Geophysical Journal International

Geophys. J. Int. (2025) **240,** 1802–1824
Advance Access publication 2025 January 17
GJI Seismology

https://doi.org/10.1093/gji/ggae458

Analysis and optimization of seismic monitoring networks with Bayesian optimal experimental design

Jake Callahan [1,2] Kevin Monogue,[3] Ruben Villarreal[1] and Tommie Catanach[1]

[1] *Computational Data Science, Sandia National Laboratories, Livermore, CA* 94550, *USA. E-mail: jpcalla@sandia.gov*
[2] *Program in Applied Mathematics, University of Arizona, Tucson, AZ* 85721, *USA*
[3] *Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA* 94305, *USA*

Accepted 2024 December 13. Received 2024 December 5; in original form 2024 September 16

SUMMARY

Monitoring networks increasingly aim to assimilate data from a large number of diverse sensors covering many sensing modalities. Bayesian optimal experimental design (OED) seeks to identify data, sensor configurations or experiments which can optimally reduce uncertainty and hence increase the performance of a monitoring network. Information theory guides OED by formulating the choice of experiment or sensor placement as an optimization problem that maximizes the expected information gain (EIG) about quantities of interest given prior knowledge and models of expected observation data. Therefore, within the context of seismo-acoustic monitoring, we can use Bayesian OED to configure sensor networks by choosing sensor locations, types and fidelity in order to improve our ability to identify and locate seismic sources. In this work, we develop the framework necessary to use Bayesian OED to optimize a sensor network's ability to locate seismic events from arrival time data of detected seismic phases at the regional-scale. This framework requires five elements: (i) A likelihood function that describes the distribution of detection and traveltime data from the sensor network, (ii) A prior distribution that describes *a priori* belief about seismic events, (iii) A Bayesian solver that uses a prior and likelihood to identify the posterior distribution of seismic events given the data, (iv) An algorithm to compute EIG about seismic events over a data set of hypothetical prior events, (v) An optimizer that finds a sensor network which maximizes EIG. Once we have developed this framework, we explore many relevant questions to monitoring such as: how to trade off sensor fidelity and earth model uncertainty; how sensor types, number and locations influence uncertainty; and how prior models and constraints influence sensor placement.

Key words: Bayesian inference; Statistical methods; Earthquake monitoring and test-ban treaty verification; Earthquake source observations; Seismic noise; Statistical seismology.

1 INTRODUCTION

Seismo-acoustic monitoring networks are central to detecting and locating earthquakes, explosions or other seismic sources. In order to improve monitoring capabilities, network designers may incorporate new sensors or data types into the network to reduce detection thresholds or improve estimate uncertainties for quantities of interest (QoIs) like location, magnitude and depth. To estimate a QoI, modern processing algorithms often employ Bayesian inference because it provides rigorous uncertainty quantification to support decision making (Myers *et al.* 2007; Arora *et al.* 2013). Therefore, when designing or analysing a monitoring network, we approach it from the philosophy of Bayesian optimal experimental design (OED) (Lindley 1956; Krause *et al.* 2008; Huan & Marzouk 2013). Within this framework, we optimize sensors of a monitoring network to reduce uncertainty about QoIs under different conditions described by a prior distribution. Therefore, with Bayesian OED we not only design an effective monitoring network, but also get an understanding of the expected performance of that network under the specified conditions. While the target application of this research is explosion monitoring, the experimental design framework we have developed applies to arbitrary seismic sources. Therefore, this framework may support other applications of seismic networks such as earthquake seismology, earthquake or tsunami early warning or exploration geophysics.

OED has been a recent active area of research in many areas of seismology including early warning, seismic source inversion, tomography and structural health monitoring. Typically these studies have focused on network design in terms of the number and location of sensors (Papadimitriou *et al.* 2005; Guest & Curtis 2011; Yuen & Kuok 2015; An *et al.* 2018; Bloem *et al.* 2020; Toledo *et al.* 2020; Böse *et al.* 2022; Yang *et al.* 2022), although some work has also explored different sensor types (Yuen & Kuok 2015). For linear inverse problems, or

those that can be linearly approximated, the alphabetic optimality criteria like D-optimal design are often used as an objective (Steinberg & Rabinowitz 2003; Coles & Curtis 2011; Burmin 2019; Bloem *et al.* 2020; Koval 2021). For nonlinear inverse problems, Bayesian methods have become popular, leveraging information-based metrics such as entropy-based design or mutual information (Maurer *et al.* 2010; Long *et al.* 2015; Bloem *et al.* 2020; Yang *et al.* 2022). However, because of the computational cost of these methods, many different approximations methods have been explored to speed up estimating the objective (Maurer *et al.* 2010; Coles & Prange 2012; Long *et al.* 2015). Finally, work has also explored different optimizers to explore the configuration space of networks ranging from genetic algorithms to gradient-based methods (Curtis *et al.* 2004; Papadimitriou *et al.* 2005; Oth *et al.* 2010; Guest & Curtis 2011; Toledo *et al.* 2020; Böse *et al.* 2022). All these considerations lead to a trade-off between computational tractability and accuracy which has started to be explored.

Our work adds to this body of recent research through the following contributions:

(i) Presenting a holistic treatment of uncertainty (e.g. model error, measurement error, sensor correlation, etc.) for the Bayesian OED problem for seismic monitoring
(ii) Studying OED in a broader context than just sensor placement, for example, relative trade-offs in model refinement versus data fidelity.
(iii) Introducing a Bayesian optimization algorithm to efficiently optimize the sensor network.
(iv) Releasing a computationally efficient grid method for fully Bayesian OED leveraging HPC that can be widely used for seismo-acoustic monitoring network design and analysis.

This work quantifies the sensitivity of a seismo-acoustic monitoring network for inferring the location and magnitude of seismic sources that include shallow earthquakes and explosions either on the surface or underground. We then present a Bayesian OED algorithm to improve the monitoring network sensitivity by optimizing the location of ground motion sensors. Our computational approach combines information and Bayesian probability theory to quantify and optimize the sensitivity of our sensor network by estimating the information gain Bayesian inference provides about QoIs. This approach includes four distinct analysis stages:

(i) Build the likelihood function to estimate the probability of data, given a seismic event.
(ii) Solve the Bayesian inference problem for locating events given data (e.g. solve for the posterior).
(iii) Estimate seismic source location sensitivity through measuring the expected information gain, with a sensor network.
(iv) Optimize the seismic monitoring network to improve the expected information gain over seismic source events.

We use observational data from the U.S. Transportable Array (IRIS Transportable Array 2003) and physics-based models to build the Bayesian-likelihood functions. These likelihood functions incorporate many sources of uncertainty and model the behaviour of the seismic sensor network. This model defines how well Bayesian inference can assimilate sensor data to locate seismic sources. For our sensor data, we consider spatially correlated traveltimes for seismic phase arrivals detected at our sensor network. We make generally justifiable assumptions on our uncertainty models that match properties of the data sets that we consider. While we present our method using seismic *P*-wave arrivals, our approach is flexible to any event or signature data because it only requires that we can construct likelihood functions.

We study how the optimized sensor configuration and network sensitivity change under different design conditions and uncertainty models. We present the dependence of our results over prior sensor distribution, sensor number and sensor fidelity. This analysis provides a framework that we can later extend to optimize sensor networks that measure other natural and explosion signatures (e.g. electromagnetic or infrasound signals), which supports a more comprehensive need for multiphenomenology explosion monitoring (e.g. Arrowsmith *et al.* 2020; Carmichael *et al.* 2020). This framework explores an alternative approach to existing tools, like Sandia National Laboratory's NetMOD (Merchant 2013), with the aim to provide a highly flexible and rigorous framework for analysing and optimizing monitoring networks. This rigour is justified through our usage of Bayesian probability theory and uncertainty quantification.

In Section 2, we describe the Bayesian inference and optimal experimental design problems in general. In Section 3 we describe the specifics of a Bayesian inference problem to identify the location and magnitude of seismic sources, using records of their *P*-phase arrivals at distributed receivers and demonstrate how to build the likelihood models from these data. Next, in Section 4 we describe the algorithms used for solving the Bayesian OED problem. Finally, in Section 5 we will describe several experiments that demonstrate the utility of this approach and identify areas for further exploration. Section 6 concludes with discussion and future work.

2 BAYESIAN METHODS

2.1 Bayesian inference

Bayesian probability theory provides a rigorous methodology to quantify and update uncertainty about beliefs (Gelman *et al.* 1995; Jaynes 2003; Beck 2010). Within this framework, uncertainty is represented using probability distributions. Therefore, within the Bayesian paradigm, probability distributions represent uncertainty about beliefs and not necessarily intrinsic stochasticity and thus are not directly tied to randomness. Uncertainty comes from both epistemic sources, when it represents a lack of knowledge about learnable phenomena (ignorance), or aleatory sources, when it represents uncertainty about inherently unknowable randomness (unresolvable uncertainty for the observer). The Bayesian perspective describes both of these sources of uncertainty using a probability distribution. Therefore, the Bayesian framework can helpfully incorporate modelling error, measurement error and parametric uncertainty.

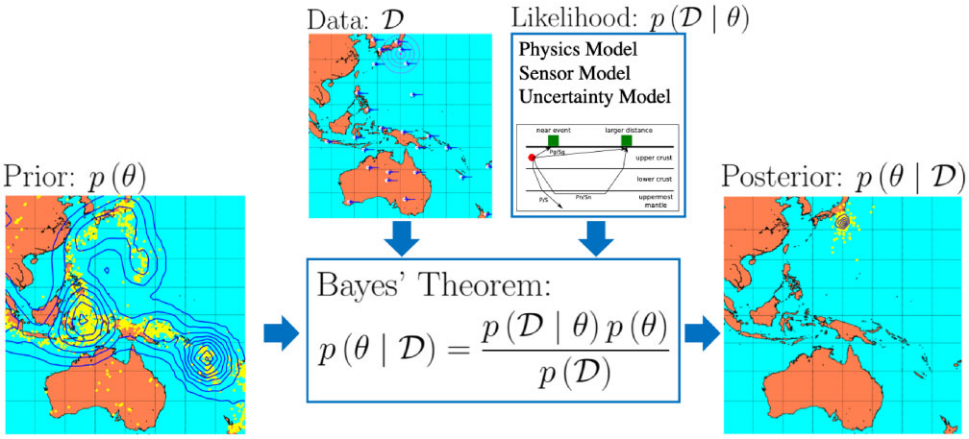

Figure 1. Illustration of a Bayesian inference process for seismic source location. Bayesian inference begins with a prior distribution for different earthquake locations θ, shown by the contour lines on the leftmost figure. As an observer collects data, they use a likelihood function model to quantify the probability of observing that data, given that an earthquake occurs at a specific location. The observer constructs this likelihood model from physical models of seismic wave propagation, models of the sensors that detect seismic signals and models of uncertainty (e.g. background noise modelling errors, etc.). The observer then applies Bayes' theorem to update the prior to assimilate this new information. The posterior distribution, shown by the contour lines in the rightmost image, then quantifies the probability that the seismic source has location θ, given the data.

As data or other sources of information become available, an observer can integrate these information into a new probability distributions to update the observer's beliefs. When the observer makes predictions, they include the uncertainty represented by these probability distributions in these predictions. The rules of Bayesian probability provide a rigorous logic for updating and propagating uncertainty just as binary logic provides rules for working with statements that are true or false.

The process of updating beliefs using data are known as Bayesian inference. Fig. 1 illustrates Bayesian inference applied to a hypothetical seismic location problem (see Myers *et al.* 2007; Arora *et al.* 2013, for some detailed applications of Bayesian inference to locating seismic sources). Bayesian inference begins by expressing prior beliefs about parameters of interest θ. For example, within the context of identifying characteristics of a seismic event, these beliefs may represent prior knowledge about the distribution of earthquake magnitudes or their locations, for example, source proximity to lithospheric faults. As an observer gathers data $\mathcal{D}$ and other information, they update the prior distribution using the rules of probability. This updated, or posterior, distribution $p(\theta \mid \mathcal{D})$ now quantifies the likelihood of the source location given the data. The observer performs this update using a likelihood function to describe the probability of the data given an event hypothesis, that is, $p(\mathcal{D} \mid \theta)$. An observer constructs such a likelihood function from a probabilistic forward model of the data observed given a set of source parameters. This means that the likelihood can equally be used to construct a generative model of the data given the source parameters. The likelihood function assumes specified source parameters that describe the seismic source and then uses physical models, sensor models and models of background signals and noise to map this source description to plausible sensor data. As an example, if the arrival time of a seismic phase at a seismometer constitutes observed data, and event parameters describe the location and origin time of an earthquake, then the likelihood uses an earth structure model to predict uncertainty in the arrival time of a seismic phase from the source to any receivers. The model of the traveltime could include (predicted) earth model uncertainty and measurement uncertainty on the sensor.

Once an observer constructs a likelihood function, they can easily construct the posterior distribution on events given data. This construction is an application of Bayes' Theorem:

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)\, p(\theta)}{p(\mathcal{D})}. \tag{1}$$

We emphasize that the probability terms in eq. (1) can be either probabilities when θ is a discrete random variable or a probability density when θ is continuous. Eq. (1) provides the foundational statement of belief about uncertainties in the model and the machinery to update these beliefs as new information becomes available. In practice, solving for the updated Bayesian posterior requires approximate computational methods since the posterior may not have an analytical expression. Common approaches generate samples representing draws from the posterior distribution and can estimate QoIs. Examples of these methods include importance sampling using Monte Carlo, Quasi Monte Carlo, meshing and Markov Chain Monte Carlo (Brooks *et al.* 2011; Owen 2013).

2.2 Bayesian OED

To quantify network performance, we require a measure of how much belief changes due to inference on observed data. This is a measure of the sensor data's utility that defines the objective for experimental design. One measure that is commonly used in information theory is the Kullback–Leibler divergence:

$$\mathrm{KL}\left[p(\theta \mid \mathcal{D}) \,||\, p(\theta)\right] = \int p(\theta \mid \mathcal{D}) \log \frac{p(\theta \mid \mathcal{D})}{p(\theta)} \mathrm{d}\theta. \tag{2}$$

The KL divergence in eq. (2) measures how many units of information (bits for $\log_2$ or nats for ln) are needed to specify a change in the distribution from $p(\theta)$ to $p(\theta \mid \mathcal{D})$. These units are related to the efficiency of encoding a probability distribution (see MacKay 2003, for discussion). A KL divergence of 0 means that the distributions are the same up to sets of measure 0. The KL divergence is always non-negative and as it increases from zero, eq. (2) implies that the distributions increasingly differ. A relatively large KL divergence therefore indicates that the data were very informative, and the prior and posterior are measurably distinct.

The Bayesian OED problem is built upon the concepts of Bayesian probability and information theory (Lindley 1956; Ginebra *et al.* 2007; Huan & Marzouk 2013). Bayesian OED assumes that the observer applies Bayes' theorem (i.e. that they are a Bayesian agent) to select a sensor configuration $\mathcal{S}$ that maximizes utility; we term $\mathcal{S}$ as the 'experiment.' Because Bayesian inference is the optimal way to assimilate information it provides, Bayesian OED defines the best case scenario for extracting information from the sensor network. The Bayesian agent optimizes a utility function that depends on the posterior. In this research, the Bayesian agent maximizes the expected information gain (EIG) from the prior to the posterior, in the view of the posterior. Notationally, the expectation $E_{\mathcal{D}|\mathcal{S}}$ indicates that the observer computes the expectation with respect to the prior distribution of hypothetical data from the experiment given by $p(\mathcal{D} \mid \mathcal{S})$. The expected information gain for a specific experimental configuration is (from eq. 2):

$$\begin{aligned}\mathcal{I}(\mathcal{S}) &= \mathrm{E}_{\mathcal{D}|\mathcal{S}}\left[\mathrm{KL}\left[p(\theta \mid \mathcal{D}, \mathcal{S}) \,||\, p(\theta)\right]\right] \\ &= \int p(\mathcal{D} \mid \mathcal{S}) \int p(\theta \mid \mathcal{D}, \mathcal{S}) \log \frac{p(\theta \mid \mathcal{D}, \mathcal{S})}{p(\theta)} \mathrm{d}\theta \mathrm{d}\mathcal{D}. \end{aligned} \tag{3}$$

The outer integral in eq. (3) is the expectation over the hypothetical data from the experiment, while the inner integral computes the KL divergence given a realization of the hypothetical data. To compute the EIG in practice, we express $p(\mathcal{D} \mid \mathcal{S})$ as the marginal distribution

$$p(\mathcal{D} \mid \mathcal{S}) = \int p\left(\mathcal{D} \mid \theta', \mathcal{S}\right) p\left(\theta'\right) \mathrm{d}\theta'$$

because the evidence is often only implicitly known by way of integrating the likelihood and prior over parameters θ'. Note here that we have assumed that $p(\theta)$ is a proper density. We then draw samples from the marginal distribution by first sampling the prior, and then sampling the data according to the likelihood. These samples allow us to compute the outer expectation.

We now maximize $\mathcal{I}(\mathcal{S})$ to estimate the best experimental design $\mathcal{S}^*$ from $\mathcal{S} \in \mathbb{S}$, where $\mathbb{S}$ is the set of possible designs under consideration:

$$\mathcal{S}^* = \underset{\mathcal{S} \in \mathbb{S}}{\operatorname{argmax}}\, \mathcal{I}(\mathcal{S}) \tag{4}$$

This optimization is generalizable to include constraints that include, for example, a sensor budget or constraints on sensor locations through methods like Lagrange multipliers or nonlinear programming. Further, while we have formulated this problem through maximizing the EIG for the posterior, we could more specifically optimize EIG about a specific quantity of interest derived from its parameters.

Solving this optimization problem is challenging because it requires solving many Bayesian inference problems for many hypothetical realizations of data from many hypothetical sensor configurations. This nested complexity means that significant care must be taken to make this approach computationally tractable.

2.3 Bayesian optimization

Greedy optimization algorithms provide an effective computational solution to sequentially place sensors in OED and for other network optimization problems (Krause *et al.* 2008; Carmichael 2020). Such greedy optimization involves sequentially adding sensors one at a time so that the optimization problem at a particular iteration is low dimensional, and therefore only requires updating the location of that particular sensor. During an iteration, the algorithm computes the EIG as an average over all possible source locations specified by the prior, and then computes an optimal location. Fig. 2 illustrates the process of adding sensors one-by-one. The optimization surfaces, shown in the top row of the figure, start out fairly symmetric with multiple optima when there are few sensors. However, as more sensors are added these symmetries are broken so there is a unique optimal location of the next sensor. The bottom row illustrates how the EIG increases with sensor density, particularly about sources that are near several sensors.

We concede that the true, optimal sensor network configuration requires that we compute sensor location solutions all at once. However, greedy optimization often does reasonably well with a significantly reduced computational cost. In fact, suboptimal bounds exist for certain classes of optimization problems approximated using greedy methods, such as the class of submodular functions. At a high level, submodular utility functions exhibit diminishing returns with each iteration, that is, adding a sensor to a smaller network yields higher gains than adding a sensor to a larger network. The EIG objective for Bayesian OED is submodular when the sensors are conditionally independent given the event description, that is, when

$$p(\mathcal{D}_i, \mathcal{D}_j \mid \theta) = p(\mathcal{D}_i \mid \theta) p(\mathcal{D}_j \mid \theta),$$

where $\mathcal{D}_i$ and $\mathcal{D}_j$ are data generated by sensors i and j, respectively. If the utility function is submodular, then we can show that the greedy optimum will be near-optimal, meaning that if the utility function is submodular, it can be shown that the greedy algorithm provides a near-optimal solution. Specifically, it offers a multiplicative approximation guarantee of $(1 - 1/e)$, where e denotes Euler's number, meaning

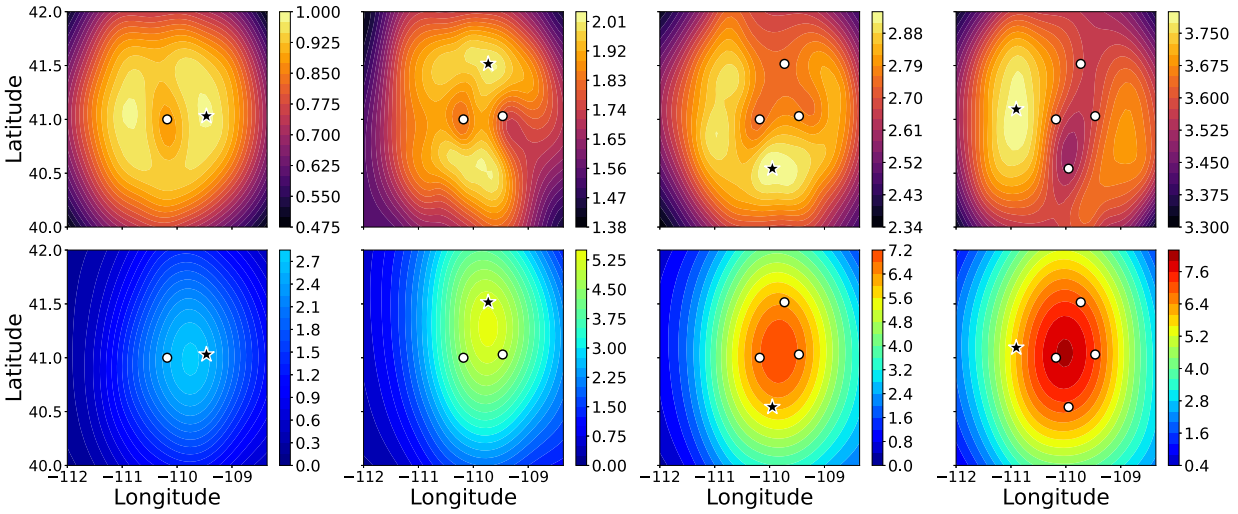

Figure 2. Illustration of greedy optimization of a sensor network with five sensors according to eq. (4). Moving from left to right, the first sensor was placed at the centre of the domain and then the subsequent four sensors are placed sequentially to maximize expected information gain (EIG). The white dots define the initial sensor configuration while the black star defines the new sensor that is being added to optimally augment the network. The top row illustrates the optimization surface, where the colour contours show how much the EIG, averaged over all event locations, would increase if a sensor were added at that location given the initial network. The bottom row shows the EIG for the augmented sensor network where the colour contours show the EIG about the location of a shallow, low-magnitude seismic source at that specific latitude and longitude, that is, for $n = 2, \ldots, 5$ it displays $\mathcal{I}(S_n \mid \theta' = [\mathcal{L}, x, m])$ for all $\mathcal{L}$ in the domain (see eq. 21).

that the greedy solution is guaranteed to be above $1 - 1/e \approx 63$ per cent of the value of the global optimum. This bound is loose in practice, and stronger assumptions on the problem structure may yield smaller departures from global optimality. We refer to Krause *et al.* (2008) for details on submodular functions and greedy optimization.

We use Bayesian optimization (Močkus 1975) to greedily optimize sensor placement locations. Instead of finding the optimal configuration of all sensors at once, we iteratively choose sensors one at a time. Given the sensors that have already been placed, we choose the location of the next sensor by using Bayesian optimization to solve the 2-D optimization problem for a single sensor placement.

Doing this requires sampling the utility function to build a surrogate model of the optimization surface from the samples, such as a Gaussian Process model (Williams & Rasmussen 2006). Using this surrogate model, we choose new points at which to evaluate the utility function according to an acquisition function. The choice of acquisition function determines how we balance the exploration of high uncertainty regions of the parameter space, improving our surrogate model, with optimizing the existing surrogate to sample new points that will be close to the predicted optimum. The residual between the optimal solution and the sampled solution improves with iteration. Details of Bayesian optimization and descriptions of acquisition functions can be found in Jones *et al.* (1998), Srinivas *et al.* (2010), Picheny *et al.* (2013) and Frazier (2018). We use the Python library SCIKIT-OPT (Head *et al.* 2020) to implement Bayesian optimization with a GP surrogate.

3 BAYESIAN SEISMIC MONITORING

3.1 General approach

As introduced in Fig. 1, Bayesian inference for seismic monitoring requires constructing a likelihood model $p(\mathcal{D} \mid \theta, \mathcal{S})$ that quantifies the likelihood of the data given an event θ with a seismic sensor network configuration $\mathcal{S}$. We assume that a source can be sufficiently defined by a vector of its origin time, location and magnitude $\theta = \{\text{Time, Lat, Long, Depth, Mag}\}$. Notationally, these parameters are epicentral location $\mathcal{L} = \{\text{Lat, Long}\}$, source depth x, event magnitude m and origin time t_o. The network $\mathcal{S}$ consists of individual stations $\mathcal{S}_i$. Such stations may have heterogenous response or sampling features but here we assume they are homogenous. Therefore, station description is sufficiently described by $\mathcal{S}_i = \{\mathcal{S}_i^{\text{Loc}}\}$ where $\mathcal{S}^{\text{Loc}}$ is the station's location in latitude and longitude.

We limit our analysis to modelling arrivals of seismic phases from their sources and leave inclusion of waveform features to future research. Therefore, our data take the form of $\mathcal{D} = \{\mathbb{D}, \mathbb{A}\}$, where $\mathbb{D}$ stores data about which stations detected different seismic phases and $\mathbb{A}$ stores information about the arrival times, t_{ij}, of the detected phases.

$$\mathbb{D}_{ij} = \begin{cases} 1 \text{ if station } i \text{ detects phase } j \\ 0 \text{ if station } i \text{ does not detect phase } j \end{cases} \tag{5}$$

$$\mathbb{A}_{ij} = \begin{cases} t_{ij} \text{ if station } i \text{ detects phase } j \\ \emptyset \text{ if station } i \text{ does not detect phase } j \end{cases} \tag{6}$$

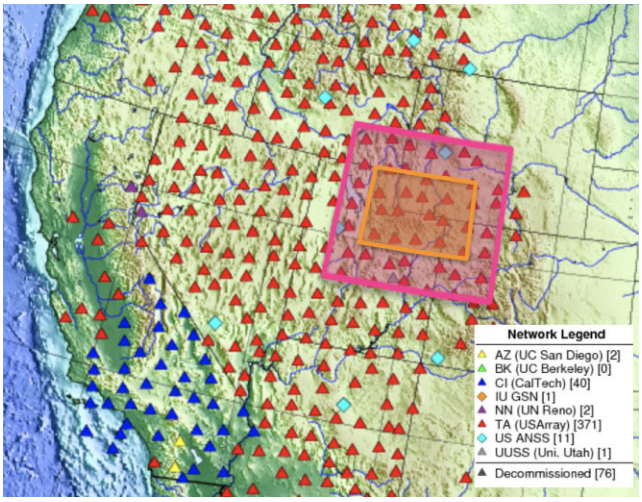

Figure 3. Map of transportable array stations from December 2007 [Map from EarthScope ANF Website (2021)]. The outer highlighted region indicates the region (Latitude $\in [39^o N, 43^o N]$, Longitude $\in [113^o W, 107.36^o W]$) that we use gathered sensor data that populates the parameters of the likelihood models. The inner highlighted region (Latitude $\in [40^o N, 42^o N]$, Longitude $\in [112^o W, 109.36^o W]$), corresponds to the monitoring region over which we build a sensor network (IRIS Transportable Array 2003).

Note that $\mathbb{A}_{ij} = \emptyset$ when no phase is detected since there is no arrival time to capture. We make the simplifying assumption that the likelihood of detection is independent of the origin time t_o. This assumption seems reasonable, but may not hold in cases that background noise is diurnally variable. Incorporating a time-dependent background would not be difficult but is left to future work. We also assume that the priors are independent, but this likewise is easy to relax as needed. The posterior then is:

$$p(\mathcal{L}, x, m, t_o \mid \mathbb{A}, \mathbb{D}, \mathcal{S}) = \frac{p(\mathbb{A}, \mathbb{D} \mid \mathcal{L}, x, m, t_o, \mathcal{S})\, p(\mathcal{L}, x, m, t_o)}{p(\mathbb{A}, \mathbb{D} \mid \mathcal{S})}$$
$$\propto p(\mathbb{A} \mid \mathcal{L}, x, m, t_o, \mathbb{D}, \mathcal{S})\, p(\mathbb{D} \mid \mathcal{L}, x, m, \mathcal{S})\, p(\mathcal{L})\, p(x)\, p(m)\, p(t_o) \quad (7)$$

To construct the likelihood $p(\mathbb{D} \mid \mathcal{L}, x, m, \mathcal{S})$ we must estimate the detection probability for a given arrival. When historic data are available, we can build a model for the detection of a phase at a station given an event at a specific location and with a specified magnitude as we discuss in Section 3.2.

We consider two separate sources of uncertainty in the arrival time likelihood $p(\mathbb{A} \mid \mathcal{L}, x, m, t_o, \mathbb{D}, \mathcal{S})$: measurement noise and model prediction uncertainty. We assume that the measurement noise distributions for each station are independent (which may not be true in situations where sensors are close, but provides a tractable simplifying assumption that is valid for sparse networks). For situations where measurement noise statistics are known *a priori* for a sensor and processing method, they be directly integrated into the likelihood model. Otherwise, the measurement noise model can be derived from data along with other assumptions that we will describe in Section 3.3.2. However, when deriving measurement error models directly from data, the effect of modelling uncertainty must also be simultaneously accounted for. Unlike for measurement uncertainty, including correlated traveltime errors across different stations in the likelihood function for modelling error is important. This correlation reflects that real Earth structure will likely be different than the modelled Earth structure and this discrepancy will induce correlated errors. We therefore model this uncertainty, and the correlation induced in the sensor network, by sampling a distribution of Earth models to estimate the distribution in arrival times as discussed in Section 3.3.1.

3.2 Detection model

We use a catalogue from the USArray Transportable Array experiment (IRIS Transportable Array 2003) to build a detection model for seismic phases, specifically the first P arrival, that is, arrivals labelled as P, Pg and Pn in the catalogue. Details of the modelling region can be seen in Fig. 3. This model is similar to the logistic regression model used by NET-VISA (Arora *et al.* 2013). In principle this method can be followed for any monitoring region with existing sensors. The USArray data set was chosen because of the homogeneity of the sensor network and its uniform coverage for a region. Sensors were deployed in this region from August 2007–August 2008 and during that time, 1089 events were registered on 45 stations. For these events, 11 487 P arrivals were detected out of the 49 005 potential detections, which corresponds to 23 per cent of potential P detections. Note we assume that every station had the potential to detect each event so the number of potential detections is just the number of events multiplied by the number of stations. Of the 1089 events, 833 had estimated magnitudes. The minimum magnitude of the data set was 0.51, maximum was 4.37 and median was 2.03.

Fig. A1 shows the mean detection probability of seismic sources in our catalogue, binned over magnitude and distance. We construct a logistic regression model using input features that include the distance between the event and the sensor (in degrees), the depth of the event and the magnitude of the event. Our catalogue data also included events with missing magnitude estimates. We therefore used an additional

Table 1. Table displaying the effect of SNR offset on measurement error for both the uniform and non-uniform prior.

SNR offset	σ_{meas}, Uniform prior	σ_{meas}, Non-uniform prior
3.5	0.1	0.1
2.91	0.1	0.1
2.32	0.13	0.15
1.73	0.23	0.38
1.14	0.41	0.75
0.55	0.62	1.19
−0.05	0.81	1.5
−0.64	0.92	1.83
−1.23	0.97	1.94
−1.82	0.99	1.98
−2.41	1.0	1.99
−3.0	1.0	2.0

Note: The first column shows the offset value, and the second two columns show the average measurement error across all events sampled from the given prior for sensors with the given offset. It appears that between values of 1.73 and −0.64, the average measurement error is more sensitive to changes in SNR. We emphasize that this is not an equivalency table, but rather a notional description of how the measurement error changes as the SNR is changed.

indicator feature to reflect the absence of magnitude information in our source vector θ. This feature is 1 when the magnitude data are absent and 0 when magnitude data are present. This feature helps us train using the data with missing magnitudes, which is critical for low magnitude events. When this model is used as part of the Bayesian OED framework, the magnitudes of hypothetical events will all be known so this indicator feature is always ignored after training. As described previously, we assume that the detection probability for each station is conditionally independent, so the likelihood model becomes:

$$p\left(\mathbb{D} \mid \mathcal{L}, x, m, \mathcal{S}\right) = \prod_i p\left(\mathbb{D}_i \mid \mathcal{L}, x, m, \mathcal{S}_i\right), \tag{8}$$

where

$$p\left(\mathbb{D}_i \mid \mathcal{L}, x, m, \mathcal{S}_i\right) = \begin{cases} \frac{\exp(\alpha \text{Dist}[\mathcal{L},\mathcal{S}_i]+\beta x+\gamma m+\delta)}{1+\exp(\alpha \text{Dist}[\mathcal{L},\mathcal{S}_i]+\beta x+\gamma m+\delta)}, & \text{if station } i \text{ detects the phase} \\ \frac{1}{1+\exp(\alpha \text{Dist}[\mathcal{L},\mathcal{S}_i]+\beta x+\gamma m+\delta)}, & \text{if station } i \text{ does not detect the phase.} \end{cases} \tag{9}$$

The coefficients $\alpha, \beta, \gamma, \delta$ in eq. (9) correspond to the regression coefficients that fit the data. $\text{Dist}\left[\mathcal{L}, \mathcal{S}_i\right]$ is the distance in degrees from $\mathcal{L}$ to $\mathcal{S}_i$, x is the depth and m is the magnitude. Since we only consider one phase, we remove the phase index in $\mathbb{D}$ hereon. With this choice, we find the distance coefficient, $\alpha = -2.82$, the depth coefficient, $\beta = -0.03$, the magnitude coefficient, $\gamma = 1.14$ and the intercept, $\delta = 1.95$. From this we see that the distance and magnitude have a much higher influence than depth on the detection probability of the first P arrival. While the distance coefficient appears larger than the depth coefficient, this primarily reflects that distance is measured in degrees while depth is in kilometres; when converted to the same units, their effects are more comparable, though depth typically has less impact on detection probability due to its smaller range of values.

3.3 Arrival time model

3.3.1 Earth model uncertainty

For our traveltime uncertainty model, first we will build an uncertainty model that captures the uncertainty due to the earth model. We then will include a conditionally independent and additive measurement uncertainty model. We take the approach of using model uncertainty over using replicate variability because we are using synthetic earth models that produce the same output (up to measurement error) for the same inputs, although these earth models will have unknown errors when compared to potentially observable 'ground truth' traveltimes. We treat this latent discrepancy between these models and the true experiment as aleatoric uncertainty since, in practice, experiments treat each event independently. It is possible to learn this discrepancy by jointly inferring events (Myers *et al.* 2007), however the added complexity is beyond the scope of this optimal experimental design study. For more details on this type of understanding, see Kennedy & O'Hagan (2001) or Maupin & Swiler (2020).

To capture earth model uncertainty, we selected 121 vertical cross-sections from Crust 1.0 (Laske *et al.* 2013) from the area around the monitoring region to get 121 different 1-D earth models with different Vp velocity profiles. These models can be seen in Fig. 4. For each of these models we used TauP (Crotwell *et al.* 1999) to compute the traveltimes for different distances, Δ, and depth, x, pairs. For a given distance and depth pair we compute the mean and variance of the traveltimes, t_i, given the traveltimes computed by TauP for the $N = 121$ models:

$$\mu\left(\Delta, x\right) = \tfrac{1}{N} \textstyle\sum_{i=1}^{N} t_i\left(\Delta, x\right) \tag{10}$$

$$\sigma\left(\Delta, x\right) = \sqrt{\tfrac{1}{N-1} \textstyle\sum_{i=1}^{N} [t_i\left(\Delta, x\right) - \mu\left(\Delta, x\right)]^2} \tag{11}$$

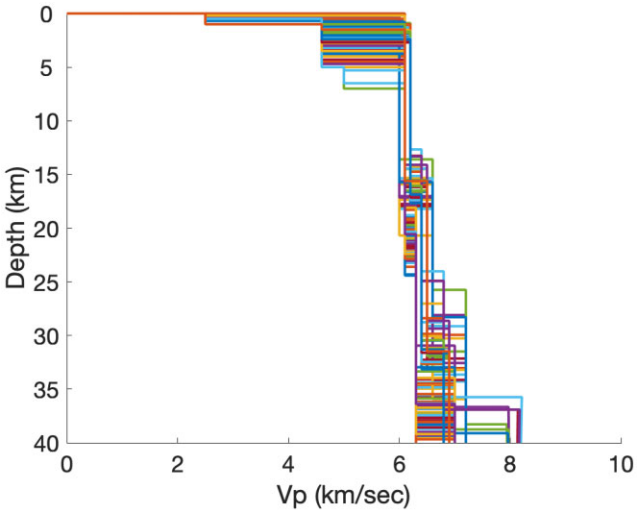

Figure 4. Illustration of the 121 1-D velocity models for Vp sampled from around the monitoring region in Fig. 3. These representative earth models are used to estimate traveltime uncertainty from earth model uncertainty.

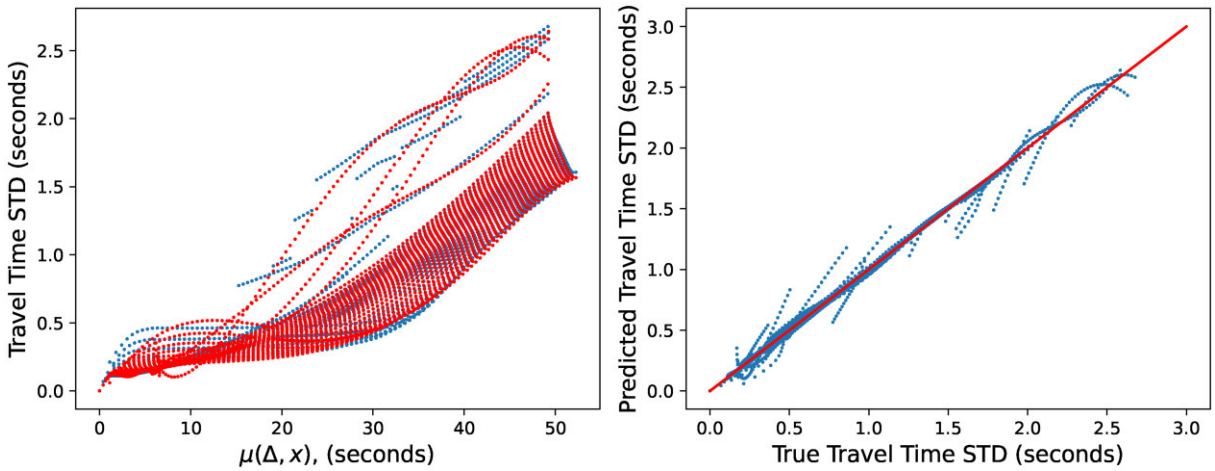

Figure 5. *Left*: A scatter plot (blue) of the estimated traveltime mean $\mu(\Delta, x)$, and estimated traveltime standard deviation, $\sigma(\Delta, x)$ for various distance and depth pairs, superimposed with a fit polynomial model (red). *Right*: A scatter plot of predicted traveltime standard deviations compared to the true value (red). A line demonstrating the performance of a perfect model is displayed in blue. The polynomial model adequately captures the bulk trend, despite some variability due to the nature of the first arriving phase.

Given the estimated mean and standard deviation pairs (Fig. 5), we derive a model for the standard deviation of the traveltime σ_{model} as a fifth degree polynomial function of depth and distance that will be used in the likelihood. This high-order polynomial sufficiently captures the major dependencies in the traveltime standard deviation over the domain of interest but would fail to extrapolate beyond that domain. Therefore, care should be taken whenever using these types of function approximations that they are trained on the domain of interest as they are not intended for extrapolation.

3.3.2 Measurement error

As described earlier, we also develop a model of measurement uncertainty for each station that we treat as conditionally independent of the other station. This takes the form of phase pick uncertainty (Velasco *et al.* 2001), σ_{meas}, that depends on a sensor's signal-to-noise ratio (SNR). We model the SNR of a sensor's detection using a linear model that is a function of log distance, $\log \Delta$, and magnitude m, given by

$$\text{SNR} = a \cdot m - b \cdot \log \Delta + c + \varepsilon, \tag{12}$$

where ε is an offset hyperparameter called the SNR offset used to account for different sensor fidelities. We fit the coefficients a, b and c again using the transportable array data set (IRIS Transportable Array 2003). We note that for our fit we found that no depth term was required which is why it was omitted but this will obviously depend on the problem context. We also add an offset term to this equation, potentially unique to each sensor, which allows us to tune sensor fidelity as we perform various experiments. As in Velasco *et al.* (2001), the σ_{meas} is thus

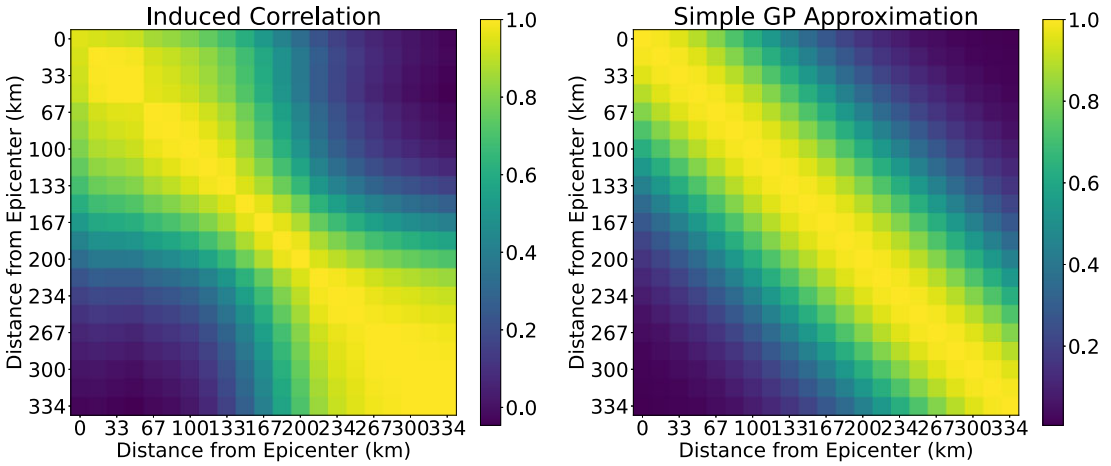

Figure 6. Visualization of the correlation matrix. The left figure is the correlation matrix, Γ, induced by the earth model uncertainty computed using eq. (14). The axes correspond to the distance along the surface from the source epicentre in km. Stations are approximately spaced 33 km apart. The right figure is a simplified model of the correlation, Γ_{GP}, found using a square-exponential kernel, which approximates Γ. The square-exponential kernel assumes that correlation is only a function of the distance between stations. This simplified model reasonably captures the correlation length scale for Γ but is unable to capture the complex, non-translation invariant block structure of Γ.

given by

$$\sigma_{\text{meas}}(\text{SNR}) = \begin{cases} \sigma_0 & \text{SNR} < t_L \\ \gamma\sigma_0 & \text{SNR} > t_U \\ \sigma_0 - \frac{\sigma_0 - \gamma\sigma_0}{\log(t_U) - \log(t_L)} \log(\frac{\text{SNR}}{t_L}) & \text{otherwise} \end{cases}, \tag{13}$$

where γ, t_U and t_L, and σ_0 may all be tuned as hyperparameters (with γ constrained to be less than 1).

Ultimately, combining the modelling and measurement error we get that the total arrival error is

$$\sigma_p^2(\Delta, x, m) = \sigma_{\text{model}}^2(\Delta, x) + \sigma_{\text{meas}}^2(\Delta, m).$$

.

3.4 Traveltime correlation

While the previous models described the magnitude of errors at a station, they have not captured any correlations between the stations. In principle, we expect that there could be signification correlation in traveltime uncertainty, particularly due to the earth model. We can compute the correlation between the traveltimes observed at two different stations at locations Δ_j and Δ_k for an event at depth x. This correlation is induced by the earth model uncertainty as:

$$\rho\left(\Delta_j, \Delta_k, x\right) = \frac{\sum_{i=1}^{N}[t_i\left(\Delta_j, x\right) - \mu\left(\Delta_j, x\right)][t_i\left(\Delta_k, x\right) - \mu\left(\Delta_k, x\right)]}{(N-1)\,\sigma\left(\Delta_j, x\right)\sigma\left(\Delta_k, x\right)}. \tag{14}$$

Here μ and σ are computed from the N earth models in different locations from Crust 1.0 as in eqs (10) and (11). For simplicity we will remove the depth dependence of the correlation by averaging the correlation over all L depths. Therefore we estimate the correlation between two sensors as $\rho\left(\Delta_j, \Delta_k\right) = \frac{1}{L}\sum_{l=1}^{L}\rho\left(\Delta_j, \Delta_k, x_l\right)$.

We define the full correlation matrix, Γ, between the stations at distances Δ_i from the source has having elements $\Gamma_{jk} = \rho\left(\Delta_j, \Delta_k\right)$. We want to fit a Gaussian process model with a square exponential kernel to this data so we can easily estimate the correlation between arbitrary sensor pairs when designing the network, that is, we want $\Gamma \approx \Gamma_{\text{GP}}$. Therefore we want to find the correlation length, ℓ, such that $\{\Gamma_{\text{GP}}\}_{jk} = \exp\left[-\frac{1}{2\ell^2}\left(\Delta_j - \Delta_k\right)^2\right]$ and Γ_{GP} minimizes the discrepancy with Γ. We, under our modelling conditions, find the correlation length scale as $\ell = 147.5$ km. The comparison of Γ and the resulting Γ_{GP} can be seen in Fig. 6. We observe that the square exponential kernel is able to capture the general length scale of the induced correlation, meaning that stations that are close together are more correlated, but does not capture its complexity. The induced correlation has a block-like structure where stations that are near to the source are highly correlated, stations far from the source are highly correlated, and stations in the transition region exhibit less strong correlation with nearby stations. This likely corresponds to the type of first arrival that is being observed at each station, where close stations observe a Pg while far stations observe a Pn. Our choice of GP kernel is translation invariant meaning that the sensor correlation is only a function of the distance between the two sensors and does not depend on the source parameters. More generally, a different GP kernel would need to be constructed for each seismic source, which is computationally challenging for the nested complexity of OED. Considering only a translation invariant GP kernel is obviously a simplification but provides a first step towards modelling station correlation which is typically very difficult and often ignored.

We can now construct the P arrival time likelihood $p(\mathbb{A} \mid \mathcal{L}, x, t_o, \mathbb{D}, \mathcal{S})$ by combining our model of the traveltime prediction, μ_p; earth-model-induced standard deviation, σ_{model}; measurement-induced standard deviation σ_{meas}; and correlation matrix, Γ_{GP}:

$$p(\mathbb{A} \mid \mathcal{L}, x, t_o, \mathbb{D}, \mathcal{S}) = \frac{1}{(2\pi)^{|\mathbb{D}|/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\left[\mathbb{A} - \mu_p(\mathcal{L}, x, \mathbb{D}, \mathcal{S}) - t_o\right]^T \Sigma^{-1}\left[\mathbb{A} - \mu_p(\mathcal{L}, x, \mathbb{D}, \mathcal{S}) - t_o\right]\right) \tag{15}$$

$$\Sigma = \sigma_{\text{model}}^T(\Delta, x, \mathbb{D})\,\Gamma_{\text{GP}}(\mathbb{D}, \mathcal{S})\sigma_{\text{model}}(\Delta, x, \mathbb{D}) + \text{diag}\left[\sigma_{\text{meas}}^2(\Delta, m, \mathbb{D}, \mathcal{S})\right]. \tag{16}$$

Here $|\mathbb{D}|$ is the number of detections, $|\Sigma|$ is the determinant, $\mu_p(\mathcal{L}, x, \mathbb{D}, \mathcal{S})$ is a vector of the predicted traveltimes for stations that had a detection, $\sigma_{\text{model}}(\Delta, x, \mathbb{D})$ is a vector of the predicted standard deviations of the traveltime to each station induced by the earth model uncertainty and $\text{diag}\left[\sigma_{\text{meas}}^2(\Delta, m, \mathbb{D}, \mathcal{S})\right]$ is a diagonal matrix of the squared predicted standard deviations of the traveltime to each station induced by the measurement uncertainty. Finally, $\Gamma_{\text{GP}}(\mathbb{D}, \mathcal{S})$ is the estimated correlation between stations using the GP model.

We further note that we marginalized our source origin time prior over t_o, assuming a uniform improper prior (meaning that an event is equally likely at any time), and therefore omit it from the model. We assume this prior since the origin time is naturally restricted by the size of the chosen domain. The improper uniform prior only behaves differently from a proper prior on the edges of the proper prior's domain, so any uniform prior that is wide enough to ensure that possible traveltimes given the domain do not occur on the edges of the domain should be functionally equivalent to an improper prior. See Fig. A4 for further details. This reduces the dimension of our seismic source parametrization space and leads to our final model of the arrival time likelihood:

$$\begin{aligned} p(\mathbb{A} \mid \mathcal{L}, x, \mathbb{D}, \mathcal{S}) &= \int_{t_o} p(\mathbb{A} \mid \mathcal{L}, x, t_o, \mathbb{D}, \mathcal{S})\, p(t_o)\, \mathrm{d}t_o \\ &= \frac{1}{(2\pi)^{(|\mathbb{D}|-1)/2}|\Sigma|^{1/2}\beta^{1/2}} \exp\left(-\frac{1}{2}\left[\mathbb{A} - \mu_p(\mathcal{L}, x, \mathbb{D}, \mathcal{S})\right]^T \Sigma^{-1}\left[\mathbb{A} - \mu_p(\mathcal{L}, x, \mathbb{D}, \mathcal{S})\right]\right) \exp\left(\frac{\alpha^2}{2\beta}\right) \end{aligned} \tag{17}$$

$$\alpha = \mathbb{1}^T \Sigma^{-1}\left[\mathbb{A} - \mu_p(\mathcal{L}, x, \mathbb{D}, \mathcal{S})\right] \tag{18}$$

$$\beta = \mathbb{1}^T \Sigma^{-1} \mathbb{1} \tag{19}$$

4 COMPUTATIONAL APPROACH

4.1 Estimating information gain

Given the Bayesian framework introduced in Section 2 and the specific models introduced in Section 3 we present a method to estimate the expected information gain, $\mathcal{I}(\mathcal{S})$, of the sensor network $\mathcal{S}$. Recall from eq. (3) that we can express EIG as:

$$\mathcal{I}(\mathcal{S}) = \int p(\theta') \int p(\mathcal{D} \mid \theta', \mathcal{S}) \int p(\theta \mid \mathcal{D}, \mathcal{S}) \log \frac{p(\theta \mid \mathcal{D}, \mathcal{S})}{p(\theta)} \mathrm{d}\theta \mathrm{d}\mathcal{D} \mathrm{d}\theta'. \tag{20}$$

We can further define $\mathcal{I}(\mathcal{S} \mid \theta')$ as the expected information gained about a specific event θ' where

$$\mathcal{I}(\mathcal{S} \mid \theta') = \int p(\mathcal{D} \mid \theta', \mathcal{S}) \int p(\theta \mid \mathcal{D}, \mathcal{S}) \log \frac{p(\theta \mid \mathcal{D}, \mathcal{S})}{p(\theta)} \mathrm{d}\theta \mathrm{d}\mathcal{D} \tag{21}$$

and thus express $\mathcal{I}(\mathcal{S})$ as:

$$\mathcal{I}(\mathcal{S}) = \int \mathcal{I}(\mathcal{S} \mid \theta')\, p(\theta')\, \mathrm{d}\theta' \tag{22}$$

$\mathcal{I}(\mathcal{S} \mid \theta')$ is an important quantity on its own as it can be used to tell how sensitive the network is to a specific event θ'. We can then produce maps of this sensitivity in order to communicate how the network performs under different conditions in order to check against requirements.

We will use the approach of estimating $\mathcal{I}(\mathcal{S} \mid \theta')$ to estimate EIG. First, we draw samples from θ' from $p(\theta')$ to construct a large set of candidate seismic events using a method like importance sampling with a Quasi-Monte Carlo (QMC) mesh (Owen 2013). QMC provides an efficient set of space-filling samples that requires significantly fewer samples than a standard uniform grid. Then, for each element in our event space, we estimate $\mathcal{I}(\mathcal{S} \mid \theta')$ and average them to estimate $\mathcal{I}(\mathcal{S})$. To estimate $\mathcal{I}(\mathcal{S} \mid \theta')$ we construct hypothetical data sets by sampling $p(\mathcal{D} \mid \theta', \mathcal{S})$. Then we will solve the Bayesian inference problem given the data sets to estimate the information gain measured via the KL divergence. We solve the Bayesian inference problem over the discrete event space instead of a continuous event space for computational efficiency, although this results in a bias. Solving the Bayesian inference problem involves sampling from the prior distribution, which we accomplish using importance sampling (discussed further in Section A3), sampling from an importance distribution $q(\theta)$ instead of the prior. This yields an estimator for the KL divergence given by

$$\text{KL}\left[p(\theta \mid D) \,||\, p(\theta)\right] \approx \sum_{i=1}^{N} \frac{w_i}{\sum_{i=1}^{N} w_i} \left(\log(p(\theta \mid D)) - \log\left(\frac{1}{N}\sum_{i=1}^{N} w_i\right)\right), \tag{23}$$

where

$$w_i = \frac{p(\theta_i)}{q(\theta_i)} p(D \mid \theta_i). \tag{24}$$

For more details on this estimator, see Section A4.

As long as enough discrete points are used, the KL divergence will converge to the same value as the continuous distribution so the bias will be small. By looking at statistics of the posterior probabilities of the discrete events, we can assess whether enough points have been used. The hypothetical data are constructed by sampling $p(\mathbb{A} \mid \mathcal{L}, x, \mathbb{D}, \mathcal{S})$ and $p(\mathbb{D} \mid \mathcal{L}, x, m, \mathcal{S})$. The resulting algorithm is summarized in Algorithm 1.

Algorithm 1 Expected Information Gain (EIG) Calculation

Input: $\mathcal{S}$ (sensor configuration), Θ (plausible events), $p(\theta')$
Result: $I(\mathcal{S}|\theta)$ for individual events θ, and total $I(\mathcal{S})$
for each event hypothesis $\theta' \in \Theta$ **do**
    simulate arrival dataset according to $D \sim p(D|L', x', m', \mathcal{S})$
    for each arrival dataset D **do**
        simulate arrival time according to $A \sim p(A|L', x', D, \mathcal{S})$
        for each simulated dataset $D = \{A, D\}$ **do**
            discretize the parameter space using N samples $\theta \sim q(\theta)$
            compute likelihoods $p(D|\theta, \mathcal{S})$ for each θ using Equations 9 and 17
            compute importance weights for each θ according to Equation 24
            compute KL divergence for information gain $I(\mathcal{S}|\theta', D)$ according to Equation 23
        end for
    end for
    compute EIG for θ', as average of $I(\mathcal{S}|\theta', D)$ across simulated data
end for
compute total EIG $I(\mathcal{S})$ as average over all event hypotheses and data

We choose to use this approach as opposed to a Markov chain Monte Carlo (MCMC) method for two reasons. First, the dimension of the sample space is small, allowing us to draw enough samples to reliably reconstruct the prior and posterior distributions. Secondly, this approach allows us to reuse likelihood computations for each sample across all steps of the algorithm, whereas an MCMC method would require computing a new likelihood at each iteration. In applications where the dimension of the sample space is higher, an MCMC method would likely be preferred. We also acknowledge that there are potential issues with this approach in cases where the importance distribution does a poor job of approximating the sampling distribution (Williams 2021). This could be particularly problematic in cases where diffuse prior samples are used for sampling a concentrated posterior. In further work we hope to explore alternative sampling methods such as MCMC, double-nested Monte Carlo and those discussed in Picard *et al.* (2019) and compare their performance to the method used in this work.

4.2 Optimization

Once we have the algorithms to estimate $\mathcal{I}(\theta' \mid \mathcal{S})$ and $\mathcal{I}(\mathcal{S})$, we can formulate the optimal experimental design problem to choose the location and type of different seismic stations. We can use the greedy Bayesian optimization method described in 2.3. We use the Python library SCIKIT-OPT (Head *et al.* 2020) to implement Bayesian optimization with a Gaussian process (GP) surrogate. This library enables us to adaptively learn hyperparameters of the GP kernel function, for example, length scales of the squared exponential kernel, magnitudes of the additive white noise, etc. Further, it can support several different criteria for Bayesian optimization that control the way in which the optimizer balances exploration versus exploitation in Bayesian optimization. This trade-off means that the Bayesian optimizer has to choose sample points that enable it to both learn the surrogate for the EIG surface and find points that optimize the EIG. The common criteria for this found in SCIKIT-OPT are the expected improvement, lower confidence bound and probability of improvement. SCIKIT-OPT also has the option to mix these criteria and choose one at random. We found that expected improvement works well but have not systematically explored all these options.

4.3 Software implementation

The models in Section 3 and algorithms from this section can be found on GitHub (Catanach *et al.* 2024). This code provides the tools necessary to analyse and optimize seismic monitoring networks. Currently we target the location problem, like those discussed in Section 5, in which we want to study how well the network will identify the location of an event and then optimize the network to provide better locations. A detailed explanation of the software implementation can be found in A5.

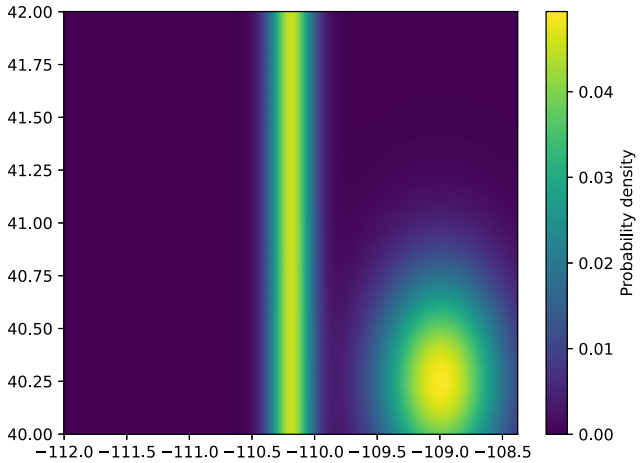

Figure 7. Simple prior distribution on latitude and longitude representing a fault and point source. It is comprised of three mixture components: a bivariate Gaussian representing the point source, a univariate Gaussian in the longitude direction multiplied by a uniform in the latitude direction representing the fault and a uniform in both directions representing the background probabilities.

5 RESULTS

Unless otherwise specified, we explore a simple model for placing sensors to monitor a square domain for latitudes between 40° N and 42° N, longitudes between 112° W and 108:36° W, magnitudes greater than 0 and depth between 0 and 40 km. For computing the EIG, 10k events were chosen using a QMC mesh defined by the Sobol sequence. For each candidate event 32 hypothetical data realizations were used. 100 steps of Bayesian optimization per sensor were used to optimize the sensor configuration.

5.1 Prior distributions

We perform our experiments with one or both of the following prior distributions on our seismic parameters.

The first prior used was a uniform prior. Under this prior, seismic sources are assumed to have a uniform prior probability in this domain. We also assume that the magnitude prior is an exponential distribution with rate parameter $\lambda = \log(10)$ and a minimum magnitude of 0.5. This prior means that the likelihood of an event of a given magnitude falls off exponentially as the magnitude increases. We assume that the origin time is a uniform improper prior meaning that all times are equally likely. The sensors are also limited to be placed in this domain.

The second distribution used a mixture distribution on latitude and longitude to very simply simulate both a fault line and a point source. It used a uniform distribution on depth and an exponential distribution with $\lambda = 10$ on magnitude.

We choose the mixture distribution on latitude and longitude to represent a fault line and a point source. The first mixture component is a bivariate Gaussian centred at $(40.25, -109)$ with covariance matrix

$$\Sigma = \begin{bmatrix} 0.125 & 0 \\ 0 & 0.125 \end{bmatrix}.$$

The second mixture component is a 1-D Gaussian in the longitude direction with mean -110.19 and standard deviation 0.125 multiplied by a uniform in the latitude direction. The final mixture component is a uniform distribution across both latitude and longitude. These components were given mixture weights 0.49, 0.49 and 0.02, respectively. See Fig. 7 for a visual representation.

The total probability for a single event under this prior is thus given by the product of the probability for each parameter. For convenience, we refer to this second prior as the fault-box prior.

5.2 Analysis results

Using our algorithm, we can perform two different types of analyses: We can design new sensor networks for a given area, and we can analyse the sensitivity of existing sensor networks to events in a given area. Fig. A5 shows what it looks like when sensors are placed sequentially in a given area. Fig. 8 shows an an analysis of network sensitivity to events in a given area. We see that the network gains more information about events that are far from the high-density areas of the prior, and less information about events that are closer to high-density areas.

5.3 Effect of sensor fidelity

We next investigate how sensor placement is affected by varying sensor fidelity conditions. We control the fidelity of a given sensor by adding an offset to its ratio of signal to measurement noise, an offset than can be thought of as corresponding to measurement noise with a given

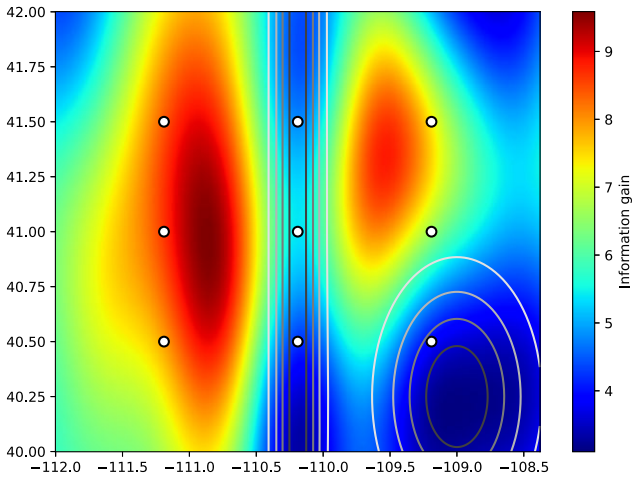

Figure 8. A network sensitivity analysis to events in a given latitude/longitude domain. The prior distribution on event location is shown by the grey contour plot. Events that are far from the high-density areas of the prior distribution contribute more information than events close to high-density areas.

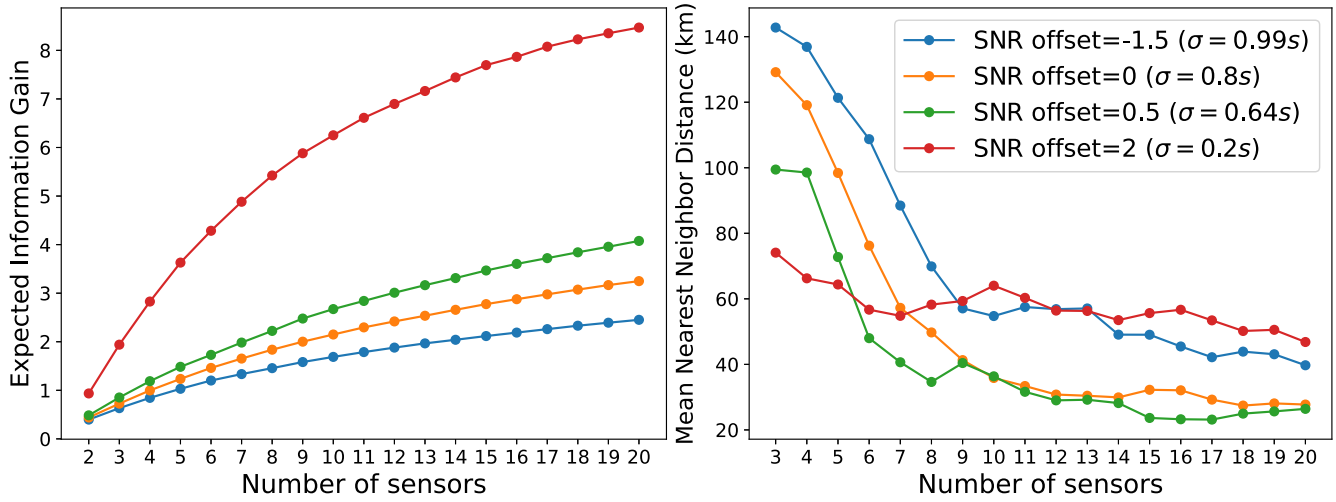

Figure 9. The left panel illustrates the change in EIG as additional sensors are placed using greedy optimization for three different networks with different sensor fidelities. These networks have signal-to-noise (SNR) ratio offsets of 2, 5, 0 and −1.5 s. The corresponding average measurement error for these networks, σ, is listed in parentheses. See Table 1 for a description of the relationship between SNR and measurement error. The right panel describes the geometry of the networks based upon how close the stations are to each other. We can see that, particularly in the beginning, the noisier the network is, the farther apart stations are added to those networks.

standard deviation. Using a uniform prior, we place 20 sensors using four different sensor fidelity values (See Fig. 9). Unsurprisingly, we see that as sensor fidelity increases, the network's information gain also increases. It is difficult to see a clear pattern in sensor proximity, but we notice that as sensor fidelity increases, sensors are generally placed closer together. This could be due to the fact that noisy sensors need to be placed farther apart from each other than less noisy sensors in order to properly triangulate events.

We next investigate the effect of sensor fidelity on information gain. We examine the effect of fidelity by comparing the information gain surface generated by a grid of 9 evenly spaced sensors across 12 different signal-to-measurement noise ratio offsets. In this experiment, these evenly spaced offsets range from −3.0 to 3.5. We perform this experiment using both a uniform prior on events and the non-uniform fault-box prior. The results of these experiments can be seen in Fig. 10, and full visualizations of how the SNR affects IG across all events in a domain can be found in Figs A6 and A7. As when controlling the measurement noise standard deviation directly, we see that below a certain fidelity offset value the measurement noise dominates the signal and as such we see minimal information gain. Once past a certain threshold (in this case a sensor fidelity of −0.045 corresponding to a measurement noise standard deviation of 1.59) the model uncertainty begins to take over and we see an increase in information gain in both the uniform and non-uniform prior cases.

5.4 Optimizing a network with boundary constraints

We also examined the behaviour of the optimization when constraints were placed on the location of the sensors according to Fig. 11. This boundary was chosen based on the boundaries of the Uinta National Forest, which was chosen because the Uinta National Forest are irregular

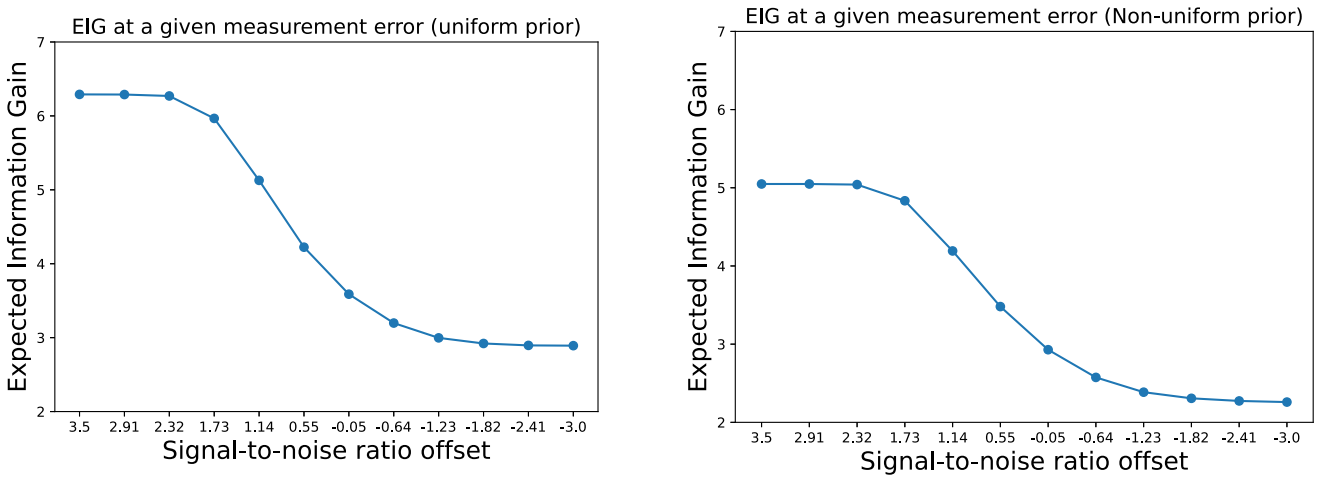

Figure 10. Illustrations of the degradation of EIG for all events on both priors as the signal-to-noise ratio is decreased. This analysis shows where measurement error dominates versus modelling error and vice versa. We see that EIG is fairly stable when SNR is offset by more than 1.73 or less than −0.64, and is greatly affected when the signal offset is between those values. See Table 1 for a description of the relationship between SNR and measurement error. For the average event, EIG is more sensitive to the measurement error when the SNR value is between 1.73 and −0.64. These larger fluctuations in measurement error correspond to the steeper curve in the plots. On the other hand, an SNR offset greater than 1.73 means that model error dominates, while an SNR offset of less than −0.64 means that measurement error dominates. This can inform where to invest effort in reducing uncertainty.

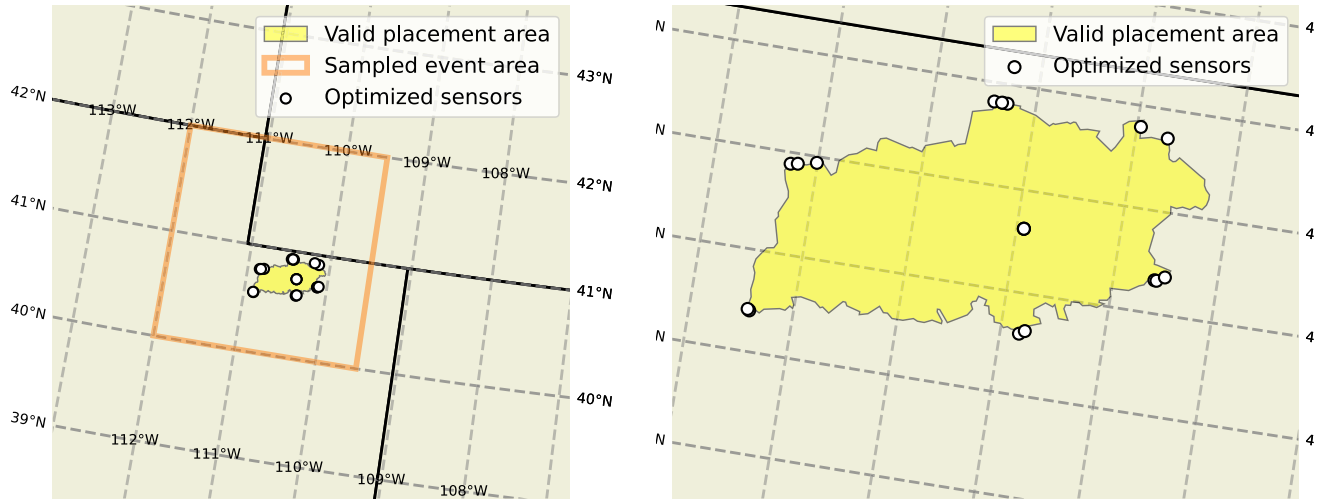

Figure 11. Network optimized under the shown boundary constraints. Twenty sensors were placed within the highlighted area. The figure on the left shows the domain on which the seismic models were trained (the outer area in Fig. 3), with the square box on the left being the area from which events were sampled (the inner area in Fig. 3). The figure on the right shows a zoomed in view of the admissible placement area shows just the area in which the events were sampled. We see that sensors were placed near the boundary of the admissible area. This is possibly done in an attempt to better capture events outside the optimization domain.

and therefore provide a good test for the bounded optimization software, and they are also entirely contained within the area on which our models were trained.

The network created by our script under our boundary constraints is shown in Fig. 11. We can see that sensors are placed on the edges of boundaries in order to gain information about the surrounding area.

5.5 Effect of correlation

We investigate the effect of station correlation on the placement of sensors. We look at three different correlation length scales, l: 14.75, 147.5 and 1475 km. We do this using both a uniform prior distribution and a fault-box prior. The signal-to-measurement-noise ratio in both cases was fixed at 0 (corresponding to a standard deviation of 1.5). Twenty stations were then placed using greedy optimization. In Fig. 12, we observe that at higher correlations EIG also increases. Since higher correlation means less information, this is what we would expect to see. However, we note that this EIG gain is relatively modest, especially as more sensors are placed. It is possible that this could mean that correlation does not have a large effect on sensor placement. The mean nearest neighbour distance is also very similar for all correlation values once the number of sensors grows large.

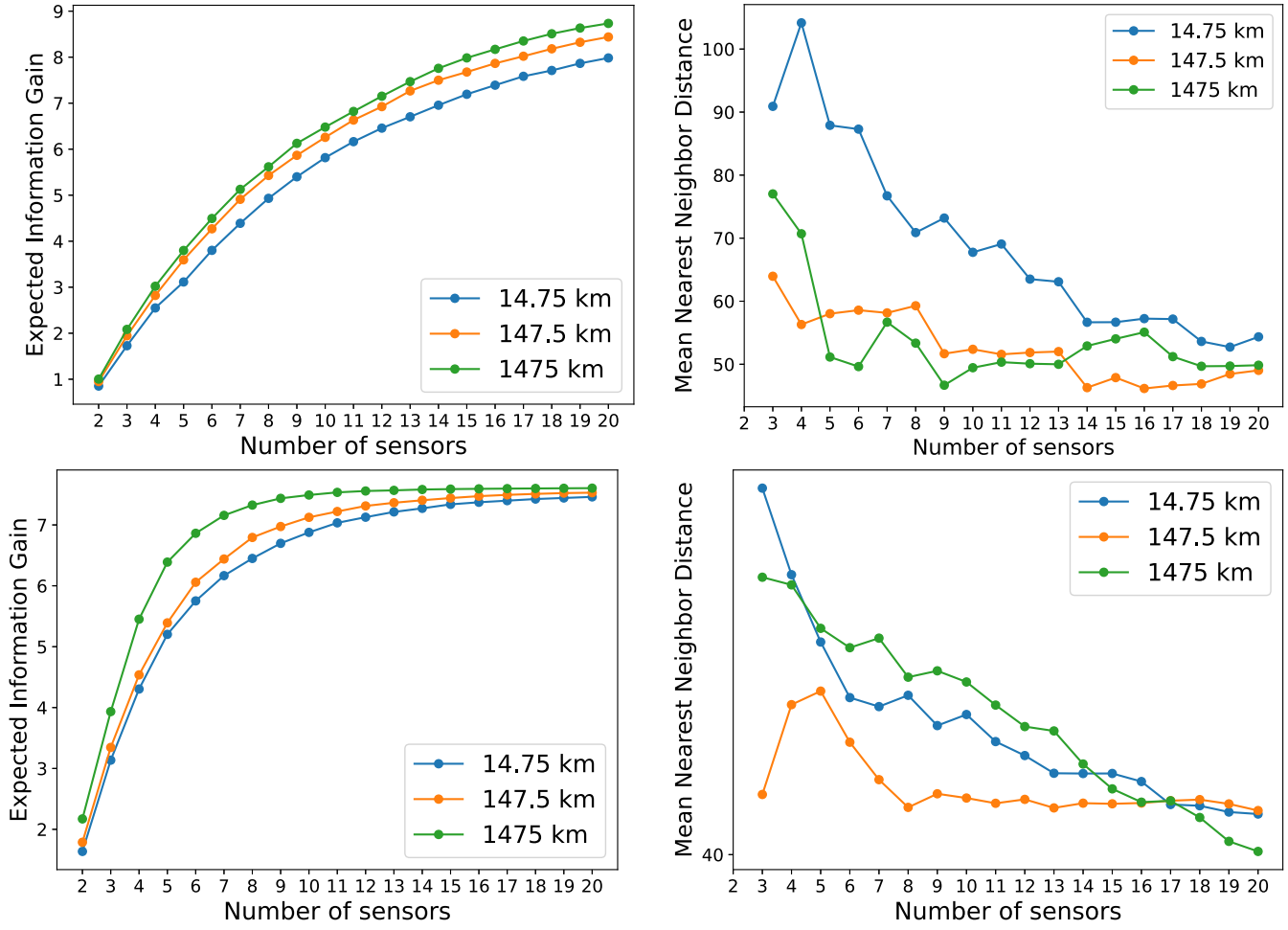

Figure 12. Analysis of the evolution of the EIG (left) and mean distance between station (right) for three different correlation length scales. Sensors were placed using both a non-uniform and uniform prior. *Top – Non-uniform prior:* We see that initially the disparity in EIG is small then increases with the number of sensors. However, after about 10 sensors, the disparity begins to decrease. Few patterns can be identified in the evolution of the geometry of the network although it may be the case that stations are initially further apart for the low correlation model. *Bottom – uniform prior:* We see that, like with the non-uniform prior, the disparity in EIG across correlation length scales starts small, grows and then shrinks again. Unlike the non-uniform prior, we see that EIG levels off sharply around six sensors. As with the non-uniform prior, few patterns can be discerned from the network geometry plot.

6 CONCLUSION

In this work, we have demonstrated and implemented a modern framework for Bayesian optimal experimental design for analysing and optimizing a seismic monitoring network. We used this framework on a seismic source location problem with uncertainty in both the detection of seismic phases and uncertainty in the arrival time. We selected these models using data from the U.S. Transportable Array and physics-based traveltime modelling with earth model uncertainty. Using these models, we capture the often-ignored influence of earth model uncertainty and station correlation on traveltimes. We further investigate the influence of station correlation, earth model uncertainty and phase-arrival pick uncertainty on the sensor placement and sensitivity of the monitoring network.

Our Bayesian OED approach will enable rigorous and flexible analysis and design of monitoring networks for applications like earthquake or explosion monitoring. When evaluating a monitoring network, decision makers in high-consequence domains can trust the rigor of the Bayesian approach to provide coherent uncertainty quantification. Further, decision makers may employ Bayesian OED to assess the monitoring network's sensitivity to different types of seismic sources and locations and therefore can certify the capabilities of the network to meet design requirements. Bayesian OED may answer other questions critical to seismic monitoring such as: how may multiphenomenology data be used to reduce uncertainty; what is the appropriate sensor fidelity or earth model resolution for estimating a QoI and how do sensor types, number and locations influence estimates of QoIs?

While this work provides a meaningful first step towards analysing and optimizing monitoring networks, many simplifications were made during this exploratory study. Based on these results we have identified several follow-on directions to increase its applicability to real monitoring problems:

(i) In this work, we used a very simple Gaussian traveltime model because it enabled marginalization of origin time and handling correlation between stations. As we saw, real data are much more complex and so more complex traveltime models should be explored.

(ii) Further, we may extend the correlation model to include more event characteristics. We ultimately assumed that the station correlation was independent of the event and was only a function of how far apart the stations were. Real data exhibits more complex correlation structures, such as depth dependence. Further, we assumed that the detections of each station were independent. Again, we would expect this not to be true.

(iii) We also assumed that the stations were identical. Studying a heterogeneous sensor network is much more realistic. Stations are heterogeneous both because of the use of different sensors but also based upon how the stations are installed, which could introduce different uncertainties and background noise environments. Modelling this heterogeneity also would enable us to better assess the trade-off between different sensor types and installation methods.

(iv) Develop methodology that is not data driven for novel sensor placements.

(v) Finally, we may incorporate many other sources of data into this analysis. We only considered P arrivals so other seismic phases should be studied using the same workflow and incorporated into the likelihood function. Also, infrasound sensors and seismic arrays could be included to make the analysis multimodal by providing directional information. This would then give us the ability to explore the utility of different sensor types as we could see how the expected information gain changes as we add sensors with these different modalities. We could also then deduce the types of seismic sources different data modalities most benefit.

ACKNOWLEDGMENTS

KM participated in this research while at Stanford's Institute for Computational & Mathematical Engineering as part of the Xplore program before joining Susquehanna International Group. The authors would like to thank Drs Brian Williams and Josh Carmichael from Los Alamos National Laboratory for their thoughtful and constructive feedback on the manuscript. This Low Yield Nuclear Monitoring (LYNM) research was funded by the National Nuclear Security Administration, Defense Nuclear Nonproliferation Research and Development (NNSA DNN R&D). The authors acknowledge important interdisciplinary collaboration with scientists and engineers from LANL, LLNL, NNSS, PNNL and SNL. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and 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, which is also archived in the Sandia report SAND2022-13022. 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. JC contributed to the algorithm design, conducted the experiments and wrote the manuscript under the supervision of TC. KM developed the initial statistical models. RV assisted with analysis of algorithm performance. TC managed the project, developed the theoretical framework and wrote the initial draft of the manuscript.

DATA AVAILABILITY

The sensor arrival and detection data used to build the likelihood models are from the IRIS Transportable Array data set collected between August 2007 and August 2008, available at https://anf.ucsd.edu/tools/events/. The earth model cross-sections used to build the traveltime models are from the Crust 1.0 data set, available at https://igppweb.ucsd.edu/~gabi/crust1.html. The accompanying software to this paper is hosted at https://github.com/sandialabs/seismic_boed.

REFERENCES

An, C., Liu, P.L. & Meng, L., 2018. A sensitivity analysis of tsunami inversions on the number of stations, *J. geophys. Int.,* **214**(2), 1313–1323.

Arora, N.S., Russell, S. & Sudderth, E., 2013. Net-visa: network processing vertically integrated seismic analysis, *Bull. seism. Soc. Am.,* **103**(2A), 709–729.

Arrowsmith, S., Park, J., Che, I.-Y., Stump, B. & Averbuch, G., 2020. Event location with sparse data: when probabilistic global search is important, *Seismol. Res. Lett.,* **92,** 976–985.

Beck, J.L., 2010. Bayesian system identification based on probability logic, *Struct. Control Health Monit.,* **17**(7), 825–847.

Bloem, H., Curtis, A. & Maurer, H., 2020. Experimental design for fully nonlinear source location problems: which method should i choose?, *J. geophys. Int.,* **223**(2), 944–958.

Böse, M., Papadopoulos, A.N., Danciu, L., Clinton, J.F. & Wiemer, S., 2022. Loss-based performance assessment and seismic network optimization for earthquake early warning, *Bull. seism. Soc. Am.,* **112**(3), 1662–1677.

Brooks, S., Gelman, A., Jones, G. & Meng, X.-L., 2011. *Handbook of Markov Chain Monte Carlo,* CRC press.

Burmin, V.Y., 2019. Optimal geometry for the seismological observation network in the Caucasus region, *Seismic Instrum.,* **55**(3), 353–362.

Carmichael, J., Nemzek, R., Symons, N. & Begnaud, M., 2020. A method to fuse multiphysics waveforms and improve predictive explosion detection: theory, experiment and performance, *J. geophys. Int.,* **222**(2), 1195–1212.

Carmichael, J.D., 2020. Hypothesis tests on Rayleigh wave radiation pattern shapes: a theoretical assessment for source screening, *Geophys. J. Int.,* **Vol. 225**(3), 1653–1671.

Catanach, T., Callahan, J., Monogue, K. & Villareal, R., 2024. https://github.com/sandialabs/seismic_boed.

Coles, D. & Curtis, A., 2011. Efficient nonlinear Bayesian survey design using DN optimization, *Geophysics,* **76**(2), Q1–Q8.

Coles, D. & Prange, M., 2012. Toward efficient computation of the expected relative entropy for nonlinear experimental design, *Inverse Probl.,* **28**(5), 055019, doi:10.1088/0266-5611/28/5/055019.

Crotwell, H.P., Owens, T.J. & Ritsema, J., 1999. The taup toolkit: flexible seismic travel-time and ray-path utilities, *Seismol. Res. Lett.,* **70**(2), 154–160.

Curtis, A., Michelini, A., Leslie, D. & Lomax, A., 2004. A deterministic algorithm for experimental design applied to tomographic and microseismic monitoring surveys, *J. geophys. Int.,* **157**(2), 595–606.
Dalcin, L. & Fang, Y.-L.L., 2021. mpi4py: Status update after 12 years of development, *Comput. Sci. Eng.,* **23**(4), 47–54.
EarthScope ANF Website, 2021. Network stations :: Monthly deployment history. Available at: http://anf.ucsd.edu/cacheimages/maps/monthly_deployment/deploymap_2007_12.rolling.png, Last accessed on 2023-09-03.
Frazier, P.I., 2018. A tutorial on Bayesian optimization, preprint(arXiv:1807.02811).
Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B., 1995. *Bayesian Data Analysis,* Chapman and Hall/CRC.
Ginebra, J. *et al.* 2007. On the measure of the information in a statistical experiment, *Bayesian Anal.,* **2**(1), 167–211.
Guest, T. & Curtis, A., 2011. On standard and optimal designs of industrial-scale 2-d seismic surveys, *J. geophys. Int.,* **186**(2), 825–836.
Head, T., Kumar, M., Nahrstaedt, H., Louppe, G. & Shcherbatyi, I., 2020. *scikit-optimize/scikit-optimize*.
Huan, X. & Marzouk, Y.M., 2013. Simulation-based optimal Bayesian experimental design for nonlinear systems, *J. Comput. Phys.,* **232**(1), 288–317.
IRIS Transportable Array, 2003. *Usarray transportable array*.
Jaynes, E.T., 2003. *Probability Theory: The Logic of Science,* Cambridge Univ. Press.
Jones, D.R., Schonlau, M. & Welch, W.J., 1998. *J. Global Optim.,* **13**(4), 455–492.
Kennedy, M.C. & O'Hagan, A., 2001. Bayesian calibration of computer models, *J. R. Stat. Soc.: Ser. B (Stat. Method.),* **63**(3), 425–464.
Koval, K., 2021. *Optimal Experimental Design for Bayesian Inverse Problems Governed by PDE Models with Uncertainty with Application to Subsurface Flow and Tsunami Equations,* Ph.D. thesis, New York University.
Krause, A., Singh, A. & Guestrin, C., 2008. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies, *J. Mach. Learn. Res.,* **9**(Feb), 235–284.
Laske, G., Masters, G., Ma, Z. & Pasyanos, M., 2013. Update on crust1.0—a 1-degree global model of earth's crust, in *Geophys. Res. Abstr,* Vol. **15,** pp. 2658.
Lindley, D.V., 1956. On a measure of the information provided by an experiment, *Ann. Math. Stat.,* **27,** 986–1005.
Long, Q., Motamed, M. & Tempone, R., 2015. Fast Bayesian optimal experimental design for seismic source inversion, *Comput. Methods Appl. Mech. Eng.,* **291,** 123–145.
MacKay, D.J., 2003. *Information Theory, Inference and Learning Algorithms,* Cambridge Univ. Press.
Maupin, K.A. & Swiler, L.P., 2020. Model discrepancy calibration across experimental settings, *Reliab. Eng. Syst. Safety,* **200,** 106818, doi:10.1016/j.ress.2020.106818.
Maurer, H., Curtis, A. & Boerner, D.E., 2010. Recent advances in optimized geophysical survey design, *Geophysics,* **75**(5), 75A177–75A194.
Merchant, B.J., 2013. *Netmod, Version 00*.
Močkus, J., 1975. On Bayesian methods for seeking the extremum, in *Optimization Techniques IFIP Technical Conference,* pp. 400–404, ed. Marchuk, G.I., Springer.
Myers, S.C., Johannesson, G. & Hanley, W., 2007. A Bayesian hierarchical method for multiple-event seismic location, *J. geophys. Int.,* **171**(3), 1049–1063.
Oth, A., Böse, M., Wenzel, F., Köhler, N. & Erdik, M., 2010. Evaluation and optimization of seismic networks and algorithms for earthquake early warning–the case of Istanbul (Turkey), *J. geophys. Res.: Solid Earth,* **115**(B10), doi:10.1029/2010JB007447.
Owen, A.B., 2013. *Monte Carlo Theory, Methods and Examples*.
Papadimitriou, C., Haralampidis, Y. & Sobczyk, K., 2005. Optimal experimental design in stochastic structural dynamics, *Probab. Eng. Mech.,* **20**(1), 67–78.
Picard, R., Williams, B. & Weaver, B., 2019. Estimating normalizing constants for posterior densities via ex post facto sampling, *J. Indian Stat. Assoc.,* **57**(2), 173–204.
Picheny, V., Ginsbourger, D., Richet, Y. & Caplin, G., 2013. Quantile-based optimization of noisy computer experiments with tunable precision, *Technometrics,* **55**(1), 2–13.
Srinivas, N., Krause, A., Kakade, S.M. & Seeger, M., 2010. Gaussian process optimization in the bandit setting: No regret and experimental design, in *Proceedings of the 27th International Conference on Machine Learning (ICML-10),* pp. 1015–1022, eds, Fürnkranz, J. & Joachims, T., Omnipress, Haifa, Israel.
Steinberg, D.M. & Rabinowitz, N., 2003. Optimal seismic monitoring for event location with application to on site inspection of the comprehensive nuclear test ban treaty, *Metrika,* **58,** 31–57.
Toledo, T., Jousset, P., Maurer, H. & Krawczyk, C., 2020. Optimized experimental network design for earthquake location problems: applications to geothermal and volcanic field seismic networks, *J. Volc. Geotherm. Res.,* **391,** 106433, doi:10.1016/j.jvolgeores.2018.08.011.
Velasco, A., Young, C. & Anderson, D., 2001. *Uncertainty in Phase Arrival Time Picks for Regional Seismic Events: An Experimental Design*.
Williams, B., 2021. *Bayesian Optimal Sensor Augmentation via Estimated Mutual Information*.
Williams, C.K. & Rasmussen, C.E., 2006. *Gaussian Processes for Machine Learning,* Vol. **2,** MIT press.
Yang, Y., Chadha, M., Hu, Z. & Todd, M.D., 2022. An optimal sensor placement design framework for structural health monitoring using Bayes risk, *Mech. Syst. Signal Process.,* **168,** 108618, doi:10.1016/j.ymssp.2021.108618.
Yuen, K.-V. & Kuok, S.-C., 2015. Efficient Bayesian sensor placement algorithm for structural identification: a general approach for multi-type sensory systems, *Earthq. Eng. Struct. Dyn.,* **44**(5), 757–774.

APPENDIX A: MODELLING DETAILS

A1 Detection model

We re-emphasize that only 23 per cent of the event-sensor pairs contained a detection. We therefore tuned our data to balance the performance of the model. We fit the logistic regression model by minimizing binary cross entropy loss across the data set. This loss comes from the KL divergence between the predicted detection probability and the realized detection. Placing higher weight in the loss function on detections biases the model to predicting detection and balances the data set composition. We experimented with different weights, Table A1 summarizes these results.

A weighting of 2 was chosen to maintain accuracy while providing a significant recall boost (catching the actual positive detections, which contribute more uniquely to the information gain).

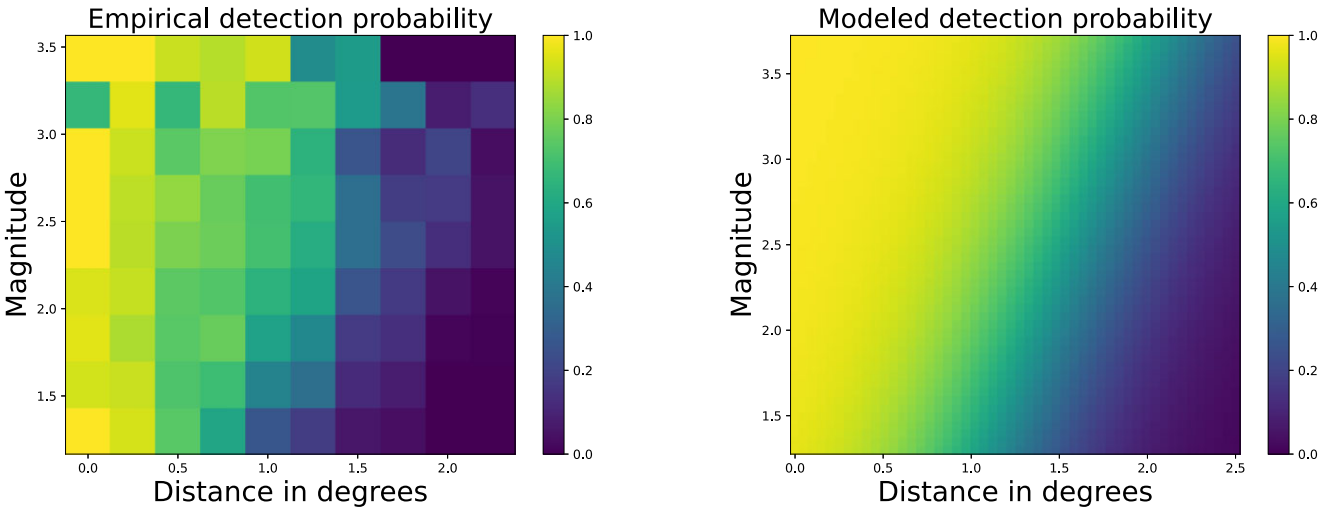

Figure A1. The left panel shows the estimated detection probability from the transportable array data set. We binned these data by distance and magnitude and then estimated the detection probability as the number of detections versus the number of potential detections for stations and events within a given distance and magnitude. The right panel shows the detection probability predicted by a logistic regression model fit to the data from the transportable array data set. The modelled detection probabilities appear as a smoother version of the empirical data histogram, indicating that the model captures the underlying distribution of the data well.

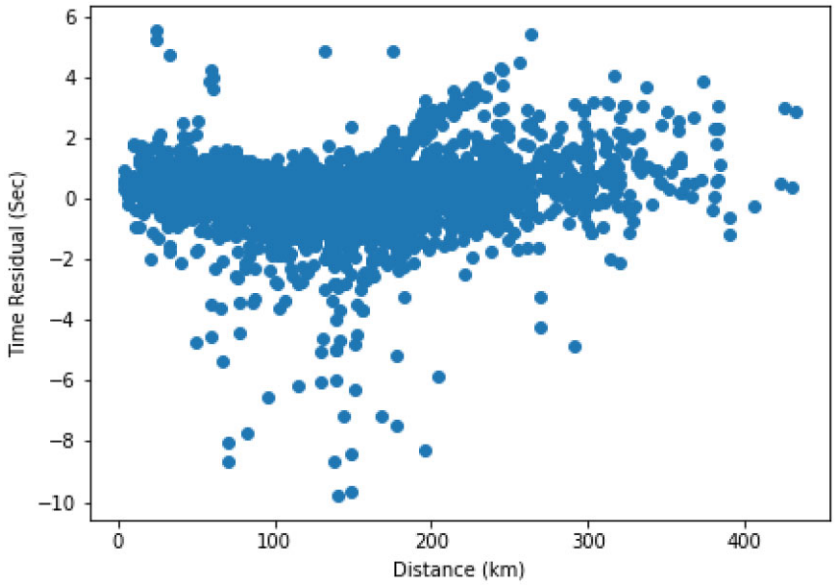

Figure A2. Arrival time residual as a function of distance in kilometres. Note that there is a bias towards positive residuals, particularly at longer distances. This bias is particularly evident in Fig. A3.

A2 Arrival time model

A2.1 Data-driven model

While the analysis in the rest of this document does not rely on a data-driven arrival time likelihood model, it is helpful to consider the complexities of real arrival time data to understand some of the modelling choices. Again, the same transportable array data set was used. For each P arrival, we predict the arrival time for a phase given the event and sensor locations and origin time in the catalogue using the IASP91 velocity model. Then, we calculated the residuals observed in the data. Fig. A2 shows a scatter plot of the residual data as a function of distance.

We see little obvious relationship between distance and the residual. This residual is probably a combination of many factors: measurement noise, traveltime model errors, phase categorization errors and location errors in the catalogue. In the histogram Fig. A3, we see that the residuals follow a heavy tailed distribution. We choose to model it with a non-centred t-distribution.

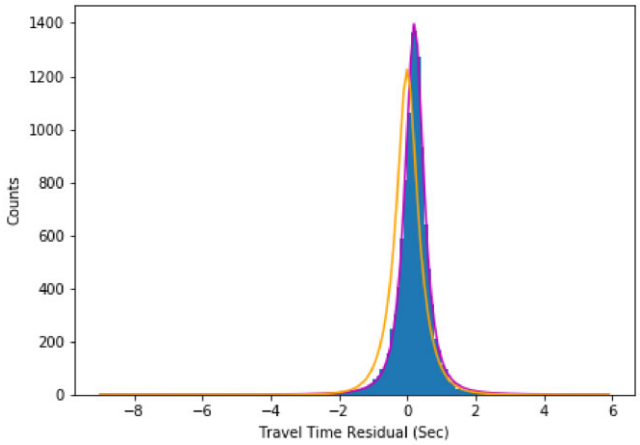

Figure A3. Arrival time residual histogram compared to the fit of different statistical models. The magenta line indicates the non-centred t-distribution fit to the data. The orange line shows the marginal residual distribution under the more tractable distance dependent Gaussian model discussed in Section 3.3.1. We see that the non-centred t t-distribution is a better fit to the data than the Gaussian model. Note that the non-data driven model does not *a priori* know the bias so it is centred at zero.

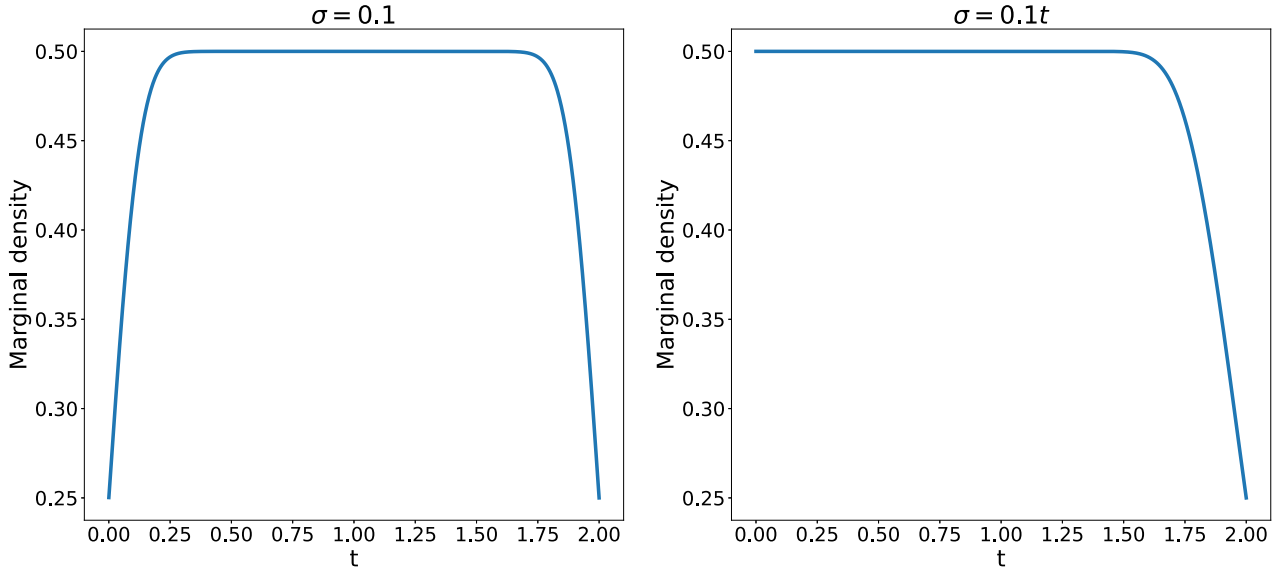

Figure A4. These plots show the 1-D marginal likelihood for a given traveltime prediction t if the origin time t_0 is in $[a - T, a]$ (meaning the origin time is before the measured arrival time). For an improper uniform prior, this marginal likelihood should be constant on $(-\infty, \infty)$. In the plot on the left, an interval length of $T = 2$ and a standard deviation of $\sigma = 0.1$ are used. The marginal likelihood matches that of an improper prior except near $t = 0$ and $t = T$. On the right, an interval length of $T = 2$ and a standard deviation of $\sigma = 0.1t$ (i.e. the error is a percentage of the mean traveltime) are used. Here differences only appear near $t = T$, and the likelihood is otherwise constant.

For the non-centred t-distribution we fit the data and found that the degrees of freedom parameter was 2.198, location parameter was 0.214 and scale parameter was 0.293. While this distribution fits the data reasonably well, it does not give us the ability to tune the various sources of uncertainty when analysing and optimizing the seismic network. Further, considering station correlation and marginalizing out origin time uncertainty is very hard for this distribution. Therefore, we instead turn to a simple Gaussian distribution because the Gaussian distribution allows us to easily model correlation and marginalize out origin time uncertainty analytically. Finally, for the purpose of Bayesian OED, ignoring the bias and choosing to use a zero mean Gaussian is justified because adding a constant, known, bias to all traveltimes would only affect the time of the arrivals but not their uncertainty and therefore the likelihood would be the same. Obviously, for inference with real data including the bias is necessary.

A2.2 Improper uniform prior

Fig. A4 shows an empirical comparison between a marginal mean traveltime likelihood using an improper prior and one using a proper prior. We see that differences between the two arise only on the boundaries of the domain of the proper prior. We note that predicted mean traveltimes are necessarily restricted by the size of the event domain, so when the proper uniform prior is wide enough to accommodate all

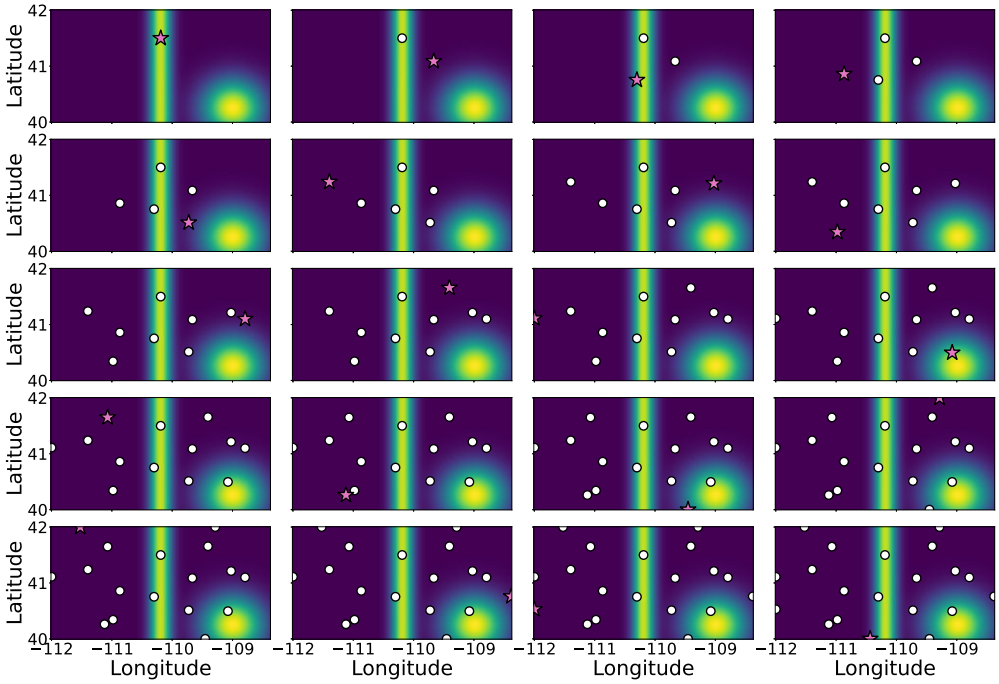

Figure A5. Illustration of the sequential placement of sensors on a square domain when using a non-uniform prior on events. Each plot shows the sensors placed in previous steps in white and the sensor placed at the current step represented as a star. The location component of the prior is depicted by the heat map underneath the sensors, with warmer areas indicating areas of higher prior density.

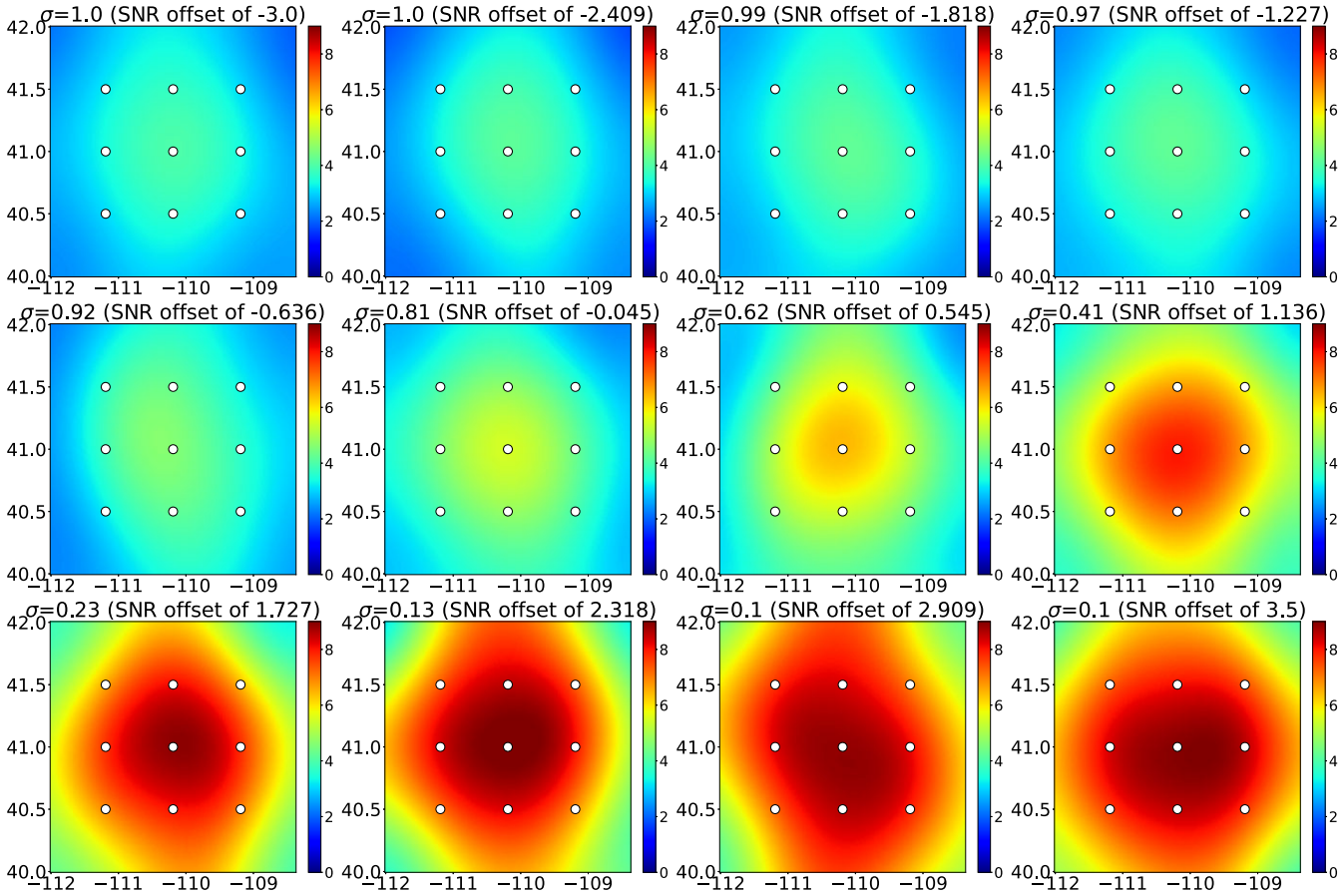

Figure A6. Illustration of the effect of changing the measurement uncertainty standard deviation over several orders of magnitude when using a uniform distribution. The colour plots illustrate the EIG of a shallow seismic source with the different stated measurement errors. For noise levels below about 0.41 s the model uncertainty dominates over measurement uncertainty so EIG is fairly stable.

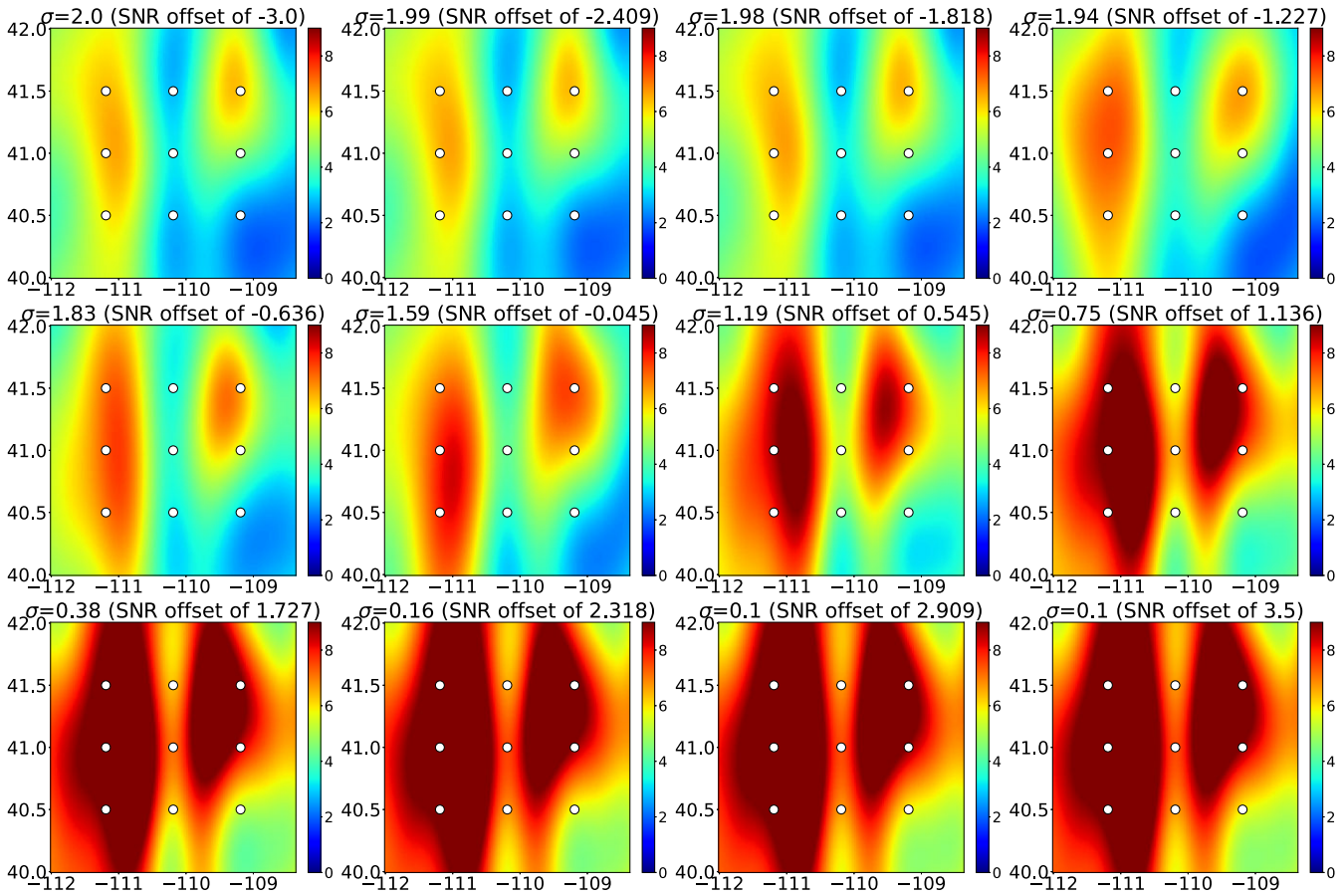

Figure A7. Illustration of the effect of changing the measurement uncertainty standard deviation over several orders of magnitude when using a non-uniform distribution. The colour plots illustrate the EIG of a shallow seismic source with the different stated measurement errors. For noise levels below about 1.59 s the model uncertainty dominates over measurement uncertainty so EIG is fairly stable.

Table A1. Over weighting of detections in the loss function was considered to correct for the data set imbalance towards non-detections. A weight of 2 was chosen to balance the different performance metrics.

Detection weight	Accuracy	Precision	Recall	AUC
1	0.870	0.728	0.686	0.92
2	0.865	0.665	0.820	0.92
3	0.850	0.617	0.874	0.92
4	0.835	0.583	0.905	0.92
5	0.819	0.553	0.920	0.92

possible traveltimes, such discrepancies will not influence the likelihoods of any observed data. In this case, the proper and improper priors are functionally equivalent.

A3 Importance sampling

For many applications of OED, there are reasonable prior distributions from which sampling is prohibitively difficult (e.g. complex fault geometries). Further, even when sampling from a prior is easy, it might not be the most computationally efficient method for estimating the integrals in (21) because events that are rare according to the prior may contribute significantly to the integral (e.g. high magnitude seismic events that are likely to cause very high information gains). Thus, to be able to fully utilize domain knowledge about areas of interest in an efficient manner it is important that we have a way to representatively sample from these challenging priors and important events. Importance sampling is one such way.

We draw samples from an importance distribution ($q(\theta')$, a distribution different from the prior but one that is possible to sample), weight those samples according to the probability density function of our target prior distribution, $p(\theta')$ and use these weighted samples to approximate our quantities of interest. This means that in Algorithm 1 instead of $\theta' \sim p(\theta')$, we have that $\theta' \sim q(\theta')$ and has a corresponding weight of $w(\theta') = p(\theta')/q(\theta')$. For a detailed discussion of importance sampling see Owen (2013).

A4 Computing EIG

The EIG is computed by first generating M event samples $\{\theta_i'\}_{i=1}^{M}$. These samples may be generated from the prior distribution or from some other distribution of interest. Then, K synthetic data observations are drawn from the likelihood for each event: $\{D_j^{(i)}\}_{j=1}^{K} \sim p(D \mid \theta_i)$. This yields $M \times K$ event-observation pairs $(\theta_i', D_j^{(i)})$.

Next, we sample N parameters, $\{\theta_n\}_{n=1}^{N} \sim q(\theta)$, where $q(\theta)$ is an important distribution for the prior distribution. This distribution is sampled using a space-filling design (e.g. a uniform distribution sampled using a QMC mesh), ensuring coverage of the parameter space. For a given data set $D_j^{(i)}$, we then compute the KL divergence between posterior $p(\theta \mid D_j^{(i)})$ and prior $p(\theta)$. The log-ratio term in the KL divergence can be rewritten using Bayes Rule:

$$\log\left(\frac{p(\theta \mid D)}{p(\theta)}\right) = \log\left(\frac{p(D \mid \theta)}{p(D)}\right),$$

so we have

$$\begin{aligned}\mathrm{KL}\left[p(\theta \mid D) \,||\, p(\theta)\right] &= \int_\theta p(\theta \mid D)\left[\log(p(D \mid \theta) - \log(p(D))\right] \\ &= \int_\theta q(\theta)\frac{p(\theta)}{q(\theta)}\frac{p(D \mid \theta)}{p(D)}\left[\log(p(D \mid \theta) - \log(p(D))\right] \\ &= \mathbb{E}_{q(\theta)}\left[\frac{p(\theta)}{q(\theta)}\frac{p(D \mid \theta)}{p(D)}\left[\log(p(D \mid \theta) - \log(p(D))\right]\right].\end{aligned} \tag{A1}$$

Here, observe that

$$\begin{aligned}p\left(D_j^{(i)}\right) &= \int p\left(D_j^{(i)} \mid \theta\right) p(\theta)\,\mathrm{d}\theta \\ &= \mathbb{E}_{q(\theta)}\left[p\left(D_j^{(i)} \mid \theta\right)\frac{p(\theta)}{q(\theta)}\right] \\ &\approx \frac{1}{N}\sum_{n=1}^{N} p\left(D_j^{(i)} \mid \theta_n\right)\frac{p(\theta_n)}{q(\theta_n)}.\end{aligned}$$

Letting

$$w_n = \tfrac{p(\theta_n)}{q(\theta_n)}\, p\left(D_j^{(i)} | \theta_n\right),$$

we can write the evidence as

$$p\left(D_j^{(i)}\right) \approx \frac{1}{N}\sum_{n=1}^{N} w_n,$$

so substituting into (A1) yields

$$\mathrm{KL}\left[p\left(\theta \mid D_j^{(i)}\right) \,||\, p(\theta)\right] \approx \sum_{n=1}^{N}\frac{w_n}{\sum_{n=1}^{N} w_n}\left(\log(p(\theta \mid D_j^{(i)})) - \log\left(\frac{1}{N}\sum_{n=1}^{N} w_n\right)\right).$$

Then, since we may treat the pair (θ_i', D_j) as a draw from the joint distribution $p(\theta', D)$, we have that

$$\begin{aligned}I(\mathcal{S}) &= \int p\left(\theta'\right)\int p\left(\mathcal{D} \mid \theta', \mathcal{S}\right)\int p\left(\theta \mid \mathcal{D}, \mathcal{S}\right)\log\frac{p\left(\theta \mid \mathcal{D}, \mathcal{S}\right)}{p\left(\theta\right)}\,\mathrm{d}\theta\,\mathrm{d}\mathcal{D}\,\mathrm{d}\theta' \\ &= \int p(\theta', \mathcal{D} \mid \mathcal{S})\mathrm{KL}\left[p(\theta \mid \mathcal{D}) \,||\, p(\theta)\right]\,\mathrm{d}\mathcal{D}\,\mathrm{d}\theta' \\ &= \mathbb{E}_{p(\theta', \mathcal{D}|\mathcal{S})}\left[\mathrm{KL}\left(p(\theta \mid \mathcal{D}) \,||\, p(\theta)\right)\right] \\ &\approx \frac{1}{MK}\sum_{i=1}^{M}\sum_{j=1}^{K}\mathrm{KL}\left[p\left(\theta \mid D_j^{(i)}\right) \,||\, p(\theta)\right]\end{aligned}, \tag{A2}$$

where $D_j^{(i)}$ denotes the j^{th} draw from $p(D \mid \theta_i)$.

Unlike the traditional double-nested Monte Carlo (DNMC) approach to computing EIG (Huan & Marzouk 2013), this method directly computes the expectation of the KL divergence by 'gridding' the parameter space and approximating the value of the posterior at each grid point. It is likely that the DNMC method is faster in general, but since we can pre-compute likelihoods and because the parameter space is small, the computation is not too expensive.

The advantage of this approach is that we have direct access to the approximate KL divergence values at each point in the domain, which allows us to generate the information surfaces shown in Figs 8, A6 and A7. It also allows us to investigate the information gain about specific events of interest in the domain since we have access to the distribution of KL divergences at each event. Given the low dimensionality of the parameter domain, this added interpretability justifies the small trade-off in computational cost.

A5 Description of software implementation

The accompanying software to this paper is hosted at https://github.com/sandialabs/seismic_boed. The user can specify models for generating synthetic data and assessing the likelihood of that synthetic data for different sensors and events in the domain of candidate events. The code is separated into two main components: analysis network and optimization network. The code is designed to use MPI so that it can run on HPC resources. The multicore parallelism through MPI is implemented by MPI4PY (Dalcin & Fang 2021). The EIG computational is highly parallelizable so it can be scaled easily to thousands of cores, which is important due to the number of computations required for robust estimates of the EIG, particularly when making the sensitivity maps to show how the network performs on specific events.

The analysis code estimates the EIG of a given seismic monitoring networks for a user-defined prior distribution of potential events. As described in Algorithm 1, the code samples these candidate events and then generates synthetic data sets that could plausibly be seen by the sensors. Likelihood models for several sensor types are provided but user-specified models can also be used. For each of the data sets the code constructs the posterior distribution and computes the information gain IG according to the KL-divergence. This information gain is averaged over all synthetic data sets to compute the EIG. The code can also return a list of the IG for different hypothetical events which can be used to generate a map of sensitivities of the network to different event locations, depths and magnitudes. See Fig. 8 for an example.

The optimization code is a wrapper around the analysis code. Given a specified initial network configuration of sensors, the code will add a desired number of sensors to the network. The goal of the optimization is to maximize the EIG of the new sensor network while respecting user-specified constraints on where sensors can be placed. This is done with a sequential (greedy) optimization that adds sensors one at a time to the initial network. Each optimization is done using a Bayesian optimization method that construct a Gaussian process surrogate model of the EIG optimization surface. This is done by evaluating many potential new sensor locations and measuring the EIG using the analysis code. These data are then used to construct the surrogate and inform new trial points to query the EIG function. The code then returns the new sensor network after the optimal sensors have been added.

Please refer to the documentation in Catanach *et al.* (2024) for a complete description of the code, capabilities and provided tutorials.

A6 More results

This section contains additional figures illustrating various results of the paper. Fig. A5 demonstrates the process of sequentially placing sensors using a non-uniform prior. Figs A6 and A7 show the effect of changing the sensor fidelity on expected information gain. The EIG was computed using 8192 sampled events to generate synthetic data over a domain discretized into 32 768 points. These numbers were chosen to ensure that the effective sample size (ESS), which measures how well the target distribution is represented by the weighted samples (Owen 2013), remained relatively large while keeping the computational cost feasible. Since the other approximations we make (e.g. greedy optimization) are likely to be more impactful than uncertainty in the EIG estimator, we prioritized computational feasibility over fine-tuning the number of samples used.