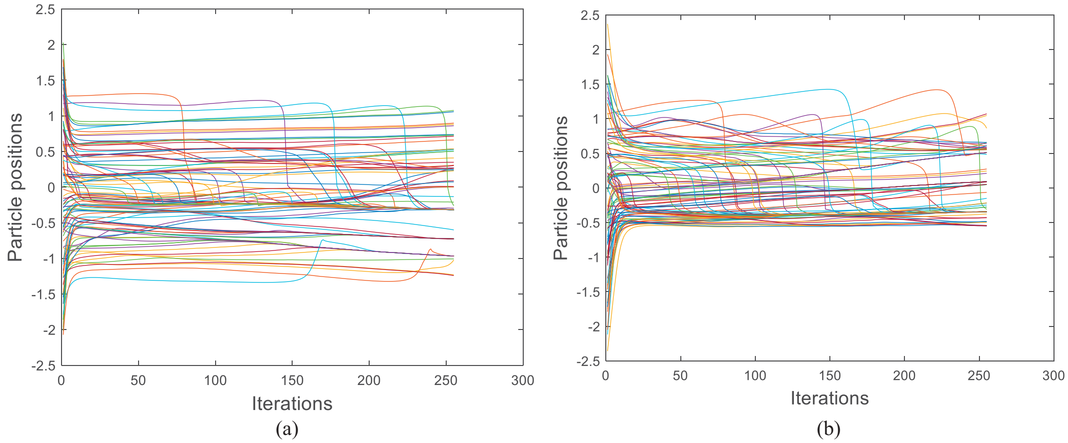


Figure 15. *Approximation to posterior particle positions at different iterations: (a) particle positions at θ_1; (b) particle positions at θ_2.*

The final particle positions of the worst-case prior distribution (black), initial particle positions (where the prior distribution and approximation to the posterior particles are the same, shown in red), and its approximation to the posterior (blue) can be seen in Figure 16. As anticipated, the layout of the final positions of the particles from the approximation to the posterior takes a shape similar to the one shown by the double-banana posterior in (Detommaso et al., 2018). The worst-case prior distribution assigns a higher density to areas of a low posterior density and vice versa. In a manner consistent with the previous statement, Figure 16 also shows that most of the final positions of the particles from the prior distribution (worst-case prior distribution) are positioned away from the final positions of the approximation to the posterior distribution.

A direct comparison of the optimal prior and worst-case prior distributions in the form of scatter plots and histograms of the latent variables is found in Figure 17. Figure 17 has been produced using the `plotmatrix` function from *MATLAB* (2022). It can be clearly seen that the optimal and worst-case prior distributions differ from the initial prior distribution and are no longer a Gaussian distribution. A very similar support can be seen of the optimal prior distribution w.r.t. the worst-case prior distribution.

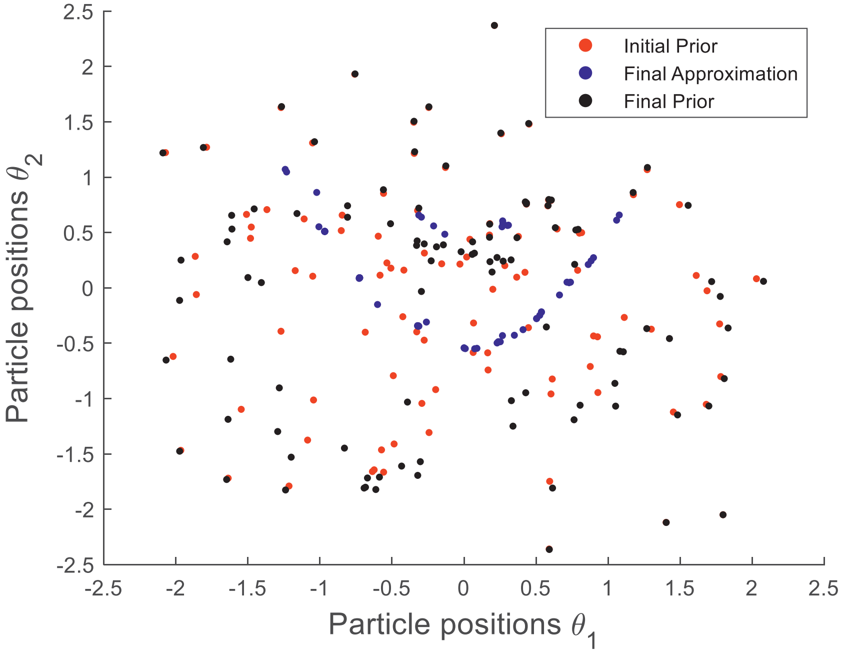


Figure 16. *Initial prior, final approximation to the posterior and final worst-case prior particle positions.*

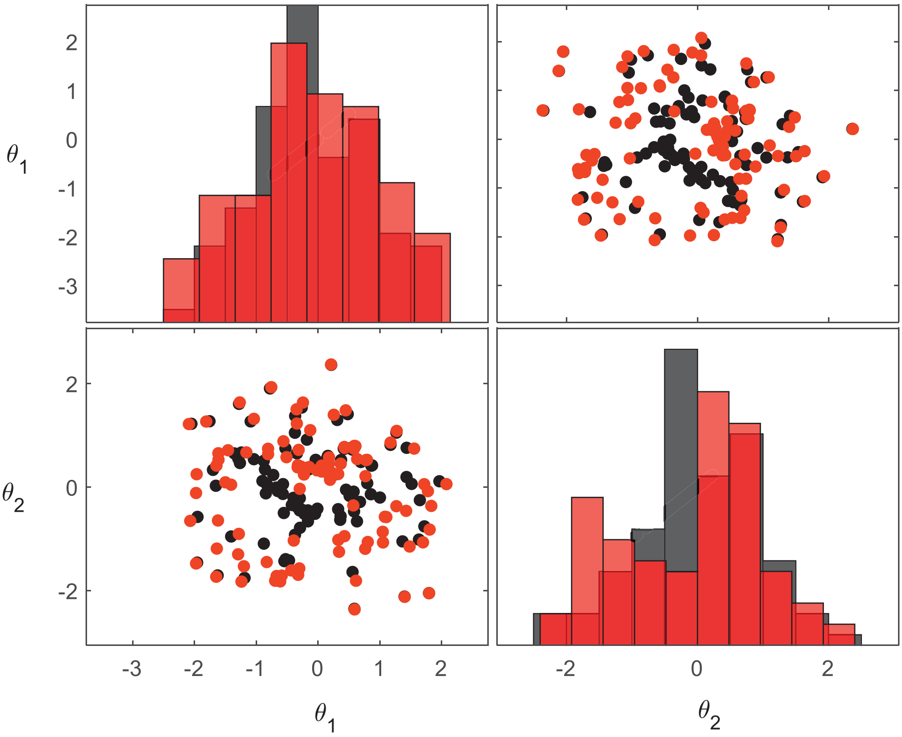


Figure 17. *Scatterplots and histograms show the prior distribution, black—optimal prior distribution case, and red—worst-case prior distribution.*

https://doi.org/10.1017/dce.2025.8 Published online by Cambridge University Press

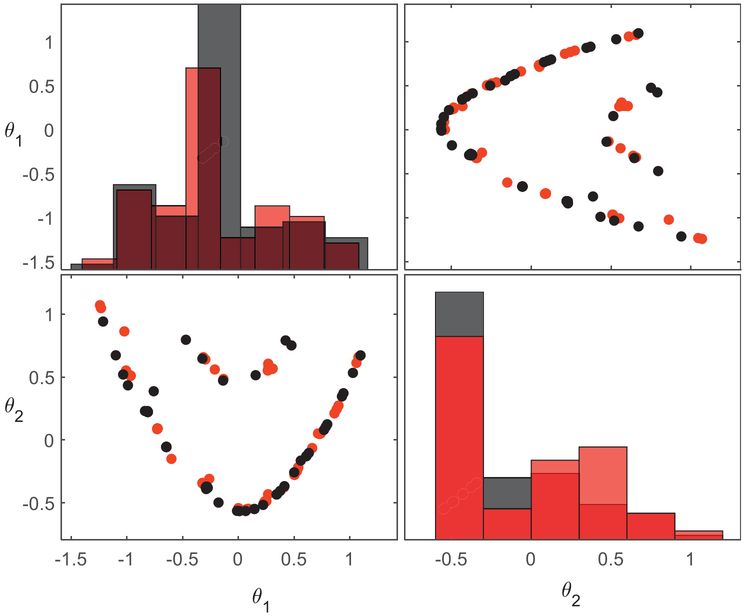


Figure 18. *Scatterplots and histograms show the approximation to the posterior distribution, black—optimal prior distribution case, and red—worst-case prior distribution.*

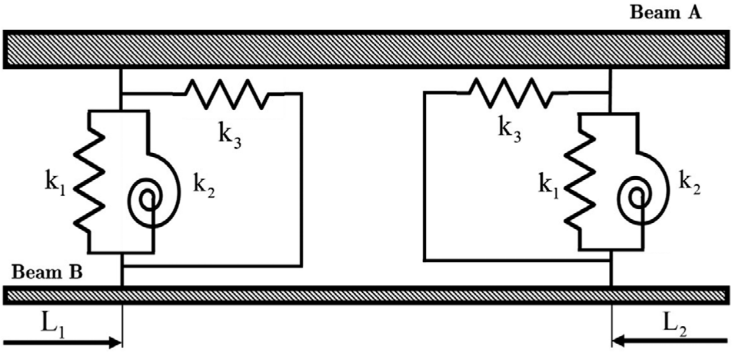


Figure 19. *Theoretical model of a coupled beam structure.*

Table 1. *Coupled beam dimensions, distances from edges to connections, and mechanical characteristics*

	Thickness	Width	Length	L_1	L_2	Young's modulus	Density
			[mm]			[GPa]	[Kg/m^3]
Beam A	6	25	600	20	20	210	7800
Beam B	3						
Springs	k_1		k_3			k_2	
	[MN/m]					[Nm/rad]	
	100		10			500	

https://doi.org/10.1017/dce.2025.8 Published online by Cambridge University Press

Table 2. *Coupled beam structure natural frequencies [Hz]*

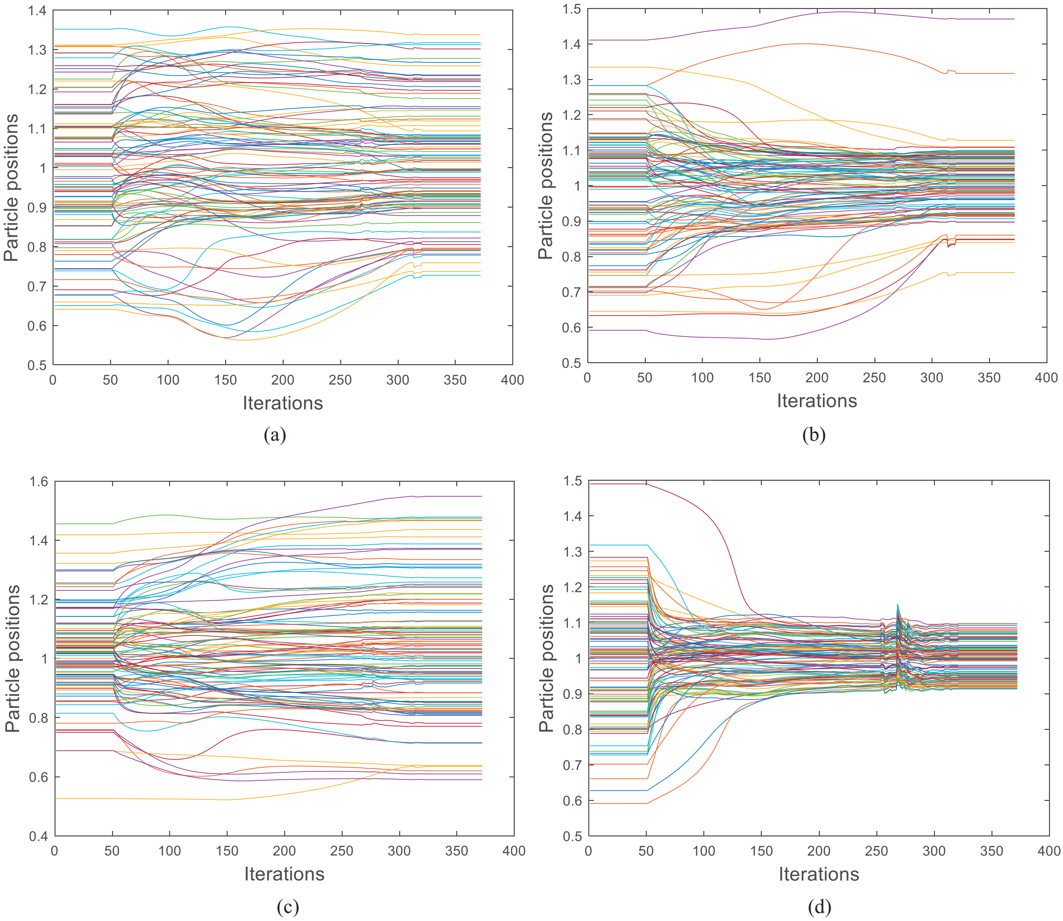
f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
16.0	50.2	92.8	134.6	245.3	260.7	428.0	478.6

Scatter plots and histograms can also be found in Figure 18. For the cases where the optimal prior and worst-case prior distributions have been estimated, the scatter plots and the histograms of the latent variables of the resulting approximation to the posterior are plotted. Very small differences are found when comparing the resulting approximations to the posterior distributions. This is a consequence of the small sensitivity of the posterior distribution to changes of the considered uncertain prior distribution.

6.2. Double-beam structure example

The model used in this second example is based on the coupled beam structure illustrated in Igea (2023).

The structure is shown in Figure 19, two connecting fixtures composed of three springs each: one translational, one shear, and one rotational that link two beams. This example shows practical interest, as it can be used to depict structural conditions where the attaching ensembles between elements show



(a)

(b)

(c)

(d)

Figure 20. *Optimal prior particle positions at different iterations for different latent parameters: (a) θ_1; (b) θ_2; (c) θ_3; and (d) θ_4.*

https://doi.org/10.1017/dce.2025.8 Published online by Cambridge University Press

uncertainty. The causes of such uncertainty can be derived from boundary conditions and manufacturing variability. More specifically, the four uncertain parameters chosen are the spring stiffnesses and the Young's modulus of both beams: the rotational springs $k_2 = 500\theta_1$ [Nm/rad], the shear springs $k_3 = 10^7\theta_2$ [N/m], the translational springs $k_1 = 10^{10}\theta_3$ [N/m], and the Young's modulus of both beams $E_1 = E_2 = 210*10^9\theta_3$ [Pa]. For those four uncertain parameters, the initial prior distribution is a multivariate Gaussian prior distribution chosen as $\mathcal{N}(I, 0.03I)$.

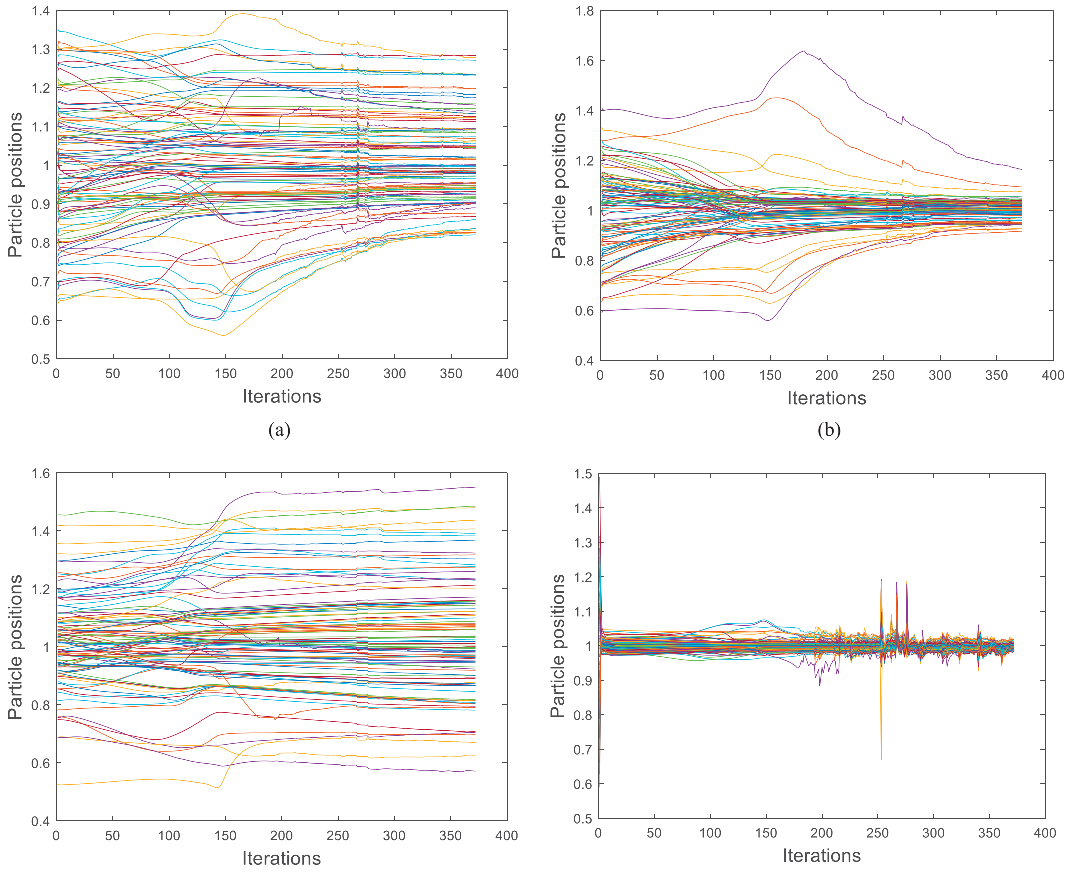
Dimensions and mechanical characteristics of the double-beam model may be found in Table 1.

Using the data on Table 1, the first eight natural frequencies of the model were assessed and introduced in Table 2.

The numerical frequencies obtained in Table 2 were produced using a Finite Element (FE) code. The code assumes a 2D Euler-Bernoulli beam model. Uniform discretization with two hundred FEs for each beam was used. Each FE has two nodes, and each node has two degrees of freedom.

The likelihood function is assumed to be a multivariate Gaussian distribution; the mean is given by the deterministic value of the eight natural frequencies in Table 2, and the covariance is assumed to be a diagonal covariance matrix that has standard deviations of 2% of their deterministic values ($\sigma_i = 0.02f_i$).

In this example, for the definition of the ambiguity set, the statistical distance used is also the 2-Wassertein distance, where the radius is $\varepsilon = 0.04$, and a nominal prior distribution equal to the initial prior distribution is selected.



(a)

(b)

(c)

(d)

Figure 21. *Approximation to posterior particle positions at different iterations for different latent parameters: (a) θ_1; (b) θ_2; (c) θ_3; and (d) θ_4.*

https://doi.org/10.1017/dce.2025.8 Published online by Cambridge University Press

The values described above are used as inputs of the algorithm, and the interacting WGFs are used to find the resulting distributions for two different cases: (a) the optimal prior distribution and its resulting approximation to the posterior; (b) the worst-case prior distribution and its resulting approximation to the posterior. In this example, the Gaussian process method described in Section 5.2.2 is used to calculate an approximation to the gradient of the log-likelihood at the particle positions evaluated.

The values of step size used in the interacting particle flow WGF algorithm for the optimal prior distribution case are $\alpha = 5*10^{-5}$ and $\tau = 2.5*10^{-3}$. The values used for the worst-case prior distribution case are $\alpha = 5*10^{-5}$ and $\tau = 5*10^{-5}$.

Figures 20 and 21, respectively, illustrate the positions of the particles from the optimal prior and from the approximation to the posterior for each iteration. In Figure 20, after iteration $i = N_a$, it can be seen that for θ_1, θ_2, and θ_4, the prior particle positions start concentrating at values close to one. It can also be seen that θ_4 has the most rapid change out of all the latent variables; this is probably due to being the latent variable, which most affects the model output. However, the opposite effect can be observed for θ_3, this is most likely due to the low sensitivity of the model output to changes of the latent variable θ_3. Figure 21 shows how the particles of the approximation to the posterior also concentrate to values closer to one as the number of iterations progresses for all the latent variables except for θ_3.

Figure 22 shows scatter plots and a histogram produced by the `plotmatrix` function of *MATLAB* (2022), of the initial particles from the prior distribution approximation to the posterior (red), the final particles from the optimal prior distribution (black), and the final particles from the approximation to the posterior distribution (blue). It can be seen that the particle positions from the optimal prior distribution and the approximation to the posterior distribution are quite similar for all latent variables except for θ_3. From the histogram, it can be also seen that for the latent variable θ_3, the optimal prior distribution has a bigger support than the initial prior distribution.

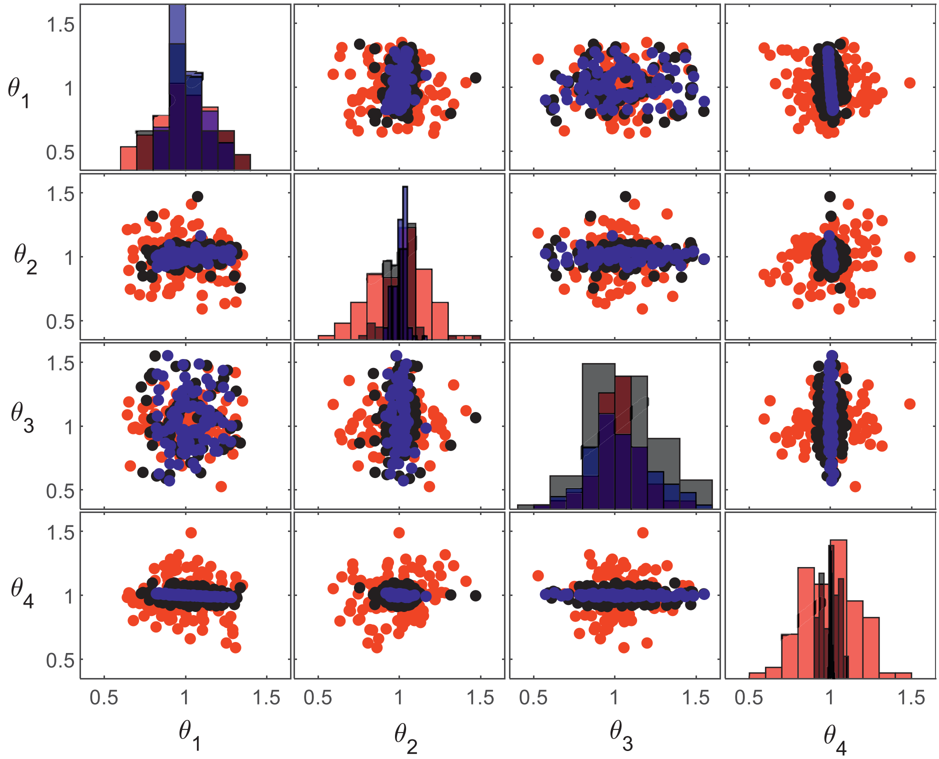
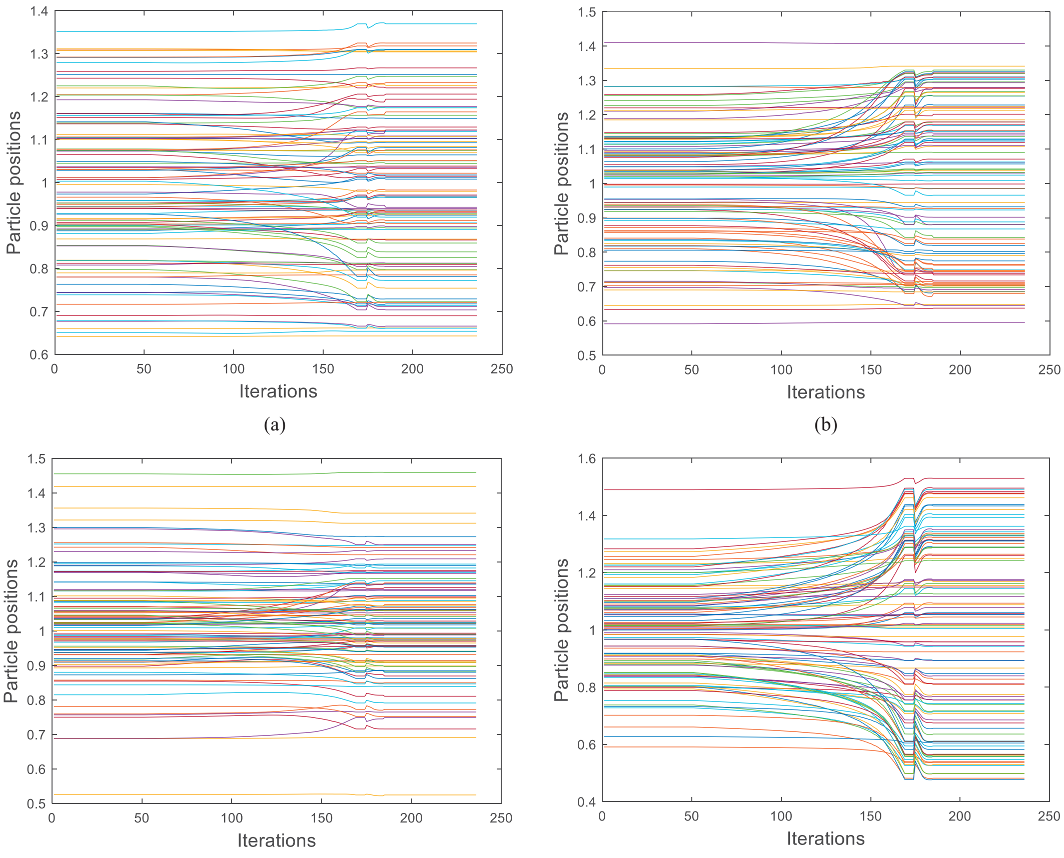


Figure 22. *Scatterplots and histograms show: red—initial prior distribution; blue—final approximation to the posterior distribution; and black—optimal prior distribution.*

The positions of the particles from the worst-case prior distribution and from the approximation to the posterior for each iteration are shown in Figures 23 and 24, respectively. Figure 23 shows that after iteration $i = N_a$ for θ_1, θ_2, and θ_4, the prior particle positions part from values close to one. This is the opposite of what occurs for the optimal prior distribution case. In a manner similar to what happens for the optimal case, the latent variable θ_4 experiments the fastest change of all the uncertain parameters. This is most likely due to the higher sensitivity of the model output to the changes of this latent variable. Figure 24 illustrates how, as the number of iterations progresses, the particles of the approximation to the posterior depart from values close to one. However, in this case, the change in the positions of the particles of the approximation to the posterior is not as significant as in the case for the optimal prior distribution.

Histograms and scatter plots produced by the `plotmatrix` function of *MATLAB* (2022), can be found in Figure 25. The graphs illustrate the positions of the particles. In blue, the final particles from approximation to the posterior. In black, the final particles from the optimal prior distribution. In red, the initial particles from the prior distribution approximation to the posterior distribution. From the histograms, it can be deduced that in the worst-case prior distribution, the supports of the graphs of all latent variables are bigger compared to the ones of the initial prior distribution. The scatterplots show that worst-case prior particles have moved in such a manner that most of their particles lie in regions of lower posterior density.

Figure 26 directly compares the optimal prior and worst-case prior distributions using the `plotmatrix` function from *MATLAB* (2022) by plotting scatter plots and the histograms of the latent variables. In general,



(a)

(b)

(c)

(d)

Figure 23. *Worst-case prior particle positions at different iterations for different latent parameters: (a) θ_1; (b) θ_2; (c) θ_3; and (d) θ_4.*

https://doi.org/10.1017/dce.2025.8 Published online by Cambridge University Press

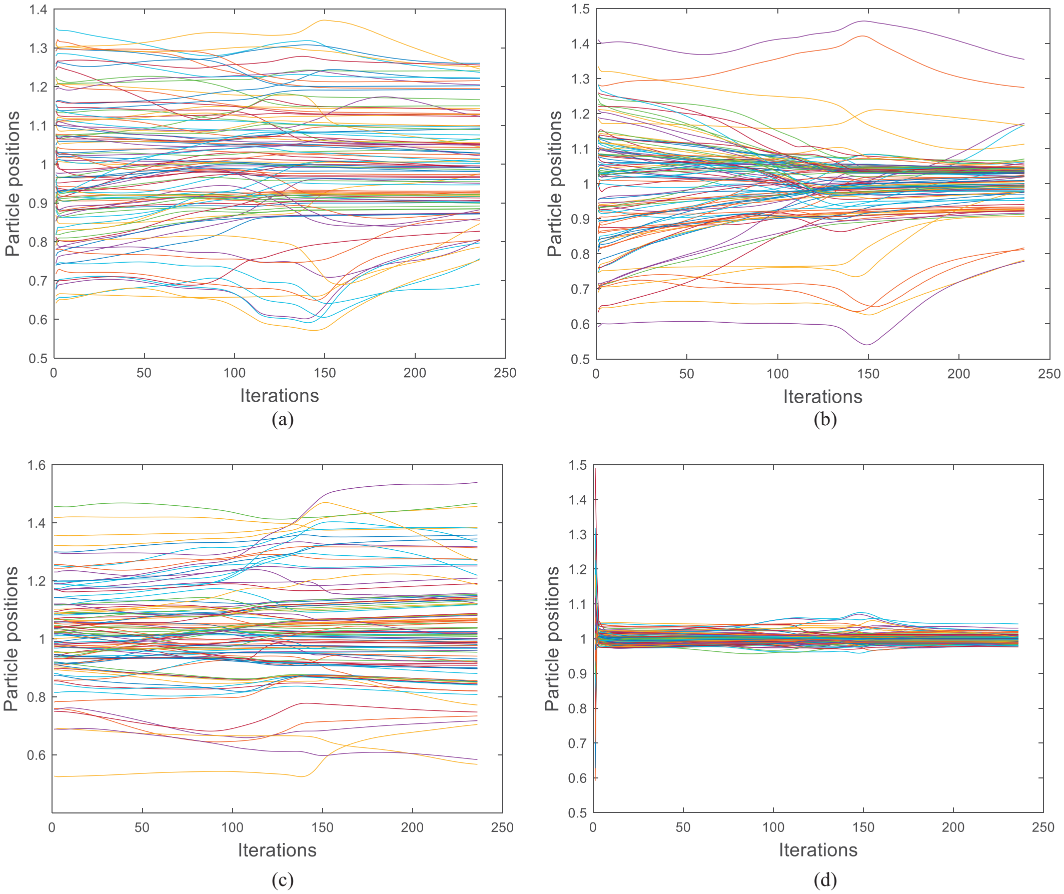


Figure 24. *Approximation to posterior particle positions at different iterations for different latent parameters: (a) θ_1; (b) θ_2; (c) θ_3; and (d) θ_4.*

for most of the latent variables, it can be seen that the support of the worst-case prior distribution is bigger than the optimal prior distribution.

Figure 27 also has scatter plots and the histograms of the latent variables of the resulting approximation to the posterior distribution when the optimal prior distribution and worst-case prior distribution have been calculated. When comparing the resulting approximations to the posterior distributions, it can be seen that for the case with the optimal prior distribution, the resulting approximation to the posterior distribution is more concentrated compared to the approximation to the posterior distribution that results from the worst-case prior distribution. In this example, it is seen that the posterior distribution is slightly sensitive to the considered uncertain prior distribution.

7. Conclusion

BMU is widely used in engineering applications for evaluating the latent parameters' posterior distribution of a physics-based model, given informative measurement data. These updated models are used to investigate the structure's dynamic behavior under various operational and environmental loading and are critical, for example, in maintenance planning, reliability analysis, and remaining useful life estimations. Therefore, the results obtained with BMU are critical in decision-making in engineering. However, they can be sensitive to: the prior distribution assumptions, especially in the presence of limited data. In this article, a robust Bayesian inference approach, based on WGFs has been proposed. This approach yields an

https://doi.org/10.1017/dce.2025.8 Published online by Cambridge University Press

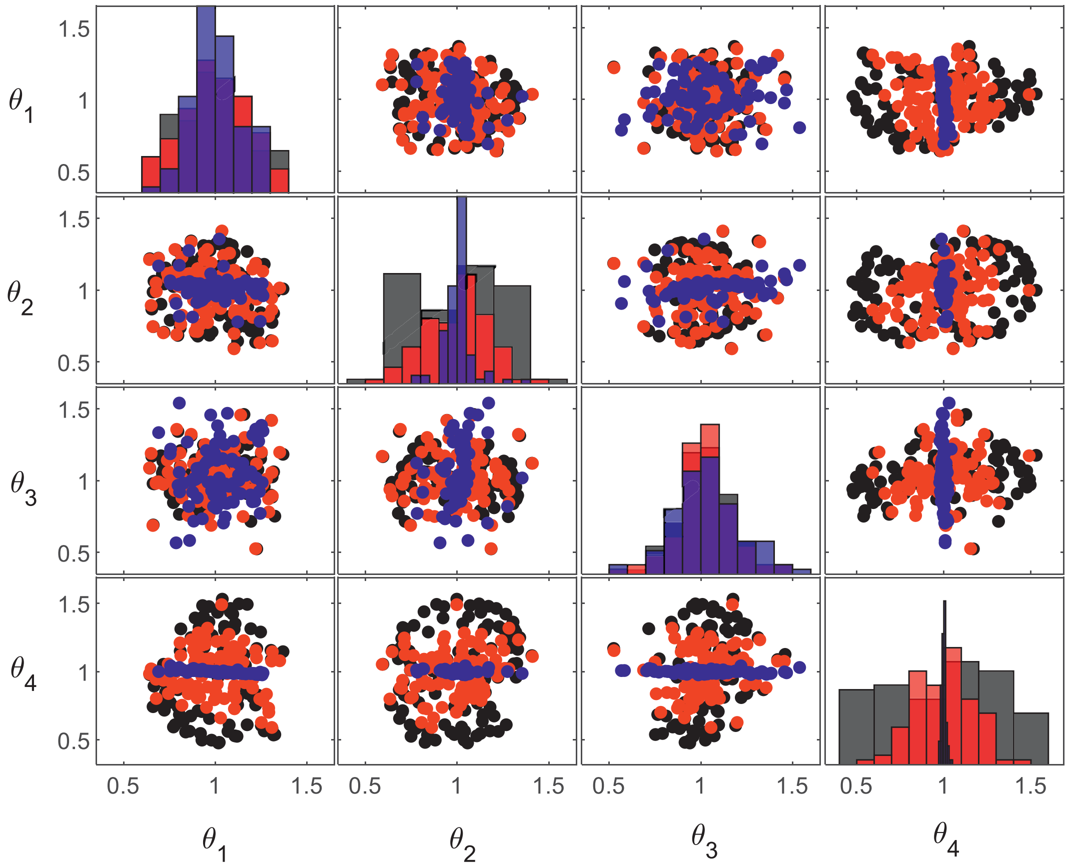


Figure 25. *Scatterplots and histograms show: red—initial prior distribution; blue—final approximation to the posterior distribution; and black—worst-case prior distribution.*

estimation of the posterior distribution of the latent parameters by finding the optimal and worst-case prior distributions. This estimation is produced by an algorithm that combines an interacting WGF formulation with an ambiguity set. The ambiguity set is defined by a nominal distribution, a statistical distance, and a radius. In this article, the 2-Wasserstein distance is used as the statistical distance. Due to the properties of the 2-Wasserstein distance, the distributions that lie inside the prescribed radius do not need to have the same support. The ambiguity set may be used to explore the sensitivity of the posterior distribution prediction of the system to uncertainty in the prior distribution. This application may be of particular interest for cases where the opinions of different experts are conflicting. The approximations to the posterior distribution found with the proposed approach can be used as lower and upper bounds on subsequent metric calculations used for decision-making. These bounds on the resulting metric can be readily used in decision-making to assess if the decisions taken are robust to prior uncertainty or otherwise.

The interacting WGF formulation is derived from first principles, obtaining particle discretization equations for the calculation of the optimal and worst-case prior distributions. The derivation of the interacting WGFs allows the development of the proposed method, which may reduce the computational cost incurred if all the possible prior distributions that lie inside the ambiguity set were to be tested directly. A KDE is used to obtain estimates of the gradient of the logarithm of the prior distribution and of the gradient of the logarithm of the approximations to the posterior distribution with respect to the particle positions.

https://doi.org/10.1017/dce.2025.8 Published online by Cambridge University Press

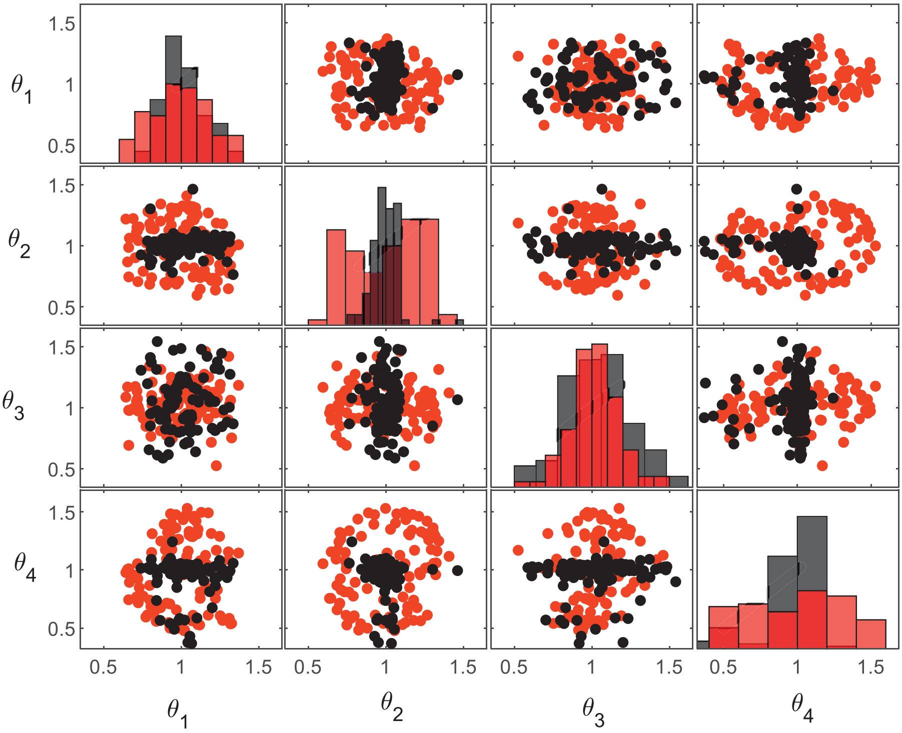


Figure 26. *Scatterplots and histograms show the prior distribution; black—optimal prior distribution case; and red—worst-case prior distribution.*

The article illustrates how the gradient of the logarithm of the likelihood may be estimated either using an ensemble method or a Gaussian process regression method. Two numerical examples have been used to show both the optimal and worst-case prior distributions and their resulting approximation to the posterior distribution. In these examples, it is shown that for the optimal prior distribution case, the particles' positions tend to be near the particles of the approximation to the posterior distribution, this means the optimal prior distribution assigns a higher prior density close to regions of high posterior density. For the worst-case prior distribution, the opposite behavior may be seen; the particles tend to move to positions far from the particles of the approximation to the posterior density. As a consequence, the worst-case prior distribution has a bigger support than the initial prior distribution. The proposed approach is general, and it may be relevant for application areas outside decision-making in engineering.

In the numerical studies, the choice of 100 samples was validated both in terms of convenience and computational cost constraints. It was chosen in terms of convenience, as using a lower number of particles, it was found easier to explain some of the key results and figures shown throughout the article. This also allowed to reduce the computational cost that would have been incurred by having a higher number of samples, as in each iteration of the algorithm, the physics-based model would have had to be run at the particle locations. Currently, both the effects of the KDE and GP approximations are assumed to be insignificant on the approximations of the gradient of the logarithm of the prior and likelihood. However, it should also be noted that the dimensionality of the posterior has a big effect on these approximations because kernel functions are affected by it. For more complex, higher-dimensionality problems, or where higher accuracy is required, a larger number of samples would typically be necessary. Future work may focus on the convergence properties of the proposed approach, increasing the sample size and the dynamic selection of step sizes, the latter would allow the proposed approach to become more computationally efficient. The method would also benefit from the development of a sample-efficient strategy, in which the reuse of samples from previous iterations may be integrated into the proposed methodology, reducing the number of simulations further. Another potential direction of interest is the

https://doi.org/10.1017/dce.2025.8 Published online by Cambridge University Press

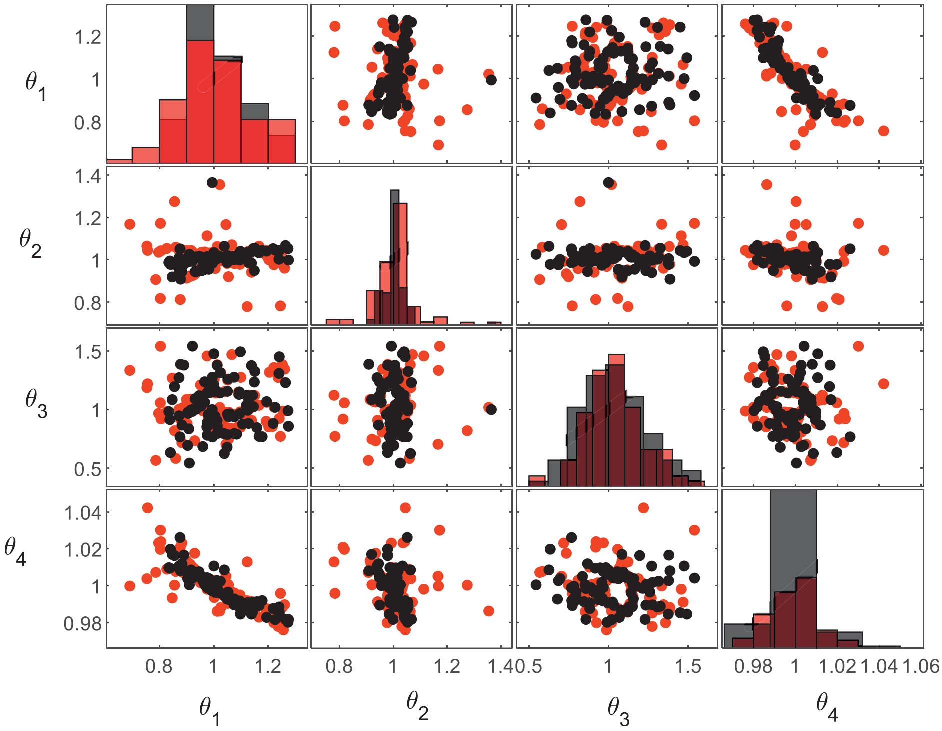


Figure 27. *Scatterplots and histograms show the approximation to the posterior distribution; black—optimal prior distribution case; and red—worst-case prior distribution.*

development of a principled approach for the selection of the nominal prior distribution and its radius, as at this stage it is assumed to be known. These topics are currently under investigation.

Data availability statement. The data that support the findings of this study will be made available upon request.

Acknowledgments. We are grateful to Youssef Marzouk for his invaluable insights and thought-provoking discussions, which were instrumental in guiding us toward this research direction.

Author contribution. Conceptualization: F.I. and A.C.; Formal analysis: F.I.; Methodology: F.I. and A.C.; Supervision: A.C.; Writing—original draft: F.I. and A.C; and Writing—review and editing: F.I. and A.C. All authors approved the final submitted draft.

Funding statement. Felipe Igea and Alice Cicirello thank the EPSRC and Schlumberger for an Industrial Case postgraduate scholarship (Grant ref. EP/T517653/1). Alice Cicirello gratefully acknowledges the financial support provided by the Alexander von Humboldt Foundation Research Fellowship for Experienced Researchers supporting part of this research.

Competing interest. None declared.

References

Acerbi L (2018) Variational Bayesian Monte Carlo. *Advances in Neural Information Processing Systems (NeurIPS 2018) 31*, 8223–8233. http://arxiv.org/abs/1810.05558.

Alvarez-Melis D, **Schiff Y and Mroueh Y** (2021) Optimizing Functionals on the Space of Probabilities with Input Convex Neural Networks. https://arxiv.org/abs/2106.00774v3.

Ambrosio L, **Gigli N and Savaré G** (2005) *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Basel Boston Berlin: Birkhäuser.

Bayraksan G and Love DK (2015) Data-driven stochastic programming using phi-divergences. *INFORMS Tutorials in Operations Research*, 1–19. https://doi.org/10.1287/EDUC.2015.0134.

Beck JL and Katafygiotis LS (1998) Updating models and their uncertainties. I: Bayesian statistical framework, *Journal of Engineering Mechanics 124*(4), 455–456.

https://doi.org/10.1017/dce.2025.8 Published online by Cambridge University Press

Berger JO, **Moreno E**, **Pericchi LR**, **Bayarri MJ**, **Bernardo JM**, **Cano JA**, **De la Horra J**, **Martín J**, **Ríos-Insúa D**, **Betrò B and Dasgupta A** (1994) An overview of robust Bayesian analysis. *TEST 3*(1), 5–124.

Blei, D. M., **Kucukelbir, A.**, & **McAuliffe, J. D.** (2017). Variational inference: A review for statisticians. In *Journal of the American Statistical Association 112*(518), 859–877). https://doi.org/10.1080/01621459.2017.1285773

Campbell T and Li X (2019) Universal Boosting Variational Inference. *ArXiv.* http://arxiv.org/abs/1906.01235

Chen C, **Zhang R**, **Wang W**, **Li B and Chen L** (2018) A unified particle-optimization framework for scalable Bayesian sampling. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, vol. 2, pp. 746–755. https://arxiv.org/abs/1805.11659v2.

Chen Y, **Huang DZ**, **Huang J**, **Reich S and Stuart AM** (2023) Gradient Flows for Sampling: Mean-Field Models, Gaussian Approximations and Affine Invariance. https://arxiv.org/abs/2302.11024v3

Cheng Z, **Zhang S**, **Yu L and Zhang C** (2023) Particle-based variational inference with generalized Wasserstein gradient flow. *Advances in Neural Information Processing Systems*, *36*. https://arxiv.org/abs/2310.16516v1.

Chérief-Abdellatif B-E and Alquier P (2019) MMD-Bayes: Robust Bayesian Estimation via Maximum Mean Discrepancy. https://arxiv.org/abs/1909.13339v2

Chewi, S., **Le Gouic, T.**, **Lu, C.**, **Maunu, T.**, & **Rigollet, P.** (2020). SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence. *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*. Article no. 177, 2098–2109 https://arxiv.org/abs/2006.02509v1.

Chizat L and Bach F (2018) On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in Neural Information Processing Systems*, 3036–3046. https://arxiv.org/abs/1805.09545v2

Detommaso G, **Cui T**, **Spantini A**, **Marzouk Y and Scheichl R** (2018) A Stein variational Newton method. *Advances in Neural Information Processing Systems*, 9169–9179. https://arxiv.org/abs/1806.03085v2

Ding S, **Dong H**, **Fang C**, **Lin Z and Zhang T** (2023) Provable Particle-based Primal-Dual Algorithm for Mixed Nash Equilibrium. https://arxiv.org/abs/2303.00970v1

Dunbar, O. R. A., **Duncan, A. B.**, **Stuart, A. M.**, & **Wolfram, M. T.** (2022). Ensemble inference methods for models with noisy and expensive likelihoods. *SIAM Journal on Applied Dynamical Systems*, *21*(2), 1539–1572. https://doi.org/10.1137/21M1410853.

Ebrahimian H, **Astroza R**, **Conte JP and de Callafon RA** (2017) Nonlinear finite element model updating for damage identification of civil structures using batch Bayesian estimation. *Mechanical Systems and Signal Processing 84*, 194–222.

Fan J, **Zhang Q**, **Taghvaei A and Chen Y** (2021) Variational Wasserstein gradient flow. *Proceedings of Machine Learning Research 162*, 6185–6215. https://arxiv.org/abs/2112.02424v3.

Farrar CR and Worden K (2013) *Structural Health Monitoring : A Machine Learning Perspective*. West Sussex, UK: Wiley.

Gao Y and Liu J-G (2020) A note on parametric Bayesian inference via gradient flows. *Annals of Mathematical Sciences and Applications 5*(2), 261–282.

Ghosh A and Basu A (2016) Robust bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics 68*(2), 413–437. https://doi.org/10.1007/S10463-014-0499-0/FIGURES/10.

Go J and Isaac T (2022) Robust expected information gain for optimal Bayesian experimental design using ambiguity sets. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence, UAI 2022*, pp. 728–737. https://arxiv.org/abs/2205.09914v1.

Green PL and Worden K (2015) Bayesian and Markov chain Monte Carlo methods for identifying nonlinear systems in the presence of uncertainty. *Philosophical Transactions of the Royal Society A 373*(2051), 20140405. https://doi.org/10.1098/rsta.2014.0405.

Hooker G and Vidyashankar AN (2011) Bayesian model robustness via disparities. *TEST 23*(3), 556–584. https://doi.org/10.1007/s11749-014-0360-z.

Igea F (2023) Probabilistic uncertainty quantification for structural dynamics under limited data [PhD thesis]. University of Oxford.

Igea F and Cicirello A (2023) Cyclical variational Bayes Monte Carlo for efficient multi-modal posterior distributions evaluation. *Mechanical Systems and Signal Processing 186*, 109868

Kamariotis A, **Straub D and Chatzi E** (2020) Optimal maintenance decisions supported by SHM: A benchmark study. In *Life-Cycle Civil Engineering: Innovation, Theory and Practice - Proceedings of the 7th International Symposium on Life-Cycle Civil Engineering, IALCCE 2020*, pp. 679–686. https://doi.org/10.1201/9780429343292-88/OPTIMAL-MAINTENANCE-DECISIONS-SUPPORTED-SHM-BENCHMARK-STUDY-KAMARIOTIS-STRAUB-CHATZI.

Kamariotis A, **Chatzi E and Straub D** (2023) A framework for quantifying the value of vibration-based structural health monitoring. *Mechanical Systems and Signal Processing 184*, 109708. https://doi.org/10.1016/J.YMSSP.2022.109708.

Kennedy MC and O'Hagan A (2001) Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(3), 425–464.

Kingma DP, **Salimans T**, **Jozefowicz R**, **Chen X**, **Sutskever I and Welling M** (2016) Improving variational inference with inverse autoregressive flow. *Advances in Neural Information Processing Systems*, 4743–4751. http://arxiv.org/abs/1606.04934.

Koune I, **Rózsás A**, **Slobbe A and Cicirello A** (2023) Bayesian system identification for structures considering spatial and temporal correlation. *Data-Centric Engineering 4*, e22.

Kuhn D, **Esfahani PM**, **Nguyen VA and Shafieezadeh-Abadeh S** (2019) Wasserstein distributionally robust optimization: Theory and applications in machine learning. *Operations Research & Management Science in the Age of Analytics*, 130–166. https://doi.org/10.1287/educ.2019.0198.

Lin T, **Jin C and Jordan MI** (2019) On gradient descent ascent for nonconvex-concave minimax problems. In *37th International Conference on Machine Learning, ICML 2020*, *PartF168147-8*, pp. 6039–6049. https://arxiv.org/abs/1906.00331v8.

Liu Q (2017) Stein variational gradient descent as gradient flow. *Advances in Neural Information Processing Systems*, 3116–3124. https://arxiv.org/abs/1704.07520v2

Liu Q and Wang D (2016) Stein variational gradient descent: A general purpose Bayesian inference algorithm. *Advances in Neural Information Processing Systems*, 2378–2386. https://arxiv.org/abs/1608.04471v3

Liu C, **Cheng P**, **Zhang R**, **Zhuo J**, **Zhu J and Carin L** (2018). Accelerated First-order Methods on the Wasserstein Space for Bayesian Inference. https://www.researchgate.net/publication/326222852.

Lu Y (2022) Two-Scale Gradient Descent Ascent Dynamics Finds Mixed Nash Equilibria of Continuous Games: A Mean-Field Perspective. https://arxiv.org/abs/2212.08791v2.

Lu J, **Lu Y and Nolen J** (2019) Scaling limit of the Stein variational gradient descent: The mean field regime. *SIAM Journal on Mathematical Analysis 51*(2), 648–671. https://doi.org/10.1137/18M1187611.

Lye A, **Cicirello A and Patelli E** (2021) Sampling methods for solving Bayesian model updating problems: A tutorial. *Mechanical Systems and Signal Processing 159*, 107760.

The MathWorks Inc. (2022). MATLAB version: 9.8.0 (R2020a), Natick, Massachusetts: The MathWorks Inc. https://www.mathworks.com

Matsubara T, **Knoblauch J**, **Briol FX and Oates CJ** (2021) Robust generalised Bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society. Series B: Statistical Methodology 84*(3), 997–1022. https://doi.org/10.1111/rssb.12500.

Mei S, **Montanari A and Nguyen PM** (2018) A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences of the United States of America 115*(33), E7665–E7671. https://doi.org/10.1073/PNAS.1806579115/SUPPL_FILE/PNAS.1806579115.SAPP.PDF.

Mottershead J and Friswell M (1993) Model updating in structural dynamics: A survey. *Journal of Sound and Vibration 167*(2), 347–375.

van Parys BPG, **Esfahani PM and Kuhn D** (2017) From data to decisions: Distributionally robust optimization is optimal. *Management Science 67*(6), 3387–3402. https://doi.org/10.1287/mnsc.2020.3678.

Ramgraber M, **Weatherl R**, **Blumensaat F and Schirmer M** (2021) Non-gaussian parameter inference for hydrogeological models using stein variational gradient descent. *Water Resources Research 57*(4), e2020WR029339. https://doi.org/10.1029/2020WR029339.

Rasmussen CE (2003) Gaussian processes to speed up hybrid Monte Carlo for expensive Bayesian integrals. *Bayesian Statistics 7*, 651–659.

Rocchetta R, **Broggi M**, **Huchet Q and Patelli E** (2018) On-line Bayesian model updating for structural health monitoring. *Mechanical Systems and Signal Processing 103*, 174–195. https://doi.org/10.1016/j.ymssp.2017.10.015.

Rytter A (1993) Vibrational Based Inspection of Civil Engineering Structures. Dept. of Building Technology and Structural Engineering, Aalborg University.

Sankararaman S (2015) Significance, interpretation, and quantification of uncertainty in prognostics and remaining useful life prediction. *Mechanical Systems and Signal Processing 52*–53, 228–247. https://doi.org/10.1016/j.ymssp.2014.05.029.

Santambrogio F (2015) *Optimal Transport for Applied Mathematicians*, Vol. *87*. Heidelberg New York Dordrecht London: Springer International Publishing. https://doi.org/10.1007/978-3-319-20828-2.

Santambrogio F (2016) {Euclidean, metric, and Wasserstein} gradient flows: An overview. *Bulletin of Mathematical Sciences 7*(1), 87–154. https://doi.org/10.1007/s13373-017-0101-1.

Sedehi O, **Papadimitriou C**, **Katafygiotis LS** (2019) Probabilistic hierarchical Bayesian framework for time-domain model updating and robust predictions. *Mechanical Systems and Signal Processing 123*, 648–673.

Simoen E, **Papadimitriou C and Lombaert G** (2013) On prediction error correlation in Bayesian model updating. *Journal of Sound and Vibration 332*(18), 4136–4152.

Simoen, E., **De Roeck, G.**, **& Lombaert, G.** (2015). Dealing with uncertainty in model updating for damage assessment: A review. In *Mechanical Systems and Signal Processing*, Vol. *56*, pp. 123–149. https://doi.org/10.1016/j.ymssp.2014.11.001.

Straub D and Papaioannou I (2015) Bayesian updating with structural reliability methods. *Journal of Engineering Mechanics 141*(3), 04014134.

Verzobio A, **Bolognani D**, **Quigley J and Zonta D** (2018) The consequences of heuristic distortions on SHM-based decision problems. *E-Journal of Nondestructive Testing 23*(11). https://www.ndt.net/search/docs.php3?id=23282

Wang Y, **Chen P and Li W** (2022) Projected Wasserstein gradient descent for high-dimensional Bayesian inference. *SIAM/ASA Journal on Uncertainty Quantification 10* (4), 1513–1532. https://doi.org/10.1137/21M1454018.

Yamada M, **Suzuki T**, **Kanamori T**, **Hachiya H and Sugiyama M** (2011) Relative density-ratio estimation for robust distribution comparison. *Neural Computation 25*(5), 1324–1370. https://doi.org/10.1162/NECO_a_00442.

Yuen KV, **Beck JL and Katafygiotis LS** (2006) Efficient model updating and health monitoring methodology using incomplete modal data without mode matching. *Structural Control and Health Monitoring 13*(1), 91–107. https://doi.org/10.1002/stc.144.

Cite this article: Igea F and Cicirello A (2025). An interacting Wasserstein gradient flow strategy to robust Bayesian inference for application to decision-making in engineering. *Data-Centric Engineering*, 6, e19. doi:10.1017/dce.2025.8