SIAM J. SCI. COMPUT.
Vol. 46, No. 5, pp. A3124–A3150

A DISTRIBUTIONS-BASED APPROACH FOR DATA-CONSISTENT INVERSION*

KIRANA O. BERGSTROM†, TROY D. BUTLER†, AND TIMOTHY M. WILDEY‡

Abstract. We formulate a novel approach to solve a class of stochastic problems, referred to as data-consistent inverse (DCI) problems, which involve the characterization of a probability measure on the parameters of a computational model whose subsequent push-forward matches an observed probability measure on specified quantities of interest (QoI) typically associated with the outputs from the computational model. Whereas prior DCI solution methodologies focused on either constructing nonparametric estimates of the densities or the probabilities of events associated with the preimage of the QoI map, we develop and analyze a constrained quadratic optimization approach based on estimating push-forward measures using weighted empirical distribution functions. The method proposed here is more suitable for low-data regimes or high-dimensional problems than the density-based method, as well as for problems where the probability measure does not admit a density. Numerical examples are included to demonstrate the performance of the method and to compare with the density-based approach where applicable.

Key words. data-consistent inversion, stochastic inverse problems, uncertainty quantification, quadratic optimization

MSC codes. 28A50, 65K10, 62G07

DOI. 10.1137/24M1641646

See reproducibility of computational results at end of the article.

1. Introduction.

The formulation and solution of an inverse problem impacts the type of information revealed about parameters of a model from observed data. For instance, inverse problems can be posed deterministically and solved using optimization methods [28, 21] to determine "parameters of best fit" in a particular sense. Some stochastic inverse problems seek a quantification of epistemic uncertainty in likely parameter values utilizing Bayesian methods [27, 16]. In this work, we seek to quantify aleatoric uncertainties about the parameters of the model from probabilistic information of the observed data. In other words, we seek a probability measure

*Submitted to the journal's Numerical Algorithms for Scientific Computing section March 5, 2024; accepted for publication (in revised form) July 2, 2024; published electronically October 3, 2024.

https://doi.org/10.1137/24M1641646

Funding: The work of the second author was supported by the National Science Foundation under grant DMS-2208460 as well as by the NSF IR/D program, while working at the National Science Foundation. However, any opinion, finding, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. The work of the third author was supported by the Office of Science, Early Career Research Program. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

†Department of Mathematical and Statistical Sciences, University of Colorado - Denver, Denver, CO 80204 USA (kirana.bergstrom@ucdenver.edu, troy.butler@ucdenver.edu).

‡Computational Mathematics Department, Center for Computing Research, Sandia National Labs, Albuquerque, NM 87185 USA (tmwilde@sandia.gov).

of the model inputs (i.e., the model parameters) given a probability measure that characterizes uncertainties of the model outputs. We require the solution to this inverse problem to be *data-consistent,* meaning that the push-forward of the probability measure on the parameter space matches a given target distribution on the observed data space. This type of inverse problem naturally arises in scenarios where observed variability in data is primarily due to intrinsic variability in the model inputs.

1.1. Comparison to related inverse problem work. A probability distribution can be characterized by its cumulative distribution function (CDF), measure, or, if it exists, its density (i.e., its Radon–Nikodym derivative). Previous approaches to solving the data-consistent inverse (DCI) problem either approximate a pullback measure directly through event approximation in both the input and output spaces as in [12] or use a nonparametric kernel density approximation in the output space [11]. These approaches are challenging to implement when the number of simulated data are limited, e.g., when the computational model is expensive to evaluate. Moreover, there is an implicit assumption that sufficient information exists on observations to specify either a measure or density. The method described in this paper circumvents these issues by directly approximating a CDF via an optimally weighted empirical distribution function (EDF). The proposed method does not require the existence of densities for any of the distributions involved, and it does not rely on kernel density estimation methods, which may be unreliable in low-data regimes and are computationally expensive in high-dimensional spaces. A key feature of DCI solutions is that all computations are performed in the output space, which is generally lower-dimensional than the parameter space, and the proposed method preserves this advantage.

1.2. Contributions. The contributions of this work are both algorithmic and theoretical. The method we propose builds upon the approach introduced and analyzed in [2, 26] for approximating push-forward measures by performing a change-of-measure objective using EDFs via solution to an optimization problem that determines optimal weights on the simulated output samples. These weights minimize the L^2-norm between the weighted EDF and a given target distribution function and are guaranteed to exist since they are the result of a strictly convex quadratic optimization problem. Thus, for the case where simulated data are limited, a solution is achievable that is guaranteed optimal in an L^2-sense. Moreover, the target distribution may be in the form of either a CDF or an EDF, which guarantees a solution to this approach even in cases further limited by availability of observational data.

The key to building upon the optimization-based approach for push-forward EDFs to solve the DCI problem is through the addition of a critical binning step in the output space. This provides a practical partitioning of the observed space prior to solution of the optimization problem that permits the proper distribution of weights in the input space. We both prove and demonstrate numerical convergence of the optimization-based solution to the DCI solution.

1.3. Organization. Background and previous approaches for DCI are described in section 2. Section 3 describes the optimization-based approach. Convergence and other theoretical results are discussed in section 4 (proofs are found in the appendices). Section 5 contains applications and examples. Conclusions and future directions follow in section 6. Section 7 provides details on obtaining the code utilized to produce the results in this work.

2. Data-consistent inversion (DCI). We summarize the concepts of DCI and direct the interested reader to [11] for more details.

2.1. Terminology and notation. Let $\boldsymbol{\lambda} \in \Lambda$ denote the uncertain parameters in a model for which the goal is to estimate a distribution given a particular distribution on quantities of interest (QoI) corresponding to model outputs. Let Q denote the QoI map from the parameter space, Λ, to the data space, $\mathcal{D}$, i.e., $Q(\Lambda) = \mathcal{D}$. We equip Λ and $\mathcal{D}$ with σ-algebras $\mathcal{B}_\Lambda$ and $\mathcal{B}_\mathcal{D}$ as well as measures μ_Λ and $\mu_\mathcal{D}$, respectively. These measures are dominating measures that allow us to express probability measures defined on these spaces as Radon–Nikodym derivatives. While not necessary, it is often the case that $\Lambda \subset \mathbb{R}^p$ and $\mathcal{D} \subset \mathbb{R}^d$, $\mathcal{B}_\Lambda$ and $\mathcal{B}_\mathcal{D}$ are Borel σ-algebras, and μ_Λ and $\mu_\mathcal{D}$ are Lebesgue measures. In that case, the Radon–Nikodym derivatives are referred to as probability density functions or more simply as densities.

Due to the variability in the parameters, the QoI follow a distribution. The distribution of these observations defines the *observed distribution* and corresponds to a probability measure, P_{obs}, on the measurable space $(\mathcal{D}, \mathcal{B}_\mathcal{D})$. The problem of determining a distribution of parameters that could have produced the observed distribution is a stochastic inverse problem that can be solved using DCI. In mathematical terms, the problem is stated as follows: given an observed measure P_{obs} on $(\mathcal{D}, \mathcal{B}_\mathcal{D})$, we seek a measure P_Λ on $(\Lambda, \mathcal{B}_\Lambda)$ such that for $A \in \mathcal{B}_\mathcal{D}$,

$$P_\Lambda(Q^{-1}(A)) = P_{\text{obs}}(A). \tag{2.1}$$

Such a measure is P_Λ is called a *pullback* of P_{obs}, and P_{obs} is the *push-forward* of P_Λ. Note that we are using the common measure-theoretic shorthand notation Q^{-1} to denote the *preimage* map (i.e., we are not assuming Q is invertible).

We emphasize that the stochastic inverse problem is often ill-posed, i.e., many pullback measures may exist. Even in cases where the dimension of $\mathcal{D}$ equals or even exceeds the dimension of Λ, nonlinear Q will often result in nonuniqueness of solutions. Section 2.2 summarizes both a theoretical and a practical approach for constructing a particular solution under some additional assumptions.

2.2. Density-based solutions. Here, we summarize the theoretical construction of the "density-based" solution in the more general context of Radon–Nikodym derivatives of probability measures. We use the term "density-based" because it is frequently the case that the derivatives are densities as stated above.

If P_{obs} admits a Radon–Nikodym derivative with respect to $\mu_\mathcal{D}$, it is denoted π_{obs}. We then reframe the stochastic inverse problem in terms of Radon–Nikodym derivatives where (2.1) is rewritten as

$$P_\Lambda(Q^{-1}(A)) = \int_{Q^{-1}(A)} \pi_\Lambda \, d\mu_\Lambda = \int_A \pi_{\text{obs}} \, d\mu_\mathcal{D} = P_{\text{obs}}(A). \tag{2.2}$$

In other words, the goal is to determine a pullback of P_{obs} in terms of a Radon–Nikodym derivative with respect to μ_Λ, which we denote here by π_Λ.

Clearly, (2.1) dictates some restrictions on the structure of a solution, π_Λ, in terms of its aggregate behavior on sets defined by Q^{-1}, but this QoI map provides no information as to how π_Λ should behave *within* a set $Q^{-1}(\boldsymbol{q})$ for a given $\boldsymbol{q} \in \mathcal{D}$.

The key theoretical tool for constructing a solution to the stochastic inverse problem is the disintegration theorem [13]. Below, we summarize a version for probability measures that describes the *unique* decomposition of any probability measure defined on $(\Lambda, \mathcal{B}_\Lambda)$ in terms of its associated push-forward and a family of conditional probability measures.

THEOREM 2.1 (disintegration theorem). *Assume $Q : \Lambda \to \mathcal{D}$ is $\mathcal{B}_\Lambda$-measurable, P_Λ is a probability measure on $(\Lambda, \mathcal{B}_\Lambda)$, and $P_\mathcal{D}$ is the push-forward measure of P_Λ on $(\mathcal{D}, \mathcal{B}_\mathcal{D})$. There exists a $P_\mathcal{D}$-almost everywhere uniquely defined family of conditional probability measures $\{P_{\boldsymbol{q}}\}_{\boldsymbol{q}\in\mathcal{D}}$ on $(\Lambda, \mathcal{B}_\Lambda)$ such that for any $A \in \mathcal{B}_\Lambda$,*

$$P_{\boldsymbol{q}}(A) = P_{\boldsymbol{q}}(A \cap Q^{-1}(\boldsymbol{q})),$$

so $P_{\boldsymbol{q}}(\Lambda \setminus Q^{-1}(\boldsymbol{q})) = 0$, and there exists the following disintegration of P_Λ:

$$P_\Lambda(A) = \int_\mathcal{D} P_{\boldsymbol{q}}(A) dP_\mathcal{D}(\boldsymbol{q}) = \int_\mathcal{D} \left(\int_{A\cap Q^{-1}(\boldsymbol{q})} dP_{\boldsymbol{q}}(\boldsymbol{\lambda}) \right) dP_\mathcal{D}(\boldsymbol{q})$$

for $A \in \mathcal{B}_\Lambda$.

From Theorem 2.1, it is self-evident that P_Λ is a solution to the stochastic inverse problem if its push-forward $P_\mathcal{D}$ is equal to P_{obs}. While this gives some mechanism for checking whether a proposed probability measure is a solution to the stochastic inverse problem, it fails to address how to construct P_Λ since the family of conditional probability measures remains undefined. To address this, we utilize an initial probability measure P_{init} on $(\Lambda, \mathcal{B}_\Lambda)$. The disintegration of this measure is used to construct the unknown family of conditional probability measures. The push-forward of P_{init} defines a measure we refer to as the *predicted* probability measure, P_{pred}.

In [11], it is shown that if the initial, observed, and predicted distributions admit Radon–Nikodym derivatives, and if the observed measure P_{obs} is absolutely continuous with respect to the predicted measure P_{pred}, then Theorem 2.1 implies a solution to the stochastic inverse problem is given by

$$P_{\text{update}}(A) := \int_\mathcal{D} \left(\int_{A\cap Q^{-1}(\boldsymbol{q})} \frac{\pi_{\text{init}}(\boldsymbol{\lambda})}{\pi_{\text{pred}}(Q(\boldsymbol{\lambda}))} d\mu_{\Lambda,\boldsymbol{q}} \right) \pi_{\text{obs}}(\boldsymbol{q}) d\mu_\mathcal{D} \tag{2.3}$$

for all $A \in \mathcal{B}_\Lambda$, where $\{\mu_{\Lambda,\boldsymbol{q}}\}_{\boldsymbol{q}\in\mathcal{D}}$ comes from the disintegration of the measure μ_Λ, π_{pred} is the Radon–Nikodym derivative of P_{pred}, and the term $\pi_{\text{init}}(\boldsymbol{\lambda})/\pi_{\text{pred}}(Q(\boldsymbol{\lambda}))$ defines the conditional density associated with $Q^{-1}(\boldsymbol{q})$ for each $\boldsymbol{q} \in \mathcal{D}$. For further details of this derivation, and proof that such a solution is data-consistent, we refer the reader to [11].

The term $\pi_{\text{obs}}(\boldsymbol{q})$ in (2.3) can be brought into the inner integral by rewriting it as $\pi_{\text{obs}}(Q(\boldsymbol{\lambda}))$. It follows that the data-consistent solution is given by integrating the following Radon–Nikodym derivative over Λ:

$$\pi_{\text{update}}(\boldsymbol{\lambda}) := \pi_{\text{init}}(\boldsymbol{\lambda}) \frac{\pi_{\text{obs}}(Q(\boldsymbol{\lambda}))}{\pi_{\text{pred}}(Q(\boldsymbol{\lambda}))} = \pi_{\text{init}}(\boldsymbol{\lambda}) r(\boldsymbol{\lambda}) \tag{2.4}$$

where, by a standard result involving changes of measures and Radon–Nikodym derivatives,[1] the ratio $r(\boldsymbol{\lambda}) = \frac{\pi_{\text{obs}}(Q(\boldsymbol{\lambda}))}{\pi_{\text{pred}}(Q(\boldsymbol{\lambda}))}$ defines the Radon–Nikodym derivative of P_{obs} with respect to P_{pred} and serves to reweight the initial likelihoods of parameters. It is worth noting that for each datum $\boldsymbol{q} \in \mathcal{D}$, $r(\boldsymbol{\lambda})$ is *constant* for all $\boldsymbol{\lambda} \in Q^{-1}(\boldsymbol{q})$. In other words, this reweighting updates the initially assumed likelihoods $Q^{-1}(\boldsymbol{q})$ for each $\boldsymbol{q} \in \mathcal{D}$ without updating the initially assumed conditional likelihoods of points within such sets. It is this observation that leads to the reference of the solution given in (2.3) and (2.4) as an *update* to the initial distribution.

[1] Cf. Proposition 3.9(a) in [19].

2.3. Numerical approximation of densities. Even when spaces are nominally infinite-dimensional, practical considerations and discretizations often result in $\Lambda \subset \mathbb{R}^p$ and $\mathcal{D} \subset \mathbb{R}^d$ for some finite p and d so that the Radon–Nikodym derivatives (with the exception of $r(\boldsymbol{\lambda})$) are densities as previously mentioned. We are often confronted with the situation where these densities must be approximated from a finite set of (typically i.i.d.) samples from the observed distribution along with a finite set of (typically i.i.d.) samples from the initial distribution and their associated predicted values. Below, we describe a typical approach based on direct estimation of the observed and predicted densities.

Suppose we have m data points $\{\boldsymbol{q}^1, \dots, \boldsymbol{q}^m\}$, equal to $\{Q(\boldsymbol{\lambda}^1), \dots, Q(\boldsymbol{\lambda}^m)\}$, where $\{\boldsymbol{\lambda}^1, \dots, \boldsymbol{\lambda}^m\}$ are unobserved. Here, we take a moment to clearly define the use of subscript and superscript indices throughout this work. Since any collection of samples may denote a collection of vectors, we denote each sample within the collection using superscripts. Any index subscript denotes a component of a vector; i.e., the ith sample of a d-dimensional QoI is denoted as $\boldsymbol{q}^i = [\boldsymbol{q}_1^i, \dots, \boldsymbol{q}_d^i]^\top$.

Any parametric or nonparametric density estimation method can be applied to estimate π_{obs} from these m samples. Some popular nonparametric density estimation techniques utilized in the literature are kernel density estimation (KDE) [17, 14], normalizing flows (NFs) [25], and Dirichlet process–based mixture models [18, 23].

We emphasize that even if the initial density is given exactly, the predicted density is not typically known except for the simplest of QoI maps and initial distributions. In practice, the predicted density, π_{pred}, is approximated by propagating n i.i.d. samples from the initial distribution, $\{\boldsymbol{\lambda}^1, \dots, \boldsymbol{\lambda}^n\}$, through the map Q, to create a set of i.i.d. samples from the (unknown) predicted distribution $\{Q(\boldsymbol{\lambda}^1) = \boldsymbol{q}^1, \dots, Q(\boldsymbol{\lambda}^n) = \boldsymbol{q}^n\}$ for which any density estimation technique can be applied. As previously mentioned, the theoretical existence of π_{update} assumes P_{obs} is absolutely continuous with respect to P_{pred}. However, in practice, we often make a stronger assumption that allows us to utilize this set of i.i.d. samples from the initial distribution to either create an estimate of π_{update} or apply rejection sampling to this set to generate an i.i.d. set of samples from the updated distribution. We refer the interested reader to [11] for more details and simply restate the stronger form below.

Assumption 2.2 (strong form of the predictability assumption). There exists a constant $C > 0$ such that for almost every $\boldsymbol{q} \in \mathcal{D}$, $\pi_{\text{obs}}(\boldsymbol{q}) \leq C\pi_{\text{pred}}(\boldsymbol{q})$.

To verify this assumption is satisfied in practice, we utilize a quantitative diagnostic that we summarize below. If the predictability assumption is satisfied, then π_{update} given in (2.4) defines a density, which implies

$$1 = \int_\Lambda \pi_{\text{update}}(\boldsymbol{\lambda})\, d\mu_\Lambda = \int_\Lambda \pi_{\text{init}}(\boldsymbol{\lambda}) r(\boldsymbol{\lambda})\, d\mu_\Lambda =: \mathbb{E}_{\text{init}}(r(\boldsymbol{\lambda})). \tag{2.5}$$

Thus, in practice, we verify the predictability assumption holds by computing the sample average of $r(\boldsymbol{\lambda})$ on a random sample drawn from the initial distribution, and comparing this to a value of unity.

2.4. An illustrative example. Consider a manufacturing process for creating metal alloy rods to be used in a high-temperature environment such as welding. Inconsistencies in the manufacturing process, such as the mechanical processes for the cutting, measuring, and smoothing of the ends of the rods, coupled with inconsistencies in the alloy, result in variations in rod lengths, ℓ, and thermal diffusivities, κ. For this example, assume that $\ell \in [1.9, 2.1]$ and $\kappa \in [0.5, 1.5]$ are both dimensionless. Further assume that the objective is to determine the distribution of ℓ and κ

prior to the utilization of the rods in an engineered system whose functional safety is dependent on these parameters. Using the notation above, we let $\boldsymbol{\lambda} = (\ell, \kappa)$ and $\Lambda = [1.9, 2.1] \times [0.5, 1.5] \subset \mathbb{R}^2$.

We now define a model of an experimental procedure that leads to a QoI map between the measure spaces $(\Lambda, \mathcal{B}_\Lambda, \mu_\Lambda)$ and $(\mathcal{D}, \mathcal{B}_\mathcal{D}, \mu_\mathcal{D})$. Assuming that the rods are long and thin, the temperature of the rod at spatial location $0 < x < \ell$ and time $t > 0$ is modeled by the 1-dimensional heat equation

$$\begin{cases} \dfrac{\partial}{\partial t} u(x,t) = \kappa \dfrac{\partial^2}{\partial x^2} u(x,t), & 0 < x < \ell, \\ u(0,t) = u(\ell,t) = 0, & t > 0, \\ u(x,0) = x, & 0 < x < \ell, \end{cases}$$

where u is (dimensionless) temperature. For the sake of illustration and simplicity, we consider homogeneous Dirichlet boundary conditions, where "0" is the ambient temperature, throughout the experiment. The initial temperature profile $u(x,0) = x$ is chosen to provide an interesting response over both space-time and the parameter space. The analytic solution for the temperature is given by

$$u(x,t;\boldsymbol{\lambda}) = \frac{2\ell^2}{\pi} \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} \exp\left[-\kappa \frac{k\pi}{\ell^2} t\right] \sin\left(\frac{k\pi x}{\ell}\right). \tag{2.6}$$

Assume for each experiment that we have a single sensor that can accurately and precisely measure the temperature at a specific point $x^\star = 1.2$ along the rod and point $t^\star = 0.01$ in time. In this case, $Q(\boldsymbol{\lambda}) = u(x^\star, t^\star; \boldsymbol{\lambda})$ defines the QoI map from Λ to $\mathcal{D} \subset \mathbb{R}$. For practical computations, we truncate this sum after the 100th term. Note that for this example the QoI is defined by a single temperature measurement in space-time so that $\mathcal{D} \subset \mathbb{R}$, whereas the parameter space $\Lambda \subset \mathbb{R}^2$. Subsequently, there are infinitely many possible combinations of ℓ and κ in the contour defined by $Q^{-1}(\boldsymbol{q})$ for a given datum $\boldsymbol{q} \in \mathcal{D}$, and the conditional probability for these (ℓ, κ) pairs is not specified by (2.1). See Figure 1 for an illustration. For the sake of simplicity, to generate the observed samples, we assume π_{obs} is normal and take m samples from it. This simulates a typical scenario where we do not know the exact observed

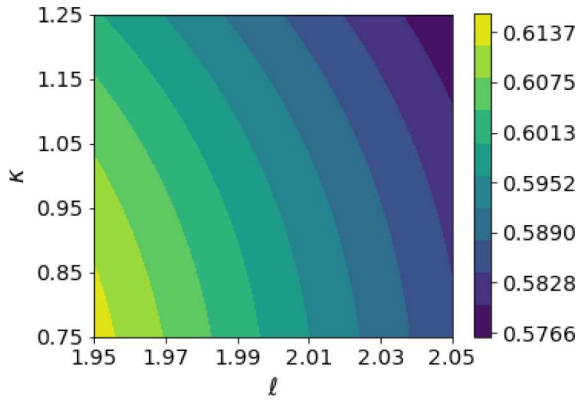

FIG. 1. *Contours for the map Q in the illustrative example. Note the distinct "contour sets" whose probabilities are uniquely defined by P_{obs}. However, the probabilities within these contour sets cannot be determined by P_{obs}.*

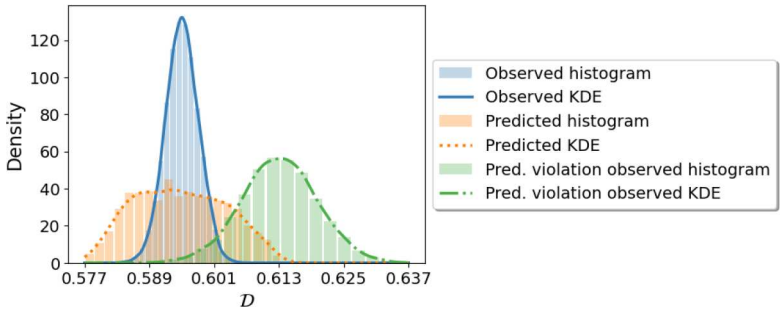

FIG. 2. *Predicted and observed histograms for the illustrative example and their estimated KDEs. We also show an alternate possible observed distribution, labeled "Pred. violation observed" because for this example, Assumption* 2.2 *is violated: this observed distribution is not absolutely continuous with respect to the predicted.*

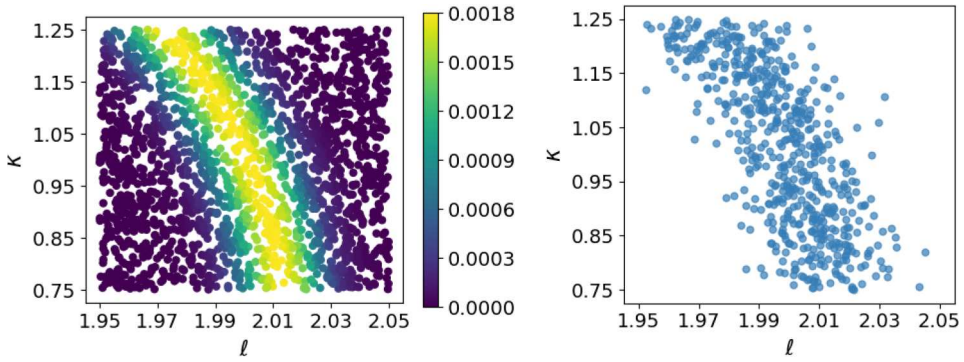

FIG. 3. *Result of density-based method for data-consistent inversion. On the left we show density-based weights plotted on the initial samples. The result of rejection sampling is shown on the right.*

distribution but instead have access to samples drawn from this distribution. Here, we utilize a Gaussian KDE to estimate π_{obs} from the data.

We assume a uniform initial distribution Λ and propagate $n = 2E3$ i.i.d. samples from this distribution to produce samples from the predicted distribution. Since $\mathcal{D}$ is dimensionless, we can either visually confirm Assumption 2.2 by comparing the observed and predicted density approximations in Figure 2 or compute the diagnostic, which to five significant digits is $\mathbb{E}_{\text{init}}(r(\boldsymbol{\lambda})) \approx 1.0064$. We pause here to demonstrate the usefulness of this diagnostic. For the predictability-violating observed distribution shown in Figure 2, the diagnostic is $\mathbb{E}_{\text{init}}(r(\boldsymbol{\lambda})) \approx 0.5118$. For more complex examples when the distributions are not easily visualized, the diagnostic provides a reliable way to verify whether Assumption 2.2 is satisfied.

With Assumption 2.2 verified, we move onto interrogating π_{update} via the weights $r(\boldsymbol{\lambda})$ available on the set of initial samples. In the left plot of Figure 3, we visualize $r(\boldsymbol{\lambda})$ as a function on the set of initial samples, while in the right plot we show an i.i.d. set of samples for π_{update} generated via rejection sampling based on these $r(\boldsymbol{\lambda})$ values.

3. Empirical distributions and optimal estimation. The density-based method for solving the DCI problem from section 2 requires the initial, predicted,

and observed distributions to all admit densities. Moreover, when any of these densities are estimated, a sufficient number of samples must exist to approximate them, e.g., using KDEs. In practice, the number of observed or simulated samples may be limited due to either high experimental or computational cost in obtaining each sample. Moreover, utilizing density estimation methods may fail if even a single one of the densities (initial, predicted, or observed) fails to exist. We therefore propose an alternative method based on empirical distribution functions (EDFs), which always exist as approximations to cumulative distribution functions (CDFs) that are in 1-to-1 correspondence with probability measures.

An attractive feature of the density method is that all of the critical computations take place in the data space, which is generally lower-dimensional than the parameter space. Specifically, the densities of the predicted and observed distributions are estimated, and a ratio of densities is computed, all in the data space. While the ratios are then applied to samples in the parameter space, the computations are restricted to the data space. We pursue a distribution-based approach with a similar feature. The specific method we propose, in section 3.3, is motivated from the empirical importance weights for a change of probability measure method described in [2]. In [2], i.i.d. samples from an unknown *proposal* distribution are reweighted using an optimization-based method to perform a change-of-measure from the proposal to a *target* distribution that itself may only be given in terms of a set of i.i.d. samples. We summarize this method, using our notation and terminology, in section 3.1.

3.1. Computation of optimal weights on the data space. For all but the simplest of QoI maps and initial distributions, the predicted distribution is unknown and must be estimated. We assume access to a set of n i.i.d. samples $\{\boldsymbol{q}^i\}_{i=1}^n$ to estimate the predicted empirical CDF as

$$F^n_{\text{pred}}(\boldsymbol{q}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\boldsymbol{q}^i \preceq \boldsymbol{q}),$$

where

$$\mathbb{I}(\boldsymbol{q}^i \preceq \boldsymbol{q}) = \begin{cases} 1 & \text{if } \boldsymbol{q}^i \preceq \boldsymbol{q}, \\ 0 & \text{else,} \end{cases}$$

which converges almost everywhere to F_{pred} as $n \to \infty$, e.g.; see Theorem 20.6 in [8]. Note that because the samples are vectors in $\mathbb{R}^d$ space, we use the symbols $\preceq$ and $\succeq$ to denote elementwise less than and elementwise greater than, respectively. For example if $\boldsymbol{x} \preceq \boldsymbol{y}$, then $\boldsymbol{x}_1 \le \boldsymbol{y}_1$, $\boldsymbol{x}_2 \le \boldsymbol{y}_2, \ldots, \boldsymbol{x}_d \le \boldsymbol{y}_d$.

An EDF over n samples can be transformed by a vector $\boldsymbol{w}$ into a *weighted EDF* as

$$F^n_{\text{pred};\boldsymbol{w}}(\boldsymbol{q}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{w}_i \mathbb{I}(\boldsymbol{q}^i \preceq \boldsymbol{q}), \tag{3.1}$$

where $\boldsymbol{w}_i \ge 0$ for $1 \le i \le n$ and $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{w}_i = 1$ so that $F^n_{\text{pred};\boldsymbol{w}}$ defines a valid probability distribution.

In [2], it is proposed that the weights should minimize the quadratic function

$$\frac{1}{2}\int_{\mathcal{D}} \left(F^n_{\text{pred};\boldsymbol{w}}(\boldsymbol{q}) - F_{\text{obs}}(\boldsymbol{q})\right)^2 d\mu_{\mathcal{D}},$$

Routine 3.1. $QP(\{\boldsymbol{q}^i, \ldots, \boldsymbol{q}^\ell\} \subset \mathbb{R}^d, F_{\text{targ}})$.

1: Set $\boldsymbol{H} \in \mathbb{R}^{\ell \times \ell}$ and $\boldsymbol{b} \in \mathbb{R}^\ell$ as

$$\boldsymbol{H}_{ij} := \frac{1}{\ell^2} \prod_{k=1}^{d} \int_{z_k^{i,j}}^{1} d\boldsymbol{q}_k, \quad b_i := \frac{1}{\ell} \prod_{k=1}^{d} \int_{\boldsymbol{q}_k^i}^{1} F_{\text{targ}}(\boldsymbol{q})\, d\boldsymbol{q}_k$$

where $z_k^{i,j} = \max\{\boldsymbol{q}_k^i, \boldsymbol{q}_k^j\}$.

2: Solve for $\boldsymbol{w}$:

$$\begin{aligned} &\text{minimize} && \boldsymbol{w}^\top \boldsymbol{H} \boldsymbol{w} - \boldsymbol{b}^\top \boldsymbol{w} \\ &\text{subject to} && \boldsymbol{w} \succeq \boldsymbol{0}, \\ & && \frac{1}{\ell} \sum_{i=1}^{\ell} \boldsymbol{w}_i = 1. \end{aligned} \tag{3.2}$$

3: Return $\boldsymbol{w}$. ▷ Output

which is the square of the L^2-norm of the difference between the weighted predicted EDF and the true observed CDF. Moreover, [2] proves that this choice of $\boldsymbol{w}$ defines a distribution with a CDF that converges in both L^1 and L^2 to the observed CDF F_{obs}.

We note here that, as in [2], we can, without loss of generality, assume that the support of the predicted sample distribution is contained within the d-dimensional hypercube. If this is not the case, we simply scale a bounding box for the samples from the predicted distribution using either the known (assumed compact) support of distribution, or the minimum and maximum componentwise values within the sample set and the integrand by the same factor. The scaled problem is equivalent to the unscaled version, and the optimal $\boldsymbol{w}$ is the same. To find the minimizing weights $\boldsymbol{w}$, we solve a standard-form quadratic program (QP) with linear constraints. This is summarized in Routine 3.1 (see [2] for derivation details). In Routine 3.1, we use ℓ instead of n for the number of samples because we consider two different algorithms in the following subsections that involve different numbers of samples. We also use F_{targ} instead of the CDF F_{obs} since an EDF that approximates this CDF may also be used as discussed below.

The matrix H constructed in Routine 3.1 is symmetric positive-definite so the QP has a unique solution [9]. We emphasize that the matrix H is dense and of order ℓ, meaning that the computational complexity of the QP increases as the number of samples does. However, the complexity of the QP is independent of the dimension of the problem. In step 2, any algorithm or package for solving QPs with linear constraints should suffice. In this work we solve the QP using the convex optimization Python package *cvxopt* [3]. Running $QP(\{\boldsymbol{q}^1, \ldots, \boldsymbol{q}^n\}, F_{\text{obs}})$ returns the vector of weights $\boldsymbol{w} \in \mathbb{R}^n$ such that $F^n_{\text{pred};\boldsymbol{w}}$ defined in (3.1) is the L^2-optimal estimate of F_{obs}. If the exact target CDF F_{targ} is not known, as in the case when we have samples from the observed distribution instead of an exact form, we can apply the method using the EDF F^m_{targ} for a set of m i.i.d. samples from the observed distribution. In that case, letting $\{\boldsymbol{y}^1, \ldots, \boldsymbol{y}^m\}$ denote the m samples of the target distribution, the b_i computation in step 1 reduces to

$$b_i = \frac{1}{\ell} \prod_{k=1}^{d} \int_{\boldsymbol{q}_k^i}^{1} F^m_{\text{targ}}(\boldsymbol{q})\, d\boldsymbol{q}_k = \frac{1}{\ell} \prod_{k=1}^{d} \int_{\boldsymbol{q}_k^i}^{1} \frac{1}{m} \sum_{j=1}^{m} \mathbb{I}(\boldsymbol{y}^j \preceq \boldsymbol{q})\, d\boldsymbol{q}_k.$$

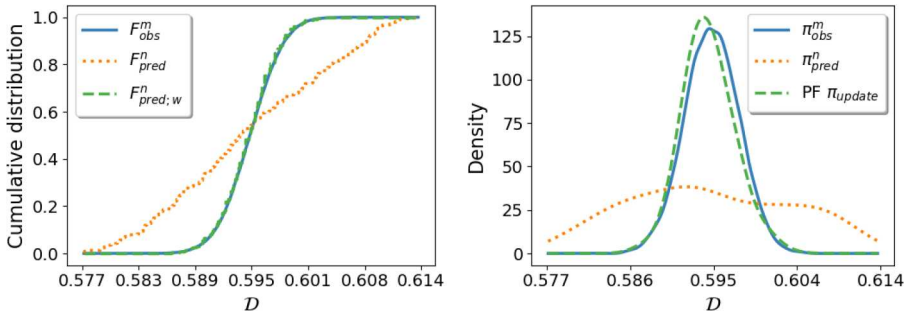

FIG. 4. *Data space comparisons of the optimal L^2 reweighting scheme with the density-based approach for the illustrative example when $n = 2E3$, $m = 1E4$. On the left we plot the EDFs; on the right we plot the estimated densities.*

It is shown in [2] that the resulting weighted EDF converges in the L^1- and L^2-norms to F_{targ} as $\ell, m \to \infty$. Applying Routine 3.1 to the illustrative example with $\ell = n = 2E2$ samples (chosen to demonstrate the step-function nature of the solution), where the target distribution is an EDF of a normal distribution (using $m = 1E4$), results in the distribution approximation shown in Figure 4. For the sake of comparison, we also show the results in the data space for the density-based method with $n = 2E2$, i.e., we plot the densities for predicted, observed, and push-forward of the updated distribution. In both plots, we observe that the estimated predicted distribution is a poor estimate of the observed distribution whether it is represented as an EDF (left plot) or density (right plot). The different reweighting schemes both produce a change of measure resulting in an improved estimate of the observed distribution. In the subsections below, we explore how to utilize the QP-based reweighting scheme to construct a weighted EDF approximation of the updated distribution.

3.2. Naïve approach for utilizing an EDF. The first method we consider for solving the DCI problem is deemed the "naïve" approach. For this approach, we use Routine 3.1 to compute a set of weights on the predicted samples (samples from the initial distribution that have been pushed forward through the map Q), and we apply these weights directly on the corresponding initial samples. This is summarized in Algorithm 3.2, which outputs $F^n_{\text{init};\boldsymbol{w}}$, defined in (3.4), as the solution. This defines the following probability measure:

$$P^n_{\text{init};\boldsymbol{w}}(B) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{w}_i \mathbb{I}(\boldsymbol{\lambda}^i \in B), \quad B \in \mathcal{B}_\Lambda. \tag{3.3}$$

By comparing $r(\boldsymbol{\lambda})$ (shown in the left plot of Figure 3) to the weights, $\boldsymbol{w}$, computed from the naïve method (shown in the left plot of Figure 5), we observe that while the geometric structures of these weights are similar, the values are not the same as shown on the right side of Figure 5. Not only does this naïve method place significantly more weight on certain samples, but the variation in the weights is also significantly greater than the variation in the density-based weights. This is due to the simple fact that the QP is solved in the data space without any controls for variability in the parameter space. Thus, while the solution is always optimal in the data space, it does not necessarily preserve any structure in the parameter space. More precisely, this naïve approach does not require the solution to conform to the well-defined conditional structure dictated by the generalized contours of the map in the parameter space. This observation motivates the binning approach in the next section.

Algorithm 3.2. Naïve approach.

1: Set $n \in \mathbb{N}$, F_{obs}, and $\{\boldsymbol{\lambda}^i\}_{i=1}^n$. ▷ Input
2: **for** $1 \le i \le n$ **do**
3: Evaluate $\boldsymbol{q}^i = Q(\boldsymbol{\lambda}^i)$.
4: **end for**
5: Use Routine 3.1 to compute $\boldsymbol{w} = QP(\{\boldsymbol{q}^i\}_{i=1}^n, F_{\text{obs}})$.
6: Apply $\boldsymbol{w}$ to initial samples $\{\boldsymbol{\lambda}^i\}_{i=1}^n$,

$$F^n_{\text{init};\boldsymbol{w}}(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{w}_i \mathbb{I}(\boldsymbol{\lambda}^i \preceq \boldsymbol{\lambda}). \tag{3.4}$$

7: Return $F^n_{\text{init};\boldsymbol{w}}$. ▷ Output

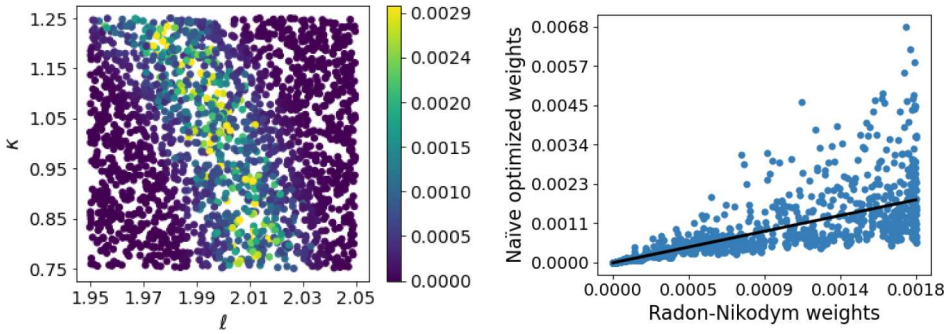

FIG. 5. *Weights resulting from naïve distribution-based method applied to initial samples on the left. Direct comparison to density-based weights on the right with the line indicating where perfect agreement occurs.*

3.3. Binning approach. We now describe a two-step method to enforce the desired structure along the so-called contours defined by the preimage map Q^{-1} for which the conditionals defined by the disintegration of the initial probability measure are desired. The key is to construct the "binning distribution" (defined below) associated with a particular partitioning of $\mathcal{D}$. Denote by $\{\mathcal{C}_k\}_{k=1}^p$ a partitioning of the data space. For each cell $\mathcal{C}_k$, a representative point, denoted by $\boldsymbol{c}^k$, is identified. These representative points could be the geometric center of the cell, the average of the predicted samples that fall into the cells, or any other suitable point in $\mathcal{C}_k$. Alternatively, we may first generate the points and then consider the partition defined (implicitly) as Voronoi cells. Each cell must be a *continuity set* (i.e., a measurable set A with $P(\partial A) = 0$ where ∂A denotes the topological boundary of A) of P_{pred}, which in turn implies that it is a continuity set of P_{obs}, by Assumption 2.2.

It is important to emphasize that the explicit construction of these cells is never actually required in either theory or practice, but we do reference the assumed continuity set property of these cells in the theoretical analysis of section 4.2. Further discussion on how to form an appropriate partition is found in section 3.4. For now, it suffices to understand this property as implying that the boundaries of such sets are both measurable and have zero-measure. We solve a QP using the representative points as input samples in Algorithm 3.2, which effectively gives us an L^2-optimized weight for each (implicitly defined) *cell*. We define the resulting weighted EDF below.

DEFINITION 3.1 (binning distribution). *The binning distribution associated with* $\{\boldsymbol{c}_k\}_{k=1}^p$ *is defined as* $F^p_{\mu_{\mathcal{D}};\boldsymbol{w}} = \frac{1}{p}\sum_{k=1}^p \boldsymbol{w}_i \mathbb{I}(\boldsymbol{c}^k \preceq \boldsymbol{q})$.

Note here that because our sample set is now composed of p samples, $\boldsymbol{w}$ is now a p-dimensional vector that is chosen separately from the n parameter samples for which the QoI map is evaluated. The subscript $\mu_{\mathcal{D}}$ in $F^p_{\mu_{\mathcal{D}},\boldsymbol{w}}$ is utilized to emphasize that the partitioning utilized to construct the bins may be done a priori to any specification of a probability measure or generation of a random sample set on $\mathcal{D}$. In other words, the measure of each bin may best be described, a priori, by utilizing $\mu_{\mathcal{D}}$.

To define a weighted EDF on Λ, we utilize a classifier, denoted Q^p, to *bin* the n samples $\{\boldsymbol{\lambda}^1, \ldots, \boldsymbol{\lambda}^n\}$ into the p (implicitly defined) sets $\{Q^{-1}(\mathcal{C}^1), \ldots, Q^{-1}(\mathcal{C}^p)\}$. It is at times conceptually convenient to also view the classifier Q^p as defining a (discrete) map between the parameter and data space so that $Q^p(\boldsymbol{\lambda}) = \boldsymbol{c}^k$ for a particular value of $1 \le k \le p$. The meaning of the notation Q^p as either a classifier or a discrete map will always be clear from the context. This results in a vector of n weights

Algorithm 3.3. Binning approach.

1: Set $p, n_{\text{batch}} \in \mathbb{N}$, F_{init}, and F_{obs}. ▷ Input
2: Partition $\mathcal{D} = \bigcup_{k=1}^p \mathcal{C}_k$.
3: Identify representative point $c^k \in \mathcal{C}_k$ for each k.
4: Use Routine 3.1 to compute $\boldsymbol{w} = QP(\{\boldsymbol{c}^k\}_{k=1}^p, F_{\text{obs}})$.
5: Use $\boldsymbol{w}$ to set $\{n_{k,\min}\}_{k=1}^p$ with $n_{k,\min} \ge 0$ for all k.
6: Set $n \leftarrow 0$, $n_{k,\text{total}} \leftarrow 0$ for $1 \le k \le p$, initialize empty array of parameter samples $\mathcal{S}$, and define classifier Q^p such that $Q^p(\lambda) = \boldsymbol{c}^k$ for some $1 \le k \le p$ for all $\lambda \in \Lambda$.
7: **while** $n_{k,\text{total}} < n_{k,\min}$ for any k **do**
8: Generate n_{batch} initial samples $\{\boldsymbol{\lambda}^i\}_{i=1}^{n_{\text{batch}}}$.
9: Append $\{\boldsymbol{\lambda}^i\}_{i=1}^{n_{\text{batch}}}$ to $\mathcal{S}$, and set $n \leftarrow n + n_{\text{batch}}$.
10: **for** $1 \le i \le n_{\text{batch}}$ **do**
11: Apply Q^p to classify to which $\{Q^{-1}(\mathcal{C}_k)\}_{k=1}^p$ sample $\boldsymbol{\lambda}^i$ belongs.
12: **end for**
13: **for** $1 \le k \le p$ **do**
14: $n_{k,\text{batch}} \leftarrow$ number of $\{\boldsymbol{\lambda}^i\}_{i=1}^{n_{\text{batch}}}$ in $Q^{-1}(\mathcal{C}_k)$.
15: $n_{k,\text{total}} \leftarrow n_{k,\text{total}} + n_{k,\text{batch}}$.
16: **end for**
17: **end while**
18: Create array of n scalar weights $(u_1, \ldots, u_n)$ as

$$u_i = \begin{cases} \dfrac{\boldsymbol{w}_k}{n_k} & \text{if } n_{k,\min} > 0, \\ 0 & \text{else,} \end{cases}$$

and set $\boldsymbol{u} = [u_1, \ldots, u_n]^\top$.
19: Construct weighted EDF, $F^{n,p}_{\text{init};\boldsymbol{u}}$, on sample set $\mathcal{S}$ as

$$F^{n,p}_{\text{init};\boldsymbol{u}}(\boldsymbol{\lambda}) = \sum_{i=1}^n \boldsymbol{u}_i \mathbb{I}(\boldsymbol{\lambda}^i \preceq \boldsymbol{\lambda}). \tag{3.5}$$

20: Return $F^{n,p}_{\text{init};\boldsymbol{u}}$. ▷ Output

$\boldsymbol{u} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n]^\top$, where $\boldsymbol{u}_i$ corresponds to sample $\boldsymbol{\lambda}^i$ and is defined as $\boldsymbol{u}_i = \frac{1}{n_k}\boldsymbol{w}_k$ when $\boldsymbol{\lambda}^i \in Q^{-1}(\mathcal{C}_k)$ and n_k is equal to the total number of parameter samples in $Q^{-1}(\mathcal{C}_k)$. We emphasize that we bin the samples in the output space, which is typically lower-dimensional than the parameter space. In addition, the computational cost of the naïve algorithm depends primarily on p, and the cost for adding samples is negligible. The method is described in detail below.

Observe that the push-forward of $F^{n,p}_{\text{init};\boldsymbol{u}}$ through the discrete map Q^p is, by construction, equal to $F^p_{\mu_{\mathcal{D}};\boldsymbol{w}}$ since we have

$$F^p_{\mu_{\mathcal{D}};\boldsymbol{w}}(\boldsymbol{q}) = \frac{1}{p}\sum_{k=1}^{p} \boldsymbol{w}_k \mathbb{I}(\boldsymbol{c}^k \preceq \boldsymbol{q}) = \frac{1}{p}\sum_{k=1}^{p} \sum_{\{i:Q(\boldsymbol{\lambda}^i)\in\mathcal{C}_k\}} \boldsymbol{u}_i \mathbb{I}(\boldsymbol{q}^i \preceq \boldsymbol{q}) = F^{n,p}_{\text{pred};\boldsymbol{u}}(\boldsymbol{q}). \tag{3.6}$$

Thus, when applying this approach, we denote by $F^{n,p}_{\text{pred};\boldsymbol{u}}$ the push-forward of $F^{n,p}_{\text{init};\boldsymbol{u}}$ via the map Q^p. As p and n increase, $F^{n,p}_{\text{pred};\boldsymbol{u}}$ and $F^{n,p}_{\text{init};\boldsymbol{u}}$ converge to the observed and update distribution, respectively, as shown in section 4.

We illustrate the weights obtained from the binning method, where the partition is generated using a regular grid with 35 bins, in the left plot of Figure 6. Comparing the right plots of Figures 5 and 6, it is clear the resulting weights from the binning method are closer to the density-based weights than the weights generated by the naïve method. Note that the binning method assigns samples in the same bin the same weights, which explains the stepwise nature of the right plot in Figure 6.

3.4. Construction of bins. The binning method we have developed is similar to the two-step nonparametric importance sampling methods in [29] and [24]. There, a proposal distribution is approximated using a nonparametric density method, and then classical importance sampling is performed. Here, we are using a histogram-like nonparametric density estimation to approximate the predicted density on the data space before performing an optimization-based change of measure.

If we construct the cells using a regular grid partitioning scheme and representative points are chosen as the geometric centers of the cells, the first step is exactly a histogram density approximation. When the support of the predicted density is irregular or the data space is very high dimensional, it may be more effective to use a data-driven partitioning scheme. For example, we can cluster the predicted samples using K-means clustering [4], and use variance-minimizing representative points. For

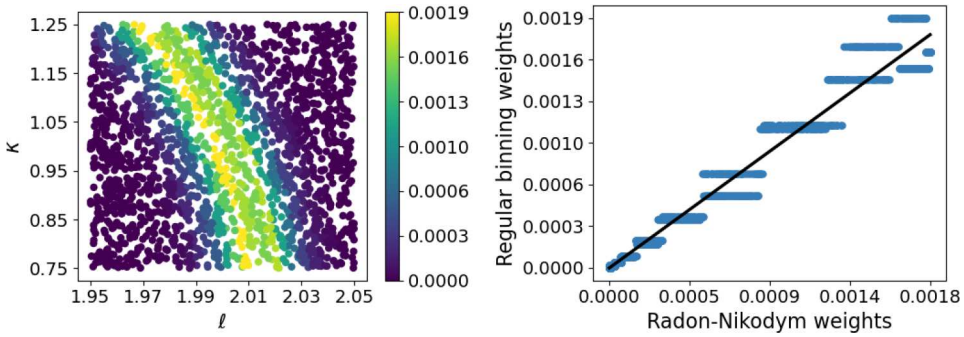

FIG. 6. *Weights resulting from the two-step partitioning method applied to initial samples on the left. Direct comparison to density-based weights on the right with the line indicating where perfect agreement occurs.*

the convergence results of this paper to hold, we require a single additional assumption on the clustering method.

Assumption 3.2. The binning distribution $F^{n,p}_{\mu_{\mathcal{D}};\boldsymbol{w}}$ converges as $n, p \to \infty$ to a distribution that is absolutely continuous with respect to the observed distribution.

For the illustrative example, Figure 7 shows the binning of the samples in the parameter space that result from using both a regular grid partition and a K-means clustering method to create the cells on the data space. The bins resulting from the K-means binning method still follow the contours of the mapping, as do the bins from the regular grid binning method, but they are more size-variable along the data-informed direction.

We plot the resulting weights obtained from the K-means clustering on the initial samples and compare to the density-based weights in the plots of Figure 8.

Visualizing the push-forward of the solution resulting from each method, as in Figure 9, we see that the binning method is not necessarily the L^2-optimal step function on the predicted samples. For this figure, in order to more clearly see differences in the results for each of the methods, we reran the heat equation experiment with $m = 10E4$ and $n = 2E2$, and p, the number of bins, equal to 10. It may, in fact, result in a greater L^2-distance between the resulting weighted EDF in the data space

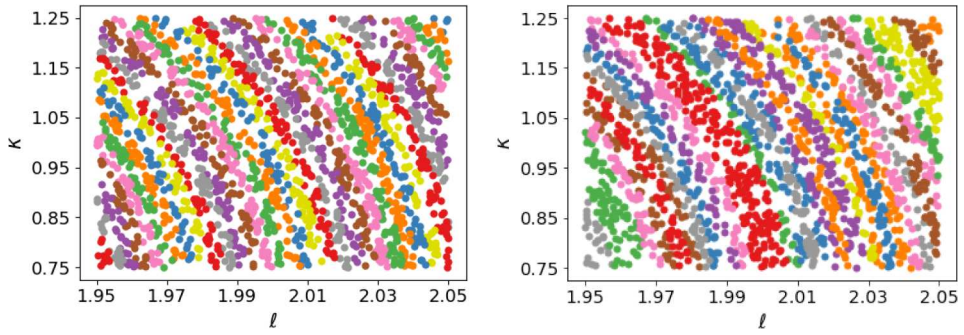

FIG. 7. *Binning of initial samples resulting from regular grid and K-means clustering partitioning methods using* 40 *bins. Although the bins are computed on the output space, here we are showing the samples binned in the input space to illustrate how samples in the same bin fall in the same contour event.*

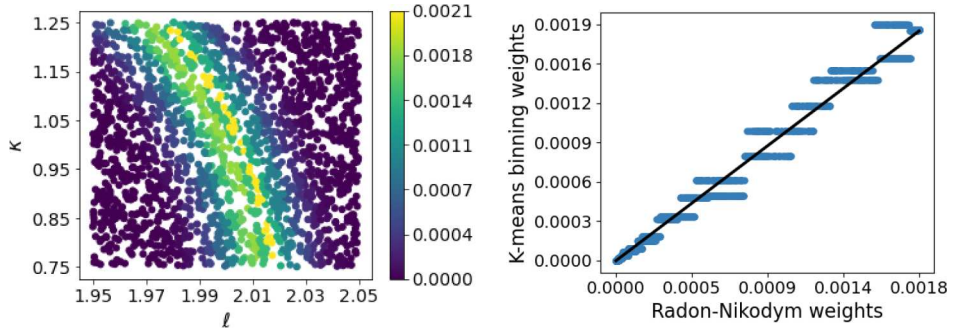

FIG. 8. *Weights resulting from the two-step partitioning method, where the partition is generated using a K-means classifier, applied to initial samples on the left. Direct comparison to density-based weights on the right.*

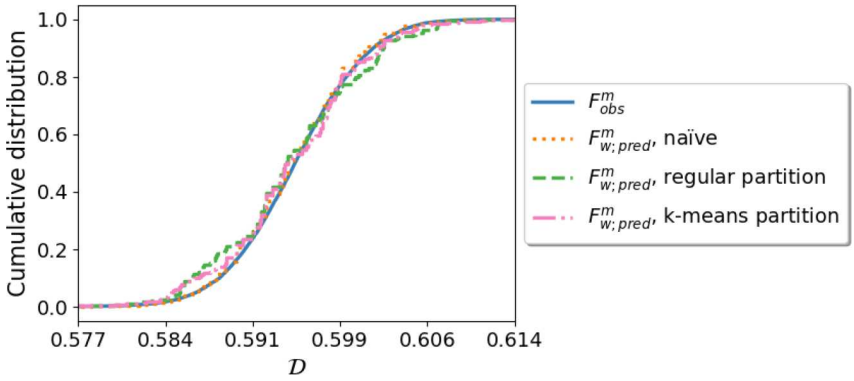

FIG. 9. *Push-forward distributions resulting from the various distribution-based methods. In both partition cases, we used* 10 *bins. In all cases,* $m = 10E4$, $n = 2E2$.

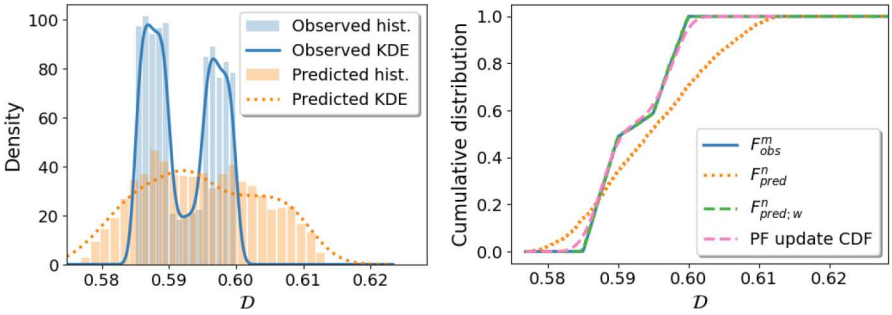

FIG. 10. *On the left: Histograms of observed and predicted samples and KDE estimations of the observed and predicted densities. On the right: DCI results for the binning and density-based methods. The CDF of the push-forward of the update computed from the density method ("PF update CDF" in the plot) is clearly less accurate than the push-forward of the binning-based solution.*

and the observed distribution than that produced by the naïve method. However, the result is still optimal for the weighted EDF over the partition centers, and as long as the partition is sufficient, it results in a relatively small loss in accuracy for the push-forward and remains data-consistent. The solution itself is stable and accurate compared to the naïve method.

Next, we consider the case where at least one of the distributions involved does not admit a density. In this case, the density-based solution does not technically exist and application of the density method will result in clear inaccuracies. For the heat equation example, we consider an observed distribution defined as a mixture of uniform distributions, $P_{\text{obs}} \sim 0.5\mathcal{U}((0.585, 0.59]) + 0.1\mathcal{U}((0.59, 0.595]) + 0.4\mathcal{U}((0.595, 0.6])$, resulting in a piecewise-linear distribution function. We use $m = 10E4$ observed samples and $n = 2E4$ input/predicted samples. The distributions involved and the result of the DCI density-based methods and binning-based methods are shown in Figure 10. The right plot illustrates the errors the density-based method produces around the endpoints of the subintervals associated with each uniform distribution (where a "density" exhibits discontinuities), whereas the binning-based method produces a solution that is indistinguishable to the observed distribution.

4. Theoretical results. The goal of this section is to summarize existing results and develop the necessary theory required to prove that the weighted EDF obtained via Algorithm 3.2 converges to the data-consistent solution. Section 4.1 provides the theoretical results for multivariate distribution functions. Specifically, we provide a theoretical connection between the L^p-convergence of multivariate distribution functions and the more commonly studied weak convergence. For the *univariate* case, L^1-convergence of *univariate* distribution functions is equivalent to convergence in the Kantorovich/Wasserstein distance metric and implies weak convergence [20]. The weak convergence of multivariate distribution functions subsequently implies convergence on the so-called P-continuity sets, where P is the probability measure associated with the limit of the distribution functions. Recalling that the binning sets utilized in Algorithm 3.2 are P_{pred}-continuity sets, we then prove that the weighted EDFs converge to the data-consistent solution in section 4.2.

4.1. Convergence of multivariate distribution functions. We start with a formal definition of weak convergence for probability measures defined on a measurable space constructed from a metric space S along with its Borel σ-algebra Σ.

DEFINITION 4.1. *A sequence of probability measures* $\{P_n\}_{n\in\mathbb{N}}$ *on* (S,Σ) *converges weakly to probability measure* P *if, for any* P*-continuity set* A, $\lim_{n\to\infty} P_n(A) \to P(A)$.

The following lemma relates this definition of weak convergence to the pointwise convergence of the corresponding multivariate distribution functions when $S=\mathbb{R}^d$.

LEMMA 4.2. *Let* F *be a distribution function on* $\mathbb{R}^d$ *corresponding to a probability measure* P, *and* $\{F_n\}_{n\in\mathbb{N}}$ *a sequence of distribution functions corresponding to probability measures* $\{P_n\}$. *If* $F_n(\boldsymbol{x}) \to F(\boldsymbol{x})$ *at all continuity points* $\boldsymbol{x}\in\mathbb{R}^d$ *of* F, *then* $P_n \to P$ *weakly.*

Proof. The proof of this is provided within Example 2.3 of [7]. ∎

The necessity of convergence at *every* continuity point of F in the above lemma hints at a technical hurdle we must overcome. Specifically, convergence of distribution functions in L^p is not pointwise convergence. However, convergence in L^p implies the existence of a subsequence $F_{n_k} \to F$ almost everywhere (a.e.). It is not immediately obvious that all the continuity points of F should belong to the a.e. set nor is it immediately clear how to relate a subsequential limit to a limit of the original sequence of distribution functions. We address these issues below.

The next lemma states that a multivariate CDF that converges a.e. must also converge at all continuity points of the limit distribution.

LEMMA 4.3. *Let* F *be a multivariate distribution function on* $\mathbb{R}^d$, *and* $\{F_n\}_{n\in\mathbb{N}}$ *a sequence of approximate distribution functions. Suppose that* $F_n \to F$ *a.e.; then* $F_n \to F$ *at all continuity points of* F.

Proof. See Appendix A. ∎

We now state the main result of this subsection that connects L^p-convergence of multivariate distribution functions to weak convergence of probability measures.

THEOREM 4.4. *Let* F *be a multivariate distribution function on* $\mathbb{R}^d$ *corresponding to a probability measure* P, *and* $\{F_n\}_{n\in\mathbb{N}}$ *a sequence of distribution functions corresponding to probability measures* $\{P_n\}_{n\in\mathbb{N}}$. *If* $F_n \to F$ *in* L^p, *where* $p\geq 1$, *then* $P_n \to P$ *weakly.*

Proof. See Appendix B. ∎

Theorem 4.4 immediately extends a result given in [2] involving the convergence of the L^2-based optimization method for reweighting the empirical distribution function in the data space that we exploit within Algorithm 3.3. Specifically, between Theorem 1 and Corollary 1 in [2], the authors demonstrate that the optimal L^2-convergence of distribution functions implies L^1-convergence. It is then commented that in the univariate case (i.e., for distribution functions defined on $\mathbb{R}$) this is equivalent to convergence in the Kantorovich/Wasserstein distance from which weak convergence follows under the additional assumption of bounded support of the proposal measure. By referring to Theorem 4.4 instead of the Kantorovich/Wasserstein distance, it immediately follows that the optimization approach developed in [2], and utilized in the QP portion of Algorithm 3.3, produces a sequence of distribution functions and corresponding probability measures that weakly converge in $\mathbb{R}^d$ for any $d \geq 1$. We summarize this observation as a corollary.

COROLLARY 4.5. *Let F be a distribution function on $\mathbb{R}^d$ corresponding to a probability measure P, and $\{F^n_{\boldsymbol{w}}\}$ a sequence of distribution functions resulting from applying Algorithm* 3.2 *to a set of n samples $\{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n\}$ from a distribution that is absolutely continuous with respect to P. Then, the probability distributions corresponding to the distribution functions $\{F^n_{\boldsymbol{w}}\}_{n\in\mathbb{N}}$ converge weakly to F as $n \to \infty$.*

4.2. Convergence of EDF approximations of solutions to the inverse problem. Returning to the inverse problem, the goal is to use Theorem 4.4 to show that under certain conditions the binning-based solution obtained from Algorithm 3.3 produces EDFs such that the associated probabilities of certain events on Λ converge to the probabilities given by the data-consistent solution P_{update}.

Recall from (3.6) that the push-forward of $F^{n,p}_{\text{init};\boldsymbol{u}}$ through the map Q^p is, by construction, equal to $F^p_{\mu_{\mathcal{D}};\boldsymbol{w}} = \frac{1}{p}\sum_{k=1}^p \boldsymbol{w}_i \mathbb{I}(\boldsymbol{c}^k \preceq \boldsymbol{q})$. Denoting by $F^{n,p}_{\text{pred};\boldsymbol{u}}$ the push-forward of $F^{n,p}_{\text{init};\boldsymbol{u}}$ via the map Q^p proves the following lemma.

LEMMA 4.6. *Let $\{\mathcal{C}_k\}_{k=1}^p$ be a P_{obs}-continuity partition of $\mathcal{D}$, with associated weights $\boldsymbol{w} \in \mathbb{R}^p$ generated by Algorithm* 3.3, *and let $\{\boldsymbol{\lambda}^j\}_{j=1}^n$ be the corresponding sequence of random samples in Λ with computed weights $\boldsymbol{u} \in \mathbb{R}^n$ from Algorithm* 3.3. *Together, these weights and samples define a predicted distribution $P^{n,p}_{\text{pred};\boldsymbol{u}}$ such that for every P_{obs}-continuity set D,*

$$\lim_{p\to\infty}\lim_{n\to\infty} P^{n,p}_{\text{pred};\boldsymbol{u}}(D) = P_{\text{obs}}(D).$$

In other words, $P^{n,p}_{\text{pred};\boldsymbol{u}} \to P_{\text{obs}}$ weakly as $p, n \to \infty$.

We now formally state the main result of this paper.

THEOREM 4.7. *Assume the predictability assumption holds, and let F_{update} and P_{update} denote the distribution and probability measure, respectively, defined by the density-based approach. Under the conditions of Lemma* 4.6, *the sequence of distribution functions $\{F^{n,p}_{\text{init};\boldsymbol{u}}\}_{n,p=1}^\infty$ and associated probability measures $\{P^{n,p}_{\text{init};\boldsymbol{u}}\}_{n,p=1}^\infty$ have the following properties.*

(i) *Defining the push-forward $F^{n,p}_{\text{pred};\boldsymbol{u}}$ as in Lemma* 4.6, *$F^{n,p}_{\text{pred};\boldsymbol{u}} \to F_{\text{obs}}$ weakly as $n, p \to \infty$.*

(ii) *Let $A \in \mathcal{B}_\Lambda$ with $Q(A) \in \mathcal{B}_{\mathcal{D}}$ a continuity set of P_{obs} and $P_{\text{obs}}(Q(A)) > 0$. If $\{\boldsymbol{\lambda}^j\}_{j=1}^n$ is an i.i.d. set drawn from P_{init}, then $P^{n,p}_{\text{init};\boldsymbol{u}}(A) \to P_{\text{update}}(A)$ as $n, p \to \infty$.*

Proof. See Appendix C. ∎

Part (i) of the above theorem states that the sequence of pullback distributions constructed via Algorithm 3.3 push forward to distributions that converge weakly to the observed distribution. Part (ii) of the above theorem describes the convergence on the parameter space of these pullback distributions. The condition that $P_{\text{obs}}(Q(A)) > 0$ is used to avoid conditioning on sets of zero measure.

5. Numerical results.

5.1. Convergence. To demonstrate convergence in both the data space $\mathcal{D}$ and the parameter space Λ, we return to the example introduced in section 2.4. We begin with a reference set in Λ, push this set through the map Q, and then consider the corresponding contour set. To this end, we take $A = [2.01, 2.02] \times [0.95, 1.0] \subset \Lambda$ to define $B = Q(A) \approx [0.59, 0.5936] \subset \mathcal{D}$, and $Q^{-1}(B) \subset \Lambda$ (see Figure 11).

Baseline estimates of probabilities for these sets are generated with the density-based DCI method and repeated trials utilizing distinct initial samples for each trial (the observed samples are not regenerated) to compute the average probabilities for a more accurate approximation. We set $m = 1E5$, $n = 1E5$, and utilize 10 trials to obtain $P_{\text{obs}}(B) \approx 0.26697$ and $P_{\text{update}}(A) \approx 0.01986$.

We now utilize the set B to illustrate the weak convergence for probabilities in $\mathcal{D}$ as $n, p \to \infty$ (see Theorem 4.7(i)). Sets of n (ranging from $1E3$ to $1E4$) initial samples are drawn. Following Algorithm 3.3, each set of initial samples is drawn by *appending* samples to the previous set. Once the initial samples are drawn, we create p bins (ranging from 20 to 160) using regular linear spacing on $\mathcal{D}$. We then compute the optimization-based weights and estimate the probability of B as the sum of weights on samples binned in B. Figures 12 and 13 summarize the mean error and standard deviation of these results taken over 20 trials. Trends are observed as we move from the upper left corner (corresponding to small n and p values) to the lower right corner (corresponding to larger n and p values) in both figures. Specifically, Figure 12 illustrates a trend toward more accurate approximations, while Figure 13 illustrates a trend towards reduced variance in these estimates. Note that the set A is quite small in comparison to the size of the parameter space, so that estimate variance is primarily driven by sample size n instead of bin number p. Consider, for example, that in one trial with $n = 1E4$, only 89 samples fall in A.

We now illustrate the weak convergence for probability in Λ in Theorem 4.7(ii) with the set A. Figure 14 shows that as $n, p \to \infty$ the absolute error decreases. The standard deviation of the results also decreases as $n, p \to \infty$ as shown in Figure 15.

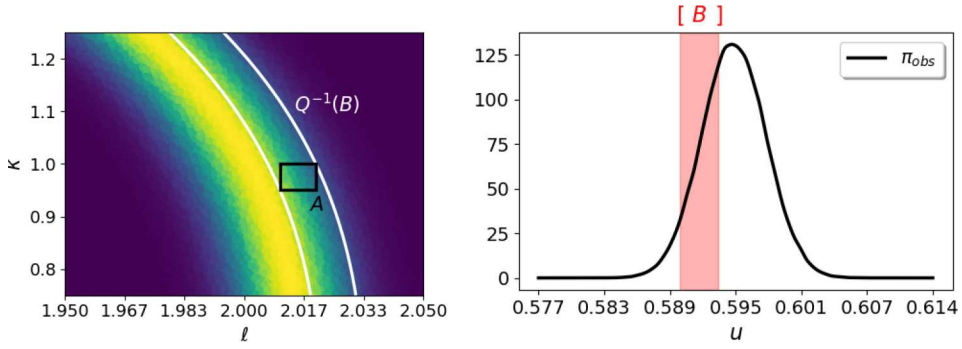

FIG. 11. *Sets A and $Q^{-1}(B)$ in Λ on the left. Set B on the right.*

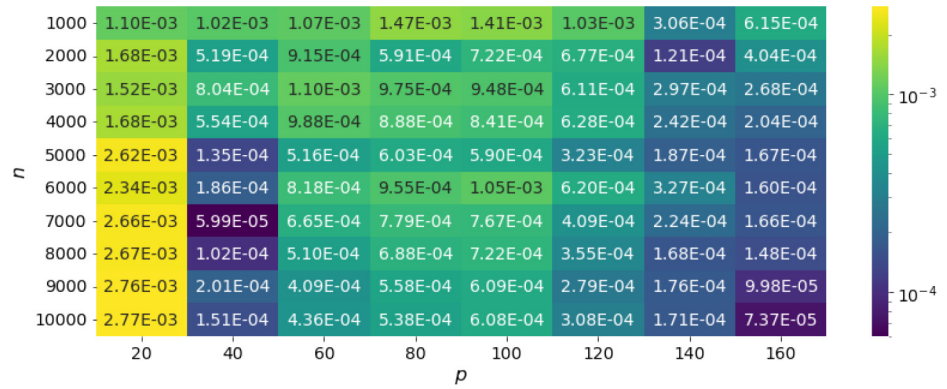

FIG. 12. *Absolute error between mean of* $P^{n,p}_{\text{pred};\boldsymbol{u}}(B)$ *over* 20 *trials and* $P_{\text{obs}}(B)$.

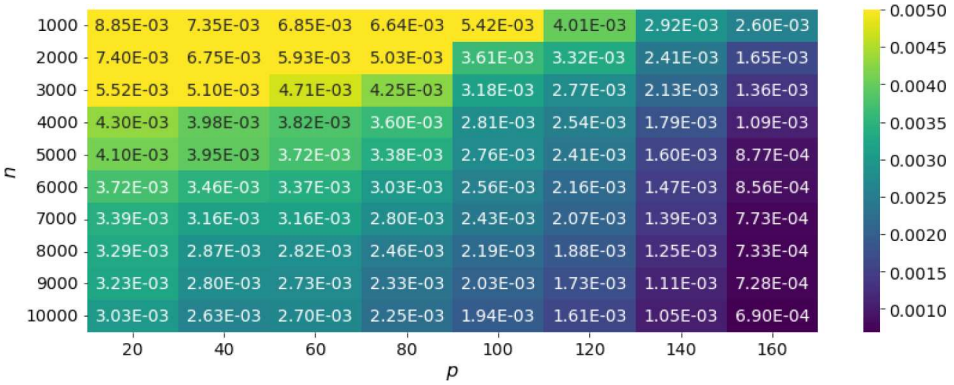

FIG. 13. *Standard deviation over* 20 *trials of* $P^{n,p}_{\text{pred};\boldsymbol{u}}(B)$.

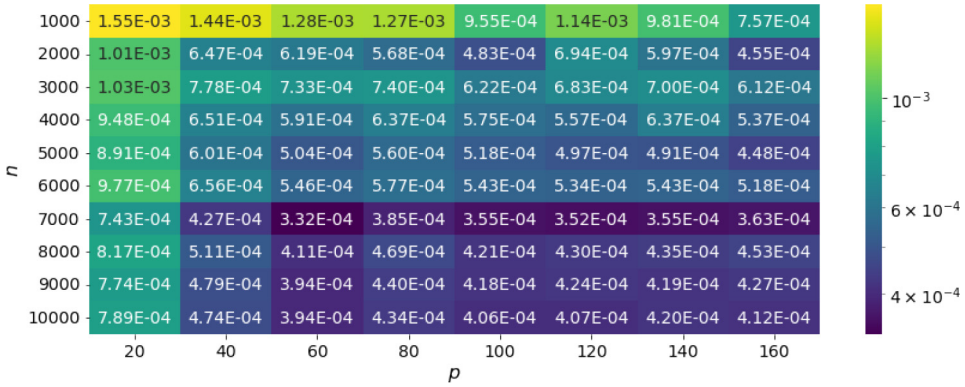

FIG. 14. *Absolute error between mean of* $P^{n,p}_{\text{init};\boldsymbol{u}}(A)$ *over* 100 *trials and* $P_{\text{obs}}(A)$.

5.2. Fluid flow through porous media. We now consider a 3-dimensional porous media model with a heterogeneous permeability field from the SPE10 data set [15]. The physical domain is a 3-dimensional slab: $\Omega = [0, 1200] \times [0, 2200] \times [0, 100]$, where the coordinates are given in feet. The model for single phase incompressible flow is given by

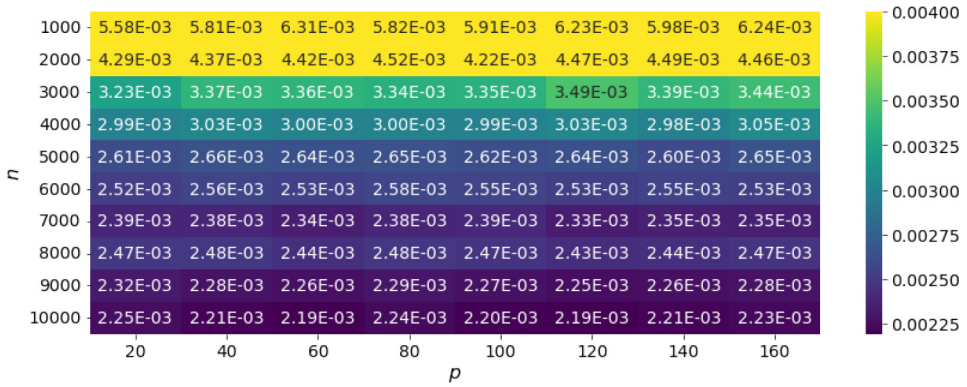

FIG. 15. *Standard deviation over* 20 *trials of* $P^{n,p}_{\text{pred};\boldsymbol{u}}(A)$.

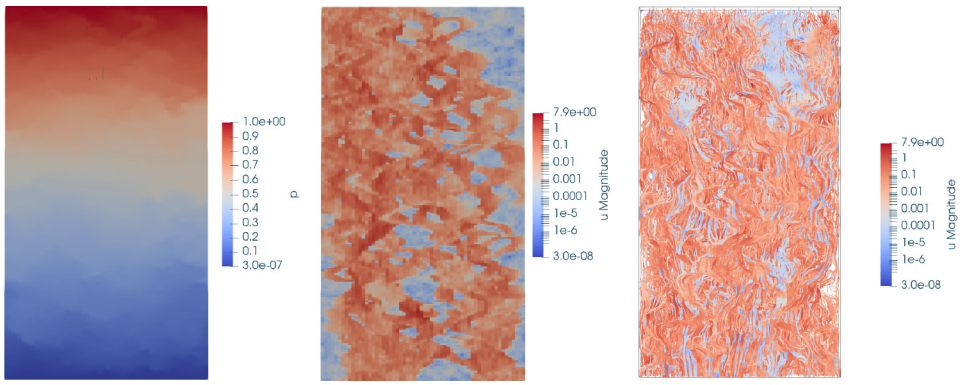

FIG. 16. *On the left: The numerical approximation of the pressure field. In the middle: The numerical approximation of the magnitude of the velocity field. On the right: The streamlines associated with the samples from the initial distribution.*

$$\begin{cases} \mathbf{u} = -\mathbf{K}\nabla p, & x \in \Omega, \\ \nabla \cdot \mathbf{u} = 0, & x \in \partial\Omega, \end{cases} \tag{5.1}$$

where $\mathbf{K}$ is the heterogeneous permeability field designed to mimic channelized flow in a subsurface reservoir. We apply a pressure drop from one side (high y values) to the other (low y values) by fixing the pressure along these faces and applying no-flow boundary conditions ($\mathbf{u} \cdot \mathbf{n} = 0$) on the remaining faces. The mesh is a uniform grid of $60 \times 220 \times 50$ elements, and we use a hybridized mixed formulation [10], which results in 6.6 million degrees of freedom, where 2.2 million are global/hybrid and 4.4 million are local/subgrid.

The input space for the inverse problem is defined by the 3-dimensional starting positions for a set of streamlines, generated using ParaView [1], which flow through the domain (generally in the decreasing y-direction). The output space is defined by the 3-dimensional final positions (we simplify the output space to 1 dimension as described below) for the streamlines. These final positions are given when the flow stops or the streamlines reach $y = 0$ (i.e., when they exit the domain). The pressure field, the magnitude of the velocity field, and the set of streamlines associated with the samples from the initial density are shown in Figure 16. The starting and ending points of

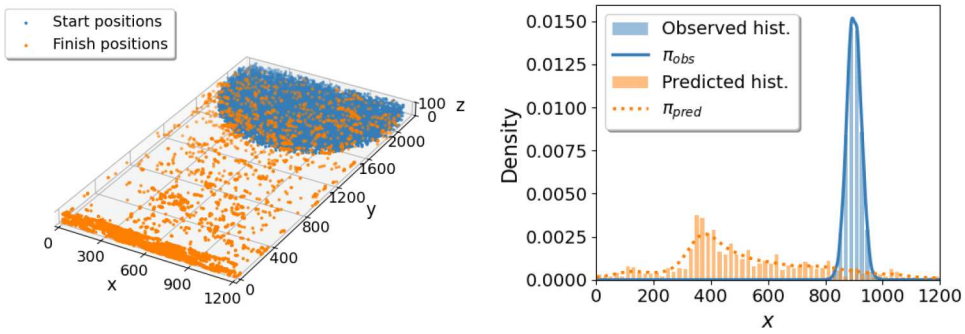

FIG. 17. *On the left: Starting positions for the initial samples shown in blue, and the endpoints for these samples shown in orange. On the right: The predicted samples and KDE shown in orange, and the observed samples and KDE shown in blue. (Color figure available online.)*

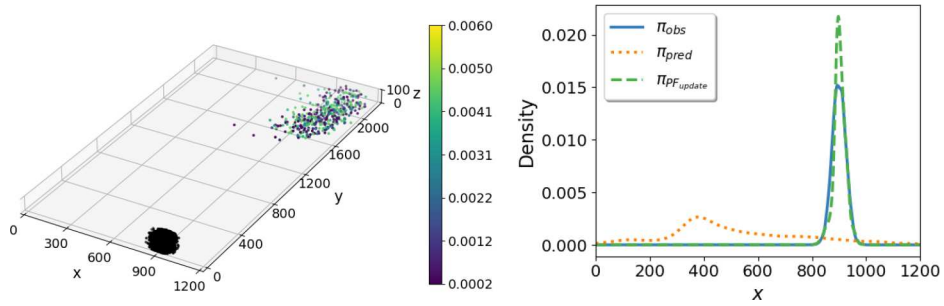

FIG. 18. *On the left: Update samples shown in color, observed shown in black. For readability, because the majority of the weights are close to zero, we remove samples whose weights are less than* 1E-4 *from the plot. The observed samples are* 1*-dimensional, so they are plotted with y and z both uniformly and randomly distributed from* 0 *to* 100*. On the right: Push-forward of results from the density method. (Color figure available online.)*

the initial samples are shown on the left of Figure 17. They are generated uniformly in the intersection of a 3-dimensional sphere and the domain. We are interested in the initial samples that either exit or get close to exiting the domain by the end of the simulation, that is, samples for which the y-values of their finish positions are less than 100. Selecting just these samples from the initial sample set leaves us with 6467 initial samples. We then focus on the x-position of the streamlines, resulting in a 1-dimensional output space. The observed distribution is $\mathcal{N}(900, 25)$, and we generate $1E4$ points from the observed distribution to serve as observed samples. We illustrate each output distribution using a KDE on the right of Figure 17.

We apply the density method along with rejection sampling to find a set of samples from the updated density. The mean r-value of the density method is approximately 1.07, indicating good, but not excellent, results. We show the accepted samples in the input space, as well as the push-forward of these samples in the data space, in Figure 18. We observe that the update samples are clustered toward a region with higher x-values. This makes sense considering the spatial properties of the problem. In the data space, we see that the results are also reasonable, although the push-forward of the update solution does not quite match the observed density.

Next, we apply the K-means binning method. Because the majority of the probability for the observed distribution is centered in a small region of the support of

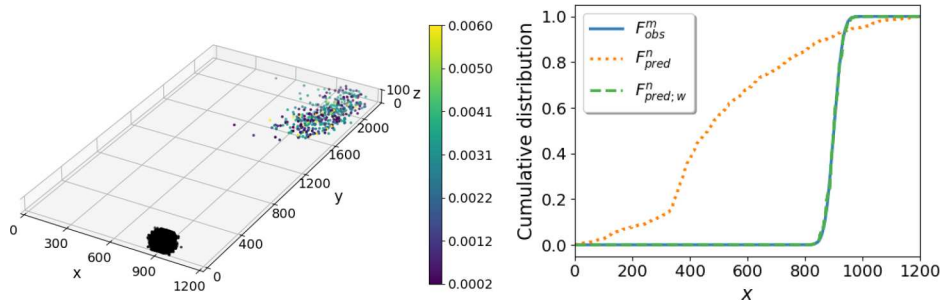

FIG. 19. *On the left: Solution weights from the K-means binning method shown in color; observed shown in black. For readability, because the majority of the weights are close to zero, we remove samples whose weights are less than* 1E-4 *from the plot. The observed samples are* 1*-dimensional, so they are plotted with* y *and* z *both uniformly and randomly distributed from* 0 *to* 100*. On the right: Push-forward of results from the K-means binning method. (Color figure available online.)*

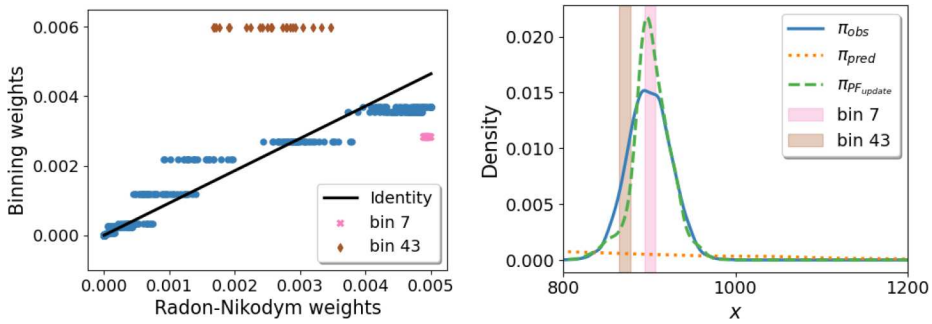

FIG. 20. *On the left: Comparison of the weights from the density and distribution-based methods, highlighting outlier bins. On the right: Zoomed-in plot of push-forward of density solution, highlighting outlier bin regions.*

the predicted, we require a relatively large number of bins, which informs our choice of $p = 100$ bins. The results, shown as weighted EDFs on the data space, as well as the weights plotted on the initial samples, are given in Figure 19. We observe that the binning approach produces larger weights for the samples that are mostly clustered toward a region of higher x-values, which is both what we expect and agrees with the results from the density solution. Unlike the density-based method that produced some clearly identifiable approximation errors in the push-forward of the updated density, the resulting CDFs show that the push-forward of the solution from the binning method closely matches the observed distribution on the data space.

We provide some more details regarding the issues with the density-based approximation seen in the right plot of Figure 18 that the binning method is able to overcome. First, we see that the majority of the predicted density is concentrated around $x = 400$ with long tails, while the observed density is narrow and centered along the far end of the right-tail of the predicted distribution. Subsequently, a large number of samples from the initial distribution need to be propagated through the model to obtain an accurate approximation of the predicted density in this region. In other words, the issue of accuracy with the density-based method in this case is essentially due to insufficient samples within the support of the observed distribution which impacts the KDE approximations. This is explored further in Figure 20. The

left plot of Figure 20 shows a comparison of the weights of the samples obtained from the two methods. The majority of the samples exhibit weights that fall roughly along the identity line, indicating a similarity of the computed weights between the methods in each region. However, there are two clear bins of samples that are outliers as highlighted in this plot. The right plot of Figure 20 zooms in on the observed density to highlight where these specific outlying bins occur in the data space. As discussed above, this is one of the tail-ends of the predicted density resulting in relatively few predicted samples falling within in the region where the observed density is concentrated. Subsequently, the predicted KDE used in the density method exhibits approximation errors in this region that are magnified by their presence in the denominator of the updated density. By comparing the KDE estimate of the push-forward of the updated density to the observed density, we see that bin 43 is associated with samples where the density-based method leads to an underestimation of the Radon–Nikodym weights, while bin 7 is associated with samples where the density-based method leads to an overestimation of the Radon–Nikodym weights. This is consistent with the position of the outlier bins relative to the identity line in the left plot of Figure 20.

6. Conclusions. In this work, we develop a distribution-based approach for solving data-consistent stochastic inverse problems. While previous approaches to solving such problems require approximations of events or densities, this work provides a rigorous methodology that applies to cases with limited observational or simulation/prediction data as well as situations where densities and events cannot be approximated effectively. This builds upon prior work by incorporating a binning approach in the data space to properly distribute probabilistic weights in the input space. Proofs of convergence of the distribution-based solution to the density-based solution as the number of samples and bins increase are provided.

In future work, we aim to formalize a theoretical proof of the stability of the pullback measure that is defined by the weighted EDF on the parameter space, as well as a theoretical proof to justify the generalization of the method to more arbitrary input sample sets (i.e., not just those generated according to a specified initial distribution). We are also interested in quantifying the effect of the model form errors on the weighted EDFs and the corresponding approximations to the pullback measure. Future directions will also include a utilization of this approach in the context of optimal experimental design and an exploration of the use of this approach to efficiently perform hierarchical Bayesian inference.

It is also worth noting that a somewhat related Bayesian approach referred to as data-free inference (DFI) is explored in [6, 22] to construct a family of posterior distributions exhibiting self-consistent correlations, in the absence of data, with respect to the available information. However, the notion of consistency in DFI centers around given statistical information and the goal is to utilize a Bayesian framework to construct an entire family of posteriors that are pooled together into a single distribution. A full comparison of the DCI and DFI solutions as well as the various numerical methods utilized to construct and interrogate these solutions is also a topic for future work.

7. Code. Code to reproduce the figures and examples in this paper is contained in the GitHub repository [5] for this paper.

Appendix A. Proof of Lemma 4.3. Before beginning the proof, we recall again the notions of "greater than" and "less than" in multivariate spaces. Specifically, we use the standard notation $\succeq$ and $\preceq$ to indicate componentwise greater than and componentwise less than, respectively; i.e., $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d) \succeq \boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_d)$ if and only if $\boldsymbol{x}_1 > \boldsymbol{y}_1, \ldots, \boldsymbol{x}_d > \boldsymbol{y}_d$. If a function is "nondecreasing," we mean this in the componentwise sense: $F(\boldsymbol{x}) \geq F(\boldsymbol{y})$ if $\boldsymbol{x} \succeq \boldsymbol{y}$.

Proof. Let $\boldsymbol{x}$ be a continuity point of F. Let $A = \{\boldsymbol{z} : F_n(\boldsymbol{z}) \to F(\boldsymbol{z})\}$, which we refer to as an "a.e. set" meaning the Lebesgue measure of the complement of A is zero. For $\delta > 0$, define $\boldsymbol{x} - \delta$ as the point $(\boldsymbol{x}_1 - \delta, \ldots, \boldsymbol{x}_d - \delta)$. Then, $\{\boldsymbol{z} : \boldsymbol{x} - \delta \preceq \boldsymbol{z} \preceq \boldsymbol{x}\}$ defines a d-dimensional cube of Lebesgue measure $\delta^d > 0$. Since A is an a.e. set, there exists an (uncountably) infinite number of points in $A \cap \{\boldsymbol{z} : \boldsymbol{x} - \delta \preceq \boldsymbol{z} \preceq \boldsymbol{x}\}$. It is therefore possible to choose a sequence $\{\boldsymbol{l}^i\}_{i=1} \subset A \cap \{\boldsymbol{z} : \boldsymbol{z} \preceq \boldsymbol{x}\}$ such that $\boldsymbol{l}^i \to \boldsymbol{x}$. An analogous argument implies we can also choose a sequence $\{\boldsymbol{u}^i\}_{i=1} \subset A \cap \{\boldsymbol{z} : \boldsymbol{z} \succeq \boldsymbol{x}\}$ such that $\boldsymbol{u}^i \to \boldsymbol{x}$. Because each F_n is nondecreasing in the componentwise sense, for any $n \in \mathbb{N}$ and any $i \in \mathbb{N}$, we have

$$F_n(\boldsymbol{l}^i) \leq F_n(\boldsymbol{x}) \leq F_n(\boldsymbol{u}^i).$$

The goal is to prove $F_n(\boldsymbol{x}) \to F(\boldsymbol{x})$. To do this, we first observe that

$$\liminf_{n\to\infty} F_n(\boldsymbol{l}^i) \leq \liminf_{n\to\infty} F_n(\boldsymbol{x}) \leq \liminf_{n\to\infty} F_n(\boldsymbol{u}^i)$$

and

$$\limsup_{n\to\infty} F_n(\boldsymbol{l}^i) \leq \limsup_{n\to\infty} F_n(\boldsymbol{x}) \leq \limsup_{n\to\infty} F_n(\boldsymbol{u}^i).$$

Now, as $n \to \infty$, $F_n(\boldsymbol{l}^i) \to F(\boldsymbol{l}_i)$ because $\boldsymbol{l}^i \in A$ for each i, and similarly $F_n(\boldsymbol{u}^i) \to F(\boldsymbol{u}^i)$ for each i. Thus,

$$F(\boldsymbol{l}^i) = \lim_{n\to\infty} F_n(\boldsymbol{l}^i) \leq \liminf_{n\to\infty} F_n(\boldsymbol{x})$$

and

$$\limsup_{n\to\infty} F_n(\boldsymbol{x}) \leq \lim_{n\to\infty} F_n(\boldsymbol{u}^i) = F(\boldsymbol{u}^i).$$

As $i \to \infty$, by construction we have $\boldsymbol{l}^i \to \boldsymbol{x}$ and $\boldsymbol{u}^i \to \boldsymbol{x}$, and $\boldsymbol{x}$ is a continuity point of F, so we have

$$\lim_{i\to\infty} F(\boldsymbol{l}^i) = F(\boldsymbol{x}) \leq \lim_{i\to\infty} \left(\liminf_{n\to\infty} F_n(\boldsymbol{x})\right) = \liminf_{n\to\infty} F_n(\boldsymbol{x})$$

and

$$\lim_{i\to\infty} \left(\limsup_{n\to\infty} F_n(\boldsymbol{x})\right) = \limsup_{n\to\infty} F_n(\boldsymbol{x}) \leq F(\boldsymbol{x}) = \lim_{i\to\infty} F(\boldsymbol{u}^i).$$

Thus, we have found

$$F(\boldsymbol{x}) \leq \liminf_{n\to\infty} F_n(\boldsymbol{x}) \leq \limsup_{n\to\infty} F_n(\boldsymbol{x}) \leq F(\boldsymbol{x})$$

which implies that $\lim_{n\to\infty} F_x(\boldsymbol{x})$ exists and is equal to $F(\boldsymbol{x})$. Since $\boldsymbol{x}$ was an arbitrary continuity point of F, the conclusion follows. ∎

Appendix B. Proof of Theorem 4.4. Before we prove this theorem, we recall a useful result connecting the convergence of subsequences to the original sequence that is straightforward to prove but can seem confusing when used without context.

LEMMA B.1. *Let $\{p_n\}_{n\in\mathbb{N}}$ be a sequence of real numbers. If there exists $p \in \mathbb{R}$ such that for any subsequence $\{p_{n_k}\}_{k\in\mathbb{N}}$ there exists a further subsequence $\{p_{n_{k_l}}\}_{l\in\mathbb{N}}$ that converges to p, then $p_n \to p$.*

While Lemma B.1 is a standard result in analysis, the proof can be difficult to find as it is often left as an exercise. We provide the proof below for ease of reference.

Proof. Suppose that (p_n) does not converge to p. Then there exists a $\varepsilon > 0$ such that for every $N \in \mathbb{N}$ there exists $n \geq N$ such that $|p_n - p| \geq \varepsilon$. Choose such an $\varepsilon > 0$, and inductively construct a subsequence (p_{n_k}) as follows. For $k = 1$, $n_1 \geq 1$ such that $|p_{n_1} - p| \geq \varepsilon$; for $k = 2$, choose $n_2 \geq n_1 + 1$ such that $|p_{n_2} - p| \geq \varepsilon$. Assume that we have chosen the first k terms in the sequence, and then choose $n_{k+1} \geq n_k$ such that $|p_{n_{k+1}} - p| \geq \varepsilon$. This defines a subsequence (p_{n_k}) such that $|p_{n_k} - p| \geq \varepsilon$ for all $k \in \mathbb{N}$. By assumption, there must exist a further subsequence $(p_{n_{k_\ell}})$ such that $p_{n_{k_\ell}} \to p$, but by construction all terms in any subsequence are at least a ε distance from p, a contradiction. ∎

We can finally proceed in proving Theorem 4.4 below.

Proof. Let $\{F_{n_k}\}_{k\in\mathbb{N}}$ be a subsequence of $\{F_n\}_{n\in\mathbb{N}}$. Since $F_n \to F$ in L^p, $F_{n_k} \to F$ in L^p as well. Thus, there exists a further subsequence $\{F_{n_{k_l}}\}_{l\in\mathbb{N}}$ that converges a.e. to F. By Lemma 4.3, for any continuity point $\boldsymbol{x}$ of F, the subsubsequence $F_{n_{k_\ell}}(\boldsymbol{x}) \to F(\boldsymbol{x})$. By Lemma B.1, $F_n(\boldsymbol{x}) \to F(\boldsymbol{x})$ at all continuity points of F. Finally, by Lemma 4.2, $P_n \to P$ weakly. ∎

Appendix C. Proof of Theorem 4.7.

Proof. The proof of (i) follows immediately from Lemma 4.6.

To prove (ii), first we let $C_A := Q^{-1}(Q(A))$ denote the contour event in $\mathcal{B}_\Lambda$ induced by the set A. We apply the law of total probability to write

$$P^{n,p}_{\text{init};\boldsymbol{u}}(A) = P^{n,p}_{\text{init};\boldsymbol{u}}(A \,|\, C_A) P^{n,p}_{\text{init};\boldsymbol{u}}(C_A) + P^{n,p}_{\text{init};\boldsymbol{u}}(A \,|\, C_A^c) P^{n,p}_{\text{init};\boldsymbol{u}}(C_A^c),$$

where C_A^c denotes the complement of C_A. Since $A \subset C_A$, A and C_A^c are disjoint, which implies $P^{n,p}_{\text{init};\boldsymbol{u}}(A \,|\, C_A^c) = 0$. Since $P^{n,p}_{\text{init};\boldsymbol{u}}(C_A) = P^{n,p}_{\text{pred};\boldsymbol{u}}(Q(A))$, we have

$$P^{n,p}_{\text{init};\boldsymbol{u}}(A) = P^{n,p}_{\text{init};\boldsymbol{u}}(A \,|\, C_A) P^{n,p}_{\text{pred};\boldsymbol{u}}(Q(A)).$$

From (i), we have $P^{n,p}_{\text{pred};\boldsymbol{u}}(Q(A)) \to P_{\text{obs}}(Q(A)) > 0$ as $n, p \to \infty$. We use this along with the method of n being chosen as a function of p in Algorithm 3.3 to observe that for sufficiently large p, $P^{n,p}_{\text{pred};\boldsymbol{u}}(Q(A)) > 0$. By the definition of C_A, it follows that both $P_{\text{update}}(C_A) > 0$ and, for sufficiently large p, $P^{n,p}_{\text{pred};\boldsymbol{u}}(C_A) > 0$. Because the samples $\{\boldsymbol{\lambda}^j\}_{j=1}^n$ used to construct $P^{n,p}_{\text{init};\boldsymbol{u}}$ are i.i.d., the strong law of large numbers implies $P^{n,p}_{\text{init};\boldsymbol{u}}(A \,|\, C_A) \to P_{\text{init}}(A \,|\, C_A)$. Since P_{update} and P_{init} have the same conditional probabilities on contour events via the disintegration theorem, we have

$$\lim_{n,p\to\infty} P^{n,p}_{\text{init};\boldsymbol{u}}(A \,|\, C_A) \to P_{\text{init}}(A \,|\, C_A) = P_{\text{update}}(A \,|\, C_A).$$

It follows that

$$P^{n,p}_{\text{init};\boldsymbol{u}}(A) \to P_{\text{update}}(A \,|\, C_A) P_{\text{obs}}(Q(A)) \text{ as } n, p \to \infty.$$

The result follows by recognizing that $P_{\text{obs}}(Q(A)) = P_{\text{update}}(C_A)$ and applying the law of total probability to P_{update} analogously to how it was applied to $P^{n,p}_{\text{init};\boldsymbol{u}}$ above. ∎

Reproducibility of computational results. This paper has been awarded the "SIAM Reproducibility Badge: Code and data available" as a recognition that the authors have followed reproducibility principles valued by SISC and the scientific computing community. Code and data that allow readers to reproduce the results in this paper are available at https://zenodo.org/doi/10.5281/zenodo.10676816.

REFERENCES

[1] J. Ahrens, B. Geveci, and C. Law, *ParaView: An end-user tool for large-data visualization*, in Visualization Handbook, Elsevier, 2005.

[2] S. Amaral, D. L. Allaire, and K. E. Willcox, *Optimal L^2-norm empirical importance weights for the change of probability measure*, Stat. Comput., 27 (2017), pp. 625–643, https://doi.org/10.1007/s11222-016-9644-3.

[3] M. S. Andersen, J. Dahl, and L. Vandenberghe, *CVXOPT: A Python Package for Convex Optimization*, version 1.3.2. Available at https://cvxopt.org, 2023.

[4] D. Arthur and S. Vassilvitskii, *k-means++: The advantages of careful seeding*, in Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '07), New Orleans, LA, 2007, ACM, New York, SIAM, Philadelphia, 2007, pp. 1027–1035, https://dl.acm.org/doi/10.5555/1283383.1283494.

[5] K. Bergstrom, T. Butler, and T. Wildey, *kirana-bergstrom/DCI-distributions: v*, 2024, https://zenodo.org/doi/10.5281/zenodo.10676816.

[6] R. D. Berry, H. N. Najm, B. J. Debusschere, Y. M. Marzouk, and H. Adalsteinsson, *Data-free inference of the joint distribution of uncertain model parameters*, J. Comput. Phys., 231 (2012), pp. 2180–2198, https://doi.org/10.1016/j.jcp.2011.10.031.

[7] P. Billingsley, *Convergence of Probability Measures*, John Wiley & Sons, 1999.

[8] P. Billingsley, *Probability and Measure*, anniversary ed., John Wiley & Sons, 2012.

[9] S. P. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.

[10] F. Brezzi, and FM. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[11] T. Butler, J. Jakeman, and T. Wildey, *Combining push-forward measures and Bayes' rule to construct consistent solutions to stochastic inverse problems*, SIAM J. Sci. Comput., 40 (2018), pp. A984–A1011, https://doi.org/10.1137/16M1087229.

[12] T. Butler, D. Estep, S. Tavener, C. Dawson, and J. J. Westerink, *A measure-theoretic computational method for inverse sensitivity problems* III*: Multiple quantities of interest*, SIAM/ASA J. Uncertain. Quantif., 2 (2014), pp. 174–202, https://doi.org/10.1137/130930406.

[13] J. Chang and D. Pollard, *Conditioning as disintegration*, Stat. Neerl., 51 (1997), pp. 287–317, https://doi.org/10.1111/1467-9574.00056.

[14] Y.-C. Chen, *A tutorial on kernel density estimation and recent advances*, Biostatist. Epidemiol., 1 (2017), pp. 161–187, https://doi.org/10.1080/24709360.2017.1396742.

[15] M. Christie and M. Blunt, *Tenth SPE comparative solution project: A comparison of upscaling techniques*, in SPE Reservoir Simulation Symposium, 2001, SPE-66599-MS, https://doi.org/doi:10.2118/66599-MS.

[16] G. Demoment and Y. Goussard, *Inversion within the probabilistic framework*, in Bayesian Approach to Inverse Problems, J. Idier, ed., John Wiley & Sons, Hoboken, ISTE, London, 2008, pp. 59–77.

[17] L. Devroye and L. Gyorfi, *Nonparametric Density Estimation: The View*, Wiley, New York, 1985.

[18] M. D. Escobar and M. West, *Bayesian density estimation and inference using mixtures*, J. Amer. Statist. Assoc., 90 (1995), pp. 577–588, https://doi.org/10.1080/01621459.1995.10476550.

[19] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*, 2nd ed., Wiley, 1999.

[20] A. L. Gibbs and F. E. Su, *On choosing and bounding probability metrics*, Int. Stat. Rev., 70 (2002), pp. 419–435, https://doi.org/10.2307/1403865.

[21] A. Kirsch, *An Introduction to the Mathematical Theory of Inverse Problems*, Springer, New York, 2011.
[22] H. N. Najm, R. D. Berry, C. Safta, K. Sargsyan, and B. J. Debusschere, *Data-free inference of uncertain parameters in chemical models*, Int. J. Uncertain. Quantif., 4 (2014), pp. 111–132, https://doi.org/10.1615/Int.J.UncertaintyQuantification.2013005679.
[23] R. M. Neal, *Markov chain sampling methods for Dirichlet process mixture models*, J. Comput. Graph. Statist., 9 (2000), pp. 249–265, https://doi.org/10.2307/1390653.
[24] J. C. Neddermeyer, *Computationally efficient nonparametric importance sampling*, J. Amer. Statist. Assoc., 104 (2009), pp. 788–802, https://doi.org/10.1198/jasa.2009.0122.
[25] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, *Normalizing flows for probabilistic modeling and inference*, J. Mach. Learn. Res., 22 (2021), 57.
[26] M. Sanghvi, P. Honarmandi, V. Attari, T. Duong, R. Arroyave, and D. L. Allaire, *Uncertainty propagation via probability measure optimized importance weights with application to parametric materials models*, in AIAA Scitech 2019 Forum, 2019, AIAA 2019-0967, https://arc.aiaa.org/doi/abs/10.2514/6.2019-0967.
[27] A. M. Stuart, *Inverse problems: A Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559, https://doi.org/10.1017/S0962492910000061.
[28] C. R. Vogel, *Computational Methods for Inverse Problems*, SIAM, Philadelphia, 2002, https://doi.org/10.1137/1.9780898717570.
[29] P. Zhang, *Nonparametric importance sampling*, J. Amer. Statist. Assoc., 91 (1996), pp. 1245–1253, https://doi.org/10.2307/2291743.