Optimal design of large-scale Bayesian linear inverse problems under reducible model uncertainty: good to know what you don't know*

Alen Alexanderian†, Noemi Petra‡, Georg Stadler§, and Isaac Sunseri†

Abstract. We consider optimal design of infinite-dimensional Bayesian linear inverse problems governed by partial differential equations that contain secondary *reducible* model uncertainties, in addition to the uncertainty in the inversion parameters. By reducible uncertainties we refer to parametric uncertainties that can be reduced through parameter inference. We seek experimental designs that minimize the posterior uncertainty in the primary parameters, while accounting for the uncertainty in secondary parameters. We accomplish this by deriving a marginalized A-optimality criterion and developing an efficient computational approach for its optimization. We illustrate our approach for estimating an uncertain time-dependent source in a contaminant transport model with an uncertain initial state as secondary uncertainty. Our results indicate that accounting for additional model uncertainty in the experimental design process is crucial.

Key words. Optimal experimental design, Bayesian inference, inverse problems, model uncertainty, sensor placement, sparsified designs.

AMS subject classifications. 65C60, 62K05, 62F15, 35R30.

1. Introduction.

An inverse problem uses measurement data and a mathematical model to estimate a set of uncertain model parameters. An experimental design specifies the strategy for collecting measurement data. For example, in inverse problems where measurement data are collected using sensors, an experimental design specifies the placement of the sensors. This is the setting considered in the present work. Optimal experimental design (OED) [5, 37] refers to the task of determining an experimental setup such that the measurements are most informative about the underlying parameters. This is particularly important in situations where experiments are costly or time-consuming, and thus only a small number of measurements can be collected. In addition to the parameters estimated by the inverse problem, the governing mathematical models often involve simplifications, approximations, or modeling assumptions, resulting in additional uncertainty. These additional uncertainties must be taken into account in the experimental design process; failing to do so could result in suboptimal designs.

We distinguish between two types of uncertainties: *reducible* and *irreducible* [32]. Reducible uncertainties, also referred to as epistemic uncertainties, are those that can be reduced through parameter inference. In contrast, irreducible uncertainties, also known as aleatoric uncertainties, are inherent to the model and are impractical or impossible to reduce through parameter inference. In this article, we aim at computing optimal experimental designs in the presence of *reducible* model uncertainty.

*Submitted to the editors June 23, 2020.

Funding: Supported in part by US National Science Foundation DMS #1723211, #1654311, and #1745654.

†Department of Mathematics, North Carolina State University, Raleigh, NC, USA

‡Department of Applied Mathematics, University of California, Merced, CA, USA

§Courant Institute of Mathematical Sciences, New York University, New York, NY, USA

In what follows, we consider the model

$$\boldsymbol{y} = \mathcal{E}(m, b) + \boldsymbol{\eta}, \tag{1.1}$$

where $\boldsymbol{y}$ is a vector of measurement data, (m, b) a pair of uncertain parameter vectors or functions, $\mathcal{E}$ a model that maps (m, b) to measurements, and $\boldsymbol{\eta}$ a random vector that models additive measurement errors. Herein, m is the parameter of primary interest, which we seek to infer, and b represents additional uncertain parameters. We assume m and b are elements of infinite-dimensional Hilbert spaces. Furthermore, we assume that the uncertainty in b is reducible. Thus, we can formulate an inverse problem to estimate both m and b. However, when designing experiments to solve the inverse problem, our main interest is reducing the uncertainty in m. We achieve this by finding sensor placements that minimize the posterior uncertainty in m, while taking into account the uncertainty in b. This results in an OED problem in which we minimize the marginal posterior uncertainty in m.

In this article, we focus on the case of a model that is linear in m and b and is of the form:

$$\mathcal{E}(m, b) = \mathcal{F}m + \mathcal{G}b. \tag{1.2}$$

Here, $\mathcal{F}$ and $\mathcal{G}$ are bounded linear transformations from suitably defined Hilbert spaces to the space of measurement data. This models, for example, a linear inverse problem with uncertain volume or boundary terms. The mathematical foundations for Bayesian inversion and design of experiments in this context are discussed in section 2.

Examples for secondary uncertainties are initial conditions, boundary conditions that are introduced into a model due to the necessity to truncate a computational domain, or unknown forcing or source terms in a real world system that are only incorporated approximately in the mathematical model. When designing experiments, failure to properly account for these secondary uncertainties may result in suboptimal experimental designs. For instance, not taking into account secondary uncertainties in the mathematical model may result in sensors being located close to uncertain sources, resulting in observations that can provide biased information on the parameter of primary interest. If one aims at finding designs that are optimal for both primary and secondary uncertain parameters, the design is likely to be suboptimal for inference of the primary parameter.

Related work. In many inverse problems, one has model uncertainties in addition to the inversion parameters. A robust parameter inversion strategy must account for such additional model uncertainties; see [4, 10, 19–21, 26, 27] for a small sample of the literature addressing such issues. This work is about A-optimal experimental design for Bayesian linear inverse problems governed by partial differential equations (PDEs) with model uncertainties. For a review of the literature on optimal design of inverse problems governed by computationally intensive models, we refer to [2]. Here, we mainly review related work on optimal design of linear inverse problems. The articles [3, 14, 15] present methods for large-scale ill-posed linear inverse problems. Specifically, the present article builds on [3], which focuses on A-optimal experimental design of infinite-dimensional Bayesian linear inverse problems.

Recent work also considers A-optimal design of infinite-dimensional Bayesian linear inverse problems with model uncertainties [22]. The key difference to the present work is that [22] considers OED for inverse problems governed by models with *irreducible* uncertainties and

formulates the OED problem as one of optimization under uncertainty. In contrast, in this work we consider OED under *reducible* model uncertainties and propose a formulation that aims at minimizing the marginal posterior variance of the primary parameters. By combining primary and secondary uncertainties, the problem considered in this work can formally be written as goal-oriented OED problem, as studied in [6]. However, taking a model uncertainty perspective and considering infinite-dimensional primary and secondary uncertain parameters require a tailored approach that distinguishes primary and secondary uncertainties.

Other related efforts include [13, 17, 30]. In [13], the authors present an adaptive A-optimal design strategy for linear dynamical systems. OED for linear inverse problems with linear equality and inequality constraints is addressed in [30]. This results in OED with an effectively nonlinear inverse problem for which the authors propose an approach based on Bayes risk minimization. In [17], the authors present an approach for A-optimal design of infinite-dimensional Bayesian linear inverse problems using ideas from randomized subspace iteration and reweighted ℓ_1-minimization.

Contributions. This article makes the following contributions to the state-of-the-art in OED for large-scale linear inverse problems. (i) We provide a mathematical formulation of OED under reducible model uncertainty and show how the OED problem can be reformulated to take advantage of the often low dimensionality of the measurement space (see section 3); in particular, our formulation eliminates the need for trace estimation in the discretized parameter space. (ii) We develop a scalable computational framework for solving the class of OED problems under study (see section 4). Specifically, the computational complexity of our methods, in terms of the number of PDE solves, does not grow with the dimension of the discretized primary and secondary parameters. (iii) We present illustrative numerical results, in context of a contaminant transport inverse problem (see section 5 and section 6) where we seek to estimate an unknown source term, but have additional uncertainty in the initial state. Our numerical experiments examine different aspects of our proposed framework, and elucidate the importance of incorporating additional model uncertainties in the OED problem.

2. Bayesian inverse problems governed by models with reducible uncertainties.

After introducing basic notation in subsection 2.1, we present preliminaries regarding Gaussian measures on Hilbert spaces in subsection 2.2. Next, we outline the setup of Bayesian linear inverse problems with additional reducible model uncertainties in infinite-dimensions (subsection 2.3) as well as in discretized form (subsection 2.4). We discuss basics on optimal design of such inverse problems in subsection 2.5.

2.1. Notation.

Herein we consider a probability space $(\Omega, \mathfrak{A}, \mathbb{P})$, where Ω is a sample space, $\mathfrak{A}$ a sigma-algebra on Ω, and $\mathbb{P}$ is a probability measure. Given a Hilbert space $\mathcal{X}$, we denote by $\mathfrak{B}(\mathcal{X})$ the Borel sigma-algebra on $\mathcal{X}$. A Gaussian measure on $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$, with mean $\bar{z} \in \mathcal{X}$ and covariance operator $\mathcal{C} : \mathcal{X} \to \mathcal{X}$, is denoted by $\mathcal{N}(\bar{z}, \mathcal{C})$. We also recall that for a random variable $Z : (\Omega, \mathfrak{A}, \mathbb{P}) \to (\mathcal{X}, \mathfrak{B}(\mathcal{X}))$, its law is a Borel measure $\mathcal{L}_Z$ on $\mathcal{Y}$, that satisfies $\mathcal{L}_Z(A) = \mathbb{P}(Z \in A)$ for every $A \in \mathfrak{B}(\mathcal{Y})$ [39]. Also, for a linear transformation $T : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y}$ is another Hilbert space, we denote the adjoint by T^*.

2.2. Marginals of Gaussian measures.

Here we discuss some preliminaries regarding Gaussian measures and Gaussian random variables taking values in Hilbert spaces. First

we record the following known result about the law of a linear transformation of a Hilbert space-valued Gaussian random variable, which we prove for completeness.

Lemma 2.1. *Let $\mathcal{X}$ and $\mathcal{Y}$ be infinite-dimensional Hilbert spaces. Suppose $Z : \Omega \to \mathcal{X}$ is a Gaussian random variable with law $\mu = \mathcal{N}(\bar{z}, \mathcal{C})$. Consider the random variable $Y = TZ$, where $T : \mathcal{X} \to \mathcal{Y}$ is a bounded linear transformation. Then, $Y : (\Omega, \mathfrak{A}, \mathbb{P}) \to (\mathcal{Y}, \mathfrak{B}(\mathcal{Y}))$ is a Gaussian random variable with law $\nu = \mathcal{N}(T\bar{z}, T\mathcal{C}T^*)$.*

Proof. Using [11, Proposition 1.18], we know that the law of the random variable $T : (\mathcal{X}, \mathfrak{B}(\mathcal{X}), \mu) \to (\mathcal{Y}, \mathfrak{B}(\mathcal{Y}))$ is given by $\mu \circ T^{-1} = \mathcal{N}(T\bar{z}, T\mathcal{C}T^*) = \nu$. To complete the proof we show $\mathcal{L}_Y = \nu$. Namely, for every $A \in \mathfrak{B}(\mathcal{Y})$,

$$\mathcal{L}_Y(A) = \mathbb{P}(Y \in A) = \mathbb{P}(TZ \in A) = \mathbb{P}(Z \in T^{-1}(A)) = \mu(T^{-1}(A)) = \nu(A). \qquad \blacksquare$$

Consider a Hilbert space $\mathcal{V} = \mathcal{V}_1 \times \mathcal{V}_2$, where $\mathcal{V}_1$ and $\mathcal{V}_2$ are real, separable, infinite-dimensional Hilbert spaces with inner products $\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$, respectively. An element $z \in \mathcal{V}$ is of the form $z = (z_1, z_2)$ with $z_1 \in \mathcal{V}_1$ and $z_2 \in \mathcal{V}_2$, respectively. We assume that $\mathcal{V}$ is equipped with the natural inner product

$$\langle\!\langle x, y \rangle\!\rangle = \langle x_1, y_1 \rangle_1 + \langle x_2, y_2 \rangle_2, \quad x, y \in \mathcal{V}.$$

Let $Z : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{V}, \mathcal{B}(\mathcal{V}), \mu)$ be a Gaussian random variable with law $\mu = \mathcal{N}(\bar{z}, \mathcal{C})$. The marginal laws of Z can be defined analogously to the finite-dimensional setting, as shown next. This shows that the familiar marginalization results for Gaussian random variables remain meaningful in infinite dimensions.

We denote realizations of Z by $z = (z_1, z_2) = (\Pi_1 z, \Pi_2 z) \in \mathcal{V}$, where Π_1 and Π_2 denote linear projection operators onto $\mathcal{V}_1$ and $\mathcal{V}_2$, respectively. The following result concerns the law of $\Pi_i Z$, $i = 1, 2$, i.e., marginal laws of Z.

Lemma 2.2. *$Z_i = \Pi_i Z$ has a Gaussian law μ_i with mean $\bar{z}_i = \Pi_i \bar{z}$ and covariance operator $\mathcal{C}_{ii}$, which satisfies*

$$\langle \mathcal{C}_{ii} u, v \rangle_i = \int_{\mathcal{V}_i} \langle s - \bar{z}_i, u \rangle_i \langle s - \bar{z}_i, v \rangle_i \, \mu_i(ds), \quad i = 1, 2, \quad \text{for all } u, v \in \mathcal{V}_i. \tag{2.1}$$

Proof. By Lemma 2.1, $\Pi_i Z$ has a Gaussian law $\mu_i = \mathcal{N}(\Pi_i \bar{z}, \mathcal{C}_{ii})$ with $\mathcal{C}_{ii} = \Pi_i \mathcal{C} \Pi_i^*$. It remains to show that $\mathcal{C}_{ii}$ satisifes (2.1). Without loss of generality, we assume $\bar{z} \equiv 0$ and show the result for $i = 1$. By definition of the covariance operator $\mathcal{C}$ of μ, we have $\langle\!\langle \mathcal{C}a, b \rangle\!\rangle = \int_{\mathcal{V}} \langle\!\langle z, a \rangle\!\rangle \langle\!\langle z, b \rangle\!\rangle \, \mu(dz)$, for $a, b \in \mathcal{V}$. Therefore, for arbitrary $u, v \in \mathcal{V}_1$, we have

$$\langle \mathcal{C}_{11} u, v \rangle_1 = \langle\!\langle \mathcal{C}\Pi_1^* u, \Pi_1^* v \rangle\!\rangle = \int_{\mathcal{V}} \langle\!\langle z, (u, 0) \rangle\!\rangle \langle\!\langle z, (v, 0) \rangle\!\rangle \mu(dz) = \int_{\mathcal{V}_1} \langle s, u \rangle_1 \langle s, v \rangle_1 \mu_1(ds). \qquad \blacksquare$$

In the present work, $\mathcal{V}_1 = L^2(\mathcal{T})$ and $\mathcal{V}_2 = L^2(\mathcal{D})$ with $\mathcal{T}$ and $\mathcal{D}$ bounded open sets in $\mathbb{R}^{d_i}$ with $d_i \in \{1, 2, 3\}$, for $i = 1, 2$. In this case, realizations of $\Pi_1 Z$ and $\Pi_2 Z$ are square-integrable functions on $\mathcal{T}$ and $\mathcal{D}$, respectively. Thus, we can also view $\Pi_i Z$ as a random field. Consider,

e.g., $Z_1 = \Pi_1 Z$. This marginalized random field has mean $\bar{z}_1(x)$ and the following covariance function (kernel):

$$c_{11}(x,y) := \int_\Omega (Z_1(x,\omega) - \bar{z}_1(x))(Z_1(y,\omega) - \bar{z}_1(y))\,\mathbb{P}(d\omega).$$

As expected, the (marginal) covariance operator $\mathcal{C}_{11}$ can be written as an integral operator with kernel c_{11}. To show this, we use (2.1) and again, for simplicity, assume $\bar{z} \equiv 0$. Note that

$$\begin{aligned}
\langle \mathcal{C}_{11}u, v\rangle_1 = \int_{\mathcal{Y}_1} \langle s,u\rangle_1\langle s,v\rangle_1 \mu_1(ds) &= \int_\Omega \langle Z_1(\omega),u\rangle_1\langle Z_1(\omega),v\rangle_1\,\mathbb{P}(d\omega)\\
&= \int_\Omega \int_{\mathcal{T}}\int_{\mathcal{T}} Z_1(x,\omega)Z_1(y,\omega)u(x)v(y)\,dx\,dy\,\mathbb{P}(d\omega)\\
&= \int_{\mathcal{T}}\left[\int_{\mathcal{T}}\left(\int_\Omega Z_1(x,\omega)Z_1(y,\omega)\,\mathbb{P}(d\omega)\right)v(y)\,dy\right]u(x)\,dx\\
&= \int_{\mathcal{T}}\left[\int_{\mathcal{T}} c_{11}(x,y)v(y)\,dy\right]u(x)\,dx,
\end{aligned}$$

where we used Fubini's theorem to change the order of the integrals. From this, we deduce

$$[\mathcal{C}_{11}v](\cdot) = \int_{\mathcal{T}} c_{11}(\cdot,y)v(y)\,dy.$$

In finite dimensions, we recover the following well-known [35] result, which we prove here for completeness:

Lemma 2.3. *Consider a Gaussian random vector*

$$\boldsymbol{Z} = \begin{bmatrix}\boldsymbol{Z}_1\\ \boldsymbol{Z}_2\end{bmatrix} \sim \mathcal{N}(\bar{\boldsymbol{z}}, \mathbf{C}) = \mathcal{N}\left(\begin{bmatrix}\bar{\boldsymbol{z}}_1\\ \bar{\boldsymbol{z}}_2\end{bmatrix}, \begin{bmatrix}\mathbf{C}_{11} & \mathbf{C}_{12}\\ \mathbf{C}_{21} & \mathbf{C}_{22}\end{bmatrix}\right),$$

where $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ denote subsets of entries of $\boldsymbol{Z}$ and the mean and covariance matrix are partitioned consistent with partitioning of $\boldsymbol{Z}$. Then, the marginals of $\boldsymbol{Z}$ are Gaussian, with $\boldsymbol{Z}_1 \sim \mathcal{N}(\bar{\boldsymbol{z}}_1, \mathbf{C}_{11})$ and $\boldsymbol{Z}_2 \sim \mathcal{N}(\bar{\boldsymbol{z}}_2, \mathbf{C}_{22})$.

Proof. Note that $\boldsymbol{Z}_1 = \mathbf{P}\boldsymbol{Z}$ with $\mathbf{P} = \begin{bmatrix}\mathbf{I} & \mathbf{0}\end{bmatrix}$, where $\mathbf{I}$ is the identity matrix of dimension equal to that of $\boldsymbol{Z}_1$ and $\mathbf{0}$ the zero matrix of the same size as $\boldsymbol{Z}_2$. Thus, $\boldsymbol{Z}_1 \sim \mathcal{N}\big(\mathbf{P}\bar{\boldsymbol{z}}, \mathbf{P}\mathbf{C}\mathbf{P}^{\mathsf{T}}\big) = \mathcal{N}(\bar{\boldsymbol{z}}_1, \mathbf{C}_{11})$. Showing the statement about the law of $\boldsymbol{Z}_2$ is analogous. ∎

2.3. Bayesian inverse problem setup. We consider a Bayesian linear inverse problem for $\theta = (m,b) \in \mathcal{V} = \mathcal{V}_1 \times \mathcal{V}_2$ and where the forward model is of the form (1.2). We assume Gaussian priors for the primary and secondary parameters, which we denote by m and b, respectively, and for simplicity of the presentation assume no prior correlation between m and b. The presented framework can be modified to allow for prior correlations between m and b. Thus, the prior law of (m,b) is the product measure $\mu_{\text{pr}} = \mu_{\text{pr,m}} \otimes \mu_{\text{pr,b}}$, with $\mu_{\text{pr,m}}$ and $\mu_{\text{pr,b}}$ each Gaussian measures on $\mathcal{V}_1$ and $\mathcal{V}_2$, i.e., $\mu_{\text{pr,m}} = \mathcal{N}(m_{\text{pr}}, \Gamma_{\text{pr,m}})$ and $\mu_{\text{pr,b}} = \mathcal{N}(b_{\text{pr}}, \Gamma_{\text{pr,b}})$. Note that $\mu_{\text{pr}} = \mathcal{N}(\theta_{\text{pr}}, \Gamma_{\text{pr}})$ with $\theta_{\text{pr}} = (m_{\text{pr}}, b_{\text{pr}})$ and $\Gamma_{\text{pr}} = \Gamma_{\text{pr,m}} \times \Gamma_{\text{pr,b}}$, where

$$(\Gamma_{\text{pr,m}} \times \Gamma_{\text{pr,b}})(u_1,u_2) = (\Gamma_{\text{pr,m}}u_1, \Gamma_{\text{pr,b}}u_2), \quad (u_1,u_2) \in \mathcal{V}.$$

The inverse problem under study considers inference of m and b using measurement data $\boldsymbol{y} \in \mathbb{R}^{n_\mathrm{d}}$ and the model

$$\boldsymbol{y} = \mathcal{F}m + \mathcal{G}b + \boldsymbol{\eta}. \tag{2.2}$$

The measurement noise vector $\boldsymbol{\eta}$ is assumed to be independent of (m, b), and we assume a Gaussian noise model, $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_{\text{noise}})$. Under these assumptions, the posterior is the Gaussian measure $\mu^{\boldsymbol{y}}_{\text{post}} = \mathcal{N}(\theta_{\text{post}}, \Gamma_{\text{post}})$ with [33]

$$\Gamma^{-1}_{\text{post}} = \mathcal{E}^*\boldsymbol{\Gamma}^{-1}_{\text{noise}}\mathcal{E} + \Gamma^{-1}_{\text{pr}}, \quad \theta_{\text{post}} = \Gamma_{\text{post}}(\mathcal{E}^*\boldsymbol{\Gamma}^{-1}_{\text{noise}}\boldsymbol{y} + \Gamma^{-1}_{\text{pr}}\theta_{\text{pr}}). \tag{2.3}$$

Note that $\mathcal{E}^*$ denotes the adjoint of the linear transformation $\mathcal{E}$. Specifically, $\mathcal{E}^*$ satisfies $\mathcal{E}^*\boldsymbol{y} = (\mathcal{F}^*\boldsymbol{y}, \mathcal{G}^*\boldsymbol{y}) \in \mathcal{V}$, for $\boldsymbol{y} \in \mathbb{R}^{n_\mathrm{d}}$.

2.4. The discretized problem. Let $\boldsymbol{m}$ and $\boldsymbol{b}$ be discretized versions of m and b. Recall that we consider a parameter space $\mathcal{V}$ of the form $\mathcal{V} = L^2(\mathcal{T}) \times L^2(\mathcal{D})$. The discretized parameter space is $\mathcal{V}_n = \mathbb{R}^{n_\mathrm{m}} \times \mathbb{R}^{n_\mathrm{b}} \cong \mathbb{R}^n$, where n_m and n_b are the dimensions of the discretized parameters $\boldsymbol{m}$ and $\boldsymbol{b}$, respectively, and $n = n_\mathrm{m} + n_\mathrm{b}$. An element $\boldsymbol{u} \in \mathcal{V}_n$, $\boldsymbol{u} = (\boldsymbol{u}_1, \boldsymbol{u}_2)$ with $\boldsymbol{u}_1 \in \mathbb{R}^{n_\mathrm{m}}$ and $\boldsymbol{u}_2 \in \mathbb{R}^{n_\mathrm{b}}$, can be represented as $\boldsymbol{u} = [\boldsymbol{u}_1^\mathsf{T} \quad \boldsymbol{u}_2^\mathsf{T}]^\mathsf{T}$. The discretized parameter space is endowed with the inner product

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\mathbb{M}} = \boldsymbol{u}_1^\mathsf{T}\mathbf{M}_1\boldsymbol{v}_1 + \boldsymbol{u}_2^\mathsf{T}\mathbf{M}_2\boldsymbol{v}_2 = \boldsymbol{u}^\mathsf{T}\mathbb{M}\boldsymbol{v}, \quad \boldsymbol{u}, \boldsymbol{v} \in \mathcal{V}_n,$$

with $\mathbb{M} = \begin{bmatrix} \mathbf{M}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_2 \end{bmatrix}$, and where the "weight" matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ are defined based on the method used to discretize the L^2-inner products on $L^2(\mathcal{T})$ and $L^2(\mathcal{D})$, respectively; see section 6 for examples. The discretized forward operator is defined by

$$\mathbf{E}\boldsymbol{\theta} = \begin{bmatrix} \mathbf{F} & \mathbf{G} \end{bmatrix} \begin{bmatrix} \boldsymbol{m} \\ \boldsymbol{b} \end{bmatrix} = \mathbf{F}\boldsymbol{m} + \mathbf{G}\boldsymbol{b},$$

where $\mathbf{F}$ and $\mathbf{G}$ are discretizations of $\mathcal{F}$ and $\mathcal{G}$ in (2.2). The respective marginal priors are $\mathcal{N}(\boldsymbol{m}_{\text{pr}}, \boldsymbol{\Gamma}_{\text{pr,m}})$ and $\mathcal{N}(\boldsymbol{b}_{\text{pr}}, \boldsymbol{\Gamma}_{\text{pr,b}})$, and the prior covariance is $\boldsymbol{\Gamma}_{\text{pr}} = \begin{bmatrix} \boldsymbol{\Gamma}_{\text{pr,m}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Gamma}_{\text{pr,b}} \end{bmatrix}$. Using (2.3), the posterior covariance operator satisfies

$$\boldsymbol{\Gamma}^{-1}_{\text{post}} = \begin{bmatrix} \boldsymbol{\Gamma}^{-1}_{\text{pr,m}} + \mathbf{F}^*\boldsymbol{\Gamma}^{-1}_{\text{noise}}\mathbf{F} & \mathbf{F}^*\boldsymbol{\Gamma}^{-1}_{\text{noise}}\mathbf{G} \\ \mathbf{G}^*\boldsymbol{\Gamma}^{-1}_{\text{noise}}\mathbf{F} & \boldsymbol{\Gamma}^{-1}_{\text{pr,b}} + \mathbf{G}^*\boldsymbol{\Gamma}^{-1}_{\text{noise}}\mathbf{G} \end{bmatrix}.$$

Computing the inverse of the block matrix on the right is facilitated by the well-known formula for the inverse of a such matrices [25, Theorem 2.1(ii)]. Specifically, we can show that the covariance operator of the marginal posterior law of $\boldsymbol{m}$ is given by

$$\boldsymbol{\Gamma}_{\text{post,m}} = \big(\boldsymbol{\Gamma}^{-1}_{\text{pr,m}} + \mathbf{F}^*\boldsymbol{\Gamma}^{-1}_{\text{noise}}\mathbf{F} - \mathbf{F}^*\boldsymbol{\Gamma}^{-1}_{\text{noise}}\mathbf{G}(\boldsymbol{\Gamma}^{-1}_{\text{pr,b}} + \mathbf{G}^*\boldsymbol{\Gamma}^{-1}_{\text{noise}}\mathbf{G})^{-1}\mathbf{G}^*\boldsymbol{\Gamma}^{-1}_{\text{noise}}\mathbf{F}\big)^{-1}. \tag{2.4}$$

Note also that for

$$\mathbf{F} : (\mathbb{R}^{n_\mathrm{m}}, \langle \cdot, \cdot \rangle_{\mathbf{M}_1}) \to (\mathbb{R}^{n_\mathrm{d}}, \langle \cdot, \cdot \rangle_{\mathbb{R}^{n_\mathrm{d}}}) \quad \text{and} \quad \mathbf{G} : (\mathbb{R}^{n_\mathrm{b}}, \langle \cdot, \cdot \rangle_{\mathbf{M}_2}) \to (\mathbb{R}^{n_\mathrm{d}}, \langle \cdot, \cdot \rangle_{\mathbb{R}^{n_\mathrm{d}}}),$$

where $\langle\cdot,\cdot\rangle_{\mathbb{R}^{n_\mathrm{d}}}$ denotes the Euclidean inner product on $\mathbb{R}^{n_\mathrm{d}}$, the respective adjoint operators are defined by (cf. e.g., [7])

$$\mathbf{F}^* = \mathbf{M}_1^{-1}\mathbf{F}^\mathsf{T} \quad \text{and} \quad \mathbf{G}^* = \mathbf{M}_2^{-1}\mathbf{G}^\mathsf{T}. \tag{2.5}$$

The optimal design approach we follow consists of minimizing the average posterior variance in $\boldsymbol{m}$ by minimizing the trace of the marginal posterior covariance operator defined in (2.4). We call the resulting OED criterion the *marginalized A-optimality criterion*. In section 3, we derive an alternative expression for the marginal posterior covariance operator, which is useful in applications which only allow low or moderate dimensional measurements.

2.5. Optimal experimental design. We formulate the sensor placement problem using the approach in [3, 15]. We assume $\boldsymbol{x}_i$, $i = 1, \ldots, n_\mathrm{d}$, represent a fixed set of candidate sensor locations. The goal is to select an optimal subset of these locations. We assign a non-negative weight $w_i \in \mathbb{R}$ to each $\boldsymbol{x}_i$, $i = 1, \ldots, n_\mathrm{d}$. An experimental design is specified by the vector $\boldsymbol{w} = [w_1, w_2, \ldots, w_{n_\mathrm{d}}]^\mathsf{T}$. As detailed in [3, 15], binary weight vectors are desirable to decide whether or not to place a sensor in each of the candidate locations. However, solving an OED problem with binary weights is challenging due to its combinatorial complexity. Thus, as in [3], we relax the problem by considering weights $w_i \in [0, 1]$, $i = 1, \ldots, n_\mathrm{d}$. Binary weights are obtained using sparsifying penalty functions, as discussed further in subsection 4.2. An alternative approach to obtaining binary weights, which can be suitable for some problems, is a greedy strategy; see subsection 4.3.

The vector $\boldsymbol{w}$ is introduced into the Bayesian inverse problem through the data likelihood [3]. We assume uncorrelated measurements; i.e., the noise covariance is diagonal, $\boldsymbol{\Gamma}_\mathrm{noise} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_{n_\mathrm{d}}^2)$, with σ_j^2 the noise level at the jth sensor. For $\boldsymbol{w} \in \mathbb{R}^{n_\mathrm{d}}$, we define the diagonal weight matrix $\mathbf{W} = \mathrm{diag}(w_1, w_2, \ldots, w_{n_\mathrm{d}})$ and the matrix $\mathbf{W}_\sigma$ as follows:

$$\mathbf{W}_\sigma := \mathrm{diag}\Big(\frac{w_1}{\sigma_1^2}, \frac{w_2}{\sigma_2^2}, \ldots, \frac{w_{n_d}}{\sigma_{n_d}^2}\Big) = \sum_{j=1}^{n_\mathrm{d}} w_j \sigma_j^{-2} \boldsymbol{e}_j \boldsymbol{e}_j^\top, \tag{2.6}$$

where $\boldsymbol{e}_j$ is the jth coordinate vector in $\mathbb{R}^{n_\mathrm{d}}$. The $\boldsymbol{w}$-dependent MAP estimator and posterior covariance operator are then given by [3]

$$\boldsymbol{\theta}_\mathrm{MAP}(\boldsymbol{w}) = \boldsymbol{\Gamma}_\mathrm{post}(\boldsymbol{w})\big(\mathbf{E}^*\mathbf{W}_\sigma \boldsymbol{y} + \boldsymbol{\Gamma}_\mathrm{pr}^{-1}\boldsymbol{\theta}_\mathrm{pr}\big) \quad \text{and} \quad \boldsymbol{\Gamma}_\mathrm{post}(\boldsymbol{w}) = (\mathbf{E}^*\mathbf{W}_\sigma\mathbf{E} + \boldsymbol{\Gamma}_\mathrm{pr}^{-1})^{-1}. \tag{2.7}$$

Optimal experimental design (OED) is the problem of finding a design that, within constraints on the number of sensors allowed, minimizes the posterior uncertainty in the estimated parameters. This is done by minimizing certain design criteria that quantify the posterior uncertainty [8, 37]. In this article, we use the A-optimal design criterion which is given by $\mathsf{tr}\big[\boldsymbol{\Gamma}_\mathrm{post}(\boldsymbol{w})\big]$; this criterion quantifies the average posterior variance of the parameter $\boldsymbol{\theta}$. Using this approach for (2.7), the OED objective is given by the sum of the average posterior variance of the primary and secondary parameters. The primary parameter being the main focus of parameter estimation, we seek sensor placements that minimize the uncertainty in the primary parameter, while being aware of the uncertainty in the secondary parameters. This is done by finding designs that minimize the average posterior variance of the primary parameters, quantified according to the corresponding marginalized posterior distribution. We call such designs marginalized A-optimal designs, which are the subject of section 3.

Note that ignoring the uncertainty in the secondary parameter and fixing $\boldsymbol{b}$ to some nominal value $\boldsymbol{b}_0$, results in the affine forward model $\mathbf{E}_0\boldsymbol{m} = \mathbf{F}\boldsymbol{m} + \mathbf{G}\boldsymbol{b}_0$. In this case, the posterior law of $\boldsymbol{m}$ is $\mathcal{N}(\boldsymbol{m}_{\text{MAP}}, \boldsymbol{\Gamma}_{\text{post,m}})$ with

$$\begin{aligned} \boldsymbol{m}_{\text{MAP}}(\boldsymbol{w}) &= \boldsymbol{\Gamma}_{\text{post,m}}(\boldsymbol{w})\big(\mathbf{F}^*\mathbf{W}_\sigma(\boldsymbol{y} - \mathbf{G}\boldsymbol{b}_0) + \boldsymbol{\Gamma}_{\text{pr,m}}^{-1}\boldsymbol{m}_{\text{pr}}\big) \quad \text{and} \\ \boldsymbol{\Gamma}_{\text{post,m}}(\boldsymbol{w}) &= (\mathbf{F}^*\mathbf{W}_\sigma\mathbf{F} + \boldsymbol{\Gamma}_{\text{pr,m}}^{-1})^{-1}, \end{aligned} \tag{2.8}$$

and an A-optimal design $\boldsymbol{w}$ is one that minimizes the classical A-optimality criterion

$$\psi(\boldsymbol{w}) := \mathsf{tr}\big[(\mathbf{F}^*\mathbf{W}_\sigma\mathbf{F} + \boldsymbol{\Gamma}_{\text{pr,m}}^{-1})^{-1}\big]. \tag{2.9}$$

Notice that the optimal design does not depend on the choice of $\boldsymbol{b}_0$. More importantly, such an optimal design is completely unaware of the uncertainty in $\boldsymbol{b}$.

3. Marginalized Bayesian A-optimality.

In this section, we present our formulation of the marginalized A-optimality criterion. We first derive a reformulation of the marginalized posterior covariance that facilitates an efficient computational procedure for computing marginalized A-optimal designs; see subsection 3.1. Then, we present the definition of the marginalized A-optimality criterion, in subsection 3.2, and prove its convexity. Finally, the formulation of the optimization problem for finding marginalized A-optimal designs is discussed in subsection 3.3.

3.1. Alternative form of the posterior.

Computing optimal designs based on the marginalized posterior covariance operator (2.4) entails traces of operators defined on the discretized parameter spaces. The corresponding expressions also include inverses of operators of dimensions n_{m} and n_{b}; see (2.4). The discretized parameter dimensions are typically large and depend on the computational grids used for discretization. In many large scale inverse problems, the dimension n_{d} of the measurement vector $\boldsymbol{y}$ is considerably smaller than the dimension of the discretized uncertain parameters. Also, in our approach, this measurement dimension is fixed a priori. Here we derive an alternative expression for the posterior covariance operator (2.7) that facilitates exploiting this problem structure. In particular, this allows reformulating the marginalized A-optimality criterion in terms of an operator defined on the measurement space, which can then be computed directly (see section 4). This is in contrast to previous works such as [3, 13–15, 17] that use randomized trace estimation (in the discretized parameter space) to compute the OED objective.

Theorem 3.1. *The following relation holds.*

$$(\mathbf{E}^*\mathbf{W}_\sigma\mathbf{E} + \boldsymbol{\Gamma}_{\text{pr}}^{-1})^{-1} = \boldsymbol{\Gamma}_{\text{pr}} - \boldsymbol{\Gamma}_{\text{pr}}\mathbf{E}^*(\mathbf{I} + \mathbf{W}_\sigma\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}\mathbf{E}^*)^{-1}\mathbf{W}_\sigma\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}. \tag{3.1}$$

Proof. First, we need to show that $\mathbf{I} + \mathbf{W}_\sigma\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}\mathbf{E}^*$ is invertible. To do this, we show that $\mathbf{W}_\sigma\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}\mathbf{E}^*$ has non-negative eigenvalues. Note that $\boldsymbol{\Gamma}_{\text{pr}} = \boldsymbol{\Gamma}_{\text{pr}}^* = \mathbb{M}^{-1}\boldsymbol{\Gamma}_{\text{pr}}^\mathsf{T}\mathbb{M}$. Moreover, we have that $\mathbf{E}^* = \mathbb{M}^{-1}\mathbf{E}^\mathsf{T}$. Thus, we have $(\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}\mathbf{E}^*)^\mathsf{T} = (\mathbf{E}^*)^\mathsf{T}\boldsymbol{\Gamma}_{\text{pr}}^\mathsf{T}\mathbf{E}^\mathsf{T} = \mathbf{E}\mathbb{M}^{-1}\boldsymbol{\Gamma}_{\text{pr}}^\mathsf{T}\mathbb{M}\mathbf{E}^* = \mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}\mathbf{E}^*$. That is, $\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}\mathbf{E}^*$ is symmetric; it is also clearly positive semidefinite.

To show that $\mathbf{W}_\sigma\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}\mathbf{E}^*$ has non-negative eigenvalues, we recall a basic result from linear algebra: if $\mathbf{A}$ and $\mathbf{B}$ are two square matrices, $\mathbf{AB}$ and $\mathbf{BA}$ have the same eigenvalues; see e.g., [28, page 249]. Applying this result with $\mathbf{A} = \mathbf{W}_\sigma^{1/2}\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}\mathbf{E}^*$ and $\mathbf{B} = \mathbf{W}_\sigma^{1/2}$, we

have that $\mathbf{W}_\sigma^{1/2}\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}\mathbf{E}^*\mathbf{W}_\sigma^{1/2}$ and $\mathbf{W}_\sigma\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}\mathbf{E}^*$ have the same eigenvalues. Therefore, since $\mathbf{W}_\sigma^{1/2}\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}\mathbf{E}^*\mathbf{W}_\sigma^{1/2}$ is symmetric positive semidefinite, it follows that $\mathbf{W}_\sigma\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}\mathbf{E}^*$ has non-negative eigenvalues. This implies that $\mathbf{I}+\mathbf{W}_\sigma\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}\mathbf{E}^*$ is invertible. The relation (3.1) is now seen as follows:

$$
\begin{aligned}
&(\mathbf{E}^*\mathbf{W}_\sigma\mathbf{E}+\boldsymbol{\Gamma}_{\text{pr}}^{-1})(\boldsymbol{\Gamma}_{\text{pr}}-\boldsymbol{\Gamma}_{\text{pr}}\mathbf{E}^*(\mathbf{I}+\mathbf{W}_\sigma\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}\mathbf{E}^*)^{-1}\mathbf{W}_\sigma\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}})\\
&\quad=\mathbf{E}^*\mathbf{W}_\sigma\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}-\mathbf{E}^*\mathbf{W}_\sigma\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}\mathbf{E}^*(\mathbf{I}+\mathbf{W}_\sigma\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}\mathbf{E}^*)^{-1}\mathbf{W}_\sigma\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}+\mathbf{I}-\mathbf{E}^*(\mathbf{I}+\mathbf{W}_\sigma\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}\mathbf{E}^*)^{-1}\mathbf{W}_\sigma\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}\\
&\quad=\mathbf{I}+\mathbf{E}^*\mathbf{W}_\sigma\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}-\mathbf{E}^*(\mathbf{W}_\sigma\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}\mathbf{E}^*+\mathbf{I})(\mathbf{I}+\mathbf{W}_\sigma\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}\mathbf{E}^*)^{-1}\mathbf{W}_\sigma\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}\\
&\quad=\mathbf{I}+\mathbf{E}^*\mathbf{W}_\sigma\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}-\mathbf{E}^*\mathbf{W}_\sigma\mathbf{E}\boldsymbol{\Gamma}_{\text{pr}}=\mathbf{I}.
\end{aligned}
$$

∎

Notice that this result is well known in the case $\mathbf{W}_\sigma=\boldsymbol{\Gamma}_{\text{noise}}^{-1}$. The challenge here is to account for the possibility of a singular $\mathbf{W}_\sigma$. Note that the expression in the left hand side of (3.1) involves the inverse of an $n\times n$ matrix, where $n=n_{\text{m}}+n_{\text{b}}$, whereas the expression on the right hand side involves the inverse of an $n_{\text{d}}\times n_{\text{d}}$ matrix. It is also worth noting that the proof of Theorem 3.1 can be simplified by the use of the Sherman–Morrison–Woodbury formula. Above, we chose to present a direct linear algebra argument instead, for clarity.

We introduce the following notations, which will be used in the remainder of this article.

$$
\mathbf{Q}(\boldsymbol{w}) := (\mathbf{I}+\mathbf{W}_\sigma\mathbf{C})^{-1}\,\mathbf{W}_\sigma, \quad \text{where} \quad \mathbf{C} := \mathbf{F}\boldsymbol{\Gamma}_{\text{pr,m}}\mathbf{F}^*+\mathbf{G}\boldsymbol{\Gamma}_{\text{pr,b}}\mathbf{G}^*. \tag{3.2}
$$

Next, we present tractable representations for the posterior mean and covariance operator in a (discretized) Bayesian linear inverse problem, as formulated in subsection 2.4. Recall that the primary parameter is $\boldsymbol{m}$ and the secondary parameter is $\boldsymbol{b}$.

Theorem 3.2. *The posterior law of* $\begin{bmatrix}\boldsymbol{m}\\ \boldsymbol{b}\end{bmatrix}$ *is* $\mathcal{N}\left(\begin{bmatrix}\boldsymbol{m}_{\text{MAP}}\\ \boldsymbol{b}_{\text{MAP}}\end{bmatrix},\begin{bmatrix}\boldsymbol{\Gamma}_{\text{post,m}}(\boldsymbol{w}) & \boldsymbol{\Gamma}_{\text{post,mb}}(\boldsymbol{w})\\ \boldsymbol{\Gamma}_{\text{post,mb}}^*(\boldsymbol{w}) & \boldsymbol{\Gamma}_{\text{post,b}}(\boldsymbol{w})\end{bmatrix}\right)$, *where*

$$
\begin{aligned}
\boldsymbol{\Gamma}_{\text{post,m}}(\boldsymbol{w}) &= \boldsymbol{\Gamma}_{\text{pr,m}}-\boldsymbol{\Gamma}_{\text{pr,m}}\mathbf{F}^*\mathbf{Q}(\boldsymbol{w})\mathbf{F}\boldsymbol{\Gamma}_{\text{pr,m}},\\
\boldsymbol{\Gamma}_{\text{post,b}}(\boldsymbol{w}) &= \boldsymbol{\Gamma}_{\text{pr,b}}-\boldsymbol{\Gamma}_{\text{pr,b}}\mathbf{G}^*\mathbf{Q}(\boldsymbol{w})\mathbf{G}\boldsymbol{\Gamma}_{\text{pr,b}},\\
\boldsymbol{\Gamma}_{\text{post,mb}}(\boldsymbol{w}) &= -\boldsymbol{\Gamma}_{\text{pr,m}}\mathbf{F}^*\mathbf{Q}(\boldsymbol{w})\mathbf{G}\boldsymbol{\Gamma}_{\text{pr,b}},\\
\boldsymbol{m}_{\text{MAP}}(\boldsymbol{w}) &= \boldsymbol{\Gamma}_{\text{post,m}}(\boldsymbol{w})(\mathbf{F}^*\mathbf{W}_\sigma\boldsymbol{y}+\boldsymbol{\Gamma}_{\text{pr,m}}^{-1}\boldsymbol{m}_{\text{pr}})+\boldsymbol{\Gamma}_{\text{post,mb}}(\boldsymbol{w})(\mathbf{G}^*\mathbf{W}_\sigma\boldsymbol{y}+\boldsymbol{\Gamma}_{\text{pr,b}}^{-1}\boldsymbol{b}_{\text{pr}}),\\
\boldsymbol{b}_{\text{MAP}}(\boldsymbol{w}) &= \boldsymbol{\Gamma}_{\text{post,b}}(\boldsymbol{w})(\mathbf{G}^*\mathbf{W}_\sigma\boldsymbol{y}+\boldsymbol{\Gamma}_{\text{pr,b}}^{-1}\boldsymbol{b}_{\text{pr}})+\boldsymbol{\Gamma}_{\text{post,mb}}^*(\boldsymbol{w})(\mathbf{F}^*\mathbf{W}_\sigma\boldsymbol{y}+\boldsymbol{\Gamma}_{\text{pr,m}}^{-1}\boldsymbol{m}_{\text{pr}}).
\end{aligned} \tag{3.3}
$$

Proof. Recall that the discretized forward operator $\mathbf{E}$ can be represented in a block matrix form $\mathbf{E}=\begin{bmatrix}\mathbf{F} & \mathbf{G}\end{bmatrix}$. Using this and the expression for $\boldsymbol{\Gamma}_{\text{post}}$ given in Theorem 3.1, we obtain

$$
\begin{aligned}
\boldsymbol{\Gamma}_{\text{post}}(\boldsymbol{w}) &= \boldsymbol{\Gamma}_{\text{pr}}-\boldsymbol{\Gamma}_{\text{pr}}\begin{bmatrix}\mathbf{F}^*\\ \mathbf{G}^*\end{bmatrix}\left(\mathbf{I}+\mathbf{W}_\sigma\begin{bmatrix}\mathbf{F} & \mathbf{G}\end{bmatrix}\boldsymbol{\Gamma}_{\text{pr}}\begin{bmatrix}\mathbf{F}^*\\ \mathbf{G}^*\end{bmatrix}\right)^{-1}\mathbf{W}_\sigma\begin{bmatrix}\mathbf{F} & \mathbf{G}\end{bmatrix}\boldsymbol{\Gamma}_{\text{pr}}\\
&= \boldsymbol{\Gamma}_{\text{pr}}-\boldsymbol{\Gamma}_{\text{pr}}\begin{bmatrix}\mathbf{F}^*\\ \mathbf{G}^*\end{bmatrix}(\mathbf{I}+\mathbf{W}_\sigma(\mathbf{F}\boldsymbol{\Gamma}_{\text{pr,m}}\mathbf{F}^*+\mathbf{G}\boldsymbol{\Gamma}_{\text{pr,b}}\mathbf{G}^*))^{-1}\mathbf{W}_\sigma\begin{bmatrix}\mathbf{F} & \mathbf{G}\end{bmatrix}\boldsymbol{\Gamma}_{\text{pr}}\\
&= \begin{bmatrix}\boldsymbol{\Gamma}_{\text{pr,m}} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Gamma}_{\text{pr,b}}\end{bmatrix}-\begin{bmatrix}\boldsymbol{\Gamma}_{\text{pr,m}} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Gamma}_{\text{pr,b}}\end{bmatrix}\begin{bmatrix}\mathbf{F}^*\\ \mathbf{G}^*\end{bmatrix}\mathbf{Q}(\boldsymbol{w})\begin{bmatrix}\mathbf{F} & \mathbf{G}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Gamma}_{\text{pr,m}} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Gamma}_{\text{pr,b}}\end{bmatrix}\\
&= \begin{bmatrix}\boldsymbol{\Gamma}_{\text{pr,m}}-\boldsymbol{\Gamma}_{\text{pr,m}}\mathbf{F}^*\mathbf{Q}(\boldsymbol{w})\mathbf{F}\boldsymbol{\Gamma}_{\text{pr,m}} & -\boldsymbol{\Gamma}_{\text{pr,m}}\mathbf{F}^*\mathbf{Q}(\boldsymbol{w})\mathbf{G}\boldsymbol{\Gamma}_{\text{pr,b}}\\ -\boldsymbol{\Gamma}_{\text{pr,b}}\mathbf{G}^*\mathbf{Q}(\boldsymbol{w})\mathbf{F}\boldsymbol{\Gamma}_{\text{pr,m}} & \boldsymbol{\Gamma}_{\text{pr,b}}-\boldsymbol{\Gamma}_{\text{pr,b}}\mathbf{G}^*\mathbf{Q}(\boldsymbol{w})\mathbf{G}\boldsymbol{\Gamma}_{\text{pr,b}}\end{bmatrix}.
\end{aligned} \tag{3.4}
$$

This establishes the representation of the posterior covariance operator. The expressions for $\boldsymbol{m}_{\mathrm{MAP}}(\boldsymbol{w})$ and $\boldsymbol{b}_{\mathrm{MAP}}(\boldsymbol{w})$ can be obtained using (2.7) and (3.4). ∎

Using Lemma 2.3 in conjunction with Theorem 3.2, the marginal posterior laws of $\boldsymbol{m}$ and $\boldsymbol{b}$ are given by $\mathcal{N}(\boldsymbol{m}_{\mathrm{MAP}}(\boldsymbol{w}), \boldsymbol{\Gamma}_{\mathrm{post,m}}(\boldsymbol{w}))$ and $\mathcal{N}(\boldsymbol{b}_{\mathrm{MAP}}(\boldsymbol{w}), \boldsymbol{\Gamma}_{\mathrm{post,b}}(\boldsymbol{w}))$, respectively. Since we target the primary parameter $\boldsymbol{m}$, we focus on the corresponding marginal posterior law $\mathcal{N}(\boldsymbol{m}_{\mathrm{MAP}}(\boldsymbol{w}), \boldsymbol{\Gamma}_{\mathrm{post,m}}(\boldsymbol{w}))$. The marginal covariance operator $\boldsymbol{\Gamma}_{\mathrm{post,m}}(\boldsymbol{w})$ will be used to define the marginal A-optimality criterion (see below). Also, note that the expression for $\boldsymbol{m}_{\mathrm{MAP}}$ in (3.3) is the sum of two terms: the first is the familiar expression for the posterior mean if $\boldsymbol{b}$ was fixed to $\boldsymbol{b} = \mathbf{0}$; the second reflects the impact of the uncertainty in $\boldsymbol{b}$.

3.2. The marginalized A-optimality criterion. The marginalized A-optimal design (mOED) criterion is given by

$$\Phi(\boldsymbol{w}) := \mathsf{tr}(\boldsymbol{\Gamma}_{\mathrm{post,m}}(\boldsymbol{w})) = \mathsf{tr}(\boldsymbol{\Gamma}_{\mathrm{pr,m}}) - \mathsf{tr}(\boldsymbol{\Gamma}_{\mathrm{pr,m}}\mathbf{F}^*\mathbf{Q}(\boldsymbol{w})\mathbf{F}\boldsymbol{\Gamma}_{\mathrm{pr,m}}). \tag{3.5}$$

Next, we show the convexity of the mOED objective. Before proving this, we consider a slightly more general result. Below, $S^{\mathbf{M}}_{++}$ denotes the cone of self-adjoint and positive definite operators on $\mathbb{R}^n$ equipped with the weighted inner product $\langle \cdot, \cdot \rangle_{\mathbb{M}}$.

Theorem 3.3. *Let the function $f : \mathbb{R}^{n_s}_{\geq 0} \to \mathbb{R}$ be given by*

$$f(\boldsymbol{w}) = \mathsf{tr}(\mathbf{R}\boldsymbol{\Gamma}_{\mathrm{post}}(\boldsymbol{w})\mathbf{R}^*),$$

where $\mathbf{R}$ is an $n \times n$ matrix and $\mathbf{R}^*$ denotes its adjoint with respect to $\langle \cdot, \cdot \rangle_{\mathbb{M}}$. Then, the function f is convex.

Proof. Let $\mathbf{A}(\boldsymbol{w}) = \boldsymbol{\Gamma}_{\mathrm{post}}(\boldsymbol{w})^{-1}$, and note that $\mathbf{A}(\boldsymbol{w}) \in S^{\mathbf{M}}_{++}$ for all $\boldsymbol{w} \in \mathbb{R}^{n_s}_{\geq 0}$. First we show the function $G(\mathbf{A}) = \mathsf{tr}(\mathbf{R}\mathbf{A}^{-1}\mathbf{R}^*)$ is convex on $S^{\mathbf{M}}_{++}$. Consider the restriction of G to a line, $\mathbf{S} + t\mathbf{B}$, where $\mathbf{S} \in S^{\mathbf{M}}_{++}$ and $\mathbf{B}$ is self-adjoint; we consider values of t for which $\mathbf{S} + t\mathbf{B} \in S^{\mathbf{M}}_{++}$. Let $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^*$ be the spectral decomposition of $\mathbf{V} = \mathbf{S}^{-1/2}\mathbf{B}\mathbf{S}^{-1/2}$; here $\boldsymbol{\Lambda}$ is a diagonal matrix with the eigenvalues $\{\lambda_i\}_{i=1}^n$ of $\mathbf{V}$ on its diagonal and $\mathbf{U}$ is a matrix with the corresponding eigenvectors $\{\boldsymbol{u}_i\}_{i=1}^n$ as its columns. Letting $\mathbf{L} = \mathbf{S}^{-1/2}\mathbf{R}^*$, we note

$$\begin{aligned} G(\mathbf{S} + t\mathbf{B}) &= \mathsf{tr}(\mathbf{R}\mathbf{S}^{-1/2}(\mathbf{I} + t\mathbf{S}^{-1/2}\mathbf{B}\mathbf{S}^{-1/2})^{-1}\mathbf{S}^{-1/2}\mathbf{R}^*) \\ &= \mathsf{tr}(\mathbf{L}\mathbf{L}^*(\mathbf{I} + t\mathbf{V})^{-1}) = \sum_{i=1}^n \left\langle \mathbf{L}\mathbf{L}^*(\mathbf{I} + t\mathbf{V})^{-1}\boldsymbol{u}_i, \boldsymbol{u}_i \right\rangle_{\mathbb{M}} = \sum_{i=1}^n (1 + t\lambda_i)^{-1} \left\langle \mathbf{L}^*\boldsymbol{u}_i, \mathbf{L}^*\boldsymbol{u}_i \right\rangle_{\mathbb{M}}. \end{aligned}$$

Thus, $G(\mathbf{S} + t\mathbf{B})$ is a linear combination of convex functions with non-negative coefficients, $\langle \mathbf{L}^*\boldsymbol{u}_i, \mathbf{L}^*\boldsymbol{u}_i \rangle_{\mathbb{M}} \geq 0$, and is thus convex. This shows that G is convex on $S^{\mathbf{M}}_{++}$. It remains to show that $f(\boldsymbol{w}) = G(\mathbf{A}(\boldsymbol{w}))$ is convex. Recall that $\mathbf{A}(\boldsymbol{w}) = \boldsymbol{\Gamma}_{\mathrm{pr}}^{-1} + \mathbf{E}^*\mathbf{W}_\sigma\mathbf{E}$; thus $\mathbf{A}$ is affine in $\boldsymbol{w}$ and therefore, for $\alpha \in [0, 1]$,

$$\begin{aligned} f(\alpha\boldsymbol{w} + (1-\alpha)\boldsymbol{v}) &= G(\mathbf{A}(\alpha\boldsymbol{w} + (1-\alpha)\boldsymbol{v})) = G(\alpha\mathbf{A}(\boldsymbol{w}) + (1-\alpha)\mathbf{A}(\boldsymbol{v})) \\ &\leq \alpha G(\mathbf{A}(\boldsymbol{w})) + (1-\alpha)G(\mathbf{A}(\boldsymbol{v})) = \alpha f(\boldsymbol{w}) + (1-\alpha)f(\boldsymbol{v}). \end{aligned}$$ ∎

Corollary 3.4. *The function $\Phi : \mathbb{R}^{n_d}_{\geq 0} \to \mathbb{R}$, defined in (3.5), is convex.*

Proof. Using (3.4), we can write $\Phi(\boldsymbol{w})$ as

$$\Phi(\boldsymbol{w}) = \mathsf{tr}(\mathbf{R}\boldsymbol{\Gamma}_{\text{post}}(\boldsymbol{w})\mathbf{R}^*) \quad \text{with} \quad \mathbf{R} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Thus, the convexity of $\Phi(\boldsymbol{w})$ can be concluded from Proposition 3.3. ∎

Consider the marginalized A-optimality criterion $\Phi(\boldsymbol{w})$ in (3.5). Since the prior covariance operator is independent of $\boldsymbol{w}$, minimizing $\Phi(\boldsymbol{w})$ is equivalent to minimizing

$$\Psi(\boldsymbol{w}) := -\mathsf{tr}(\mathbf{F}\boldsymbol{\Gamma}_{\text{pr,m}}^2\mathbf{F}^*\mathbf{Q}(\boldsymbol{w})). \tag{3.6}$$

This is the objective function we use in finding a marginalized A-optimal design. Henceforth, we refer to this objective function as the mOED objective or the mOED criterion.

3.3. Computing optimal designs. Here we describe the optimization problem for computing mOEDs. The ultimate goal is to find a binary optimal design vector that minimizes the mOED objective Ψ, defined in (3.6). That is, letting $\mathcal{X} = \{0,1\}^{n_{\text{d}}}$, we would like to solve

$$\min_{\boldsymbol{w}\in\mathcal{X}} \Psi(\boldsymbol{w}), \quad \text{s.t.} \sum_{i=1}^{n_{\text{d}}} w_i = N, \tag{3.7}$$

where N is a desired number of sensors. However, as mentioned above, solving such a binary optimization problem can be intractable due to its combinatorial complexity. One possibility to find an approximate solution to this problem is via a greedy procedure, i.e., place sensors one-by-one. This method does not require derivatives of the objective with respect to weights. Greedy approaches result, in general, in suboptimal solutions, which, in practice, are often quite good. Computational details of this approach are discussed in subsection 4.3. We also compare, in subsection 6.1, the performance of the greedy approach against the approach described next.

As an alternative to the greedy approach, one can consider a relaxation of the problem and allow for design weights in the interval $[0,1]$. Binary weights are then obtained using sparsifying penalty functions. Specifically, we consider an optimization problem of the form

$$\min_{\boldsymbol{w}\in\mathcal{W}} \Psi(\boldsymbol{w}) + \gamma P(\boldsymbol{w}), \tag{3.8}$$

where $\mathcal{W} = [0,1]^{n_{\text{d}}}$, $\Psi(\boldsymbol{w})$ is the mOED objective, $\gamma > 0$ is a penalty parameter, and $P(\boldsymbol{w})$ is a penalty function. Minimization of (3.8) usually requires gradients of the objective. Key computational aspects are discussed in the next section where we outline computational methods for tackling the mOED problem.

4. Computational methods. In this section, we present a computational framework for computing mOEDs.

4.1. Efficient computation of mOED objective and its gradient. Consider the objective function $\Psi(\boldsymbol{w})$ defined in (3.6). We note that the argument of the trace in (3.6) is an operator

defined on $\mathbb{R}^{n_\mathrm{d}\times n_\mathrm{d}}$, where n_d is the number of candidate sensor locations (i.e., the dimension of the measurement vector). This objective function can be computed as follows:

$$\Psi(\boldsymbol{w}) = -\sum_{i=1}^{n_\mathrm{d}} \boldsymbol{e}_i^\mathsf{T}\mathbf{D}\mathbf{Q}(\boldsymbol{w})\boldsymbol{e}_i = -\sum_{i=1}^{n_\mathrm{d}} \boldsymbol{e}_i^\mathsf{T}\mathbf{D}\boldsymbol{q}_i, \quad \text{where } \mathbf{D} = \mathbf{F}\boldsymbol{\Gamma}_{\mathrm{pr,m}}^2\mathbf{F}^*, \tag{4.1}$$

$\boldsymbol{q}_i = \mathbf{Q}(\boldsymbol{w})\boldsymbol{e}_i$ with $\mathbf{Q}(\boldsymbol{w})$ given in (3.2), and $\boldsymbol{e}_i$ is the ith standard basis vector in $\mathbb{R}^{n_\mathrm{d}}$, $i = 1, \ldots, n_\mathrm{d}$. Note that

$$\boldsymbol{q}_i = (\mathbf{I} + \mathbf{W}_\sigma\mathbf{C})^{-1}\mathbf{W}_\sigma\boldsymbol{e}_i = \sigma_i^{-2}w_i(\mathbf{I} + \mathbf{W}_\sigma\mathbf{C})^{-1}\boldsymbol{e}_i. \tag{4.2}$$

To derive the expression for the gradient of Ψ, we first need the following derivative:

$$\frac{\partial}{\partial w_j}\mathbf{Q}(\boldsymbol{w}) = -\sigma_j^{-2}\left(\mathbf{I} + \mathbf{W}_\sigma\mathbf{C}\right)^{-1}(\boldsymbol{e}_j\boldsymbol{e}_j^\mathsf{T})\mathbf{C}\left(\mathbf{I} + \mathbf{W}_\sigma\mathbf{C}\right)^{-1}\mathbf{W}_\sigma + \sigma_j^{-2}\left(\mathbf{I} + \mathbf{W}_\sigma\mathbf{C}\right)^{-1}\boldsymbol{e}_j\boldsymbol{e}_j^\mathsf{T}.$$

Thus,

$$\begin{aligned}
\frac{\partial\Psi}{\partial w_j} &= -\frac{\partial}{\partial w_j}\mathsf{tr}(\mathbf{Q}(\boldsymbol{w})\mathbf{D}) \\
&= \mathsf{tr}\left[\sigma_j^{-2}\left(\mathbf{I} + \mathbf{W}_\sigma\mathbf{C}\right)^{-1}\boldsymbol{e}_j\boldsymbol{e}_j^\mathsf{T}\mathbf{C}\left(\mathbf{I} + \mathbf{W}_\sigma\mathbf{C}\right)^{-1}\mathbf{W}_\sigma\mathbf{D}\right] - \mathsf{tr}\left[\sigma_j^{-2}\left(\mathbf{I} + \mathbf{W}_\sigma\mathbf{C}\right)^{-1}\boldsymbol{e}_j\boldsymbol{e}_j^\mathsf{T}\mathbf{D}\right] \\
&= \sigma_j^{-2}\boldsymbol{e}_j^\mathsf{T}\mathbf{C}(\mathbf{I} + \mathbf{W}_\sigma\mathbf{C})^{-1}\mathbf{W}_\sigma\mathbf{D}(\mathbf{I} + \mathbf{W}_\sigma\mathbf{C})^{-1}\boldsymbol{e}_j - \sigma_j^{-2}\boldsymbol{e}_j^\mathsf{T}\mathbf{D}(\mathbf{I} + \mathbf{W}_\sigma\mathbf{C})^{-1}\boldsymbol{e}_j \\
&= \sum_{i=1}^{n_\mathrm{d}} w_i\sigma_i^{-2}\sigma_j^{-2}\boldsymbol{e}_j^\mathsf{T}\mathbf{C}(\mathbf{I} + \mathbf{W}_\sigma\mathbf{C})^{-1}\boldsymbol{e}_i\boldsymbol{e}_i^\mathsf{T}\mathbf{D}(\mathbf{I} + \mathbf{W}_\sigma\mathbf{C})^{-1}\boldsymbol{e}_j - \sigma_j^{-2}\boldsymbol{e}_j^\mathsf{T}\mathbf{D}(\mathbf{I} + \mathbf{W}_\sigma\mathbf{C})^{-1}\boldsymbol{e}_j,
\end{aligned}$$

where we have used the cyclic property of the trace and the definition of $\mathbf{W}_\sigma$ in (2.6). Letting $\boldsymbol{y}_i = (\mathbf{I} + \mathbf{W}_\sigma\mathbf{C})^{-1}\boldsymbol{e}_i$, $i = 1, \ldots, n_\mathrm{d}$, and substituting in the above expression, leads to

$$\frac{\partial\Psi}{\partial w_j} = \sum_{i=1}^{n_\mathrm{d}} w_i\sigma_i^{-2}\sigma_j^{-2}(\boldsymbol{e}_j^\mathsf{T}\mathbf{C}\boldsymbol{y}_i)\boldsymbol{e}_i^\mathsf{T}\mathbf{D}\boldsymbol{y}_j - \sigma_j^{-2}\boldsymbol{e}_j^\mathsf{T}\mathbf{D}\boldsymbol{y}_j, \quad j = 1, \ldots, n_\mathrm{d}. \tag{4.3}$$

Note that the vectors $\boldsymbol{q}_i$ in (4.2) and vectors $\boldsymbol{y}_i$ in the definition of the gradient are related according to $\boldsymbol{q}_i = w_i\sigma_i^{-2}\boldsymbol{y}_i$, $i = 1, \ldots, n_\mathrm{d}$.

The matrices $\mathbf{C}$ and $\mathbf{D}$ in (4.2) and (4.3) are of size $n_\mathrm{d} \times n_\mathrm{d}$. As mentioned previously, in many cases, the measurement dimension n_d is considerably smaller than the dimension of the discretized primary and secondary parameters. This case typically arises in inverse problems governed by PDEs, where the dimension of the discretized parameters grow upon grid refinements, while the measurement dimension n_d is fixed a priori.

The matrices $\mathbf{C}$ and $\mathbf{D}$ can be built in a precomputation step, as outlined in Algorithm 4.1. The computational cost to build $\mathbf{C}$ and $\mathbf{D}$ is $3n_\mathrm{d}$ forward and $2n_\mathrm{d}$ adjoint PDE solves. Once the matrices $\mathbf{C}$ and $\mathbf{D}$ are computed, the OED objective and gradient evaluation can be performed without further PDE solves and require only linear algebra operations; see Algorithm 4.2. The cost of evaluating the objective function is dominated by the cost of steps 1–3, which amount to computing $\mathbf{Y} = (\mathbf{I} + \mathbf{W}_\sigma\mathbf{C})^{-1}$; this can be done in $\mathcal{O}(n_\mathrm{d}^3)$ arithmetic

Algorithm 4.1 Computing matrices $\mathbf{C}$ in (3.2) and $\mathbf{D}$ in (4.1) needed for mOED objective and gradient evaluation.

1: **for** $i = 1$ to n_d **do**
2: Compute $\boldsymbol{a}_i = \boldsymbol{\Gamma}_{\mathrm{pr,m}}\mathbf{F}^*\boldsymbol{e}_i$
3: Compute $\boldsymbol{d}_i = \mathbf{F}\boldsymbol{\Gamma}_{\mathrm{pr,m}}\boldsymbol{a}_i$ $\{\text{columns of } \mathbf{D} = \mathbf{F}\boldsymbol{\Gamma}^2_{\mathrm{pr,m}}\mathbf{F}^*\}$
4: Compute $\boldsymbol{c}_i = \mathbf{F}\boldsymbol{a}_i + \mathbf{G}\boldsymbol{\Gamma}_{\mathrm{pr,b}}\mathbf{G}^*\boldsymbol{e}_i$ $\{\text{columns of } \mathbf{C} = \mathbf{F}\boldsymbol{\Gamma}_{\mathrm{pr,m}}\mathbf{F}^* + \mathbf{G}\boldsymbol{\Gamma}_{\mathrm{pr,b}}\mathbf{G}^*\}$
5: **end for**
6: Build $\mathbf{C} = [\boldsymbol{c}_1 \quad \cdots \quad \boldsymbol{c}_{n_\mathrm{d}}]$ and $\mathbf{D} = [\boldsymbol{d}_1 \quad \cdots \quad \boldsymbol{d}_{n_\mathrm{d}}]$

operations, by precomputing an LU factorization of $\mathbf{I} + \mathbf{W}_\sigma\mathbf{C}$ and then performing triangular solves to compute columns of $\mathbf{Y}$. We also need the matrix-matrix product $\mathbf{DY}$ (see step 5 of Algorithm 4.2), which requires an additional $\mathcal{O}(n_d^3)$ operations. The additional effort in computing the gradient is dominated by one matrix-matrix product, $\mathbf{CY}$, amounting to $\mathcal{O}(n_\mathrm{d}^3)$ arithmetic operations.

Algorithm 4.2 Computing $\Psi(\boldsymbol{w})$ and its gradient $\nabla\Psi(\boldsymbol{w})$.

Input: Design vector $\boldsymbol{w}$.
Output: $\Psi = \Psi(\boldsymbol{w})$ and $\nabla\Psi = \nabla\Psi(\boldsymbol{w})$
1: `/* evaluation of the objective function */`
2: **for** $i = 1$ to n_d **do**
3: Solve the system $(\mathbf{I} + \mathbf{W}_\sigma\mathbf{C})\boldsymbol{y}_i = \boldsymbol{e}_i$
4: **end for**
5: Compute $\Psi = -\sum_{i=1}^{n_\mathrm{d}} w_i\sigma_i^{-2}\boldsymbol{e}_i^\mathsf{T}\mathbf{DY}\boldsymbol{e}_i$ $\{\mathbf{Y} = [\boldsymbol{y}_1 \quad \boldsymbol{y}_2 \quad \cdots \quad \boldsymbol{y}_{n_\mathrm{d}}]\}$
6: `/* evaluation of the gradient */`
7: **for** $j = 1$ to n_d **do**
8: Compute $\dfrac{\partial\Psi}{\partial w_j} = \sum_{i=1}^{n_\mathrm{d}} w_i\sigma_i^{-2}\sigma_j^{-2}(\boldsymbol{e}_j^\mathsf{T}\mathbf{CY}\boldsymbol{e}_i)(\boldsymbol{e}_i^\mathsf{T}\mathbf{DY}\boldsymbol{e}_j) - \sigma_j^{-2}\boldsymbol{e}_j^\mathsf{T}\mathbf{DY}\boldsymbol{e}_j$
9: **end for**

4.2. Sparsity control. Here we discuss several options for choosing the penalty function $P(\boldsymbol{w})$ in (3.8). A straightforward choice for $P(\boldsymbol{w})$ is the ℓ_1-norm of $\boldsymbol{w}$; see e.g., [14, 15]. As is well-known, the ℓ_1-penalty promotes sparsity, but not necessarily a binary structure, in the computed design vectors. Another option is to solve a sequence of optimization problems where penalty functions approximating ℓ_0-"norm" (the number of nonzero elements in a vector) are used. An example is the so-called regularized ℓ_0-sparsification approach proposed in [3]; in this approach, which we use in the present work, a continuation approach is used, and a sequence of optimization problems, with non-convex penalty functions approaching the ℓ_0-norm, are solved. A related approach is the use of reweighted ℓ_1-minimization, as done in [17]. Solving optimization problems with continuous weights, combined with a suitable penalty method, enables the use of powerful gradient-based optimization algorithms to explore the set of admissible designs. The effectiveness of such approaches in obtaining optimal sensor

placements has been demonstrated in a number of previous works; see e.g., [3, 14, 15, 17].

4.3. Greedy sensor placement. An alternative approach for finding sparse mOEDs is to use a greedy strategy. Greedy approaches have been used successfully in many sensor placement applications to obtain designs that, while suboptimal, provide near optimal performance; see e.g., [9, 18, 23, 31]. In a greedy approach, we place sensors one at a time: in each step, we select a sensor that provides the largest decrease in the design criterion. A greedy approach can be attractive due to its simplicity and the fact that it does not require the gradient of the design criterion. However, the computational complexity of greedy sensor placement, in terms of function evaluations, scales with the number of candidate sensor locations and the number of the sensors in the optimal design. Note that the computational cost, in terms of function evaluations, of finding a greedy sensor placement (in its most basic form) with K sensors is

$$C(K, n_{\mathrm{d}}) = Kn_{\mathrm{d}} - (K-1)K/2. \tag{4.4}$$

5. Model problem setup. To illustrate our approach for computing optimal designs under reducible uncertainty, we consider a linear inverse problem governed by a time-dependent advection-diffusion equation with two sources of uncertainty: the parameter of primary interest is a time-dependent scalar-valued function $m = m(t)$, which models the time amplitude of a source entering on the right hand side of the equation. The second uncertain parameter is the spatially distributed initial condition $b = b(\boldsymbol{x})$. Specifically, we consider:

$$u_t - \kappa\Delta u + \boldsymbol{v}\cdot\nabla u = \delta(\boldsymbol{x})m(t) \quad \text{in } \mathcal{D}\times\mathcal{T}, \tag{5.1a}$$

$$u(\cdot, 0) = b(\boldsymbol{x}) \quad \text{in } \mathcal{D}, \tag{5.1b}$$

$$\kappa\nabla u\cdot\boldsymbol{n} = 0 \quad \text{on } \partial\mathcal{D}\times\mathcal{T}. \tag{5.1c}$$

Here, $\mathcal{D}$ is a bounded open set in $\mathbb{R}^2$, the time interval $\mathcal{T} = (0, T)$, where $T > 0$ is a final time, $\kappa > 0$ is the diffusion coefficient, and $\boldsymbol{v}$ is a given velocity field. Note that the solution $u(\boldsymbol{x}, t)$, which can be interpreted as concentration, depends affinely on m and b. In our numerical experiments, $\kappa = 0.001$ and $\mathcal{D}$ is a unit square with two cutouts as shown in Figure 1 (left). If (5.1a) models the flow of a contaminant in a region, the cutouts could represent buildings, for instance. The velocity field $\boldsymbol{v}$ (shown in Figure 1) is obtained by solving Navier-Stokes equations with no-outflow boundary conditions and non-zero tangential boundary conditions as in [3]. The function δ in the source term is given by a mollified delta-function:

$$\delta(\boldsymbol{x}) = \left(\frac{1}{2\pi L} e^{-\frac{1}{2L^2}\|\boldsymbol{x}-\boldsymbol{x}_0\|^2}\right), \tag{5.2}$$

where the "correlation length" L is 0.05 in our experiments, and $\boldsymbol{x}_0 = (0.5, 0.35)$ as indicated by the red dot in Figure 1 (left).

5.1. Parameter-to-observable map. The parameter-to-observable map maps the time evolution of the right hand side amplitude, $m \in L^2(\mathcal{T})$ and the initial condition $b \in L^2(\mathcal{D})$ to point measurements of the solution of the advection-diffusion equation (5.1). To write it in the form (1.2), we define the continuous linear operators $\mathcal{S}_1$ and $\mathcal{S}_2$ as follows: $\mathcal{S}_1$ maps m to the PDE solution u, with $b = 0$, and $\mathcal{S}_2$ maps b to the PDE solution u, with $m = 0$. Then,

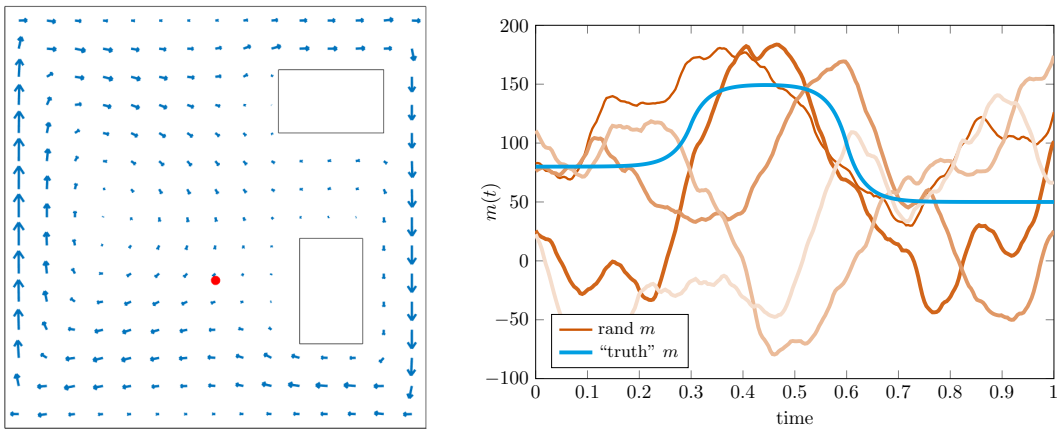

Figure 1. *Left: Sketch of domain $\mathcal{D}$ and velocity field $\boldsymbol{v}$ in (5.1). The red dot indicates the location $\boldsymbol{x}_0 = (0.5, 0.35)$ where the source term (5.2) is centered. Right: the "truth" source term m and five samples from the prior distribution of m shown in cyan and various shades of orange, respectively.*

the solution to the initial-boundary value problem (5.1) can be written as $u = \mathcal{S}_1 m + \mathcal{S}_2 b$; see [36, p.152]. Next, let $\mathcal{B}$ be a linear observation operator that extracts the values of $u(\boldsymbol{x}, t)$ on a set of sensor locations $\{\boldsymbol{x}_1, \boldsymbol{x_2}, \ldots, \boldsymbol{x}_{n_\mathrm{d}}\} \in \mathcal{D}$, and takes an average of u over the time interval $[0.95, 0.99]$. Then $\mathcal{F} = \mathcal{B}\mathcal{S}_1$ and $\mathcal{G} = \mathcal{B}\mathcal{S}_2$ map the primary inference parameter m and the additional uncertain parameter b to measurement $\boldsymbol{y} \in \mathbb{R}^{n_\mathrm{d}}$:

$$\mathcal{F} : m(t) \overset{\mathcal{S}_1}{\longmapsto} u(\boldsymbol{x}, t) \overset{\mathcal{B}}{\longmapsto} \boldsymbol{y}, \qquad \mathcal{G} : b(\boldsymbol{x}) \overset{\mathcal{S}_2}{\longmapsto} u(\boldsymbol{x}, t) \overset{\mathcal{B}}{\longmapsto} \boldsymbol{y}. \tag{5.3}$$

The corresponding discrete parameter-to-observable maps $\mathbf{F}$ and $\mathbf{G}$ are obtained through discretization using, for instance, finite elements.

Computations of derivatives of an objective that involves the parameter-to-observable map requires the adjoint operators $\mathcal{F}^*$ and $\mathcal{G}^*$. These can be derived using the formal Lagrangian method, resulting in the following adjoint equations [36]. Given a vector of observations $\boldsymbol{y} \in \mathbb{R}^{n_\mathrm{d}}$, we first solve the adjoint equation (see [1, 3]) for the adjoint variable $p = p(\boldsymbol{x}, t)$

$$-p_t - \nabla \cdot (p\boldsymbol{v}) - \kappa \Delta p = -\mathcal{B}^* \boldsymbol{y} \quad \text{in } \mathcal{D} \times \mathcal{T}, \tag{5.4a}$$

$$p(\cdot, T) = 0 \quad \text{in } \mathcal{D}, \tag{5.4b}$$

$$(\boldsymbol{v}p + \kappa \nabla p) \cdot \boldsymbol{n} = 0 \quad \text{on } \partial\mathcal{D} \times \mathcal{T}, \tag{5.4c}$$

and obtain the action of the adjoint operators as $\mathcal{F}^* \boldsymbol{y} = -\int_{\mathcal{D}} f(\boldsymbol{x}) p(\boldsymbol{x}, \cdot) d\boldsymbol{x}$ and $\mathcal{G}^* \boldsymbol{y} = -p(\cdot, 0)$.

5.2. Prior laws of m and b. To complete the definition of the Bayesian inverse problem, we specify the prior laws for m and b. We assume both to be Gaussian random fields, and thus it is sufficient to specify the mean and covariance operator. For the primary parameter m, which is a function of time only, we choose the mean to be the constant function $m_\mathrm{pr} \equiv 65$, and specify the covariance operator $\Gamma_{\mathrm{pr,m}}$ according to

$$[\Gamma_{\mathrm{pr,m}} z](t) = \int_{\mathcal{T}} c(s, t) z(t)\, dt, \quad z \in L^2(\mathcal{T}),$$

where we chose the Matérn-3/2 covariance kernel

$$c(s,t) = \sigma^2 \left(1 + \frac{\sqrt{3}|s-t|}{\ell}\right) \exp\left(-\frac{\sqrt{3}|s-t|}{\ell}\right). \tag{5.5}$$

This covariance function ensures that draws from the prior law of m are (almost surely) contintinuously differentiable; see, e.g., [16, 24, 38]. In our numerical experiments, we use the parameters $\sigma = 80$ and $\ell = 0.17$ in (5.5). Samples from the resulting distribution are shown in Figure 1 (right).

The realizations of the secondary parameter b are functions defined over the spatial domain $\mathcal{D}$. For the distribution of b we choose a Gaussian with mean $b_{\text{pr}} \equiv 50$, and a Laplacian-like covariance operator of the form $(-\epsilon\Delta + \alpha I)^{-2}$ [33], with $\epsilon = 4.5 \times 10^{-3}$ and $\alpha = 2.2 \times 10^{-1}$. We equip the Laplace operator with homogeneous Robin boundary conditions with constant coefficient. We do this to mitigate undesired boundary effects that can arise when PDE operators are used to define covariance operators [12, 29].

5.3. Discretization. We discretize the forward problem using linear finite elements on triangular meshes in space and use the implicit Euler method in time. This guides the discretization of the primary and secondary uncertainties m and b. Specifically, the discretized uncertain source terms is the vector $\boldsymbol{m}$ whose entries are the values of m at the time-steps used by the forward solver. We discretize the $L^2(\mathcal{T})$ inner product using quadrature. That is, for $f, g \in L^2(\mathcal{T})$,

$$\langle f, g\rangle_1 = \int_{\mathcal{T}} f(t)g(t)\,dt \approx \sum_{j=1}^{n_{\text{m}}} \nu_j f(t_j)g(t_j) = \boldsymbol{f}^{\mathsf{T}}\mathbf{M}_1\boldsymbol{g} =: \langle \boldsymbol{f}, \boldsymbol{g}\rangle_{\mathrm{M}_1},$$

where $\{\nu_j\}_{j=1}^{n_{\text{m}}}$ are quadrature weights, $\boldsymbol{f}$ and $\boldsymbol{g}$ are vectors (in $\mathbb{R}^{n_{\text{m}}}$) of function values at the time-steps, and $\mathbf{M}_1 = \text{diag}(\nu_1, \nu_2, \ldots, \nu_{n_{\text{m}}})$. In the present work, we use the composite trapezoid rule to discretize the $L^2(\mathcal{T})$ inner product.

The uncertain initial state b is discretized using finite element Lagrange nodal basis functions, $\varphi_1(\boldsymbol{x})$, $\ldots$, $\varphi_{n_{\text{b}}}(\boldsymbol{x})$. This leads to the discretization $b(\boldsymbol{x}) \approx b_h(\boldsymbol{x}) = \sum_{j=1}^{n_{\text{b}}} b_i\varphi_i(\boldsymbol{x})$. The discretized initial state is given by the vector $\boldsymbol{b}$ of finite-element coefficients. This finite element method is also used to discretize the PDE operator $(-\epsilon\Delta + \alpha I)$, which is the square root of the covariance operator of the distribution of $\boldsymbol{b}$. The covariance operator is thus defined as the square of the finite element operator, corresponding to a mixed discretization of the 4th-order covariance operator [7]. Also, note that the discretized $L^2(\mathcal{D})$-inner product is given by $\langle \boldsymbol{u}, \boldsymbol{v}\rangle_{\mathrm{M}_2} = \boldsymbol{u}^{\mathsf{T}}\mathbf{M}_2\boldsymbol{v}$, for $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{n_{\text{b}}}$, where $\mathbf{M}_2$ is the finite-element mass matrix.

In the numerical experiments below, we use a discretization with $n_{\text{m}} = 257$ time steps and $n_{\text{b}} = 1{,}529$ spatial degrees of freedom. The "truth" primary parameter m is shown in Figure 1 (right), and the "truth" secondary parameter b is given by a random draw from the prior law of b, depicted in Figure 2 (top left). For computing solutions for the inverse problem, we synthesize data using "truth" parameters b and m, and add Gaussian noise with standard deviation $\sigma_{\text{noise}} = 0.25$ to each data point. That is, we assume $\boldsymbol{\Gamma}_{\text{noise}} = \sigma_{\text{noise}}^2\mathbf{I}$, with $\sigma_{\text{noise}} = 0.25$. Notice that the sensor measurements obtained from the model range approximately in the interval $[51, 54]$; see e.g., Figure 2 (top right). Thus, a noise standard deviation of 0.25 is significant compared to the variations of model output at the sensors.

5.4. Illustrating the impact of the secondary uncertainty. To depict the impact of the secondary uncertainty on the solution of the forward problem, in Figure 2 we show snapshots of the solution of the state equation. Here, we use two random draws from the prior distribution of b, i.e., the secondary uncertainty, as initial conditions. Recall that the initial condition used for the first row is also used as "truth" secondary parameter. For the primary uncertainty, the time evolution of the right hand side source, the "truth" parameter (see Figure 1 (right)) is used. Note that even at the final snapshot, around which measurements are taken for inference, distinct differences caused by the different initial conditions are visible. This indicates that the uncertainty in the initial state cannot be ignored.

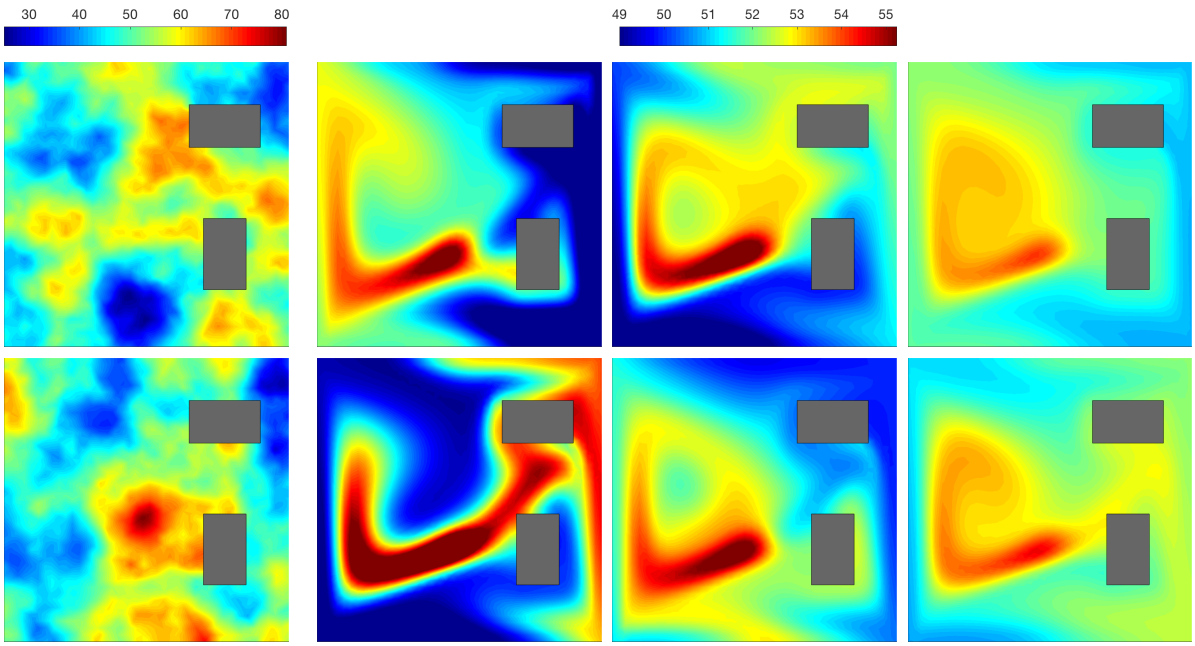

Figure 2. *Shown in each row are snapshots of the concentration at times $t = 0, 0.4, 0.6, 1$ (from left to right). For the primary parameter m entering on the right hand side of (5.1a), the "truth" parameter shown in Figure 1 (right) is used. For the secondary parameter b, i.e., the initial condition, two different realizations from the distribution of b are used. Note that a different colorbar is used for the initial conditions than for the other snapshots.*

6. Computational results. In this section, we present numerical results for the model problem described in section 5. In subsection 6.1, we compare the performance of regularized ℓ_0-sparsification and greedy approaches for computing mOEDs. Then, in subsections 6.2 and 6.3, we demonstrate the importance of taking the additional model uncertainty into account for computing sensor placements.

6.1. Comparison of sparsification algorithms. Here, we compare the two different approaches to obtain binary mOEDs discussed in section 4. As discussed in subsection 4.2, when using ℓ_0-sparsification we solve a sequence of optimization problems with non-convex penalty functions using a gradient-based optimization algorithm. Here, we use MATLAB's

interior point quasi-Newton solver provided by the `fmincon` function, which we supply with routines implementing the mOED objective and its gradient. In contrast, the greedy approach only requires the mOED objective. As can be seen in Figure 3 (left), the greedy and the ℓ_0-sparsified designs perform similary. While in this figure the ℓ_0-sparsification finds slightly lower objective values, we have also observed tests where the objective values are identical or the greedy approach is slightly better.

It is also important to consider the computational cost of these algorithms. We do so by recording the number of mOED objective function evaluations required by the two algorithms in Figure 3 (right). Note that the cost of greedy sensor placement scales with the number of sensors in the optimal design, see also (4.4). The cost of the ℓ_0-sparsification, in terms of function evaluations, remains nearly constant. Of course, the regularized ℓ_0-sparsification method requires gradients additionally to objective evaluations. However, as discussed in subsection 4.1, the additional cost of computing the gradient is small compared to the cost of mOED objective function evaluation. Therefore, the number of objective function evaluations is a reasonable measure to compare the cost of the two algorithms.

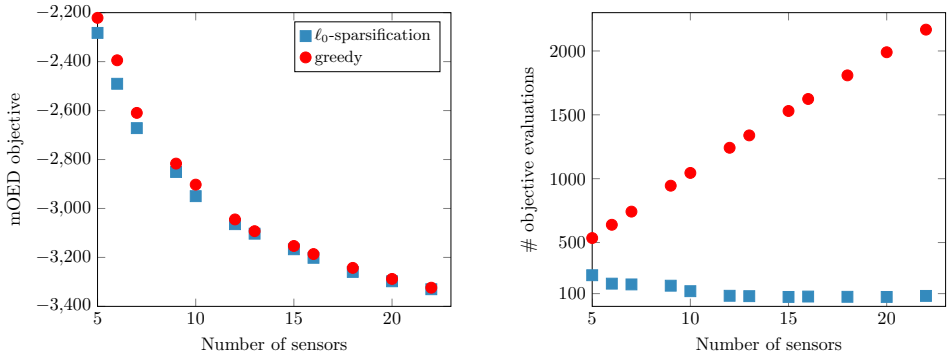

Figure 3. *Left: mOED objective values (y-axis) plotted against number of sensors (x-axis) for the greedy (red dots) and the ℓ_0-sparsification approaches (blue dots). Right: Number of mOED objective evaluations required to converge for computing greedy (red) and ℓ_0-sparsified (blue) designs.*

In the remainder of this section, where we compare the performance of designs obtained with and without marginalization, we use the greedy approach to find optimal designs. This is motivated by the fact that the greedy approach facilitates computing (near) optimal designs with a desired number of sensors, while the ℓ_0-sparsification approach only provides indirect control on the number of sensors by changing the penalty parameter γ.

6.2. Studying the posterior uncertainty. Next, we compare the performance of designs obtained by performing mOED against those using OED with no marginalization in terms of the resulting marginal posterior uncertainty. Note that designs obtained without marginalization, which we simply refer to as OED, minimize the classical A-optimality criterion ψ in (2.9) whereas designs with marginalization minimize the mOED criterion in (3.6).

Figure 4 shows two designs with 20 sensors, one taking into account the secondary uncertainty through marginalization, and one assuming that there is no secondary uncertainty. On the right panel of Figure 4, the pointwise standard deviation of the marginalized posterior

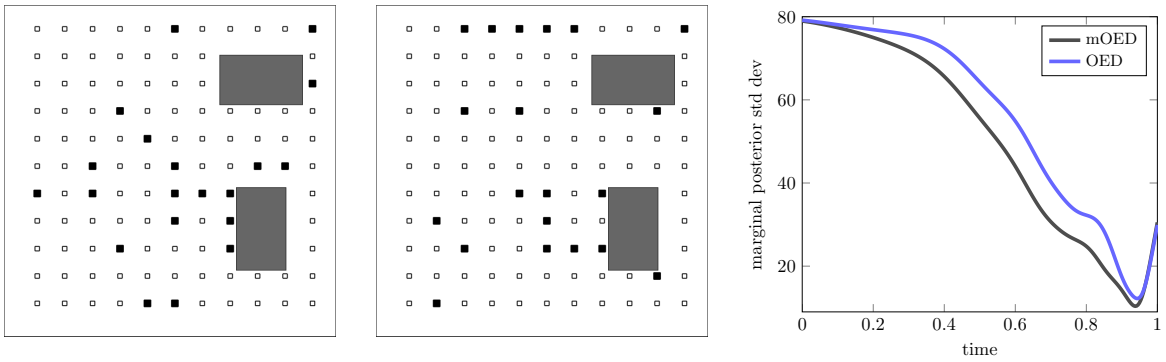

Figure 4. *Shown are A-optimal designs with 20 sensors (filled squares) using mOED (left) and OED without marginalization (center), i.e., the design obtained with OED neglecting secondary uncertainties. Inactive sensors are shown as empty squares. On the right, the marginal posterior standard deviation field (i.e., square root of the diagonal of* $\mathbf{\Gamma}_{\text{post,m}}(\boldsymbol{w})$ *in (3.3)) is shown for the two designs.*

distribution are shown for the two sensor placements. The following conclusions can be drawn. First, note that mOED is superior, with respect to the marginalized posterior variance, to the design computed without taking the secondary uncertainty into account. Of course, this is by construction of mOEDs. However, the difference is significant and exists for all times $t \in \mathcal{T}$. Second, since measurements are taken around the final time, the uncertainty is more reduced for later times. However, close to the final time T, the uncertainty increases again as there is not enough time for the concentration field to propagate to and be picked up by sensors.

6.3. Study of MAP points. Next, we compare MAP points computed with the mOED and OED designs shown in Figure 4. Note that the MAP point for mOED does not depend on a realization of the secondary parameter (see (3.3)), while it does for OED without marginalization (see (2.8)). In Figure 5, we show the MAP point for the mOED, which recovers features from the "truth" parameter but resorts to the prior mean when little information can be gathered from observations.

As mentioned above, we need a realization of the secondary parameter b when computing the MAP point using the classical OED. If we knew the "truth" b, the additional uncertainty would vanish and the problem reduces to an inverse (and OED) problem with fully specified model as, e.g., in [3]. The corresponding MAP point, shown in blue in Figure 5, slightly improved compared to the MAP point from the mOED formulation. However, in general the "truth" secondary parameter is unknown, and we only know its distribution. If random draws from the secondary parameter distribution are used in the MAP computation, the model error is underestimated and the corresponding MAP points may be poor. This can be seen in Figure 5, where MAP points obtained with random draws from the distribution of b are shown in red.

The above discussed difference between mOED and OED without marginalization is summarized in Figure 6. On the left, we plot the relative $L^2(\mathcal{T})$-error between the MAP point and the "truth" primary parameter versus the mOED objective. Using OED with random draws for b result in MAP points that tend to be further from the "truth" parameter than

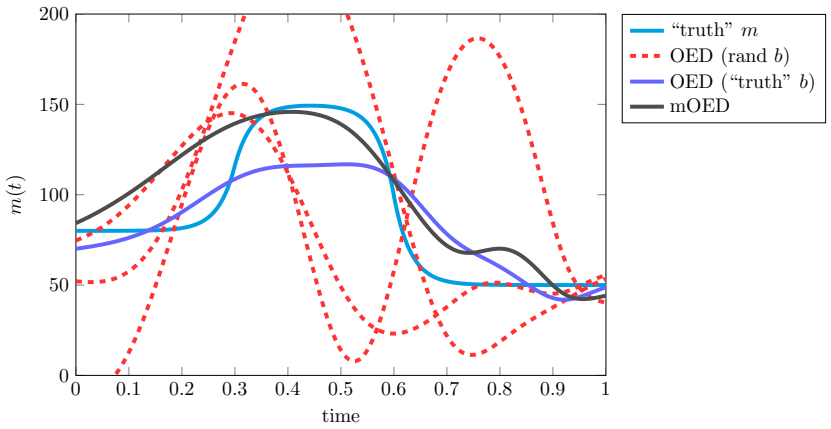

Figure 5. *Comparison of MAP estimates computed with mOED and OED without marginalization. Shown are the MAP estimates computed using sensor placements obtained via mOED (black solid line), OED with the secondary parameter b set to the "truth" (blue solid line), and OED with b taken as realizations from corresponding prior distribution (red dotted lines).*

the mOED MAP point. If the "truth" secondary parameter is used in the computation of the MAP point using OED, the reconstruction is slightly better than the result of mOED. It can also be seen that the mOED objective is independent from draws of the secondary parameter, as also discussed above. The results in Figure 6 (left) depend on the noise realizations in the synthetic data. In Figure 6 (right), we show the probability density function of the error between the MAP point and the "truth" primary parameter for random observation noise. As can be seen, it is slightly more likely to obtain a better MAP point when using OED with the "truth" parameter than with mOED. However, it can clearly be seen that mOED MAP points significantly outperform OED MAP points with random realizations from the prior distribution of b.

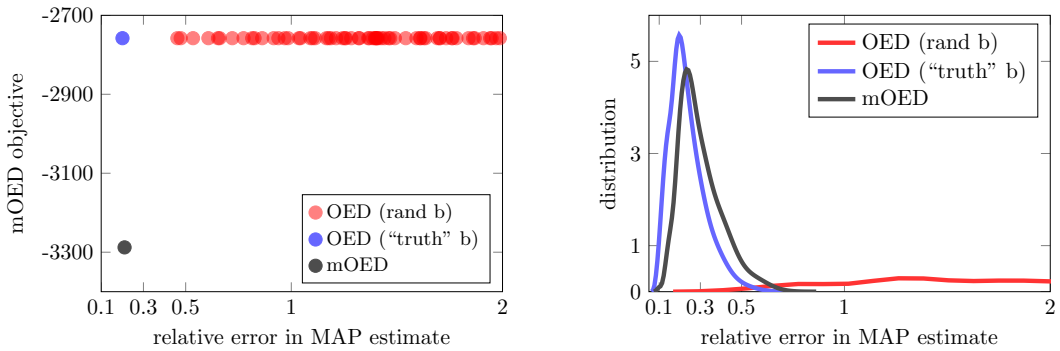

Figure 6. Left: *Relative error in the MAP estimate (x-axis) and reduction in the objective (y-axis) for mOED (black dot), OED with the secondary parameter b set to the "truth" (blue dot), and OED with b taken as different realizations of b (red dots).* Right: *The distribution of the errors with various realizations of the noise in the data. Note that the x-axis is cut at 2 due to the long tail of the error distribution corresponding to OED with b taken as different realizations of b. In this study, we used* 200 *samples of the secondary parameter, and* 500 *samples of measurement noise.*

7. Conclusion. In this article, we have considered linear inverse problems with reducible model uncertainty and presented a mathematical and computational framework for computing marginalized A-optimal sensors placements. Our results show that it is important to take into account additional sources of model uncertainty for the optimal design and the inverse problem in general. The designs computed by minimizing the marginalized A-optimality criterion are superior compared to classical A-optimal designs, in terms of the quality of the estimated primary parameters: the marginalized optimal designs result in optimal uncertainty reduction as well as more accurate MAP estimates. The overall conclusions support the claim made in this article's title, namely that in the context of design of inverse problems, it is good to know what you don't know. This information should be used when computing optimal designs.

An important direction for future work is design of nonlinear inverse problems under model uncertainty. A related direction is a sensitivity analysis framework for detecting sources of model uncertainty that are most important to the solution of the inverse problem. This would enable incorporating only the most important sources of model uncertainty in the OED problem, hence reducing the computational complexity of the problem. For deterministic inverse problems, first steps in this direction are presented in [34].

REFERENCES

[1] V. Akçelik, G. Biros, A. Drăgănescu, O. Ghattas, J. Hill, and B. van Bloemen Waanders, *Dynamic data-driven inversion for terascale simulations: Real-time identification of airborne contaminants*, in Proceedings of SC2005, Seattle, 2005.

[2] A. Alexanderian, *Optimal experimental design for Bayesian inverse problems governed by PDEs: A review*, Preprint, (2020). https://arxiv.org/abs/2005.12998.

[3] A. Alexanderian, N. Petra, G. Stadler, and O. Ghattas, *A-optimal design of experiments for infinite-dimensional Bayesian linear inverse problems with regularized ℓ_0-sparsification*, SIAM J. Sci. Comput., 36 (2014), pp. A2122–A2148.

[4] A. Y. Aravkin and T. Van Leeuwen, *Estimating nuisance parameters in inverse problems*, Inverse Problems, 28 (2012), p. 115016.

[5] A. C. Atkinson and A. N. Donev, *Optimum Experimental Designs*, Oxford, 1992.

[6] A. Attia, A. Alexanderian, and A. K. Saibaba, *Goal-oriented optimal design of experiments for large-scale Bayesian linear inverse problems*, Inverse Problems, 34 (2018), p. 095009.

[7] T. Bui-Thanh, O. Ghattas, J. Martin, and G. Stadler, *A computational framework for infinite-dimensional Bayesian inverse problems. Part I: The linearized case, with application to global seismic inversion*, SIAM J. Sci. Comput., 35 (2013), pp. A2494–A2523.

[8] K. Chaloner and I. Verdinelli, *Bayesian experimental design: A review*, Statist. Sci., 10 (1995), pp. 273–304.

[9] L. Chamon and A. Ribeiro, *Approximate supermodularity bounds for experimental design*, in Advances in Neural Information Processing Systems, 2017, pp. 5403–5412.

[10] E. M. Constantinescu, N. Petra, J. Bessac, and C. G. Petra, *Statistical treatment of inverse problems constrained by differential equations-based models with stochastic terms*, SIAM/ASA J. Uncertain. Quantif., 8 (2020), pp. 170–197.

[11] G. Da Prato, *An introduction to infinite-dimensional analysis*, Springer Science & Business Media, 2006.

[12] Y. Daon and G. Stadler, *Mitigating the influence of boundary conditions on covariance operators derived from elliptic PDEs*, Inverse Probl. Imaging, 12 (2018), pp. 1083–1102.

[13] J. Fohring and E. Haber, *Adaptive A-optimal experimental design for linear dynamical systems*, SIAM/ASA J. Uncertain. Quantif., 4 (2016), pp. 1138–1159.

[14] E. Haber, L. Horesh, and L. Tenorio, *Numerical methods for experimental design of large-scale linear ill-posed inverse problems*, Inverse Problems, 24 (2008), pp. 125–137.

[15] E. Haber, Z. Magnant, C. Lucero, and L. Tenorio, *Numerical methods for A-optimal designs with a sparsity constraint for ill-posed inverse problems*, Comput. Optim. Appl., (2012), pp. 1–22.
[16] M. S. Handcock and M. L. Stein, *A Bayesian analysis of kriging*, Technometrics, 35 (1993), pp. 403–410.
[17] E. Herman, A. Alexanderian, and A. K. Saibaba, *Randomization and reweighted ℓ_1-minimization for A-optimal design of linear inverse problems*, SIAM J. Sci. Comput., accepted (2020). https://arxiv.org/abs/1906.03791.
[18] J. Jagalur-Mohan and Y. Marzouk, *Batch greedy maximization of non-submodular functions: Guarantees and applications to experimental design*, Preprint, (2020). https://arxiv.org/abs/2006.04554.
[19] J. Kaipio and V. Kolehmainen, *Approximate marginalization over modeling errors and uncertainties in inverse problems*, Bayesian Theory and Applications, (2013), pp. 644–672.
[20] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, vol. 160 of Applied Mathematical Sciences, Springer-Verlag, New York, 2005.
[21] V. Kolehmainen, T. Tarvainen, S. R. Arridge, and J. P. Kaipio, *Marginalization of uninteresting distributed parameters in inverse problems-application to diffuse optical tomography*, Int. J. Uncertain. Quantif., 1 (2011).
[22] K. Koval, A. Alexanderian, and G. Stadler, *Optimal experimental design under irreducible uncertainty for linear inverse problems governed by PDEs*, Inverse Problems, accepted (2020). https://arxiv.org/abs/1912.08915.
[23] A. Krause, A. Singh, and C. Guestrin, *Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies*, J. Mach. Learn. Res., 9 (2008), pp. 235–284.
[24] F. Lindgren, H. Rue, and J. Lindström, *An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach*, J. R. Stat. Soc. Ser. B. Stat. Methodol., 73 (2011), pp. 423–498.
[25] T.-T. Lu and S.-H. Shiou, *Inverses of 2×2 block matrices*, Comput. Math. Appl., 43 (2002), pp. 119–129.
[26] J. B. Nagel, *Bayesian techniques for inverse uncertainty quantification*, PhD thesis, ETH Zurich, 2017.
[27] R. Nicholson, N. Petra, and J. P. Kaipio, *Estimation of the Robin coefficient field in a Poisson problem with uncertain conductivity field*, Inverse Problems, 34 (2018), p. 115005.
[28] J. M. Ortega, *Matrix theory: a second course*, The University Series in Mathematics, Plenum Press, New York, 1987.
[29] L. Roininen, J. M. Huttunen, and S. Lasanen, *Whittle-Matérn priors for Bayesian statistical inversion with applications in electrical impedance tomography*, Inverse Probl. Imaging, 8 (2014), p. 561.
[30] L. Ruthotto, J. Chung, and M. Chung, *Optimal experimental design for inverse problems with state constraints*, SIAM J. Sci. Comput., 40 (2018), pp. B1080–B1100.
[31] G. Shulkind, L. Horesh, and H. Avron, *Experimental design for nonparametric correction of misspecified dynamical models*, SIAM/ASA J. Uncertain. Quantif., 6 (2018), pp. 880–906.
[32] R. C. Smith, *Uncertainty quantification: Theory, implementation, and applications*, vol. 12 of Computational Science and Engineering Series, SIAM, 2013.
[33] A. M. Stuart, *Inverse problems: A Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.
[34] I. Sunseri, J. Hart, B. van Bloemen Waanders, and A. Alexanderian, *Hyper-differential sensitivity analysis for inverse problems constrained by partial differential equations*, Preprint, (2020). https://arxiv.org/abs/2003.00978.
[35] Y. L. Tong, *The multivariate normal distribution*, Springer Science & Business Media, 2012.
[36] F. Tröltzsch, *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*, vol. 112 of Graduate Studies in Mathematics, American Mathematical Society, 2010.
[37] D. Uciński, *Optimal measurement methods for distributed parameter system identification*, CRC Press, Boca Raton, 2005.
[38] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*, vol. 2, MIT press Cambridge, MA, 2006.
[39] D. Williams, *Probability with martingales*, Cambridge university press, 1991.