
FNOPE: Simulation-based inference on function spaces with Fourier Neural Operators

Guy Moss^{1,2,*}

Leah Sophie Muhle³

Reinhard Drews³

Jakob H. Macke^{1,2,4,†}

Cornelius Schröder^{1,2,†,*}

¹Machine Learning in Science, University of Tübingen, Tübingen, Germany

²Tübingen AI Center, Tübingen, Germany

³Department of Geosciences, University of Tübingen, Tübingen, Germany

⁴Department Empirical Inference, Max Planck Institute for Intelligent Systems, Tübingen, Germany

[†]Joint supervision

Abstract

Simulation-based inference (SBI) is an established approach for performing Bayesian inference on scientific simulators. SBI so far works best on low-dimensional parametric models. However, it is difficult to infer function-valued parameters, which frequently occur in disciplines that model spatiotemporal processes such as the climate and earth sciences. Here, we introduce an approach for efficient posterior estimation, using a Fourier Neural Operator (FNO) architecture with a flow matching objective. We show that our approach, FNOPE, can perform inference of function-valued parameters at a fraction of the simulation budget of state of the art methods. In addition, FNOPE supports posterior evaluation at arbitrary discretizations of the domain, as well as simultaneous estimation of vector-valued parameters. We demonstrate the effectiveness of our approach on several benchmark tasks and a challenging spatial inference task from glaciology. FNOPE extends the applicability of SBI methods to new scientific domains by enabling the inference of function-valued parameters.

1 Introduction

Probabilistic inference of mechanistic parameters in numerical models is a ubiquitous task across many scientific and engineering disciplines. Among methods for Bayesian inference, simulation-based inference (SBI, [1–6]) has emerged as a powerful approach for performing inference without requiring explicit formulation or evaluation of the likelihood. Instead, SBI only requires a simulator model which can sample from the likelihood. By training a generative model on pairs of parameters and simulation outputs, SBI can directly estimate probability distributions such as the posterior distribution.

However, existing SBI methods are designed to infer a limited number of vector-valued parameters, which strongly limits their use for inferring spatially and/or temporally varying, function-valued parameters. In these cases, parameters are commonly inferred on fixed discretizations of the domain. Despite some recent advances leveraging generative models to infer higher-dimensional posterior

*{firstname.secondname}@uni-tuebingen.de

Code available at <https://github.com/mackelab/fnope>

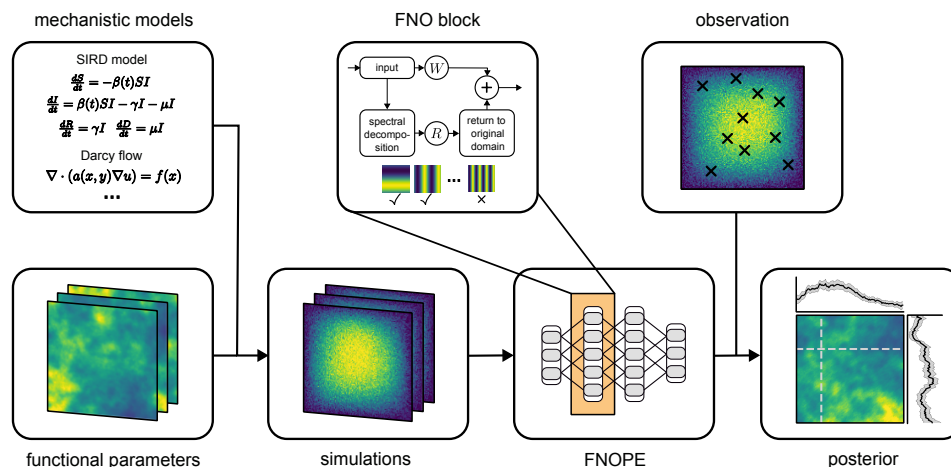


Figure 1: **Overview.** FNOPE approximates the posterior over function-valued parameters of a mechanistic model conditioned on function-valued observations. We use a FNO architecture with a flow matching objective to efficiently represent the function-valued parameters, enabling us to estimate extremely high dimensional posterior distributions at arbitrary discretizations of the domain.

distributions [7–11], the high-dimensional inference problems that arise from such approaches remain a challenge.

Furthermore, current models need to be retrained for new discretizations of the parameters or the observations. This is particularly challenging in fields like the geosciences, where observations cannot always be made at the same locations. An alternative to using fixed discretizations is to represent the functions using a fixed set of basis functions, where the inference problem becomes inferring the basis function coefficients, as used, e.g., in [12]. However, these approaches require a good selection of basis functions and suffer from a trade-off between choosing sufficiently expressive basis sets, while maintaining a tractable number of parameters to infer.

To overcome these limitations, we require methods that are capable of modeling and inferring function-valued data. Here, we propose to make use of the Fourier Neural Operator (FNO, [13]) architecture, which operates on function-valued data, for performing SBI on function-valued parameters. Neural operators [14–16] combine operations on global features of function-valued data with local (typically pointwise) operations, thus capturing both global and local structures. In particular, FNOs use Fourier features to model the global structure. For smoothly varying data, the spectral power is concentrated in the lower frequency components of the spectral decomposition. This allows for a compact representation of the global structure of the data, and hence for the inference of function-valued data on high resolution discretizations.

We present FNOPE (Fig. 1), an inference method for *function-valued parameters*: It trains **FNOs** for **Posterior Estimation** using a flow-matching objective [9, 17]. FNOPE is capable of solving inference problems for spatially and temporally varying parameters, and can generalize to posterior evaluations on non-uniform, previously unseen discretizations of the parameter and observation domains. Furthermore, FNOPE can estimate additional, vector-valued parameters of the simulator. We demonstrate these features on a collection of benchmark tasks, as well as a challenging real-world task from glaciology. We compare the performance of FNOPE to SBI approaches that use fixed-discretization or basis-function representation of the parameters, and show that FNOPE outperforms these methods, especially for low simulation budgets. Thus, FNOPE enables efficient inference for high dimensional inference problems that were previously challenging or even intractable.

2 Preliminaries

2.1 Simulation-based inference

SBI is designed to solve stochastic inverse problems: Given a simulator parametrized by θ , a known prior distribution $p(\theta)$ and an observation $x \in \mathbb{R}^{N_x}$, the goal is to infer the posterior $p(\theta | x)$ for

(typically) vector-valued parameters $\theta \in \mathbb{R}^{N_\theta}$. The simulator implicitly defines the model likelihood $p(x | \theta)$ by allowing us to sample $x \sim p(x | \theta)$. We can construct a training dataset by sampling from the joint $p(\theta)p(x | \theta)$ to construct a dataset of simulations $S = \{(\theta_i, x_i)\}_{i=1}^K$ for a number of simulations K . Standard approaches in neural posterior estimation (NPE) approximate the posterior $q^\phi(\theta | x)$ with a normalizing flow, which is trained by minimizing the negative log-likelihood $-\mathbb{E}_{(\theta,x) \sim S} \log q^\phi(\theta | x)$ [1, 3]. In contrast to this, flow-matching posterior estimation (FMPE) [9] learns a conditional velocity field $v_{t,x}^\phi(\theta_t)$ to iteratively denoise samples from a base distribution (typically a Gaussian distribution) to the posterior distribution $p(\theta | x)$. The velocity $v_{t,x}^\phi$ is trained via the flow matching objective

$$\mathcal{L}_{\text{FMPE}} = \mathbb{E}_{t \sim \mathcal{U}[0,1], (\theta,x) \sim S, z_t \sim p_t(z_t | \theta)} \|v_{t,x}^\phi - u_t(z_t | \theta)\|^2, \quad (1)$$

where $p_t(z_t | \theta)$ are the sample-conditional flow paths for z_t , and $u_t(z_t | \theta)$ are the true velocity fields. The sample-conditional paths are chosen so that p_t and u_t are analytically tractable.

2.2 Fourier Neural Operators

We use FNOs [13] to efficiently learn the posterior distribution of function-valued parameters. FNOs are a class of neural operators using the Fourier basis as an intermediate representation of functional data to learn mappings between function spaces. We assume to have a bounded domain $D \subset \mathbb{R}^d$, on which we define function spaces $\mathcal{A}(D; \mathbb{R}^{d_a})$ and $\mathcal{B}(D; \mathbb{R}^{d_b})$. The goal of neural operators is to approximate some given operator $\mathcal{G} : \mathcal{A} \rightarrow \mathcal{B}$ by a learnable operator $\tilde{\mathcal{G}}^\phi : \mathcal{A} \rightarrow \mathcal{B}$. In practice, the function-valued data is represented as discretizations of sample functions $a_i \in \mathcal{A}, b_i \in \mathcal{B}$ on the domain D . A single-layer FNO, $\tilde{\mathcal{G}}^\phi : \mathcal{A} \rightarrow \mathcal{B}$, is defined by

$$b(x) = \sigma(W^\phi a(x) + (\mathcal{K}^\phi a)(x)) \quad \forall x \in D, \quad (2)$$

where W_ϕ is a learnable linear operator, σ is a (pointwise) non-linearity, and

$$(\mathcal{K}^\phi a)(x) = \mathcal{F}^{-1}(R^\phi(\mathcal{F}a))(x) \forall x \in \mathcal{D}.$$

Here, $\mathcal{F}$ and $\mathcal{F}^{-1}$ refer to the Fourier and inverse Fourier transformation, and R^ϕ refers to some operator acting on the Fourier modes of a . Typically, R^ϕ is a linear transformation, and therefore corresponds to a convolution in real space with $\mathcal{K}^\phi$. But typically R^ϕ only acts on the lower Fourier modes, discarding higher ones, and therefore gives rise to a compact representation of high-resolution data. However, as Eq. 2 includes the linear operator $W^\phi a(x)$, FNOs are still able to capture local structures.

3 Method

To extend the standard SBI setting to inferring function-valued parameters, we develop FNOPE by extending FMPE with FNOs as backbone (Figs. 1,2). FNOPE takes the function-valued parameters θ and observations x as input, and estimates the FMPE flow-field v^ϕ for function-valued parameters using a combination of several FNO blocks.

We assume that θ as well as x are evaluated on discretizations specified by positions l^θ and l^x , which means we choose it from some set of $(l^\theta, l^x) \in \mathcal{D}_\theta \times \mathcal{D}_x$. Here, the parameter positions l^θ are independent of the observation positions l^x and can additionally vary between samples i . To adapt the parameter prior to function-valued parameters, we define a prior draw as an evaluation of an underlying measure μ (e.g., a Gaussian Process) at specific locations l^θ : $\theta \sim p_{l^\theta}$. The simulator then returns observations x at locations l^x following the likelihood $p_{l^x}(x | \theta, l^\theta)$. Many such simulations create a dataset $S = \{(l_i^\theta, \theta_i, l_i^x, x_i)\}_{i=1}^K$ for a number of simulations K . The explicit usage of the positions l^x and l^θ allows for flexible conditioning of the posterior.

3.1 Function-valued FMPE objective

To learn the velocity field v^ϕ , we adapt the FMPE objective function [9] for the function-valued setting. Given a discretized observation (x_o, l_o^x) , and a desired parameter discretization l^θ , we want to sample $\theta \sim p_{l^\theta}(\theta | x_o, l_o^x)$. This is done by first sampling from a base distribution $\xi_1 \sim p_{l^\theta}(\xi)$ and

then learning the velocity field $v_{l^\theta}^\phi(t, \xi_t, x_o, l_o^x)$ for ξ_t . In the following, we omit the arguments of $v_{l^\theta}^\phi$ for clarity. The learned velocity field allows us to iteratively denoise ξ_1 into a sample ξ_0 from the target posterior distribution. Note that the noise distribution is discretized on the same positions l^θ as the parameter θ . Similarly to Eq. 1, the velocity field $v_{l^\theta}^\phi$ is optimized via the loss function

$$\mathcal{L}_1 = \mathbb{E}_{t \sim \mathcal{U}[0,1], (l^\theta, \theta, l^x, x) \sim S, \xi_t \sim p_{l^\theta}(\xi_t | \theta)} \|v_{l^\theta}^\phi - u_t(\xi_t | \theta)\|^2. \quad (3)$$

Here, $p_{t, l^\theta}(\xi_t | \theta_t)$ describes a known noising process such that $\xi_1 | \theta$ is (approximately) drawn from the base distribution $p_{l^\theta}(\xi)$. Furthermore, $u_t(\xi_t | \theta)$ is the true vector field of the path defined by $p_{t, l^\theta}(\xi_t | \theta_t)$. We use the rectified flows formulation [18], such that $u_t(\xi_t | \theta_t) = (\xi_t - \theta)$.

The noise distribution, $\xi_1 \sim p_{l^\theta}(\xi)$, is commonly defined to be independent Gaussian white noise, $\xi \sim \mathcal{N}(0, I)$. Such distributions give rise to samples with a uniform power spectrum of ξ . As FNOs typically operate on the lower frequency modes of their inputs, independent Gaussian white noise would not be a suitable base distribution choice for our application. Instead, we sample noise from a Gaussian Process $\xi \sim \mathcal{GP}(0, k(\cdot, \cdot))$ [19, 20], where $k(\cdot, \cdot)$ is the square exponential kernel with lengthscale l . Using Bochner's theorem (Appendix S3.1) [21, 22], the spectral density of samples is

$$P(f) = (2\pi l^2)^{d_\theta/2} \exp(-2\pi^2 l^2 f^2),$$

where d_θ is the domain dimension of θ (and therefore of ξ). We choose l to be dependent on the highest Fourier mode M used by the FNO. This ensures that the majority of the signal power in the noise samples ξ_1 is conserved by the FNO block. We use the heuristic $l = \frac{2}{\pi(M/2+1)}$, which in expectation assures that $> 99\%$ of the spectral density of samples ξ is in the lower M frequency modes (derivation in Appendix S3.2). The covariance kernel k of the Gaussian Process is scaled to have unit marginal variance and defines the FMPE noise sampling during training via

$$p_{l^\theta}(\xi_t | \theta) = \mathcal{N}((1-t)\theta, t^2 k(l^\theta, l^\theta)).$$

3.2 Adapting to non-uniform, unseen discretizations of the domain

To operate on non-uniform discretizations, we adopt the work of Lingsch et al. [23] and use a FNO with a type II non-uniform fast Fourier transform (NUFFT, [24]). The NUFFT allows us to approximate the first M spectral modes of data or parameters discretized on any N points in the domain. We additionally add the positions l^θ and l^x to the input of the FNO blocks through multilayer perceptrons (MLPs, omitted in Fig. 2 for clarity) [25, 26].

In addition to non-uniform discretizations, FNOPE is also able to deal with distinct data and parameter discretizations at training and evaluation time. If we can query the simulator for arbitrary discretizations l^θ, l^x , we can generate the training data with mixed discretizations. However, this is not the case for all simulators. To mitigate errors when evaluating the posterior at non-uniform discretizations unseen during training, we perform data augmentation during training. First, we independently mask parts of the parameters and observations by randomly removing entries of θ_i and x_i and the corresponding positions l_i^θ and l_i^x . Second, we add small, independent Gaussian noise to the remaining positions (Appendix S3.3).

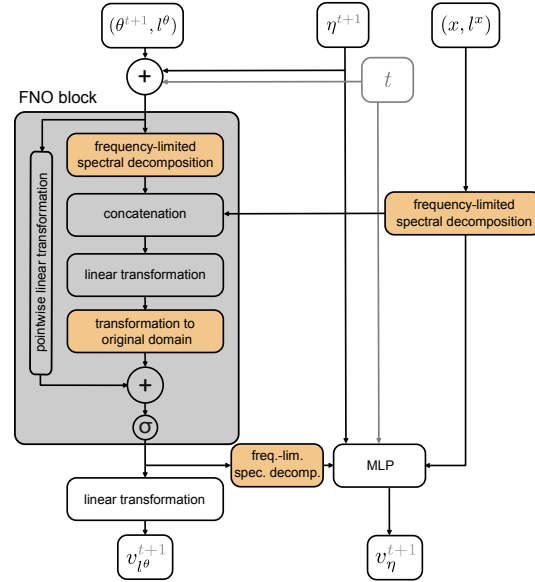


Figure 2: **FNOPE architecture.** FNOPE is based on several FNO blocks (gray): A FNO block receives the discretization-dependent spectral features of the function-valued parameters and observations as an input and processes them in a linear layer before transforming it back to the original domain via the approximate inverse transformation. A pointwise linear operation on the input is added, along with embeddings of the flow time, point positions, and vector-valued parameters. We expand this setup to several parallel channels and stack layers. The vector-valued velocities are separately estimated via a MLP.

In addition to this flexible implementation, we provide and evaluate **FNOPE (fix)**, a variant of FNOPE which uses the fast Fourier transform (FFT) in the FNO blocks and can be used for applications which exclusively consider parameters and observations discretized on uniform grids. The FFT computes spectral components in $O(N \log N)$ in the number of points in the discretization N , compared to $O(N \cdot M^D)$ for the NUDFT, where D is the domain dimension. Therefore, FNOPE scales favorably with the *parameter* dimensionality N (the number of points in the discretization), but FNOPE (fix) scales favorably with the *domain* dimensionality D (details in Appendix S3.4).

3.3 Inferring additional parameters

As most real world simulators have additional vector-valued parameters η , we extend the FNOPE architecture to also infer their posterior distribution. Vector-valued parameters are drawn from a known prior distribution $p(\eta)$ and the model likelihood becomes $p_{l^\theta}(x_i | \theta_i, l_i^\theta, \eta)$. The resulting inference problem is to estimate the posterior distribution $p_{l^\theta}(\theta, \eta | x_o, l_o^x)$. Hence, we now condition the velocity field $v_{l^\theta}^\phi$ additionally on the vector-valued parameters η by embedding them into the channel-dimension and subsequently adding them to the input of the FNO blocks (Fig. 2).

To estimate the velocity field of the vector-valued parameters v_η^ϕ (Fig. 2), we use a multilayer perceptron (MLP) to process the spectral features of the output of the final FNO block together with the vector-valued parameters and the spectral features of the observation. This approach results in a network which targets the combined velocity $v^\phi = [v_{l^\theta}^\phi, v_\eta^\phi]$. The combined network can be trained by an extension of the loss in Eq. 3,

where the noise $p(z_t | \eta)$ of the vector-valued parameters is given by a normal distribution $\mathcal{N}(z_t; (1-t)\eta, t^2\mathbf{I})$, and $u_t = [u_t(\xi_t | \theta), u_t(z_t | \eta)]$. The vector field $u_t(z_t | \eta)$ for the vector-valued parameters is analogously defined as $u_t(z_t | \eta) = (z_t - \eta)$. In practice, we separately normalize the loss for $v_{l^\theta}^\phi$ and v_η^ϕ by the number of parameters (Appendix S3.5).

4 Experiments

We apply FNOPE to four simulators: a Gaussian linear toy example, the SIRD model from epidemiology, the Darcy flow inverse problem and a real world application from glaciology (details in Appendix S5).

For the linear Gaussian simulator, we can analytically compute the posterior distribution, allowing us to compare the estimated posterior distributions to this ground truth using the Sliced-Wasserstein Distance (SWD) [27]. However, as is common in SBI applications, we do not have access to the ground truth posterior for the other simulators. Instead, we use a combination of two metrics to measure the quality of our posteriors: First, we report the predictive mean square error (Pred. MSE) between ground truth observations and predictive simulations from posteriors conditioned on those observations. We complement this metric with simulation-based calibration (SBC) [28] on the posterior marginal distributions. We quantify posterior calibration using the Error of Diagonal (EoD), measuring the average distance of the calibration curve of the estimated posterior from a perfectly calibrated posterior. Good performance on both of these metrics is not a sufficient condition to indicate a correctly estimated posterior, but healthy posteriors typically achieve good performance on these metrics. All evaluation metrics are averaged over three runs and we report mean $\pm$ standard error. We provide more details on all evaluation metrics in Appendix S2. We provide an overview of training and sampling times, as well as network sizes and computational resources used for all tasks, in Appendix S1.

4.1 Baseline methods

We compare FNOPE to three baseline methods: NPE (with normalizing flows) [3] and FMPE [9] on the coefficients of the spectral basis functions of the parameters (NPE/FMPE (spectral) respectively, details in Appendix S4). We also compare to FMPE with a fixed parameter discretization (FMPE (raw)).

For all baseline methods, we use the *sbi toolbox* [29]. For the SIRD simulator we compare to Simformer [30], a transformer-based amortized inference approach that is also capable of flexible

discretization of function-valued parameters. The other baselines cannot be applied in their basic version to this task because they do not support non-uniform discretizations of both parameters and observations.

4.2 Linear Gaussian

We first show the ability of FNOPE to approximate the true posterior of a linear Gaussian, as commonly done in SBI benchmarks [31]. To illustrate FNOPE's ability to infer a large number of parameters, we increase the dimensionality to 1000. We also replace the independent Gaussian prior in this task with a Gaussian process to model smoothly-varying function-valued parameters.

FNOPE clearly outperforms all benchmark methods on this problem (Fig. 3). With a training dataset of 10^2 simulations the SWD is close to zero for both FNOPE and FNOPE (fix). In contrast, both NPE and FMPE based on spectral features need as many as 10^5 simulations to achieve similar performance. Furthermore, this example shows that the data augmentation applied in training FNOPE, results in a small difference between FNOPE and FNOPE (fix), which is an effect of the introduced positional noise. FNOPE learns a posterior under a slightly broader likelihood than what is defined by the model and for very constrained posteriors the posterior quality is slightly poorer. However, we will see that for more challenging tasks, this is an acceptable trade off, as we gain flexibility on evaluation points.

We perform ablation experiments (Appendix S7.1), and observe that the performance of FNOPE is dependent on using sufficiently many Fourier modes in the FNO blocks (Fig. S1a,b). However, other hyperparameters show less influence on the performance (Fig. S1c-e).

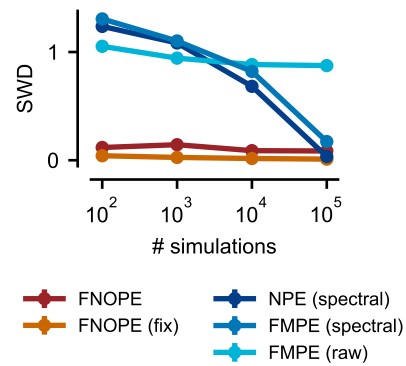


Figure 3: **Linear Gaussian simulator.** Sliced-Wasserstein distance (SWD) to ground truth posterior.

4.3 SIRD: Inference on unseen, non-uniform discretizations

Next, we consider the Susceptible-Infected-Recovered-Deceased (SIRD) model [32] to demonstrate the ability of FNOPE to solve inference problems on non-uniform discretizations of the parameters and observations that were not seen in the training data. In addition, we also show its ability to simultaneously infer vector-valued parameters. The model has three parameters: recovery rate, death rate, and contact rate [33, 34]. We use the same setup as in Gloeckler et al. [30], where we assume that the contact rate varies over time, but recovery and death rates are constant in time. We sample training simulations on a dense uniform grid for both parameters and observations. For evaluation we sample 100 observations, each discretized on a different set of 40 randomly sampled time points in $[0, 50]$, using contact rates defined on a distinct set of 40 randomly sampled times.

FNOPE, as well as Simformer, can reliably infer the posterior distribution (Fig. 4a) and the observations lie close to the mean of the posterior predictive (Fig. 4b). Both methods are comparable in terms of MSE of posterior predictive samples to the observations, as well as producing well-calibrated posteriors (Fig. 4c). When we use only 20 timepoints to condition on, the performance of FNOPE slightly decreases (Fig. S2). This highlights the necessity of the FNO block to have enough observation points to perform a reliable (approx.) Fourier transformation. We also observe that FNOPE performs robustly across the base distribution lengthscale (Fig. S3). This experiment shows that FNOPE is on par with the state of the art on this low dimensional problem: It successfully infers function-valued parameters together with vector-valued parameters and can be conditioned on arbitrary discretizations of the observations. However, FNOPE can also be applied to very high dimensional problems, as shown in the following experiment.

4.4 Darcy flow: Scalable inference in high dimensions

The Darcy flow is defined by a second order elliptic PDE and has been used to model many processes including the deformation of linearly elastic materials, or the electric potential in conductive materials.

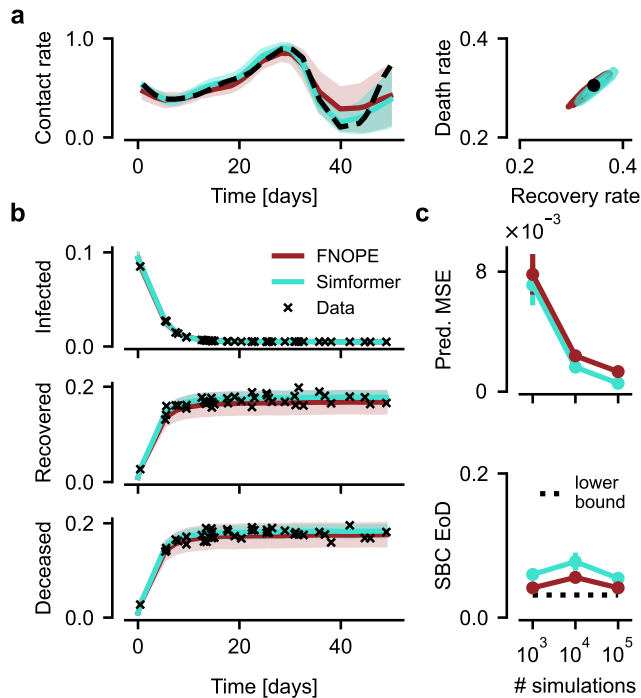


Figure 4: **SIRD model.** (a) Posterior conditioned on 40 time points. *left:* Posterior (mean $\pm$ std.) of the time-varying parameter and ground truth parameters (dashed). *right:* Two dimensional posterior of vector-valued parameters and ground truth parameters (dot). (b) Posterior predictive (mean $\pm$ std.) of infected, recovered and deceased populations with observations marked. (c) *upper:* MSE of posterior predictive samples to observations. *lower:* Simulation-based calibration error of diagonal (SBC EoD). ‘Lower bound’ refers to the SBC EoD for uniformly sampled posterior ranks (details in Appendix S2).

In the geosciences, the Darcy equation is used to describe the distribution of groundwater as a function of the spatially variable hydraulic permeability, which can be inferred from point observations in wells [35]. The Darcy flow is a common benchmark model for FNO applications, especially in the context of training PDE emulator models [13, 36, 37].

We consider the steady-state of the two dimensional Darcy flow equation on a unit square:

$$\begin{aligned} -\nabla \cdot (a(x) \nabla u(x)) &= 1 & x &\in (0, 1)^2 \\ u(x) &= 0 & x &\in \partial(0, 1)^2, \end{aligned}$$

where $a(x) \geq 0$ is the permeability we want to infer and $u(x)$ is the hydraulic potential. We adapt the implementation from [38], which provides a GPU-optimized solver. We use a log-normal prior distribution for the permeability, similar to Lim et al. [37]: $b = \log(a) \sim N(0, (-\Delta + \tau I)^{-2})$, where Δ is the Laplacian operator and $\tau = 9$. We sample the prior on a 129×129 grid which results in $\approx 16k$ parameters. For FNOPE, we use the first 32 Fourier modes in both spatial dimensions and for spectral NPE/FMPE we used the first 16 modes, resulting in $2 \cdot 16^2 = 512$ parameter dimensions. For all methods, we infer the log-permeability and evaluate in the original space (as in [37]).

Samples from the posterior inferred with FNOPE closely resemble the ground truth (Fig. 5a and Fig. S6). Both FNOPE and FNOPE (fix) correctly capture the fine-structure of the posterior samples and reproduce parameters at much higher fidelity than all baseline methods. While the spectral methods learn oversmoothed posteriors that do not capture local structures, the posterior samples from FMPE (raw) are much noisier and only capture the rough global structure. The posterior means show a similar trend (Fig. S7), and the standard deviations of the baseline methods are higher compared to FNOPE (Fig. S8).

The MSEs between posterior predictive samples and ground truth observations of FNOPE and FNOPE (fix) are consistently better compared to the spectral baseline methods, especially at lower simulation budgets, and are in the same range as FMPE (raw) (Fig. 5b). While all methods are reasonably well-calibrated (Fig. 5c), the visual appearance is vastly different. We additionally measure the posterior quality in terms of posterior log-probability (normalized by the number of pixels) of the associated ground truth parameter θ [31]. FNOPE has a much higher log probability (Fig. 5d) compared to FMPE (raw). FNOPE (fix) also achieves strong performance for a sufficient number of simulations. As the spectral methods do not model the parameters directly, we cannot calculate the log-probabilities they assign to the ground truth parameters. Overall, FNOPE is the only

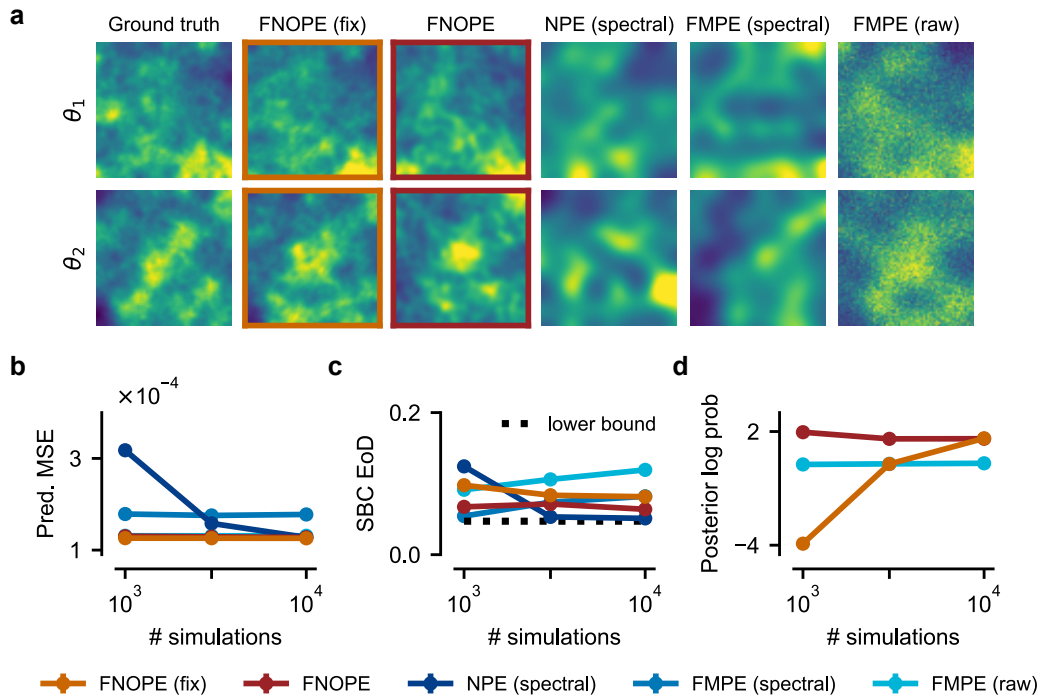


Figure 5: **Darcy flow.** (a) Ground truth parameter and posterior samples for a simulation budget of 10^4 training samples (more posterior samples in Fig. S6). (b) MSE of posterior predictives to the ground truth observation (FNOPE, FNOPE (fix) and FMPE (raw) visually overlap). (c) Simulation-based calibration Error of Diagonal (EoD) for 50 dimensions. (d) Posterior log-probability of ground truth samples normalized by the number of dimensions (higher is better).

method that consistently performs well on all presented metrics. In additional ablation experiments (Fig. S4), we show that FNOPE attains strong performance across different hyperparameter choices, and that our lengthscale heuristic (Sec. 3.1) is an appropriate choice for this task.

4.5 Mass balance rates of Antarctic ice shelves: Real world application

Finally, we turn to a real-world task from glaciology: Inference of snow accumulation and basal melt rates of Antarctic ice shelves from radar internal reflection horizons (IRHs) [39–41]. Snow continuously accumulates on top of the ice shelf. Over time, it is transported to larger depths where the former surfaces are further deformed by ice flow and form internal layers of constant age, which are measured by radar (Fig. 6a,b). The inference of accumulation and melt rates is a challenging SBI task, where the model is misspecified, as it cannot account for all real-world effects. We use an isochronal advection scheme forward model as described in [39]. In this work, the authors consider simulations on a grid of 500 points along a one-dimensional spatial domain and directly infer 50 parameters on a fixed downsampling of this domain using NPE. We refer to this approach as NPE (raw) and compare it to FNOPE and the baseline methods.

First, we evaluate all methods on a test set of simulations (Fig. 6c). As with the previous tasks, the performance of FNOPE at 10^3 simulations is comparable to the performance of the other methods at 10^5 simulations in terms of predictive MSE, and is only marginally worse at 10^2 simulations. All methods show a reasonable calibration in terms of SBC EoD at all simulation budgets (Fig. 6d). We then test the performance of all methods on real data (as in [39]). Posterior predictive samples from FNOPE match the observation very well (Fig. 6b), and while FNOPE still performs better at low simulation budgets than the other methods, the relative improvement compared to the baselines is smaller than the one observed on synthetic data (Fig. 6e). We note that this modeling problem was explicitly set up by Moss et al. [39] so that NPE (raw) can infer the posterior using a feasible number of simulations. FNOPE achieves the same performance with two orders of magnitude fewer

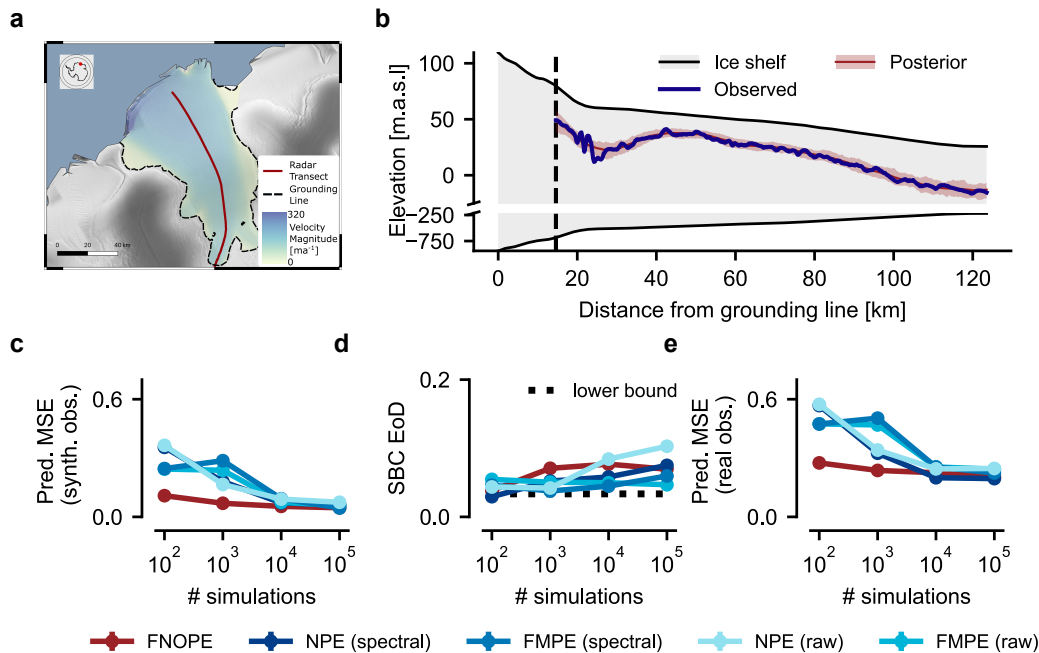


Figure 6: **Mass balance rates of ice shelves.** (a) Measurement transect of the radar data in Ekström Ice Shelf, Antarctica (adapted from [39], published under Creative Commons CC BY license). (b) Posterior predictive results obtained with FNOPE (trained on 10⁵ simulations), compared to radar-based observation on one layer in the ice shelf. (c)-(e) Performance measures on test simulations and the real observation, where NPE (raw) refers to the method used in [39].

simulations. Additionally, FNOPE is able to infer the full parameter dimensionality (500) instead of downsampling to 50 dimensions (Fig. S5).

5 Discussion

We present FNOPE, a simulation-based inference method using Fourier Neural Operators to efficiently infer function-valued parameters. On a variety of task, we showed that FNOPE can infer posteriors for function-valued data at very small simulation budgets compared to baseline methods, especially for high-dimensional problems. In addition, by building upon existing work for FNOs on non-uniform discretizations of the domain, FNOPE can generate samples from posteriors and observations defined on any discretizations of the domain, even if these discretizations were never seen during training.

Related work Scaling SBI methods to high-dimensional parameter spaces has been the focus of many works which make use of state of the art generative modeling techniques such as generative adversarial networks [7], diffusion models [8, 10, 11] and flow matching [9]. In particular, recent works [30, 42] use a transformer architecture to tokenize function-valued parameters, allowing for complete flexibility in estimating conditional distributions. However, as these methods explicitly model each point for discretized function-valued parameters, they are limited in terms of scalability. Our FNO-based approach allows us to compactly represent the parameters, significantly lowering the computational costs as the number of parameters grows. Finally, for applications where only the one-dimensional marginal distributions of the Bayesian posterior are needed, it is possible to scale SBI methods to higher dimensions, as the correlation between the parameters does not need to be captured [43]. However, in most scientific applications the correlation structure is an essential object of interest and marginal distributions are only of limited use.

Estimating function-valued parameters using FNOs [44–46] and other neural operators [47–49] has been explored in previous work. A majority of these approaches consider deterministic inversion, or estimating a single value for the parameters as opposed to targeting the Bayesian posterior. While

probabilistic generative models such as invertible Fourier Neural Operators (iFNOs, Long et al. [50]) estimate a conditional distribution, they do not explicitly target the Bayesian posterior. We compare the performance of iFNO on the Darcy task, and while it has comparable performance in terms of predictive MSE, the uncertainty estimates are not well-calibrated and the conditional distribution collapses to a tiny parameter region (Appendix S8).

The closest neighbours to our work are Lingsch et al. [51], who use a FNO architecture with an FMPE objective to learn vector-valued parameters and, additionally, learn an emulator producing function-valued observations. The crucial difference to our approach is that their simulators are deterministic and the inferred parameters are not function-valued. Recently, Lim et al. [37] developed an approach for score-based modeling in function spaces using FNOs. This enables high-dimensional posterior inference, but their approach is limited to uniform grids and does not allow for flexible conditioning.

Another related approach is diffusion posterior sampling [52–54], which seeks to learn a high-dimensional prior distribution from samples using score-based models. The learned priors can then be used to generate samples from the posterior using analytically tractable model likelihoods [55, 56]. Other works extended such approaches for intractable likelihoods [57, 58]. Similarly to diffusion posterior sampling, we only require prior samples. However, instead of learning the prior distribution, we learn the posterior distribution directly.

Limitations The FNO-backbone used by FNOPE inherently makes assumptions about the structure of the parameters. These assumptions enable computationally efficient inference, but result in some limitations. First, the FNO assumes limited high-frequency information in the parameter and observation domains. Therefore, they are ill-suited to infer parameters with high power in higher frequencies—for example, parameters with discontinuities. This could potentially be addressed by neural operators using other transforms, such as wavelet transforms [59]. Furthermore, to (accurately) compute the FFT or NUDFT of the observations, we require sufficiently many points in their discretization. Therefore, unlike other flexible methods [30, 42], our approach cannot perform inference using extremely sparse observations. Still, the SIRD experiment shows that even for 20 points we get reasonable estimations (Fig. S2). Finally, the computational complexity of our approaches still scales exponentially with the domain dimension, which could be challenging in high-dimensional domains.

Conclusion We presented FNOPE, a simulation-based inference approach for inferring function-valued data. FNOPE can be applied to non-uniform, unseen discretizations of the domain, can scale to large parameter dimensions, and can be trained using comparatively small simulations budgets. As we show in various experiments, FNOPE can therefore tackle spatiotemporal inference problems that were previously challenging or even intractable for simulation-based inference.

Acknowledgments and Disclosure of Funding

This work was funded by the German Research Foundation (DFG) under Germanys Excellence Strategy EXC number 2064/1 390727645 and SFB 1233 Robust Vision (276693517) and DFG (DR 822/3-1) and the Heinrich-Böll-Stiftung. Data collection was supported by Alfred Wegener Institute through logistic grants AWI_ANT_18. The authors acknowledge support by the state of Baden-Württemberg through bwHPC. GM is a member of the International Max Planck Research School for Intelligent Systems (IMPRS-IS). We thank Manuel Gloeckler for providing the training data and Simformer results for the SIRD experiment. We thank Daniel Gedon and Julius Vetter, and all members of Mackelab for discussions and feedback on the manuscript.

References

- [1] George Papamakarios and Iain Murray. Fast ϵ -free inference of simulation models with bayesian conditional density estimation. In *Advances in Neural Information Processing Systems*, volume 29, 2016.
- [2] Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural dynamics. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

- [3] David S. Greenberg, Marcel Nonnenmacher, and Jakob H. Macke. Automatic posterior transformation for likelihood-free inference. In *International Conference on Machine Learning*, 2019.
- [4] George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd international conference on artificial intelligence and statistics*. PMLR, 2019.
- [5] Conor Durkan, Iain Murray, and George Papamakarios. On contrastive learning for likelihood-free inference. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119. PMLR, 2020.
- [6] Stefan T. Radev, Ulf Kai Mertens, Andreas Voss, Lynton Ardizzone, and U. Köthe. Bayesflow: Learning complex stochastic models with invertible neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33, 2020.
- [7] Poornima Ramesh, Jan-Matthis Lueckmann, Jan Boelts, Álvaro Tejero-Cantero, David S. Greenberg, Pedro J. Goncalves, and Jakob H. Macke. GATSBI: Generative adversarial training for simulation-based inference. In *International Conference on Learning Representations*, 2022.
- [8] Tomas Geffner, George Papamakarios, and Andriy Mnih. Compositional score modeling for simulation-based inference. In *International Conference on Machine Learning*, 2022.
- [9] Jonas Wildberger, Maximilian Dax, Simon Buchholz, Stephen Green, Jakob H Macke, and Bernhard Schölkopf. Flow matching for scalable simulation-based inference. In *Advances in Neural Information Processing Systems*, volume 36, 2023.
- [10] Marvin Schmitt, Valentin Pratz, Ullrich Koethe, Paul-Christian Bürkner, and Stefan T. Radev. Consistency models for scalable and fast simulation-based inference. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [11] Julia Linhart, Gabriel Cardoso, Alexandre Gramfort, Sylvain Le Corff, and Pedro L. C. Rodrigues. Diffusion posterior sampling for simulation-based inference in tall data settings. *ArXiv*, 2024.
- [12] R. Hull, E. Leonarduzzi, L. De La Fuente, H. Viet Tran, A. Bennett, P. Melchior, R. M. Maxwell, and L. E. Condon. Simulation-based inference for parameter estimation of complex watershed simulators. *Hydrology and Earth System Sciences*, 28, 2024.
- [13] Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.
- [14] Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deepnet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3, 2019.
- [15] Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24, 2023.
- [16] Nikola B. Kovachki, Samuel Lanthaler, and Andrew M. Stuart. Operator learning: Algorithms and analysis. *arXiv*, 2024.
- [17] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.
- [18] Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.
- [19] Marin Biloš, Kashif Rasul, Anderson Schneider, Yuriy Nevmyvaka, and Stephan Günnemann. Modeling temporal data as continuous functions with stochastic process diffusion. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202. PMLR, 2023.
- [20] Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguët, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.
- [21] M.L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer New York, 2012. ISBN 9781461214946.

- [22] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.
- [23] Levi Lingsch, Mike Y. Michelis, Emmanuel De Bézenac, Sirani M. Perera, Robert K. Katzschmann, and Siddhartha Mishra. Beyond regular grids: Fourier-based neural operators on arbitrary domains. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.
- [24] Leslie Greengard and June-Yub Lee. Accelerating the nonuniform fast fourier transform. *SIAM Review*, 46, 2004.
- [25] Boris Bonev, Thorsten Kurth, Christian Hundt, Jaideep Pathak, Maximilian Baust, Karthik Kashinath, and Anima Anandkumar. Spherical fourier neural operators: Learning stable dynamics on the sphere. In *International Conference on Machine Learning*, 2023.
- [26] Zongyi Li, Nikola Borislavov Kovachki, Chris Choy, Boyi Li, Jean Kossaifi, Shourya Prakash Otta, Mohammad Amin Nabian, Maximilian Stadler, Christian Hundt, Kamyar Azizzadenesheli, and Anima Anandkumar. Geometry-informed neural operator for large-scale 3d PDEs. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [27] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51, 2015.
- [28] Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating bayesian inference algorithms with simulation-based calibration. *ArXiv*, 2018.
- [29] Jan Boelts, Michael Deistler, Manuel Gloeckler, Álvaro Tejero-Cantero, Jan-Matthis Lueckmann, Guy Moss, Peter Steinbach, Thomas Moreau, Fabio Muratore, Julia Linhart, Conor Durkan, Julius Vetter, Benjamin Kurt Miller, Maternus Herold, Abolfazl Ziaemehr, Matthijs Pals, Theo Gruner, Sebastian Bischoff, Nastya Krouglova, Richard Gao, Janne K. Lappalainen, Bálint Mucsányi, Felix Pei, Auguste Schulz, Zinovia Stefanidi, Pedro Rodrigues, Cornelius Schröder, Faried Abu Zaid, Jonas Beck, Jaivardhan Kapoor, David S. Greenberg, Pedro J. Gonçalves, and Jakob H. Macke. sbi reloaded: a toolkit for simulation-based inference workflows. *Journal of Open Source Software*, 10, 2025.
- [30] Manuel Gloeckler, Michael Deistler, Christian Weilbach, Frank Wood, and Jakob H. Macke. All-in-one simulation-based inference. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.
- [31] Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. Benchmarking simulation-based inference. In *International conference on artificial intelligence and statistics*, 2021.
- [32] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115, 1927.
- [33] Yi-Cheng Chen, Ping-En Lu, Cheng-Shang Chang, and Tzu-Hsuan Liu. A time-dependent sir model for covid-19 with undetectable infected persons. *Ieee transactions on network science and engineering*, 7, 2020.
- [34] Jonathan Schmidt, Nicholas Krämer, and Philipp Hennig. A probabilistic state space model for joint inference from differential equations and data. *Advances in neural information processing systems*, 34, 2021.
- [35] Wolfgang Nowak and Olaf A. Cirpka. Geostatistical inference of hydraulic conductivity and dispersivities from hydraulic heads and tracer data. *Water Resources Research*, 42, 2006.
- [36] Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *ACM/JMS Journal of Data Science*, 1, 2024.
- [37] Jae Hyun Lim, Nikola B Kovachki, Ricardo Baptista, Christopher Beckham, Kamyar Azizzadenesheli, Jean Kossaifi, Vikram Voleti, Jiaming Song, Karsten Kreis, Jan Kautz, et al. Score-based diffusion models in function space. *ArXiv*, 2025.
- [38] PhysicsNeMo Contributors. Nvidia physicsnemo: An open-source framework for physics-based deep learning in science and engineering. version: 1.2.0, 2023.
- [39] Guy Moss, Vjeran Vinjevi, Olaf Eisen, Falk M. Oraschewski, Cornelius Schröder, Jakob H. Macke, and Reinhard Drews. Simulation-based inference of surface accumulation and basal melt rates of an antarctic ice shelf from isochronal layers. *Journal of Glaciology*, 71, 2025.

- [40] Edwin D. Waddington, Thomas A. Neumann, Michelle R. Koutnik, Hans Peter Marshall, and David L. Morse. Inference of accumulation-rate patterns from deep layers in glaciers and ice sheets. *Journal of Glaciology*, 53, 2007.
- [41] M. J. Wolovick, J. C. Moore, and L. Zhao. Joint inversion for surface accumulation rate and geothermal heat flow from ice-penetrating radar observations at dome a, east antarctica. part i: Model description, data constraints, and inversion results. *Journal of Geophysical Research: Earth Surface*, 126, 2021.
- [42] Paul E. Chang, Nasrullo Loka, Daolang Huang, Ulpu Remes, Samuel Kaski, and Luigi Acerbi. Amortized probabilistic conditioning for optimization, simulation and inference. *ArXiv*, 2024.
- [43] Luca Ambrogioni, Umut Güçlü, Julia Berezutskaya, Eva van den Borne, Yağmur Güçlütürk, Max Hinne, Eric Maris, and Marcel van Gerven. Forward amortized inference for likelihood-free variational marginalization. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 777–786. PMLR, 16–18 Apr 2019.
- [44] Md Ashiqur Rahman, Manuel A Florez, Anima Anandkumar, Zachary E Ross, and Kamyar Azizzadenesheli. Generative adversarial neural operators. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.
- [45] Jacob H Seidman, Georgios Kissas, George J. Pappas, and Paris Perdikaris. Variational autoencoding neural operators. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202. PMLR, 2023.
- [46] Abdolmehdi Behroozi, Chaopeng Shen, and Daniel Kifer. Sensitivity-constrained fourier neural operators for forward and inverse problems in parametric differential equations. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [47] Zhongkai Hao, Zhengyi Wang, Hang Su, Chengyang Ying, Yinpeng Dong, Songming Liu, Ze Cheng, Jian Song, and Jun Zhu. GNOT: A general neural operator transformer for operator learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202. PMLR, 2023.
- [48] Tian Wang and Chuang Wang. Latent neural operator for solving forward and inverse pde problems. In *Advances in Neural Information Processing Systems*, volume 37, 2024.
- [49] Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A fast transformer solver for pdes on general geometries. In *International Conference on Machine Learning*, 2024.
- [50] Da Long, Zhitong Xu, Qiwei Yuan, Yin Yang, and Shandian Zhe. Invertible fourier neural operators for tackling both forward and inverse problems. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258. PMLR, 2025.
- [51] Levi E. Lingsch, Dana Grund, Siddhartha Mishra, and Georgios Kissas. FUSE: Fast unified simulation and estimation for PDEs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [52] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.
- [53] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023.
- [54] Hyungjin Chung, Jeongsol Kim, Michael Thompson McCann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023.
- [55] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *The Eleventh International Conference on Learning Representations*, 2023.
- [56] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022.
- [57] Gabriel Cardoso, Yazid Janati el idrissi, Sylvain Le Corff, and Eric Moulines. Monte carlo guided denoising diffusion models for bayesian linear inverse problems. In *The Twelfth International Conference on Learning Representations*, 2024.

- [58] Zihui Wu, Yu Sun, Yifan Chen, Bingliang Zhang, Yisong Yue, and Katherine Bouman. Principled probabilistic imaging using diffusion models as plug-and-play priors. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [59] Tapas Tripura and Souvik Chakraborty. Wavelet neural operator for solving parametric partial differential equations in computational mechanics problems. *Computer Methods in Applied Mechanics and Engineering*, 404, 2023.

Supplementary Material

S1 Software and Computational Resources

For all baseline SBI methods, we use the `sbi` toolbox [29], for the Simformer baseline we use the publicly available code from Gloeckler et al. [30]. We use an optimized solver to solve the Darcy Flow PDE [38].

We use various compute resources for the different experiments. For each experiment, we run the training and evaluation for each method, for each simulation budget, and for each of the three random seeds separately. The summary of network sizes and compute times for all tasks are provided in Tab. S1,S2,S3.

For the Linear Gaussian and SIRD experiments, we perform our experiments on Nvidia RTX 2080ti GPU nodes. Both simulators have negligible wall-clock costs on these GPU nodes. The Darcy flow experiment required GPUs with higher VRAM to accommodate the large ($\approx 16k$ dimensional) parameters and observations. We performed these experiments on Nvidia A100 GPUs. For the Antarctic Ice experiment, we perform training and evaluation on CPU, namely Intel Xeon Gold 16 cores, 2.9GHz. We perform this experiment on CPU as the main computation cost is in running the simulations for the predictive MSE check, and the implementation of the model is not accelerated by use of GPUs. The simulation costs are described in Moss et al. [39].

Table S1: **Network sizes and training times for Gaussian Linear (GL) and Darcy Flow (Darcy) tasks.** We report the mean over 3 runs with a training budget of 10k simulations. Baselines estimating the "raw" (untransformed) parameters are denoted (R) and those estimating Fourier coefficients are denoted (S). Linear Gaussian task was run on a 2080ti GPU, Darcy flow on an A100 GPU.

Task	Metric	FNOPE	FNOPE (fix)	NPE (S)	FMPE (S)	FMPE (R)
GL	# params.	110K	109K	567K	86.3K	299K
GL	train (Tot.) [m]	3.12	1.98	6.77	2.46	5.14
GL	train (/epoch) [s]	0.99	0.87	2.17	0.24	0.31
GL	sample (/sample) [ms]	2.16	1.06	0.05	0.24	0.20
Darcy	# params.	11.6M	11.6M	3.54M	898K	9.18M
Darcy	train (Tot.) [m]	72.2	41.8	20.5	10.9	12.2
Darcy	train (/epoch) [s]	20.3	26.2	2.82	0.66	0.73
Darcy	sample (/sample) [ms]	280	35.2	0.21	1.95	2.70

Table S2: **Network sizes and training times for SIRD.** We report the mean over 3 runs with a training budget of 10k simulations. Both methods were trained and evaluated on a 2080ti GPU.

Task	Metric	FNOPE	Simformer
SIRD	# params.	117K	286K
SIRD	train (Tot.) [m]	6.95	47.17
SIRD	train (/epoch) [s]	1.19	9.43
SIRD	sample (/sample) [ms]	0.22	0.22

Table S3: **Network sizes and training times for Antarctic Ice (Ice) task.** We report the mean over 3 runs with a training budget of 10k simulations. Baselines estimating the "raw" (untransformed) parameters are denoted (R) and those estimating Fourier coefficients are denoted (S). All methods were trained and evaluated on CPU.

Task	Metric	FNOPE	NPE (S)	FMPE (S)	NPE (R)	FMPE (R)
Ice	# params.	25.3K	236K	92.3K	361K	96.2K
Ice	train (Tot.) [m]	47.4	19.7	31.6	6.80	33.5
Ice	train (/epoch) [s]	9.82	4.21	1.92	5.64	2.02
Ice	sample (/sample) [ms]	22.8	1.50	16.6	1.53	23.9

S2 Evaluation details

We here describe more details about our evaluation procedures. We evaluate on a heldout test set $\{(\theta_j^o, l_j^\theta, \eta_j^o, x_j^o, l_j^x)\}_{j=1}^{J_{\text{test}}}$, where J_{test} is the number of test simulations. Given an approximate posterior distribution $q_{l^\theta}^\phi(\theta, \eta \mid x, l^x)$ and a test observation (x_j^o, l_j^x) , we draw K_{post} posterior samples $(\theta_{kj}, \eta_{kj}) \sim q_{l^\theta}^\phi(\theta, \eta \mid x_j^o, l_j^x)$. In the case where no vector-valued parameters η are present, they can be omitted. Similarly, for methods which do not explicitly use the positions l^θ, l^x (e.g. FNOPE (fix)), the positions can be omitted, as we do not apply these methods to tasks where we consider arbitrary discretizations.

We report the average and standard error over all J_{test} test simulations.

S2.1 Sliced Wasserstein Distance

Following Bonneel et al. [27], we define the (empirical) sliced Wasserstein(-2) distance (SWD) between N samples from two probability distributions p and q as

$$\text{SWD}(p, q) = \mathbb{E}_{u \sim U(\mathbb{S}^{D-1})} \left[\left(\frac{1}{K} \sum_{i=1}^K \|x_u^{(k)} - y_u^{(k)}\|_2^2 \right)^{1/2} \right], \quad (4)$$

where $x_k \sim p(x), y_k \sim q(y)$ are samples from the two distributions, u are uniformly randomly sampled vectors on the unit sphere $\mathbb{S}^{D-1}$, and $x_u^{(k)}, y_u^{(k)}$ are the 1-dimensional i -th order statistics of the projections $u^\top x_k, u^\top y_k$ respectively. We calculate the SWD with 50 random projections u and $K = 1000$ posterior samples.

S2.2 Simulation-based Calibration Error of Diagonal

Simulation-based calibration (SBC) [28] is a standard measure of the calibration of approximate posterior distributions (in terms of over- or underconfidence). We obtain ranks r_{ij} for each sample (θ_j^o, x_j^o) in the test set using SBC with the 1-dimensional marginal distributions used as the reducing functions. That is, for each of the dimensions i of θ , the rank r_{ij} is an integer in $(1, K_{\text{post}} + 1)$. This results in J_{test} ranks. The cumulative distribution function of ranks is therefore

$$\text{CDF}_i(\alpha) = \frac{1}{J_{\text{test}}} \sum_j \mathbb{I}[r_{ij}/K_{\text{post}} < \alpha].$$

The SBC Error of Diagonal (SBC EoD) is then the mean absolute distance between this cumulative distribution and the cumulative distribution function of a uniform distribution,

$$\text{SBC EoD}(i) = \int_0^1 |\text{CDF}_i(\alpha) - \alpha| d\alpha.$$

In contrast to the SBC area under the curve (SBC AUC), the EoD will detect poor calibrations for posteriors that are overconfident at low confidence levels α , and underconfident at high α (or vice-versa). Finally, we report the average SBC EoD across the dimensions of θ ,

$$\text{SBC EoD} = \frac{1}{N_\theta} \text{SBC EoD}(i).$$

For SIRD, since the posterior dimensionalities are low, we compute an average SBC EoD over all one-dimensional marginals of the posterior. For the Darcy Flow and Antarctic Ice tasks, we select a subset of 50 marginal distribution for computing the SBC EoD, regularly spread across the domain.

S2.3 Predictive MSE

To calculate the MSE for posterior predictive samples, we run for each posterior sample (θ_{kj}, η_{kj}) , and each true observation x_j^o , the simulator $x_{kj} \sim p_{l_j^x}(x \mid \theta_{kj}, l_j^\theta, \eta_{kj})$. We then compute the average mean square error of the simulation x_{kj} to the corresponding observations x_j^o ,

$$\text{MSE} = \frac{1}{J_{\text{test}} K_{\text{post}}} \sum_{k=1, j=1}^{J_{\text{test}}, K_{\text{post}}} \frac{1}{|l_j^x|} \|x_{kj} - x_j^o\|_{L^2}^2,$$

where $|l_j^x|$ is the number of points in the discretization l_j^x and therefore the dimensionality of x_j^o . We use this metric since the simulators considered in this work correspond to (unknown) unimodal likelihood functions—a correctly estimated posterior will produce simulations clustered around the true observation. We opt for this metric to quantify predictive performance due its clear interpretability. However, for multimodal likelihood functions, this metric can be replaced with a scoring rule.

S2.4 Posterior Log Probability

For the Darcy Flow task, we additionally report the posterior log-probability of the true parameters [31], normalized by the number of pixels:

$$\text{log-probability per pixel} = \frac{1}{|l_j^\theta|} \log q_{l_j^\theta}^\phi(\theta_j^o | x_j^o, l_j^x).$$

For the spectral methods NPE/FMPE (spectral), we cannot directly compute the posterior-log-probabilities, as we can only compute the posterior-log-probabilities of the first M modes of the spectral decomposition of the ground truth parameters θ_j^o . However, by discarding the information of the higher modes, we remove the information which these baseline methods cannot capture, thus biasing the resulting log-probabilities in favor of these baselines.

S3 FNOPE details

S3.1 Bochner’s Theorem

We state Bochner’s theorem following Williams and Rasmussen [22]. A complex-valued function k on $\mathbb{R}^d$ is the covariance function of a weakly stationary mean continuous complex-valued random process on $\mathbb{R}^D$ if and only if it can be represented as

$$k(\tau) = \int_{\mathbb{R}^D} \exp^{2\pi i f \cdot \tau} d\mu(s) \quad (5)$$

for some positive finite measure μ . Crucially, in the less general but relevant case that μ admits a density $P(f)$, the integral is a Fourier transform between the kernel $k(\tau)$ and the spectral density $P(f)$. We apply this result to relate the lengthscale of the square exponential kernel to the spectral density of its Fourier decomposition in Sec. 3.1.

S3.2 Kernel lengthscale heuristic

The spectral density of samples from a Gaussian Process with a square exponential kernel of lengthscale l is stated in Sec. 3.1 as

$$P(f) = (2\pi l^2)^{d_\theta/2} \exp(-2\pi^2 l^2 f^2).$$

This is also a Gaussian density, and trivially we see that the full spectral power is

$$\bar{P} = \int_{\mathbb{R}^{d_\theta}} P(f) df = 1$$

We consider discretizations normalized to $[0, 1]$ in each dimension, and so the power contained in the first M spectral modes, $\bar{P}_M$, corresponds to the above integral within the domain $\|f\|_\infty \leq M/2$, i.e. where all the components of f are within $[-M/2, M/2]$. Therefore $\bar{P}_M$ simplifies to the product of the Gaussian integrals

$$\begin{aligned} \bar{P}_M &= \left(\int_{-M/2}^{M/2} (2\pi l^2)^{1/2} \exp(-2\pi^2 l^2 f_i^2) df_i \right)^{d_\theta} \\ &= \left(\text{erf} \left[\frac{\pi l M}{\sqrt{2}} \right] \right)^{d_\theta} \\ &= \left(\text{erf} \left[\frac{M\sqrt{2}}{M/2 + 1} \right] \right)^{d_\theta}, \end{aligned}$$

where erf is the Gauss error function, and in the last line we substituted our heuristic $l = \frac{2}{\pi(M/2+1)}$. This value saturates the error function and produces values very close to 1. For example, for the Darcy Flow example ($d_\theta = 2$), we set $M = 32$. The resulting spectral power in the first 32 modes is $\bar{P}_{32} \approx 0.9997$. While individual samples from the Gaussian process can result in discrete Fourier transforms where this spectral property is not fulfilled, it is clear that the majority of the spectral power will be contained in the first M modes for all samples.

S3.3 Flexible discretization

We provide further detail of the data augmentation scheme introduced in Sec. 3.2. First, we describe why this is necessary despite the use of the non-uniform fast fourier transform (NUDFT). The NUDFT is applied as a matrix multiplication, $\Theta = V(l^\theta)\theta$, where Θ is a vector containing the first M spectral components of θ , and V is the discretization-dependent transformation matrix. The inverse NUDFT is similarly implemented as a matrix multiplication $\theta = \bar{V}^\top(l^\theta)\Theta$, where $\bar{V}^\top$ is the conjugate transpose of V . This approach enables the computational efficiency of the NUDFT, as the exact inverse matrix does not need to be computed at runtime. However, $\bar{V}^\top$ is only an approximate inverse of V , with the approximation error increasing for increasing non-uniformity of the discretization.

Consider the common case where the simulation dataset S (Sec. 3) provides parameters θ_i and observations x_i always discretized on the same, uniform simulation domain. Without data augmentation, we always apply the NUDFT and its inverse without approximation error. However, if we wish to condition a posterior on x^o measured at some non-uniform discretization l^x , then the NUDFT and its inverse will produce some error, which was unseen during training. This could lead to unpredictable, out of distribution errors at evaluation time. By explicitly passing the positions l^θ and l^x to the network, as well as augmenting them to ensure the network is not always applied on uniform discretizations during training, we give FNOPE capacity to learn to counteract these approximation errors.

Masking We define a uniform distribution over the binary mask vectors with a fixed number of nonzero entries, N_{ds} . Suppose we are given a simulation $(\theta, l^\theta, x, l^x)$, where θ, l^θ consist of N_θ points, and x, l^x consist of N_x points. We construct two random binary masks, $\mathbf{M}^\theta \in \{0, 1\}^{N_\theta}$, $\mathbf{M}^x \in \{0, 1\}^{N_x}$, each with exactly N_{ds} nonzero entries. We then remove the corresponding elements of θ, l^θ where $\mathbf{M}_\theta$ is zero, and similarly remove the corresponding elements of x, l^x where $\mathbf{M}_x$ is zero. For a minibatch of simulations, we independently sample the masks $\mathbf{M}_i^\theta, \mathbf{M}_i^x$ for each simulation. If $N_{\text{ds}} > N_\theta$ or $N_{\text{ds}} > N_x$, we leave the corresponding value and position vector unchanged. The value of N_{ds} used in our work is reported for each experiment in Appendix S6.

Positional noise We additionally add small, independent gaussian noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ to each point in l^θ and l^x in the unmasked positions. This reduces generalization error for simulation datasets where the discretization of θ and x is fixed. The value of σ^2 can be set according to the spacing of the discretization. In our experiments, we always normalize the simulation domain to $[0, 1]$ in all dimensions, and set $\sigma = 10^{-3}$.

S3.4 Fixed discretization

For applications with uniform grids, we provide FNOPE (fix). Here, we use the FFT instead of the NUDFT in the FNO blocks to transform the data from physical to spectral space and back. In addition, we do not mask any parameters during training and do not add any positional noise. We expect this method to have improved performance on uniform grids as we do not introduce additional noise through the data augmentation process described above.

Another potential advantage of FNOPE (fix) is its computational efficiency. Consider the case of a parameter discretized uniformly in a D -dimensional domain, with L points per dimension, leading to a total of L^D points. The computational cost of computing the spectral decomposition of the parameters using the FFT is $O(L^D \log L^D) = O(DL^D \log L)$. Assuming the maximum number of modes in each dimension modeled by the FNO is M , this leads to M^D total modes to compute. Therefore, the computational cost of the frequency-limited NUDFT is $O(M^D L^D)$. For one-dimensional domains, the NUDFT may well be faster to compute than the FFT. However, for

higher dimensions, the NUDFT scales exponentially with both M and L . This discrepancy increases with both the dimensionality of the domain, and the number of modes M modeled by the FNO blocks.

S3.5 Additional Parameters

The naïve extension of the FMPE objective as stated in Sec. 3.3 is to minimize the L^2 loss $\|v^\phi - u_t\|^2$, where $v^\phi = [v_{l^\theta}^\phi, v_\eta^\phi]$ and $u_t = [u_t(\xi_t|\theta), u_t(z_t|\eta)]$. However, the scale of the loss for the continuous parameters, $\|v_{l^\theta}^\phi - u_t(\xi_t|\theta)\|^2$ varies with the number of points in the discretization l^θ , which we denote N_θ . To ensure that the loss is balanced for the function- and vector-valued parameters, we in practice add their L^2 losses, normalized by their respective vector dimensionalities. That is, we minimize

$$\mathbb{E} \left[\frac{1}{N_\theta} \|v_{l^\theta}^\phi - u_t(\xi_t|\theta)\|^2 + \frac{1}{N_\eta} \|v_\eta^\phi - u_t(z_t|\eta)\|^2 \right],$$

where N_η is the fixed dimensionality of η . The expectation is over the same random variables as for the statement of the loss $\mathcal{L}_2$ in Sec. 3.3.

S4 Baseline Methods

S4.1 Spectral preprocessing

For spectral NPE/FMPE we first apply a Fourier transformation to the parameters, take the first M Fourier modes and expand these M complex values to a real vector of dimension $2M - 1$ representing the real and imaginary parts (the imaginary part of the first component is always 0, hence it is discarded). For two dimensional data, the same preprocessing results in $2M^2 - 1$ parameters. After inferring the posterior of the parameters in Fourier space and sampling from it, we apply the inverse Fourier transform to get samples in the spatial domain.

For the one dimensional problems (Linear Gaussian and Ice Shelf) we first pad the data by replicating the first/last value and perform a real FFT with `torch.fft.rfft`. We then use the first M Fourier components and expand these complex numbers to a real tensor of dimension $2M$, which we use as input to NPE/FMPE. For the Linear Gaussian and Ice Shelf we use $M = 50, 10$, respectively. We then revert this process for samples from the posterior with `torch.fft.irfft` with the corresponding settings.

For the two dimensional Darcy flow, we use the two dimensional FFT implemented in *pytorch* on the padded data (in mode replicate). We then center the frequencies before cropping to the first M Fourier components in both dimensions, and expanding it to a real tensor of dimension $2M^2$. For posterior samples we again revert this process with the corresponding settings.

S4.2 NPE (spectral)

For NPE (spectral) we infer the posterior over the coefficients of the first M Fourier modes following the spectral preprocessing described above. We use NPE [3] with normalizing flows. We do not apply spectral preprocessing to the observations but pass raw observations through an embedding net as it is common practice in NPE.

S4.3 NPE (raw)

For the mass balance experiment (Sec. 4.5), we also compare to the approach of Moss et al. [39], which we refer to as NPE (raw). This approach infers the mass balance parameters on a fixed discretization of 50 gridpoints. The authors use a Neural Spline Flow with 5 transformations, two residual blocks of 50 hidden units each, ReLU nonlinearities and 10 bins. The embedding net used to embed the 441-dimensional observation is a CNN with two convolutional layers of kernel size 5, with ReLU activations and max pooling of kernel size 2. The convolutional layers are followed by two linear layers with 50 hidden units and output dimension 50. The same settings are used in the 500-dimensional experiment (Appendix S7.4). Training is performed with a batch size of 200 and an Adam optimizer with learning rate of 0.0005.

S4.4 FMPE (spectral)

As with NPE (spectral), in this approach we apply spectral preprocessing to the parameters, and infer the coefficients of the top M Fourier modes, but process the observations directly using an embedding net. We use MLPs to estimate the flows, as in Wildberger et al. [9], as implemented in the `sbi` toolbox [29]. This implementation also uses the rectified flow [18] objective for FMPE, and we use independent Gaussian noise as the noise distribution. The flow networks are conditioned on time by concatenating the time to the inputs.

S4.5 FMPE (raw)

In this approach, we infer the parameters directly on a fixed discretization of the domain. We again use embedding nets to encode the observations, and MLPs to learn the flows. As with FMPE (spectral), we use a rectified flow objective, and independent Gaussian noise as the noise distribution, as in the `sbi` toolbox. The flow networks are conditioned on time by concatenating the time to the inputs.

S4.6 Simformer

For the SIRD experiment, we apply Simformer with the same settings as in Gloeckler et al. [30]. That is, we use a transformer model with a token dimension of 50, 8 layers, and 4 heads. The widening factor is 3, and the training was performed with a batch size of 1000 and an Adam optimizer. We train Simformer to learn all conditionals, and so uniformly draw between the posterior, joint, and likelihood masks, as well as two random masks drawn from Bernoulli distributions with $p = 0.3$ and $p = 0.7$ respectively. For both SIRD experiments, where we evaluate on 20 and 40 time points respectively, we use the same Simformer model which is trained using 20 randomly sampled time points.

S5 Simulators

S5.1 Linear Gaussian model

The Gaussian Simulator is inspired by Lueckmann et al. [31], but instead of a 10 dimensional Gaussian distribution with independent dimensions we expanded the problem to 1000 dimensions and use a Gaussian Process prior (see below). Draws from the simulator are still drawn independently per dimension as $x \sim \mathcal{N}(\theta, \sigma^2 \mathbf{I})$, where $\sigma^2 = 0.1$ (as in [31]).

Prior The prior is defined as a Gaussian process $\mathcal{GP}$ on $[0, 1]$ with an equidistant discretization with 1000 timepoints. A draw from the prior is therefore defined as $\theta \sim \mathcal{GP}(0, k(\cdot, \cdot))$, where k is the squared exponential kernel, $k(t, t') = \exp(-\frac{(t-t')^2}{2l^2})$. We set the lengthscale $l = 0.05$ and variance to 1. Only in the ablation experiments (Fig. S1a,b) we changed the lengthscale l to 0.005.

Evaluation parameters The results for Fig. 3 is based on 100 observations and 1000 posterior samples for each observation.

S5.2 SIRD model

Similar to Gloeckler et al. [30] we extend the SIRD (Susceptible, Infected, Recovered, Deceased) model to have a time-dependent contact rate. Compared to the classical SIR framework the model additionally incorporates a deceased (D) population. Similar models were explored by Chen et al. [33], Schmidt et al. [34]. This addition is important for modeling diseases with significant mortality rates. The SIRD model, including a time-dependent contact rate $\beta(t)$, is defined by the following set

of differential equations:

$$\begin{aligned}\frac{dS}{dt} &= -\beta(t)SI, \\ \frac{dI}{dt} &= \beta(t)SI - \gamma I - \mu I, \\ \frac{dR}{dt} &= \gamma I, \\ \frac{dD}{dt} &= \mu I.\end{aligned}$$

Here, S , I , R and D are the susceptible, infected, recovered, and deceased population, $\beta(t)$ is the time dependent contact rate, and γ and μ are the recovery and mortality rates among the infected population. We simulate on a dense uniform grid of 100 time points for parameters and observations. The simulations are additionally contaminated by an observation noise model, which is described by a log-normal distribution with mean $S(t)$ and standard deviation $\sigma = 0.05$.

Prior We impose the same prior as in Gloeckler et al. [30]: the global variables γ and μ are drawn from a uniform distribution, $\gamma, \mu \sim \text{Unif}(0, 0.5)$. For the time-dependent contact rate we define a Gaussian process prior which is further transformed by a sigmoid function to ensure that $\beta(t) \in [0, 1]$ for all t . For the Gaussian process we use a RBF kernel k defined as $k(t, t') = \exp(-\frac{\|t-t'\|^2}{2 \cdot 7^2})$.

Evaluation parameters MSE as well as SBC EoD is based on 100 observations with 1000 posterior samples each (Fig. 4c and Fig. S2c). To calculate the posterior predictive, we sample the initial condition $I(0)$ from the Simformer prediction for both methods, as opposed to the prior defined above, to match the setting of Gloeckler et al. [30].

S5.3 Darcy Flow

Details for the Darcy model are already given in the main text. We additionally scale the log-permeability a by the scale factor of 1000 before taking the exponential - this results in permeabilities which produce sufficiently variable solutions using the Darcy flow simulator, and the permeabilities on the same scale as reported in Lim et al. [37]. The simulation output is additionally corrupted by an independent Gaussian observational noise per pixel $\zeta_i \sim \mathcal{N}(0, \sigma_i^2)$. We set $\sigma_i = \mathbb{E}[u_i^2/30]$, where the expectation is per simulation batch, resulting in a signal to noise ratio (SNR) of 30.

Evaluation Parameters All metrics are calculated over a test set of 10 observations (Fig. 5 b-d). For MSE and SBC EoD, 100 posterior samples were used for each observation and for each method. Finally, for SBC EoD, we use a subset of 50 pixels of the full 129×129 as the marginals used for the reducing functions. The same pixels were used across all methods and all observations.

S5.4 Mass balance rates of Antarctic Ice Shelves

We use the same simulator as described in Moss et al. [39]. This model takes in spatially varying surface accumulation rates $\dot{a}(l)$, which are related to the basal melt rates $\dot{b}(l)$ through the total balance condition $\dot{m}(l) = \dot{a}(l) - \dot{b}(l)$. $\dot{m}(l)$ is known and fixed across all simulations.

The layer prediction model consists of a set of isochronal layer with prescribed thicknesses $\{h_1(l), h_2(l), \dots, h_K(l)\}$ such that $\sum_{k=1}^K h_k(l) = h(l)$, where $h(l)$ is the known and fixed total ice shelf thickness. At each time step, the thickness of the layers are simultaneously updated through an advection equation,

$$\frac{\partial h_k}{\partial t} = -\nabla \cdot (h_k u),$$

where $u(l)$ is the known velocity profile of the ice shelf which is fixed across simulations. Additional layers are added at the top of the ice shelf and removed from the bottom according to $\dot{a}(l)$, $\dot{b}(l)$ accordingly. The noise model approximates the observation noise of the radar measurement given an assumed density profile of the ice shelf. At the final timestep T , the layer that most closely matches

the ground truth observation x^o according to the L^2 -norm is selected as the simulator output. Full details are described in Moss et al. [39].

Prior The prior is defined over the accumulation rate parameter $\dot{a}(l)$, and is motivated by physical observations at Ekström ice shelf. Prior samples are drawn using

$$\dot{a}(l) = \sigma_{\text{sc}} \dot{\alpha} + \mu_{\text{off}}, \quad (6)$$

where $\mu_{\text{off}} \sim \mathcal{N}(0.5, 0.25^2)$, $\sigma_{\text{sc}} \sim \mathcal{U}([0.1, 0.3])$, and $\dot{\alpha}$ is drawn from a Gaussian Process with a unit-variance, zero-mean Matérn- ν kernel of lengthscale 2500 and $\nu = 2.5$.

Evaluation Parameters For SBC EoD, as well as the predictive MSE on synthetic test simulations, we use 100 test observations and sample 10 posterior samples for each observations and for each method. The real test data consists of one field observation (shown in Fig. 6a-b), and the posterior predictive was estimated using 1000 posterior samples.

S6 Experimental details

S6.1 Linear Gaussian

For all the baseline methods, we train the networks using an Adam optimizer with a learning rate of 0.0001, and a batch size of 200. For NPE/FMPE (spectral), we use 50 modes, leading to 100 parameters to learn, and a pad width of 20 for the spectral preprocessing (Appendix S4.1). For NPE (spectral) the density estimator is a Neural Spline Flow (NSF) with 2 residual blocks with 50 hidden dimensions each, 5 transforms, with RELU activations. For FMPE (spectral), we use an MLP with 5 linear layers with 64 hidden dimensions to estimate the flow, with ELU activations. In both cases, we embed the 1000 dimensional observation into a 40-dimensional vector using an MLP with 2 layers and 50 hidden units and RELU activations. For FMPE (raw), we use an MLP with 5 layers and 64 hidden features, with ELU activations.

For FNOPE and FNOPE (fix) we use 50 Fourier modes for the FNO blocks. We use 5 FNO blocks with 16 channels, while the context is embedded into 8 channels. We train for a maximum of 500 epochs with an early patience of 50. We used a training batch size of 512 and a learning rate of 0.001. For FNOPE, we use 4 channels each for the positional and time embeddings and the target gridsize $N_{\text{ds}} = 256$ (for FNOPE (fix), no positional embedding is included). All nonlinearities are GELUs.

S6.2 SIRD

For FNOPE we use 32 Fourier modes for the FNO blocks. We use 5 FNO blocks with 16 channels, while the context is embedded into 8 channels. We train for a maximum of 1000 epochs with an early patience of 50. We use a training batch size of 200 and a learning rate of 0.001. The discretization positions and flow times are embedded into 4 channel dimensions each, and the target gridsize $N_{\text{ds}} = 40$. This experiment additionally included vector-valued parameters in $\mathbb{R}^2$. These are embedded into a 16-dimensional vector using a 1-layer MLP with a hidden dimension of 64. The flow for the vector-valued parameters is estimated using an 1-layer MLP with a hidden dimension of 64. The spectral decomposition of the output of the FNO blocks, as well as the spectral decomposition of the observation, are embedded into a 32-dimensional vector and concatenated to the input of the MLP. All nonlinearities were GELUs.

The training hyperparameters for simformer are described in S4.6.

S6.3 Darcy Flow

For all the baseline methods, we train the network using an Adam optimizer with a learning rate of 0.0001, and a batch size of 200. For NPE/FMPE (spectral), we use 16 modes, leading to $2 \times 16^2 = 512$ parameters to learn, and a pad width of 20 in each dimension for the spectral preprocessing (Appendix S4.1). For NPE (spectral) the density estimator is a NSF with 2 residual blocks with 50 hidden dimensions each, 5 transforms, with RELU activations. For FMPE (spectral), we use an MLP with 8 layers with 256 hidden dimensions to estimate the flow, with ELU activations. For FMPE (raw), we use an MLP with 8 layers and 256 hidden features, with ELU activations. All

baseline methods embed the observation with a CNN embedding net into a 100-dimensional vector using 4 convolutional layers with kernel size 5 followed by max pooling of kernel size 2, followed by a 4-layer MLP with 100 hidden units, with RELU nonlinearities throughout.

For FNOPE and FNOPE (fix) we use 32 Fourier modes for the FNO blocks. The network is made of 5 FNO blocks with 32 channels, while the context is embedded into 32 channels. We train for a maximum of 300 epochs with an early patience of 50. We use a training batch size of 200 and a learning rate of 0.0005. For FNOPE, we set the target gridsize $N_{\text{ds}} = 2048$. The architecture includes 8 channels for positional embedding, and 8 channels for time embedding. All nonlinearities are GELUs.

S6.4 Mass balance rates of Antarctic Ice Shelves

For all the baseline methods, we train the network using an Adam optimizer with a learning rate of 0.0001, and a batch size of 200. For NPE/FMPE (spectral), we use 10 modes, leading to 20 parameters to learn, and a pad width of 20 for the spectral preprocessing (Appendix S4.1). For NPE (spectral) the density estimator is a NSF with 2 residual blocks with 50 hidden dimensions each, 5 transforms, with RELU activations. For FMPE (spectral), we use an MLP with 5 linear layers with 64 hidden dimensions to estimate the flow, with ELU activations. For FMPE (raw), we use an MLP with 5 layers and 64 hidden features, with ELU activations. For all baseline methods, we used the same embedding as Moss et al. [39], which was a CNN embedding the 441-dimensional observation into a 50-dimensional vector using 2 convolutional layers with kernel size 5 followed by max pooling of kernel size 2, followed by a 2-layer MLP with 50 hidden units, with RELU nonlinearities throughout. The configuration of NPE (raw) is described in Appendix S4.3.

For FNOPE we use 10 Fourier modes for the FNO blocks. The network is made of 5 FNO blocks with 16 channels, while the context is embedded into 8 channels. We train for a maximum of 1000 epochs with an early patience of 50. We use a training batch size of 200 and a learning rate of 0.001. We do not include the data augmentation procedure for this experiment, as the discretizations of both observations and parameters was fixed to the setting of [39]. We still include positional embedding due to the parameter and observations being discretized differently to one another: the architecture included 4 channels for positional embedding, and 4 channels for time embedding. All nonlinearities are GELUs.

For the experiments on 500 gridpoints (Fig. S5) we use the same hyperparameters.

S7 Ablation experiments

S7.1 Linear Gaussian

To investigate the influence of different hyperparameters on the performance of FNOPE, we ran several ablation experiments. First, we studied the performance of FNOPE in the deliberately insufficient setting where the prior distribution over the parameter contains higher-frequency modes than what is modeled by the FNO blocks. To this end, we changed the lengthscale of the prior distribution for the Linear Gaussian task (Sec. 4.2) by a factor of 10 to 0.005, resulting in higher frequency components for the parameter θ . We then trained FNOPE with a varying number of spectral modes in the FNO blocks. With a lower number of modes, the performance of FNOPE degrades, and the posterior samples clearly miss the high frequency components present in the ground truth observations (Fig. S1a,b). Second, we varied the number of unmasked points in the FNOPE training procedure (Sec. 3.1). While the SWD decreases with the number of unmasked points, it saturates at 512 unmasked points, which is approximately half of the observed data (Fig. S1c). We then consider changing the lengthscale heuristic used to define the base distribution for FNOPE (Sec. 3.1), by defining $l = L_0/M$ for the number of modes M and some lengthscale scaling factor L_0 . In the extreme case $L_0 = 0$, this corresponds to the base distribution sampling uncorrelated white noise (WN). The lengthscale heuristic used in our work corresponds approximately to $L_0 = 4/\pi \approx 1.27$. In contrast to the other hyperparameter ablation experiments, varying the lengthscale scaling factor does not seem to impact the performance of FNOPE considerably for this task. FNOPE performs well over a wide range of lengthscale scaling factors (Fig. S1d). Finally, we consider a version of FNOPE which keeps the masking scheme as described in Sec. 3.1, but without adding positional noise to the remaining positions, which effectively only sees positions on the simulation grid during training

(FNOPE (no jitter)). On this task, we evaluate on an equispaced grid, and we see that FNOPE (no jitter) performs similarly to FNOPE (Fig. S1e). Therefore, the performance of FNOPE does not degrade from the inclusion of positional noise in this task.

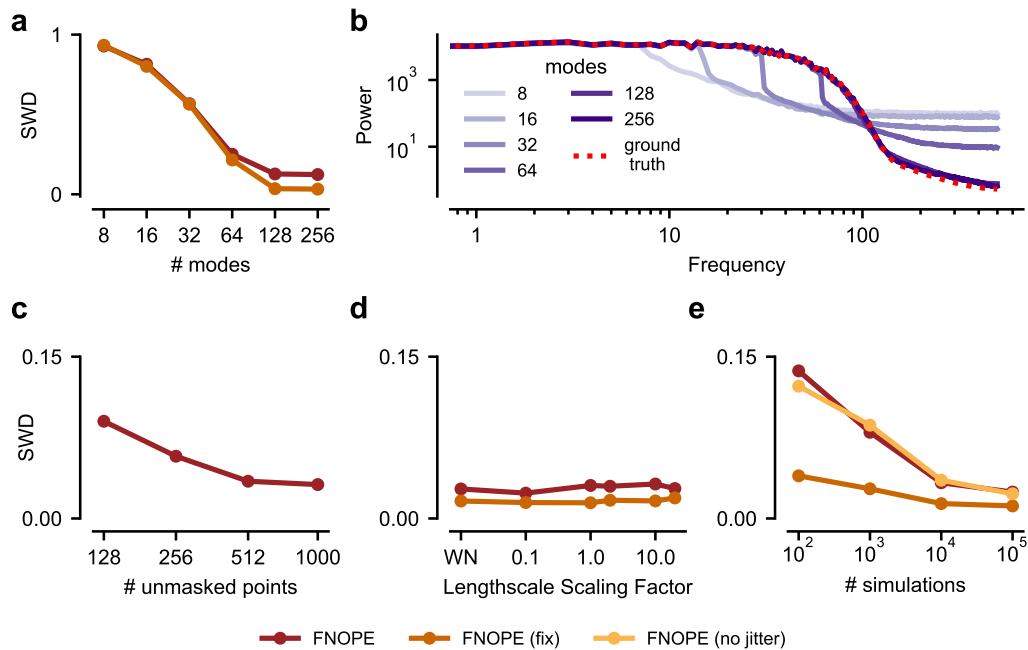


Figure S1: **Linear Gaussian ablation experiments.** (a) Performance for 10⁴ training samples in terms of SWD for the Linear Gaussian experiments with varying number of Fourier modes in the FNO block. Note that we changed the lengthscale of the simulator prior by a factor of 10 to 0.005 (Appendix S5). (b) Power analysis of FNOPE (fix) samples and ground truth x for different number of used modes in the FNO block. (c) Performance in terms of SWD for different numbers of masked points in FNOPE. (d) Influence of the lengthscale of the noise process on the SWD. “WN” corresponds to white noise (an uncorrelated Gaussian distribution). (e) Same as Fig. 3 with a version of FNOPE in which we omit the adding of positional noise (FNOPE (no jitter)).

S7.2 SIRD

First, we compare the performance of FNOPE to Simformer on the SIRD task when using observations at 20 (instead of 40) randomly sampled times (Fig. S2). The performance of FNOPE slightly degrades, while Simformer still performs robustly. Second, we adapt the lengthscale heuristic used to define the base distribution for FNOPE (Sec. 3.1), by defining $l = L_0/M$ for the number of modes M and some lengthscale scaling factor L_0 . The lengthscale heuristic used in our work corresponds approximately to $L_0 = 4/\pi \approx 1.27$. We observe that FNOPE performs robustly for a wide range of lengthscale scaling factors.

S7.3 Darcy Flow

In the Darcy task, we ablate the number of modes used in the FNO block for FNOPE. We see that while the MSE of posterior predictive simulation and posterior calibration as measured by the SBC EoD of marginal distributions is not strongly affected by the number of modes used, the posterior log-probability per pixel increases with the increased model capacity (Fig. S4a). This is expected, as more modes allow FNOPE to estimate a more constrained posterior distribution, leading to higher posterior densities. Second, we adapt the lengthscale heuristic used to define the base distribution for FNOPE (Sec. 3.1), by defining $l = L_0/M$ for the number of modes M and some lengthscale scaling factor L_0 . The lengthscale heuristic used in our work corresponds approximately to $L_0 = 4/\pi \approx 1.27$. We see that increasing the lengthscale scaling factor improves the calibration

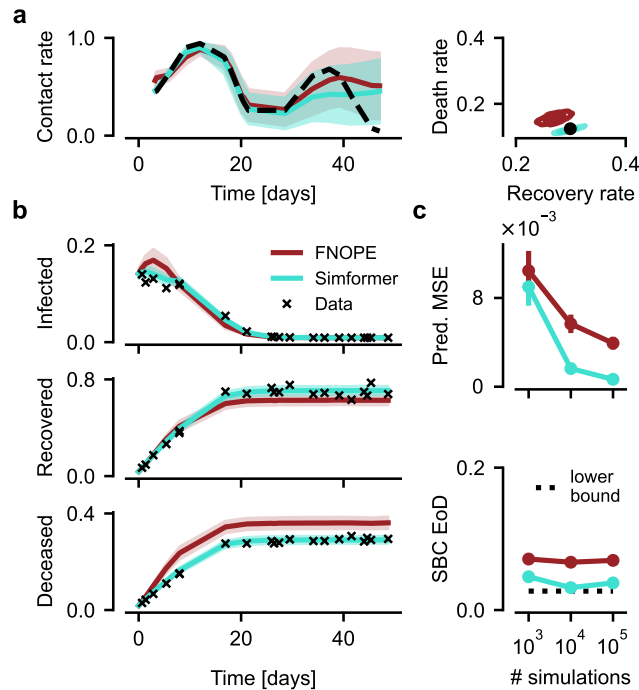


Figure S2: **SIRD model on 20 conditioning points.** (a) Posterior conditioned on 20 time points. *left:* Posterior (mean $\pm$ std.) of the time-varying parameter and ground truth parameters (dashed). *right:* Two dimensional posterior of vector-valued parameters and ground truth parameters (dot). (b) Posterior predictive (mean $\pm$ std.) of infected, recovered and deceased populations with observations marked. (c) *upper:* MSE of posterior predictive samples to observations. *lower:* Simulation-based calibration error of diagonal (SBC EoD). ‘Lower bound’ refers to the SBC EoD for uniformly sampled posterior ranks (details in Appendix S2). See Fig. 4 for the results with 40 conditioning points.

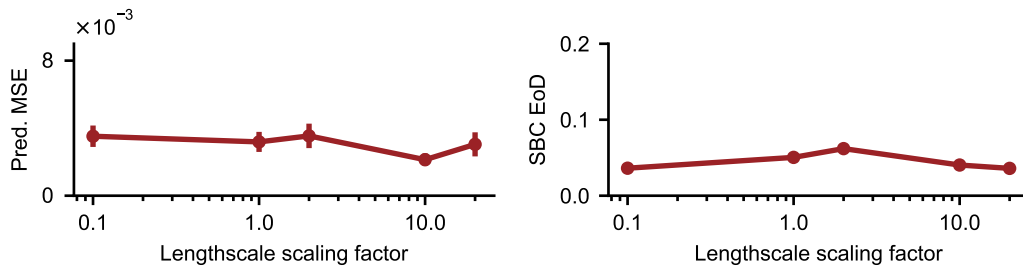


Figure S3: **SIRD ablation experiments.** Influence of the lengthscale of FNOPE’s base distribution (Sec. 3.1) on posterior quality in terms of MSE of posterior predictive simulations and Simulation-based calibration error of diagonal (SBC EoD).

of the posterior learned with FNOPE. We observe that FNOPE achieves its best performance in terms of posterior log-probability when the base distribution lengthscale scaling factor is set to a value around our lengthscale heuristic (Fig. S4b).

S7.4 Mass balance rates of Antarctic Ice Shelves

We repeat the Antarctic ice mass balance experiment (Sec. 4.5), without downsampling the parameter space from simulation to inference, i.e. inferring the full 500-dimensional posterior distribution over the mass balance parameters. Overall, we observed that FNOPE maintains its performance on this higher-dimensional problem (Fig. S5). For low simulation budgets, the performance of

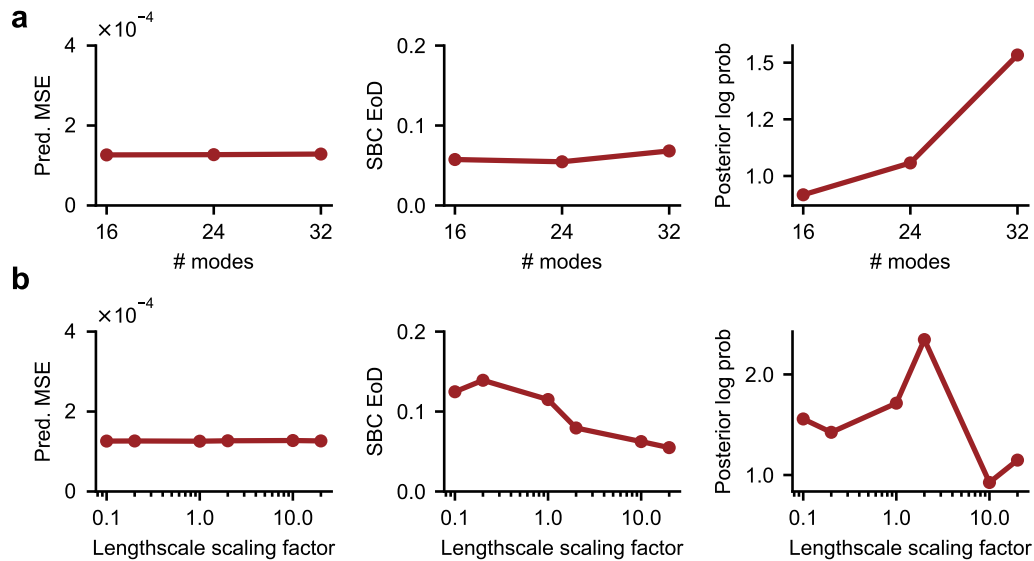


Figure S4: **Darcy ablation experiments** (a) Influence of number of used modes in the FNO block on different performance measures. The original experiments used 32 modes. Measures are the same as in Fig. 5b-d. (b) Influence of the noise length scale on different performance measures. The original experiments used a lengthscale of ≈ 1.27 . Measures are the same as in Fig. 5b-d.

FNOPE exceeds the other methods in terms of predictive MSE for both synthetic and real observations. As expected, spectral baseline methods significantly outperform the other baselines, especially at low simulation budgets.

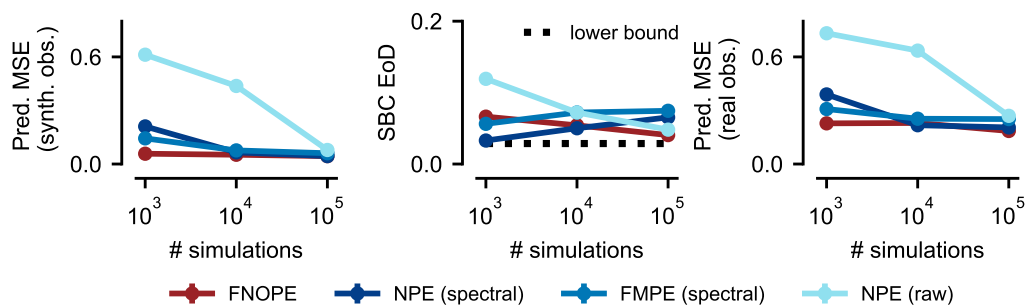


Figure S5: **Mass balance rates of ice shelves.** Inference of mass balance rates where the parameter discretization uses 500 gridpoints (observation discretization remains unchanged). Performance measures on test simulations and the real observation, where NPE (raw) refers to the method used in Moss et al. [39]. Results for FMPE (raw) omitted as this baseline was not able to always produce samples within the prior bounds across all test observations. See Fig. 6 for the results based on 50 grid points.

S8 Additional Results

S8.1 Additional Darcy results

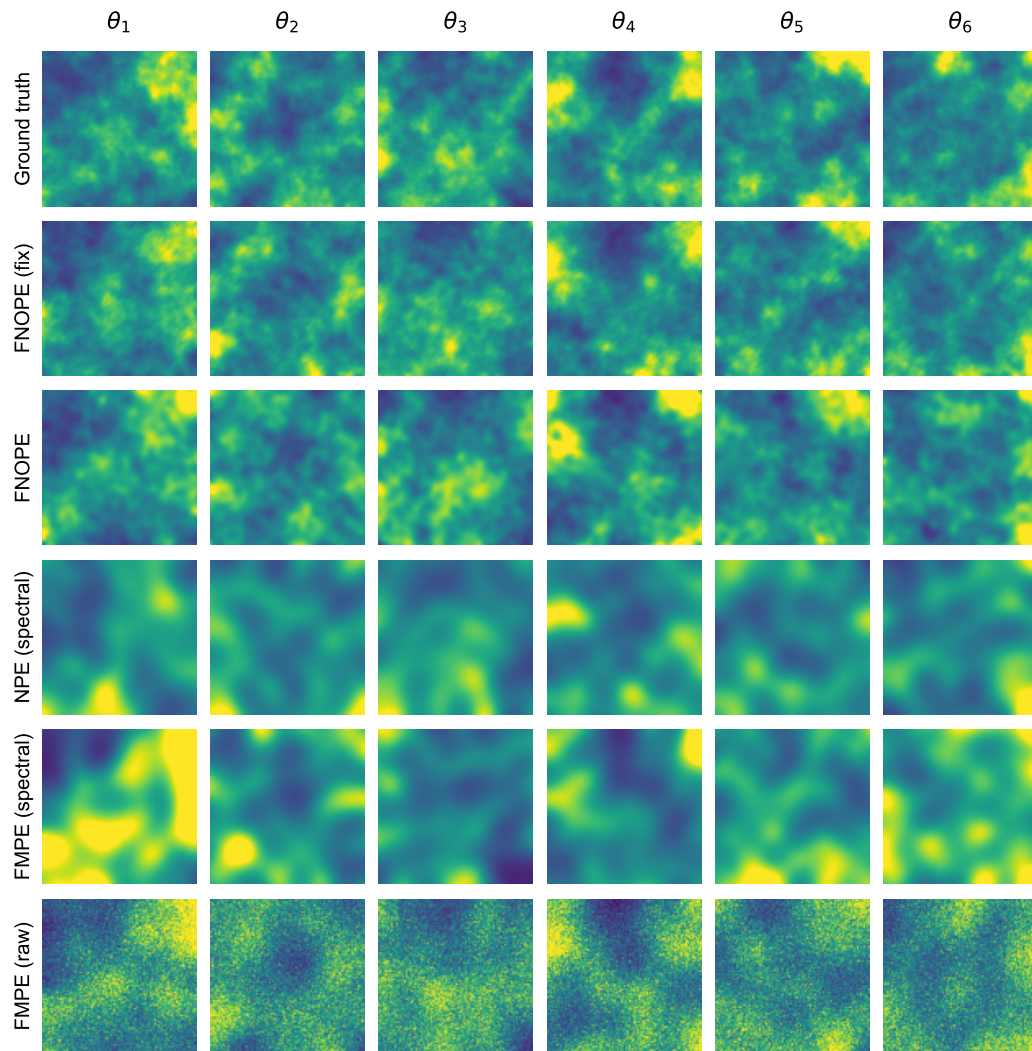


Figure S6: **Posterior samples Darcy flow experiment.** Additional posterior samples for the Darcy flow experiment for six distinct observations simulated from ground truth parameter θ_i .

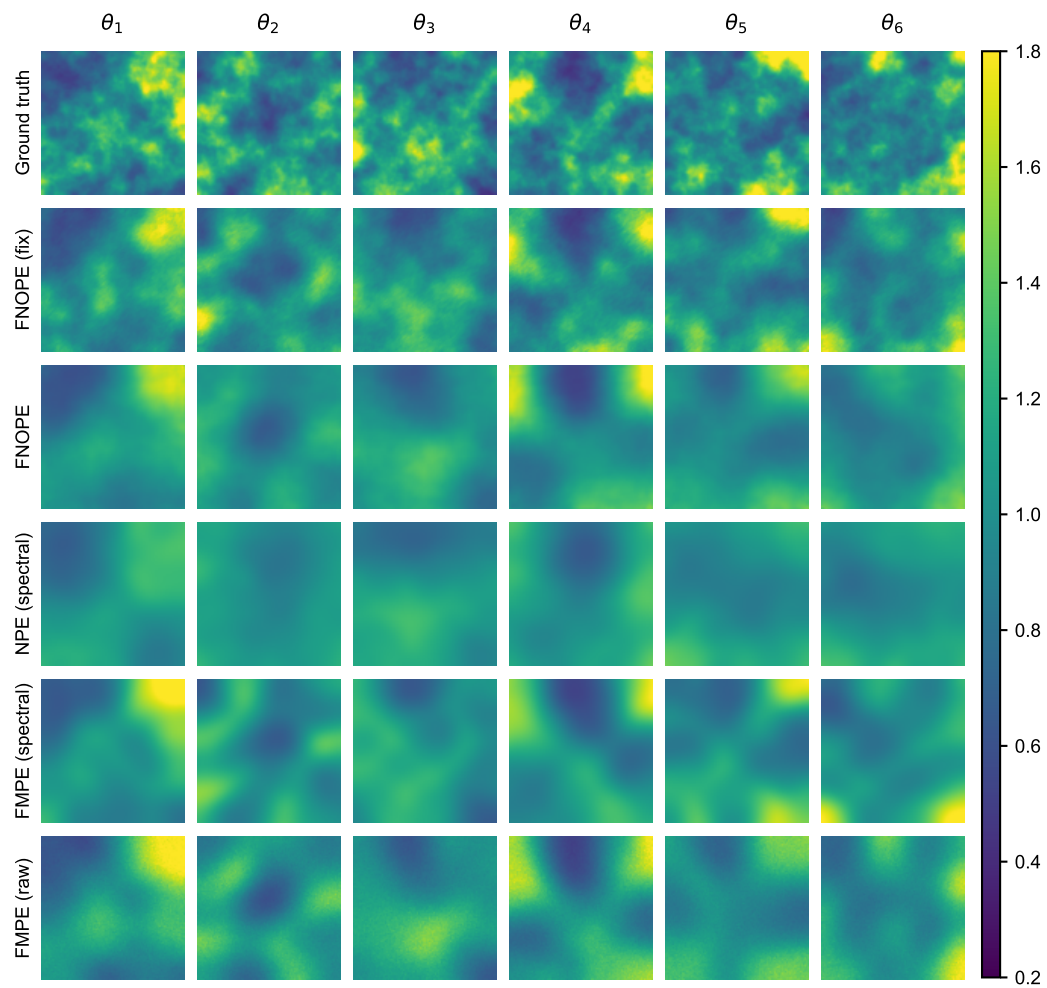


Figure S7: **Posterior means for Darcy flow experiment.** Posterior means (based on 100 samples) for the Darcy flow experiment for six distinct observations simulated from ground truth parameter θ_i .

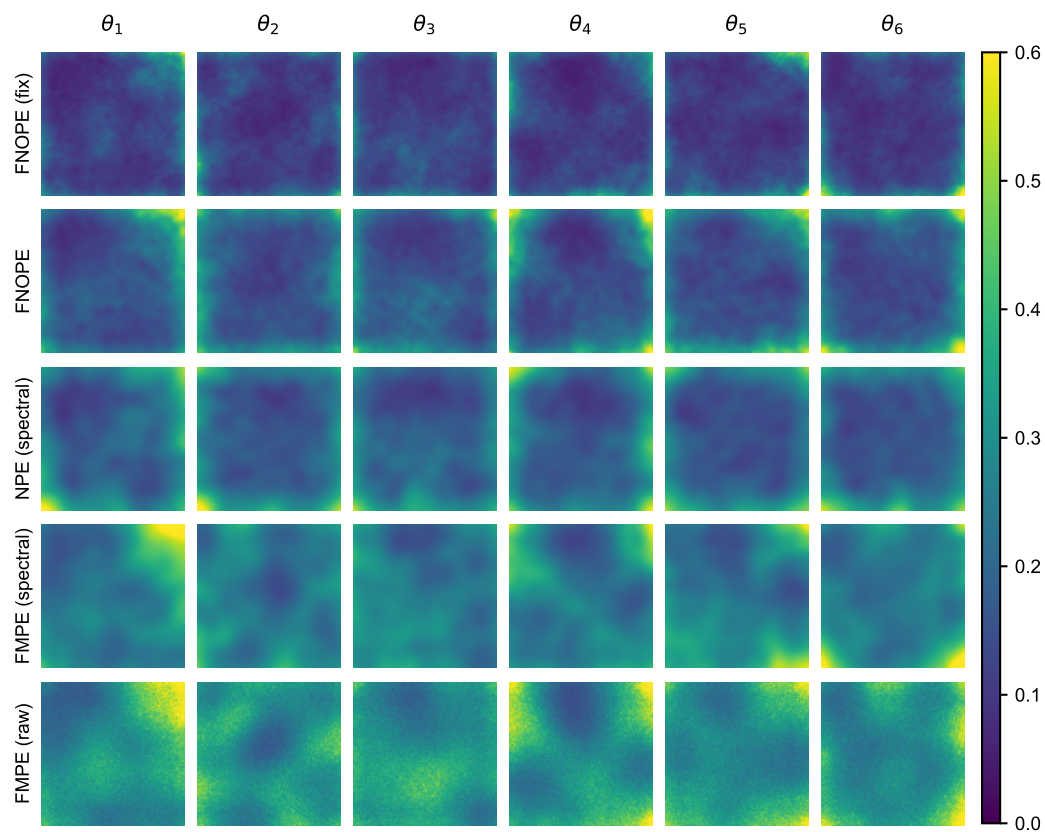


Figure S8: **Posterior standard deviation for Darcy flow experiment.** Posterior standard deviations (based on 100 samples) for the Darcy flow experiment for six distinct observations simulated from ground truth parameter θ_i .

S8.2 Comparison to invertible Fourier Neural Operator

As an additional comparison, we apply the work of Long et al. [50] and train an invertible Fourier Neural Operator (iFNO) to solve the Darcy Flow inverse problem, using the same training data and simulation budgets. We train the iFNO with the same settings as described in Long et al. [50] for the D-CURV experiment, summarized below. We observe that while iFNO achieves a comparable performance to FNOPE in terms of its predictive MSE, the posterior distribution is not calibrated as measured by the SBC EoD (Fig. S9a,b) and essentially collapsed to a point estimate. The SBC EoD value measured for iFNO (around 0.25) is consistent with this point estimate, because the one-dimensional marginal of a point mass distribution either always overestimates or underestimates the ground truth value. Hence, recalling the definition of SBC EoD (Sec. S2.2), the ground truth will have rank $r_{ij} = 1$ or $r_{ij} = K_{\text{post}} + 1$. Supposing that over different observations x_j^o , the point mass distribution is equally likely to underestimate or overestimate the ground truth, the cumulative distribution of the ranks will be given by $\text{CDF}_i(\alpha) = 0.5$ for all significance levels $\alpha \in (0, 1)$. Given the definition of the SBC EoD, a point mass distribution results in a SBC EoD value of

$$\int_0^1 |\text{CDF}_i(\alpha) - \alpha| d\alpha = 0.25 \quad (7)$$

for each dimension i . This overconfidence is also reflected in the standard deviations of the estimated distributions (Fig. S9c), which show that iFNO essentially estimates a point mass for this task.

iFNO hyperparameters We trained iFNO with 4 FNO blocks and 16 Fourier modes. The number of training epochs for the invertible Fourier blocks was set to 100, for the β -VAE 1000, and for joint training 100. The architecture of the β -VAE was the same as in Long et al. [50] with rank 32, as well as the value $\beta = 10^{-6}$. We used a minibatch size of 10 for joint training and 20 for the VAE and iFNO pretraining.

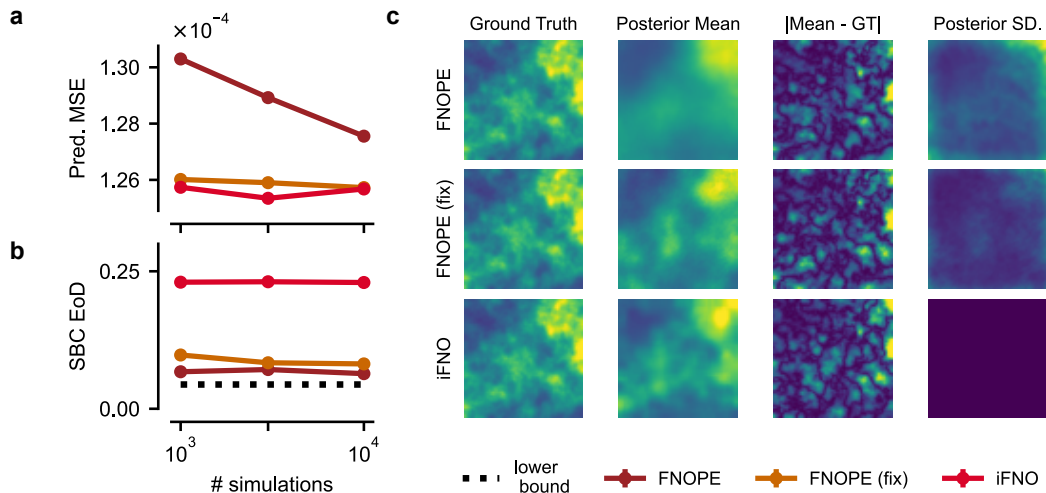


Figure S9: **Darcy Flow, comparison of FNOPE to iFNO [50].** (a) MSE of posterior predictives to the ground truth observation (zoomed in relative to Fig. 5). (b) Simulation-based calibration Error of Diagonal (SBC EoD) for different training budgets. (c) Ground truth θ , Posterior means, pixelwise error of means relative to ground truth, and posterior standard deviations. The color bars of the means match Fig. S7, and for errors and standard deviations they match Fig. S8.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: We state the contributions of our work: We introduce FNOPE, a simulation-based inference method for inferring function-valued parameters. We demonstrate on synthetic and real world tasks its ability to generalize to new and non-uniform discretizations of the parameter and observation domains, as well as its scalability to large numbers of parameters.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We explicitly mark a discussion on the limitations of our approach, including limitations of the low-frequency assumption of Fourier Neural Operators, and limitations in higher-dimensional parameter domains than those considered in our experiments.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Our work includes a lengthscale selection heuristic for the FNOPE noise distribution, for which we provide a derivation for in Appendix S3.2, and appropriately refer the reader to this derivation in the main text.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide all experimental details in appropriately labelled subsections of the appendix, including simulator configurations, model hyperparameters for each experiment, and details on the evaluation metrics. In addition, we provide our code to reproduce all experiments in the supplementary materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide our code and document how to run the experiments included in our paper. These include the simulators, as well as the random seeds used to generate the training data from the simulators.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We report all experimental details, including data splits, model hyperparameters, and optimizers in Appendix S6. We provide the random seeds used to run our training and evaluation scripts to allow full reproducibility of our experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report the standard errors in our results over 3 independent runs for each experiment. The size of the test set is provided for each experiment in the appendix and in the configuration files provided with the code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: We provide the specifications of our compute nodes as well as further information in Appendix S1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [\[Yes\]](#)

Justification: We confirm that this work conforms to the NeurIPS Code of Ethics in all respects.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: Our work presents a methodological development in performing simulation-based inference, with the goal of enabling scalable and flexible inference in various scientific domains. We do not target applications with direct societal impact, and do not foresee potential for misuse.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our work uses only synthetic or publicly available data. Our work is methodological and empirical, and therefore we do not see a risk of misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [\[Yes\]](#)

Justification: We credit all external asserts in our work, including datasets. We reproduce a figure component from existing published work which is correctly attributed and licensed, with some details omitted for the review stage to preserve anonymity.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: We provide the code for our method with appropriate documentation. No other assets are produced in this work.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#)

Justification: This work does not include crowdsourcing or research with human experts.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No human subjects were involved in this study, nor data involving human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs were not used in the development, implementation, or evaluation of the proposed method according to the NeurIPS LLM policy.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.