The Tip of the Iceberg: Enabling Scalable Simulation-Based Inference in Geoscience

DISSERTATION

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Guy Moss
aus Ramat Gan/Israel

Tübingen
2025

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation: 19.12.2025

Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Jakob H. Macke
2. Berichterstatter:	Jun.-Prof. Dr. Nicole Ludwig

Abstract

Geoscientists use computer simulations to understand and model a myriad of processes taking place on our planet. Bayesian inference is a statistical framework that identifies parameter configurations that make the simulator's predictions consistent with experimental observations. Traditional Bayesian inference approaches make use of the likelihood function, which quantifies the probability of an observation given a particular parameter configuration. However, this likelihood function cannot be evaluated directly for many computer simulators due to their complexity. This limitation necessitates new algorithms for performing inference. One recently proposed solution to this problem is Simulation-Based Inference (SBI).

This thesis consists of three publications that develop and apply new SBI approaches to enable statistical inference for larger, more complex geoscientific problems that were previously infeasible. In the first publication, we introduce a particular inference problem in the field of glaciology. This problem pertains to the inference of surface accumulation and basal melt rates of Antarctic ice shelves from radar measurements of their internal layering structure. We present a statistical framework to describe this inference problem, and apply an existing SBI method to provide an uncertainty-aware solution to this problem for the first time. In the second publication, we develop a new approach to simulation-based inference of function-valued parameters. Current SBI methods require a very large number of simulations to solve geoscientific inference problems, which can be computationally demanding or even unattainable. This is because geoscientific parameters are function-valued, describing quantities that vary in space and/or time and resulting in many values to infer. Our method exploits the spatial and temporal correlations in geoscientific parameters to perform inference using a much smaller number of simulations than existing SBI methods. We apply this approach to the Antarctic ice shelf case study, and show that it reduces the computational cost of inference by two orders of magnitude. In the third publication, we tackle a distinct type of inference problem known as source distribution estimation. This aims to identify a distribution over the parameters that is consistent with a dataset of observations, as opposed to a single or repeated measurements. This inference paradigm is also ubiquitous in geoscientific applications, for instance in extreme event modeling. We develop a new simulation-based approach to estimating source distributions and demonstrate its applicability to challenging scientific tasks.

Overall this thesis develops new statistical inference methods and applies them to solve challenging problems in various geoscience domains. It thus provides a vital connection between machine learning methodology and scientific practice, enabling statistical inference for complex and high-dimensional simulators in geoscience.

Zusammenfassung

Geowissenschaftler verwenden Computersimulationen, um eine Vielzahl von Prozessen auf unserem Planeten zu verstehen und zu modellieren. Mit Bayes'scher Inferenz, einem statistischen Ansatz, können Parameterkonfigurationen identifiziert werden, die die Vorhersagen des Simulators mit experimentellen Beobachtungen in Einklang bringen. Traditionelle Bayes'sche Inferenzansätze verwenden die Likelihood-Funktion, die die Wahrscheinlichkeit einer Beobachtung für eine bestimmten Parameterkonfiguration quantifiziert. Diese Wahrscheinlichkeitsfunktion kann jedoch aufgrund ihrer Komplexität für viele Computersimulatoren nicht direkt ausgewertet werden. Diese Einschränkung erfordert neue Algorithmen für die Durchführung der Inferenz. Eine kürzlich vorgeschlagene Lösung für dieses Problem ist die simulationsbasierte Inferenz (SBI).

Diese Arbeit besteht aus drei Publikationen, in denen neue SBI-Ansätze entwickelt und angewendet werden, um statistische Inferenz für größere, komplexere geowissenschaftliche Probleme zu ermöglichen, die bisher nicht realisierbar waren. In der ersten Publikation stellen wir ein spezielles Inferenzproblem aus dem Bereich der Glaziologie vor. Hier sollen die Oberflächenmassenbilanz und die basalen Schmelzraten von antarktischen Eisschelfen aus Radarmessungen ihrer inneren Schichtung abgeleitet werden. Wir stellen einen statistischen Ansatz zur Beschreibung dieses Inferenzproblems vor und wenden eine bestehende SBI-Methode an, um erstmals eine Lösung für dieses Problem zu finden, die Unsicherheiten miteinbezieht. In der zweiten Veröffentlichung entwickeln wir einen neuen Ansatz für die simulationsbasierte Inferenz von funktionswertigen Parametern. Aktuelle SBI-Methoden erfordern eine sehr große Anzahl von Simulationen, um geowissenschaftliche Inferenzprobleme zu lösen, was rechnerisch aufwendig oder sogar unmöglich sein kann. Dies liegt daran, dass geowissenschaftliche Parameter funktionswertig sind und Größen beschreiben, die sich räumlich und/oder zeitlich ändern und zu vielen zu inferierenden Werten führen. Unsere Methode nutzt die räumlichen und zeitlichen Korrelationen in geowissenschaftlichen Parametern, um die Inferenz mit einer viel geringeren Anzahl von Simulationen als bei bestehenden SBI-Methoden durchzuführen. Wir wenden diesen Ansatz auf die Fallstudie zum antarktischen Schelfeis an und zeigen, dass er den Rechenaufwand für die Inferenz um zwei Größenordnungen reduziert. In der dritten Veröffentlichung befassen wir uns mit einer besonderen Art von Inferenzproblem, der sogenannten Source Distribution Estimation. Dabei geht es darum, eine Verteilung über die Parameter zu identifizieren, die mit einem Datensatz von Beobachtungen übereinstimmt, im Gegensatz zu einzelnen Messungen. Dieses Inferenzproblem ist auch in geowissenschaftlichen Anwendungen, beispielsweise in der Modellierung von Extremereignissen. Wir entwickeln einen neuen simulationsbasierten Ansatz zur Schätzung von Source Distributions und demonstrieren seine Anwendbarkeit auf anspruchsvolle wissenschaftliche Aufgaben.

Insgesamt entwickelt diese Arbeit neue statistische Methoden und wendet sie zur Lösung anspruchsvoller geowissenschaftlicher Probleme an. Sie stellt eine wichtige Verbindung zwischen maschinellem Lernen und wissenschaftlicher Praxis her und ermöglicht statistische Inferenz für komplexe Simulatoren.

Acknowledgements

Working towards a PhD can be a disorienting experience, and it's easy to lose your way. I owe a debt of gratitude to many people who helped me find my way to the other side.

Firstly, I would like to thank Prof. Dr. Jakob Macke and Prof. Dr. Reinhard Drews. Thank you for your mentorship, for your words of encouragement, and for fostering incredible work environments which made me excited to do research even on the hardest days. Thank you also to Prof. Dr. Nicole Ludwig for evaluating this thesis, and to Prof. Dr. Georg Martius for serving on my examination committee.

I want to thank all my colleagues in the Macke lab and in the Glaciology and Geophysics group, past and present, for letting me be part of such fun, collaborative teams. Special thanks go to Dr. Cornelius Schröder, who went above and beyond and supported me along every step of this journey, and to Dr. Vjeran Višnejvić, for helping me become a glaciologist. Thank you to Jaivardhan Kapoor and Sebastian Bischoff, who for almost four years shared their office with me and endured my less-thought-out ideas. To Dr. Jan Teusen and Dr. Michael Deistler, who taught me their trade secrets for writing good research code. Thanks also to Franziska Weiler for her administrative support and for generally keeping me out of trouble.

I am grateful to my wonderful colleagues who provided feedback on this thesis. Thank you to Dr. Cornelius Schröder, Dr. Daniel Gedon, Julius Vetter, Leah Mühle, and Stefan Wahl.

Finally, I would like to thank my friends, family, and partner for keeping me (mostly) sane over the past few years. To Dr. Diganta, who is right about most things, except the power consumption of televisions. Thank you also to Alex, Cat, Diganta, Elena, and Paul for coming all the way to visit me in Tübingen, and to Charly who I visit in Tübingen full-time. To Dana, who will ensure that any PhD awarded refers to me as "Dana's brother". To Caro, who makes my days brighter, literally as well as figuratively. And last but not least, to my parents, Anna and Alon, for everything.

List of Publications

Primary Contributions

Guy Moss, Vjeran Višnjević, 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 (2025).

Guy Moss, Leah Sophie Muhle, Reinhard Drews, Jakob H. Macke and Cornelius Schröder. “FNOPE: Simulation-based inference on function spaces with Fourier Neural Operators.” Advances in Neural Information Processing Systems, volume 38 (NeurIPS 2025).

Julius Vetter*, **Guy Moss***, Richard Gao, Cornelius Schröder and Jakob H. Macke “Sourcerer: Sample-based Maximum Entropy Source Distribution Estimation.” Advances in Neural Information Processing Systems, volume 37 (NeurIPS 2024).

Co-Author Contributions

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 Ziaeemehr, 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, Jakob H. Macke. “sbi reloaded: a toolkit for simulation-based inference workflows.” Journal of Open Source Software (2024).

Vjeran Višnjević, **Guy Moss**, A. Clara J. Henry, Christian T. Wild, Daniel Steinhage, Reinhard Drews. “Mapping the composition of Antarctic ice shelves as a metric for their susceptibility to future climate change.” Geophysical Research Letters (2025).

Michael Deistler, Jan Boelts, Peter Steinbach, **Guy Moss**, Thomas Moreau, Manuel Gloeckler, Pedro L. C. Rodrigues, Julia Linhart, Janne K. Lappalainen, Benjamin Kurt Miller, Pedro J Gonçalves, Jan-Matthis Lueckmann, Cornelius Schröder, Jakob H. Macke. “Simulation-Based Inference: A Practical Guide.” arXiv e-Print (2025).

Contents

Acronyms

ABC Approximate Bayesian Computation.

CNN Convolution Neural Network.

ELBO Evidence Lower Bound.

EoD Error of Diagonal.

FMPE Flow Matching Posterior Estimation.

FNO Fourier Neural Operator.

FNOPE Fourier Neural Operators for Posterior Estimation.

GP Gaussian Process.

GPR Ground Penetrating Radar.

i.i.d Independently and Identically Distributed.

IRH Internal Reflection Horizon.

KL Kullback-Leibler (divergence).

MCMC Markov Chain Monte Carlo.

MLE Maximum Likelihood Estimation.

MSE Mean Square Error.

NLE Neural Likelihood Estimation.

NPE Neural Posterior Estimation.

NRE Neural Ratio Estimation.

SBC Simulation-Based Calibration.

SBI Simulation-Based Inference.

SWD Sliced-Wasserstein Distance.

Chapter 1

Introduction

The study of the Earth is inextricably linked to inference. The goal of inference is to identify unknown properties of a system based on observations of quantities that depend on them. Geoscientists have tackled inference problems for millennia: As early as the 3rd century BCE, Erastothenes was able to reason about the climate as a consequence of the Earth's inclination to the Sun [Edwards, 2011]. He is also credited with solving one of the world's first inference problems—by deducing the circumference of the Earth from measurements of the length of a shadow cast by a vertical rod. With increasing mathematical sophistication, more complex inferences became possible. Newton's development of calculus[1] [Newton, 1726] allowed him to estimate the Earth's gravitational constant from pendulum deflection measurements. Inference can also be a first step in our understanding of real-world phenomena. By measuring the flow and deformation of glaciers, Weinberg [1907] was able to estimate the viscosity of ice. Thus, when Antarctica was first discovered, inference problems were solved to develop the first ice flow models used to describe the continent. In the present, inference problems are also used to incorporate large datasets such as satellite observations into our models of the world [Goldberg and Sergienko, 2011, Mauritsen et al., 2012]. Inference problems are typically split into two categories. In *deterministic* inference, we estimate a single best "guess" of the quantity of interest given available data. However, the best estimate may not always correspond to reality. Hence, we often want to quantify the uncertainty in our estimates, by capturing all possible solutions and determining the confidence in those solutions. To obtain such uncertainties, we must turn to *statistical* inference.

One powerful framework for performing statistical inference is Bayesian inference. Bayesian inference requires two components: (i) A *prior* distribution, quantifying our existing knowledge about the unknown quantities, known as parameters, and (ii) A *likelihood function*, describing the probability of the dependent variable, known as the observation, given a specific value for the parameters. Bayes' theorem provides a formula for combining these components into the *posterior* distribution, which represents the updated belief about the parameters given the observation. Bayesian inference is a tool in countless disciplines, from particle physics [Feroz et al., 2009] to medicine [Ashby, 2006].

In many scientific applications, the likelihood function is described in terms of a computer simulator derived from theoretical principles. Simulators take the parameters as an input, and output a possible observation, i.e., a sample from the likelihood function. When the likelihood function is sufficiently complex, the Bayesian posterior cannot be calculated analytically and needs to be estimated. The advent of computational methods has enabled a renaissance in Bayesian statistics

[1] At last, this discussion can be put to rest.

through Monte Carlo methods [Metropolis and and, 1949, Rubinstein and Kroese, 2016], such as Markov Chain Monte Carlo (MCMC Metropolis et al. [1953], Duane et al. [1987], Hoffman and Gelman [2014], Betancourt [2019]) and particle filtering [Gordon et al., 1993, Liu et al., 1998, Doucet et al., 2001]. These algorithms enable Bayesian inference by performing repeated simulations. All these methods are based on the assumption that the simulator can evaluate the likelihood of the simulations it generates. However, for increasingly complex simulators, evaluating the likelihood of a simulation requires computing challenging integrals over all the possible ways the simulation could have unfolded. Such likelihood functions are *intractable*. Sampling an observation given a parameter configuration from an intractable likelihood is possible, but evaluating the likelihood function to compute the probability its probability given the parameter configuration is impractical. This is a crucial distinction, as the traditional algorithms for estimating the Bayesian posterior distribution require this probability.

Simulation-based inference (SBI) [Wood, 2010, Papamakarios and Murray, 2016, Lueckmann et al., 2017, Papamakarios et al., 2019, Durkan et al., 2020, Greenberg et al., 2019, Cranmer et al., 2020, Radev et al., 2022, Deistler et al., 2025] approaches seek to approximate the Bayesian posterior for computational simulators where the likelihood is intractable. In particular, *neural* SBI methods build on advances in machine learning to approximate Bayesian inference. SBI performs simulations for many different values of the parameters. These simulations form a training dataset, with which an artificial neural network can be trained to estimate Bayesian quantities, such as to approximate the unknown likelihood function (NLE, Wood [2010], Papamakarios et al. [2019]), the likelihood-to-evidence ratio (NRE, Durkan et al. [2020], Miller et al. [2022]), or even the posterior distribution directly (NPE, Papamakarios and Murray [2016], Lueckmann et al. [2017], Greenberg et al. [2019]). Neural SBI methods have been successfully applied to enable statistical inference in a variety of disciplines, such as neuroscience [Gonçalves et al., 2020], astronomy [Dax et al., 2021], spectroscopy [Dingeldein et al., 2023], and many others. In addition, toolboxes implementing a variety of SBI methods facilitate applications of these methods without in-depth machine learning expertise [Tejero-Cantero et al., 2020, Radev et al., 2023, Dirmeier et al., 2024, Boelts et al., 2025].

In geoscience, there exists a tension between the development of computational simulators of increasing complexity and fidelity, and statistical inference. As our understanding of physical systems improves, geoscientists develop simulators that more accurately model the real world [Winkelmann et al., 2011, Megann et al., 2014, Danabasoglu et al., 2020]. However, these simulators come with a large number of parameters, and at increasing computational costs. The result is that the required simulation budget to perform inference can increase exponentially with the number of parameters due to the curse of dimensionality [Bellman, 1966]. This is especially a challenge for geoscientific models, where the parameters are typically not just scalars, but rather functions that themselves vary in space and/or time. These challenges have so far rendered Bayesian inference and SBI infeasible for many geoscientific applications.

In this thesis, I bridge the gap between SBI and geoscience by developing SBI methodologies tailored to the unique challenges posed by geoscientific inference problems. In particular, this thesis consists of three publications tackling distinct parts of the inference workflow.

In the first publication, we provide for the first time an uncertainty-aware solution to a well-known inference problem in the field of glaciology: Inferring surface accumulation and basal melting rate histories of ice over an Antarctic ice shelf. Ice shelves in Antarctica are the floating ice masses surrounding the Antarctic continent, and are crucial in holding back, or "buttressing" ice from the continent, thus improving its stability and resistance to climate change [Reese et al., 2018, Greene et al., 2022]. When making future predictions of the ice shelves, it is crucial to know

their atmospheric and oceanic boundary conditions, namely the rate of ice accumulation through snowfall, and the rate of ice melt due to interaction with the ocean. Furthermore, uncertainties in these quantities will be propagated through to uncertainties in our prediction of the future of the ice shelf, for instance as in the Intergovernmental Panel on Climate Change (IPCC) reports [IPCC, 2023]. These rates cannot be measured directly, and so many works focused on inferring the accumulation and melting rates of ice from radar measurements of the internal layering structure of the ice [Waddington et al., 2007, Catania et al., 2010, Wolovick et al., 2021]. However, previous works do not provide the crucial uncertainty estimates. In this first work, we develop a statistical framework to perform Bayesian inference of the accumulation and melt rates. This work not only constitutes one of the first applications of SBI in geoscience, but also highlights the potential gain through a statistical treatment of glaciological inference problems, as we reveal that many distinct values for the accumulation and melting rates are indeed possible.

The SBI application in our first work required a large number ($\sim$200,000) of simulations, which may not be feasible for other models arising in geoscience. A primary reason for this requirement is that the mass balance parameters are *function-valued*, as they vary spatially across the ice shelf, resulting in a large number of values to infer. In a follow up work we develop a new SBI approach, tailored to inferring function-valued parameters with small simulation budgets. When inferring function-valued parameters, which occur frequently in geoscientific inference problems, an intuitive approach is to fix a discretization of the spatial/temporal domain, and treat the value of the function at each point as a parameter. This approach has two limitations: First, the number of parameters increases exponentially with the dimensionality of the domain, and increases linearly with the resolution of the discretization. The resulting inference problems are therefore extremely high-dimensional, and consequently are very challenging. Second, the trained SBI model is then fixed to the chosen discretization: The values of the parameter cannot be queried by the model at other points of the domain without retraining the model. To overcome these challenges, we develop an approach which makes use of Fourier Neural Operators (FNOs) [Li et al., 2021]. These neural networks are designed to operate on function-valued data by operating on the low-frequency components of the parameters. This enables us to exploit the long-range correlations common to many geoscientific parameters. We apply our new approach on the Antarctic ice shelf case study, and achieve comparable results using two orders of magnitude less simulations than what was required for the existing SBI methods used in our first publication.

Our last work presents a new approach to a related problem—source distribution estimation, also known as Empirical Bayes. Recall that in Bayesian inference, we define the prior distribution using existing knowledge about the parameters. Often such knowledge is not available, and so practitioners opt to use so-called uninformative prior distributions, such as uniform or Gaussian distributions. Using such prior distributions can introduce an error to the resulting posterior distribution, as well as making the inference problem much harder to solve by introducing a very large hypothesis space. One way to address this challenge is to estimate prior distributions from experimental data—which is a source distribution estimation problem. Unlike the Bayesian inference problem, in source distribution estimation we seek to find a distribution over the parameters which is consistent with a *dataset* of experimental observations, as opposed to just one observation, or many independent and identically distributed (i.i.d) observations. While existing methods address source distribution estimation, only a few are applicable to simulators which can only sample from the likelihood function. Furthermore, the source distribution estimation problem is known to be ill-posed, having non-unique solutions. We address this problem by employing the maximum entropy principle [Jaynes, 1968]. The maximum entropy principle selects the "maximally ignorant" distribution from those which satisfy given constraints, which intuitively assumes the least about

the parameters. Furthermore, we propose a simulation-based approach to estimating source distributions. We optimize a Sliced-Wasserstein distance (SWD, Bonneel et al. [2015], Nadjahi et al. [2020]), which is completely sample-based, meaning our approach does not require evaluating the likelihood function. We show experimentally that our approach efficiently estimates high-entropy source distributions. This approach enables practitioners to estimate informed prior distributions, and thus can significantly reduce the computational burden of downstream inference problems.

The remainder of this thesis is structured as follows. I provide a general background to inference problems and Bayesian inference (Sec. 2.1-2.2). I give an overview of the relevant methods used in this thesis (Sec. 2.3-2.4). I then provide background on the main application—inferring surface accumulation and basal melt rates of Antarctic Ice Shelves (Sec. 2.5). In Sec. 3, I summarize the three main publications. I discuss the significance of this work and avenues for future work in Sec. 4. Overall, this thesis presents how the framework of SBI can be extended and applied to challenging problems in geoscience.

Chapter 2

Background

This chapter provides the general background and context for the contributions made in this thesis. I begin by discussing approaches for deterministic and statistical inference. I then provide an overview of existing work for simulation-based inference. I provide a description of concepts in modeling spatial and temporal data relevant to this work, concerning inference of function-valued parameters. Finally, I provide some background of the main application domain in glaciology.

2.1 Deterministic inverse problems

The most common form used to describe a model of a physical system in science is a function $f : \Theta \to \mathcal{X}$, taking a known vector of quantities $\theta \in \Theta$ (known as the *parameters*), and outputting some vector $x \in \mathcal{X}$ of *predictions*[1]. If the properties of the system described by θ are known, such a model allows to make predictions about (other) properties of the system, described by x, using

$$x = f(\theta). \tag{2.1}$$

A desirable model is one that perfectly matches real-world observations. That is, if we know that the system is described by θ, and we experimentally measure an *observation* x_o, then $x_o = f(\theta)$ is satisfied.

Inherently, the ability of scientific theory to derive models predicting x from θ is independent to the questions of which of θ or x are more important to quantify or which is easier to measure experimentally. Rather, such models derive from and reflect our understanding of the mechanisms of this physical system. For many applications, therefore, predicting θ from x is also an important and relevant problem. However, the existence of the model $f : \Theta \to \mathcal{X}$ does not imply that a function $g : \mathcal{X} \to \Theta$ can be derived, or that it even exists. Given sufficiently simple functions f, it may be possible to analytically derive the inverse function, $f^{-1} : \mathcal{X} \to \Theta$, satisfying

$$\begin{aligned} f^{-1}(f(\theta)) &= \theta \quad \forall \theta \in \Theta \\ f(f^{-1}(x)) &= x \quad \forall x \in \mathcal{X}. \end{aligned} \tag{2.2}$$

In the majority of scientific research, however, this is not the case. Thus, the question: *"If I observe x_o, what is the corresponding value for the parameters θ for this system?"* becomes nontrivial to answer, and the field of methods to answer such questions is known as (deterministic) inverse problems.

[1]The case where θ and x are function-valued is discussed in Sec. 2.4.2.

One of the most common approaches is *regression*[2]. In regression, a mismatch function $\mathcal{L}(\theta, x) : \Theta \times \mathcal{X} \to \mathbb{R}$ is defined to measure the "goodness of fit" of any possible value of θ to any possible value of x. Given an observation x_o, the goal of regression is to find the value of θ which minimizes the defined mismatch $\mathcal{L}$, i.e. to solve

$$\theta^* = \arg\min_{\theta \in \Theta} \mathcal{L}(\theta, x_o). \tag{2.3}$$

For example, a common choice for $\mathcal{L}$ is the mean square error (or L^2 loss) between predictions $x = f(\theta)$ and the observation x_o, resulting in least-squares regression,

$$\theta^* = \arg\min_{\theta \in \Theta} ||f(\theta) - x_o||_{L^2}. \tag{2.4}$$

Therefore, solving inverse problem can be seen as finding the "best estimate" of the parameters θ to explain the observation x_o, with respect to some pre-defined measure of goodness of fit[3]. However, obtained solutions θ^*, whether optimal with respect to $\mathcal{L}$ or not, are typically not exact, i.e. it is typically not true that $f(\theta^*) = x_o$. This is due to a variety of reasons, chief among them is capture by the famous quote by George Box: "all models are wrong". While deterministic models f are very powerful, there will always be sources of error in the real world. These can be caused not only by the insufficient precision of our measuring devices, but also due to incomplete understanding of the physical system. In order to quantify this error, we turn from deterministic inverse problems to statistical inverse problems, also known as statistical inference.

2.2 Statistical inference

Instead of making singular predictions on the values of x, statistical models instead opt to define a probability measure $\mathcal{P}_{\mathcal{X}}$ over x. This probability measure corresponds to how likely arbitrary predictions x are to be observed under the model. As in deterministic inverse problems, there is no special significance assigned to x over θ, and so a probability measure P_Θ can also be defined, and indeed the joint measure $\mathcal{P}_{\Theta \times \mathcal{X}}$. This joint measure describes not only the possible values of θ and x individually, but also how they depend on one another. While it is possible to extend the notion of statistical inference to probability measures which do **not** admit density functions, we focus in this thesis on measures which do admit probability densities $p(x)$, $p(\theta)$ and $p(\theta, x)$.

Knowledge of the joint density $p(\theta, x)$ in fact corresponds to full knowledge of the physical system under the statistical lens. It fully quantifies all possible outcomes of the system. More commonly, and in parallel to the deterministic setting, scientific theory leads to a partial model that allows us to *generate* predictions of x, *given* knowledge of θ. Such a generative model in this case takes the form of the *likelihood function*,

$$x \sim p(x|\theta = \theta_o),^{4} \tag{2.5}$$

which is a conditional distribution over the random variable x, given a fixed value θ_o of the parameter θ_o. As opposed to Eq. (2.1), the likelihood function can generate many distinct predictions for the same value of the parameters θ, with the value of the likelihood function corresponding to the relative frequency with which the respective predictions will be produced.

[2] Also known as parameter *fitting*.

[3] The goodness of fit can include hard constraints on θ, such as bounds, or soft constraints, known as *regularization*.

[4] I will use the shorthand $p(x|\theta)$ to refer to the likelihood function for an arbitrary parameter throughout this thesis.

Frequentist and Bayesian statistical inference

In the *frequentist* view of statistical inference, it is assumed that only x is a random variable, and has corresponding distribution $p(x)$ or, when conditioned on θ, $p(x|\theta)$. However, the parameters θ are assumed to be fixed, instead of random variables. Similarly to the deterministic view, the goal is to find an optimal value of θ^* which best explains the data. The optimization problem becomes to identify the value θ^* which maximizes the likelihood of the observation x_o,

$$\theta^* = \arg\max_{\theta \in \Theta} p(x_o|\theta). \tag{2.6}$$

Thus, frequentist inference is commonly referred to as *maximum likelihood estimation* (MLE). MLE is widely used to find the "best" model to describe observed data. The MLE approach is especially potent for inference problems where the parameters θ are only of interest insofar as their ability to predict x. On the other hand, reasoning about uncertainty in the parameter estimate itself is crucial for many inference problems. First, likelihood models can be degenerate, meaning that several different values of θ can produce the same x - and no unique solution x^* exists. Second, the parameter θ can be of scientific interest in its own right, as it represents physical quantities of system that are immeasurable, or otherwise difficult to measure experimentally. In such cases, it is insufficient to only provide a best estimate, but we must also provide the uncertainty associated with this estimate.

In *Bayesian* statistical inference, the parameter vector θ is also assumed to be a random variable. In the Bayesian perspective, we assume to have some prior belief about what the parameters could be, defined in the form of the prior distribution $p(\theta)$. The goal is to find the conditional distribution (also known as the *posterior* distribution),

$$p(\theta|x = x_o). \tag{2.7}$$

As in the deterministic case, our ability to define a likelihood function $p(x|\theta)$ does not imply that we can equivalently derive a model for the conditional distribution $p(\theta|x = x_o)$. Bayes' theorem gives a principled way of updating our prior belief about the parameters given an observation. If we measure an observation x_o, Bayes' theorem states that the resulting posterior distribution satisfies

$$p(\theta|x_o) = \frac{p(x_o|\theta)p(\theta)}{p(x_o)}, \tag{2.8}$$

where $p(x_o)$ is the *marginal likelihood* (also known as the model evidence). The marginal likelihood is fully determined given the assumed knowledge of the prior $p(\theta)$ and the likelihood $p(x|\theta)$, since

$$p(x_o) = \int_\Theta p(x_o|\theta)p(\theta)d\theta. \tag{2.9}$$

Therefore, the challenge of performing Bayesian inference is to solve the integral in Eq.(2.9). There exist a remarkable amount of non-trivial distributions for which a closed form solution of this integral is possible. A particularly potent example of closed-form Bayesian inference is generalized linear models [Bishop, 2006], which appear in a variety of scientific applications [Atkinson et al., 1998, Lane, 2002, Allard et al., 2012].

2.2.1 Computational Bayesian inference

When the integral Eq. (2.9) cannot be computed analytically, it may also be challenging to approximate numerically, as θ may be a high-dimensional vector. A crucial observation, however, is that to perform inference, it may be sufficient to only generate *samples* from the posterior distribution $p(\theta|x)$, as opposed to evaluate the probability density itself. This is because samples from the distribution correspond to possible models of the observed data x_o. Furthermore, quantities of interest about the distribution such as the mean, covariance matrix, and quantiles of the distribution $p(\theta|x_o)$ may be estimated from samples via the Monte Carlo approach.

This insight is the main force behind classical approaches to numerically approximate Bayesian inference. Chief among these classical approaches is Markov Chain Monte Carlo (MCMC, Metropolis et al. [1953], Duane et al. [1987], Hoffman and Gelman [2014], Betancourt [2019]). In MCMC, the goal is to sequentially generate samples $\theta_1, \theta_2, ... \sim p(\theta|x_o)$. First, a (Markovian) proposal distribution, $q(\theta'|\theta_n)$ is defined, and a sample θ' drawn from this distribution. The proposal distribution can be chosen independently of the inference problem at hand, as long as it can theoretically assign a nonzero probability to any $\theta \in \Theta$ in some finite number of steps N. For this reason, typically Gaussian proposal distributions are used. To make the Markov chain dependent on the true posterior distribution, the next value θ_{n+1} is only set to θ' with a given probability $a(\theta_n, \theta')$. In the earliest (Metropolis) formulation of MCMC, this acceptance probability is

$$a(\theta', \theta) = \min\left(1, \frac{p(\theta'|x_o)}{p(\theta|x_o)}\right) = \min\left(1, \frac{p(\theta')(p(x_o|\theta')}{p(\theta)(p(x_o|\theta)}\right), \tag{2.10}$$

where we made use of the fact that $p(x_o)$ is independent of θ to eliminate the model evidence term in Bayes' theorem. With probability $(1 - a(\theta, \theta'))$, the next sample in the chain is simply set to the current state, $\theta_{n+1} = \theta_n$. It is well-known that in the limit $N \to \infty$, the distribution of samples θ_i converges to the target posterior distribution $p(\theta|x_o)$ [Meyn et al., 2009]. Many variations of MCMC have been proposed and studied, where the proposal distribution $q(\theta'|\theta)$ and acceptance probability $a(\theta, \theta')$ are altered.

MCMC methods, as well as other sequential Monte Carlo approaches [Gordon et al., 1993, Liu et al., 1998, Doucet et al., 2001], have been widely and successfully applied to solve challenging Bayesian inference problems. Despite this, these approaches face some limitations. First, while these approaches are known to converge to the target posterior in the limit of infinite sampling steps $N \to \infty$, in practice the number of steps required for a good approximation of the posterior distribution to be reached can be very large. Thus, approximating the posterior distribution can be computationally expensive or even infeasible, especially when computing the likelihood function $p(x|\theta)$ is computationally expensive. In addition, classical methods estimate the posterior distribution $p(\theta|x_o)$ for a fixed value of x_o. In other words, the inference algorithm needs to be restarted from scratch for each new observation $x_o^{(1)}$. This can be computationally infeasible in applications where there are many observations for which we wish to estimate posterior distributions.

The advent of machine learning, and in particular of artificial neural networks, has revolutionized computational Bayesian inference. The study of training artificial neural networks to approximate probability distributions is known as *variational inference* [Hoffman et al., 2013, Blei et al., 2017]. Given an unknown distribution $p(z)$ over a random variable $z \in \mathcal{Z}$, the goal of variational inference is to train a variational distribution, $q_\phi(z)$, parameterized by $\phi \in \Phi$, to approximate $p(z)$ by minimizing the Kullback-Leibler (KL) divergence,

$$\phi^* = \arg\min_{\phi \in \Phi} \mathrm{KL}(q_\phi(z)||p(z)). \tag{2.11}$$

As an intuitive example, $q_\phi(z)$ can be a Gaussian distribution, $\mathcal{N}(z; \mu_\phi, \Sigma_\phi)$, with mean and covariance parameterized by ϕ. In practice, q_ϕ is a deep neural network, with a particular relevant architecture being a normalizing flow, which we describe in more detail in Sec. 2.3.2.

In the context of Bayesian inference, the corresponding problem becomes to estimate a distribution $q_\phi(\theta)$ to match the posterior distribution $p(\theta|x)$. However, the KL divergence $D_{\text{KL}}(q_\phi(\theta)||p(\theta|x_o))$ cannot be computed as the distribution $p(\theta|x_o)$ is not known. However, the KL divergence can be decomposed as

$$\begin{aligned} D_{\text{KL}}(q_\phi(\theta)||p(\theta|x_o)) &= \int_\Theta q_\phi(\theta) \log\left(\frac{q_\phi(\theta)}{p(\theta|x_o)}\right) d\theta \\ &= \int_\Theta q_\phi(\theta) \left[\log(q_\phi(\theta) - \log(p(x_o|\theta)p(\theta)) + \log(p(x_o))\right] d\theta, \end{aligned} \tag{2.12}$$

where we made use of Bayes' theorem to decompose the posterior $p(\theta|x_o)$. Note that the final term, $\int_\Theta q_\phi(\theta) \log(p(x_o)) d\theta = \log p(x_o)$ is a constant with respect to ϕ, and thus can be omitted from the optimization objective without changing the optimization problem for ϕ. Thus we can again ignore the probelmatic evidence term and instead optimize the tractable remaining terms in Eq. (2.12), known as the Evidence Lower BOund (ELBO)[5].

The generalization properties of neural networks mean that variational distributions as described above can produce good approximations of the posterior distribution with a relatively small number of evaluations of the likelihood function $p(x|\theta)$, and thus offer a powerful alternative to classical inference algorithms. Furthermore, the variational distribution can be made explicitly dependent on x, $q_\phi(\theta|x)$, and an equivalent ELBO loss can be derived for the expectation of the KL divergence over $p(x)$ [Rezende and Mohamed, 2015], so that once trained, the variational distribution $q_\phi(\theta|x)$ can be evaluated for any observation x_o without any further training. This property is known as *amortization*, and in certain applications is a significant advantage of variational methods for posterior inference.

2.2.2 Intractable likelihood functions

The statistical inference methods described previously all rely on the numerical evaluation of the likelihood function. In many cases, scientific models are defined in terms of computational simulators: Computer programs which take θ as input and output a sample x from the likelihood function. The ability of the simulator to also return the value of the likelihood function for this sample is not guaranteed. In general, a simulator produces a sequence of latent random variables $z_1, ..., z_K$ depending on θ[6] These can represent random observational noise, random initial or boundary conditions, or nuisance parameters which we do not wish to infer. The resulting likelihood can then be evaluated as

$$p(x|\theta) = \int p(x|z_{1:N}, \theta) p(z_{1:N}|\theta) dz_{1:N}. \tag{2.13}$$

This integral may be solved in special cases, namely the case where the only source of noise is an observational noise of known form. However, in many cases, this integral cannot be evaluated in closed form, and the resulting likelihood is said to be intractable. In such cases, to solve the inference problem, we must use *likelihood-free* methods, also known as simulation-based inference.

[5] The ELBO is defined as the negative of the remaining terms, as it is a lower bound of the model evidence. It is defined as ELBO $= -\mathbb{E}_{q_\phi}[\log q_\phi(\theta) - \log p(\theta|x_o)] = \mathbb{E}_{q_\phi}[\log p(\theta|x_o) - \log q_\phi(\theta)] \leq p(x_o)$

[6] $K = 0$ is possible and corresponds to a deterministic simulator, and K can also be dependent on θ.

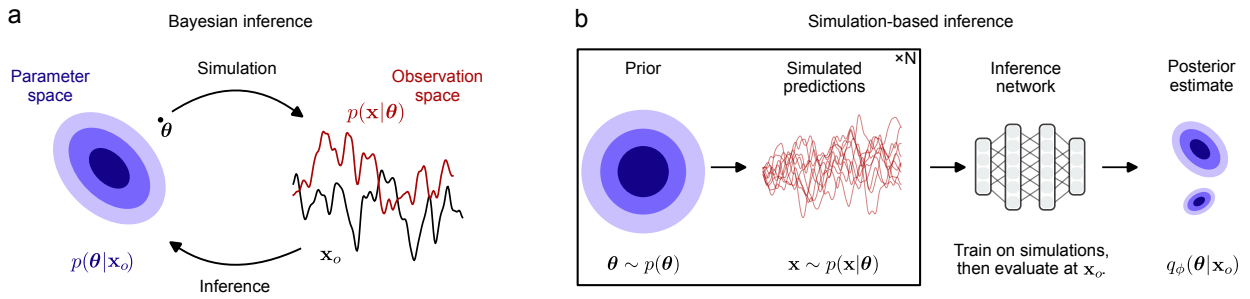

Figure 2.1: Overview of Bayesian inference and simulation-based inference (SBI). **a:** Bayesian inference is an approach for solving inverse problems. We are given a statistical model defining the likelihood $p(x|\theta)$ of some predictions x given any values of the model parameters θ, as well as a prior distribution $p(\theta)$ encoding the prior knowledge over the probably values of θ. Bayesian inference returns the posterior distribution $p(\theta|x_o)$ of the parameters given a particular observation x_o. **b:** SBI approximates Bayesian inference by estimating the posterior distribution $p(\theta|x_o)$. SBI generates a training dataset of parameter-prediction pairs (θ_i, x_i) by sampling the prior and simulating. This training dataset is used to train a conditional generative model q_ϕ. Given any observation x_o, the generative model predicts the posterior distribution $p(\theta|x_o)$ directly, or by enabling other computational Bayesian methods (Sec. 2.2.1). After Deistler et al. [2025].

2.2.3 Approximate Bayesian computation

An intuitive approach for statistical inference when we can only sample from the likelihood function (but not evaluate it) is to use alternative notions of "goodness-of-fit", as in the deterministic view of inverse problems (Sec. 2.1). By defining a mismatch $\mathcal{L}(\theta, x_o)$ that we can compute from simulations, we can replace the likelihood function by $\mathcal{L}$ in inference algorithms. This is known as Approximate Bayesian Computation (ABC) [Sisson et al., 2018].

The classical approach of Rejection ABC is to sample $\theta' \sim p(\theta)$ and $x' \sim p(x|\theta')$, and define the mismatch metric as some distance $d(x', x_o)$ (e.g. the L^2 norm).[7] The sample x' is accepted (meaning it is considered a sample from the approximate posterior) if $d(x', x_o) < \epsilon$ for some predefined threshold ϵ. Repeating this process yields an approximate posterior $\hat{p}(\theta|x_o)$, however, this procedure can yield very low acceptance rates, and therefore many simulations are required in order to estimate the approximate posterior. Other ABC approaches build on ideas from computational Bayesian inference to improve the simulation-cost of approximate inference [Marjoram et al., 2003, Sisson et al., 2007, Beaumont et al., 2008].

2.3 Neural simulation-based inference

An alternative and scalable approach for SBI is to use simulations as training data in order to train a variational distribution, q_ϕ. This variational distribution can be used to either estimate the likelihood itself, $q_\phi(x|\theta) \approx p(x|\theta)$, after which the computational Bayesian inference approaches (Sec. 2.2.1) can be used to estimate the posterior distribution. Alternatively, the variational distribution can estimate the posterior distribution directly, $q_\phi(\theta|x) \approx p(\theta|x)$. In both cases, the challenge is to design a suitable architecture for q_ϕ, combined with a suitable loss function $\mathcal{L}$, such that q_ϕ indeed converges to the desired distribution. Such approaches are similar to variational inference, but crucially, the loss function cannot make use of the now intractable likelihood function $p(x|\theta)$.

[7] note that in this case, the mismatch metric $\mathcal{L}(\theta, x_o)$ is a random variable as it depends on the random sample x'.

I begin by describing *Neural Posterior Estimation* methods (Sec. 2.3.1). I then describe two particularly powerful generative modeling frameworks for representing the variational distribution q_ϕ, namely *Normalizing flows* (Sec. 2.3.2) and *score-/flow-matching* (Sec. 2.3.3). These models can also be applied for *Neural Likelihood Estimation* (Sec. 2.3.4). Finally, I describe *Source Distribution Estimation* (Sec. 2.3.5), which is a distinct yet related problem to likelihood-free inference.

2.3.1 Neural posterior estimation

Neural Posterior Estimation (NPE, Papamakarios and Murray [2016], Lueckmann et al. [2017], Greenberg et al. [2019], Radev et al. [2022]) describes a family of methods to estimate the posterior distribution, $p(\theta|x)$. NPE defines a variational distribution, $q_\phi(\theta|x_o)$, which is optimized to approximate the true posterior distribution $p(\theta|x_o)$. A key observation made by Papamakarios and Murray [2016] is that by maximizing the *reverse* KL divergence between the variational and true posterior distributions, one can derive a tractable loss function. This follows from the following argument:

$$\begin{aligned} D_{\text{KL}}(p(\theta,x)|q_\phi(\theta|x)p(x)) &= \mathbb{E}_{(\theta,x)\sim p(\theta)p(x|\theta)}\left[\frac{\log(p(\theta)p(x|\theta))}{\log(q_\phi(\theta|x)p(x))}\right] \\ &= \mathbb{H}[p(\theta,x)] - \mathbb{E}_{(\theta,x)\sim p(\theta)p(x|\theta)}[\log p(x)] - \mathbb{E}_{(\theta,x)\sim p(\theta)p(x|\theta)}[q_\phi(\theta|x)] \\ &= -\mathbb{E}_{(\theta,x)\sim p(\theta)p(x|\theta)}[q_\phi(\theta|x)] + \text{const}, \end{aligned} \tag{2.14}$$

where the constant in the final line is with respect to the variational parameters ϕ, and $\mathbb{H}$ denotes the entropy. Thus, maximizing the probability the variational distribution assigns to samples $(\theta, x) \sim p(\theta)p(x|\theta)$ in fact minimizes the KL divergence which is zero if and only if $q_\phi(\theta|x) = p(\theta|x)$. Additionally, the loss function

$$\mathcal{L}_{\text{NPE}}(\phi) = -\mathbb{E}_{(\theta,x)\sim p(\theta)p(x|\theta)}[q_\phi(\theta|x)] \tag{2.15}$$

can be minimized without knowing the likelihood $p(x|\theta)$, making it suitable for simulation-based inference tasks. In practice, the expected value in Eq. (2.15) is approximated using a fixed dataset of draws from the simulator, $\{\theta_i, x_i\}_{i=1}^{N_{\text{sim}}}$.

One of the key advantages of NPE is that it is an amortized inference method. However, in order to train the variational posterior distribution q_ϕ to convergence, many simulations are required. Thus, a fruitful line of research has developed in improving the performance of NPE methods, primarily with the goal of reducing the number of simulations required to accurately estimate posterior distributions. Notable advancements have been the adoption of normalizing flows (Sec. 2.3.2) as the variational distribution, as well as embedding high-dimensional predictions x using an embedding network to learn effective summary statistics [Lueckmann et al., 2017, Greenberg et al., 2019]. There has also been fruitful work in developing interpretable evaluation metrics to measure the quality of the learned posterior distributions, many of which are reliant on the amortization property of NPE [Talts et al., 2020, Hermans et al., 2022, Lemos et al., 2023, Linhart et al., 2023, Cabezas et al., 2025]. When amortization of the posterior is not needed, *sequential* NPE methods can improve the simulation-efficiency of inference by using intermediate estimates of the posterior distribution to generate simulations as opposed to the prior distribution [Lueckmann et al., 2017, Greenberg et al., 2019, Deistler et al., 2022a].

2.3.2 Normalizing flows

The key insight behind many generative models, including normalizing flows [Kobyzev et al., 2019, Papamakarios et al., 2021], is that in order to sample from an unknown distribution, $\theta \sim p(\theta)$[8], it is sufficient to learn a transformation, $\theta = T(z)$, where z follows some simple distribution from which we can sample from, such as the unit Gaussian $z \sim \mathcal{N}(0, I)$. If the transformation is correct, then $p(\theta)$ is the *pushforward* distribution of $p(z)$ under T. To sample from $p(\theta)$, one can sample $z \sim p_0(z)$ and compute $T(z)$. To evaluate the probability of a given sample θ under $p(\theta)$, one computes $z = T^{-1}(x)$, and then evaluates

$$p(\theta) = |\det J_{T^{-1}}| p_0(z), \tag{2.16}$$

where $J_{T^{-1}}$ is the Jacobian matrix of the inverse transformation T^{-1}, and the absolute determinant $|\det J_{T^{-1}}|$ is a necessary volume correction term for the equality to hold.

In the deep learning context, we parameterize the transformation with some learnable weights, T_ϕ, and wish to optimize ϕ. This can be done via maximum likelihood training[9] [Papamakarios et al., 2021],

$$\phi^* = \arg\min_\phi -\mathbb{E}_{\theta \sim p(\theta)} |\det J_{T_\phi^{-1}}| p_0(T_\phi^{-1}(x)). \tag{2.17}$$

The remaining challenges to training T_ϕ is to ensure that T_ϕ is invertible with inverse T_ϕ^{-1} for all values of ϕ, and furthermore that both the inverse T_ϕ^{-1} and the absolute determinant of the Jacobian of that inverse, $|\det J_{T_\phi^{-1}}|$, can be efficiently computed. If they cannot be efficiently computed, then training can become prohibitively slow. Overcoming these challenges is the main contribution of normalizing flows.

Normalizing flows parameterize the transformation T_ϕ as a sequence of transformations,

$$T_\phi = T_{\phi_1} \circ T_{\phi_2} \circ \ldots \circ T_{\phi_L}. \tag{2.18}$$

Each transformation T_{ϕ_i} is defined as a strictly monotonic function such that it always has a lower-triangular Jacobian, with resulting determinant $\det J_i = \prod_{j=1}^N J_i^{jj}$, and that the overall Jacobian determinant is the product of the Jacobian determinants of all individual transformations,

$$\det J_{T_\phi^{-1}} = \prod_{i=1}^{L} \det J_i. \tag{2.19}$$

2.3.3 Score matching and flow matching

Neural SBI is a conditional distribution estimation task. Therefore, advances in conditional distribution estimation, and generative modeling more broadly, have often led to new methods for SBI [Geffner et al., 2022, Wildberger et al., 2023, Linhart et al., 2024, Gloeckler et al., 2024, Vetter et al., 2025]. Particularly of note is the work of Ramesh et al. [2022], who used generative adversarial networks to perform simulation-based inference for the shallow water equations in oceanography. Recently, score-based models [Hyvärinen, 2005, Sohl-Dickstein et al., 2015, Song et al., 2021] (also known as diffusion models), and flow matching models [Lipman et al., 2023] have become go-to

[8] Or conditional distributions, such as posteriors $p(\theta|x)$

[9] Thus, normalizing flows can be trained directly using the NPE loss in Eq. (2.15)

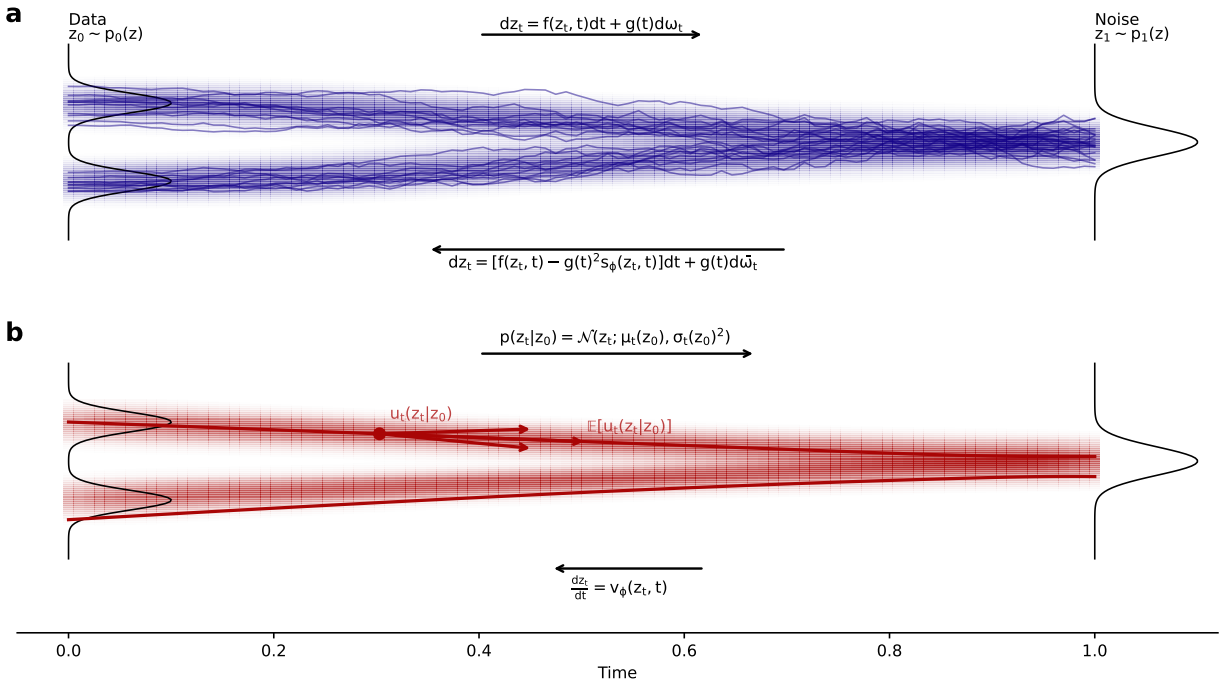

Figure 2.2: Denoising generative models. a: In diffusion models, the data $z_0 = \theta \sim p(z_0)$ is gradually degraded into a pure noise distribution through the stochastic differential equation (2.20). By learning the score function $s_\phi(z_t, t) \approx \nabla_{z_t} p(z_t)$, the reverse stochastic differential equation (2.21) can be used to map pure noise samples z_1 back into data. **b:** Flow Matching also defines a (possibly identical) time-dependent forward process to map z_0 to z_1. However, flow matching seeks to learn the unconditional velocity field $v_\phi(z_t, t) \approx u_t(z_t)$ defined by the forward process. This velocity is learned by minimizing a mean square error loss to the *conditional* velocity $u_t(z_t|z_0)$ (Eq. 2.25). Once the velocity field is trained, data samples can be generated from pure noise using an ordinary differential equation.

generative modeling approaches due to their ability to estimate high-dimensional probability distributions including timeseries [Rasul et al., 2021, Vetter et al., 2024a], images [Ramesh et al., 2021, Rombach et al., 2021, Blattmann et al., 2022], videos [Ho et al., 2022] and more. Extensive work has been dedicated to different formulations and representations of diffusion models [Salimans and Ho, 2022, Karras et al., 2022, Song et al., 2023, Meng et al., 2023, Yang et al., 2024]. Here, I provide a brief introduction to diffusion models as described in Song et al. [2021] and of flow matching as in Lipman et al. [2023]. Note that to simplify notation and to match the descriptions of these works, I describe the unconditional generation case, where we wish to model a distribution $p(\theta)$. However, the work is trivially extended to conditional generation $p(\theta|x)$ by making learnable quantities also conditioned on any condition x.

Diffusion models define a mapping from the nontrivial data distribution over $z_0 = \theta$, $p(z_0)$, to a simple (almost always Gaussian) distribution over noise $z_1 = \epsilon \sim p(\epsilon) = \mathcal{N}(0, I)$. This mapping is achieved through a continuous-time stochastic process, defined as

$$dz_t = f(z_t, t)dt + g(t)dw_t, \tag{2.20}$$

where $f(z_t, t)$ and $g(t)$ are the drift and diffusion functions of the diffusion process, and w_t is Brownian noise. A classical result [Anderson, 1982] states that given the diffusion process of Eq. (2.20) mapping $p(z_0)$ to $p(z_1)$, there exists a reverse diffusion process mapping $p(z_1)$ to $p(z_0)$,

which follows the reverse diffusion process

$$dz_t = [f(z_t, t) - g(t)^2 \nabla_{z_t} \log p(z_t)]dt + g(t)d\bar{w}, \tag{2.21}$$

where $\bar{w}$ is also Brownian motion for time flowing from $t = 1$ to $t = 0$. Crucially, this process is fully known given access to the *score function* $\nabla_{z_t} \log p(z_t)$, which is typically intractable. Thus, the goal of diffusion models is to estimate the score function with a parameterized model, $s_\phi(z, t)$, which is learned by minimizing the loss

$$\mathcal{L}_{\text{score-matching}} = \mathbb{E}_{t\sim\mathcal{U}([0,1]),z_0\sim p(\theta),z_t\sim p(z_t|z_0)} \left[\lambda(t)||s_\phi(z_t, t) - \nabla_{z_t} \log p(z_t|z_0)||^2\right], \tag{2.22}$$

where $\lambda(t)$ is a weighting function (which can be freely chosen). Note that while the score function $\nabla_{z_t} \log p(z_t)$ is intractable, the marginal score $\nabla_{z_t} \log p(z_t|z_0)$ can be trivially computed as the marginal distribution $p(z_t|z_0)$ follows from the definition of the drift and diffusion function in Eq. (2.20). In practice, different choices for the drift and diffusion function exist, but they are always chosen to be sufficiently simple such that the marginal distributions $p(z_t|z_0) = \mathcal{N}(z_t; \alpha_t z_0, \sigma_t^2 I)$, where different choices exist for the schedules α_t, σ_t, corresponding to different underlying properties of the stochastic differential equation. Despite not regressing on the actual score function $\nabla_{z_t} \log p(z_t)$, the loss $\mathcal{L}_{\text{score-matching}}$ from samples is known to be minimized by the desired score function [Hyvärinen, 2005].

Often, diffusion models are parameterized not in terms of learning the score function $\nabla_{z_t} \log p(z_t)$, but rather by predicting the denoised z_0 from z_t directly, which is linked to the score function via Tweedie's formula [Tweedie, 1947, Efron, 2011, Kim and Ye, 2021]:

$$\mathbb{E}[z_0|z_t] = z_t + \sigma_t^2 \nabla_{z_t} \log p(z_t). \tag{2.23}$$

Using Tweedie's formula, $z_\phi(z_t, t) \approx \mathbb{E}[z_0|z_t]$ is optimized by minimizing the loss

$$\mathcal{L}_{x\text{-prediction}} = \mathbb{E}_{t\sim\mathcal{U}([0,1]),z_0\sim p(\theta),z_t\sim p(z_t|z_0)} \left[\frac{\lambda(t)}{\sigma_t^4}||z_\phi(z_t, t) - z_0||^2\right].^{10} \tag{2.24}$$

Flow matching [Lipman et al., 2023] methods are very closely related to diffusion models. We again consider a continuous process $\{z_t\}_{t\in[0,1]}$ with marginal distributions $z_0 \sim p(\theta)$ and $z_1 \sim \mathcal{N}(0, I)$. Instead of estimating a score function, or indeed a denoised sample z_0, the goal of flow matching is to learn a velocity field $v_\phi(z_t, t) \approx u_t(z_t)$, where $u_t(z_t)$ corresponds to the mean paths taken by samples z_t. While this quantity is not tractable, the conditional flow velocity $u_t(z_t|z_0)$ is analytically tractable for many definitions of the paths z_t. For example, it is common to define z_t as linear interpolations between the data and Gaussian noise, i.e. $z_t = (1 - t)z_0 + t\epsilon$, in which case $u_t(z_t|z_0) = z_t - z_0$. It has been shown [Lipman et al., 2023] that regressing on this conditional velocity with the conditional flow matching loss,

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t\sim\mathcal{U}([0,1]),z_0\sim p(\theta),z_t\sim p(z_t|z_0)} \left[||v_\theta(z_t, t) - u_t(z_t|z_0)||^2\right] \tag{2.25}$$

is indeed minimized by the $u_t(z_t)$. Once $v_\phi(z_t, t)$ has been learned, samples can be generated by solving the resulting ordinary differential equation from $t = 1$ to $t = 0$, starting from samples $\epsilon \sim \mathcal{N}(0, 1)$.

[10] Another common formulation is "ϵ prediction", where the network approximates the noise added to the original data z_0 to get the current state, z_t. It is related to x-prediction, since $\epsilon = (z_t - \alpha_t z_0)/\sigma_t$.

While the formulations are motivated differently, there are many equivalences between flow matching and score matching [Lipman et al., 2023, Albergo et al., 2023]. Heuristically, denoising models are powerful as they do not require us to predict data θ from noise ϵ in one step. Instead, denoising models learn to predict θ from noisy versions of θ, which is an easier task, making these models easier to train. However, the sequential sampling procedure can itself be computationally expensive, and is a limitation of these models. Alleviating this limitation is a key research area, with many solutions being proposed. Distillation and reflow models [Salimans and Ho, 2022, Liu et al., 2023] which train a second model to sample more efficiently than the original model. Consistency and mean flow models [Song et al., 2023, Geng et al., 2025] train denoising models on varying time domains to facilitate better long-range denoising. Conditional flow matching with optimal transport [Tong et al., 2023, 2024], aims to estimate the optimal transport mapping between noise and data distributions, making the resulting trajectories z_t smoother and thus easier to integrate.

2.3.4 Neural likelihood estimation

Neural Likelihood Estimation (NLE) methods train probabilistic generative models to learn the likelihood function, $p(x|\theta)$. By considering the likelihood as a distribution over x, as opposed to a function of θ, generative models can trivially be used in the likelihood-learning scenario—that is, the probabilistic generative model is now $q_\phi(x|\theta)$ [Wood, 2010, Papamakarios et al., 2019, Lueckmann et al., 2019, Boelts et al., 2022].

There are several reasons to learn the likelihood instead of the posterior distribution directly. First, a learned neural likelihood is in fact a surrogate model of the simulator. In many cases in geoscience, this equates to having a surrogate model that can be evaluated faster than the simulator, which is of scientific value in itself [Brinkerhoff et al., 2021, Jouvet et al., 2022, Jouvet, 2023, Finn et al., 2025]. Furthermore, neural likelihood models are often differentiable, which can be of additional use compared to the simulator itself, e.g. by enabling direct optimization with respect to the forward model. Second, the neural likelihood can be used to perform likelihood-based Bayesian inference, for example with MCMC or variational inference as described in Sec. 2.2. If only the latter application is of interest, a further alternative is Neural Ratio Estimation (NRE), which learns a model of the likelihood ratio, $r(x|\theta) = \frac{p(\theta|x)}{p(\theta)}$. [Mohamed and Lakshminarayanan, 2017, Thomas et al., 2016, Hermans et al., 2020, Durkan et al., 2020, Miller et al., 2022].

2.3.5 Source distribution estimation

Recall that given prior knowledge over the parameters θ and a likelihood model $p(x|\theta)$, the goal of Bayesian inference is to quantify the posterior distribution over the parameters given a *single* observation of the system, x_o. Most Bayesian inference methods can be extended to the case where we also have multiple observations x_o^i [Bardenet et al., 2017], but under the crucial assumption that these observations come from the same underlying parameter, θ_o. Conceptually, this represents repeating an observation under the same experimental conditions.

There are many scenarios, however, where this assumption is broken. That is, we have a collection of observations, $x_o^{(i)}$, each coming from the same likelihood model $p(x|\theta)$, but generated under (potentially) different θ_i. A typical setting in which this scenario arises is in the context of *populations*. Each individual i from the population follows the same mechanistic model, described by the likelihood $p(x|\theta)$, but is described by different parameters $\theta_i \sim p(\theta_i)$, where the distribution $p(\theta)$ is unknown. One such example arises in the context of detecting gravitational waves [Thrane and Talbot, 2019]. Each binary system follows the same underlying physics encoding the likelihood

function, but each system is described by different parameters θ describing the masses of the two black holes, as well as their position in the sky. Thus, by measuring observations $x_o^{(i)}$ from each individual i, we wish to infer the distribution of the parameters for the entire population, $p(\theta)$, which we refer to as the *source distribution*.

The Bayesian approach to solving such inference problems is to describe a hyperprior, $p(\xi)$, where the hyperparameters ξ describe the possible source distributions, $s(\theta|\xi)$. The resulting likelihood of x given ξ is described by

$$p(x|\xi) = \int p(x|\theta,\xi)d\theta = \int p(x|\theta)p(\theta|\xi)d\theta. \tag{2.26}$$

Since ξ describes the population, as opposed to individuals, now the observations $x_o^{(i)}$ are generated using the same ξ. Therefore, given the observations $\{x_o^{(i)}\}_{i=1}^N$, the posterior $p(\xi|\{x_o^{(i)}\}_{i=1}^N)$ can be computed, resulting in a distribution over θ,

$$p(\theta|\{x_o^{(i)}\}_{i=1}^N) = \int p(\theta|\xi)p(\xi|\{x_o^{(i)}\}_{i=1}^N)d\xi. \tag{2.27}$$

This approach, known as *Hierarchical Bayes* [Malinverno and Briggs, 2004, Bishop, 2006, Teh and Jordan, 2010], is only tractable for simpler models. A more common approach is to estimate the maximum likelihood value for the population parameters, ξ^*, using Eq. (2.26), giving rise to $p(\theta|\{x_o^{(i)}\}_{i=1}^N) \approx p(\theta|\xi^*)$. This approach is commonly referred to as *Empirical Bayes* [Robbins, 1956, Efron and Morris, 1972, Bishop, 2006].

As for Bayesian inference, traditional methods of source distribution estimation require a tractable likelihood to evaluate the objective function (Eq. (2.26)), and are hence not available for the likelihood-free simulator scenarios. One approach to mitigate this is to first train a likelihood surrogate as in NLE (Sec. 2.3.4), which can then be used to estimate a source distribution [Wang et al., 2019, Vandegar et al., 2020].

2.4 Function-valued statistics

Many quantities in geoscience are defined not as vectors $\theta \in \mathbb{R}^n$, but rather as functions that vary with spatial positions $\mathbf{X}$ and/or time t. In general, most functions arising in geoscientific applications can be written as functions $\theta : \mathbb{X} \to \mathbb{Y}$, where the domain $\mathbb{X}$ is either (a subset of) $\mathbb{R}^D$ for $D \in \{1, 2, 3\}$ or (a subset of) $\mathbb{R}^D \times \mathbb{R}^+$. $\mathbb{Y}$ can also be multi-dimensional—for example velocities in all spatial directions would correspond to a vector in $\mathbb{R}^3$. We denote the function space where θ is defined as $\mathcal{A}(\mathbb{X}; \mathbb{Y}) := \mathcal{A}$.

Such function spaces on continuous domains can be infinite-dimensional[11], in which case it is not possible to fully describe θ with a finite-dimensional vector of numbers. In this section I briefly introduce relevant concepts in machine learning for working with function-valued data.

2.4.1 Gaussian processes

In order to perform probabilistic inference on infinite-dimensional, function-valued variables, we must also define the notion of a distribution over infinite-dimensional quantities. In particular, a prominent class of distributions over function-valued variables is the *Gaussian Process* (GPs,

[11] For example, if $\mathbb{A}$ is the space of continuous functions

Williams and Rasmussen [2006]). For notational clarity, I only consider *single-output* GPs, that is, distributions over functions $f : \mathbb{X} \to \mathbb{R}$.

Gaussian Processes are defined in terms of two quantities:

1. The *mean* function $\mu : \mathbb{X} \to \mathbb{R}$ defines the mean of the distribution at any given point $\mathbf{X} \in \mathbb{X}$, and
2. The *kernel function* $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, which defines the covariance between the random variables $f(\mathbf{X})$ and $f(\mathbf{X}')$.

Gaussian Processes are denoted as $\theta(\sim \mathcal{GP}(\mu(\cdot), k(\cdot, \cdot))$. GPs then define a distribution over the function θ in the sense that given some locations $X = [\mathbf{X_1}, \mathbf{X}_2, \dots \mathbf{X}_N]$ in the domain, the vector of values of θ at these positions, $\theta(X) = [\theta(\mathbf{X}_1), \dots, \mathbf{X}_N]$ is normally distributed. More precisely,

$$\theta(X) \sim \mathcal{N}(\theta(X); \mu(X), k(X, X)), \tag{2.28}$$

where $m(X) = [\mu(\mathbf{X}_1), \dots, m(\mathbf{X}_N)]$, and

$$k(X, X) = \begin{bmatrix} k(\mathbf{X}_1, \mathbf{X}_1) & \dots & k(\mathbf{X}_1, \mathbf{X}_N) \\ \vdots & \ddots & \vdots \\ k(\mathbf{X}_N, \mathbf{X}_1) & \dots & k(\mathbf{X}_N, \mathbf{X}_N) \end{bmatrix}. \tag{2.29}$$

Intuitively for geophysical functions, the kernel $k(\mathbf{X}, \mathbf{X}')$ typically captures the continuity, or variability, of the function. Typical choices for the kernel function are the square exponential kernel

$$k(\mathbf{X}, \mathbf{X}') = \exp\left(\frac{-||\mathbf{X} - \mathbf{X}'||_2^2}{2l^2}\right), \tag{2.30}$$

for some length scale l, and the Matérn$-\nu$ kernel

$$k(\mathbf{X}, \mathbf{X}') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}||\mathbf{X} - \mathbf{X}'||_\mathbf{2}}{l}\right)^\nu K_\nu \left(\frac{\sqrt{2\nu}||\mathbf{X} - \mathbf{X}'||_\mathbf{2}}{l}\right), \tag{2.31}$$

where l is the lengthscale, ν is a positive real number, Γ is the Gamma function, and K_ν is the modified Bessel function. In general, the only restriction is that the kernel must be positive semi-definite, so it always yields valid covariance matrices for any vector of positions X.

Gaussian Processes maintain many of the useful properties of Gaussian distributions, such as closed-form marginalization and conditioning. Additionally, the flexibility in the choice of the kernel allows for a variety of geophysical prior knowledge to be incorporated into the distribution. In the context of this thesis, GPs will be used as prior distributions for geophysical inverse problems, as well as tractable base distributions for generative modeling of function-valued parameters. However, note that the useful properties of GPs also mean that they have wide-ranging applications. Classical applications in the geosciences include regression and classification[12] [Oliver and Webster, 1990, Cressie, 1990, Christianson et al., 2022]. Approximations of Gaussian Processes, for example by finding sparse approximations [Snelson and Ghahramani, 2005], low-rank approximations via inducing points [Titsias, 2009], state-space approaches Sarkka et al. [2013], or variational inference [Hensman et al., 2013] have enabled the application of GPs even to extremely high-data regimes. GPs can also be extended to the multi-output scenario, either by treating the individual outputs independently, or by constructing covariance kernels that also explicitly model how the different outputs covary [Álvarez and Lawrence, 2011].

[12] where Gaussian Processes are also referred to as kriging

2.4.2 Deep learning architectures for function-valued data

Many geophysical simulators are defined in terms of ordinary or partial differential equations. Mathematically, these simulators operate on functions, however in practice they are typically solved using approximate computational methods such as finite difference or finite element solvers. Such computational approaches only require the value of the function $\theta(\mathbf{X}, t)$ on a predefined grid, $\{\mathbf{X}_i, t_i\}_{i=1}^M$. Therefore, the function is computationally represented using a fine-dimensional quantity $\theta_i \in \mathbb{Y}^M$. In principle, any neural network architecture can be applied to this discretized representation, such a discretization scheme typically yields large vectors, matrices, or tensors. Therefore, specialized architectures employing inductive biases are employed to reduce the network sizes required to model these data.

Convolutional neural networks and tokenization

Convolutional Neural Networks (CNNs, LeCun et al. [1989], Krizhevsky et al. [2012], LeCun et al. [2015]) are prominently used in the context of functions defined on a fixed discretization of the domain. CNNs perform convolutions using small, learnable kernels k. This produces a feature map that highlights the presence of learned patterns at different spatial locations. Because the same kernel is applied everywhere, CNNs use far fewer parameters, making them more efficient for large, structured inputs such as continuous functions. CNNs stack several learnable convolution operations, with initial layers learning the fine-structure of the data, and the latter learning large-scale patterns. However, a disadvantage of CNNS is that they require a fixed input shape: A neural network trained using one discretization cannot be applied to the same parameter discretized differently. This is not a strong limitation in the context of many machine learning problems, e.g. image data, where the data format is highly standardized. However, geoscientific domains can be vastly different. Different measurements of the same system cannot always be made at the same positions in space and/or time due to operational constraints.

An alternative approach to convolutions is *tokenization* of the function. For example, Deep Sets [Zaheer et al., 2017] embed each value $\theta(\mathbf{X}_i)$, together with its position $\mathbf{X}_i$ using some embedding network $h_\phi(\theta, \mathbf{X}) : \mathbb{Y} \times \mathbb{X} \to \mathbb{R}^L$, where L is a fixed latent dimension. The embeddings of all the points in the discretization are then aggregated into one vector $\rho \in \mathbb{R}^L$[13]. The aggregate latent embedding ρ can then be used to predict desired outputs using a decoder network g_ψ. For example, it can be used to predict the value of θ at an unseen location X_j ($g_\psi(\rho, X_j) \in \mathbb{Y}$, i.e. interpolation), or to predict a class label for the parameter ($g_\psi(\rho) \in \{0, 1\}$, i.e. classification). Neural Processes [Garnelo et al., 2018a,b] extend the Deep Set formalism to probabilistic interpolation, by using a generative model for the decoder g_ϕ, with the latent embedding ρ as a context.

A more expressive alternative to Deep Sets and Neural Processes are Transformers [Vaswani et al., 2017]. Similarly to the former, Transformers embed each value $\theta(\mathbf{X}_i)$ (optionally together with its position $\mathbf{X}_i$) into a latent vector ρ_i of fixed size L. However, as opposed to aggregating the resulting tokens, transformers operate on all the tokens ρ_i in an *attention block*. In the case that we only have one input stream[14], the learned embeddings ρ_i are further transformed into key, query, and value matrices (K, Q, V respectively). The output of the attention block is then defined as

$$\text{Attention} = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{2.32}$$

[13] The aggregation function is typically taken to be an average over the individual embeddings.

[14] This is known as self-attention, which is distinct from cross-attention, where there are two input streams.

where d_K is the dimension of the key embeddings. Transformers stack several of these attention blocks. Between each block there is a nonlinear, token-wise transformation (normally a multilayer perceptron). Note that the attention operation (Eq.(2.32)) can be applied independently of the number of tokens in the key, query, and attention matrices, and the rest of the transformer architecture operates token-wise. Therefore, the transformer can also be applied flexibly to varying discretizations of the spatial domain. Furthermore, since it operates on all the tokens, transformers are more flexible models than aggregator-based models such as deep sets. However, this comes at increased computation and memory cost, as the attention operation scales with the number of tokens as $O(M^2)$. One way to resolve this limitation is to tokenize *features* of the function, as opposed to each point individually, for example by tokenizing patches [Han et al., 2021] or spectral powers [Buchholz and Jug, 2022].

Fourier neural operators

In general, neural network architectures for mapping between infinite-dimensional function spaces are known as neural operators [Kovachki et al., 2023, 2024]. Fourier Neural Operator (FNO, Li et al. [2021]) are particularly applicable to continuous functions. Unlike CNNs, which operate in physical space with local kernels, FNOs perform global convolutions in the Fourier domain. FNOs typically operate on a much smaller latent representation of the data, by simply discarding the high-frequency components of the function. While this is a strong inductive bias, it is satisfied by many function-valued data in the geosciences, which satisfy strong smoothness constraints. Given an input function θ discretized on some grid $\mathbf{X}$, a single FNO block operates on $\theta(\mathbf{X})$ applies a learnable linear transformation to the lower frequency modes, adds a learnable linear residual connection, and applies a nonlinearity. Mathematically:

$$z = \mathcal{F}(\theta(\mathbf{X})), \tag{2.33}$$

$$z'_k = \begin{cases} R_k\, z_k, & \text{if } |k| \le K, \\ z_k, & \text{otherwise} \end{cases} \tag{2.34}$$

$$\xi(\mathbf{X}) = \mathcal{F}^{-1}(z'), \tag{2.35}$$

$$\theta'(\mathbf{X}) = \sigma\left(\xi(\mathbf{X}) + W\theta(\mathbf{X})\right) \tag{2.36}$$

Here K is the number of retained modes, R_k are learnable complex matrices, and W is a linear map in physical space[15]. Multiple FNO blocks are stacked to learn complex mappings, with the Fourier-domain multiplication giving efficient global receptive fields. The Fourier basis is essentially a fixed and compact basis to represent functions, which can be efficiently and differentiably computed using the Fast Fourier Transform [Cooley and Tukey, 1965]. However, other basis functions can be used, especially when the Fourier basis may not be sufficient, such as using Wavelet operators [Tripura and Chakraborty, 2023], or using learnable latent spaces, such as with transformers [Hao et al., 2023].

2.5 Modeling ice sheets and ice shelves

The Antarctic Ice Sheet is a dynamic and active component of the global climate. It is crucial in maintaining the global circulation of ocean currents [Goosse and Fichefet, 1999], as well as

[15] it is common to set W as a pointwise (1×1) transformation

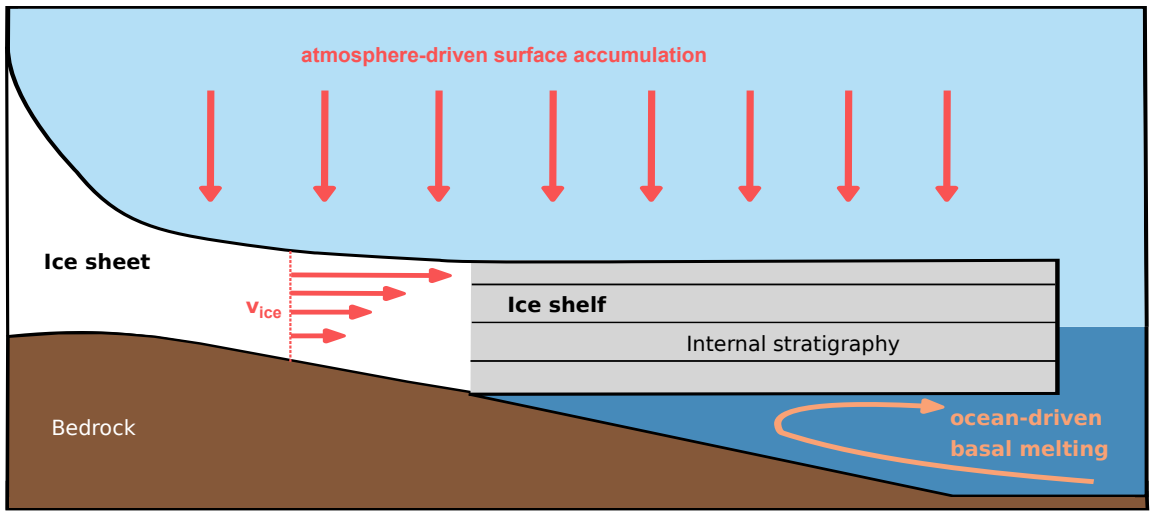

Figure 2.3: Cross-section of an Antarctic ice shelf. Ice flows outwards from the continental ice sheet, buttressed by the ice shelf. The ice flow influences the internal stratigraphy of the ice shelf and is itself influenced by atmosphere-driven surface accumulation and ocean-driven basal melt. The stratigraphy, consisting of layers of constant age, can be detected as Internal Reflection Horizons (IRHs) in Ground-Penetrating Radar (GPR) measurements.

influencing global temperatures through high albedo snow coverage [Munneke et al., 2011]. The Antarctic Ice Sheet is vulnerable to rising global temperatures and contains ice equivalent to approximately 58 meters of sea level rise [Morlighem et al., 2020]. Finally, the Antarctic Ice Sheet is also of interest as it is an archive of past climate information, and current estimates suggest that the oldest ice in Antarctica is over 10^6 years old [Parrenin et al., 2017]. The study of Antarctica is therefore multifaceted, with much attention given to the projection of its future extent, identification of areas which are vulnerable to climate change, the interaction of Antarctica with the atmosphere and ocean, and the probing of past climates using Antarctic ice records.

Modeling the dynamics of the Antarctic Ice Sheet and its ice shelves enables to better understand its past, as well as make predictions about its future. The primary dynamics of ice are those of fluid flow. Specifically, ice is assumed to be an incompressible, viscous, non-Newtonian fluid. Due to the slow speed of ice flow, second order effects of inertia and acceleration can be neglected, resulting in the Stokes equation of ice flow,

$$\nabla \cdot \sigma = -\mathbf{F}. \tag{2.37}$$

Here, σ is the **stress tensor** of the ice, and $\mathbf{F}$ are the external forces acting on the ice, which is typically $\mathbf{F} = \rho_{\text{ice}}\mathbf{g}$, where ρ_{ice} is the density of the ice and $\mathbf{g}$ is gravitational acceleration [Greve and Blatter, 2009, Hooke, 2019]. The stress tensor of ice can be related to the ice velocity, which allows ice flow to be modeled computationally. Namely, the stress tensor is given by $\sigma = -pI + \tau$, where p is the known isotropic pressure of ice, and τ is the deviatoric stress. Using Glen's flow law, the deviatoric stress is given by

$$\dot{\epsilon} = A\tau^n, \tag{2.38}$$

where A is the fluidity, n is Glen's exponent [16]. The strain rate tensor, $\dot{\epsilon}$, has components

$$\dot{\epsilon}_{ij} = \frac{1}{2}\left(\frac{\partial U_i}{\partial X_j} + \frac{\partial U_j}{\partial X_i}\right), \tag{2.39}$$

where $\mathbf{U}$ is the ice velocity and $\mathbf{X}$ are the Cartesian coordinates.

Despite knowledge of these governing equations, modeling ice flow can still be a significant challenge. In part, this is because of the sheer extent of Antarctic ice sheets, spanning thousands of kilometers horizontally, and up to four kilometers vertically. To model ice flow on a sufficiently fine resolution on such a spatial extent is difficult even for comparatively short timescales. For this reason, significant work has been devoted to the computational modeling of the Stokes equation [Winkelmann et al., 2011, Larour et al., 2012, Gagliardini et al., 2013]. Still, a single simulation of the Stokes flow equation with these models can span several days on supercomputers [Fischler et al., 2022, Bueler, 2023], which poses a challenge to exploring various parameter configurations. Concretely, statistical inference with models of this complexity is typically intractable. Several approximations of the Stokes flow, such as the Shallow Ice Approximation and the Shallow Shelf Approximation [Greve and Blatter, 2009], have been successfully applied to speed up ice flow models where these approximations are valid. Additionally, recent works focus on designing easy to use and interpretable models. For example, Shapero et al. [2021] design a flexible and composable framework allowing for more user experimentation, whereas Verjans et al. [2022] incorporate uncertainties into a stochastic ice flow model.

The challenging nature of modeling ice flow is not only due to computational complexity, but also to unknown variables. Our knowledge of the current state of the ice sheet is mostly limited on values at the ice surface, which are accessible through remote sensing [Morlighem, 2022, Gardner et al., 2022]. However, ice flow depends also on sub-surface quantities, such as the temperature profile $T(z)$ of the ice, the roughness of the ground at the ice base, or the melting of ice at its interface with the ocean [Greve and Blatter, 2009, Hooke, 2019]. These quantities can typically only be measured pointwise[17], as opposed to throughout the ice sheet, and therefore need to be estimated in order to perform an ice flow simulation, which can lead to downstream errors. Thus, there exists a large body of works into the inference of ice-dynamic quantities that are not directly measurable [Neckel et al., 2012, Riel et al., 2021, Brinkerhoff et al., 2021, Jezek et al., 2022, Hoffman et al., 2022, Riel and Minchew, 2023].

[16] Glen's exponent is typically assumed to be $n = 3$, Greve and Blatter [2009], but recent work argues it should be $n = 4$, e.g. Bons et al. [2018]

[17] For example, through ice core drilling [Brook and Buizert, 2018]

Chapter 3

Publications

This thesis develops and applies new techniques to enable simulation-based inference for geoscientific inference problems. In this chapter, I discuss the three main publications that are part of this thesis, with the full publications provided in the Appendix. All these works were in collaboration with different colleagues, and so I also provide a breakdown of my individual author contributions for each work separately.

3.1 Statement of contributions

Simulation-based inference of surface accumulation and basal melt rates of an Antarctic ice shelf from isochronal layers

Published as: "Simulation-based inference of surface accumulation and basal melt rates of an Antarctic ice shelf from isochronal layers"—**Guy Moss**, Vjeran Višnjević, Olaf Eisen, Falk M. Oraschewski, Cornelius Schröder, Jakob H. Macke and Reinhard Drews in the Journal of Glaciology (2025).

This publication was co-authored by **GM**, VV, OE, FMO, CS, JHM and RD. The idea to apply SBI to infer basal melting rates of ice shelves was conceptualized by JHM and RD. **GM** developed this idea further. The forward model was implemented by **GM**, with input from VV and RD. The SBI application was implemented by **GM**, with input from CS and JHM. The main experiments were run by **GM**. Data processing and curation was done by FMO, OE and RD. **GM**, CS and RD wrote the first draft of the manuscript. All coauthors contributed in editing and refining the manuscript. The main supervision in this project was from CS, JHM and RD.

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

Published as: "FNOPE: Simulation-based inference on function spaces with Fourier Neural Operators"—**Guy Moss**, Leah Sophie Muhle, Reinhard Drews, Jakob H. Macke and Cornelius Schröder in the 39th Conference on Neural Information Processing Systems (NeurIPS 2025).

This publication was co-authored by **GM**, LSM, RD, JHM and CS. **GM** initialized the idea of performing SBI for function-valued parameters by training Fourier Neural Operators in a flow-matching approach. The method was implemented by **GM** and LSM. Experiments were run by **GM**, LSM, and CS. The initial draft was written by **GM**, LSM, RD, and CS. RD and JHM provided feedback throughout the project, as well as editing of the manuscript. The main supervision of this project was from JHM and CS.

Sourcerer: Sample-based Maximum Entropy Source Distribution Estimation

Published as: "Sourcerer: Sample-based Maximum Entropy Source Distribution Estimation"—Julius Vetter*, **Guy Moss***, Richard Gao, Cornelius Schröder and Jakob H. Macke in the 38th Conference on Neural Information Processing Systems (NeurIPS 2024)

This paper was co-authored by JV, **GM**, CS, RG, and JHM. JV had the initial idea to target maximum entropy source distributions using a sample-based approach. JV and **GM** together formalized this idea in terms of a constrained optimization problem, and proved the uniqueness of the maximum entropy source distribution. **GM** worked out the connection to the average posterior and proved that average posterios were not always valid source distributions. JV performed the initial implementation of the code and the analysis pipeline for the synthetic tasks. JV and **GM** implemented the application to the Hodgkin-Huxley model. JV and **GM** wrote the initial draft of the paper. CS, RG, and JHM provided the feedback on the project throughout, and reviewed and edited the initial version of the paper.

3.2 Simulation-based inference of surface accumulation and basal melt rates of an Antarctic ice shelf from isochronal layers

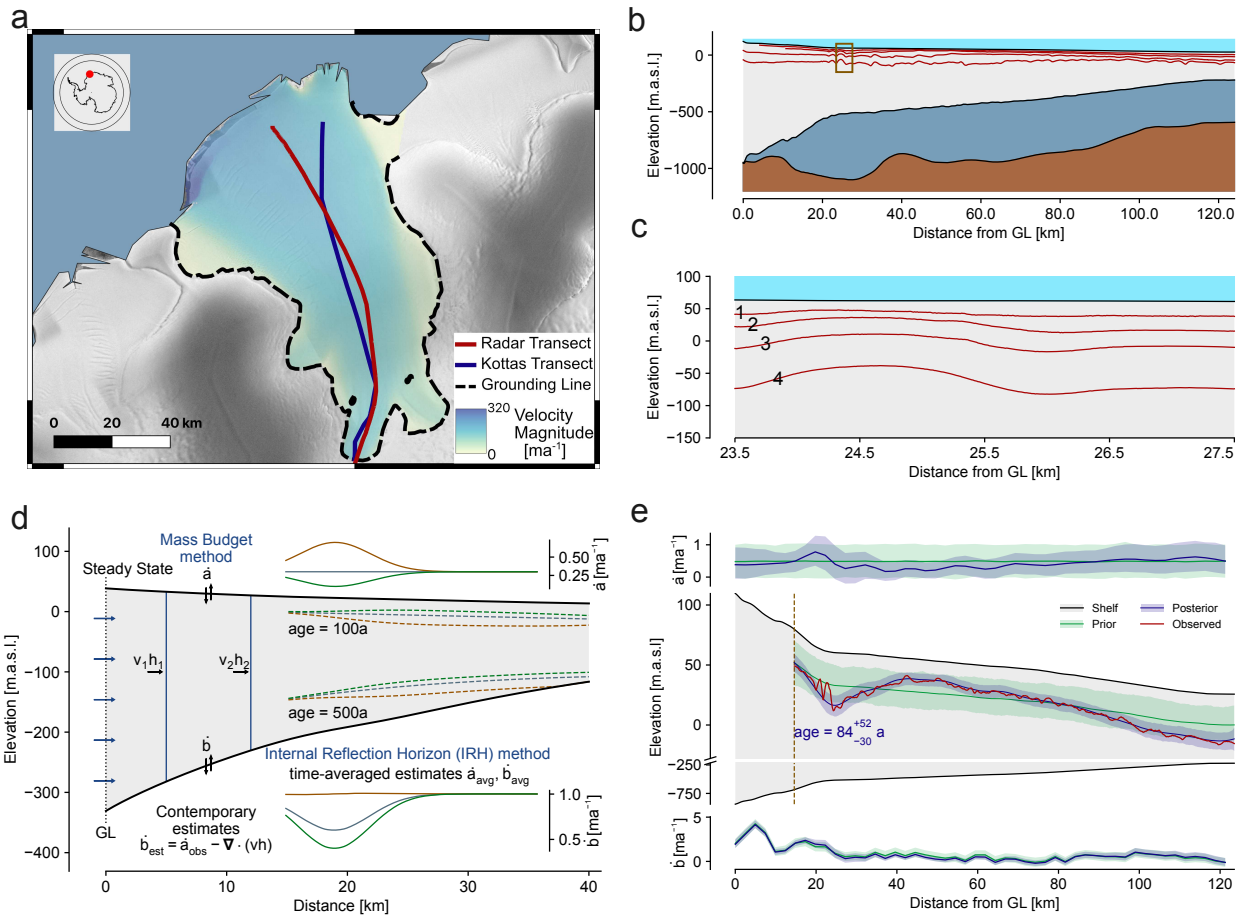

Figure 3.1: Inferring the Mass Balance Parameters of Ekström Ice Shelf, Antarctica. a: Radar transect where internal reflection horizon (IRH) data was collected on Ekström Ice Shelf, Antarctica. **b** Visualization of labeled IRHs. **c:** Zoom in on box in panel b. **d:** Schematic of different methods of estimating mass balance parameters of ice shelves. Traditional methods infer the total mass balance (surface accumulation rate - basal melt rate) from the steady state thickness of the ice shelf. Knowledge of the IRHs can separate the effects of surface accumulation rate and basal melt rate. **e:** Prior and learned posterior over the surface accumulation and basal melt rates (top and bottom, respectively). Posterior is conditioned on IRH 2. Middle panel shows prior and posterior predictive distributions compared to IRH 2. Panels adapted from Moss et al. [2025b].

Motivation

Antarctic ice shelves are floating ice masses surrounding the continent. While they only account for 11% of the total Antarctic surface area [Andreasen et al., 2023], they buttress the grounded ice sheet, thereby increasing its stability [Bindschadler et al., 2011]. The extent of the ice shelf itself is dependent on the incoming ice flow from the continent, the accumulation of ice through snowfall

on the ice shelf surface, and the ocean-driven melting of ice at the base. The latter atmospheric- and oceanic-driven factors are particularly of interest as they are in effect unknown boundary conditions on ice flow [Winkelmann et al., 2012, Gudmundsson et al., 2019], which can have a significant effect on predictions of the future state of the ice shelf [Bindschadler et al., 2013].

Inferring the surface accumulation rate and basal melt rate, collectively known as the mass balance parameters, from empirical observations has been the subject of many existing works [Waddington et al., 2007, Catania et al., 2010, Steen-Larsen et al., 2010, Wolovick et al., 2021, Višnjević et al., 2022, Theofilopoulos and Born, 2023]. In particular, the mass balance parameters can be inferred from the internal stratigraphy of the ice shelves. The internal stratigraphy is the internal structure of the shelves, which consists of layers of constant age of deposition. These layers are shaped by ice flow (Sec 2.5), which is in turn dependent on the mass balance parameters. Notably, the internal stratigraphy can be measured using ground-penetrating radar (GPR, Schroeder et al. [2020]), where layers of constant deposition appear as internal reflection horizons (IRHs). Therefore, to infer the mass balance parameters of an ice shelf, we should first define a simulator predicting linking them to the IRHs (Fig. 3.1d).

A limitation of existing methods of inferring the mass balance parameters is that they all produce point estimates, i.e., they do not quantify the uncertainty in the estimated mass balance parameters. These uncertainty estimates are vital as they also influence uncertainty estimates of projections of the future state of the ice shelf. In this work, we address this limitation by developing a framework to approach this task statistically, and then solve the resulting inference problem with SBI.

Methods

Our first contribution is to develop a simulator which is capable of generating ice shelf internal stratigraphies of ice shelves given prescribed surface accumulation rates $\dot{a}$ and basal melt rates $\dot{b}$. We circumvent computationally expensive ice flow models (Sec. 2.5) by adapting a tracer method [Born, 2017, Born and Robinson, 2021]). In this work, we consider a vertical cross-section of the ice shelf following an ice flowline (Fig. 3.1a-c), and parameterize the horizontal coordinate along the ice shelf by X. The tracer method splits this vertical cross-section into segments, each with varying thickness along the cross-section, $\{H_1(X), ..., H_L(X)\}$, for some number of layers L. Each layer $H_l(X)$ is then evolved using the advection equation. The advection equations are independent and can therefore be solved in parallel. The mass balance parameters are used to add and remove mass from the top and bottom layers respectively at each time step. We make the forward model stochastic by defining a physically-grounded noise model. This model accounts for uncertainty from observational noise, but crucially also for the high-frequency features of internal stratigraphies which are not captured by the tracer method, which produces smooth internal layers. This model was able to produce realistic stratigraphies in approximately 1 minute of computation time on a single CPU, enabling us to perform many simulations to tackle the inference problem.

Using this model, we define the resulting statistical inference problem. We define the parameters, which are the values of the surface accumulation rate $\dot{\mathbf{a}}(X)$ on a predefined grid of 50 positions along the flowline domain. Assuming the ice shelf is in steady state, we can compute the resulting basal melt rate values $\dot{\mathbf{b}}(X)$ corresponding to any particular sample of $\dot{a}$. For a given radar measurement of an ice shelf, the IRHs are labeled manually. Since the labeling process is time consuming, only a few IRHs are labeled, and hence they cover the vertical span of the ice shelf. As a result, we define a separate inference problem per labeled IRH x_j^o. We define a prediction from our forward model to be the layer elevations of the closest layer (in terms of Euclidean distance)

to x_j^o. This choice has the distinct advantage that solving the inference problem for the different observed IRHs also provides a time signal of how the mass balance parameters changed over time.

We define a prior distribution over the surface accumulation rates $\dot{a}$ as a Gaussian Process with a Matérn kernel (Sec. 2.4.1). The mean function and kernel of the Gaussian process are motivated by past observations of surface accumulation rates on Antarctic Ice Shelves. We then use NPE (Sec. 2.3.1) to solve the resulting inference problems. In particular, we first design a synthetic experiment on an idealized ice shelf to validate that our approach can indeed recover the ground truth. We then use this approach to infer the surface accumulation rates and basal melt rates on a flowline in Ekström Ice Shelf, Antarctica (Fig. 3.1), using new measurements of its internal stratigraphy.

Results

We first validate our approach on the synthetic ice shelf example. We observe that the ground truth surface accumulation and basal melt rates are indeed within the 95% confidence intervals of our posterior distribution. Furthermore, the posterior predictive simulations produce internal layers which are much closer to the observation than prior predictive simulations. Despite this strong reduction in predictive uncertainty, the distribution over the mass balance parameters is notably still broad. This shows that there is no unique solution, and even constrained by IRH measurements there can be significant uncertainty in the mass balance parameters of an ice shelf. We also estimate the age of the synthetic IRHs by simulating predictive simulations and measuring the ages of the best-fit layers. In the synthetic example, we observe that our posterior predictive ages are well-calibrated to the ground truth ages. This supports the claim that our approach can additional be used to estimate the ages of the IRHs without time-intensive ice coring.

We then move on to real IRH data from Ekström Ice Shelf, Antarctica. Here, we also observe a significant decrease in the predictive error from the prior predictive to the posterior predictive distributions (Fig. 3.1e). While no ground truth mass balance parameters are known for Ekström Ice Shelf, we compare our results to previous studies of its mass balance parameters, and note that our posteriors exhibit similar trends in the spatial dependence of both the surface accumulation and basal melt rates as what was reported in those studies. We predict the ages of the observed IRHs using our posterior predictive distribution, with the deepest layer estimated at ≈ 200a.

In conclusion, we develop a framework to infer the mass balance parameters of Antarctic ice shelves from measurements of their internal stratigraphies. In contrast to existing works, our approach also estimates the uncertainties in the mass balance parameters. *Our results support the claim that there is a need for these uncertainties*, as there exists a wide range of mass balance parameters that are consistent with the observed internal stratigraphies of ice shelves. Due to the large number of parameters (50 values of $\dot{a}$), our SBI approach requires a large number of simulations: $189,000$ training simulations were generated for each of the synthetic and Ekström ice shelf tasks. This large simulation budget is enabled by the forward model developed in our work. An advantage of our NPE approach, compared to methods that require sequential sampling, is that the majority of the computational cost is amortized. The same simulations could be reused in solving the inverse problems for each of the individual observed IRHs. This reduced the computational cost of our approach by four (the number of observed IRHs in the dataset).

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

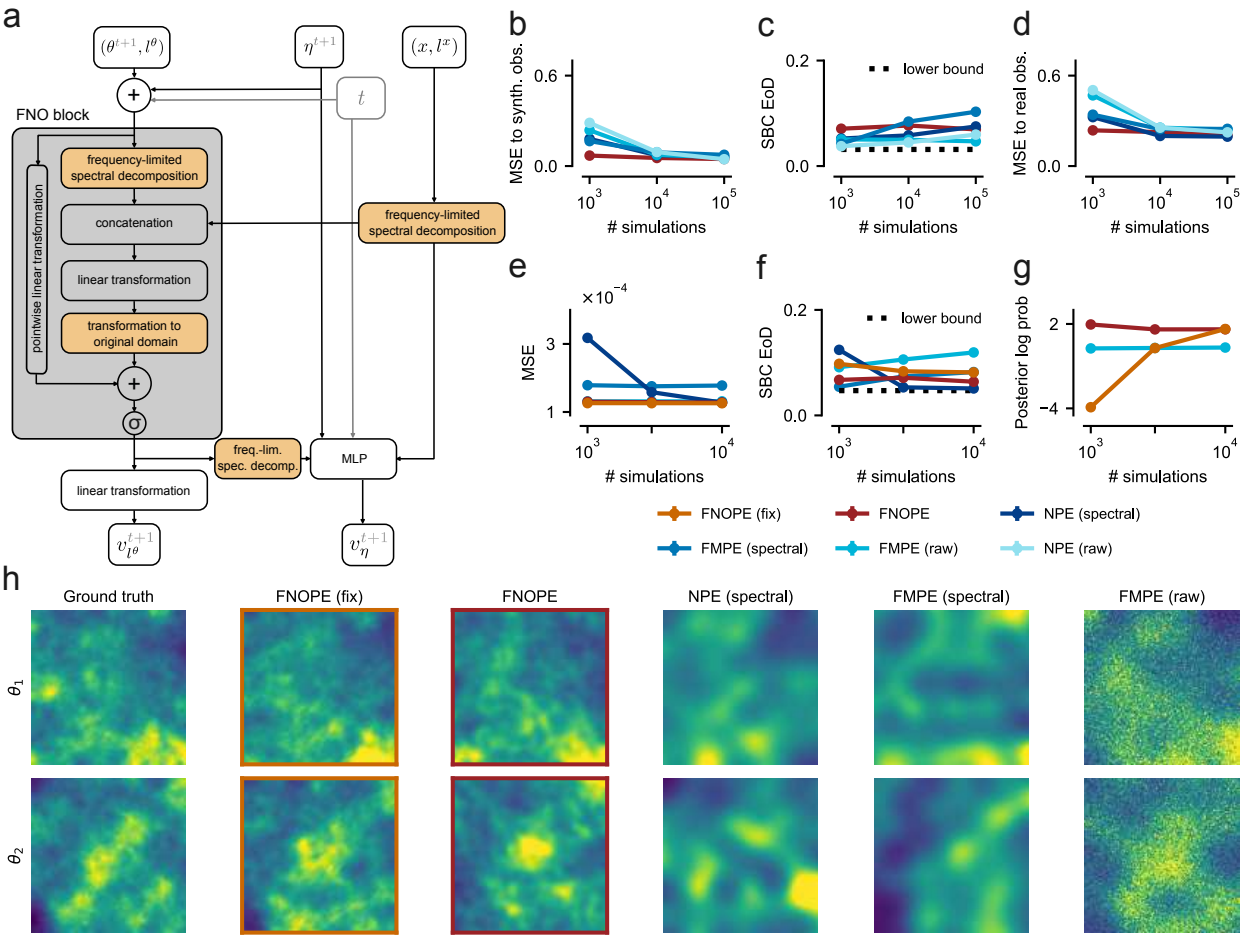

Figure 3.2: FNOPE: Simulation-based inference of function-valued parameters. a: Schematic of FNOPE architecture. The network estimates the flow field for both function-valued and vector-valued parameters, which is used to sample from the posterior distribution using flow matching. **b-d:** Benchmarking of FNOPE and baseline methods on mass balance parameter task for several simulation budgets. Note that FNOPE (fix) is not applied on this task as the parameter and observation domains are distinct. **e-g:** Benchmarking of FNOPE and FNOPE (fix) on Darcy flow task for several simulation budgets. Note that NPE (raw) is not applied on this task due to computational constraints. **h:** Samples from the posterior over Darcy flow permeability parameters, trained with 10^4 simulations for each of the methods. Panels adapted from Moss et al. [2025a].

Motivation

In the geosciences, the inference of function-valued parameters is a common task. Typically, the function domains are spatial, temporal, or both. Applying existing SBI methods to the inference of function-valued parameters suffers from several disadvantages. First, they require fixing a representation of the function-valued parameter for both training and sampling stages. This necessitates

a predetermined basis expansion, which may not be sufficiently expressive, or a fixed discretization, meaning the parameters cannot be inferred at other locations without introducing additional interpolation error. Second, these fixed representations of the function-valued parameters typically yield many coefficients to infer, and therefore require a (prohibitively) large number of training simulations.

One property of function-valued parameters that may be used to alleviate these challenges is that these parameters often exhibit structure, specifically strong spatial/temporal correlations. This structure implies that while the parameters are high-dimensional, the latent dimensionality is lower. Exploiting this latent dimensionality can therefore make the inference easier than operating on a predetermined representation of the parameters. In this work, we seek to develop a new approach for simulation-based inference tailored to inferring function-valued parameters. Our approach operates in function space, thereby avoiding the challenges that come from choosing a fixed representation of these parameters. Furthermore, by incorporating knowledge of the smoothness constraints on function-valued parameters, we develop an efficient approach that facilitates inference using considerably fewer simulations than previous work. Our approach trains **F**ourier **N**eural **O**perators for **P**osterior **E**stimation (FNOPE).

Methods

We formulate the task of learning a distribution over function-valued θ as learning a distribution over the value of θ evaluated at an arbitrary set of points $l^\theta \in \mathcal{D}_\theta$, where $\mathcal{D}_\theta$ is the function domain. We also assume that x is function-valued and observed on a set of points $l^x \in \mathcal{D}_x$, where the domain $\mathcal{D}_x$ may differ from D_θ. Therefore, l^x may differ from l^θ and may contain a different number of points. This formulation is particularly useful for geophysical measurements, where operational constraints mean that the positions at which x can be measured (l^x) are unpredictable and may differ from the positions at which we wish to infer the parameters. We extend the definition of the prior distribution as the conditional distribution of a measure on the function space, μ, at the finite set of positions, l^θ, i.e., $\theta \sim p_{l^\theta}$. Similarly, the simulator returns predictions following the likelihood $p_{l^x}(x|\theta, l^\theta)$. Given this problem formulation, the goal is to estimate the posterior, $p_{l^\theta}(\theta|x, l^x)$, given sets of positions, l^θ and l^x.

We address this problem by using Fourier Neural Operators (FNOs, Sec. 2.4.2). Since FNOs operate on a fixed number of spectral modes of the input data, they can be applied regardless of the number of points in the discretizations l^θ and l^x of the parameters and observations, respectively. In addition, to ensure that we can apply our approach to non-uniform discretizations l^θ, l^x of the respective domains, we augment the FNOs by using the type II non-uniform discrete Fourier transform [Greengard and Lee, 2004, Lingsch et al., 2024]. Furthermore, we design an FNO architecture that explicitly conditions on the discretizations l^θ, l^x (Fig. 3.2a), which allows the network to learn to compensate for errors in the non-uniform discrete Fourier transform arising from different discretizations for test and training samples.

To train the FNO architecture as a probabilistic generative model, we use a flow matching posterior estimation (FMPE) scheme (Sec. 2.3), where the FNO parameterizes the learnable vector field $v^\phi_{l^\theta}$. We make further additions to the FMPE scheme to extend the applicability, robustness, and performance of FNOPE. First, we add masking and positional noise during training to the discretizations l^θ and l^x of the simulations. This improves the robustness of FNOPE when sampling posterior distributions at different discretizations to those used in simulations. Second, we change the base distribution in the flow-matching scheme from a unit Gaussian to a Gaussian Process (GP) with a square exponential kernel (Sec. 2.4.1). We set the length scale of this GP to produce

samples with smoothness matching the smoothness implied by the highest FNO spectral mode. This modification reduces the amount of white noise artifacts in generated samples. Finally, to make our method applicable to simulators containing both function- and vector-valued parameters, we extend the framework and architecture to also allow for the estimation of vector-valued parameters η. This results in the combined velocity field, $v^\phi = [v^\phi_{l^\theta}, v^\phi_\eta]$, which is trained with the loss:

$$\mathcal{L}_{\text{FNOPE}} = \mathbb{E}_{t\sim\mathcal{U}[0,1],(l^\theta,\theta,l^x,x,\eta)\sim S,\xi_t\sim p_{l^\theta}(\xi_t|\theta),z_t\sim p(z_t|\eta)}||v^\phi - u_t||^2. \tag{3.1}$$

Here, $S = \{\theta_i, l^\theta_i, x_i, l^x_i, \eta_i\}_{i=1}^N$ is the simulation dataset and t is the flow time, z_t is the linear interpolation between the vector-valued parameters η and a sample from a unit Gaussian $z_1 \sim \mathcal{N}(0,1)$, ξ_t is a convex combination between the real parameter θ and a sample from the Gaussian process ξ_1. Following Liu et al. [2023], the true velocity of the rectified flow between the base and data distribution is

$$u_t = \begin{bmatrix} \xi_t - \theta \\ z_t - \eta \end{bmatrix}. \tag{3.2}$$

Results

We benchmark FNOPE against established methods, namely NPE with normalizing flows, and a Multilayer Perceptron-based FMPE as a flow matching baseline without the modifications made in our approach. Since we consider tasks with high numbers of parameters, we anticipate that these baselines will struggle to correctly estimate the posteriors, especially in the low-simulation-budget regime. We therefore also compare against spectral variants of NPE and FMPE, where the parameters are represented by their first M spectral modes. The spectral NPE and FMPE variants then estimate the posterior over the coefficients of these modes, thus significantly reducing the dimensionality of the parameter space.

We also demonstrate the improvement of our approach on a real-world inference task by revisiting the Ekström Ice Shelf inference task from Sec. 3.2 (Fig. 3.2b-d). We first evaluate the Mean Square Error (MSE) between the posterior predictive simulations from each method to a synthetic IRH. We observe that all methods converge to a similar performance given sufficiently many simulations. However, for smaller simulation budgets the methods inferring the parameters on a fixed discretization (denoted "raw") perform significantly worse than the methods inferring the spectral coefficients (denoted "spectral", here 10 modes are used). Furthermore, FNOPE significantly outperforms even these spectral baselines, appearing to converge to the posterior using only 10^3 simulations. We also verify that the learned posteriors are calibrated using Simulation-based calibration (SBC, [Talts et al., 2020]) on the posterior marginal distributions. The calibration is summarized for each method and simulation budget by the error of the diagonal (EoD). We see that all methods are reasonably well-calibrated, indicating that the posterior learned using FNOPE is not significantly more over- or underconfident than the baseline methods. Finally, we also measure the posterior predictive MSE when conditioning the trained networks on a real IRH measured in Ekström Ice Shelf, where the model may be misspecified. As for the synthetic observation, we observe that FNOPE outperforms all methods at low simulation budgets.

We also explore the performance of FNOPE on a higher-dimensional problem, by inferring a function-valued parameter that varies in two spatial dimension in the Darcy Flow model. The Darcy Flow model is a PDE with many applications. One application is to model the spatial distribution of groundwater as a function of the hydraulic permeability. 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 and $u(X)$ is the hydraulic potential. We simulate the forward model on a resolution of 129×129, leading to a total of $>$16k parameters to infer. Despite the large number of parameters, we observe (Fig. 3.2h) that FNOPE produces visually faithful samples to ground truth parameters, while the baseline methods cannot. As for the previous task, we again observe (Fig. 3.2e-g) that in terms of posterior predictive MSE, FNOPE outperforms the baseline methods, especially for lower simulation budgets. The posterior learned by FNOPE is reasonably well-calibrated. Finally, we also measure the log-probability (per pixel) of the ground truth samples under the posterior learned by FNOPE, as another measure of posterior performance [Papamakarios and Murray, 2016, Greenberg et al., 2019, Durkan et al., 2020, Lueckmann et al., 2021]. Here, we can only compare against the non-spectral FMPE method, as the spectral methods do not model the parameters directly and thus cannot assign them a log-probability. In terms of this metric, we again see that FNOPE outperforms the baseline method. Overall, our results demonstrate that FNOPE is a flexible, simulation-efficient approach for simulation-based inference of function-valued parameters.

3.4 Sourcerer: Sample-based Maximum Entropy Source Distribution Estimation

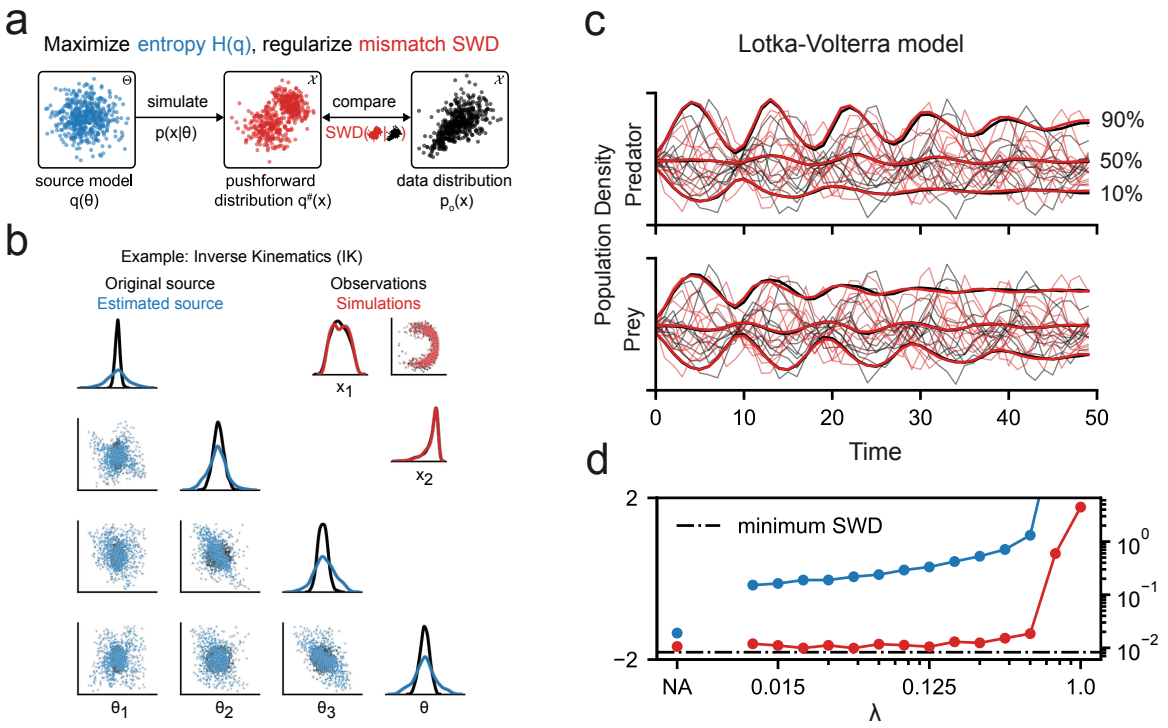

Figure 3.3: Sourcerer: Sample-based, maximum entropy source distribution estimation. a: Method schematic. We estimate a source distribution q_ϕ with maximum entropy such that its pushforward $q_\phi^\#$ matches a given observed data distribution $p_o(x)$. **b:** Sourcerer finds a higher-entropy source distribution than that used to generate the data for a benchmark task, even though the pushforward distributions match. **c:** Predictive distributions on the Lotka-Volterra task from source distributions estimated using Sourcerer and the ground truth source. **d:** Sliced-Wasserstein distance (lower is better) between simulations and observations, and entropy of estimated sources (higher is better) for the Lotka-Volterra task. Sources are estimated using different regularization strengths λ and without the entropy regularization (NA). Sourcerer is able to learn higher-entropy sources without sacrificing predictive quality. Panels adapted from Vetter et al. [2024b].

Motivation

The source distribution estimation problem (Sec. 2.3.5), also known as Empirical Bayes, is a related yet distinct paradigm to Bayesian inference. The source distribution estimation problem is to infer a distribution of parameters consistent with a dataset of observations. This is in contrast to a single observation or independently and identically distributed (i.i.d.) observations. A distinctive challenge inherent in the source distribution estimation problem is its ill-posedness. For a sufficiently degenerate forward model, different source distributions over the parameters can give rise to the same distribution over the predictions. Consequently, the source distribution consistent with some observed dataset may not be unique.

A common application of source distribution estimation is in the estimation of prior distributions from data. When such datasets are available, they can be used to learn informed prior distributions,

thereby facilitating downstream inference problems. The non-uniqueness of the source distribution estimation problem is a particular concern in this case, as the learned prior should still have coverage of all feasible parameters that can be consistent with the observations. In this work, we propose to estimate the maximum entropy source distribution. The maximum entropy principle [Good, 1963, Jaynes, 1968] is a principled approach to the selection of distributions, as it aligns with the intuitive concept that the solution should be "maximally ignorant", that is to say, as broad as possible. We show that the maximum entropy source distribution is unique, and propose a simulation-based method to estimate it.

Methods

Given a distribution $q(\theta)$ over the parameters, and the (potentially intractable) likelihood model $p(x|\theta)$ we define the *pushforward* of q as

$$q^{\#}(x) = \int p(x|\theta)q(\theta)d\theta. \tag{3.3}$$

For the source distribution estimation task, we are given a dataset $\mathcal{D} = \{x_i\}_{i=1}^N$ from some unknown distribution over the observations, $x_i \sim p_o(x)$. A source distribution q is defined as a distribution which satisfies $q^{\#} = p_o$. This condition can alternatively be stated as $D(q^{\#}, p_o) = 0$ for any valid distance metric $D(\cdot, \cdot)$. In this work, we want to find the source distribution q which also maximizes the entropy $H(q)$ among all source distributions (Fig. 3.3a). In practice, we consider a variational family of source distributions q_ϕ which is trained by the regularized objective function

$$\mathcal{L}_{\text{sourcerer}} = \lambda H(q_\phi) - (1 - \lambda) \log(D(q_\phi^{\#}, p_o)) \tag{3.4}$$

for some regularization parameter $\lambda \in (0, 1)$. This regularized objective is only an approximation of the constrained optimization problem we try to solve. The minimizer of Eq. (3.4) may not satisfy the hard constraint $D(q_\phi^{\#}, p_o) = 0$, and therefore may only be an approximate source distribution. However, due to challenges such as model misspecification, and a limited variational family, the hard constraint cannot be satisfied for any distribution in practice. Therefore, we opt for this relaxation of the constrained optimization problem.

We choose the Sliced-Wasserstein Distance (SWD, Bonneel et al. [2015], Kolouri et al. [2019], Nadjahi et al. [2020]),

$$\text{SWD}(q_\phi^{\#}, p_o) = \mathbb{E}_{u \sim U(\mathbb{S}^{n-1})} \left[\frac{1}{N} \sum_{i=1}^{N} ||u^\top x_i - u^\top y_i||_2^2 \right]^{1/2} \tag{3.5}$$

as the distance metric D. Here, $x_i \sim p_o(x)$, $y_i \sim q_\phi^{\#}(y)$ are samples from the probability distributions and u are uniformly randomly sampled vectors on the unit sphere $\mathbb{S}^{D-1}$. One advantage of using the SWD is that we do not need to evaluate the likelihood $p(x|\theta)$, and thus this loss is appropriate for simulation-based methods. We additionally estimate the entropy $H(q_\phi)$ in a sample-based manner using the Kozachenko-Leonenko estimator [Kozachenko and Leonenko, 1987, Berrett et al., 2019]. Estimating the entropy from samples removes the constraint from the variational distribution q_ϕ to admit tractable log-probability computation. Overall, our sample-based approach requires only differentiable simulators, or, if the simulator is not differentiable, a differentiable surrogate model. This is in contrast to previous state-of-the-art approaches [Vandegar et al., 2020], which require a tractable and differentiable likelihood model (or surrogate). Therefore, Sourcerer can more readily be applied to a larger class of source distribution estimation problems.

Results

We compare Sourcerer against Neural Empirical Bayes [Vandegar et al., 2020], a state-of-the-art approach for source distribution estimation, on several benchmark tasks. Both approaches successfully estimate source distributions as measured by the pushforwards of the learned sources. However, Sourcerer consistently identifies higher entropy source distributions, thus highlighting both the non-uniqueness of the source distribution estimation problem and the ability of Sourcerer to find higher entropy distributions that are still consistent with the observations.

We also apply Sourcerer to simulators that produce high-dimensional observations in the form of timeseries. These settings are more similar to geoscientific applications than the low-dimensional benchmark tasks. Training a surrogate model to generate timeseries would be challenging in itself, but if the simulators are differentiable, Sourcerer circumvents the need to train a surrogate model altogether. In particular, we consider the Lotka-Volterra predator-prey model from ecology (Fig. 3.3c). We measure the quality of the learned source (Fig. 3.3d) via the SWD between the pushforward and the true observed distribution (red), and the entropy of the learned source (blue). We observe that the pushforward SWD is near the theoretical minimum, indicating that Sourcerer estimates a valid source distribution. Furthermore, when we set the regularization strength $\lambda > 0$ in the Sourcerer objective (Eq. (3.4)), meaning that we also try to maximize the entropy of the learned source, we observe that Sourcerer estimates higher entropy source distributions without considerably reducing predictive fidelity.

Overall, our results demonstrate that Sourcerer is a robust method for estimating high-entropy source distributions. An advantage of Sourcerer is that it does not require likelihood evaluations, so it can be applied more readily in likelihood-free scenarios. We apply Sourcerer to a synthetic example from ecology, and note its general applicability to various other domains. Learning source distributions from past experimental datasets allows us to eliminate many unrealistic parameter settings from our prior knowledge. Thus, using the learned sources as informed prior distributions for new experimental data, we can reduce the difficulty of downstream inference problems for these simulators.

Chapter 4

Discussion

Statistical inference is one of the primary tools for integrating experimental observations into simulators of geoscientific phenomena. As more experimental data becomes available and simulators become more complex, geoscientists will need a variety of algorithms to solve the increasingly challenging inference problems that arise. A remarkable variety of algorithms already exist for statistical inference, and they are tailored to different settings and model configurations. Among these, simulation-based inference (SBI) is a useful and broadly applicable recent addition capable of solving many inference tasks. SBI makes few assumptions about the simulator or the data and can learn high-dimensional, multimodal posterior distributions for simulators without a tractable likelihood function.

This thesis demonstrates how SBI can solve a significant and challenging problem in glaciology and provide meaningful scientific insights into the ice dynamic history of ice shelves (Sec. 3.2). This has the potential to improve future projections of Antarctic ice shelf progression. Furthermore, this thesis expands the applicability of SBI methods in geoscience. First, we introduce a new SBI method designed for function-valued parameters and demonstrate its superiority over traditional SBI approaches for such problems (Sec. 3.3). This work extends the applicability of SBI to geoscientific inference problems involving parameters that vary spatially and/or temporally. Second, we introduce a new simulation-based approach for estimating source distributions, a different type of inference problem (Sec. 3.4). One notable application of this method is learning informed prior distributions of geoscientific parameters, making downstream inference problems easier to solve. These advancements highlight SBI's potential to solve many previously intractable geoscience inference problems. However, challenges remain to be overcome for this promise to be fulfilled.

First, the fundamental challenge of SBI is to train (potentially large) generative models in the low-data regime, since we are often restricted by the simulator's computational cost. One way to address this challenge is to improve our generative models. There are many examples of advancements in generative models being successfully applied to SBI, including normalizing flows [Papamakarios et al., 2019, Greenberg et al., 2019], score- and flow-matching models [Geffner et al., 2022, Wildberger et al., 2023, Linhart et al., 2024, Gloeckler et al., 2024], and prior-data fitted networks [Vetter et al., 2025]. However, this is only a partial answer. Solving real-world problems presents new challenges. Some of these challenges are general, such as high-dimensional parameter spaces (as in the ice shelf case study, Sec. 3.2), or high simulator degeneracy, which leads to nontrivial correlations in the posterior distribution [Deistler et al., 2022b]. Conversely, these challenges may be specific to the inference task. For example, observations may contain missing or corrupted data [Lueckmann et al., 2017, Verma et al., 2025]. Additionally, simulators may

encounter numerical issues for certain parameter configurations [Bernaerts et al., 2025]. When aiming to solve real-world inference problems, it is important to develop a thorough understanding of the problem domain. Shifting the focus from general methodologies to specific inference problems can also be advantageous. Focusing on specific inference problems often allows us to make additional assumptions that can significantly improve the simulation-efficiency of our methods. In geoscientific inference problems, these assumptions can take many forms. For example, we employed a smoothness assumption in FNOPE (Sec. 3.3).

Second, a notable obstacle in inference is model misspecification. SBI methods in particular may produce unpredictable and inaccurate results under model misspecification [Ward et al., 2022, Cannon et al., 2022, Montel et al., 2024]. This is especially true for likelihood misspecification, where the implicit likelihood defined by the simulator does not match the actual relationship between the parameters and observations. One way to address this issue is through Generalized Bayesian Inference [Bissiri et al., 2013, Knoblauch et al., 2022]. These methods replace the likelihood function with an alternative loss function quantifying the fit of parameters to the observations. This targets a different distribution to the Bayesian posterior, which can nevertheless be a meaningful measure of uncertainty in the parameters. These generalized loss functions may focus on features that are less sensitive to the model misspecification, and thus be less affected by it. Generalized Bayesian Inference has also been applied in simulation-based scenarios [Gao et al., 2023, Matsubara et al., 2024, Kelly et al., 2025], where the generalized loss function is estimated from simulations. Another approach to alleviating likelihood misspecification involves regularizing the training of embedding networks to map simulations and observations to similar summary statistics [Huang et al., 2023]. This approach is particularly powerful when a calibration dataset is available from a higher-fidelity simulator or experimental measurements. In this case, optimal transport in the embedding space can be used to estimate a mapping between misspecified simulations and real observations. Thus, SBI models trained on misspecified simulations can be calibrated [Wehenkel et al., 2025, Senouf et al., 2025]. While these approaches are promising, they require strong assumptions about the inference problem. For example, one must choose an appropriate generalized loss function or have a high fidelity simulator. This makes these methods difficult to apply in many settings. More research is therefore needed to enable a general approach to SBI under model misspecification.

Third, a relatively under-explored line of research is improving the performance of SBI methods using auxiliary information from simulators. The core motivation behind SBI is to perform inference for simulators where the likelihood function is intractable. However, there may be auxiliary information that can be extracted from the simulator and that should, in principle, be useful for solving inference problems with fewer simulations. For example, Brehmer et al. [2020] showed that, if the entire latent trace z of the simulator is known, together with the conditioned likelihood $p(x|z, \theta)$, the training loss for NRE and NLE methods can be augmented to increase their simulation-efficiency. Zeghal et al. [2022] also used this information to improve the efficiency of NPE. Other works have explored improving SBI for multi-fidelity simulators, which allow for both high-fidelity, expensive simulations and lower-cost, less accurate simulations [Prescott and Baker, 2020, Warne et al., 2022, Krouglova et al., 2025]. Furthermore, an increasing number of computational simulators are written in automatically differentiable frameworks, which provide access to the gradient of the simulator prediction with respect to the parameters. Likelihood-based algorithms often use differentiable models to perform inference, for example in variational inference [Rezende and Mohamed, 2015, Kucukelbir et al., 2017] and particle-based variational inference [Liu and Wang, 2016, Liu et al., 2019]. Simulator gradients can also be incorporated into particle-based samplers for SBI [Simons et al., 2022, Dellaporta et al., 2022].Our approach (Sec. 3.4) for source distribution estimation also benefits from differentiable simulators because they enables it to estimate source

distributions without first training a surrogate model. Relatively few works, however, have explored the use of differentiable simulators for neural SBI without making additional assumptions (e.g. Brehmer et al. [2020]). Accelerating neural SBI with simulator gradients could enable inference for many tasks that are currently infeasible and should be the subject of future investigations.

Finally, once a posterior is learned, how can we incorporate the learned uncertainty into large-scale models? For instance, in our Antarctic ice shelf case study (Sec. 3.2), we inferred a posterior distribution over the mass balance parameters of Ekström Ice Shelf. These parameters in turn affect large scale models of the Antarctic ice sheet. However, it may be computationally infeasible to evaluate the resulting uncertainty in the model's predictions by simulating many samples from the inferred posterior. This is because large-scale models typically use many parameters and have high computational costs. Additionally, other sources of uncertainty exist, such as errors arising from finite discretizations, modeling approximations, and observational errors in input data, which are not always accounted for. Propagating uncertainties in large-scale geoscientific models is generally challenging. Estimating uncertainties in crucial quantities, such as sea level and global temperature projections, is typically done through large intercomparison projects that compare predictions from different modeling frameworks under different scenarios [Asay-Davis et al., 2015, Griffies et al., 2016, Dunne et al., 2024]. These projects aim to capture the range of possibilities, as opposed to making principled probabilistic predictions. On the other hand, recent works have developed fully stochastic simulators. These simulators can be derived from physical principles, typically at the cost of reduced model complexity or additional physical assumptions [Nicholls et al., 2021, Verjans et al., 2022, Madsen et al., 2022]. Alternatively, generative models can be trained to emulate stochastic simulators [Schmidt et al., 2025, Finn et al., 2025], but these models require a lot of data to train, and scaling them to larger domains is more difficult compared to their physically-derived counterparts. This may result in a scenario where scientists are presented with a binary choice between cheaper simulators that allow for uncertainty quantification, and large-scale and high-fidelity simulators that are deterministic by design. Many modeling frameworks allow scientists to increase model fidelity by incurring higher computational costs. Similarly, much can be gained from designing frameworks that allow scientists to flexibly trade computational cost for improved uncertainty quantification.

In conclusion, in this thesis I show that simulation-based inference is a tool with the potential to unearth a wealth of scientific insights in geoscientific applications. I demonstrate what can be achieved with available methods, as well as how these methods can be pushed further to reach this potential. I believe that future research will continue developing improved methods and solving real, meaningful inference problems in geoscience. This will make simulation-based inference a standard tool in the geoscientist's toolkit.

Bibliography

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *CoRR*, abs/2303.08797, 2023. doi: 10.48550/arXiv.2303.08797.

Denis Allard, Alessandro Comunian, and Philippe Renard. Probability aggregation methods in geoscience. *Mathematical Geosciences*, 44(5):545–581, 2012. doi: 10.1007/s11004-012-9396-3.

Mauricio A. Álvarez and Neil D. Lawrence. Computationally efficient convolved multiple output gaussian processes. *Journal of Machine Learning Research*, 12(41):1459–1500, 2011.

Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. doi: 10.1016/0304-4149(82)90051-5.

Julia R. Andreasen, Anna E. Hogg, and Heather L. Selley. Change in antarctic ice shelf area from 2009 to 2019. *The Cryosphere*, 17(5):2059–2072, 2023. doi: 10.5194/tc-17-2059-2023.

Xylar S. Asay-Davis, Stephen L. Cornford, Gaël Durand, Benjamin K. Galton-Fenzi, Rupert M. Gladstone, G. Hilmar Gudmundsson, Tore Hattermann, David M. Holland, Denise Holland, Paul R. Holland, Daniel F. Martin, Pierre Mathiot, Frank Pattyn, and Hélène Seroussi. Experimental design for three interrelated marine ice-sheet and ocean model intercomparison projects. *Geoscientific Model Development Discussions*, 8:9859–9924, 2015.

Deborah Ashby. Bayesian statistics in medicine: a 25 year review. *Statistics in Medicine*, 25(21):3589–3631, 2006. doi: 10.1002/sim.2672.

Peter Atkinson, Hester Jiskoot, Remo Massari, and Tavi Murray. Generalized linear modelling in geomorphology. *Earth Surface Processes and Landforms*, 23(13):1185–1195, 1998. doi: 10.1002/(SICI)1096-9837(199812)23:13<1185::AID-ESP928>3.0.CO;2-W.

Rémi Bardenet, Arnaud Doucet, and Chris Holmes. On markov chain monte carlo methods for tall data. *Journal of Machine Learning Research*, 18(47):1–43, 2017.

Mark A. Beaumont, Jean-Marie Cornuet, Jean-Michel Marin, and Christian P. Robert. Adaptive approximate bayesian computation. *Biometrika*, 96:983–990, 2008.

Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966. doi: 10.1126/science.153.3731.34.

Yves Bernaerts, Michael Deistler, Pedro J. Gonçalves, Jonas Beck, Marcel Stimberg, Federico Scala, Andreas S. Tolias, Jakob H. Macke, Dmitry Kobak, and Philipp Berens. Combined statistical-biophysical modeling links ion channel genes to physiology of cortical neuron types. *Patterns*, 2025. doi: 10.1016/j.patter.2025.101323.

Thomas B. Berrett, Richard J. Samworth, and Ming Yuan. Efficient multivariate entropy estimation via k-nearest neighbour distances. *The Annals of Statistics*, 47(1):288 – 318, 2019. doi: 10.1214/18-AOS1688.

Michael Betancourt. The convergence of markov chain monte carlo methods: From the metropolis method to hamiltonian monte carlo. *Annalen der Physik*, 531(3):1700214, 2019. doi: 10.1002/andp.201700214.

R. Bindschadler, H. Choi, A. Wichlacz, R. Bingham, J. Bohlander, K. Brunt, H. Corr, R. Drews, H. Fricker, M. Hall, R. Hindmarsh, J. Kohler, L. Padman, W. Rack, G. Rotschky, S. Urbini, P. Vornberger, and N. Young. Getting around antarctica: New high-resolution mappings of the grounded and freely-floating boundaries of the antarctic ice sheet created for the international polar year. *Cryosphere*, 5:569–588, 2011. doi: 10.5194/TC-5-569-2011.

Robert A. Bindschadler, Sophie Nowicki, Ayako Abe-Ouchi, Andy Aschwanden, Hyeungu Choi, Jim Fastook, Glen Granzow, Ralf Greve, Gail Gutowski, Ute Herzfeld, and et al. Ice-sheet model sensitivities to environmental forcing and their use in projecting future sea level (the searise project). *Journal of Glaciology*, 59(214):195–224, 2013. doi: 10.3189/2013JoG12J125.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Pier Giovanni Bissiri, Chris C. Holmes, and Stephen G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 78: 1103 – 1130, 2013.

Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773.

Jan Boelts, Jan-Matthis Lueckmann, Richard Gao, and Jakob H Macke. Flexible and efficient simulation-based inference for models of decision-making. *Elife*, 11, 2022. doi: 10.7554/eLife.77220.

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 Ziaeemehr, 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(108), 2025. doi: 10.21105/joss.07754.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015. doi: 10.1007/s10851-014-0506-3.

Paul D. Bons, Thomas Kleiner, Maria-Gema Llorens, David J. Prior, Till Sachau, Ilka Weikusat, and Daniela Jansen. Greenland ice sheet - higher non-linearity of ice flow significantly reduces estimated basal motion. *Geophysical Research Letters*, 45(13):6542–6548, 2018. doi: 10.1029/2018GL078356.

Andreas Born. Tracer transport in an isochronal ice-sheet model. *Journal of Glaciology*, 63:22–38, 2017. doi: 10.1017/JOG.2016.111.

Andreas Born and Alexander Robinson. Modeling the greenland englacial stratigraphy. *Cryosphere*, 15:4539–4556, 2021. doi: 10.5194/TC-15-4539-2021.

Johann Brehmer, Gilles Louppe, Juan Pavez, and Kyle Cranmer. Mining gold from implicit models to improve likelihood-free inference. *Proceedings of the National Academy of Sciences*, 117(10): 5242–5249, 2020. doi: 10.1073/pnas.1915980117.

Douglas Brinkerhoff, Andy Aschwanden, and Mark Fahnestock. Constraining subglacial processes from surface velocity observations using surrogate-based bayesian inference. *Journal of Glaciology*, 67(263):385–403, 2021. doi: 10.1017/jog.2020.112.

Edward J. Brook and Christo Buizert. Antarctic and global climate history viewed from ice cores. *Nature*, 558(7709):200–208, 2018. doi: 10.1038/s41586-018-0172-5.

Tim-Oliver Buchholz and Florian Jug. Fourier image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022.

Ed Bueler. Performance analysis of high-resolution ice-sheet simulations. *Journal of Glaciology*, 69 (276):930–935, 2023. doi: 10.1017/jog.2022.113.

Luben M. C. Cabezas, Vagner S. Santos, Thiago R. Ramos, Pedro L. C. Rodrigues, and Rafael Izbicki. Cp4sbi: Local conformal calibration of credible sets in simulation-based inference. *arXiv*, 2025. doi: 10.48550/arXiv.2508.17077.

Patrick Cannon, Daniel Ward, and Sebastian M. Schmon. Investigating the Impact of Model Misspecification in Neural Simulation-based Inference. *arXiv*, 2022. doi: 10.48550/arXiv.2209.01845.

Ginny Catania, Christina Hulbe, and Howard Conway. Grounding-line basal melt rates determined using radar-derived internal stratigraphy. *Journal of Glaciology*, 56(197):545–554, 2010. doi: 10.3189/002214310792447842.

Ryan B. Christianson, Ryan M. Pollyea, and Robert B. Gramacy. Traditional kriging versus modern gaussian processes for large-scale mining data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 16:488–506, 2022.

James W. Cooley and John W. Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19(90):297–301, 1965.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020. doi: 10.1073/pnas.1912789117.

Noel Cressie. The origins of kriging. *Mathematical Geology*, 22(3):239–252, 1990. doi: 10.1007/BF00889887.

Gokhan Danabasoglu, Jean-François Lamarque, Julio T. Bacmeister, D. A. Bailey, A. K. DuVivier, James Edwards, Louisa K. Emmons, John T. Fasullo, R. Garcia, Andrew Gettelman, Cécile Hannay, Marika M. Holland, William Large, P. H. Lauritzen, D. M. Lawrence, J. T. M. Lenaerts, K. Lindsay, William H. Lipscomb, M. J. Mills, Richard B. Neale, K. W. Oleson, Bette L. Otto-Bliesner, Adam S. Phillips, William J. Sacks, Simone Tilmes, Leo Kampenhout, Mariana Vertenstein, Alessia Bertini, J. Dennis, Clara Deser, Claude Fischer, Baylor Fox-Kemper, Jennifer E. Kay, Douglas E. Kinnison, P. J. Kushner, Vincent E. Larson, M. C. Long, S. Mickelson, James K. Moore, Eric Nienhouse, Lorenzo M. Polvani, P. J. Rasch, and Warren G. Strand. The community earth system model version 2 (cesm2). *Journal of Advances in Modeling Earth Systems*, 12(2), 2020. doi: 10.1029/2019MS001916.

Maximilian Dax, Stephen R. Green, Jonathan Gair, Jakob H. Macke, Alessandra Buonanno, and Bernhard Schölkopf. Real-time gravitational wave science with neural posterior estimation. *Physical Review Letters*, 127:241103, 2021. doi: 10.1103/PhysRevLett.127.241103.

Michael Deistler, Pedro J Goncalves, and Jakob H Macke. Truncated proposals for scalable and hassle-free simulation-based inference. In *Advances in Neural Information Processing Systems*, volume 35, 2022a.

Michael Deistler, Jakob H. Macke, and Pedro J. Gonçalves. Energy-efficient network activity from disparate circuit parameters. *Proceedings of the National Academy of Sciences*, 119(44), 2022b. doi: 10.1073/pnas.2207632119.

Michael Deistler, Jan Boelts, Peter Steinbach, Guy Moss, Thomas Moreau, Manuel Gloeckler, Pedro L. C. Rodrigues, Julia Linhart, Janne K. Lappalainen, Benjamin Kurt Miller, Pedro J. Gonçalves, Jan-Matthis Lueckmann, Cornelius Schröder, and Jakob H. Macke. Simulation-based inference: A practical guide. *arXiv*, 2025. doi: 10.48550/arXiv.2508.12939.

Charita Dellaporta, Jeremias Knoblauch, Theodoros Damoulas, and Francois-Xavier Briol. Robust bayesian inference for simulator-based models via the mmd posterior bootstrap. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, 2022.

Lars Dingeldein, Pilar Cossio, and Roberto Covino. Simulation-based inference of single-molecule force spectroscopy. *Machine Learning: Science and Technology*, 4(2):025009, 2023. doi: 10.1088/2632-2153/acc8b8.

Simon Dirmeier, Simone Ulzega, Antonietta Mira, and Carlo Albert. Simulation-based inference with the python package sbijax. *arXiv*, 2024. doi: arXiv.2409.19435.

Arnaud Doucet, Nando de Freitas, and Neil Gordon. *An Introduction to Sequential Monte Carlo Methods*. Springer, 2001.

Simon Duane, A.D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics Letters B*, 195(2):216–222, 1987. doi: 10.1016/0370-2693(87)91197-X.

John P. Dunne, Helene T. Hewitt, Julie Arblaster, Frédéric Bonou, Olivier Boucher, Tereza Cavazos, Paul J. Durack, Martin Juckes Birgit Hassler, Tomoki Miyakawa, Matthew Mizielinski, Vaishali Naik, Zebedee Nicholls, Eleanor O'Rourke, Robert Pincus, Isla R. Simpson Benjamin M. Sanderson, , and Karl E. Taylor. An evolving coupled model intercomparison project phase 7 (cmip7) and fast track in support of future climate assessment. *EGUsphere*, 2024:1–51, 2024. doi: 10.5194/egusphere-2024-3874.

Conor Durkan, Iain Murray, and George Papamakarios. On contrastive learning for likelihood-free inference. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Paul N. Edwards. History of climate modeling. *WIREs Climate Change*, 2(1):128–139, 2011. doi: 10.1002/wcc.95.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Bradley Efron and Carl Morris. Limiting the risk of Bayes and empirical Bayes estimators, part ii: The empirical Bayes case. *Journal of the American Statistical Association*, 67(337):130–139, 1972.

Farhan Feroz, Michael P. Hobson, and Michael Bridges. Multinest: an efficient and robust bayesian inference tool for cosmology and particle physics. *Monthly Notices of the Royal Astronomical Society*, 398(4):1601–1614, 2009. doi: 10.1111/j.1365-2966.2009.14548.x.

Tobias S. Finn, Marc Bocquet, Pierre Rampal, Charlotte Durand, Flavia Porro, Alban Farchi, and Alberto Carrassi. Generative ai models enable efficient and physically consistent sea-ice simulations. *arXiv*, 2025. doi: 10.48550/arXiv.2508.14984.

Yannic Fischler, Martin Rückamp, Christian Bischof, Vadym Aizinger, Mathieu Morlighem, and Angelika Humbert. A scalability study of the ice-sheet and sea-level system model (issm, version 4.18). *Geoscientific Model Development*, 15(9):3753–3771, 2022. doi: 10.5194/gmd-15-3753-2022.

Olivier Gagliardini, Thomas Zwinger, Fabien Gillet-Chaulet, Gérard Durand, Lionel Favier, Basile de Fleurian, Ralf Greve, Mika Malinen, Carlos Martín, Peter Råback, Juha Ruokolainen, Michel Sacchettini, Martin Schäfer, Hakime Seddik, and Jonas Thies. Capabilities and performance of elmer/ice, a new-generation ice sheet model. *Geoscientific Model Development*, 6(4):1299–1318, 2013. doi: 10.5194/gmd-6-1299-2013.

Richard Gao, Michael Deistler, and Jakob H Macke. Generalized bayesian inference for scientific simulators via amortized cost estimation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Alex S. Gardner, Mark Fahnestock, and Ted Scambos. Measures its_live landsat image-pair glacier and ice sheet surface velocities, version 1. 2022. doi: 10.5067/IMR9D3PEI28U.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and S. M. Ali Eslami. Conditional neural processes. In *Proceedings of the 35th International Conference on Machine Learning*, 2018a.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J. Rezende, S. M. Ali Eslami, and Yee Whye Teh. Neural processes. *arXiv*, 2018b. doi: 10.48550/arXiv.1807.01622.

Tomas Geffner, George Papamakarios, and Andriy Mnih. Compositional score modeling for simulation-based inference. In *Proceedings of the 32nd International Conference on Machine Learning*, 2022.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv*, 2025. doi: 10.48550/arXiv.2505.13447.

Manuel Gloeckler, Michael Deistler, Christian Dietrich Weilbach, Frank Wood, and Jakob H. Macke. All-in-one simulation-based inference. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Daniel N. Goldberg and Olga V. Sergienko. Data assimilation using a hybrid ice flow model. *The Cryosphere*, 5(2):315–327, 2011. doi: 10.5194/tc-5-315-2011.

Pedro J Gonçalves, Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F Podlaski, Sara A Haddad, Tim P Vogels, et al. Training deep neural density estimators to identify mechanistic models of neural dynamics. *Elife*, 9, 2020. doi: 10.7554/eLife.56261.

I. J. Good. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Annals of Mathematical Statistics*, 34:911–934, 1963.

Hugues Goosse and Thierry Fichefet. Importance of ice-ocean interactions for the global ocean circulation: A model study. *Journal of Geophysical Research: Oceans*, 104(C10):23337–23355, 1999. doi: 10.1029/1999JC900215.

Neil J. Gordon, David Salmond, and Adrian F. M. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, 140:107–113, 1993. doi: 10.1049/ip-f-2.1993.0015.

David Greenberg, Marcel Nonnenmacher, and Jakob Macke. Automatic posterior transformation for likelihood-free inference. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Chad A. Greene, Alex S. Gardner, Nicole-Jeanne Schlegel, and Alexander D. Fraser. Antarctic calving loss rivals ice-shelf thinning. *Nature*, 609(7929):948–953, 2022. doi: 10.1038/s41586-022-05037-w.

Leslie Greengard and June-Yub Lee. Accelerating the nonuniform fast fourier transform. *SIAM Review*, 46(3), 2004. doi: 10.1137/S003614450343200X.

Ralf Greve and Heinz Blatter. *Dynamics of Ice Sheets and Glaciers*. Springer, 2009.

Stephen M. Griffies, Gokhan Danabasoglu, Paul J. Durack, Alistair J. Adcroft, Venkatramani Balaji, Claus W. Böning, Eric P. Chassignet, Enrique N. Curchitser, Julie Deshayes, Helge Drange, Baylor Fox-Kemper, Peter J. Gleckler, Jonathan M. Gregory, Helmuth Haak, Robert Hallberg, Patrick Heimbach, Helene Theresa Hewitt, David M. Holland, Tatiana Ilyina, Johann H. Jungclaus, Yoshiki Komuro, John P. Krasting, William G. Large, Simon J. Marsland, Simona Masina, Trevor J. McDougall, A. J. George Nurser, James C. Orr, Anna Pirani, Fangli Qiao, Ronald J. Stouffer, Karl E. Taylor, Anne Marie Treguier, Hiroyuki Tsujino, Petteri Uotila, Maria Valdivieso, Qiang Wang, Michael Winton, and Stephen G. Yeager. Omip contribution to cmip6: experimental and diagnostic protocol for the physical component of the ocean model intercomparison project. *Geoscientific Model Development*, 9(9):3231–3296, 2016. doi: 10.5194/gmd-9-3231-2016.

G. Hilmar Gudmundsson, Fernando S. Paolo, Susheel Adusumilli, and Helen A. Fricker. Instantaneous antarctic ice sheet mass loss driven by thinning ice shelves. *Geophysical Research Letters*, 46:13903–13909, 2019. doi: 10.1029/2019GL085027.

Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

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*, 2023.

James Hensman, Nicolò Fusi, and Neil D. Lawrence. Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, 2013.

Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free MCMC with amortized approximate ratio estimators. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Joeri Hermans, Arnaud Delaunoy, François Rozet, Antoine Wehenkel, Volodimir Begy, and Gilles Louppe. A crisis in simulation-based inference? beware, your posterior approximations can be unfaithful. *Transactions on Machine Learning Research*, 2022.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Andrew O. Hoffman, Knut Christianson, Nicholas Holschuh, Elizabeth Case, Jonathan Kingslake, and Robert Arthern. The impact of basal roughness on inland thwaites glacier sliding. *Geophysical Research Letters*, 49(14), 2022. doi: 10.1029/2021GL096564.

Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(40):1303–1347, 2013.

Roger LeB. Hooke. *Principles of Glacier Mechanics*. Cambridge University Press, 2019.

Daolang Huang, Ayush Bharti, Amauri Souza, Luigi Acerbi, and Samuel Kaski. Learning robust statistics for simulation-based inference under model misspecification. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

IPCC. Climate change 2023: Synthesis report, 2023. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland. doi: 10.59327/IPCC/AR6-9789291691647.

Edwin T. Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4(3): 227–241, 1968. doi: 10.1109/TSSC.1968.300117.

Kenneth C. Jezek, Caglar Yardim, Joel T. Johnson, Giovanni Macelloni, and Marco Brogioni. Analysis of ice-sheet temperature profiles from low-frequency airborne remote sensing. *Journal of Glaciology*, 68(271):1027–1037, 2022. doi: 10.1017/jog.2022.19.

Guillaume Jouvet. Inversion of a stokes glacier flow model emulated by deep learning. *Journal of Glaciology*, 69(273):13–26, 2023. doi: 10.1017/jog.2022.41.

Guillaume Jouvet, Guillaume Cordonnier, Byungsoo Kim, Martin Lüthi, Andreas Vieli, and Andy Aschwanden. Deep learning speeds up ice flow modelling by several orders of magnitude. *Journal of Glaciology*, 68(270):651–664, 2022. doi: 10.1017/jog.2021.120.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Ryan P. Kelly, David J. Warne, David T. Frazier, David J. Nott, Michael U. Gutmann, and Christopher Drovandi. Simulation-based bayesian inference under model misspecification. *arXiv*, 2025. doi: 10.48550/arXiv.2503.12315.

Kwanyoung Kim and Jong Chul Ye. Noise2score: Tweedie's approach to self-supervised image denoising without clean images. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. An optimization-centric view on bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23(132):1–109, 2022.

Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43: 3964–3979, 2019. doi: 10.1109/tpami.2020.2992934.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized Sliced Wasserstein distances. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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(89):1–97, 2023.

Nikola B. Kovachki, Samuel Lanthaler, and Andrew M. Stuart. Operator learning: Algorithms and analysis. *arXiv*, 2024. doi: 10.48550/arXiv.2402.15715.

L. Kozachenko and N. Leonenko. A statistical estimate for the entropy of a random vector. *Problems of Information Transmission*, 23:95–101, 1987.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Anastasia N. Krouglova, Hayden R. Johnson, Basile Confavreux, Michael Deistler, and Pedro J. Gonçalves. Multifidelity simulation-based inference for computationally expensive simulators. *arXiv*, 2025. doi: 10.48550/arXiv.2502.08416.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(14):1–45, 2017.

Peter W. Lane. Generalized linear models in soil science. *European Journal of Soil Science*, 53(2): 241–251, 2002. doi: 10.1046/j.1365-2389.2002.00440.x.

Eric Y. Larour, Hélène Seroussi, Mathieu Morlighem, and E. Rignot. Continental scale, high order, high spatial resolution, ice sheet modeling using the ice sheet system model (issm). *Journal of Geophysical Research: Earth Surface*, 117(F1), 2012. doi: 10.1029/2011JF002140.

Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. doi: 10.1162/neco.1989.1.4.541.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539.

Pablo Lemos, Adam Coogan, Yashar Hezaveh, and Laurence Perreault-Levasseur. Sampling-based accuracy testing of posterior estimators for general inference. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

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 *The 9th International Conference on Learning Representation*, 2021.

Levi E. Lingsch, Mike Yan Michelis, Emmanuel de Bezenac, Sirani M. Perera, Robert K. Katzschmann, and Siddhartha Mishra. Beyond regular grids: Fourier-based neural operators on arbitrary domains. In *Forty-first International Conference on Machine Learning*, 2024.

Julia Linhart, Alexandre Gramfort, and Pedro Rodrigues. L-c2st: Local diagnostics for posterior approximations in simulation-based inference. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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. doi: 10.48550/arXiv.2404.07593.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The 11th International Conference on Learning Representations*, 2023.

Chang Liu, Jingwei Zhuo, Pengyu Cheng, Ruiyi Zhang, and Jun Zhu. Understanding and accelerating particle-based variational inference. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Jun S. Liu, Rong Chen, and Wing Hung Wong. Rejection control and sequential importance sampling. *Journal of the American Statistical Association*, 93(443):1022–1031, 1998.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Xingchao Liu, Chengyue Gong, and qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The 11th International Conference on Learning Representations*, 2023.

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.

Jan-Matthis Lueckmann, Giacomo Bassetto, Theofanis Karaletsos, and Jakob H. Macke. Likelihood-free inference with emulator networks. In *Proceedings of The 1st Symposium on Advances in Approximate Bayesian Inference*, volume 96 of *Proceedings of Machine Learning Research*, 2019.

Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. Benchmarking simulation-based inference. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 2021.

Rasmus Bødker Madsen, Anne-Sophie Høyer, Lærke Therese Andersen, Ingelise Møller, and Thomas Mejer Hansen. Geology-driven modeling: A new probabilistic approach for incorporating uncertain geological interpretations in 3d geological modeling. *Engineering Geology*, 309: 106833, 2022. doi: 10.1016/j.enggeo.2022.106833.

Alberto Malinverno and Victoria A. Briggs. Expanded uncertainty quantification in inverse problems: Hierarchical bayes and empirical bayes. *Geophysics*, 69(4):1005–1016, 2004. doi: 10.1190/1.1778243.

Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003. doi: 10.1073/pnas.0306899100.

Takuo Matsubara, Jeremias Knoblauch, François-Xavier Briol, and Chris. J. Oates. Generalized bayesian inference for discrete intractable likelihood. *Journal of the American Statistical Association*, 119(547):2345–2355, 2024. doi: 10.1080/01621459.2023.2257891.

Thorsten Mauritsen, Bjorn Stevens, Erich Roeckner, Traute Crueger, Monika Esch, Marco Giorgetta, Helmuth Haak, Johann Jungclaus, Daniel Klocke, Daniela Matei, Uwe Mikolajewicz, Dirk Notz, Robert Pincus, Hauke Schmidt, and Lorenzo Tomassini. Tuning the climate of a global model. *Journal of Advances in Modeling Earth Systems*, 4(3), 2012. doi: 10.1029/2012MS000154.

Alex Megann, Dave Storkey, Yevgeny Aksenov, Steven G. Alderson, Daley Calvert, Tim Graham, Patrick Hyder, John Siddorn, and Bablu Sinha. Go5.0: the joint nerc–met office nemo global ocean model for use in coupled and forced applications. *Geoscientific Model Development*, 7(3): 1069–1092, 2014. doi: 10.5194/gmd-7-1069-2014.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Nicholas Metropolis and S. Ulam and. The monte carlo method. *Journal of the American Statistical Association*, 44(247):335–341, 1949. doi: 10.1080/01621459.1949.10483310.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953. doi: 10.1063/1.1699114.

Sean Meyn, Richard L. Tweedie, and Peter W. Glynn. *Markov Chains and Stochastic Stability*. Cambridge Mathematical Library. Cambridge University Press, 2009.

Benjamin K Miller, Christoph Weniger, and Patrick Forré. Contrastive neural ratio estimation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv*, 2017. doi: 10.48550/arXiv.1610.03483.

Noemi Anau Montel, J.B.G. Alvey, and Christoph Weniger. Tests for model misspecification in simulation-based inference: from local distortions to global model checks. *arXiv*, 2024. doi: 10.48550/arXiv.2412.15100.

Mathieu Morlighem. Measures bedmachine antarctica, version 3. 2022. doi: 10.5067/FPSU0V1MWUB6.

Mathieu Morlighem, Eric Rignot, Tobias Binder, Donald Blankenship, Reinhard Drews, Graeme Eagles, Olaf Eisen, Fausto Ferraccioli, René Forsberg, Peter Fretwell, Vikram Goel, Jamin S. Greenbaum, Hilmar Gudmundsson, Jingxue Guo, Veit Helm, Coen Hofstede, Ian Howat, Angelika Humbert, Wilfried Jokat, Nanna B. Karlsson, Won Sang Lee, Kenichi Matsuoka, Romain Millan, Jeremie Mouginot, John Paden, Frank Pattyn, Jason Roberts, Sebastian Rosier, Antonia Ruppel, Helene Seroussi, Emma C. Smith, Daniel Steinhage, Bo Sun, Michiel R. van den Broeke, Tas D. van Ommen, Melchior van Wessem, and Duncan A. Young. Deep glacial troughs and stabilizing ridges unveiled beneath the margins of the antarctic ice sheet. *Nature Geoscience*, 13(2): 132–137, 2020. doi: 10.1038/s41561-019-0510-8.

Guy Moss, Leah Sophie Muhle, Reinhard Drews, Jakob H. Macke, and Cornelius Schröder. FNOPE: Simulation-based inference on function spaces with Fourier Neural Operators. In *Advances in Neural Information Processing Systems*, volume 35, 2025a.

Guy Moss, Vjeran Višnjević, 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, 2025b. doi: 10.1017/jog.2025.13.

Peter K. Munneke, Michiel van den Broeke, Jan T. M. Lenaerts, Mark G. Flanner, Alex S. Gardner, and Willem Jan van de Berg. A new albedo parameterization for use in climate models over the antarctic ice sheet. *Journal of Geophysical Research: Atmospheres*, 116(D5), 2011. doi: 10.1029/2010JD015113.

Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli. Statistical and topological properties of sliced probability divergences. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Niklas Neckel, Reinhard Drews, Wolfgang Rack, and Daniel Steinhage. Basal melting at the Ekström Ice Shelf, Antarctica, estimated from mass flux divergence. *Annals of Glaciology*, 53(60):294–302, 2012. doi: 10.3189/2012AOG60A167.

Isaac Newton. *Philosophiæ naturalis principia mathematica*. 1726.

Zebedee R. J. Nicholls, Malte Meinshausen, J Lewis, Maisa Rojas Corradi, Kalyn Dorheim, Thomas Gasser, Robert Gieseke, Austin Patrick Hope, Nicholas James Leach, Laura Anne McBride, Yann Quilcaille, Joeri Rogelj, Ross J. Salawitch, Bjørn Hallvard Samset, Marit Sandstad, Alexey N. Shiklomanov, Ragnhild Bieltvedt Skeie, Christopher Smith, Steve Smith, Xuanming Su, Junichi

Tsutsui, Benjamin Aaron Vega-Westhoff, and Dawn L. Woodard. Reduced complexity model intercomparison project phase 2: Synthesizing earth system knowledge for probabilistic climate projections. *Earth's Future*, 9(6), 2021. doi: 10.1029/2020EF001900.

Margaret A. Oliver and Richard Webster. Kriging: a method of interpolation for geographical information systems. *International journal of geographical information systems*, 4(3):313–332, 1990. doi: 10.1080/02693799008941549.

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.

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*, 2019.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Frédéric Parrenin, Marie G. P. Cavitte, Donald D. Blankenship, Jérôme Chappellaz, Hubertus Fischer, Olivier Gagliardini, Valérie Masson-Delmotte, Olivier Passalacqua, Catherine Ritz, Jason L. Roberts, Martin J. Siegert, and Duncan A. Young. Is there 1.5-million-year-old ice near dome c, antarctica? *The Cryosphere*, 11(6):2427–2437, 2017. doi: 10.5194/tc-11-2427-2017.

Thomas P. Prescott and Ruth E. Baker. Multifidelity approximate bayesian computation with sequential monte carlo parameter sampling. *SIAM/ASA Journal on Uncertainty Quantification*, 9: 788–817, 2020. doi: 10.1137/20M1316160.

Stefan T. Radev, Ulf K. Mertens, Andreas Voss, Lynton Ardizzone, and Ullrich Kothe. BayesFlow: Learning Complex Stochastic Models With Invertible Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(4):1452–1466, 2022. doi: 10.1109/TNNLS.2020.3042395.

Stefan T. Radev, Marvin Schmitt, Lukas Schumacher, Lasse Elsemüller, Valentin Pratz, Yannik Schälte, Ullrich Köthe, and Paul-Christian Bürkner. BayesFlow: Amortized Bayesian workflows with neural networks. *Journal of Open Source Software*, 8(89):5702, 2023. doi: 10.21105/joss.05702.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

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 *The 10th International Conference on Learning Representations*, 2022.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Ronja Reese, G. Hilmar Gudmundsson, Anders Levermann, and Ricarda Winkelmann. The far reach of ice-shelf thinning in antarctica. *Nature Climate Change*, 8(1):53–57, 2018. doi: 10.1038/s41558-017-0020-x.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Bryan Riel and Brent Minchew. Variational inference of ice shelf rheology with physics-informed machine learning. *Journal of Glaciology*, 69(277):1167–1186, 2023. doi: 10.1017/jog.2023.8.

Bryan Riel, Brent M. Minchew, and Tobias Bischoff. Data-driven inference of the mechanics of slip along glacier beds using physics-informed neural networks: Case study on rutford ice stream, antarctica. *Journal of Advances in Modeling Earth Systems*, 13(11), 2021. doi: 10.1029/2021MS002621.

Herbert E. Robbins. An empirical bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and basic theory*. Springer, 1956.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Reuven Y. Rubinstein and Dirk P. Kroese. *Simulation and the Monte Carlo Method*. Wiley Publishing, 2016.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *The 10th International Conference on Learning Representations*, 2022.

Simo Sarkka, Arno Solin, and Jouni Hartikainen. Spatiotemporal learning via infinite-dimensional bayesian filtering and smoothing: A look at gaussian process regression through kalman filtering. *IEEE Signal Processing Magazine*, 30(4):51–61, 2013. doi: 10.1109/MSP.2013.2246292.

Jonathan Schmidt, Luca Schmidt, Felix M. Strnad, Nicole Ludwig, and Philipp Hennig. A generative framework for probabilistic, spatiotemporally coherent downscaling of climate simulation. *npj Climate and Atmospheric Science*, 8(1):270, 2025. doi: 10.1038/s41612-025-01157-y.

Dustin M. Schroeder, Robert G. Bingham, Donald D. Blankenship, Knut Christianson, Olaf Eisen, Gwenn E. Flowers, Nanna B. Karlsson, Michelle R. Koutnik, John D. Paden, and Martin J. Siegert. Five decades of radioglaciology. *Annals of Glaciology*, 61(81):1–13, 2020. doi: 10.1017/aog.2020.11.

Ortal Senouf, Antoine Wehenkel, Cédric Vincent-Cuaz, Emmanuel Abbé, and Pascal Frossard. Inductive domain transfer in misspecified simulation-based inference. *arXiv*, 2025. doi: 10.48550/arXiv.2508.15593.

Daniel R. Shapero, Jessica A. Badgeley, Andrew Hoffmann, and Ian R. Joughin. icepack: a new glacier flow modeling package in python, version 1.0. *Geoscientific Model Development*, 14(7):4593–4616, 2021. doi: 10.5194/gmd-14-4593-2021.

Jack Simons, Song Liu, and Mark Beaumont. Variational likelihood-free gradient descent. In *Fourth Symposium on Advances in Approximate Bayesian Inference*, 2022.

Scott A. Sisson, Yanan Fan, and Mark A. Beaumont. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007. doi: 10.1073/pnas.0607208104.

Scott A. Sisson, Yanan Fan, and Mark M. Tanaka. Overview of abc. In *Handbook of Approximate Bayesian Computation*, chapter 1. CRC Press, Taylor & Francis Group, 2018.

Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems*, volume 18, 2005.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *The 9th International Conference on Learning Representations*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Hans Christian Steen-Larsen, Edwin D. Waddington, and Michelle R. Koutnik. Formulating an inverse problem to infer the accumulation-rate pattern from deep internal layering in an ice sheet using a monte carlo approach. *Journal of Glaciology*, 56(196):318–332, 2010. doi: 10.3189/002214310791968476.

Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating bayesian inference algorithms with simulation-based calibration. *arXiv*, 2020. doi: 10.48550/arXiv.1804.06788.

Yee Whye Teh and Michael I. Jordan. *Hierarchical Bayesian nonparametric models with applications*, page 158–207. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010.

Alvaro Tejero-Cantero, Jan Boelts, Michael Deistler, Jan-Matthis Lueckmann, Conor Durkan, Pedro J. Gonçalves, David S. Greenberg, and Jakob H. Macke. sbi: A toolkit for simulation-based inference. *Journal of Open Source Software*, 5(52):2505, 2020. doi: 10.21105/joss.02505.

Alexios Theofilopoulos and Andreas Born. Sensitivity of isochrones to surface mass balance and dynamics. *Journal of Glaciology*, 69(274):311–323, 2023. doi: 10.1017/jog.2022.62.

Owen Thomas, Ritabrata Dutta, Jukka Corander, Samuel Kaski, and Michael U Gutmann. Likelihood-free inference by ratio estimation. *Bayesian Analysis*, 17(1):1 – 31, 2016. doi: 10.1214/20-BA1238.

Eric Thrane and Colm Talbot. An introduction to Bayesian inference in gravitational-wave astronomy: Parameter estimation, model selection, and hierarchical models. *Publications of the Astronomical Society of Australia*, 36, 2019. doi: 10.1017/pasa.2019.2.

Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, 2009.

Alexander Tong, Nikolay Malkin, Kilian Fatras, Lazar Atanackovic, Yanlei Zhang, Guillaume Huguet, Guy Wolf, and Yoshua Bengio. Simulation-free schrödinger bridges via score and flow matching. In *International Conference on Artificial Intelligence and Statistics*, 2023.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, 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.

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*, 404:115783, 2023. doi: 10.1016/j.cma.2022.115783.

M. C. K. Tweedie. Functions of a statistical variate with given means, with special reference to laplacian distributions. *Mathematical Proceedings of the Cambridge Philosophical Society*, 43(1): 41–49, 1947. doi: 10.1017/S0305004100023185.

Maxime Vandegar, Michael Kagan, Antoine Wehenkel, and Gilles Louppe. Neural empirical Bayes: Source distribution estimation and its applications to simulation-based inference. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Vincent Verjans, Alexander A. Robel, Hélène Seroussi, Lizz Ultee, and Andrew F. Thompson. The stochastic ice-sheet and sea-level system model v1.0 (stissm v1.0). *Geoscientific Model Development*, 15(22):8269–8293, 2022. doi: 10.5194/gmd-15-8269-2022.

Yogesh Verma, Ayush Bharti, and Vikas Garg. Robust simulation-based inference under missing data via neural processes. In *The Thirteenth International Conference on Learning Representations*, 2025.

Julius Vetter, Jakob H. Macke, and Richard Gao. Generating realistic neurophysiological time series with denoising diffusion probabilistic models. *Patterns*, 5(9), 2024a. doi: 10.1016/j.patter.2024.101047.

Julius Vetter, Guy Moss, Cornelius Schröder, Richard Gao, and Jakob H. Macke. Sourcerer: Sample-based maximum entropy source distribution estimation. In *Advances in Neural Information Processing Systems*, volume 37, 2024b.

Julius Vetter, Manuel Gloeckler, Daniel Gedon, and Jakob H. Macke. Effortless, simulation-efficient bayesian inference using tabular foundation models. *arXiv*, 2025. doi: doi.org/10.48550/arXiv.2504.17660.

Vjeran Višnjević, Reinhard Drews, Clemens Schannwell, Inka Koch, Steven Franke, Daniela Jansen, and Olaf Eisen. Predicting the steady-state isochronal stratigraphy of ice shelves using observations and modeling. *The Cryosphere*, 16:4763–4777, 2022. doi: 10.5194/tc-16-4763-2022.

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*, 53:694–712, 2007. doi: 10.3189/002214307784409351.

Yixin Wang, Andrew C. Miller, and David M. Blei. Comment: Variational Autoencoders as Empirical Bayes. *Statistical Science*, 34(2), 2019.

Daniel Ward, Patrick Cannon, Mark Beaumont, Matteo Fasiolo, and Sebastian Schmon. Robust neural posterior estimation and statistical model criticism. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

David J. Warne, Thomas P. Prescott, Ruth E. Baker, and Matthew J. Simpson. Multifidelity multilevel monte carlo to accelerate approximate bayesian parameter inference for partially observed stochastic processes. *Journal of Compuational Physiscs*, 469(C), 2022. doi: 10.1016/j.jcp.2022.111543.

Antoine Wehenkel, Juan L. Gamella, Ozan Sener, Jens Behrmann, Guillermo Sapiro, Joern-Henrik Jacobsen, and marco cuturi. Addressing misspecification in simulation-based inference through data-driven calibration. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Boris Weinberg. über die innere reibung des eises. ii. *Annalen der Physik*, 327(2):321–332, 1907. doi: 10.1002/andp.19073270208.

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.

Christopher K. I. Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006.

Ricarda Winkelmann, Maria A. Martin, Marianne Haseloff, Torsten Albrecht, Ed Bueler, Constantine Khroulev, and Anders Levermann. The potsdam parallel ice sheet model (pism-pik) – part 1: Model description. *The Cryosphere*, 5(3):715–726, 2011. doi: 10.5194/tc-5-715-2011.

Ricarda Winkelmann, Anders Levermann, Maria A. Martin, and Katja Frieler. Increased future ice discharge from antarctica owing to higher snowfall. *Nature*, 492(7428):239–242, 2012. doi: 10.1038/nature11616.

Michael J. Wolovick, John Christopher Moore, and Liyun 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(5), 2021. doi: 10.1029/2020JF005937.

Simon N Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466 (7310):1102–1104, 2010. doi: 10.1038/nature09319.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):105:1–105:39, 2024. doi: 10.1145/3626235.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Justine Zeghal, Franccois Lanusse, Alexandre Boucaud, Benjamin Remy, and Eric Aubourg. Neural posterior estimation with differentiable simulators. In *39th International Conference on Machine Learning Workshop on Machine Learning for Astrophysics*, 2022.

Appendices

Journal of Glaciology

Simulation-based inference of surface accumulation and basal melt rates of an Antarctic ice shelf from isochronal layers

Guy Moss[1], Vjeran Višnjević[2], Olaf Eisen[3,4], Falk M. Oraschewski[2], Cornelius Schröder[1,†], Jakob H. Macke[1,5,†] and Reinhard Drews[2,†]

[1]Machine Learning in Science, University of Tübingen and Tübingen AI Center, Tübingen, Germany; [2]Department of Geosciences, University of Tübingen, Tübingen, Germany; [3]Alfred-Wegener-Institut, Helmholtz-Zentrum für Polar und Meeresforschung, Bremerhaven, Germany; [4]Faculty of Geosciences, University of Bremen, Bremen, Germany and [5]Max Planck Institute for Intelligent Systems, Tübingen, Germany

Article

Cite this article: Moss G, Višnjević V, Eisen O, Oraschewski FM, Schröder C, Macke JH, Drews R (2025) Simulation-based inference of surface accumulation and basal melt rates of an Antarctic ice shelf from isochronal layers. *Journal of Glaciology* **71**, e44, 1–21. https://doi.org/10.1017/jog.2025.13

Received: 18 November 2024
Revised: 27 January 2025
Accepted: 11 February 2025

Keywords:
Bayesian inference; Ice shelves; Melt - basal; Mass-balance reconstruction; Machine learning; Simulation-based inference

Corresponding author: Guy Moss;
Email: guy.moss@uni-tuebingen.de

[†]Joint supervision

Abstract

The ice shelves buttressing the Antarctic ice sheet determine the rate of ice-discharge into the surrounding oceans. Their geometry and buttressing strength are influenced by the local surface accumulation and basal melt rates, governed by atmospheric and oceanic conditions. Contemporary methods quantify one of these rates, but typically not both. Moreover, information about these rates is only available for recent time periods, reaching at most a few decades back since measurements are available. We present a new method to simultaneously infer the surface accumulation and basal melt rates averaged over decadal and centennial timescales. We infer the spatial dependence of these rates along flow line transects using internal stratigraphy observed by radars, using a kinematic forward model of internal stratigraphy. We solve the inverse problem using simulation-based inference (SBI). SBI performs Bayesian inference by training neural networks on simulations of the forward model to approximate the posterior distribution, therefore also quantifying uncertainties over the inferred parameters. We validate our method on a synthetic example, and apply it to Ekström Ice Shelf, Antarctica, for which independent validation data are available. We obtain posterior distributions of surface accumulation and basal melt averaging over up to 200 years before 2022.

1. Introduction

The majority of the Antarctic ice sheet is buttressed by floating ice shelves (Bindschadler and others, 2011) which provide large contact areas for ice–ocean interactions. Approximately half of the ice shelves' total mass loss is attributed to ocean-induced melting at the underside of ice shelves (Depoorter and others, 2013), and its spatiotemporal variability imprints ice flow dynamics farther upstream (Reese and others, 2017; Gudmundsson and others, 2019). Consequently, ice flow and ocean models need to be coupled for future projections; frameworks (Goldberg and others, 2019; Gladstone and others, 2021), parameterizations (Burgard and others, 2022; Goldberg and Holland, 2022) and benchmarks (Asay-Davis and others, 2016) for this task have been developed. Similarly, the local snow accumulation is influenced by atmospheric conditions and is crucial in determining ice shelf thickness (Winkelmann and others, 2012). As a result, ice flow models are also coupled to climate models for future projections (Goelzer and others, 2016; Pattyn and others, 2017). It is crucial to confront ice flow models with observations to validate them and investigate their ability to explain observed phenomena. Here, we present a new method that infers surface accumulation (also known as 'surface mass balance' (Lenaerts and others, 2019)) and basal melt rates (collectively, the mass-balance parameters) from the ice shelves' internal stratigraphy, which can be routinely mapped by radio-echo sounding.

Typically, surface accumulation is the more accessible mass-balance parameter (Eisen and others, 2008); it can be measured in situ using stake farms and can also be derived from multiple firn cores (Lenaerts and others, 2019). Many of these observations validate atmospheric models such as RACMO (van Wessem and others, 2018) and MAR (Gallée and Schayes, 1994; Agosta and others, 2019), which estimate surface accumulation on 35 km grids (Lenaerts and others, 2019) (with few locations being estimated at a higher resolution of 5.5 km). Estimating the basal melt is more challenging and is typically dependent on knowledge of surface accumulation. For example, estimates of surface accumulation have been used along with mass conservation arguments to estimate basal melt (Neckel and others, 2012; Depoorter and others, 2013; Berger and others, 2017; Adusumilli and others, 2020). These approaches have provided Antarctic-wide time series of the last few decades of basal melt rates (Adusumilli and others, 2020). The spatial resolution is currently limited to the kilometer scale, which may miss fine grained processes occurring within ice shelf channels (Drews, 2015; Marsh and others, 2016) or near basal terraces

https://doi.org/10.1017/jog.2025.13 Published online by Cambridge University Press

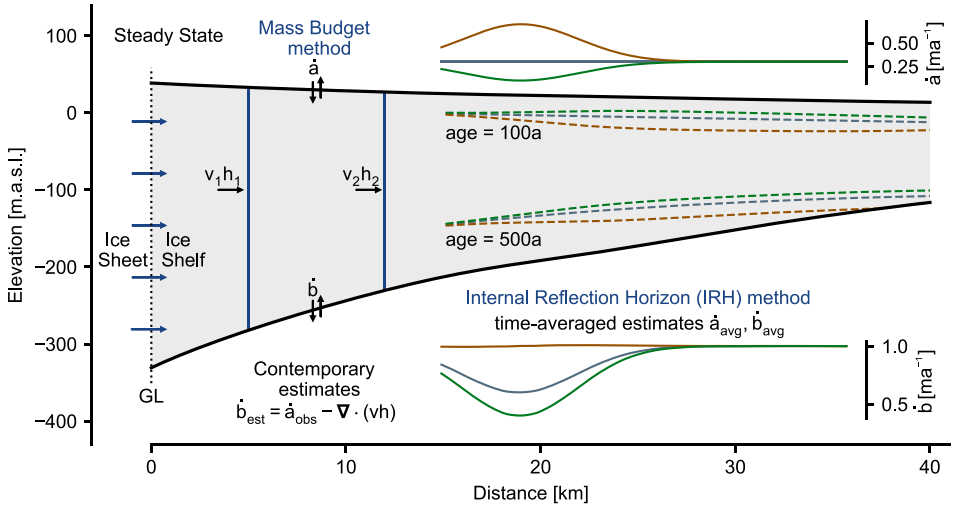

Figure 1. Estimation of mass-balance parameters from a steady-state ice shelf with two methods. The Eulerian Mass Budget method (left) detects the difference of surface accumulation and basal melt within two flux gates (blue vertical lines) by considering flux divergence $\nabla \cdot (\mathbf{vh})$. Often, the basal melt rates $\dot{b}$ are inferred assuming that the surface accumulation ($\dot{a}_{\text{obs}}$) is known. In the internal reflection horizon (IRH) method, we are given information on the internal stratigraphy of the shelf. This information is used to separate the known total mass balance into individual estimates of surface accumulation and basal melt ($\dot{a}_{\text{avg}}$, $\dot{b}_{\text{avg}}$ respectively). These estimates correspond to the time-averaged value over the age of the IRH to the present. The inset plots show different surface accumulation and basal melt parameterizations which give rise to the same total mass balance and overall shape of the ice shelf, but different internal stratigraphy.

(Dutrieux and others, 2014). Measurements of basal melt which are independent of the surface accumulation are also available, but typically only on short temporal scales, for example, with time-lapse radar measurements of ice thickness change (Zeising and others, 2022). Using phase-coherent data acquisition, these measurements can disentangle the observed thickness change into strain thinning and basal melt (Nicholls and others, 2015). This has provided much insights, e.g., in terms of relevant tidal (Sun and others, 2019) and seasonal timescales (Vankova and Nicholls, 2022).

Here, we investigate to what extent the radar-imaged isochronal ice stratigraphy (Eisen and others, 2004) can provide additional information for inferring mass-balance parameters. On grounded ice, radar-imaged internal reflection horizons (IRHs) have been used in multiple ways, for example, to infer the surface accumulation history (Waddington and others, 2007; MacGregor and others, 2009; Catania and others, 2010; Steen-Larsen and others, 2010; Wolovick and others, 2021; Theofilopoulos and Born, 2023), velocity patterns of the ice flow (Eisen, 2008; Holschuh and others, 2017), ice-rise evolution (Drews and others, 2015; Henry and others, 2023) or large-scale model calibration (Leysinger Vieli and others, 2011; Sutter and others, 2021). On ice shelves, surface accumulation rate can also be derived from the radar-measured shallow stratigraphy (Pratap and others, 2022), but not from intermediate depths and below where the stratigraphy is also influenced by basal melt and ice flow. The stratigraphy of ice shelves differs for various combinations of surface accumulation and basal melt rates (Visnjevic and others, 2022). This suggests that given an ice flow model of the internal stratigraphy that accounts for the surface accumulation and basal melt rates, we can use observed IRHs to recover the surface accumulation and basal melt rate histories (Fig. 1). Thus, our goal is to solve the inverse problem of inferring the surface accumulation and basal melt rates that can explain the observed IRHs under the physical constraints of the ice flow model.

Inverse problems, also known as *inversion*, *data assimilation* or *inference* problems in the literature, denote the task of finding the model parameters that are compatible both with empirical observations and prior knowledge. This problem is widespread in the geosciences, e.g., in hydrogeology (Linde and others, 2015), seismology (Symes, 2009) or in climate science (Tebaldi and Sansó, 2008). Bayesian inference provides a powerful framework for solving inference problems, but conventional Bayesian approaches are restricted to models for which the so-called 'likelihood function' is tractable. A tractable likelihood function is one that can be efficiently evaluated (see Appendix C.1 for examples of tractable and intractable likelihood functions). However, this is not the case in our setup. We therefore use simulation-based inference (SBI, Papamakarios and Murray (2016); Lueckmann and others (2017); Cranmer and others (2020)) to solve the inverse problem presented in this work. In SBI, we evaluate the forward model under different values of the model's parameters from a prior distribution. We use the resulting simulated dataset to train a neural network that performs conditional density estimation. In the neural posterior estimation (NPE) variant, the network approximates the Bayesian posterior distribution. A key advantage of NPE is the amortization of simulation cost. An amortized inference framework is one that, once trained, can be instantly applied to find the posterior distribution for any new observation without requiring more simulations or training. Importantly, SBI does not require the forward model to be differentiable and can also work with 'blackbox' models. Therefore, our approach can be extended to a variety of preexisting forward models. To the authors' knowledge, this work is the first application of SBI in glaciology, but we note that it has already been applied in other geoscientific disciplines such as geothermics (Omagbon and others, 2021), hydrogeology (Allgeier and Cirpka, 2023), hydrology (Hull and others, 2022) and molecular ecology (Overcast and others, 2021).

In this study, we consider steady-state ice shelves and IRHs in the local meteoric ice (LMI) body of ice shelves (Das and others, 2020). This work is a test case for inferring atmospheric and oceanographic boundary conditions from the ice stratigraphy with a novel inference technique that provides uncertainty estimates.

https://doi.org/10.1017/jog.2025.13 Published online by Cambridge University Press

Our approach can be transferred to other ice flow regimes (e.g. flank flow on grounded ice) where similar scientific questions can be explored. Our approach can similarly be adapted to ice shelves exhibiting marine ice formation. Moreover, the isochronal stratigraphy of ice shelves and ice sheets (including the neighboring ice rises) is currently the only archive of surface accumulation and basal melt over the past hundreds of years. Our approach is capable of testing this archive. Thus, this study provides one link between observational initiatives (such as AntArchitecture, Bingham and others, 2024) for Antarctica-wide internal stratigraphy datasets and the modeling community.

The paper is structured as follows: In Section 2, we describe our forward model of the internal stratigraphy of an ice shelf and introduce our inference approach. In Section 3, we detail the synthetic ice shelf construction. We also present the results of inferring the mass-balance parameters from this synthetic stratigraphy and compare the posterior distribution to a known ground truth. In Section 4, we describe the setting of the Ekström Ice Shelf (EIS) and the dataset of observed IRHs along the central flow line transect. We then provide the results of our inference framework and compare them to independent measurements of surface accumulation uniquely available in this location for the periods 1996–2005 and 2014–23. In Section 5, we interpret our results and evaluate our approach. We finally conclude and discuss future perspective in Section 6.

2. Methodology

2.1. Forward model

We denote spatially varying parameters as functions, e.g. $\dot{a}(x)$ or at times $\dot{a}$ for brevity, while bold-faced characters denote the discretized values of this function on a specified grid, e.g. $\dot{\mathbf{a}} = [\dot{a}(x_1), \dots, \dot{a}(x_n)]^\top$.

2.1.1. Ice flow model

We model ice shelves using the shallow shelf approximation (SSA) (Morland, 1984). Throughout this study, we consider ice shelves in steady state. Consequently, the ice surface s, base f, thickness $h = s - f$ and velocity v are all fixed throughout our simulations. We assume plug flow for the ice shelf regime, meaning that the horizontal velocity profile does not change in the vertical direction z. These assumptions results in the mass-balance condition

$$\nabla \cdot (hv) = \dot{m}, \tag{1}$$

where hv is the total mass flux, $\nabla\cdot$ is the divergence operator and $\dot{m} = \dot{a} - \dot{b}$ is the total mass-balance rate. Here we use the convention that the surface accumulation rate $\dot{a}$ is positive for mass gain of the ice shelf and the basal melt rate $\dot{b}$ is positive for mass loss. In this exploratory study, we focus on flow lines. We parameterize our domain such that x denotes the distance along the flow line, and v_x now denotes the velocity parallel to the flow line. The two-dimensional (2-D) geometry is only valid for observations located on flow lines and in the absence of lateral compression and extension. While the former is approximately true in our case, the latter is unrealistic for most Antarctic ice shelves. To account for ice flux into or out of our modeling domain, we, therefore, include the ice flux component normal to the flow line as an additional, spatially variable term to the total mass-balance rate $\dot{m}$ (Appendix A). We test the validity of this approach in a 2-D synthetic example (Section 3) that includes a spatially variable total mass balance and lateral compression. For the real-world scenario, we estimate the the normal ice flux component from satellite velocities.

We seek to predict the steady-state internal stratigraphy for a given flow line and possible surface accumulation and basal melting rate profiles. We define the internal stratigraphy to be a set of isochronal layer elevations $\{e_1(x), \dots e_L(x)\}$, with $f(x) \le e_1(x) \le e_2(x) \le \cdots \le e_L(x) \le s(x)$. One approach to calculate the internal stratigraphy uses the SSA expression for the vertical component of the velocity (Greve and Blatter, 2009) to have a fully specified velocity field. This can then be used to calculate the age field $\mathcal{A}(x, z)$ of the shelf. Contours of constant age (isochrones) then define the internal stratigraphy. However, these methods suffer from numerical diffusion and can be computationally expensive (Visnjevic and others, 2022).

The computational efficiency of the forward model is crucial for our inference method, as we need to evaluate the forward model many times. As a result, we opt instead to use an implementation of the tracer method (Born, 2017; Born and Robinson, 2021). The model is seeded with vertical segments each with a thickness profile $\{h_1(x), \dots, h_L(x)\}$, such that the sum matches the ice geometry $\sum_{l=1}^{L} h_l(x) = h(x)$. The horizontal velocity $v_x(x)$ is used to advect mass within segments and to thin or thicken the segments as a function of the prescribed strain rates. The accumulation $\dot{a}(x)$ and melt $\dot{b}(x)$ rates are used to add new segments or take away mass from the two boundary segments at the top and bottom of the shelf respectively. The (isochronal) layer elevations are then the boundaries between our modeled segments. We use the convention that e_l corresponds to the top of segment l, which can be calculated using the cumulative thicknesses of the segments below,

$$e_l(x) = f(x) + \sum_{l'=1}^{l} h_{l'}(x). \tag{2}$$

In our simulations, we used a high temporal resolution of one isochronal layer per year. Despite the high resolution, the layer tracing method allows for determining the internal stratigraphy in a computationally efficient manner. For the domains and timescales considered in our study, the complete forward model can be evaluated on the order of 60 s on a single CPU core, enabling the application of SBI methods (see Appendix E for details).

To uniquely determine the layer thicknesses in such a scheme, we need to specify the boundary conditions on the layer thicknesses h_l at the inflow boundary $x = 0$ (here corresponding to the grounding line). The true boundary conditions are typically not known. However, the stratigraphy in a large part of the domain is still independent of the boundary conditions. This zone corresponds to the LMI body of ice shelves (Das and others, 2020). When inferring from observed stratigraphy data, we use only data within the LMI body. We detail our model of the LMI body in Appendix A.

2.1.2. Noise model

The ice flow model predicts isochronal layers with varying depth over spatial scales of kilometers. Observed IRHs, however, also show variability on sub-kilometer scales. This systematic model-data misfit is caused by errors in input datasets (such as surface velocity, geometry), coarse resolution of the forward model and omission of higher order processes that are not included in the forward model, such as the effect of rheology. For inference, it is important that the predicted isochrones have consistent statistical properties with the observed IRHs. This is achieved by the definition of an appropriate noise model.

https://doi.org/10.1017/jog.2025.13 Published online by Cambridge University Press

The ice flow model predicts isochronal layer elevations $\{\mathbf{e}_1(\mathbf{x}), \dots, \mathbf{e}_L(\mathbf{x})\}$ on a fixed grid $\mathbf{x} \in \mathbb{R}^N$ where N is the number of grid points. Guided by empirical observations, the noise model should have the property that the errors of different between modeled layers l at different depths are spatially correlated and amplified for deeper layers. We, therefore, define a layer-wise noise model as the product of an x-dependent baseline noise function and a z-dependent vertical amplification factor. More precisely, the additive noise $\boldsymbol{\delta}_l \in \mathbb{R}^N$ of layer l is defined as

$$\boldsymbol{\delta}_l = \boldsymbol{\epsilon} \odot \mathbf{T}(\mathbf{e}_l), \tag{3}$$

where $\boldsymbol{\epsilon} = [\epsilon_1, \dots, \epsilon_N]^\top$ is a $\mathbf{x}$-dependent noise profile, which is shared for all layers, $\mathbf{T}(\mathbf{e}_l) = [T(e_{l,1}), \dots, T(e_{l,N})]^\top$ is a deterministic function of elevation (increasing with depth), and $\odot$ denotes an element-wise product. The vertical scaling $\mathbf{T}(\cdot)$ mimics uncertainties in the travel time-to-depth conversion which depend on the density $\rho(z)$. Here, this is done using $\rho(z)$ as in Drews and others (2016) and an empirical density-permittivity relation (Looyenga, 1965) to calculate the radio-wave speed $c(z)$. This results in the factor

$$T(z) = \int_z^s \frac{dz'}{c(z')}, \tag{4}$$

which we then discretize on the set of layer elevations.

The sub-kilometer variability of the observed IRHs are modeled with power spectral densities $\boldsymbol{\epsilon}$:

$$\boldsymbol{\epsilon} = A_{\boldsymbol{\epsilon}} \sum_{n=1}^{N} \sqrt{\exp^{\beta_n}} \cos(2\pi\omega_n + \chi_n), \tag{5}$$

where the log power spectral densities β_n and offsets χ_n are randomly sampled from normal and uniform distributions respectively: $\beta_n \sim \mathcal{N}(\mu_{\beta_n}, \sigma^2_{\beta_n})$ and $\chi_n \sim U([-\pi, \pi))$. The frequencies ω_n are the corresponding Fourier frequencies of the simulation grid $\mathbf{x}$ and $A_{\boldsymbol{\epsilon}}$ is a global scale factor (set to 4×10^{-10}). In the synthetic ice shelf (Section 3), we define the distribution of the log power densities using $\sigma^2_{\beta_n} = 0.5$ and

$$\mu_{\beta_n} = -8\,(1 - \exp^{-200\omega_n})\,. \tag{6}$$

For EIS, the distribution means μ_{β_n} and variances $\sigma^2_{\beta_n}$ were calibrated given the observed IRHs on a separate set of calibration simulations (full details in Appendix B). We emphasize that this representation of the noise model is a choice—we define a mathematical model of the mismatch, rather than model a physical effect directly. Thus, other choices are possible. We choose this representation of the noise model for its flexibility and interpretability.

By combining the ice flow model with the empirically guided noise model, we have arrived at a physically motivated forward model to sample a plausible observed internal stratigraphy of an ice shelf from the mass-balance rate parameters $\dot{a}$ and $\dot{b}$.

2.2. Inference

Having established the forward model, we arrive at the *inverse problem* of finding the surface accumulation and basal melt rates that best explain the observed internal stratigraphy. We use Bayes theorem with model parameters θ and outcomes X:

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}. \tag{7}$$

Here, $p(\theta|X)$ is the posterior distribution of the parameters given a particular outcome X, $p(X|\theta)$ is the likelihood function of the model, $p(\theta)$ is the prior distribution encoding our existing knowledge on the plausible values of θ and $p(X)$ is the model evidence. The goal of Bayesian inference is to find the posterior $p(\theta|X_o)$, where X_o is *observed data* which has the same form as X, but is measured, instead of simulated.

2.2.1. Simulation-based inference

It is generally not possible to analytically solve for the Bayesian posterior distribution (Eqn (7)), as the evidence term $p(X)$ cannot be computed. Approximate methods exist to solve Eqn (7) using knowledge of only the likelihood function and prior distribution. In this work, we deploy SBI, an approximate Bayesian inference and *likelihood-free* approach, using only samples from our forward model. In SBI, we use artificial neural networks to approximate conditional probability distributions. While there exist different variants of SBI which target either the likelihood $p(X|\theta)$ or the likelihood ratio (see Cranmer and others (2020) for an overview), we focus on NPE, which approximates the posterior distribution directly (Papamakarios and Murray, 2016; Greenberg and others, 2019).

In NPE, we generate a training dataset $\{(\theta_k, X_k)\}_{k=1}^K$ (Fig. 2) by sampling parameters from the prior $\theta_k \sim p(\theta)$ and sampling from the forward model $X_k \sim p(X|\theta_k)$. To approximate the posterior distribution, a variational family of distributions $q_\phi(\theta|X)$ is typically defined in terms of a neural network with learnable weights ϕ. We represent q_ϕ as a normalizing flow (Durkan and others, 2019; Kobyzev and others, 2019; Papamakarios and others, 2019a). Normalizing flows are flexible generative models, which, once trained, can be used either to sample or evaluate the (log-) probability density function of the conditional distribution $q_\phi(\theta|X)$, for any outcome X in the support of the training dataset. We provide a brief description of normalizing flows in Appendix C.2 and refer the reader to Papamakarios and others (2019a) for a review.

In NPE, the neural network is trained by minimizing the expected negative log-probability

$$\mathcal{L}(\phi) = \mathbb{E}_{\theta_k \sim p(\theta), X_k \sim p(X|\theta_k)}[-\log q_\phi(\theta_k|X_k)] \tag{8}$$

on the training dataset. More intuitively, this loss seeks to maximize the probability assigned to the training data. It can be trivially shown that minimizing this loss is equivalent to minimizing the (forward) Kullback–Leibler (KL) divergence between the variational distribution and the true posterior distribution (see C.3).

It has been shown that, if there exists a set of weights ϕ such that $q_\phi(\theta|X)$ is the true posterior distribution, and in the limit of infinite training samples $K \to \infty$, the minimum of the loss in Eqn (8) is reached when $q_\phi(\theta|X) = p(\theta|X)$ for all X—i.e. when our estimated distribution matches the true posterior (see Proposition 1 of Papamakarios and Murray 2016 for full statement and proof).

We additionally make use of an *embedding network*, which are commonly used in SBI workflows to improve performance. Embedding networks learn *summary statistics* $Y(X)$, which are lower-dimensional representations of the outcomes X. Using the embedding $Y(X)$ as an input to the normalizing flow instead of X itself reduces the model complexity. The embedding network is trained jointly with the normalizing flow. In our setting, X_k are spatially varying IRH elevations, and so we choose a 1-dimensional (1-D) convolutional neural network as our embedding net, resulting in 50-dimensional embeddings on which the posterior network is conditioned (full details in Appendix D). Throughout this work, we use the `sbi` package for Python (Tejero-Cantero and others, 2020) to perform inference.

https://doi.org/10.1017/jog.2025.13 Published online by Cambridge University Press

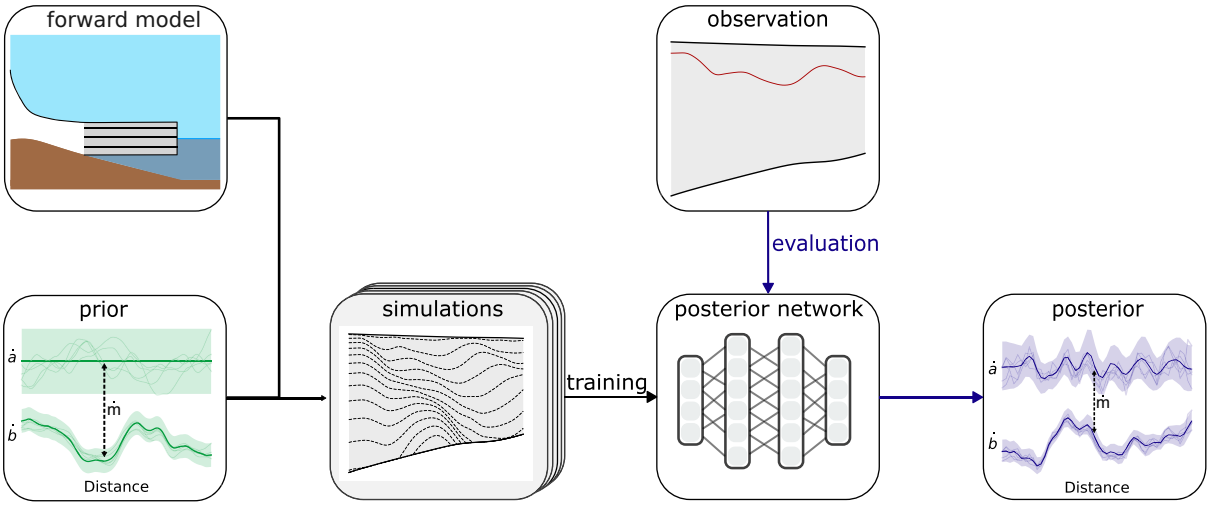

Figure 2. Simulation-based inference workflow. SBI has two primary phases: training and evaluation. In the training phase, accumulation rates are randomly sampled from a prior distribution, the corresponding basal melt rates are obtained using total mass balance, and the resulting internal stratigraphy is calculated using the forward model. These simulations from the prior are used to train a neural network which parameterizes conditional distributions. In the evaluation phase, the trained network is conditioned on the observed IRH and outputs the Bayesian posterior distribution over the parameters (without any additional calls to the forward model).

2.2.2. Definitions of model parameters, outputs and observations

We define $\theta = \dot{\mathbf{a}}_{\text{inf}} = [\dot{a}_{i_1}, \dots, \dot{a}_{i_J}]^\top$, the values of the surface accumulation rate on a discretized grid $\tilde{\mathbf{x}}$. In our experiment, we choose the number of inference grid points $J = 50$ as a compromise between computational complexity of the inference problem while still inferring accumulation rate at a high resolution of ~2.5 km. This is smaller than the discretized grid $\mathbf{x}$ we use for our simulations, which has 500 gridpoints in our experiments. In practice, we take $\tilde{\mathbf{x}}$ to be a regularly spaced subset of $\mathbf{x}$, so that $\dot{\mathbf{a}}_{\text{inf}}$ can also be taken as a subset of $\mathbf{a}$. However, $\tilde{\mathbf{x}}$ can be any discretization of the flow line and need not be a subset of $\mathbf{x}$. Furthermore, despite defining θ to only represent the surface accumulation, any inference of the surface accumulation rate automatically extends to inference of the basal melt rates. This is because for any probability distribution $q(\dot{\mathbf{a}}_{\text{inf}})$, the total mass-balance relationship implies that $\dot{\mathbf{b}}_{\text{inf}} \sim q(\dot{\mathbf{a}}_{\text{inf}} - \dot{\mathbf{m}}_{\text{inf}})$, where $\dot{\mathbf{b}}_{\text{inf}}, \dot{\mathbf{m}}_{\text{inf}}$ are the respective discretizations of $\dot{b}, \dot{m}$ onto $\tilde{\mathbf{x}}$.

We now turn to describing the observation, X_o, and forward model outcomes X_k. The observed data are a set of different IRHs, $\{\mathbf{e}_m(\mathbf{x})\}_{m=1}^M$, where $e_m(x_i)$ is the elevation of the mth IRH in our dataset at grid position i. The IRH elevations need not and typically are not observed at the same locations as the simulation gridpoints; and so we first interpolate the IRH elevations onto the simulation grid $\mathbf{x}$ using linear spline interpolation (as implemented in `Scipy` (Virtanen and others, 2020)). Therefore, we assume $\{\mathbf{e}_m(\mathbf{x})\}_{m=1}^M$ is already defined on $\mathbf{x}$. One reasonable choice is to define the observation X_o as the entire set of all measured IRH elevations. However, in our work, we choose to separately infer the mass balance from each IRH in our observed dataset. This choice has two advantages: first, ordering IRHs by depth also corresponds to their reverse age order, with the oldest IRHs being the deepest. Thus, inferring the surface accumulation and basal melt rates for deeper IRHs corresponds to inferring the average rates over longer periods of time. By comparing the inferred mass-balance parameters obtained with different IRHs, we can reason whether or not our steady-state assumption is valid. The second advantage is practical—we seek a consistent representation of the observations that can be applied across ice shelves. Given a different ice shelf, there will be a different number of IRHs at different depths. Therefore, the embedding net for these data will have to have a different architecture for each ice shelf. In our representation, the embedding net can always be a 1-D convolutional net, as the observations are always 1-D vectors.

Thus, given a dataset of M observed IRHs, we have M inference problems to solve, where each observation corresponds to one IRH. It is, therefore, reasonable to take one isochronal layer of the simulated stratigraphy as the output of the forward model. For the mth inference problem, we define the outcome of the forward model as the isochronal layer $\mathbf{e}_l$ that is closest to IRH m (in the mean square sense). More precisely, for inference problem m and simulation k, we define the observation of the forward model to be $X_k^m = \mathbf{e}_{l^*}(\mathbf{x}_{i \geqslant i(m)})$, where

$$l^* = \arg\min_l ||\mathbf{e}_l(\mathbf{x}_{i \geqslant i(m)}) - \mathbf{e}_m(\mathbf{x}_{i \geqslant i(m)})||_2^2. \quad (9)$$

Here, $i(m)$ is the index of the boundary of the LMI body for IRH $\mathbf{e}_m(\mathbf{x})$. For $i < i(m)$, the IRH $e_m(x_i)$ is outside the LMI body and for $i \geqslant i(m)$ within the LMI body (see Appendix B for details). We further define $\mathbf{x}_{i \geqslant i(m)} = [x_{i(m)}, x_{i(m)+1}, \dots, x_N]^\top$ as the restriction of the gridpoints $\mathbf{x}$ to within the LMI body of $\mathbf{e}_m(\mathbf{x})$. We correspondingly set the observation for IRH m to $X_o^m = \mathbf{e}_m(\mathbf{x}_{i \geqslant i(m)})$. Our choice to select the simulated isochronal layer that most closely matches the IRH is due to the true age of the IRH being unknown. This introduces degeneracy into the forward model—two simulations with different surface accumulation and basal melt rate parameterizations can produce isochronal layers with a similar geometry but different ages. It is, therefore, important to define the prior distribution appropriately, which we do in the following section.

2.2.3. Choice of prior distribution

We aim to approximate the posterior distribution

$$p(\dot{\mathbf{a}}|X_o^m) \propto p(X_o^m|\dot{\mathbf{a}})p(\dot{\mathbf{a}}). \quad (10)$$

The likelihood $p(X_o^m|\dot{\mathbf{a}})$ is not tractable but can be sampled from using the forward model. To specify the prior, we use the long-term snow accumulation observations of the Neumayer stations

https://doi.org/10.1017/jog.2025.13 Published online by Cambridge University Press

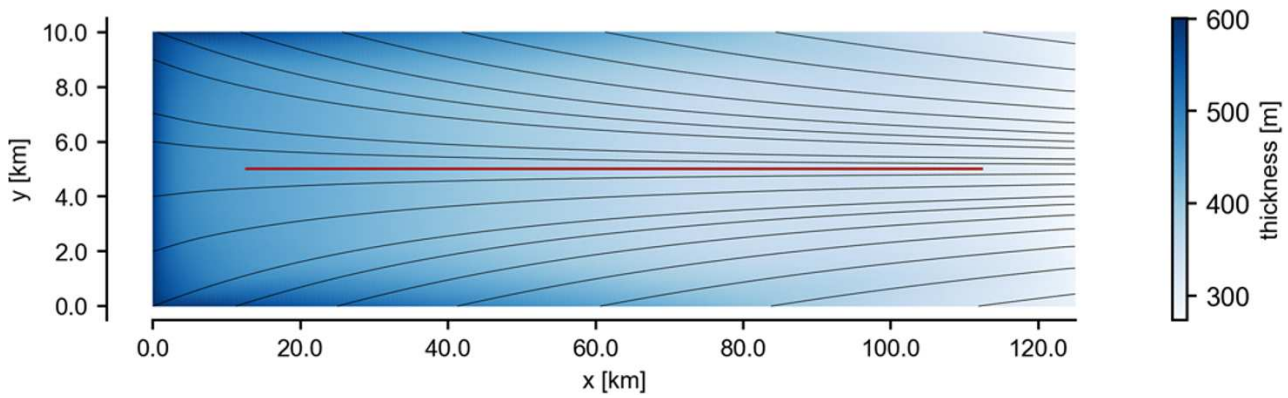

Figure 3. Two-dimensional flow tube domain setup for the synthetic example. Map view of the simulated ice shelf's surface. Flow lines (gray lines) converge to the central flow line (red). Color indicates ice thickness. The input variables for the internal stratigraphy model are evaluated on the central flow line.

(Wesche and others, 2016; Wesche and Regnery, 2022) over more than 30 years to define an empirically motivated prior for EIS, which we also use for the synthetic ice shelf. We first assume that localized surface melt ($\dot{a} < 0$) is possible, but rare. We also observe that average rate of accumulation is $\sim$0.5 m a^{-1}, and that the accumulation rate is almost everywhere under 2 m a^{-1}. Finally, we take the accumulation rate to vary smoothly in space. We define a prior distribution that satisfies these criteria, while still allowing for a broad range of surface accumulation rate profiles. We define the following generative process for $\dot{\mathbf{a}}$: first, we draw a sample $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_N]^\top$ from a Gaussian process with mean function $\mu = 0$ and a Matérn kernel with a Matérn-ν of 2.5 and a length scale of 2500 m (Rasmussen and Williams, 2005). We then independently sample an offset $\mu_{\text{off}} \sim \mathcal{N}(0.5, 0.25^2)$ and scale $\sigma_{\text{sc}} \sim U(0.1, 0.3)$ parameter. Finally, we set $\dot{\mathbf{a}} = \sigma_{\text{sc}}\dot{\boldsymbol{\alpha}} + \mu_{\text{off}}\mathbf{1}$. We inspect the implicit prior this defines over the basal melt rates in Section 4.3.

Defining the prior in this way is sufficiently expressive to capture numerous accumulation rate profiles, while also restricting the samples to conform to empirical knowledge. Additionally, the prior is shared for all M inference problems we have defined, and one evaluation of the forward model provides an observation X_k^m for each of the inference problems. Thus, the same training dataset can be used for all posterior networks in our SBI approach, significantly reducing the computational costs.

3. Synthetic test case

Before we apply the presented workflow to EIS, we showcase its applicability in a synthetic test case in which all parameters are known.

3.1. Configuration of shelf and flow line

We test our workflow on a 2-D flow tube geometry from which we extract a flow line to infer the prescribed surface and basal accumulation rates as done later in the case of EIS. The flow tube is modeled using `icepack` (Shapero and others, 2021) on a grid $L_x = 125\,\text{km} \times L_y = 10\,\text{km}$, with the along-flow direction x and across-flow direction y. We prescribe a Dirichlet boundary condition at the inflow and lateral boundaries, with a constant thickness of h_0, and a constant along-flow velocity of v_{0x}. The outflow boundary is set to be a static calving front. We initialize with a zero centered, longitudinally symmetric across-flow velocity v_{0y} on the lateral boundaries, resulting in a flow field that has convergence (i.e. mass input) on the center flow line. We prescribe a spatially variable total mass balance $\dot{m}$: In our experiments, we set $v_{ox} = 100\,\text{m a}^{-1}$, $v_{0y} = \pm 20\,\text{m a}^{-1}$ at $y = 0$ and $y = L_y$, respectively,

$$\dot{m} = -0.6 - 0.05\frac{x}{L_x} + 0.3\exp\left(-\left(\frac{x - 0.7L_x}{0.1L_x}\right)^2\right). \quad (11)$$

We let the geometry evolve under the SSA approximation until steady state is reached. From the steady-state ice shelf, we choose a discretization of the central flow line, $\mathbf{x}$, and extract the relevant variables along this flow line to define the internal stratigraphy model (Fig. 3). The numerical values for additional parameters for the spin-up are given in Appendix D. The variables we need are the surface $\mathbf{s}$ and base $\mathbf{f}$ elevations, the along-flow velocities $\mathbf{v}_x$, and the along- and across-flow flux divergences $d(\mathbf{v}_x\mathbf{h})/dx$, $d(\mathbf{v}_y\mathbf{h})/dy$. These define the total mass balance, since:

$$\dot{\mathbf{a}} - \dot{\mathbf{b}} = \frac{d(\mathbf{v}_x\mathbf{h})}{dx} + \frac{d(\mathbf{v}_y\mathbf{h})}{dy}. \quad (12)$$

We then solve the inverse problem which accounts in this case for mass gain through lateral compression. We choose a random sample from the prior distribution as the ground truth, $\dot{\mathbf{a}}_{\text{GT}}$, from which $\dot{\mathbf{b}}_{\text{GT}}$ follows accordingly. The forward model is then sampled to obtain a set of ground truth layer elevations, $\mathbf{e}_o(\mathbf{x})$. From these layer elevations, we choose to perform inference for four layers of ages 50, 100 and 150, and 300 years (labeled 1–4 in ascending order of age). These ages roughly correspond to the range of ages of the IRHs that we expect to observe on ice shelves.

3.2. Inference results

We evaluate the trained neural posterior network on the ground truth isochronal layer of age 50 years. The inferred posterior mean for the surface accumulation rate parameter is close to the ground truth accumulation rate (Fig. 4a,c) with the ground truth lying within the 95% confidence intervals of the posterior distribution.

Next, we evaluate the forward model on samples from the posterior (and prior) distribution to get the respective *predictive distributions*. The prior predictive distribution (Fig. 4b, green) is the distribution over the internal layers generated by simulating the forward model with mass-balance parameters drawn from the prior distribution. The posterior predictive distribution (Fig. 4b, blue) is defined similarly by simulating with mass-balance parameters from the posterior distribution. The posterior predictive

https://doi.org/10.1017/jog.2025.13 Published online by Cambridge University Press

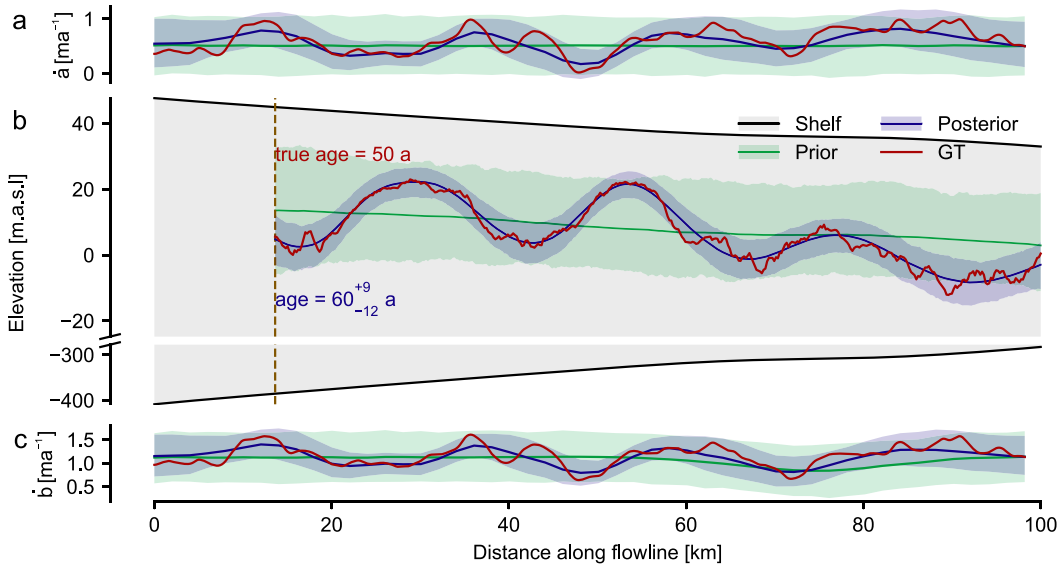

Figure 4. Prior and posterior (predictive) for the synthetic dataset. (a and c) Prior and posterior over surface accumulation and basal melt rates respectively for layer 1 of the synthetic ice shelf, of age 50 years. Solid line is the distribution mean, the shaded region represents the 5th and 95th percentiles. The ground truth (GT) parameters used to generate the reference isochronoal layer are shown in red. (b) Cross section of the ice shelf. Prior and posterior predictive distributions for the layer closest matching the ground truth isochronal layer. The vertical dashed line represents the LMI boundary for this isochronal layer. The posterior predictive reconstructs the observed layer with higher accuracy and lower uncertainty. The posterior predictive distribution of the age of the isochronal layer is 60^{+9}_{-12} years (meaning a median of 60 years, and 16th and 84th percentiles of 48 and 69 years, respectively). The average root-mean-square error (RMSE) relative to the GT isochronal layer is $3.9\,\mathrm{m}$ for the posterior predictive distribution and $11.5\,\mathrm{m}$ for the prior predictive distribution.

matches the ground truth isochronal layer with high fidelity. We calculate the RMSE of the predictive simulations relative to the ground truth layer elevations for 1000 simulations using prior and posterior samples. The average RMSE for the posterior predictive distribution is 3.9 m, compared to 11.5 m for the prior predictive distribution. Uncertainties in the layer elevations are much smaller than those of the prior predictive distribution. This is in contrast to the posterior uncertainty over the mass-balance rates, which is still considerable. This showcases the importance of our uncertainty-aware approach: there is more than one parameterization of accumulation and basal melt rates that can lead to similar isochronal layers.

The posterior uncertainty is also reflected in the inferred age of the isochronal layer. We infer an age of 60^{+9}_{-12} years for this layer (meaning a median of 60 years, and 16th and 84th percentiles of 48 and 69 years, respectively). This value closely matches the age of the ground truth isochronal layer, which was not used during inference. Thus, we have produced an estimate of the age of the layer without requiring time intensive measurements such as ice cores. We report the posterior distributions for deeper synthetic layers in Appendix F.

Finally, while we do not use the isochronal layer elevations outside the LMI boundary for inference, we can still infer the surface accumulation and basal melt rates at these locations. This is because the values of surface accumulation rate and basal melt rate still affect the downstream isochronal layer elevations, and so the observed elevations in the LMI body still contain information about the mass-balance rates upstream of the LMI boundary. Thus, we are still able to infer the mass-balance rates for $x < 15\,\mathrm{km}$.

4. Ekström Ice Shelf

EIS is a medium-sized ice shelf located between the Sörasen and Halvfarryggen Ice Rises in Dronning Maud Land, East Antarctica (Fig. 5c). EIS makes for an appropriate study site since the steady-state assumption likely holds (Drews and others, 2013; Schannwell and others, 2019). Moreover, because of the proximity of the Neumayer station III, numerous observations are available, e.g. ice thickness, surface velocities and most importantly surface accumulation rates, which we will use later for validation.

4.1. Data preprocessing

First, we used `Antarctic Mapping Tools` (Greene and others, 2017), BedMachine Antarctica (Morlighem and others, 2017) and `ITS_LIVE` (Gardner and others, 2018; 2022) to obtain the surface elevation s, thickness h and velocity $\mathbf{v}$ for EIS. In order to define the flow tube domain for EIS, we also used the `itslive_flowline` tool to find two flow lines which formed the side-boundaries of the domain. The other two boundaries of the domain were the grounding line, and a straight line connecting the two flow lines. The straight line was chosen to ensure that the radar transect where data were measured is wholly contained within the flow tube domain.

We preprocessed the raw ice shelf geometry and velocity data prior to evaluating the model. This ensured numerical stability of the forward model. Using the `icepack` package for Python (Shapero and others, 2021), we first smoothed the raw thickness data by solving a regularized minimization problem. We then solved for the best-fitting velocity by fitting a fluidity parameter in an SSA model to the observed velocity and smoothed thickness. The hyperparameters used for preprocessing are given in Appendix D.

4.2. Radar measurements of internal stratigraphy

Internal stratigraphy data along the central flow line of EIS (Fig. 5a) were acquired using a ground-based ground-penetrating radar

https://doi.org/10.1017/jog.2025.13 Published online by Cambridge University Press

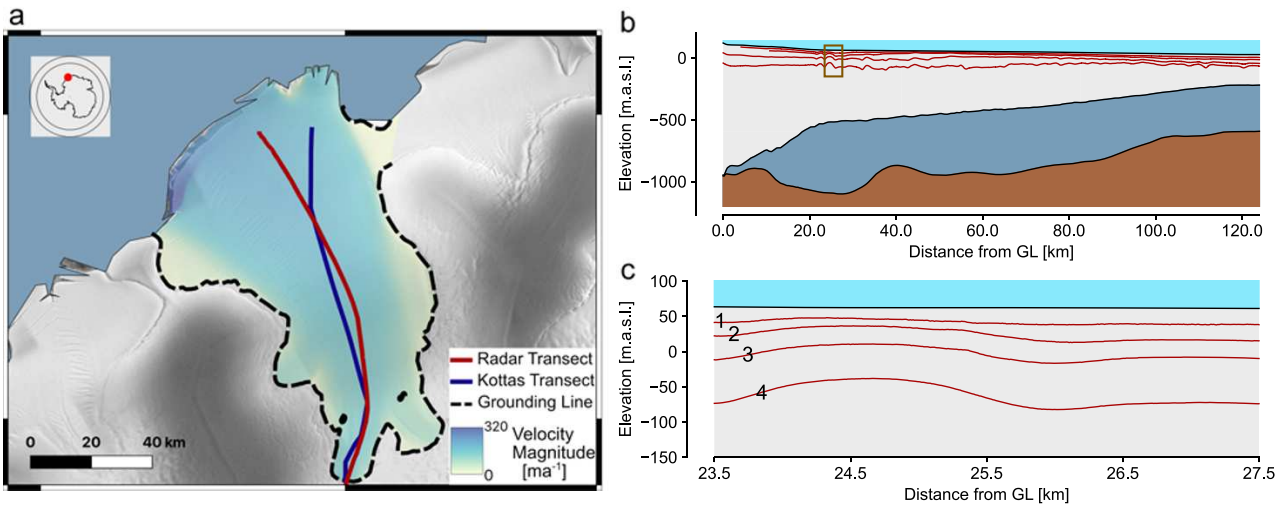

Figure 5. Overview of the Ekström Ice Shelf. (a) Satellite view of Ekström Ice Shelf along with location of the radar transect along the central flow line (red line) and the Kottas traverse (blue line). An independent estimate of surface accumulation via stake arrays is available on Kottas traverse, which we use to validate our results. In our model, we use the velocity data from `ITS_LIVE` (Gardner and others, 2018; 2022). (b) Vertical cross-section view of the radar transect, along with ice surface and base take from BedMachine Antarctica (Morlighem and others, 2017), starting at the grounding line (GL). Red lines indicate four picked internal reflection horizons (IRHs). (c) Zoom in on box in B. The IRHs are numbered 1–4 in order of increasing depth. This plot is shown with the radar data used to label the IRHs in Figure I1.

with a center frequency of 50 MHz (pulseEKKO^{TM} from Sensors & Software) in two consecutive field seasons (2021/22 and 2022/23) with logistic support from the Neumayer III station (Wesche and others, 2016; Wesche and Regnery, 2022). Radar processing was done with ImpDAR (Lilien and others, 2020) and included trace averaging to equidistant spacing (10 m), bandpass filtering (with cutoff frequencies of 20 and 75 MHz), and a topographic correction using the REMA surface elevation (Howat and others, 2019). The latter provides observations consistent with the modeling setup. The radar detects the ice–ocean interface and continuous IRHs down to ∼200 m depth (Fig. 5b and c). Four IRHs were digitized along the entire 130 km long profile using a semi-automatic maximum tracking scheme. The vertical offset off IRHs at the profile junction in the mid-shelf region between both years is much smaller than the radar system's wavelength in ice (∼3.4 m). Consequently, IRHs were connected without adjustments. For the travel time-to-depth conversion, we used a depth-density profile representative for ice shelves of the Dronning Maud Land Coast (Hubbard and others (2013), eqn (1)).

4.3. Inference results

We inspect the prior over the basal melt rates as a validation of our modeling choices. The implicit prior is the same as the prior defined for the surface accumulation, with the mean shifted by the total mass balance on the flow line, $\dot{\mathbf{m}}$. The basal melt rate is larger (up to 4 m a^{-1}) near the grounding line and gradually stabilizes in the along-flow direction to values between 0 and 1 m a^{-1} downstream. This is in agreement with previous estimates for basal melt profiles on this particular ice shelf (Neckel and others, 2012).

We infer the surface accumulation and basal melt rates from IRH 2 in our dataset, which has an average (ice equivalent) depth of 30 m (Fig. 6). The posterior over the surface accumulation rate has uncertainty comparable to that of the prior. However, there is a shift in the overall spatial trend of the accumulation rate; particularly, there is higher surface accumulation rate at ∼20 km from the grounding line. Accumulation rate also increases steadily downstream the flow line. As in the synthetic case, the posterior predictive distribution reproduces the observed IRH with much higher fidelity and confidence than the prior predictive distribution. The average RMSE relative to the observed IRH is 4.6 m for 1000 posterior predictive simulations, compared to 11.8 m for 1000 prior predictive simulations. The posterior predictive produces an independent estimate of the unknown age of the IRH of 84^{+52}_{-30} years (meaning a median of 84 years, and 16th and 84th percentiles of 54 and 136 years, respectively).

Our method can use much deeper IRHs for the inference of accumulation and basal melt rates. For IRH 4 of the observed dataset (of average depth 131 m), the proportion of the IRH that is within the LMI body is smaller. This is due to the unknown boundary condition influencing the IRH elevation at much further points along the flow line. This discarding of data has visible effects on the posteriors over the mass-balance parameters (Fig. 7). These rates are now more similar to the priors for the first 60 km of the transect and only diverge at points further down the ice shelf, where the values of accumulation and basal melt rates affect the dynamics of the IRH. Regardless, the posterior predictive distribution resembles the observed IRH at higher fidelity and precision than the prior predictive. The average RMSE relative to the observed IRH is 10.0 m for the posterior predictive distribution and 16.4 m for the prior predictive distribution. The estimated age of this IRH by our method is 188^{+96}_{-49} years. The uncertainty of the age estimates reasonably increases for deeper IRHs.

5. Discussion

5.1. Posterior mass-balance rates are consistent between IRHs

We compare the four posteriors over the surface accumulation obtained from the Ekström IRH dataset (Fig. 8). The posteriors for the shallower IRHs 1–3 all show a similar qualitative relationship: a local maximum of the accumulation at a distance of ∼20 km from the grounding line, followed by a steady increase in the accumulation downstream. The increase in accumulation at ∼20 km is even identified in the posterior for IRH 3, despite

https://doi.org/10.1017/jog.2025.13 Published online by Cambridge University Press

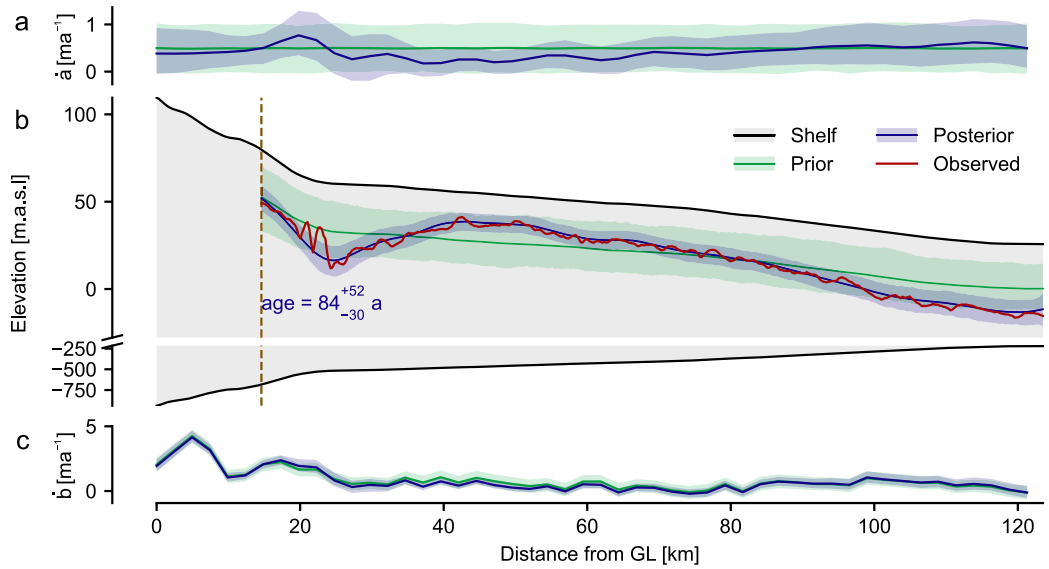

Figure 6. Prior and posterior (predictive) for the Ekström dataset, IRH 2, of average depth 30 m. (a and c) Prior and posterior over surface accumulation and basal melt rates respectively, starting at the grounding line (GL). Solid line is the distribution mean, the shaded region represents the 5th and 95th percentiles. (b) Cross section of the ice shelf. Prior and posterior predictive distributions for the layer closest matching the observed IRH. The vertical dashed line represents the LMI boundary for this IRH. The posterior predictive reconstructs the observed IRH with higher accuracy and lower uncertainty. The posterior predictive distribution of the age of the IRH is 84^{+52}_{-30} years (meaning a median of 84 years, and 16th and 84th percentiles of 54 and 136 years, respectively). The average RMSE relative to the observed IRH is $4.6\,\mathrm{m}$ for the posterior predictive distribution and $11.8\,\mathrm{m}$ for the prior predictive distribution.

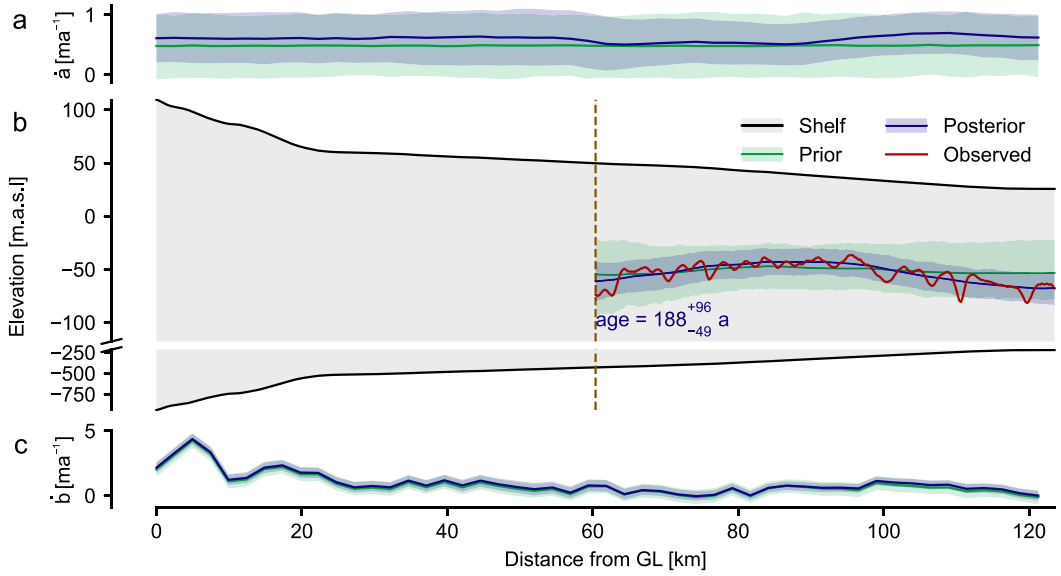

Figure 7. Prior and posterior (predictive) for the Ekström dataset, IRH 4, of average depth 113 m. Same as Figure 6 for the deeper IRH. The posterior predictive distribution of the age of the IRH is 188^{+96}_{-49} years. The average root-mean-square error relative to the observed IRH is $10.0\,\mathrm{m}$ for the posterior predictive distribution and $16.4\,\mathrm{m}$ for the prior predictive distribution.

the LMI boundary being downstream of it, at ∼30 km from the grounding line. This is reasonable, as the mass-balance parameters at a given location affect the flow field downstream of this location, and consequently, the formation of isochronal layers. For IRH 4, the LMI boundary is much further downstream at ∼60 km. Thus, the local surface accumulation maximum at ∼20 km is not found; however, the overall trend of increasing surface accumulation downstream is still identified. There is a corresponding trend in the basal melt rate, as the local basal melt rate still exhibits a maximum at ∼20 km. The reason for this is unknown, but the location corresponds both with the seaward limit of the tidal flexure zone and with the confluence region of ice originating from the eastern tributary. One or both of these factors could alter the basal melt rates inferred at this location. As we will show later (Section 5.3), this local maximum also appears in independent remote-sensing estimates.

The inferred posteriors also allow us to estimate the age of the IRHs. By sampling from the posterior distribution, and evaluating the forward model with the resulting mass-balance parameter samples, we obtain a distribution of isochronal layers similar to

https://doi.org/10.1017/jog.2025.13 Published online by Cambridge University Press

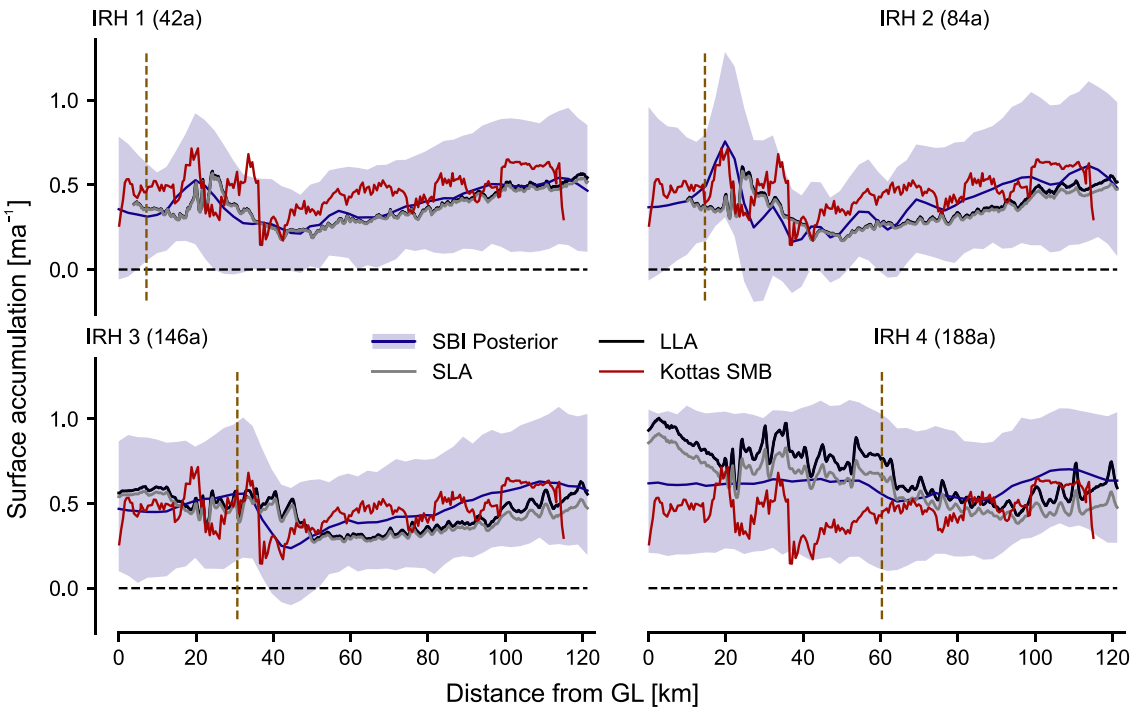

Figure 8. Ekström Ice Shelf—dependence of posterior surface accumulation rate on depth of IRH used for inference. The posteriors are compared to the shallow layer approximation (SLA) and local layer approximation (LLA) (Waddington and others, 2007), and an estimate of the distribution of the accumulation rate based on measurements along the Kottas traverse. See Figure G1 for yearly Kottas measurements. As the real age of the IRHs is not known, the SBI-derived median age is used for the SLA and LLA approximations. Median ages for IRH 1–4 are 42^{+32}_{-12}, 84^{+52}_{-30}, 146^{+52}_{-38} and 188^{+96}_{-49} years. The LMI boundary, representing where the IRH data were masked, is shown with the brown dashed lines.

the observed IRH, with known ages. Thus, we estimate the ages of the four IRHs as 42^{+32}_{-12}, 84^{+52}_{-30}, 146^{+52}_{-38} and 188^{+96}_{-49} years. This is an important finding because IRH age is otherwise only accessible through ice coring. Our results, however, depend on a realistic prior for the surface accumulation and basal melt rates, as defined in Section 2.2.3. Given a miscalibrated prior, the estimated ages would not be reliable (see Fig. H1 for an example). We hypothesize that, given an independent measurement of the IRH age, our approach could further constrain the posterior distributions over the mass-balance parameters.

The consistent spatial patterns of magnitudes of accumulation rates inferred from IRHs 1–4 are supportive of EIS being in steady state over the last hundreds of years but given that steady-state is one of our model assumptions this interpretation needs to be considered with care.

5.2. Comparison to shallow and local layer approximations

To validate our approach, we compare the inferred surface accumulation rate of our experiments with estimates from other methods. First, we computed the shallow layer approximation (SLA) and local layer approximation (LLA) as described in Waddington and others (2007). Given the depth and age of IRH m, the SLA and LLA approximations for the accumulation rate $\dot{\mathbf{a}}$ are defined as

$$\begin{aligned}\dot{\mathbf{a}}^m_{\mathrm{SLA}} &= \frac{1}{\mathcal{A}_m}(\mathbf{s}-\mathbf{e}_m(\mathbf{x})),\\ \dot{\mathbf{a}}^m_{\mathrm{LLA}} &= -\ln\left(1-\frac{\mathbf{s}-\mathbf{e}_m(\mathbf{x})}{\mathbf{h}}\right)\frac{\mathbf{h}}{\mathcal{A}_m},\end{aligned} \tag{13}$$

where $\mathcal{A}_m$ is the age of IRH m. Intuitively, the SLA takes the ice thickness above layer m and divides it by the layer age, whereas LLA accounts for strain thinning assuming a linear vertical velocity profile (which is often the case for ice-shelf flow). Since the age of the observed IRHs is not known, we use the median age of the posterior predictive distribution results. As expected, we observe that both SLA and LLA closely match the SBI posterior mean accumulation rate for the shallow IRHs of median estimated ages 42 and 84 years (Fig. 8). As the strain rates of the flow are small, the relatively shallow IRHs (mean ice equivalent depth of 30 m) have not notably deformed, and hence the assumptions of SLA and LLA are appropriate. However, for the deeper IRHs 3 and 4 of estimated ages 146 and 188 years, we see that both SLA and LLA estimates diverge from our posterior mean accumulation rate. This shows that more involved approaches are required when using deeper IRHs for inference. For deeper IRHs where the SLA and LLA no longer applied, Steen-Larsen and others (2010) inferred the surface accumulation rates on grounded ice using a Monte Carlo approach. By treating the age of the IRH as an additional parameter to infer, they were able to identify the age of the IRH with high confidence. Extensions of our approach could incorporate this parameterization to reduce the uncertainty of the inferred IRH.

5.3. Comparison with independent estimates of surface accumulation and basal melting

For the Ekström transect comparison data are provided by repeat readings of accumulation stakes in 500 m spacing along the nearby Kottas traverse (Fig. 8). Yearly readings are available in the period 1996–2005 and on a yearly to three-yearly interval between 2014 and 2023 (Mengert, 2018). We use this dataset to construct a direct estimate of time-averaged surface accumulation rate along the central flow line transect over these periods. For this, we project

https://doi.org/10.1017/jog.2025.13 Published online by Cambridge University Press

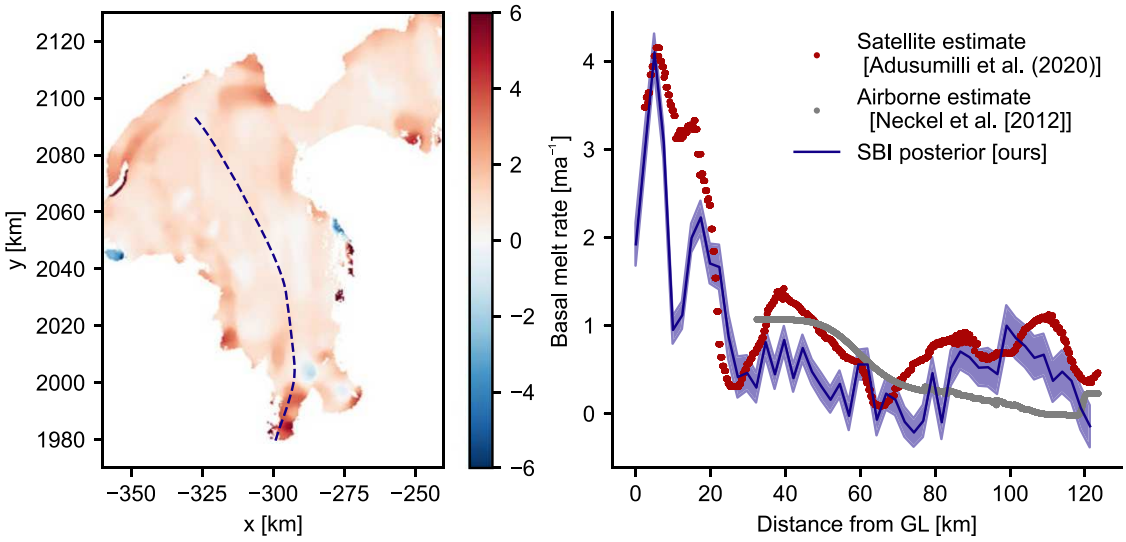

Figure 9. Basal melting rates comparison. (a) Map of basal melt rates for Ekström Ice Shelf, using data from Adusumilli and others (2020). (b) Comparison of inferred basal melt rates from IRH 1 to independent estimates of basal melt rates, calculated on the flow line transect.

the measurements from the Kottas traverse to the flow line transect, taking into account an increased uncertainty for increasing projection distance (see Appendix G for details).

The Kottas traverse accumulation measurements closely match the posterior means of our approach (Fig. 8) for IRHs 1 and 2. As the accumulation rate measurements on the Kottas traverse span the past 26 years, it may not be a good validation for the deeper IRHs. Regardless, the Kottas accumulation rate measurements lie within the posterior uncertainty for IRHs 3 and 4. These comparisons further corroborate our approach and highlight the advantages of uncertainty-aware methods, especially as the measured accumulation rates also varied considerably year-to-year (Appendix G).

We also compare our inferred basal melt rates with independent measurements of basal melt rates. In Fig. 9, we show the basal melt rates inferred from IRH 1 for the Ekström dataset, in comparison to independent estimates of basal melt rates through satellite altimetry data (Adusumilli and others, 2020), and through airborne radar measurements of the ice-shelf thickness (Neckel and others, 2012). We observe a quantitative match between our posterior basal melting rate and the estimates from Adusumilli and others (2020). The ice-shelf wide melt rate estimates show overall comparatively little spatial variability. The most notable difference is that our basal melt rates show more variability in the first 20 km of the profile, and this could be due to the proximity of the grounding zone where the SSA approximation does not hold. However, we note that also the satellite-derived estimates show this oscillation in basal melt rates albeit with a smaller magnitude. The estimates from Neckel and others (2012) excluded the grounding zone area but otherwise show a good match in magnitude but with much less spatial variability. This is because they decided to apply spatial smoothing the degree of which could be revisited given the new results derived here.

The good quantitative match with independently collected data both for surface accumulation and basal melting increases our confidence for our inferred surface accumulation and basal melt rates. However, these results are limited by some of our modeling assumptions, which we discuss next.

5.4. Limitations of modeling approach

The fidelity at which the posterior predictive distributions reproduce the observed IRHs of EIS (Figs. 6 and 7) supports our modeling choices for this ice shelf, as the combination of the forward model and accumulation rate prior distribution are sufficiently expressive to reproduce the IRHs.

However, our inferred surface accumulation and basal melt rates rely on the modeling assumption that the internal radar data are collected on a flow-line transect. This assumption is required to conclude that the ice observed in the internal stratigraphy is indeed the accumulated ice modeled in our domain, as commonly assumed in the literature (Waddington and others, 2007; Steen-Larsen and others, 2010; Theofilopoulos and Born, 2023). Similarly, in the context of ice rises, it is often assumed that the transverse velocity is negligible relative to the vertical velocity of the ice, so that the surface accumulation rate can be inferred along the same transect as the IRH data (Callens and others, 2016; Koch and others, 2024). However, many of the available IRH measurements do not align with flow lines of the ice sheet. In this case, our assumption would not be valid.

Because our observations are on a flow line transect only, it is difficult to judge to what extent unidentified three-dimensional (3-D) effects overprint our analysis. Previous approaches (Pattyn and others, 2012) have had similar limitations because ice-flow divergence and/or convergence could not be predicted by their 2-D forward model. They concluded that their inferred basal melt rates which best matched the radar stratigraphy would be a lower boundary because ice flow on the ice shelf was convergent. In our case, we do correct for the observed convergence from observed surface velocities along the flow tube. The normal component of ice-flow is always $< 20\,\mathrm{m\,a^{-1}}$ and often $< 1\,\mathrm{m\,a^{-1}}$. This is small compared to the along-flow velocities and also compared to the total surface mass balance accumulated along the flow tube. Together with the empirical validation with independently collected surface accumulation and basal melt rates, this increases our confidence that our modeling approach yields trustworthy results. Yet, a more rigorous quantification of 3-D effects, for example, in a

https://doi.org/10.1017/jog.2025.13 Published online by Cambridge University Press

synthetic study using a 3-D forward model can provide further validation. Others have made progress in this direction, for example, Wolovick and others (2021) consider a 3-D steady-state ice sheet and jointly infer the temporally averaged accumulation rate and geothermal heat flow. This is done by using more radar attributes in addition to stratigraphy such as existence or absence of subglacial water and/or basal freeze-on.

Common to most previous studies is the steady-state assumption which is imposed because the inclusion of transient ice thickness changes increases the model parameter space to a degree which cannot easily be solved in the inverse problem, particularly with quantified uncertainties over the inferred parameters. Ways forward in this regard could be deterministic gradient descent schemes with explicitly calculating sensitivity matrices as suggested by Theofilopoulos (2022); Theofilopoulos and Born (2023) and this will be an important step forward to better exploit the growing IRH archive for ice-sheet modeling.

5.5. SBI as a tool for geoscientific inversion problems

The inverse problem tackled in this work typifies geoscientific inverse problems, as the forward model is defined in terms of a partial differential equation, and the parameters are high dimensional and vary in space. Hence, it is valuable to compare the SBI approach in this case to the wide variety of methods and algorithms that have been developed to solve geoscientific inverse problems. In the remainder of the section, we discuss NPE as used in our work. However, there exist other variants of SBI with relative advantages and disadvantages, depending on the problem setting (Cranmer and others, 2020; Lueckmann and others, 2021). In particular, we provide a brief discussion of the neural likelihood estimation (NLE) variant in Appendix C.4.

The SBI approach as presented here has two key features. First, we estimate the Bayesian posterior distribution, providing quantitative uncertainty estimates. Modeling uncertainty is important as it can highly influence and propagate to future modeling predictions. Additionally, locations of high uncertainty show areas requiring further study, helping to guide future work. This is in contrast to deterministic inversion methods, which do not estimate uncertainty, or likelihood-based inference methods which are not possible when the likelihood defined by the simulator is not known. Thus, approximate Bayesian methods and SBI in particular can be applied to a larger class of inference problems. Second, a unique advantage of *single round* SBI methods (Cranmer and others, 2020) such as NPE (as used in our study) is *amortization*. Our method as presented here is not yet fully amortized, instead amortizing the vast majority of the computational cost, as preprocessing relies on the observed value of X. In order to train the density estimator $q_\phi(\dot{\mathbf{a}}|X_k^m)$, we first calculate X_k^m for each simulation dependent on the value of X_o^m. Our method still amortizes the cost of simulating the forward model many times, which is by far the largest computational cost in the approach. In the Ekström example, we have evaluated the forward model a total of 190 000 times, accounting for ∼99% of the total computation cost (Appendix E). This amortization is specific to the geometry and velocity of the EIS; different geometries and velocities change the dependence of the internal layers on the mass-balance parameters, which would require simulating from a new model.

On the other hand, SBI faces some limitations as an inference tool. Primarily, SBI methods are known to require a large number of simulations to be trained (Lueckmann and others, 2021). This problem suffers from the curse of dimensionality—the number of simulations required scales exponentially with the number of parameters we are trying to infer (in this work, we limited the number of parameters to 50). This is particularly challenging for geoscientific problems, where typically the parameters of interest vary spatially (and temporally), and thus the number of parameters can grow very large. The SBI approach needs to be adapted to more efficiently represent high-dimensional, spatially varying parameters θ at high resolutions. Some potential approaches are polynomial or spectral representations. Future work should also explore variants of SBI that are better suited to high-dimensional or even continuous parameters (Ramesh and others, 2022; Geffner and others, 2023). Finally, SBI works under the assumption that the forward model is well-specified, meaning that given samples from the prior, it can generate simulations closely resembling the observation. The posteriors obtained by SBI can be strongly biased when this is not the case (Cannon and others, 2022). Work to address this concern has been done, e.g. by incorporating the model mismatch into the forward model (Ward and others, 2022), as done in our work using the calibrated noise model. However, designing and calibrating such noise models for each inference task are challenging, and a standard approach for addressing model mismatch does not yet exist.

6. Conclusions

We presented a novel approach for inferring the spatially varying surface accumulation and basal melt rates along ice-shelf flow lines from radar measurements of their internal stratigraphy. We validated the method on a synthetic ice shelf example and inferred the surface accumulation and basal melt rates along a flow line in EIS, Antarctica. We separately inferred the mass-balance parameters from four different IRHs. The inferred distributions were further validated by independent stake array measurements of surface accumulation rates uniquely available in Ekström Ice Shelf. Using our approach, we were able to estimate the otherwise unknown age of the IRHs as 42^{+32}_{-12}, 84^{+52}_{-30}, 146^{+52}_{-38} and 188^{+96}_{-49} years. The presented approach can be transferred to other Antarctic ice shelves and also to other flow regimes such as grounded ice. A strength of our approach is the principled uncertainty estimates in the inferred surface accumulation and basal melt rates. These uncertainty estimates can be integrated in future projections of the Antarctic ice sheet (Verjans and others, 2022; Ultee and others, 2024). We identified avenues for future work as more can be learned by relaxing the steady-state assumption on the ice shelf. The forward model and inference framework should be adapted to account for potential transient signals in the mass-balance parameters.

This work was an example use case of SBI for a geoscientific inverse problem. We showcased the strengths of SBI as a likelihood-free approach to approximate the Bayesian posterior, amortizing the cost of simulating the forward model many times. SBI can become more applicable to such inverse problems involving spatially (and temporally) varying parameters if it can be extended to deal with the challenge of high-dimensional parameter inference.

Finally, our approach highlights the value of internal stratigraphy measurements. Initiatives to map the Antarctic-wide internal stratigraphy (e.g. Bingham and others, 2024) can provide invaluable data toward uncovering the history of the Antarctic ice sheet. Sophisticated inference methods could be combined with such a dataset to provide a new, independent, Antarctica-wide parameterization of accumulation and basal melt rate histories.

https://doi.org/10.1017/jog.2025.13 Published online by Cambridge University Press

Data availability statement. The extracted IRH elevations along the Ekström transect are available in Oraschewski and others (2024a). The processed radar data are additionally available in Oraschewski and others (2024b). Simulation data available in Moss and others (2023).

Software availability. Code for preprocessing Ekström Ice Shelf data and generating synthetic ice shelf data is available in Moss and others (2024a). Code for layer tracing forward model and simulation-based inference workflow is available in Moss and others (2024b).

Acknowledgements. The authors would like to thank Daniel Shapero for his inputs on use of icepack. The authors would also like to thank Andreas Born and Therese Riekch for insightful discussions on the implementation of the layer tracing solver for calculating internal stratigraphy. We acknowledge excellent logistic support from staff at Neumayer Station III and the GrouZe team on-site.

This work was funded by the German Research Foundation (DFG) under Germany's Excellence Strategy—EXC number 2064/1—390727645 and SFB 1233 'Robust Vision' (276693517) and the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. Reinhard Drews and Vjeran Višnjević were supported by an Emmy Noether grant of the Deutsche Forschungsgemeinschaft (DR 822/3-1). We acknowledge the support by the German Academic Scholarship Foundation to Falk M. Oraschewski. Field observations were supported by the Alfred Wegener Institute through logistic grants AWI_ANT_23 (Drews) and AWI_ANT_8 (Eisen). We acknowledge support from the Open Access Publication Fund of the University of Tübingen. Guy Moss is a member of the International Max Planck Research School for Intelligent Systems (IMPRS-IS).

References

Adusumilli S, Fricker HA, Medley B, Padman L and Siegfried MR (2020) Interannual variations in meltwater input to the Southern Ocean from Antarctic ice shelves. *Nature Geoscience 2020 13:9* **13**, 616–620. doi: 10.1038/s41561-020-0616-z

Agosta C and 10 others (2019) Estimation of the Antarctic surface mass balance using the regional climate model MAR (1979–2015) and identification of dominant processes. *The Cryosphere* **13**, 281–296. doi: 10.5194/tc-13-281-2019

Allgeier J and Cirpka OA (2023) Surrogate-model assisted plausibility-check, calibration, and posterior-distribution evaluation of subsurface-flow models. *Water Resources Research* **59**, 1–18. doi: 10.1029/2023WR034453

Asay-Davis XS and 13 others (2016) Experimental design for three interrelated marine ice sheet and ocean model intercomparison projects: MISMIP v. 3 (MISMIP+), ISOMIP v. 2 (ISOMIP+) and MISOMIP v. 1 (MISOMIP1). *Geoscientific Model Development* **9**, 2471–2497. doi: 10.5194/GMD-9-2471-2016

Berger S, Drews R, Helm V, Sun S and Pattyn F (2017) Detecting high spatial variability of ice shelf basal mass balance, Roi Baudouin Ice Shelf, Antarctica. *The Cryosphere* **11**, 2675–2690. doi: 10.5194/tc-11-2675-2017

Bindschadler R and 17 others (2011) Getting around Antarctica: New high-resolution mappings of the grounded and freely-floating boundaries of the Antarctic ice sheet created for the International Polar Year. *Cryosphere* **5**, 569–588. doi: 10.5194/TC-5-569-2011

Bingham RG and 53 others (2024) Review Article: Antarctica's internal architecture: Towards a radiostratigraphically–informed age–depth model of the Antarctic ice sheets. *EGUsphere* **2024**, 1–66. doi: 10.5194/egusphere-2024-2593

Blei DM, Kucukelbir A and McAuliffe JD (2017) Variational inference: A review for statisticians. *Journal of the American Statistical Association* **112**, 859–877. doi: 10.1080/01621459.2017.1285773

Born A (2017) Tracer transport in an isochronal ice-sheet model. *Journal of Glaciology* **63**(237), 22–38. doi: 10.1017/JOG.2016.111

Born A and Robinson A (2021) Modeling the greenland englacial stratigraphy. *Cryosphere* **15**, 4539–4556. doi: 10.5194/TC-15-4539-2021

Brinkerhoff D, Aschwanden A and Fahnestock M (2021) Constraining subglacial processes from surface velocity observations using surrogate-based Bayesian inference. *Journal of Glaciology* **67**(263), 385–403. doi: 10.1017/jog.2020.112

Burgard C, Jourdain NC, Reese R, Jenkins A and Mathiot P (2022) An assessment of basal melt parameterisations for Antarctic ice shelves. *Cryosphere* **16**, 4931–4975. doi: 10.5194/TC-16-4931-2022

Callens D, Drews R, Witrant E, Philipp M and Pattyn F (2016) Temporally stable surface mass balance asymmetry across an ice rise derived from radar internal reflection horizons through inverse modeling. *Journal of Glaciology* **62**(233), 525–534. doi: 10.1017/jog.2016.41

Cannon P, Ward D and Schmon SM (2022) Investigating the impact of model misspecification in neural simulation-based inference. *ArXiv e-prints, arXiv:2209.01845*. doi: 10.48550/arXiv.2209.01845

Catania G, Hulbe C and Conway H (2010) Grounding-line basal melt rates determined using radar-derived internal stratigraphy. *Journal of Glaciology* **56**(197), 545–554. doi: 10.3189/002214310792447842

Cranmer K, Brehmer J and Louppe G (2020) The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences* **117**, 30055–30062. doi: 10.1073/pnas.1912789117

Das I and 11 others (2020) Multidecadal basal melt rates and structure of the Ross Ice Shelf, Antarctica, using airborne ice penetrating radar. *Journal of Geophysical Research: Earth Surface* **125**, e2019JF005241. doi: 10.1029/2019JF005241

Depoorter MA and 6 others (2013) Calving fluxes and basal melt rates of Antarctic ice shelves. *Nature* **502**, 89–92. doi: 10.1038/nature12567

Drews R (2015) Evolution of ice-shelf channels in Antarctic ice shelves. *The Cryosphere* **9**, 1169–1181. doi: 10.5194/TC-9-1169-2015

Drews R and 6 others (2016) Constraining variable density of ice shelves using wide-angle radar measurements. *The Cryosphere* **10**, 811–823. doi: 10.5194/tc-10-811-2016

Drews R, Martín C, Steinhage D and Eisen O (2013) Characterizing the glaciological conditions at Halvfarryggen ice dome, Dronning Maud Land, Antarctica. *Journal of Glaciology* **59**(213), 9–20. doi: 10.3189/2013JoG12J134

Drews R, Matsuoka K, Martín C, Callens D, Bergeot N and Pattyn F (2015) Evolution of Derwael ice rise in Dronning Maud Land, Antarctica, over the last millennia. *Journal of Geophysical Research: Earth Surface* **120**, 564–579. doi: 10.1002/2014JF003246

Durkan C, Bekasov A, Murray I and Papamakarios G (2019) Neural Spline Flows. Advances in Neural Information Processing Systems. In. Curran Associates Inc., Volume 32.

Dutrieux P and 6 others (2014) Basal terraces on melting ice shelves. *Geophysical Research Letters* **41**, 5506–5513. doi: 10.1002/2014GL060618

Eisen O and 15 others (2008) Ground-based measurements of spatial and temporal variability of snow accumulation in East Antarctica. *Reviews of Geophysics* **46**, 2. doi: 10.1029/2006RG000218

Eisen O (2008) Inference of velocity pattern from isochronous layers in firn, using an inverse method. *Journal of Glaciology* **54**(187), 613–630. doi: 10.3189/002214308786570818

Eisen O, Nixdorf U, Wilhelms F and Miller H (2004) Age estimates of isochronous reflection horizons by combining ice core, survey, and synthetic radar data. *Journal of Geophysical Research: Solid Earth* **109**, B4. doi: 10.1029/2003JB002858

Gallée H and Schayes G (1994) Development of a three-dimensional meso-γ primitive equation model: Katabatic winds simulation in the area of Terra Nova Bay, Antarctica. *Monthly Weather Review* **122**, 671–685.

Gardner A and 6 others (2018) Increased West Antarctic and unchanged East Antarctic ice discharge over the last 7 years. *The Cryosphere* **12**, 521–547. doi: 10.5194/tc-12-521-2018

Gardner A, Fahnestock M and Scambos T (2022) *MEaSUREs ITS_LIVE Landsat Image-Pair Glacier and Ice Sheet Surface Velocities*, version 1. doi: 10.5067/IMR9D3PEI28U

Geffner T, Papamakarios G and Mnih A (2023) Compositional score modeling for simulation-based inference. *Proceedings of Machine Learning Research*. PMLR, pp. 11098–11116, Volume 202.

Gladstone R and 12 others (2021) The framework for ice sheet-ocean coupling (fisoc) v1.1. *Geoscientific Model Development* **14**, 889–905. doi: 10.5194/GMD-14-889-2021

Goelzer H, Huybrechts P, Loutre MF and Fichefet T (2016) Last interglacial climate and sea-level evolution from a coupled ice sheet–climate model. *Climate of the Past* **12**, 2195–2213. doi: 10.5194/cp-12-2195-2016

Goldberg DN, Gourmelen N, Kimura S, Millan R and Snow K (2019) How accurately should we model ice shelf melt rates? *Geophysical Research Letters* **46**, 189–199. doi: 10.1029/2018GL080383

Goldberg DN and Holland PR (2022) The relative impacts of initialization and climate forcing in coupled ice sheet-ocean modeling: Application to Pope, Smith, and Kohler Glaciers. *Journal of Geophysical Research: Earth Surface* **127**, e2021JF006570. doi: 10.1029/2021JF006570

Greenberg D, Nonnenmacher M and Macke J (2019) Automatic posterior transformation for likelihood-free inference. *Proceedings of Machine Learning Research*. PMLR, pp. 2404–2414, Volume 97.

Greene CA, Gwyther DE and Blankenship DD (2017) Antarctic mapping tools for MATLAB. *Computers & Geosciences* **104**, 151–157. doi: 10.1016/j.cageo.2016.08.003

Greve R and Blatter H (2009) *Dynamics of Ice Sheets and Glaciers*. Berlin Heidelberg: Springer. doi: 10.1007/978-3-642-03415-2

Gudmundsson GH, Paolo FS, Adusumilli S and Fricker HA (2019) Instantaneous Antarctic ice sheet mass loss driven by thinning ice shelves. *Geophysical Research Letters* **46**, 13903–13909. doi: 10.1029/2019GL085027

Henry ACJ and 6 others (2023) Predicting the three-dimensional age-depth field of an ice rise. *Authorea*. doi: 10.22541/essoar.169230234.44865946/v1

Holschuh N, Parizek BR, Alley RB and Anandakrishnan S (2017) Decoding ice sheet behavior using englacial layer slopes. *Geophysical Research Letters* **44**, 5561–5570. doi: 10.1002/2017GL073417

Howat IM, Porter C, Smith BE, Noh MJ and Morin P (2019) The reference elevation model of Antarctica. *The Cryosphere* **13**, 665–674. doi: 10.5194/tc-13-665-2019

Hubbard B and 6 others (2013) Ice shelf density reconstructed from optical televiewer borehole logging. *Geophysical Research Letters* **40**, 5882–5887. doi: 10.1002/2013GL058023

Hull R and 7 others (2022) Using simulation-based inference to determine the parameters of an integrated hydrologic model: A case study from the upper Colorado River basin. *Hydrology and Earth System Sciences Discussions* **2022**, 1–38. doi: 10.5194/hess-2022-345

Kingma DP and Ba J (2015) Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Kobyzev I, Prince SJ and Brubaker MA (2019) Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43**, 3964–3979. doi: 10.1109/TPAMI.2020.2992934

Koch I and 9 others (2024) Radar internal reflection horizons from multi-system data reflect ice dynamic and surface accumulation history along the Princess Ragnhild Coast, Dronning Maud Land, East Antarctica. *Journal of Glaciology* **70**, e18. doi: 10.1017/jog.2023.93

Lenaerts JTM, Medley B, van den Broeke MR and Wouters B (2019) Observing and modeling ice sheet surface mass balance. *Reviews of Geophysics* **57**, 376–420. 10.1029/2018RG000622

Leysinger Vieli GJMC, Hindmarsh RCA, Siegert MJ and Bo S (2011) Time-dependence of the spatial pattern of accumulation rate in East Antarctica deduced from isochronic radar layers using a 3-D numerical ice flow model. *Journal of Geophysical Research: Earth Surface* **116**, F2. doi: 10.1029/2010JF001785

Lilien DA, Hills BH, Driscol J, Jacobel R and Christianson K (2020) ImpDAR: An open-source impulse radar processor. *Annals of Glaciology* **61**(81), 114–123. doi: 10.1017/aog.2020.44

Linde N, Renard P, Mukerji T and Caers J (2015) Geological realism in hydrogeological and geophysical inverse modeling: A review. *Advances in Water Resources* **86**, 86–101. doi: 10.1016/j.advwatres.2015.09.019

Looyenga H (1965) Dielectric constants of heterogeneous mixtures. *Physica* **31**, 401–406. doi: 10.1016/0031-8914(65)90045-5

Lueckmann JM, Boelts J, Greenberg D, Goncalves P and Macke J (2021) Benchmarking simulation-based inference. *Proceedings of Machine Learning Research*. PMLR, 130, 343–351.

Lueckmann JM, Goncalves PJ, Bassetto G, Öcal K, Nonnenmacher M and Macke JH (2017) Flexible statistical inference for mechanistic models of neural dynamics. *Advances in Neural Information Processing Systems*. In Curran Associates Inc., Volume 30.

MacGregor JA, Matsuoka K, Koutnik MR, Waddington ED, Studinger M and Winebrenner DP (2009) Millennially averaged accumulation rates for the Vostok Subglacial Lake region inferred from deep internal layers. *Annals of Glaciology* **50**(51), 25–34. doi: 10.3189/172756409789097441

Marsh OJ and 6 others (2016) High basal melting forming a channel at the grounding line of Ross Ice Shelf, Antarctica. *Geophysical Research Letters* **43**, 250–255. doi: 10.1002/2015GL066612

Mengert M (2018) Spatial and Temporal Variation of Snow Accumulation Along Kottas Traverse, Dronning Maud land, East Antarctica. doi: 10013/epic.29be472c-08be-495a-ba18-8c8e295101a5

Morland LW (1984) Thermomechanical balances of ice sheet flows. *Geophysical & Astrophysical Fluid Dynamics* **29**, 237–266. doi: 10.1080/03091928408248191

Morlighem M and 31 others (2017) BedMachine v3: Complete bed topography and ocean bathymetry mapping of Greenland from multibeam echo sounding combined with mass conservation. *Geophysical Research Letters* **44**, 11,051–11,061. doi: 10.1002/2017gl074954

Moss G and 6 others (2023) Assets for "Simulation-Based Inference of Surface Accumulation and Basal Melt Rates of an Antarctic Ice shelf from Isochronal Layers". doi: 10.5281/zenodo.10245153

Moss G and 6 others (2024a) preprocessing-ice-data. doi: 10.5281/zenodo.11440869

Moss G and 6 others (2024b) sbi-ice. doi: 10.5281/zenodo.11440807

Neckel N, Drews R, Rack W and Steinhage D (2012) Basal melting at the Ekström Ice Shelf, Antarctica, estimated from mass flux divergence. *Annals of Glaciology* **53**(60), 294–302. doi: 10.3189/2012AoG60A167

Nicholls KW, Corr HF, Stewart CL, Lok LB, Brennan PV and Vaughan DG (2015) A ground-based radar for measuring vertical strain rates and time-varying basal melt rates in ice sheets and shelves. *Journal of Glaciology* **61**(230), 1079–1087. doi: 10.3189/2015JOG15J073

Omagbon J and 8 others (2021) Case studies of predictive uncertainty quantification for geothermal models. *Geothermics* **97**, 102263. doi: 10.1016/j.geothermics.2021.102263

Oraschewski FM, Moss G, Koch I, Ershadi MR, Eisen O and Drews R (2024a) *Ground-Penetrating Radar Data (50MHz) and Internal Reflection Horizons Along the Central Flowline of Ekström Ice Shelf, Dronning Maud Land, East Antarctica*. doi: 10.1594/PANGAEA.965143

Oraschewski FM, Moss G, Koch I, Ershadi MR, Eisen O and Drews R (2024b) *Ground-Penetrating Radar Data (50MHz) and Internal Reflection Horizons Along the Central Flowline of Ekström Ice Shelf, Dronning Maud Land, East Antarctica - NetCDF File*. doi: 10.1594/PANGAEA.965141

Overcast I and 17 others (2021) A unified model of species abundance, genetic diversity, and functional diversity reveals the mechanisms structuring ecological communities. *Molecular Ecology Resources* **21**, 2782–2800. doi: 10.1111/1755-0998.13514

Papamakarios G and Murray I (2016) Fast ϵ -free Inference of Simulation Models with Bayesian Conditional Density Estimation. *Advances in Neural Information Processing Systems*. In Curran Associates Inc., Volume 29.

Papamakarios G, Nalisnick ET, Rezende DJ, Mohamed S and Lakshminarayanan B (2019a) Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research* **22**, 1–64.

Papamakarios G, Sterratt D and Murray I (2019b) Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* **89**. In PMLR

Pattyn F and 8 others (2012) Melting and refreezing beneath Roi Baudouin Ice Shelf (East Antarctica) inferred from radar, GPS, and ice core data. *Journal of Geophysical Research: Earth Surface* **117**, F4. doi: 10.1029/2011JF002154

Pattyn F, Favier L, Sun S and Durand G (2017) Progress in numerical modeling of Antarctic ice-sheet dynamics. *Current Climate Change Reports* **3**, 174–184. doi: 10.1007/s40641-017-0069-7

Pratap B and 7 others (2022) Three-decade spatial patterns in surface mass balance of the Nivlisen Ice Shelf, central Dronning Maud Land, East Antarctica. *Journal of Glaciology* **68**(267), 174–186. doi: 10.1017/JOG.2021.93

https://doi.org/10.1017/jog.2025.13 Published online by Cambridge University Press

Ramesh P and 6 others (2022) GATSBI: Generative adversarial training for simulation-based inference. In *International Conference on Learning Representations*.
Rasmussen CE and Williams CKI (2005) *Gaussian Processes for Machine Learning*. The MIT Press.
Reese R, Gudmundsson GH, Levermann A and Winkelmann R (2017) The far reach of ice-shelf thinning in Antarctica. *Nature Climate Change* **8**, 53–57. doi: 10.1038/s41558-017-0020-x
Schannwell C, Drews R, Ehlers TA, Eisen O, Mayer C and Gillet-Chaulet F (2019) Kinematic response of ice-rise divides to changes in ocean and atmosphere forcing. *The Cryosphere* **13**, 2673–2691. doi: 10.5194/tc-13-2673-2019
Shapero DR, Badgeley JA, Hoffman AO and Joughin IR (2021) icepack: a new glacier flow modeling package in Python, version 1.0. *Geoscientific Model Development* **14**, 4593–4616. doi: 10.5194/gmd-14-4593-2021
Steen-Larsen HC, Waddington ED and Koutnik MR (2010) Formulating an inverse problem to infer the accumulation-rate pattern from deep internal layering in an ice sheet using a Monte Carlo approach. *Journal of Glaciology* **56**(196), 318–332. doi: 10.3189/002214310791968476
Sun S, Hattermann T, Pattyn F, Nicholls KW, Drews R and Berger S (2019) Topographic shelf waves control seasonal melting near Antarctic ice shelf grounding lines. *Geophysical Research Letters* **46**, 9824–9832. doi: 10.1029/2019GL083881
Sutter J, Fischer H and Eisen O (2021) Investigating the internal structure of the Antarctic ice sheet: The utility of isochrones for spatiotemporal ice-sheet model calibration. *Cryosphere* **15**, 3839–3860. doi: 10.5194/tc-15-3839-2021
Symes WW (2009) The seismic reflection inverse problem. *Inverse Problems* **25**, 123008. doi: 10.1088/0266-5611/25/12/123008
Tarasov L, Dyke AS, Neal RM and Peltier W (2012) A data-calibrated distribution of deglacial chronologies for the North American ice complex from glaciological modeling. *Earth and Planetary Science Letters* **315–316**, 30–40. doi: 10.1016/j.epsl.2011.09.010
Tebaldi C and Sansó B (2008) Joint projections of temperature and precipitation change from multiple climate models: A hierarchical Bayesian approach. *Journal of the Royal Statistical Society Series A: Statistics in Society* **172**, 83–106. doi: 10.1111/j.1467-985X.2008.00545.x
Tejero-Cantero A and 7 others (2020) SBI: A toolkit for simulation-based inference. *Journal of Open Source Software* **5**, 2505. doi: 10.21105/joss.02505
Theofilopoulos A (2022) *Reconstructing surface mass balance from the englacial stratigraphy of the Greenland Ice Sheet*. Ph.D. thesis, The University of Bergen.
Theofilopoulos A and Born A (2023) Sensitivity of isochrones to surface mass balance and dynamics. *Journal of Glaciology* **69**(274), 311–323. doi: 10.1017/jog.2022.62
Ultee L, Robel AA and Castruccio S (2024) A stochastic parameterization of ice sheet surface mass balance for the Stochastic Ice-Sheet and Sea-Level System Model (StISSM v1.0). *Geoscientific Model Development* **17**, 1041–1057. doi: 10.5194/gmd-17-1041-2024
Vankova I and Nicholls KW (2022) Ocean variability beneath the Filchner-Ronne ice shelf inferred from basal melt rate time series. *Journal of Geophysical Research: Oceans* **127**, e2022JC018879. doi: 10.1029/2022JC018879
van Wessem JM and 18 others (2018) Modelling the climate and surface mass balance of polar ice sheets using RACMO2 – Part 2: Antarctica (1979–2016). *The Cryosphere* **12**, 1479–1498. doi: 10.5194/tc-12-1479-2018
Verjans V, Robel AA, Seroussi H, Ultee L and Thompson AF (2022) The stochastic ice-sheet and sea-level system model v1.0 (STISSM v1.0). *Geoscientific Model Development* **15**, 8269–8293. doi: 10.5194/gmd-15-8269-2022
Virtanen P and 34 others (2020) SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–272. doi: 10.1038/s41592-019-0686-2
Visnjevic V and 6 others (2022) Predicting the steady-state isochronal stratigraphy of ice shelves using observations and modeling. *The Cryosphere* **16**, 4763–4777. doi: 10.5194/tc-16-4763-2022
Waddington ED, Neumann TA, Koutnik MR, Marshall HP and Morse DL (2007) Inference of accumulation-rate patterns from deep layers in glaciers and ice sheets. *Journal of Glaciology* **53**(183), 694–712. doi: 10.3189/002214307784409351
Ward D, Cannon P, Beaumont M, Fasiolo M and Schmon S (2022) Robust Neural Posterior Estimation and Statistical Model Criticism. *Advances in Neural Information Processing Systems*. In Curran Associates Inc., pp. 33845–33859, Volume 35.
Wesche C and 6 others (2016) Neumayer III and Kohnen station in Antarctica operated by the Alfred Wegener Institute. *Journal of Large-Scale Research Facilities JLSRF* **2**, A85–A85. doi: 10.17815/jlsrf-2-152
Wesche C and Regnery J (2022) Expeditions to Antarctica: Ant-Land 2021/22 Neumayer Station III, Kohnen Station, Flight Operations and Field Campaigns. *Berichte zur Polar- und Meeresforschung = Reports on polar and marine research*. doi: 10.57738/bzpm_0767_2022
Winkelmann R, Levermann A, Martin MA and Frieler K (2012) Increased future ice discharge from Antarctica owing to higher snowfall. *Nature* **492**, 239–242. doi: 10.1038/nature11616
Wolovick MJ, Moore JC and Zhao L (2021) 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**, e2020JF005937. doi: 10.1029/2020JF005936
Wood SN (2010) Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* **466**, 1102–1104. doi: 10.1038/nature09319
Zeising O, Steinhage D, Nicholls KW, Corr HFJ, Stewart CL and Humbert A (2022) Basal melt of the southern Filchner Ice Shelf, Antarctica. *The Cryosphere* **16**, 1469–1482. doi: 10.5194/tc-16-1469-2022

Appendix A. Forward model details

The layer tracing scheme described in Section 2.1 is equivalent to solving a set of advection equations for a set of layers. Here, we explicitly write down the advection equations solved and the boundary conditions defined. We correct for 2-D effects in the advection equations by adding a term for the normal flow into the flow line. We then account for the inflow boundary condition at the grounding line $x = 0$.

A 1-D advection equation for a layer on a flow line reads

$$\frac{\partial h_l}{\partial t} = v_x \frac{\partial h_l}{\partial x}. \tag{A1}$$

In practice, real flow lines have some incoming or outgoing (normal) flux, q_y, where y denotes the horizontal direction perpendicular to x. We account for this normal ice flux and instead solve

$$\frac{\partial h_l}{\partial t} = v_x \frac{\partial h_l}{\partial x} + r_l(x) \frac{\partial q_y(x)}{\partial y}, \tag{A2}$$

where $r_l(x) = h_l(x)/h(x)$ is the ratio of thickness of layer l to the total thickness of the shelf. This equation holds due to the plug flow assumption, in which the flux divergence $\partial q_y/\partial y$ is independent of the depth z. The quantity $\partial q_y/\partial y$ is constant for all layers and independent of the layer thickness. This normal flux component accounts for lateral compression or extension of the flowtube centered on the flow line, and in our case we estimate it from satellite inferred velocities. For EIS, the normal flow component is small compared to the total mass balance along the flow lines, but for other cases this correction might be much more significant.

In order to define the inflow boundary condition, we would need to know the relative thickness of the incoming layers (or alternatively, the vertical age distribution at x_0). This is typically not available in radar measurements of the stratigraphy. It is, therefore, important to use only the IRH elevation information within the *LMI body* of the flow line. This is the region of our domain which is independent of the boundary condition we chose at x_0 (Fig. A1). This region can be found by tracking the trajectory traced by a particle initially at $(x = x_0, z = s(x_0))$. The LMI body is the region of the domain above this path. As a consequence of this consideration, we need to discard more of the IRH elevation data for the deeper IRHs in the dataset.

For a complete definition of the simulator, we still need to define some boundary condition at x_0. Since the true boundary condition is not known, we

https://doi.org/10.1017/jog.2025.13 Published online by Cambridge University Press

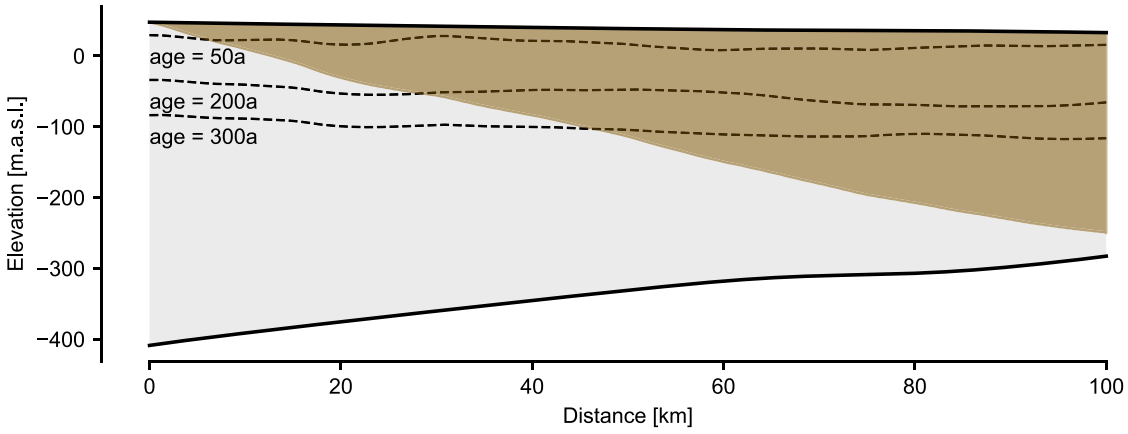

Figure A1. Local meteoric ice (LMI) body. The layer elevations in the LMI body (shaded region) are independent of the inflow boundary conditions. Outside the LMI body, the layer elevations are dependent on this boundary condition, and so using IRH observations within this region would require assuming the internal stratigraphy of the incoming ice.

choose an inflow boundary condition which improves the numerical stability of the layer tracing model,

$$\frac{\partial h_i}{\partial t} = \frac{\partial q_x}{\partial x}\bigg|_{x=0} + r_i(x)\frac{\partial q_y(x)}{\partial y}. \tag{A3}$$

This boundary condition makes the implicit assumption that near the inflow boundary, the layer thickness profile is a scalar multiple of the total thickness,

$$h(x)\frac{\partial h_i(x)}{\partial x} = h_i(x)\frac{\partial h(x)}{\partial x}. \tag{A4}$$

We define a similar boundary condition at $x = L_x$; however, this boundary condition does not have a similar effect on the applicability of the IRH data.

Appendix B. Calibrating the simulator

We calibrate hyperparameters of the noise model using a set of 1000 simulations for the same ice shelf geometry, velocity and mass-balance parameters prior. This set of calibration simulations is not used again to train NPE (Section 2.2.1) to avoid overfitting. We use the same set of calibration simulations to calibrate the per-layer LMI boundary mask (Section 2.2.2). Algorithm 1 defines our calibration procedure for the parameters of the noise distribution for EIS.

In Section 2.2.3, we used the LMI boundary $i(m)$ for each IRH m in order to define the domain $\mathbf{x}_{i\geqslant i(m)}$ on which we compare the observed IRH data to the simulated isochronal layers. The physical interpretation of the LMI boundary is found in Appendix A, and here we specify the exact definition of $i(m)$.

For each simulation k in the calibration set, we define the trajectory of a particle starting at the surface of the inflow boundary ($x = 0, z = s(0)$) by $p^k(t) = (x^k(t), z^k(t))$. In practice, since we are on a flow line, $v_x > 0$ everywhere, and thus x^k increases monotonically with t. Therefore, the trajectory traces out a unique curve $z^k(x)$ in the domain.

Using this path, we define the LMI boundary $i(m, k)$ for simulation k and IRH m to be the first point in the x domain such that the path is below the IRH elevation,

$$i(m,k) = \min_i\{i : z^k(x_i) < e_m(x_i)\}. \tag{B1}$$

In the case that the path stays above the IRH for the entire domain, we define $i(m, k) = L_x$, meaning that the IRH is entirely outside the LMI body.

This definition gives a different LMI body for each simulation. In order to perform inference, we require one a fixed LMI body boundary to use across all simulations. We define $i(m)$ 'pessimistically' as the 75th percentile of the calculated $i(m, k)$ boundaries in the calibration set. Therefore, some of the simulated layer elevations we use will be dependent on the unknown boundary conditions but a small amount that should not affect the inferred results.

Appendix C. Additional information on simulation-based inference

C.1. Tractable and intractable likelihood functions

Here, we provide intuition about the difference of forward models which have intractable and tractable likelihood functions. Forward models which have *tractable* likelihood functions are ones where the density $p(X|\theta)$ can be explicitly evaluated for a given (X, θ) pair, where X refers to the observed data and θ are the parameters of the model. As an example, a common setting in glaciology is a deterministic forward model with additive observational noise. Such forward models are of the form

$$X = f(\theta) + \epsilon, \tag{C1}$$

where $f(\theta)$ is a deterministic function, and ϵ follows some distribution, $\epsilon \sim p_{\epsilon(\epsilon)}$. In such models, the likelihood $p(X|\theta)$ can be evaluated as

$$p(X|\theta) = p_\epsilon(X - f(\theta)). \tag{C2}$$

A more general setting considers nondeterministic forward models, where latent variables z are sampled, and then affect further computations in the forward model. As a motivational example, consider a stochastic differential equation solved with an Euler–Maruyama scheme, that is, an iterative update

$$X_{t+1} = X_t + f(X_t, \theta, t)dt + g(X_t, \theta, t)z_t\sqrt{dt}, \tag{C3}$$

where f is now a deterministic function dependent on the state X_t and parameters θ and t, g is a deterministic function, and $z_t \sim \mathcal{N}(0, I)$ is random noise. Suppose the outcome from this model is the state at the end of the simulation, $X = X_T$. In order to compute the likelihood of this model, we need to integrate over all the noise samples $z = [z_1, \ldots, z_{T-1}]$,

$$p(X|\theta) = \int p(X, z|\theta)dz. \tag{C4}$$

In general, approximating this integral is significantly more expensive than generating a sample from $p(X|\theta)$ by simulating, and so we call this likelihood intractable

For our forward model, the intractability of the likelihood follows from the fact that we select the closest layer to the IRH after sampling the noise. Therefore, the noise model cannot be efficiently calculated as in Eqn (C2). Note that, even in the case of tractable likelihood functions, it might be beneficial to use 'likelihood-free' inference methods, for example in the case that the forward model is computationally expensive to compute.

C.2. Normalizing flows

While NPE can be performed with any valid parameterization of the variational distribution $q_\phi(\theta|X)$, in this work, we make the common choice to use a normalizing flow model. In the following, we adapt the notation of Papamakarios and others (2019a).

https://doi.org/10.1017/jog.2025.13 Published online by Cambridge University Press

Algorithm 1: Noise model calibration

Inputs: (Noiseless) simulated layer elevations $\mathbf{e}_l^{(k)}(\mathbf{x})$ for $l = 1, ...L$, $k = 1..., K_{\text{cal}}$ and simulation grid $\mathbf{x}$, IRH interpolated on the same simulation grid, $\mathbf{e}_m(\mathbf{x})$
Outputs: Mean and standard deviation $\mu_{\beta_n}, \sigma_{\beta_n}$ defining noise model
for $k \leftarrow 1$ *to* K_{cal} **do**
  $l^* \leftarrow \arg\min_l ||\mathbf{e}_l^{(k)}(\mathbf{x}) - \mathbf{e}_m(\mathbf{x})||_2^2$; `# best fitting layer to IRH`
  $\tilde{\mathbf{e}}_{l*}^{(k)}(\mathbf{x}) \leftarrow \mathbf{e}_{l*}^{(k)}(\mathbf{x}) - \texttt{lowpass\_filter}(\mathbf{e}_{l*}^{(k)}(\mathbf{x}))$; `# detrend best layer`
  $\tilde{\mathbf{e}}_m(\mathbf{x}) \leftarrow \mathbf{e}_m((\mathbf{x}) - \texttt{lowpass\_filter}(\mathbf{e}_m(\mathbf{x}))$; `# detrend IRH`
  $\gamma_1^{(k)}, ..., \gamma_n^{(k)} \leftarrow \text{PSD}\left([\mathbf{e}_m(\mathbf{x}) - \tilde{\mathbf{e}}_m(\mathbf{x})] - \left[\mathbf{e}_{l*}^{(k)}(\mathbf{x}) - \tilde{\mathbf{e}}_{l*}^{(k)}(\mathbf{x})\right]\right)$; `# power spectral density of difference`
  $\beta_1^{(k)}, ..., \beta_n^{(k)} \leftarrow \log(\gamma_1^{(k)}), ..., \log(\gamma_n^{(k)})$
for $n \leftarrow 1$ *to* N **do**
  $\mu_{\beta_n} \leftarrow \frac{1}{K_{\text{cal}}} \sum_{k=1}^{K_{\text{cal}}} \beta_n^{(k)}$; `# empirical mean of coefficients`
  $\sigma_{\beta_n} \leftarrow \sqrt{\frac{1}{K_{\text{cal}}-1} \sum_{k=1}^{K_{\text{cal}}} (\beta_n^{(k)} - \mu_{\beta_n})^2}$; `# empirical standard deviation of coefficients`
return $\mu_{\beta_1}, \sigma_{\beta_1}, ..., \mu_{\beta_N}, \sigma_{\beta_N}$

The goal of normalizing flows is to learn a map from a distribution $u \sim p_u(u)$ to a distribution of the same dimensionality $x \sim p_x(x)$. Here, $p_x(x)$ is the target distribution, and $p_u(u)$ is a simple distribution that can easily be evaluated and sampled from (e.g. a multivariate Gaussian). If we can express x as a differentiable, invertible transformation of u, i.e. $x = T(u)$, then we can also evaluate the probability

$$p_x(x) = p_u(u)|\det J_T(u)|^{-1}, \tag{C5}$$

where $u = T^{-1}(x)$ and J_T is the Jacobian matrix of the transformation T.

Normalizing flows are then defined by a sequence of transformations $T_1, \dots, T_K$ which transform $z_0 = u$ to $z_K = T_K \circ \cdots \circ T_1(u) = x$. Each of the transformations is learnable. The composition of differentiable, invertible transformations is also invertible and differentiable, with the determinant of the Jacobian satisfying

$$\det J_{T_2 \circ T_1}(u) = \det J_{T_2}(T_1(u)) \cdot \det J_{T_1}(u). \tag{C6}$$

One common way to parameterize the individual transformations T_i is through *coupling transforms*, where the input z is split into two parts, $z = [z_{1:d}, z_{d+1:D}]$, where D is the dimensionality of $z \in \mathbb{R}^D$. The first part, $z_{1:d}$, remains unchanged. The second part is transformed elementwise, $z_i' = \tau(z_i; h_i)$, where τ is some monotonic function of z_i conditioned on $h_i = F(z_{1:d})$, for some learnable function F of the first d components of z. The vector is permuted between each layer, so that not the same components are mapped through the identity with each transformation. An advantage of this parameterization is that the Jacobian matrix can be calculated as the product of its diagonal components, making the evaluation of the log-probabilities significantly faster. In this work, we use neural spline flows (NSFs, Durkan and others 2019). In NSFs, τ are monotonic spline functions, which are analytically invertible, yet highly flexible.

C.3. Connection between NPE loss and Kullback-Leibler divergence

We note that the expected forward KL divergence between the true posterior $p(\theta|X)$ and the variational distribution $q_\phi(\theta|X)$ can be decomposed as

$$\begin{aligned}\mathbb{E}_{p(x)}[D_{\text{KL}}(p(\theta|X)||q_\phi(\theta|X)))] &= \mathbb{E}_{p(X)p(\theta|X)}\left[\log \frac{p(\theta|X)}{q_\phi(\theta|X)}\right] \\ &= -\mathbb{E}_{p(X)p(\theta|X)}[\log q_\phi(\theta|X)] + \mathbb{E}_{p(X)p(\theta|X)}[\log p(\theta|X)],\end{aligned} \tag{C7}$$

where the first term on the right hand side is the negative of the loss in Eqn (8), and the second term is independent of the variational parameters ϕ and is thus a constant with respect to the variational parameters ϕ which we optimize. Therefore, minimizing the NPE loss (Eqn (8)) is equivalent to minimizing the expected forward KL divergence between the variational and true posterior distributions.

C.4. Neural Likelihood Estimation

In our work, we use NPE to directly approximate the posterior distribution. However, other variants of SBI exist. In particular, a prominent method of SBI is NLE (Wood, 2010; Papamakarios and others, 2019b). In NLE, the goal is to approximate the likelihood function $p(X|\theta)$ from a dataset of simulations $\{\theta_k, X_k\}_{k=1}^K$. This likelihood can then be used to perform Bayesian inference using the existing set of Bayesian inference methods, such as Markov chain Monte Carlo (MCMC), or variational inference (Blei and others, 2017).

Similarly to NPE, we define a variational distribution (e.g. a normalizing flow), $q_\phi(X|\theta)$, which can be trained by minimizing the negative log-probabilities of the model outcomes X_k given model parameters θ_k

$$\mathcal{L}(\phi) = \mathbb{E}_{\theta_k \sim p(\theta), X_k \sim p(X|\theta_k)}[-\log q_\phi(X_k|\theta_k)]. \tag{C8}$$

Notice that the conditioning of θ_k, X_k is reversed relative to Eqn (8).

The likelihood model learned in NLE can also be thought of as an emulator of the forward model, as it allows us to draw samples $X \sim p(X|\theta)$. This makes NLE methods advantageous also for problems where the likelihood is tractable, as the learned emulators can typically be evaluated faster than the original forward model. Related ideas have been explored for glaciological problems. For example, Tarasov and others (2012) use Bayesian neural networks as a surrogate model for particular parameter-output pairs in a glacial systems model, which they then use to perform inference with MCMC. Similarly, Brinkerhoff and others (2021) use a deterministic residual neural network as a surrogate model to predict a low-dimensional representation of outcomes from a subglacial hydrology model. They then derive an approximation of the uncertainty in this model to arrive at a nondeterministic likelihood which can then be used for inference.

Appendix D. Further implementation details

For completeness, we give the values of the important hyperparameters involved in the workflow. We provide the details of the spatial and temporal resolutions of our various simulations, along with the regularization strengths of the smoothing of the EIS data (Tables D1, D2 and D3).

During inference, we applied NPE as implemented in the `sbi` package (Tejero-Cantero and others, 2020) to obtain the results for both the synthetic (Section 3) and Ekström (Section 4) ice shelves. We used the NSF (Durkan and others, 2019) as implemented in Lueckmann and others (2021), and with the same architecture of five flow transformations, two residual blocks of 50 hidden units each, ReLU nonlinearity and 10 bins. We also embedded the 500-dimensional observation of layer elevations to a 50-dimensional summary statistic used as the condition for the NSF. The embedding network consisted of two convolutional layers with kernel size 5, each followed by ReLU activations and max pooling with kernel size 2. The number of output channels for the two convolutional layers were 6 and 12, respectively. The output channels of the second convolutional layer were then concatenated and fed through two

https://doi.org/10.1017/jog.2025.13 Published online by Cambridge University Press

Table D1. Hyperparameters for synthetic ice shelf spin-up modeling

Parameter	Value
Mesh resolution (x, y)	(310 m, 250 m)
Spin-up duration	1000 years
Spin-up time step	1.0 years
Boundary conditions	Dirichlet inflow and side boundaries

Table D2. Hyperparameters for the preprocessing of the data for Ekström Ice Shelf

Parameter	Value
Mesh resolution	300 m
Thickness smoothness reg. penalty	1000
Log fluidity reg. penalty	1000
Boundary conditions	Dirichlet inflow and side boundaries

Table D3. Layer tracer forward model simulation configuration

Parameter	Value
Simulation time	1000 years
Time step	0.5 years

fully connected linear layers, each followed by ReLU activations. The number of hidden units was set to 50. Training was done as in Lueckmann and others (2021), with the exception that the batch size was set to 1000 (default is 50). For each NPE run, we train five networks initialized with different random seeds and report in our results the run with the best validation loss.

Following the work of Lueckmann and others (2021), we split the simulation dataset into training and validation datasets with a 90–10 split. We optimize the loss in Eqn (8) using an Adam Optimizer (Kingma and Ba, 2015) with a batch size of 50, a learning rate of 0.0005, with the maximum gradient norm clipped to 5.0. For each training epoch, we calculate the validation loss on the entire validation dataset. If the validation loss has not surpassed its best value for 20 training epochs, we assume convergence and stop training.

Appendix E. Computational costs

We provide a breakdown of the approximate computational costs of the different stages in our workflow for both synthetic and Ekström Ice Shelves in Tables E1 and E2, respectively. These are dependent the hardware used and vary stochastically as a result of random number generators. This section is intended to provide intuition into the relative scales of the different stages of the workflow, rather than exact measurements. We had access to 16-core Intel Xeon Gold 2.9 GHz CPU nodes and Nvidia RTX 2080ti GPU nodes. While the large number evaluations of the forward layer tracing model were by far the most computationally intensive section of the workflow in both cases, these simulations were trivially performed in parallel across CPU cores, thus reducing the wall-clock time of the workflow.

This analysis highlights the advantages of our amortized approached to inference. For EIS, the total computational time of the noiseless simulations accounted for $\approx 99.8\%$ of the total computational time of the workflow. The simulations need not be repeated when we infer from other IRHs, greatly benefiting the computational efficiency of inference as the number of IRHs in the dataset increases. This advantage is slightly reduced when considering the parallelization we have used (Table E2), as the training of the probability density

Table E1. Synthetic ice shelf approximate computational cost breakdown. Some tasks are embarrassingly parallelizable—parallel resources and times are shown in square brackets. All times reported in minutes

Task	Node [parallel]	Time [parallel]
Spin-up of ice shelf	CPU	120
1000 calibration simulations	CPU [100 cores]	1000 [10]
189 000 noiseless simulations	CPU [100 cores]	2×10^5 [2000]
Noise and layer selection	CPU [100 cores]	400 [20]
Training NPE for 1 IRH	GPU	30

Note: Only the last two tasks need to be repeated for each IRH measurement.

Table E2. Ekström Ice Shelf approximate computational cost breakdown. Some tasks are embarrassingly parallelizable—parallel resources and times are shown in square brackets. All times reported in minutes

Task	Node [parallel]	Time [parallel]
Generating mesh	CPU	< 1
Data preprocessing	CPU	10
1000 calibration simulations	CPU [100 cores]	1000 [10]
189 000 noiseless simulations	CPU [100 cores]	2×10^5 [2000]
Noise and layer selection	CPU [100 cores]	400 [20]
Training NPE for 1 IRH	GPU	30

Note: Only the last two tasks need to be repeated for each IRH measurement.

estimator is not easily parallelizable across GPU nodes. Accounting for this still results in $\approx 97.6\%$ of the computational cost being amortized.

Appendix F. Additional results

We show the posterior and posterior predictive distributions when inferring from isochronal layer 3 in the synthetic ice shelf dataset, of age 150 years (Fig. F1). This isochronal layer has average depth of 120 m, comparable to the deepest IRH in the Ekström dataset. While the uncertainty is much higher than for the shallow layer, the posterior over the surface accumulation still shows a higher mean than the prior in the downstream section of the ice shelf. Additionally, the posterior predictive reconstructs the ground truth isochronal layer better than the prior predictive. The mean RMSE of the posterior predictive layers relative to the ground truth is 13.6 m, compared to 19.8 m for the prior. Therefore, we are still able to reconstruct additional information about the mass balance parameters, even from much deeper layers.

We also explore the dependence of the inferred posterior over surface accumulation rate on the depth of the layer used for inference in the synthetic case (Fig. F2), similar to the analysis done in Section 5.2 for EIS. For the synthetic ice shelf, the surface accumulation and basal melt rates were held constant for the entire simulation time, and hence the increased uncertainty with depth seen in Fig. F2 highlights that information about the mass balance parameters is gradually lost with time as a result of the action of the simulator. Indeed, for the deepest ground truth isochronal layer of average depth 183 m, the posterior distribution is almost identical to the prior distribution.

Finally, we report the RMSE in the predicted isochronal layer elevations, relative to the true IRH (for EIS) or the ground truth isochronal layer (for the synthetic ice shelf). This is done across simulations from 1000 simulations for each of the prior and posterior distributions, for each IRH. The RMSE is consistently lower for the posterior predictive distribution for all depths, for both the synthetic ice shelf (Table F1) and EIS (Table F2).

Appendix G. Kottas traverse data

Here we describe the mapping of the surface accumulation measurements on Kottas traverse to the flow line transect. For each measurement year, and each location $\tilde{x}_i$ on the Kottas traverse, we find the nearest point x_i on the flow line transect. We assume the accumulation rate at this point to be normally distributed, with mean $\tilde{a}_i$ (the Kottas traverse measurement at $\tilde{x}_i$), and variance $\sigma_a^2 ||\tilde{x}_i - x_i||_2^2 / l_a$. We set the length scale $l_a = 2.5$ km and the accumulation rate variance $\sigma_a^2 = 0.25^2\ \mathrm{m}^2\ \mathrm{m}^{-2}$. These values are chosen in accordance to the definition of the surface accumulation rate prior distribution (Section 2.2.3).

The yearly variations of the estimated surface accumulation as measured using the stake line along the Kottas traverse (Fig. G1) reasonably agree with the posterior inferred using IRH 2. These show that there is a high year-to-year variability in the surface accumulation (Mengert 2018, Fig. 13) even in this steady-state region in East Antarctica. However, the mean of these yearly measurements matches the inferred posterior mean closely.

Appendix H. Synthetic results with miscalibrated prior

An additional outcome of our approach is the estimation of the age of isochronal layers. However, the validity of this estimate depends on the prior distribution containing the true mass-balance parameters. When the true mass-balance

https://doi.org/10.1017/jog.2025.13 Published online by Cambridge University Press

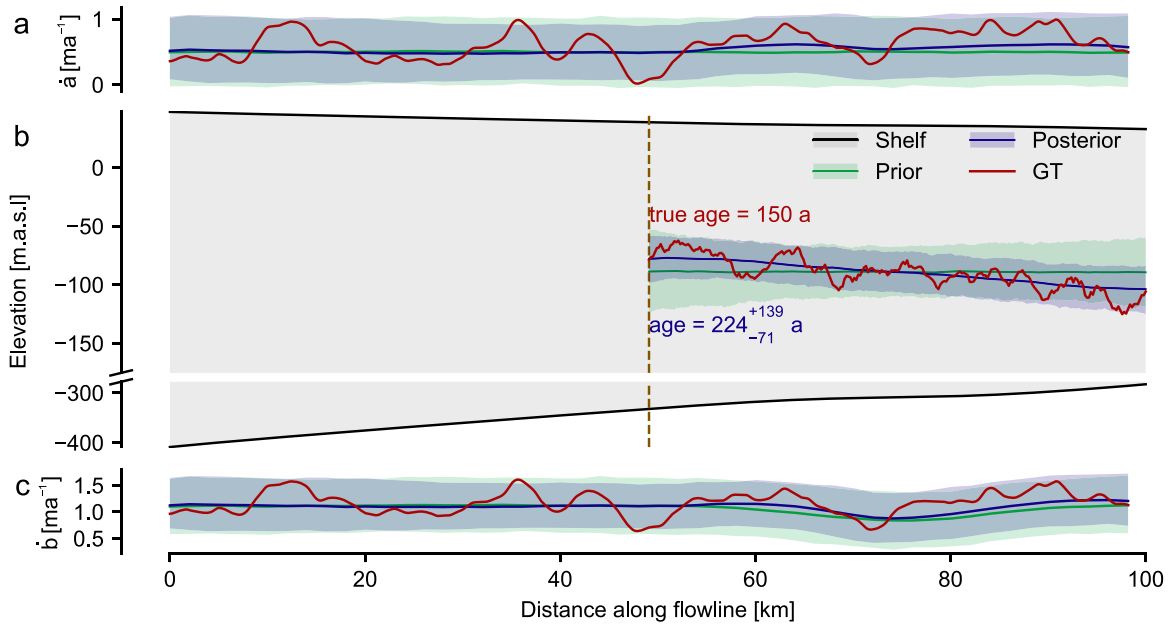

Figure F1. Prior and posterior (predictive) for the synthetic dataset. (a and c) Prior and posterior over surface accumulation and basal melt rates respectively for layer 3 of the synthetic ice shelf, of age 150 years. Solid line is the distribution mean, the shaded region represents the 5th and 95th percentiles. The ground truth (GT) parameters used to generate the reference isochronoal layer are also shown. (b) Cross section of the ice shelf. Prior and posterior predictive distributions for the layer closest matching the ground truth isochronal layer. The vertical dashed line represents the LMI boundary for this isochronal layer. The posterior predictive reconstructs the observed layer with higher accuracy and lower uncertainty. The posterior predictive distribution of the age of the isochronal layer is 224^{+139}_{-71} years.

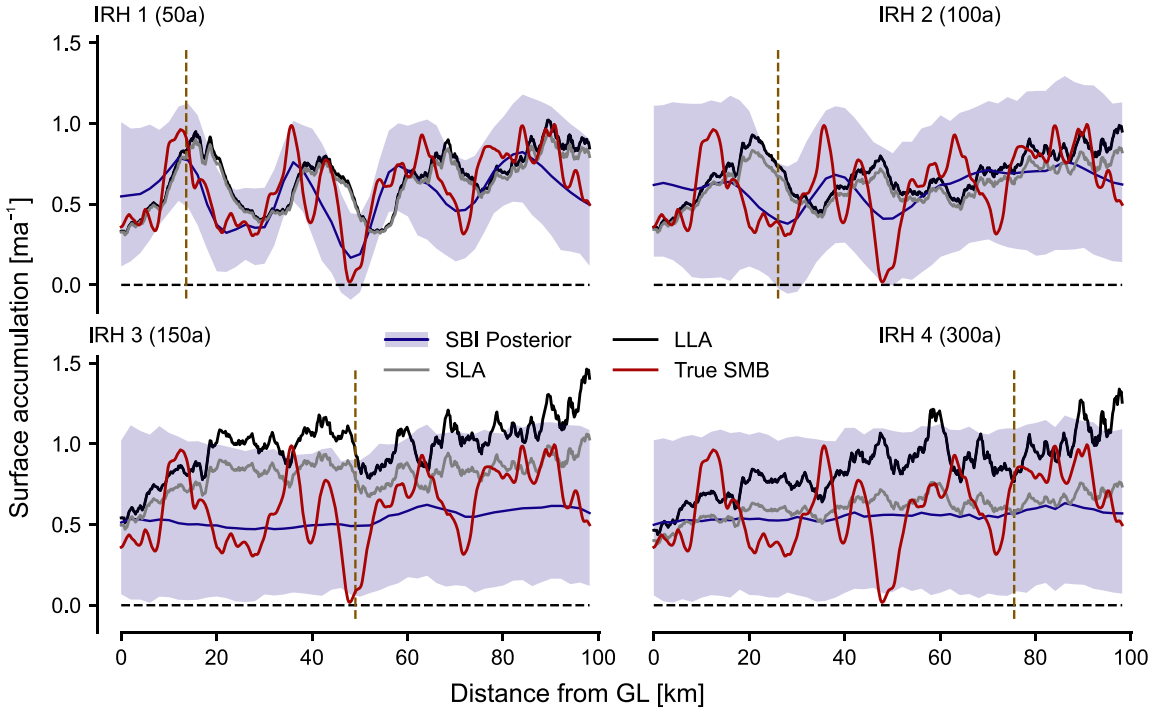

Figure F2. Synthetic shelf: dependence of posterior surface accumulation rate on depth of layer used for inference.

parameters have low probability under the prior distribution, the resulting estimate for the age of the isochronal layer can be wrong. We test this on the synthetic ice shelf by choosing a ground truth surface accumulation $\dot{\mathbf{a}}_{\text{mis}}$ that has low probability under the prior distribution. We calculate the isochronal layer of age 100 years under this ground truth and obtain the posterior distribution as before, using the same set of simulations as in the main text. The posterior does not capture that the mean surface accumulation rate should be higher than what is defined in the prior (Fig. H1a). However, this is not a failure of the inference method, as we can see that the posterior predictive still reconstructs the ground truth isochronal layer at higher fidelity than the prior (Fig. H1b). The predicted age of the isochronal layer 164^{+101}_{-44} years, which greatly overestimates the true age of 100 years.

https://doi.org/10.1017/jog.2025.13 Published online by Cambridge University Press

Table F1. Synthetic ice shelf—prior and posterior predictive distribution root-mean-square error (RMSE) relative to ground truth IRH, estimated from 1000 samples. The mean and standard deviations (SDs) in the RMSE are reported. All values are in meters

	Prior		Posterior	
IRH number	RMSE mean	RMSE SD	RMSE mean	RMSE SD
1	11.5	3.9	3.9	0.5
2	16.0	7.6	7.3	1.2
3	19.8	8.4	13.6	3.5
4	22.1	7.9	19.8	6.8

Table F2. Ekström Ice Shelf—prior and posterior predictive distribution root-mean-square error (RMSE) relative to ground truth IRH, estimated from 1000 samples. The mean and standard deviations (SDs) in the RMSE are reported. All values are in meters

	Prior		Posterior	
IRH number	RMSE mean	RMSE SD	RMSE mean	RMSE SD
1	6.8	2.1	3.0	0.9
2	11.8	3.4	4.6	1.3
3	17.0	6.6	6.8	1.4
4	16.4	7.6	10.0	2.1

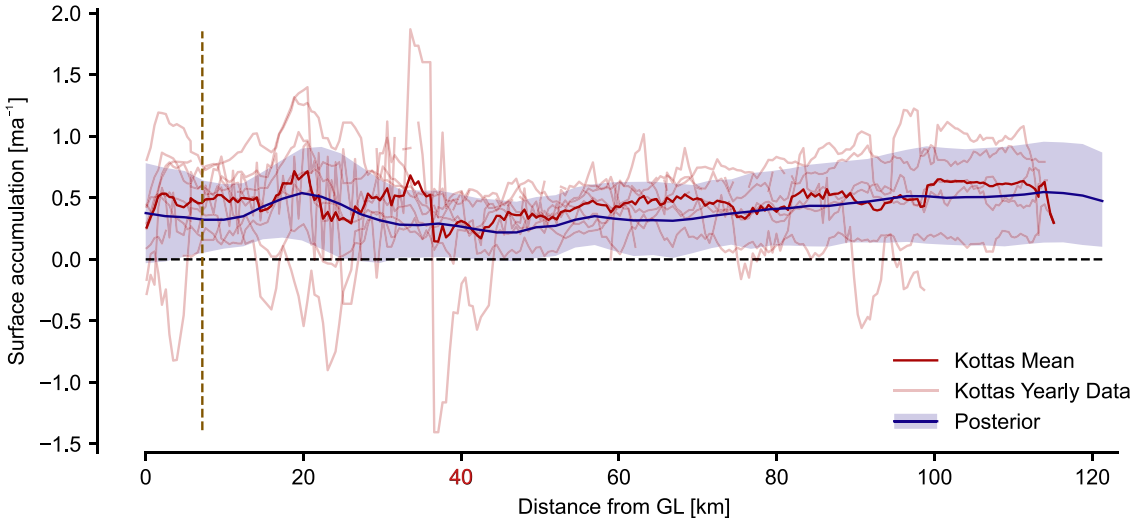

Figure G1. Yearly variations of the Kottas surface accumulation stakes measurement dataset. Years shown are 1995–2005 and 2017–19. These are compared with the posterior distribution inferred using IRH 1 of the Ekström IRH dataset.

https://doi.org/10.1017/jog.2025.13 Published online by Cambridge University Press

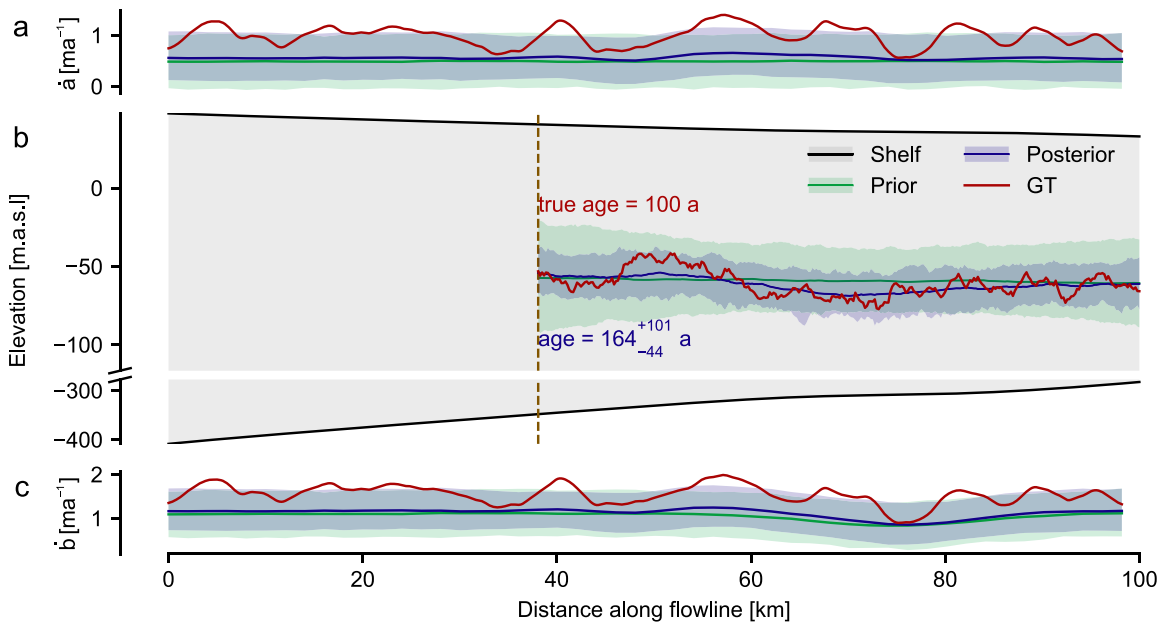

Figure H1. Prior and posterior (predictive) for the synthetic ice shelf with the low-probability ground truth. (a and c) Prior and posterior over surface accumulation and basal melt rates respectively for an isochronal layer of age 100 years. Solid line is the distribution mean, the shaded region represents the 5th and 95th percentiles. The ground truth (GT) parameters used to generate the reference isochronal layer are also shown. (b) Cross section of the ice shelf. Prior and posterior predictive distributions for the layer closest matching the ground truth isochronal layer. The vertical dashed line represents the LMI boundary for this isochronal layer. The posterior predictive reconstructs the observed layer with higher accuracy and lower uncertainty. The posterior predictive distribution of the age of the isochronal layer is 164^{+101}_{-44} years.

Appendix I. Radar data

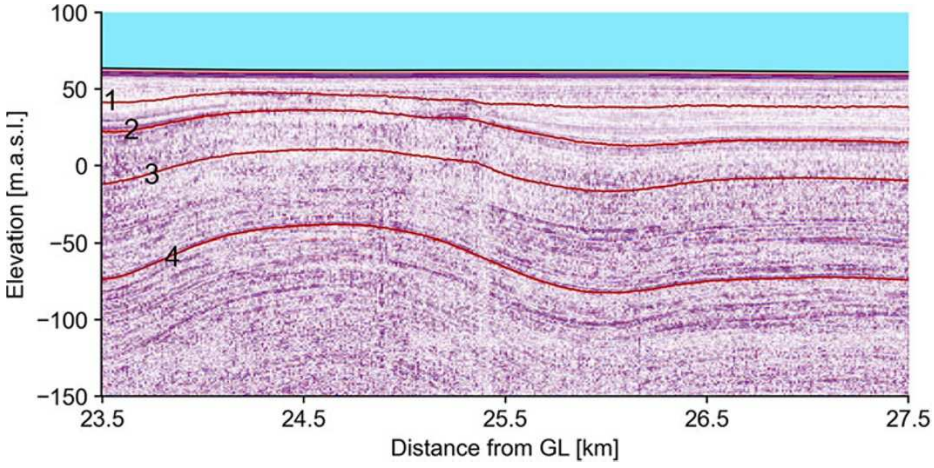

Figure I1. Radargram along transect. Zoom in on section of vertical cross-section view of the radar transect (Figure 5c). Color gradient indicates radargram data from radar survey of transect. The four labeled IRHs picked from this radargram are labeled in order of depth.

https://doi.org/10.1017/jog.2025.13 Published online by Cambridge University Press

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

Guy Moss[1,2]* **Leah Sophie Muhle**[3] **Reinhard Drews**[3]

Jakob H. Macke[1,2,4,†] **Cornelius Schröder**[1,2,†*]

[1]Machine Learning in Science, University of Tübingen, Tübingen, Germany
[2]Tübingen AI Center, Tübingen, Germany
[3]Department of Geosciences, University of Tübingen, Tübingen, Germany
[4]Department 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 temporarily 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

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

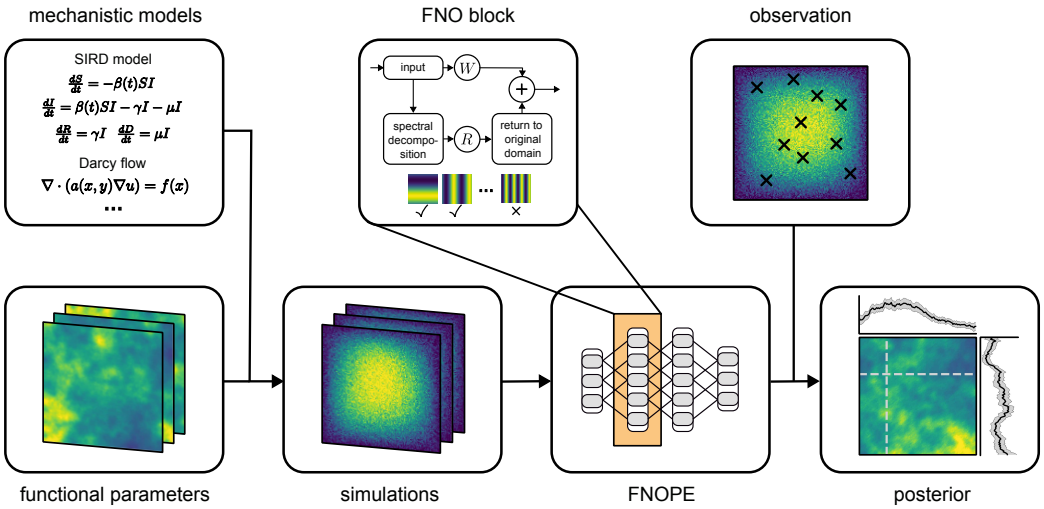

Figure 1: **Overview.** FNOPE approximates the posterior over function-valued parameters of a mechanistic model conditioned on function-valued observations. We use a FNO architecture with a flow matching objective to efficiently represent the function-valued parameters, enabling us to estimate extremely 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 **FNO**s for **P**osterior **E**stimation 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 $p(\theta \mid x)$ for

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

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

where $p_t(z_t \mid \theta)$ are the sample-conditional flow paths for z_t, and $u_t(z_t \mid \theta)$ 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 $\mathcal{G} : \mathcal{A} \to \mathcal{B}$ by a learnable operator $\tilde{\mathcal{G}}^\phi : \mathcal{A} \to \mathcal{B}$. 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, $\tilde{G}^\phi : \mathcal{A} \to \mathcal{B}$, is defined by

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

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 $p_{l^x}(x \mid \theta, l^\theta)$. Many such simulations create a dataset $S = \{(l^\theta_i, \theta_i, l^x_i, x_i)\}_{i=1}^K$ 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^x_o), and a desired parameter discretization l^θ, we want to sample $\theta \sim p_{l^\theta}(\theta \mid x_o, l^x_o)$. This is done by first sampling from a base distribution $\xi_1 \sim p_{l^\theta}(\xi)$ and

then learning the velocity field $v^{\phi}_{l^\theta}(t, \xi_t, x_o, l^x_o)$ for ξ_t. In the following, we omit the arguments of $v^{\phi}_{l^\theta}$ 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 $v^{\phi}_{l^\theta}$ is optimized via the loss function

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

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

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

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

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

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

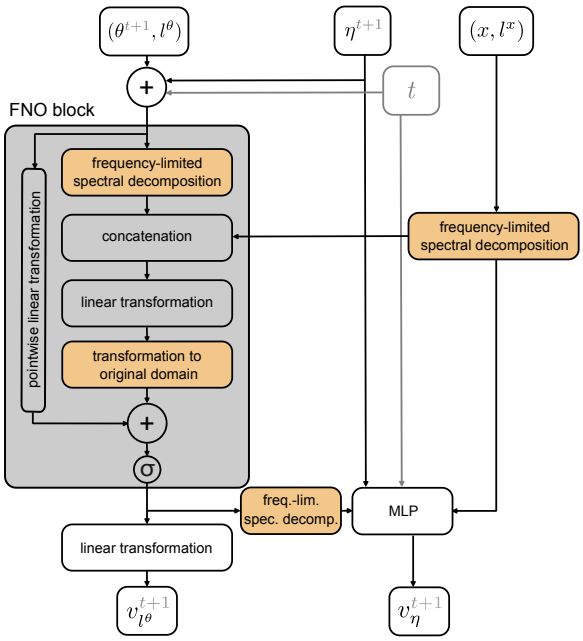

Figure 2: **FNOPE architecture.** FNOPE is based on several FNO blocks (gray): A FNO block receives the discretization-dependent spectral features of the function-valued parameters and observations as an input and processes them in a linear layer before transforming it back to the original domain via 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.

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 (NUDFT, [24]). The NUDFT 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^θ_i and l^x_i. Second, we add small, independent Gaussian noise to the remaining positions (Appendix S3.3).

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 $p_{l_i^x}(x_i \mid \theta_i, l_i^\theta, \eta)$. The resulting inference problem is to estimate the posterior distribution $p_{l^\theta}(\theta, \eta \mid x_o, l_o^x)$. 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 $p(z_t \mid \eta)$ of the vector-valued parameters is given by a normal distribution $\mathcal{N}(z_t; (1-t)\eta, t^2\mathbf{I})$, and $u_t = [u_t(\xi_t \mid \theta), u_t(z_t \mid \eta)]$. The vector field $u_t(z_t \mid \eta)$ for the vector-valued parameters is analogously defined as $u_t(z_t \mid \eta) = (z_t - \eta)$. 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.

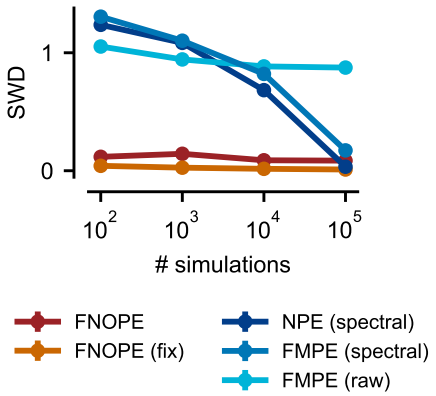

Figure 3: **Linear Gaussian simulator.** Sliced-Wasserstein distance (SWD) to ground truth posterior.

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).

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.

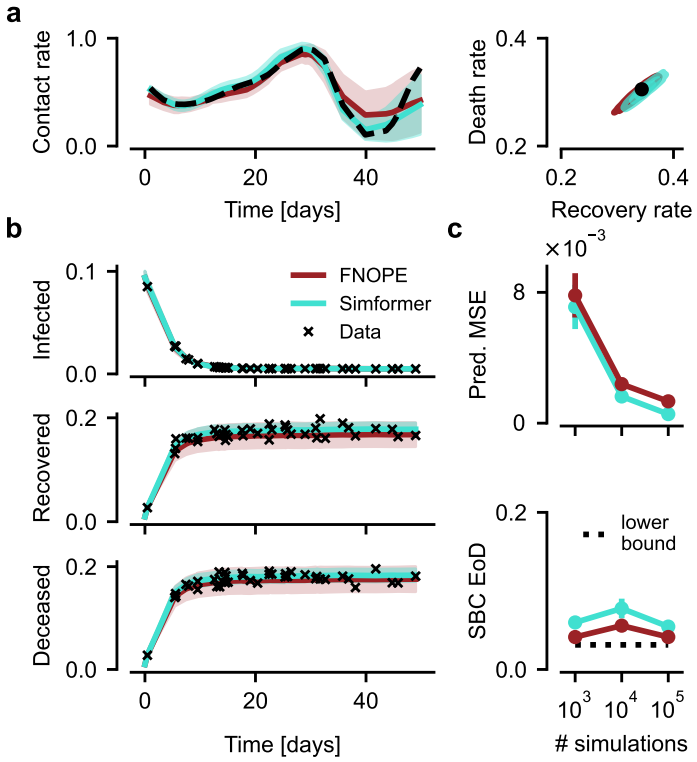

Figure 4: **SIRD model.** **(a)** Posterior conditioned on 40 time points. *left:* Posterior (mean ± std.) of the time-varying parameter and ground truth parameters (dashed). *right:* Two dimensional posterior of vector-valued parameters and ground truth parameters (dot). **(b)** Posterior predictive (mean ± 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

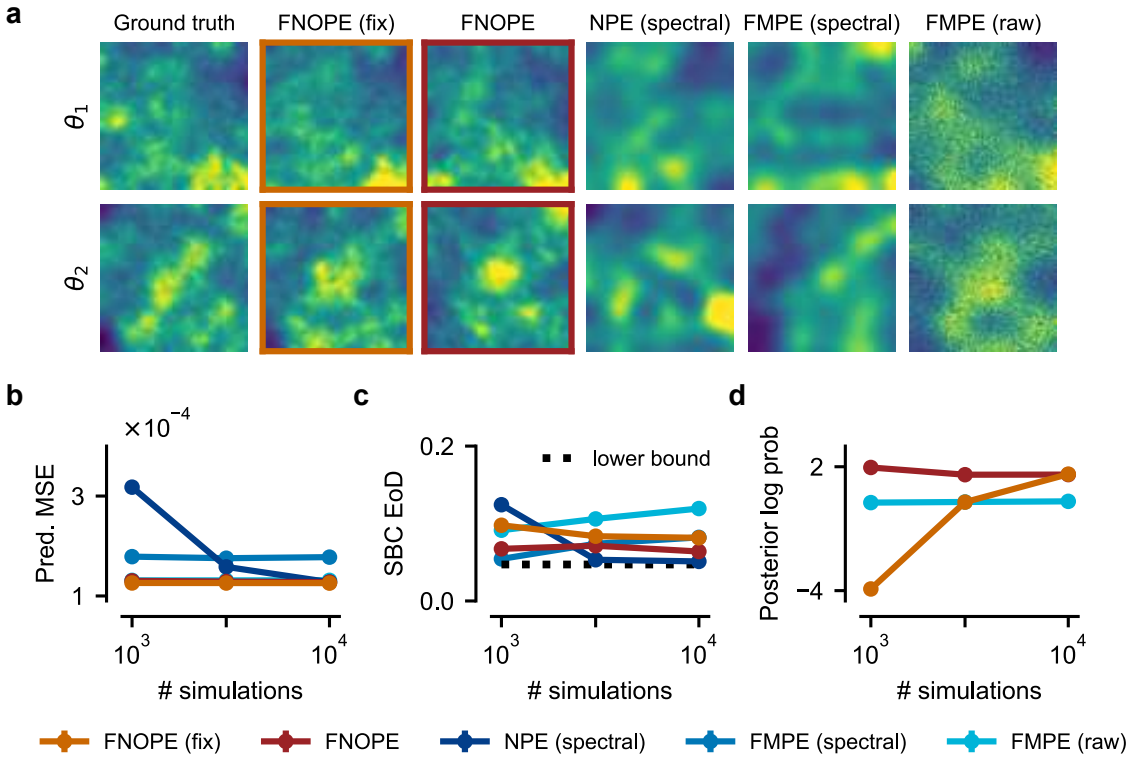

Figure 5: **Darcy flow. (a)** Ground truth parameter and posterior samples for a simulation budget of 10^4 training samples (more posterior samples in Fig. S6). **(b)** MSE of posterior predictives to the ground truth observation (FNOPE, FNOPE (fix) and FMPE (raw) visually overlap). **(c)** Simulation-based 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 accout 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

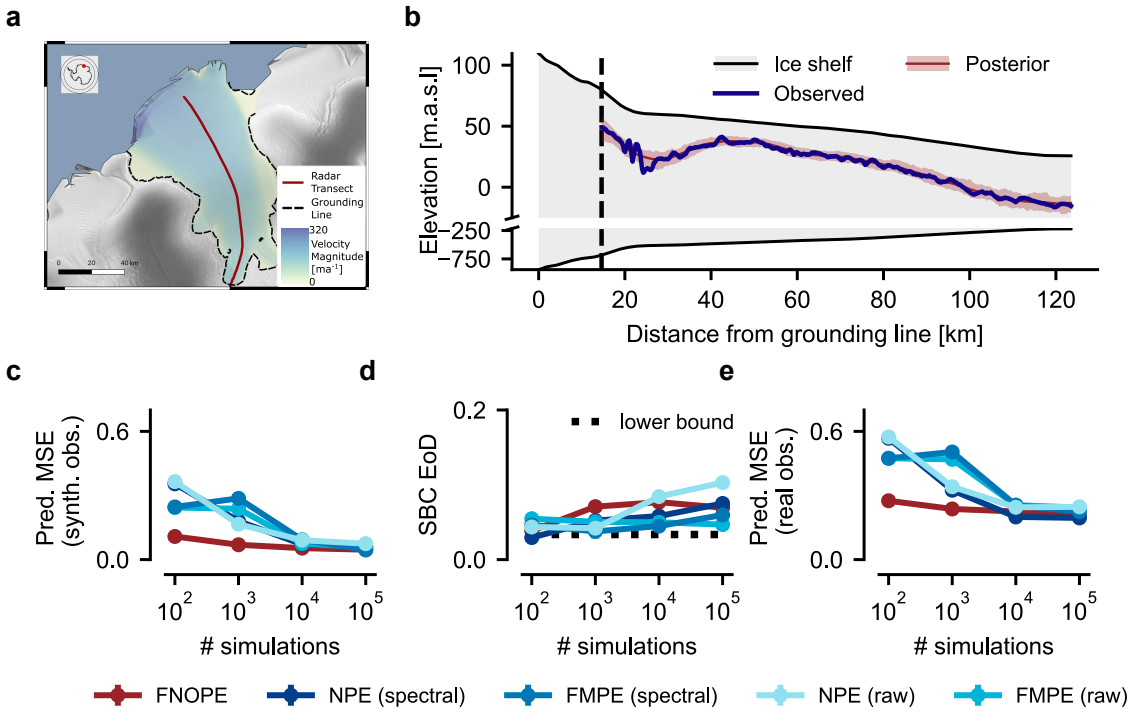

Figure 6: **Mass balance rates of ice shelves. (a)** Measurement transect of the radar data in Ekström Ice Shelf, Antarctica (adapted from [39], published under Creative Commons CC BY license). **(b)** Posterior predictive results obtained with FNOPE (trained on 10^5 simulations), compared to radar-based 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 deeponet 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 Huguet, 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 Ziaeemehr, 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, Nasrulloh 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

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

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 $x_{kj} \sim p_{l_j^x}(x \mid \theta_{kj}, l^\theta, \eta_{kj})$. 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:

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

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

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

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

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

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 näive extension of the FMPE objective as stated in Sec. 3.3 is to minimize the L^2 loss $||v^\phi - u_t||^2$, where $v^\phi = [v^\phi_{l^\theta}, v^\phi_\eta]$ and $u_t = [u_t(\xi_t|\theta), u_t(z_t|\eta)]$. However, the scale of the loss for the continuous parameters, $||v^\phi_{l^\theta} - u_t(\xi_t|\theta)||^2$ 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

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

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 respective, 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 $x \sim \mathcal{N}(\theta, \sigma^2 \boldsymbol{I})$, 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 lenghtscale 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-permebility 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 $\{h_1(l), h_2(l), \ldots, h_K(l)\}$ 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

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

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.

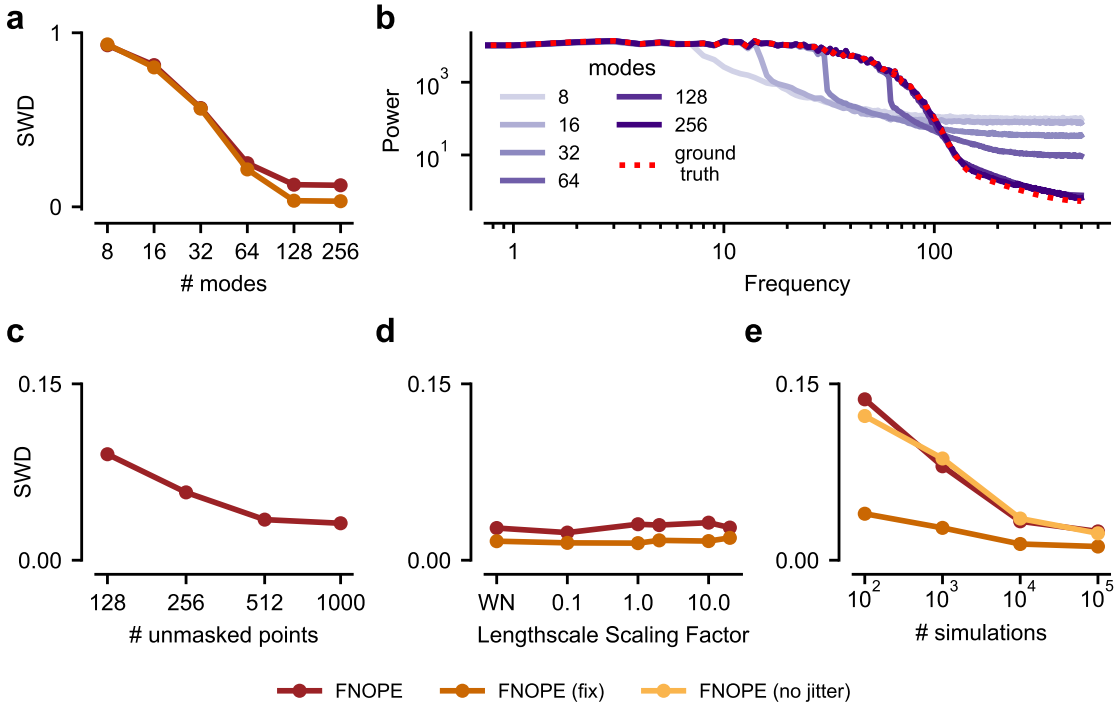

Figure S1: **Linear Gaussian ablation experiments.** **(a)** Performance for 10^4 training samples in terms of SWD for the Linear Gaussian experiments with varying number of Fourier modes in the FNO block. Note that we changed the lengthscale of the simulator prior by a factor of 10 to 0.005 (Appendix 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

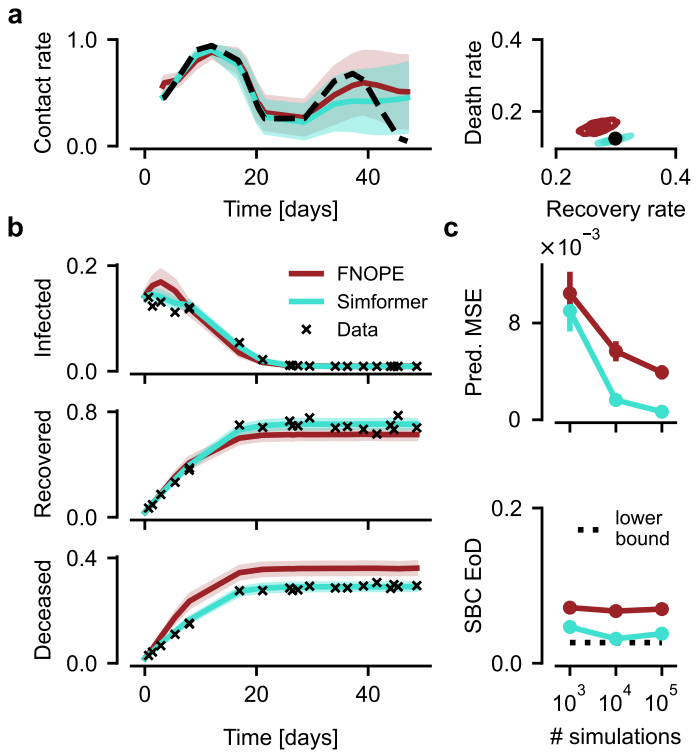

Figure S2: **SIRD model on 20 conditioning points.** **(a)** Posterior conditioned on 20 time points. *left:* Posterior (mean ± std.) of the time-varying parameter and ground truth parameters (dashed). *right:* Two dimensional posterior of vector-valued parameters and ground truth parameters (dot). **(b)** Posterior predictive (mean ± 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.

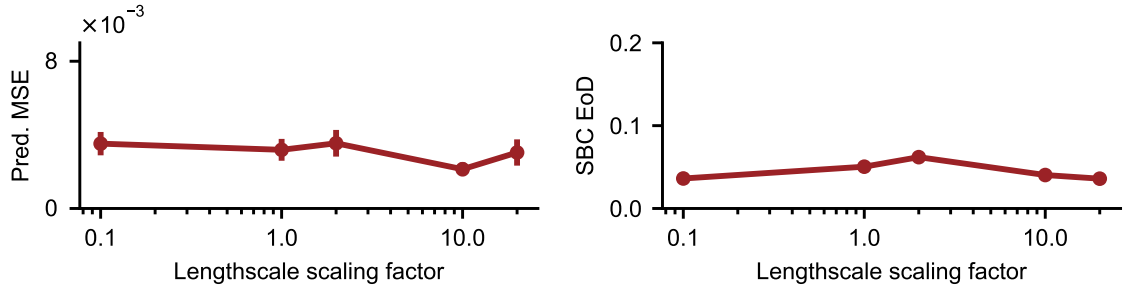

Figure S3: **SIRD ablation experiments.** Influence of the lengthscale of FNOPE's base distribution (Sec. 3.1) on posterior quality in terms of MSE of posterior predictive simulations and Simulation-based calibration error of diagonal (SBC EoD).

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

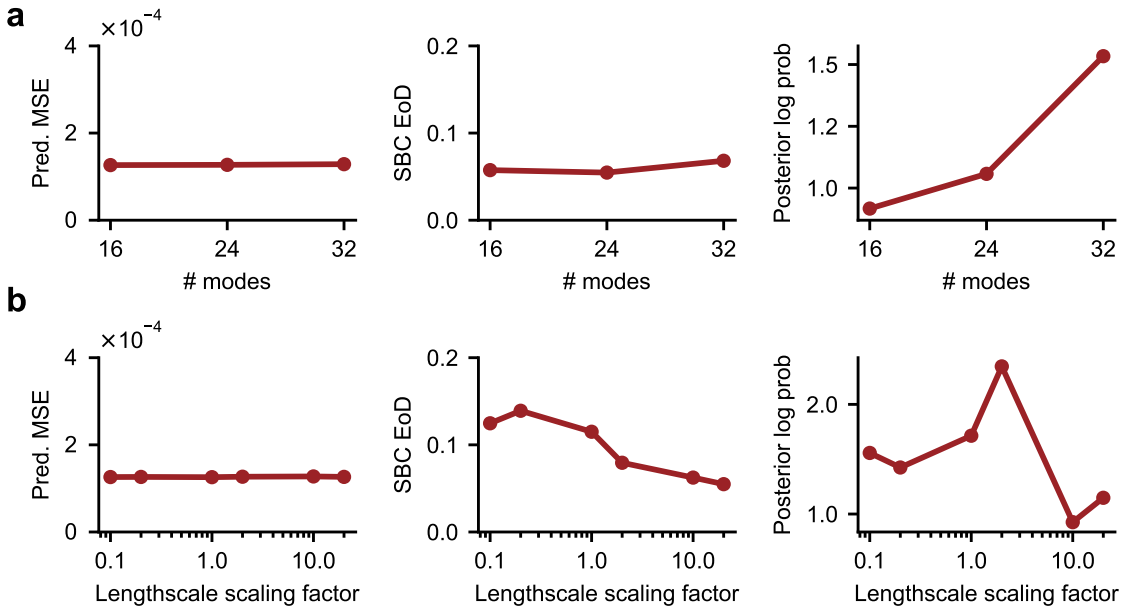

Figure S4: **Darcy ablation experiments (a)** Influence of number of used modes in the FNO block on different performance measures. The original experiments used 32 modes. Measures are the same as in Fig. 5b-d. **(b)** Influence of the noise length scale on different performance measures. The original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

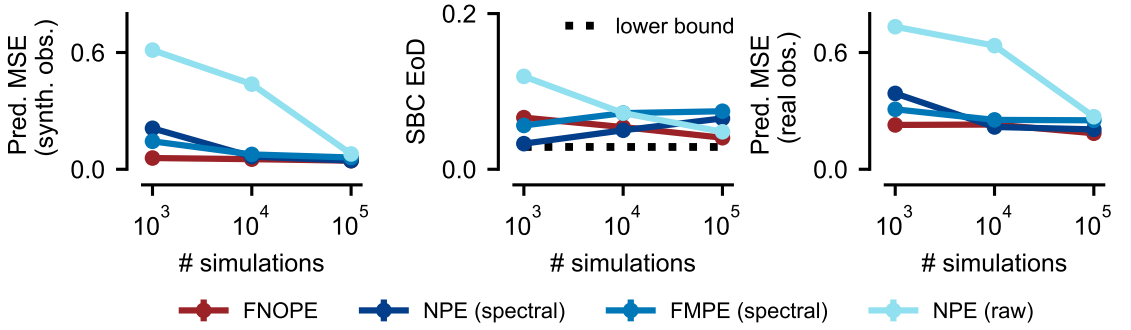

Figure S5: **Mass balance rates of ice shelves.** Inference of mass balance rates where the parameter discretization uses 500 gridpoints (observation discretization remains unchanged). Performance measures on test simulations and the real observation, where NPE (raw) refers to the method used in Moss 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

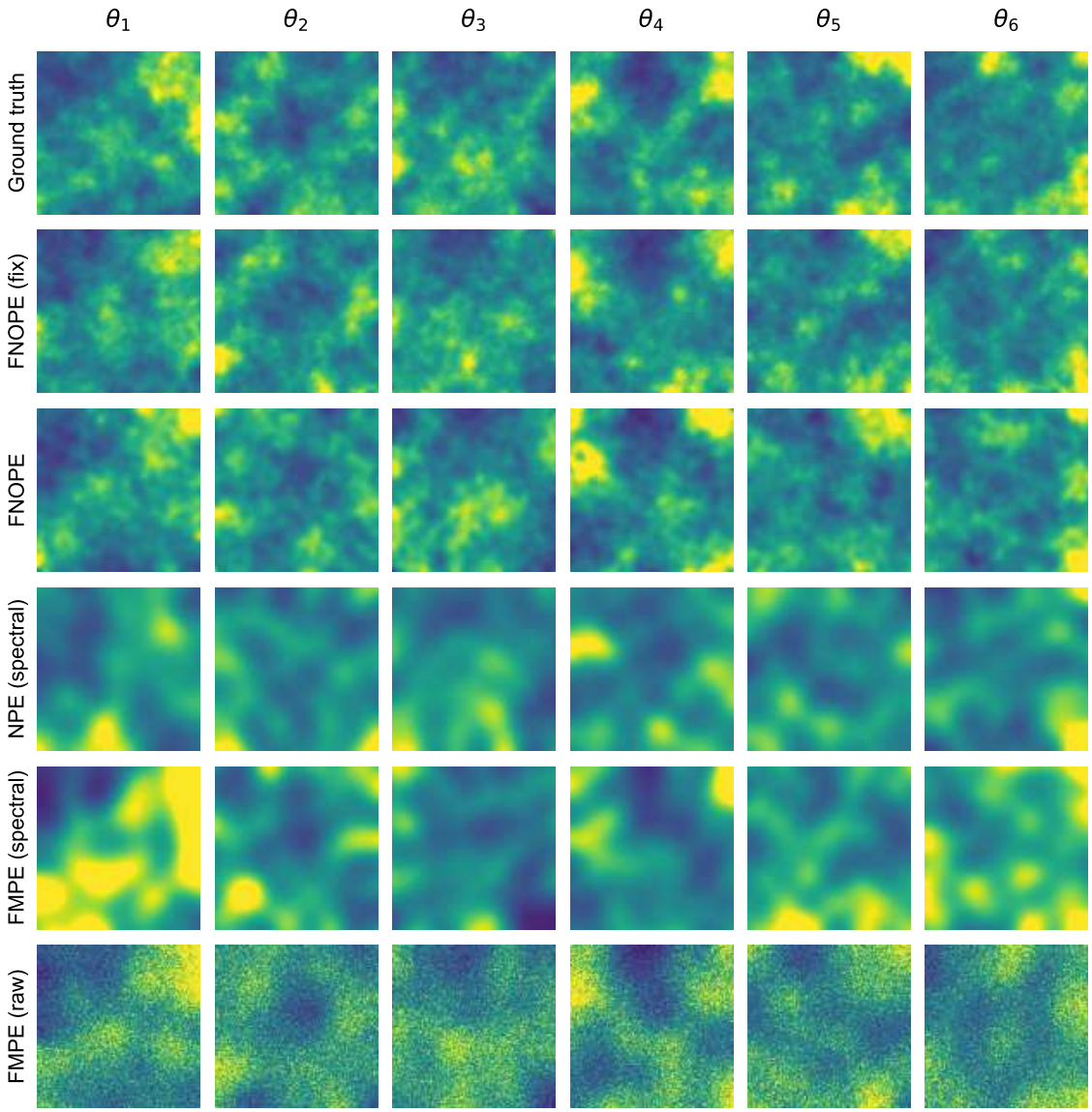

Figure S6: **Posterior samples Darcy flow experiment.** Additional posterior samples for the Darcy flow experiment for six distinct observations simulated from ground truth parameter θ_i.

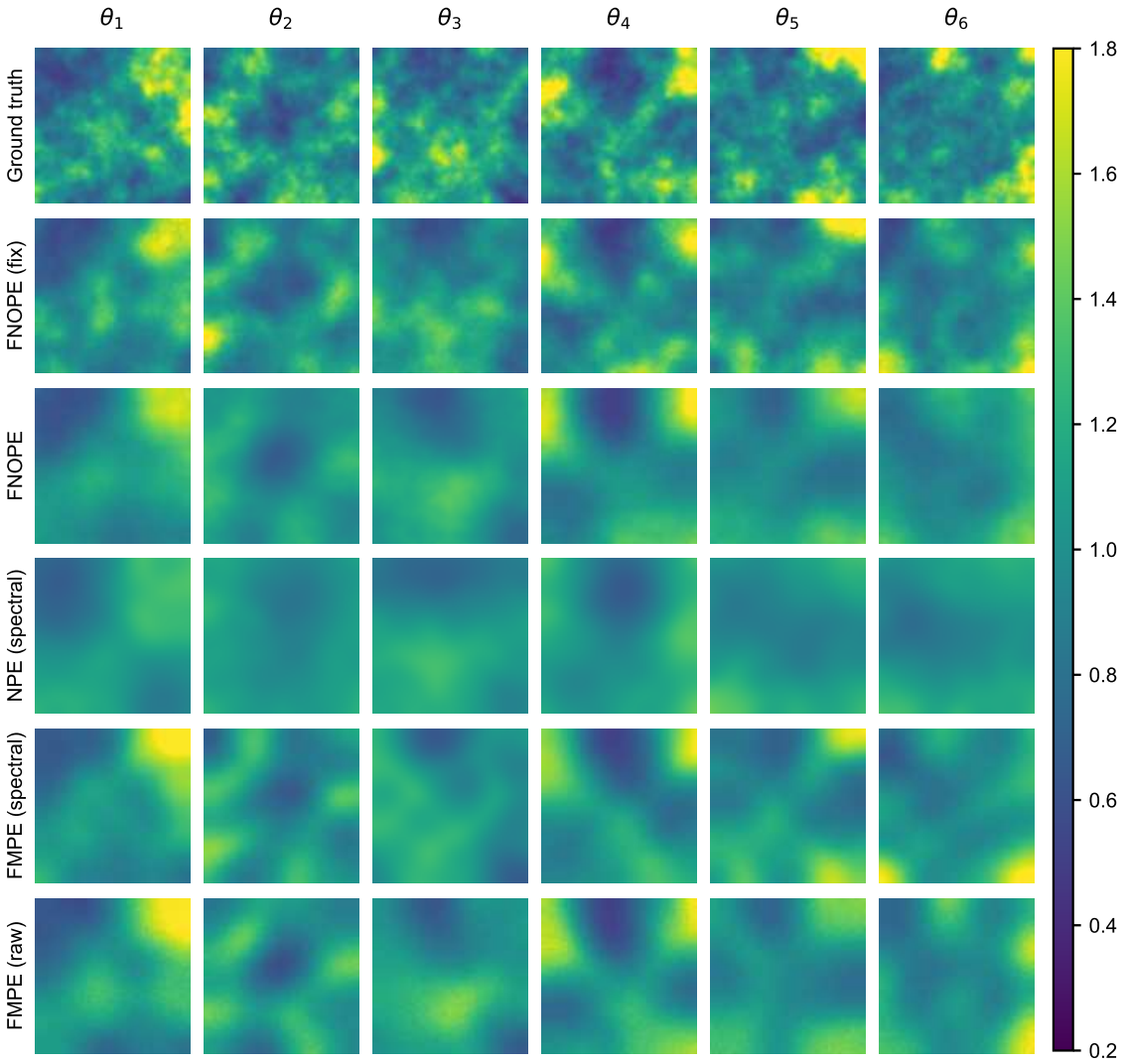

Figure S7: **Posterior means for Darcy flow experiment.** Posterior means (based on 100 samples) for the Darcy flow experiment for six distinct observations simulated from ground truth parameter θ_i.

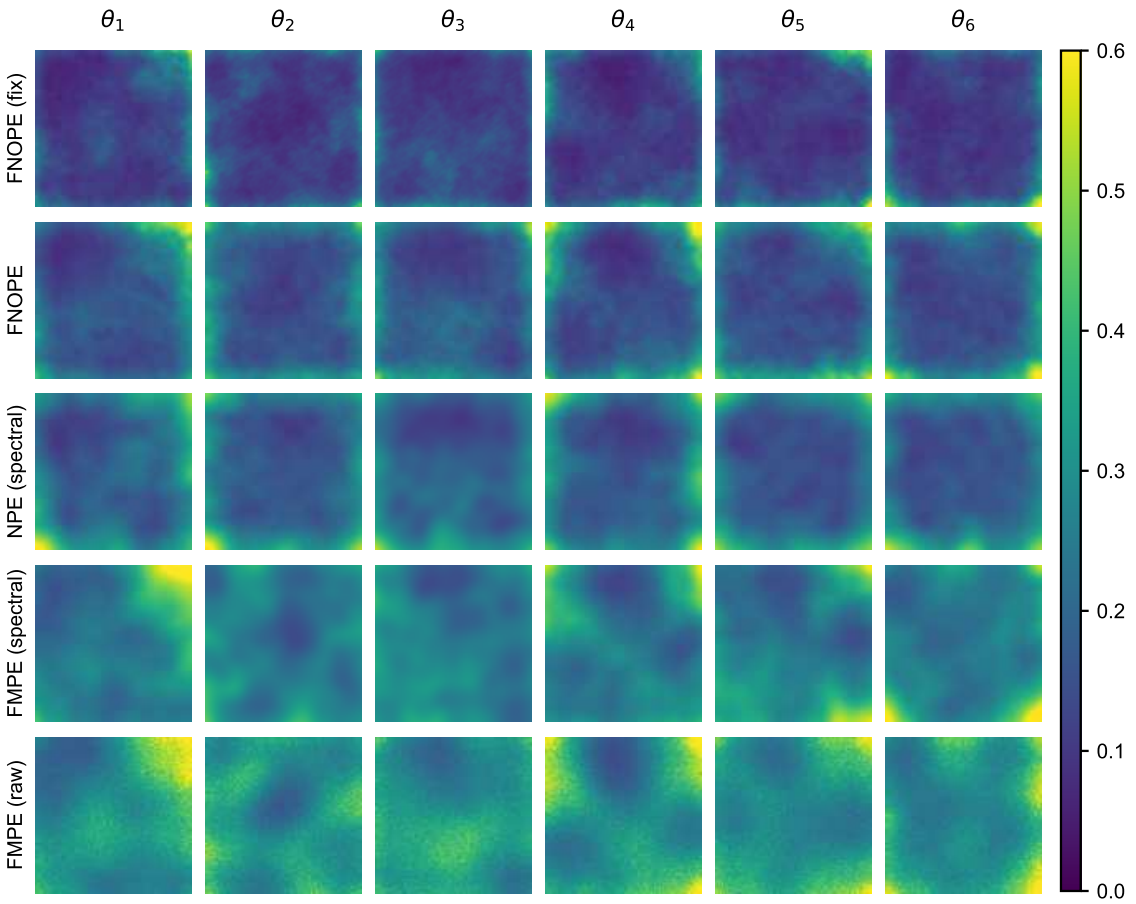

Figure S8: **Posterior standard deviation for Darcy flow experiment.** Posterior standard deviations (based on 100 samples) for the Darcy flow experiment for six distinct observations simulated from ground truth parameter θ_i.

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

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

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.

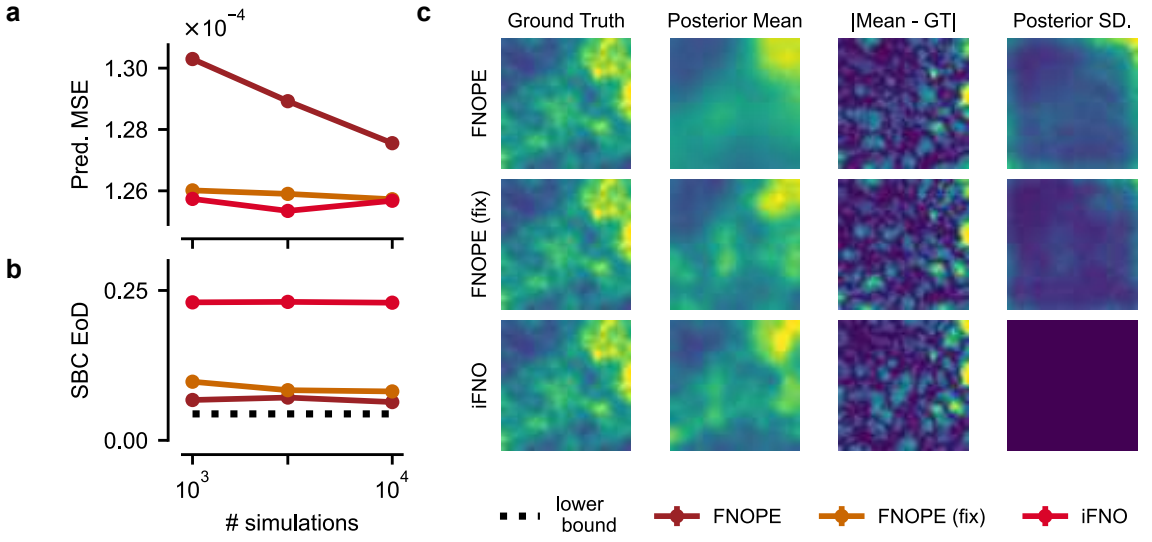

Figure S9: **Darcy Flow, comparison of FNOPE to iFNO [50].** **(a)** MSE of posterior predictives to the ground truth observation (zoomed in relative to Fig. 5). **(b)** Simulation-based calibration Error of Diagonal (SBC EoD) for different training budgets. **(c)** Ground truth θ, Posterior means, pixelwise 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 rigorousness, 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.

Sourcerer: Sample-based Maximum Entropy Source Distribution Estimation

Julius Vetter[†,1,2,*] **Guy Moss**[†,1,2,*]

Cornelius Schröder[1,2] **Richard Gao**[1,2]

Jakob H. Macke[1,2,3,*]

[1]Machine Learning in Science, Excellence Cluster Machine Learning, University of Tübingen
[2]Tübingen AI Center
[3]Department Empirical Inference, Max Planck Institute for Intelligent Systems
Tübingen, Germany
[†]Equal contribution.

Abstract

Scientific modeling applications often require estimating a distribution of parameters consistent with a dataset of observations—an inference task also known as source distribution estimation. This problem can be ill-posed, however, since many different source distributions might produce the same distribution of data-consistent simulations. To make a principled choice among many equally valid sources, we propose an approach which targets the maximum entropy distribution, i.e., prioritizes retaining as much uncertainty as possible. Our method is purely sample-based—leveraging the Sliced-Wasserstein distance to measure the discrepancy between the dataset and simulations—and thus suitable for simulators with intractable likelihoods. We benchmark our method on several tasks, and show that it can recover source distributions with substantially higher entropy than recent source estimation methods, without sacrificing the fidelity of the simulations. Finally, to demonstrate the utility of our approach, we infer source distributions for parameters of the Hodgkin-Huxley model from experimental datasets with hundreds of single-neuron measurements. In summary, we propose a principled method for inferring source distributions of scientific simulator parameters while retaining as much uncertainty as possible.

1 Introduction

In many scientific and engineering disciplines, mathematical and computational simulators are used to gain mechanistic insights. A common challenge is to identify parameter settings of such simulators that make their outputs compatible with a set of empirical observations. For example, by finding a distribution of parameters that, when passed through the simulator, produces a distribution of outputs that matches that of the empirical dataset of observations.

Suppose we have a stochastic simulator with input parameters θ and output x, which allows us to generate samples from the forward model $p(x|\theta)$ (which is usually intractable). We have acquired a dataset $\mathcal{D} = \{x_1, ..., x_n\}$ of observations with empirical distribution $p_o(x)$, and want to identify

[*]{firstname.secondname}@uni-tuebingen.de
Code available at https://github.com/mackelab/sourcerer

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

a distribution $q(\theta)$ over parameters that, once passed through the simulator, yields a "pushforward" distribution of simulations $q^{\#}(x) = \int p(x|\theta)q(\theta)d\theta$ that is indistinguishable from the empirical distribution. This setting is known by different names in different disciplines, for example as *unfolding* in high energy physics [10], *stochastic inverse problems* in various disciplines [7], *population of models* in electrophysiology [30] and *population inference* in gravitational wave astronomy [55]. Adopting the terminology of Vandegar et al. [58], we refer to this task as *source distribution estimation*.

A common approach to source distribution estimation is empirical Bayes [51, 15]. Empirical Bayes uses hierarchical models in which each observation is modeled as arising from different parameters $p(x_i|\theta_i)$. The hyper-parameters of the prior (and thus the source q_ϕ) are found by optimizing the marginal likelihood $p(D) = \prod_i \int p(x_i|\theta)q_\phi(\theta)d\theta$ over ϕ. Empirical Bayes has been successfully applied to a range of applications [31, 32, 55]. However, empirical Bayes is typically not applicable to models with intractable likelihoods, which is usually the case for scientific simulators. Using surrogate models for such likelihoods, empirical Bayes has been extended to increasingly more complicated parameterizations ϕ of the source distribution, including neural networks [59, 58].

A more general issue, however, is that the source distribution problem can often be ill-posed without the introduction of a hyper-prior or other regularization principles, as also noted in Vandegar et al. [58]: Distinct source distributions $q(\theta)$ can give rise to the same data distribution $q^{\#}(x)$ when pushed through the simulator $p(x|\theta)$ (Fig. 1, illustrative example in Appendix A.7).

We here propose to use the maximum entropy principle, i.e., choosing the "maximum ignorance" distribution within a class of distributions to resolve the ill-posedness of the source distribution problem [19, 24]. The maximum entropy principle formalizes the notion that a good choice for distributions should "assume less". It has been applied to specific source distribution estimation problems in scientific disciplines such as cosmology [23] and high-energy physics [10].

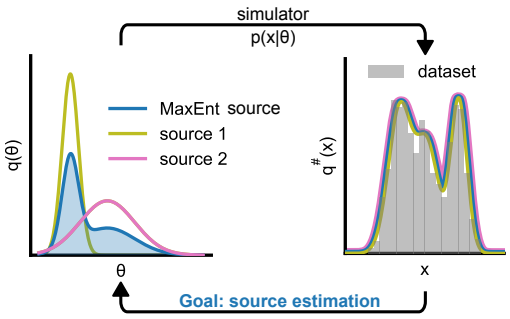

Figure 1: **Maximum entropy source distribution estimation.** Given an observed dataset $\mathcal{D} = \{x_1, \ldots, x_n\}$ from some data distribution $p_o(x)$, the *source distribution estimation* problem is to find the parameter distribution $q(\theta)$ that reproduces $p_o(x)$ when passed through the simulator $p(x|\theta)$, i.e. $q^{\#}(x) = \int p(x|\theta)q(\theta)d\theta = p_o(x)$ for all x. This problem can be ill-posed, as there might be more than one distinct source distribution. We resolve this by targeting the maximum entropy distribution, which is unique.

Our contributions We introduce *Sourcerer*, a general method for source distribution estimation, providing two key innovations: First, we target the maximum entropy source distribution to obtain a well-posed problem, thereby increasing the entropy of the estimated source distributions at no cost to their fidelity. Second, we use general distance metrics between distributions, in particular the Sliced-Wasserstein distance, instead of maximizing the marginal likelihood as in empirical Bayes. This allows evaluation of the objective using *only samples* from differentiable simulators, removing the requirement to have tractable likelihoods. We validate our method on multiple tasks, including tasks with high-dimensional observation space, which are challenging for likelihood-based methods. Finally, we apply our method to estimate the source distribution over the mechanistic parameters of the Hodgkin-Huxley model from a large (~ 1000 samples) dataset of electrophysiological recordings.

2 Methods

We formulate the source distribution estimation problem in terms of the maximum entropy principle. The (differential) entropy $H(p)$ of a distribution $p(\theta)$ is defined as

$$H(p) = -\int p(\theta) \log p(\theta) d\theta. \tag{1}$$

2.1 Data-consistency and regularized objective

For a given distribution $q(\theta)$ and a simulator with (possibly intractable) likelihood $p(x|\theta)$, the *pushforward* of q is given by $q^{\#}(x) = \int p(x|\theta)q(\theta)d\theta$. The distribution $q(\theta)$ is a source distribution if its pushfoward matches the observed data distribution $p_o(x)$, that is, $q^{\#} = p_o$ almost everywhere. Equivalently, given a distance metric $D(\cdot, \cdot)$ between probability distributions $P(\mathcal{X})$ over the data space $\mathcal{X}$, a source distribution q is one which satisfies $D(q^{\#}, p_o) = 0$. In general, for a given distribution of observations $p_o(x)$ and likelihood $p(x|\theta)$, the source distribution problem is ill-posed as there are possibly many different source distributions. The maximum entropy principle can be employed to resolve this ill-posedness:

Proposition 2.1. *Let $Q = \{q|q^{\#} = p_o\}$ be the set of source distributions for a given likelihood $p(x|\theta)$ and data distribution p_o. Suppose that Q is non-empty and compact. Then $q^* = \arg\max_{q \in Q} H(q)$ exists and is unique.*

This proposition follows from the fact that the set of source distributions is convex and that the (differential) entropy $H(q)$ is a strictly concave functional. See Appendix A.7 for a proof and additional assumptions.

Proposition 2.1 suggests to solve the constrained optimization problem

$$\max_{\phi} \quad H(q_\phi) \quad \text{s.t.} \quad D(q_\phi^{\#}, p_o) = 0, \qquad (2)$$

where q_ϕ is some parametric family of distributions.

Practically, however, a solution might not exist, for example due to simulator misspecification. Furthermore, even if a solution exists, it is difficult to obtain since we only have a fixed number of samples from p_o and can thus only estimate $D(q_\phi^{\#}, p_o)$. We therefore propose a *regularized* approximation of Eq. (2) and solve

$$\max_{\phi} \quad \lambda H(q_\phi) - (1-\lambda)\log(D(q_\phi^{\#}, p_o)) \qquad (3)$$

instead, where λ is a parameter determining the strength of the data-consistency term and the logarithm is added for numerical stability. This regularized objective is related to the Lagrangian relaxation of Eq. (2), where now $\log D(q^{\#}, p_o) \leq \log \epsilon$ for some $\epsilon > 0$ and the dual variable is $(1-\lambda)/\lambda$.

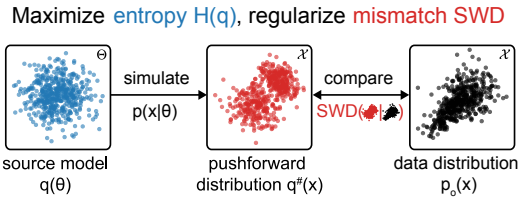

Figure 2: **Overview of Sourcerer.** Given a source distribution $q(\theta)$, we sample $\theta \sim q$ and simulate using $p(x|\theta)$ to obtain samples from the pushforward distribution $q^{\#}(x) = \int p(x|\theta)q(\theta)d\theta$. We maximize the entropy of the source distribution $q(\theta)$ while regularizing with a Sliced-Wasserstein distance (SWD) term between the pushforward of $q^{\#}$ and the data distribution $p_o(x)$ (Eq. (3)). Θ and $\mathcal{X}$ in top right corner of boxes denote parameter space and data/observation space, respectively.

For $\lambda \to 1$, the loss in Eq. (3) is dominated by the entropy term, and for $\lambda \to 0$ by the data-consistency term. We apply ideas from constrained optimization and reinforcement learning [49, 4, 1] and use a dynamical schedule during training. We initialize training with $\lambda_{t=1} = 1$, and decay this value linearly to a final value $\lambda_{t=T} = \lambda > 0$ over the course of training. This dynamical schedule encourages the variational source model to first explore high-entropy distributions, and later increase consistency with the data between high-entropy distributions. Pseudocode and details of the schedule in Appendix A.3.

2.2 Reference distribution

For many tasks, there is an additional constraint in terms of a reference distribution $p(\theta)$. For example, in the Bayesian inference framework, it is common to have a prior distribution $p(\theta)$, encoding existing knowledge about the parameters θ from previous studies. In such cases, a distribution with higher entropy than $p(\theta)$, even if it is a source distribution, is not always desirable. We therefore adapt our objective function in Eq. (3) to minimize the Kullback-Leibler (KL) divergence between the source $q(\theta)$ and the reference $p(\theta)$:

$$\min_{\phi} \quad \lambda D_{KL}(q||p) + (1-\lambda)\log(D(q^{\#}, p_o)). \qquad (4)$$

The KL divergence term can be rewritten as $D_{KL}(q||p) = -H(q) + H(q,p)$, where $H(q,p) = -\int \log(p(\theta))q(\theta)d\theta$ is the cross-entropy between q and p. Thus, provided we can evaluate the density $p(\theta)$, we can obtain a sample-based estimate of the loss in Eq. (4). In our work, we consider $p(\theta)$ to be the uniform distribution over some bounded domain B_Θ (and hence the maximum entropy distribution on this domain). This "box prior" is often used as the naive estimate from literature observations in inference studies. More specifically, in this case, $H(q,p) = -1/|B_\Theta|$, where $|B_\Theta|$ is the volume of B_Θ. Therefore, it is independent of q, and hence minimizing the KL divergence is equivalent to maximizing $H(q)$ on B_Θ. In the case where $p(\theta)$ is non-uniform (e.g., Gaussian) the cross-entropy term regularizes the loss by penalizing large $q(\theta)$ when $p(\theta)$ is small.

2.3 Sliced-Wasserstein as a distance metric

We are free to choose any distance metric $D(\cdot,\cdot)$ for the loss function Eq. (4). In this work, we use the fast, sample-based, and differentiable Sliced-Wasserstein distance (SWD) [6, 27, 42] of order two. The SWD is defined as the expected value of the one-dimensional Wasserstein distance between the projections of the distribution onto uniformly random directions u on the unit sphere $\mathbb{S}^{d-1}$ in $\mathbb{R}^d$. More precisely, the SWD is defined as

$$\text{SWD}_m(p,q) = \mathbb{E}_{u\sim\mathcal{U}(\mathbb{S}^{d-1})}[W_m(p_u, q_u)]\,, \tag{5}$$

where p_u is the one-dimensional distribution with samples $u^\top x$ for $x \sim p(x)$, and W_m is the one-dimensional Wasserstein distance of order m. In the empirical setting, where we are given n samples each from p_u and q_u respectively, the one-dimensional Wasserstein distance is computed from the order statistics as

$$W_m(p_u, q_u) = \left(\sum_{i=1}^{n} ||x_p^{(i)} - x_q^{(i)}||_m^m\right)^{1/m}, \tag{6}$$

where $x_p^{(i)}$ denotes the i-th order statistic of the samples from p_u (and similarly for $x_q^{(i)}$), and $||\cdot||_m$ denotes the L^m distance on $\mathbb{R}$ [47]. The time complexity of computing the sample-based one-dimensional Wasserstein distance is thus the time complexity of computing the order statistics, which is $\mathcal{O}(n \log n)$ in the number of datapoints n [6]. This is significantly faster than computing the multi-dimensional Wasserstein distance ($\mathcal{O}(n^3)$, 29), or the commonly used Sinkhorn algorithm for approximating the Wasserstein distance ($\mathcal{O}(n^2)$ 47). While the SWD is not the same as the multi-dimensional Wasserstein distance, it is still a valid metric on the space of probability distributions. In particular, the SWD converges quickly with rate $O(\sqrt{n})$ to its true value [41, 42].

2.4 Differentiable simulators and surrogates

Our method only requires that sampling from the simulator $p(x|\theta)$ is a differentiable operation. In practice, however, many simulators do not satisfy this property. For such simulators, we first train a surrogate model. In particular, our method can make use of surrogates that model the likelihood only implicitly. Such surrogate models can be easier to train and evaluate in practice. This is a distinct requirement from likelihood-based approaches such as Vandegar et al. [58], which require that the likelihood $p(x|\theta)$ can be evaluated explicitly *and* is differentiable. This means that our sample-based approach can be readily applied to a larger set of simulators than likelihood-based approaches.

2.5 Source model and entropy estimation

In this work we use neural samplers as proposed in Vandegar et al. [58] to parameterize a source model q_ϕ. These samplers employ unconstrained neural network architectures (in our case a multi-layer perceptron) to transform a random sample from $z \in \mathcal{N}(0, I)$ into a sample from q_ϕ. While neural samplers do not have a tractable likelihood, they are faster to evaluate than models with tractable likelihoods. Furthermore, by using unconstrained network architectures, neural samplers are flexible and additional constraints (e.g., symmetry, monotonicity) are easy to introduce.

To use likelihood-free source parameterizations, we require a purely sample-based estimator for the entropy $H(q_\phi)$. This can be done using the *Kozachenko-Leonenko* entropy estimator [28, 3], which is based on a nearest-neighbor density estimate. We use the Kozachenko-Leonenko estimator in this work for its simplicity, but note that sample-based entropy estimation is an active area of research, and other choices are possible [48]. Details about the Kozachenko-Leonenko estimator can be found in Appendix A.6.

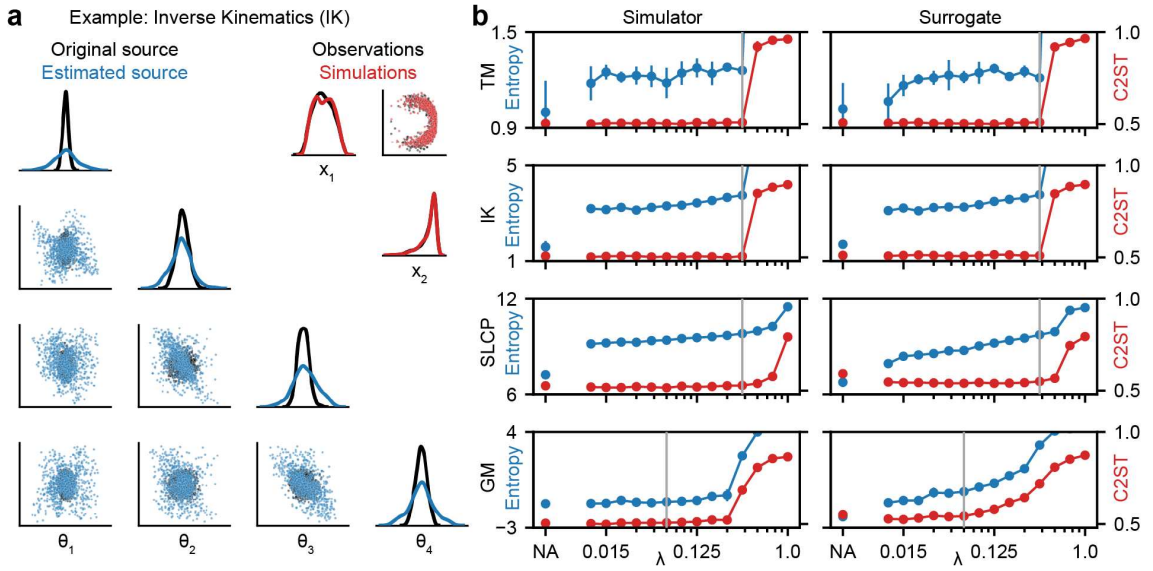

Figure 3: **Results for the source estimation benchmark. (a)** Original and estimated source and corresponding pushforward for the differentiable IK simulator ($\lambda = 0.35$). The estimated source has higher entropy than the original source that was used to generate the data. The observations (simulated with parameters from the original source) and simulations (simulated with parameters from the estimated source) match. **(b)** Performance of our approach for all four benchmark tasks (TM, IK, SLCP, GM) using both the original (differentiable) simulators, and learned surrogates. Source estimation is performed without (NA) and with entropy regularization for different choices of λ. For all cases, mean C2ST accuracy between observations and simulations (lower is better) as well as the mean entropy of estimated sources (higher is better) over five runs are shown together with the standard deviation. The gray line at $\lambda = 0.35$ ($\lambda = 0.062$ for GM) indicates our choice of final λ for the numerical benchmark results (Table 1).

3 Experiments

To evaluate the data-consistency and entropy of source distributions estimated by Sourcerer, we benchmark our method against Neural Empirical Bayes (NEB) [58], a state-of-the-art approach to source distribution estimation. The benchmark comparison is performed on four source distribution estimation tasks including three presented in Vandegar et al. [58]. We then demonstrate the advantage of Sourcerer in the case of differentiable simulators with a high-dimensional data domain, where likelihood-based empirical Bayes approaches would require training a likelihood surrogate. Finally, we use Sourcerer to estimate the source distribution for a Hodgkin-Huxley simulator of single-neuron voltage dynamics from a large dataset of experimental electrophysiological recordings. For all tasks except the Hodgkin-Huxley task (where the observed dataset is experimentally measured), we generate two datasets of observations of equal size from the same reference source distribution. The first is used to train the source model, and the second is used to evaluate the quality of the learned source.

3.1 Source Estimation Benchmark

Benchmark tasks The source estimation benchmark contains four simulators: two moons (TM), inverse kinematics (IK), simple likelihood complex posterior (SLCP), and Gaussian Mixture (GM) (details about simulators and source distributions are in Appendix A.2). Notably, all four simulators are differentiable. Therefore, we can evaluate our method directly on the simulator as well as trained surrogates. For all four simulators, source estimation is performed on a synthetic dataset of 10000 observations that were generated by sampling from a pre-defined original source distribution and evaluating the resulting pushforward distribution using the corresponding simulator. The quality of the estimated source distributions is measured using a classifier two sample test (C2ST) [33] between the observations and simulations from the source. We also report the entropy of the estimated sources. Given two sources with the same C2ST accuracy, the higher entropy source is preferable. We compare

Table 1: **Numerical benchmark results for Sourcerer.** We show the mean and standard deviation over five runs for differentiable simulators and surrogates of Sourcerer on the benchmark tasks, and compare to NEB. All approaches achieve C2ST accuracies close to 50%. For the Sliced-Wasserstein-based approach, the entropies of the estimated sources are substantially higher (bold) with the entropy regularization ($\lambda = 0.35$ for TM, IK, SLCP, $\lambda = 0.062$ for GM, gray line in Fig. 3).

Method		Sourcerer Sim. (with reg.)	Sourcerer Sim. (w/o reg.)	Sourcerer Sur. (with reg.)	Sourcerer Sur. (w/o reg.)	NEB
TM	C2ST acc.	0.51 (0.004)	0.5 (0.008)	0.51 (0.003)	0.51 (0.006)	0.53 (0.005)
	Entropy	**1.26** (0.022)	1.0 (0.198)	**1.21** (0.054)	1.02 (0.162)	1.13 (0.093)
IK	C2ST acc.	0.51 (0.002)	0.51 (0.005)	0.51 (0.005)	0.51 (0.01)	0.6 (0.014)
	Entropy	**3.75** (0.066)	1.59 (0.246)	**3.78** (0.022)	1.7 (0.165)	0.82 (0.712)
SLCP	C2ST acc.	0.53 (0.005)	0.53 (0.006)	0.55 (0.003)	0.59 (0.017)	0.53 (0.006)
	Entropy	**9.81** (0.039)	7.23 (0.052)	**9.74** (0.039)	6.76 (0.302)	7.56 (0.097)
GM	C2ST acc.	0.51 (0.005)	0.5 (0.006)	0.54 (0.006)	0.55 (0.005)	0.52 (0.004)
	Entropy	**-1.12** (0.083)	-1.25 (0.106)	**-0.36** (0.095)	-2.19 (0.212)	-1.5 (0.052)

to the NEB estimator with the same parameterization of the source model and 1024 Monte Carlo samples to estimate the marginal likelihood (details in Appendix A.3).

Benchmark performance We first check whether minimizing the Sliced-Wasserstein distance without any entropy regularization finds good source distributions. This corresponds to the case $\lambda = 0$ in Eq. (3) without any decay. In this way, we compare the data-consistency objective in Eq. (4) to the NEB objective of maximizing the marginal likelihood. We find that for the differentiable simulators, the Sliced-Wasserstein-based approach is able to find good source distributions with C2ST accuracies close to 50% for all benchmark tasks (Fig. 3, labeled NA). This also applies when we use surrogate models to generate the pushforward distributions. In particular, the quality of the estimated source distributions matches those found by NEB (Table 1).

We then apply entropy regularization as defined in Eq. (3) for all benchmark tasks. The entropy of the estimated sources is drastically increased *without* any cost in the quality of the simulations (Fig. 3b). While C2ST accuracy remains close to 50% across all benchmark tasks, the entropy of estimated sources is substantially higher than that of sources estimated with NEB, or when minimizing only the data-consistency term (Table 1). We also explore the dependence of the results on the final regularization strength λ (Fig. 3b). We observe a sharp trade-off: above a critical value of λ, the SWD term becomes too weak, and the fidelity of the simulations rapidly declines. However, below this critical value of λ, the results are robust relative to λ: the estimated sources produce simulations that match the observations, and have comparable entropy.

Additionally, for both IK and SLCP simulators, the entropy of the sources estimated by our method is higher than the entropy of the original source distribution (Fig. 3a and Fig. A7) despite the simulations and observations being indistinguishable from each other (C2ST accuracy: 50%). This does not contradict our approach: The original source distribution just happens not to be the maximum entropy source for these simulators.

We also investigate the robustness of our approach to the choice of the differentiable, sample-based distance by repeating all experiments for these benchmark tasks using the Maximum Mean Discrepancy (MMD, 22) and find comparable results (Fig. A4). Finally, we demonstrate (Fig. A5) the robustness of our approach for small dataset sizes by repeating the Two Moons task with ($N = 100$) observations (as opposed to 10000), and for high-dimensional parameter spaces by repeating the Gaussian Mixture task with $D = 25$ dimensions (as opposed to 2).

3.2 High-dimensional observations: Lotka-Volterra and SIR

Since our method is sample-based and does not require likelihoods, it is possible to estimate sources by back-propagating through the differentiable simulators directly. This is advantageous especially for simulators with high-dimensional outputs, as we no longer require to first train a surrogate likelihood model, which can be challenging when faced with high-dimensional data such as time series. Here, we

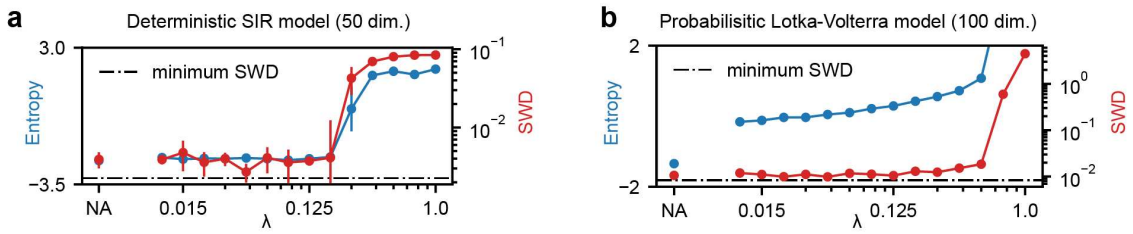

Figure 4: **Source estimation on differentiable simulators.** For both the deterministic SIR model **(a)** and probabilistic Lotka-Volterra model **(b)**, the Sliced-Wasserstein distance (lower is better) between observations and simulations as well as entropy of estimated sources (higher is better) for different choices of λ and without the entropy regularization (NA) are shown. Mean and standard deviation are computed over five runs.

highlight this capability of our method by estimating source distributions for two high-dimensional, differentiable simulators: The Lotka-Volterra model and the SIR (Susceptible, Infectious, Recovered) model. The Lotka-Volterra model is used to model the density of two populations, predators and prey. The SIR model is commonly used in epidemiology to model the spread of disease in a population (details about both models and source distributions in Appendix A.2). Compared to the benchmark tasks in Sec. 3.1, the dimensionality of the data space is much larger: Both the Lotka-Volterra and the SIR model are simulated for 50 time points resulting in a 100 and 50 dimensional time series, respectively.

Furthermore, to show that unlike NEB (which maximizes the marginal likelihood), our sample-based approach is applicable to deterministic simulators, we use a deterministic version of the SIR model with no observation noise. Similarly to the benchmark tasks, we define a source, and simulate 10000 observations using samples from this source to define a synthetic dataset on which to perform source distribution estimation. Here, we directly evaluate the quality of the estimated source distributions using the Sliced-Wasserstein distance. We compare this distance to the minimum expected distance, which is the distance between simulations of different sets of samples from the same original source. For a comparison with NEB, we train surrogate models with a reduced dimensionality and again compute C2ST accuracies and entropies of the estimated sources (see Appendix A.5 and Fig. A3 for details on surrogate training and pushforward plots).

Source estimation for the deterministic SIR model Our method is able to estimate a good source distribution for the deterministic SIR model: The Sliced-Wasserstein distance between simulations and observations is close to the minimum expected distance (Fig. 4a). In contrast to the benchmark tasks, estimating sources with entropy regularization does not lead to an increase in entropy for the SIR model, and the quality of the estimated source remains constant for various choices of λ. A possible explanation for this is that there is no degeneracy in the parameter space of the deterministic simulator, and there exists only one source distribution.

Source estimation for the probabilistic Lotka-Volterra model For the probabilistic Lotka-Volterra model, our method is also capable of estimating source distributions. As for the SIR model, the Sliced-Wasserstein distance between simulations and observations is close to the minimum expected distance (Fig. 4b). However, unlike the SIR model, estimating the source with entropy regularization yields a large increase in entropy compared to when not using the regularization. For the Lotka-Volterra model, our method yields a substantially higher entropy at no additional cost in terms of source quality.

When using the surrogate models with reduced dimensionality to estimate the source distributions, we find that Sourcerer achieves better C2ST accuracies than NEB. Furthermore, for the Lotka-Volterra model, the entropy regularization again leads to a substantial increase in the entropy of the estimated sources (Table 2). In summary, the experiments on the SIR and Lotka-Volterra models show that our approach is able to scale to higher dimensional problems and can use gradients of complex simulators to estimate source distributions directly from a set of observations.

Table 2: **Numerical results for the SIR and Lotka-Volterra model** We show the mean and standard deviation over five runs for differentiable simulators and surrogates of Sourcerer on the high-dimensional SIR and Lotka-Volterra (LV) models, and compare to NEB. For the comparison with NEB, we train the required surrogate models with reduced dimensionality (25 dimensions instead of 50 or 100). Sourcerer achieves C2ST accuracies close to 50%. For NEB, the C2ST accuracies are worse. For the LV model, the entropies of the estimated sources are higher with the entropy regularization ($\lambda = 0.015$ for SIR, $\lambda = 0.125$ for LV).

Method		Sourcerer Sim. (with reg.)	Sourcerer Sim. (w/o reg.)	Sourcerer Sur. (with reg.)	Sourcerer Sur. (w/o reg.)	NEB
SIR	C2ST acc.	0.56 (0.013)	0.56 (0.015)	0.55 (0.005)	0.55 (0.005)	0.76 (0.024)
	Entropy	-2.3 (0.079)	-2.37 (0.169)	-2.29 (0.076)	-2.5 (0.05)	-0.63 (0.174)
LV	C2ST acc.	0.57 (0.009)	0.52 (0.001)	0.56 (0.005)	0.54 (0.009)	0.62 (0.011)
	Entropy	**0.29** (0.017)	-1.34 (0.087)	**0.34** (0.05)	-1.01 (0.13)	-1.28 (0.073)

3.3 Estimating source distributions for a single-compartment Hodgkin-Huxley model

Single-compartment Hodgkin-Huxley simulator and summary statistics The single-compartment Hodgkin-Huxley model consists of a system of coupled ordinary differential equations simulating different ion channels in a neuron. We use the simulator described in Bernaerts et al. [2] with 13 parameters. In data space, we use five commonly used summary statistics of the observed and simulated spike trains. These are the (log of the) number of spikes, the mean of the resting potential, and the mean, variance and skewness of the voltage during external current stimulation. As the internal noise in the simulator has little effect on the summary statistics, we train a simple multi-layer perceptron as surrogate on 10^6 simulations. The parameters used to generate these training simulations were sampled from a uniform distribution that was used as the prior in Bernaerts et al. [2] (details on simulator, choice of surrogate and the surrogate training in Appendix A.9).

Using this surrogate, we estimate source distributions from a real-world dataset of electrophysiological recordings. The dataset [52] consists of 1033 electrophysiological recordings from the mouse motor cortex. In general, parameter inference for Hodgkin-Huxley models can be challenging as models are often misspecified [56, 2]. Thus, estimating the source distribution for this task is useful for downstream inference tasks, as the prior knowledge gained can significantly constrain the parameters of interest.

Source estimation for the Hodgkin-Huxley model On visual inspection, simulations from the estimated source look similar to the original recordings (all observations spike at least once, spikes have similar magnitudes) and show none of the unrealistic properties (e.g., spiking before the stimulus is applied) that can be observed in some of the box uniform prior simulations (Fig. 5a). This match is also confirmed by the distribution of summary statistics, which match closely between simulations and observations (Fig. 5b). Furthermore, our method achieves good C2ST accuracy of $\approx 61\%$ for different choices of λ (Fig. 5d), as well as a small Sliced-Wasserstein distance of ≈ 0.08 in the standardized space of summary statistics (Fig. 5e). While the source estimated without entropy regularization also achieves good fidelity, its entropy is significantly lower than any of the source distributions estimated with entropy regularization (Fig. 5d/e, example source distribution in Fig. 5c, full source in Fig. A11).

Overall, these results demonstrate the importance of estimating source distributions using the entropy regularization, especially on real-world datasets: Estimating the source distribution without any entropy regularization can introduce severe bias, since the estimated source may ignore entire regions of the parameter space. In this example, the parameter space of the single-compartment Hodgkin-Huxley model is known to be highly degenerate, and a given observation can be generated by multiple parameter configurations [14, 39].

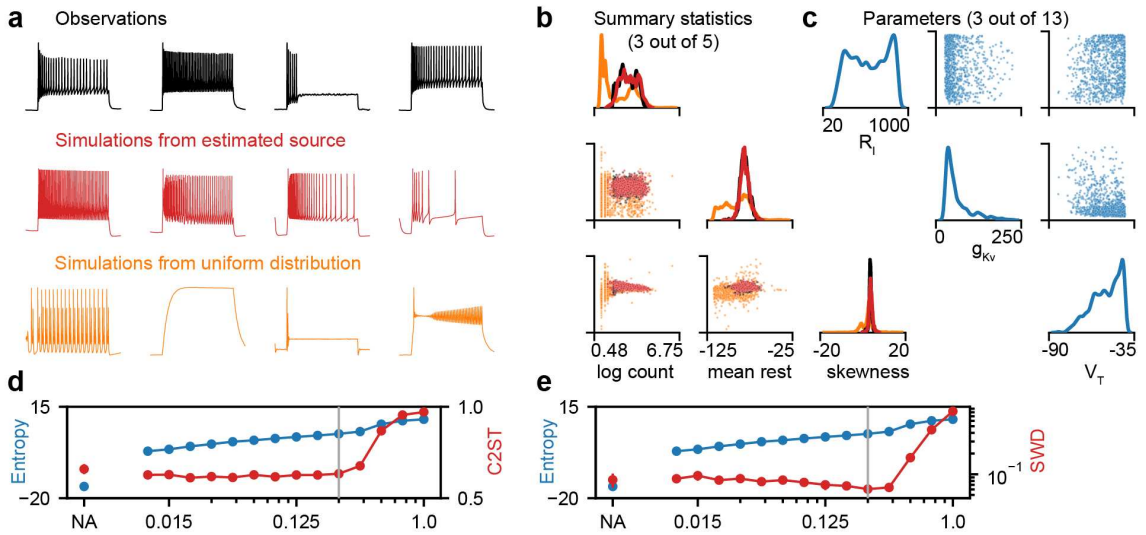

Figure 5: **Source estimation for the single-compartment Hodgkin-Huxley model. (a)** Example voltage traces of the real observations of the motor cortex dataset, simulations from the estimated source ($\lambda = 0.25$), and samples from the uniform distribution used to train the surrogate. **(b)** 1D and 2D marginals for three of the five summary statistics used to perform source estimation. **(c)** 1D and 2D marginal distributions of the estimated source for three of the 13 simulator parameters. **(d)** and **(e)** C2ST accuracy and Sliced-Wasserstein distance (lower is better) as well as entropy of estimated sources (higher is better) for different choices of λ including $\lambda = 0.25$ (gray line) and without entropy regularization (NA). Mean and standard deviation over five runs are shown.

4 Related Work

Neural Empirical Bayes High-dimensional source distributions have been estimated through variational approximations to the empirical Bayes problem. Louppe et al. [34] train a generative adversarial network (GAN) [20] q_ψ to approximate the source. The use of a discriminator to compute an implicit distance makes this approach purely sample-based as well. In order to find the optimal ψ^* of the true data-generating process, they augment the adversarial loss with a small entropy penalty on the source q_ψ. This penalty encourages low entropy, point mass distributions, which is the *opposite* of our approach. Vandegar et al. [58] take an empirical Bayes approach, and use normalizing flows for both the variational approximation of the source and as a surrogate for the likelihood $p(x|\theta)$. This allows for direct regression on the marginal likelihood, as all likelihoods can be computed directly. Finally, the empirical Bayes problem is also known as "unfolding" in the particle physics literature [10], "population inference" in gravitational wave astronomy [55], and "population of models" in electrophysiology [30]. Approaches have been developed to identify the source distribution, including classical approaches that seek to increase the entropy of the learned sources [50].

Simulation-Based Inference The use of variational surrogates of the likelihood of a simulator with intractable likelihood is known as *Neural Likelihood Estimation* in the simulation-based inference (SBI) literature [60, 45, 36, 11]. In neural posterior estimation [44, 35, 21], an *amortized* posterior density estimate is learned, which can be applied to evaluate the posterior of a single observation $x_i \in \mathcal{D}$, if a prior distribution $p(\theta)$ is already known. An intuitive but incorrect approach to source distribution estimation would be to take the *average posterior* distribution over the observations $\mathcal{D}$,

$$G_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} p(\theta|x_i). \tag{7}$$

The average posterior does not always (and typically does not) converge to a source distribution in the infinite data limit, as shown for simple examples in Appendix A.8. Intuitively, the average posterior becomes a worse approximation of a source distribution for simulators that have broader likelihoods. Instead, SBI can be seen as a downstream task of source distribution estimation; once a prior has been learned from the dataset of observations with source estimation, the posterior can be estimated for each new observation individually.

Generalized Bayesian Inference Another field related to source estimation is Generalized Bayesian Inference (GBI) [5, 40, 26]. GBI performs distance-based inference, as opposed to targeting the exact Bayesian posterior. Similarly to our work, the distance function used in GBI can be arbitrarily chosen for different tasks. However, GBI is used for single-parameter inference tasks, as opposed to the source distribution estimation task considered in this work. Similarly, Bayesian non-parametric methods [43, 38, 12] learn a posterior directly on the data space which can then be used to sample from a posterior distribution over the parameter space.

5 Summary and Discussion

In this work, we introduced Sourcerer as a method to estimate source distributions of simulator parameters given datasets of observations. This is a common problem setting across a range of scientific and engineering disciplines. Our method has several advantages: first, we employ a maximum entropy approach, improving reproducibility of the learned source, as the maximum entropy source distribution is unique while the traditional source distribution estimation problem can be ill-posed. Second, our method allows for sample-based optimization. In contrast to previous likelihood-based approaches, this scales more readily to higher dimensional problems, and can be applied to simulators without a tractable likelihood. We demonstrated the performance of our approach across a diverse suite of tasks, including deterministic and probabilistic simulators, differentiable simulators and surrogate models, low- and high-dimensional observation spaces, and a contemporary scientific task of estimating a source distribution for the single-compartment Hodgkin-Huxley model from a dataset of electrophysiological recordings. Throughout our experiments, we have consistently found that our approach yields higher entropy sources without reducing the fidelity of simulations from the learned source.

Limitations In this work, we used the Sliced-Wasserstein distance (and MMD) for the data-consistency term between simulations and observations. In practice, different distance metrics can lead to different estimated sources, depending on its sensitivity to different features. While our method is compatible with any sample-based differentiable distance metric between two distributions, there is still an onus on the practitioner to carefully select a reasonable distance metric for the data at hand. For example, in some cases, it might be appropriate to use a combination of several distance metrics for different modalities of the data. Similarly, there is a dependence on the final regularization strength λ. Principled methods for defining the regularization strength are desirable, though as we demonstrate, our results are robust to a large range of λ.

In addition, the method requires a differentiable simulator, which in practice may require the training of a surrogate model, for example, when dealing with a (partially) discrete simulator. While this is a common requirement for simulation-based methods, this could present a challenge for some applications. Finally, in our work, we enforce the maximum entropy principle on the entire (parameter) source distribution. In practice, for example when constructing prior distributions for Bayesian inference, there are other choices, such as the Jeffrey's prior [9].

Acknowledgements

This work was funded by the German Research Foundation (DFG) under Germany's Excellence Strategy – EXC number 2064/1 – 390727645 and SFB 1233 'Robust Vision' (276693517). This work was co-funded by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A and the European Union (ERC, DeepCoMechTome, 101089288). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. JV is supported by the AI4Med-BW graduate program. JV and GM are members of the International Max Planck Research School for Intelligent Systems (IMPRS-IS). We would like to thank Jonas Beck, Sebastian Bischoff, Michael Deistler, Manuel Glöckler, Jaivardhan Kapoor, Auguste Schulz, and all members of Mackelab for feedback and discussion throughout the project.

References

[1] Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In *International conference on machine learning*, 2019.

[2] Yves Bernaerts, Michael Deistler, Pedro J Goncalves, Jonas Beck, Marcel Stimberg, Federico Scala, Andreas S Tolias, Jakob H Macke, Dmitry Kobak, and Philipp Berens. Combined statistical-mechanistic modeling links ion channel genes to physiology of cortical neuron types. *bioRxiv*, 2023.

[3] Thomas B. Berrett, Richard J. Samworth, and Ming Yuan. Efficient multivariate entropy estimation via k-nearest neighbour distances. *The Annals of Statistics*, 2019.

[4] D.P. Bertsekas and W. Rheinboldt. *Constrained Optimization and Lagrange Multiplier Methods*. Computer science and applied mathematics. Elsevier Science, 2014.

[5] Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2016.

[6] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 2015.

[7] T. Butler, J. Jakeman, and T. Wildey. Combining push-forward measures and bayes' rule to construct consistent solutions to stochastic inverse problems. *SIAM Journal on Scientific Computing*, 2018.

[8] E.K.P. Chong, W.S. Lu, and S.H. Zak. *An Introduction to Optimization: With Applications to Machine Learning*. Wiley, 2023.

[9] Guido Consonni, Dimitris Fouskakis, Brunero Liseo, and Ioannis Ntzoufras. Prior Distributions for Objective Bayesian Analysis. *Bayesian Analysis*, 2018.

[10] G. Cowan. *Statistical Data Analysis*. Oxford science publications. Clarendon Press, 1998.

[11] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 2019.

[12] Charita Dellaporta, Jeremias Knoblauch, Theodoros Damoulas, and François-Xavier Briol. Robust Bayesian inference for simulator-based models via the MMD posterior bootstrap. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022.

[13] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *International Conference on Learning Representations*, 2017.

[14] Gerald M Edelman and Joseph A Gally. Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences*, 2001.

[15] Bradley Efron and Carl Morris. Limiting the risk of Bayes and empirical Bayes estimators, part ii: The empirical Bayes case. *Journal of the American Statistical Association*, 1972.

[16] Philip E. Gill, Walter Murray, and Margaret H. Wright. *Practical Optimization*. Society for Industrial and Applied Mathematics, 2019.

[17] Manuel Glöckler, Michael Deistler, and Jakob H. Macke. Adversarial robustness of amortized Bayesian inference. In *International Conference on Machine Learning*, 2023.

[18] Pedro J Gonçalves, Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F Podlaski, Sara A Haddad, Tim P Vogels, et al. Training deep neural density estimators to identify mechanistic models of neural dynamics. *Elife*, 2020.

[19] I. J. Good. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Annals of Mathematical Statistics*, 1963.

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

[21] David S. Greenberg, Marcel Nonnenmacher, and Jakob H. Macke. Automatic posterior transformation for likelihood-free inference. In *International Conference on Machine Learning*, 2019.

[22] A Gretton, KM. Borgwardt, MJ. Rasch, B Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 2012.

[23] Will Handley and Marius Millea. Maximum-entropy priors with derived parameters in a specified distribution. *Entropy*, 2018.

[24] Edwin T. Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 1968.

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[26] Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. An optimization-centric view on bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 2022.

[27] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized Sliced Wasserstein distances. In *Advances in Neural Information Processing Systems*, 2019.

[28] L. Kozachenko and N. Leonenko. A statistical estimate for the entropy of a random vector. *Problems of Information Transmission*, 1987.

[29] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 1955.

[30] Brodie A. J. Lawson, Christopher C. Drovandi, Nicole Cusimano, Pamela Burrage, Blanca Rodriguez, and Kevin Burrage. Unlocking data sets by calibrating populations of models to data density: A study in atrial electrophysiology. *Science Advances*, 2018.

[31] Tai Sing Lee and David Mumford. Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A*, 2003.

[32] Ning Leng, John A. Dawson, James A. Thomson, Victor Ruotti, Anna I. Rissman, Bart M. G. Smits, Jill D. Haag, Michael N. Gould, Ron M. Stewart, and Christina Kendziorski. EBSeq: an empirical bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, 2013.

[33] David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017.

[34] Gilles Louppe, Joeri Hermans, and Kyle Cranmer. Adversarial variational optimization of non-differentiable simulators. In *International Conference on Artificial Intelligence and Statistics*, 2019.

[35] 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*, 2017.

[36] Jan-Matthis Lueckmann, Giacomo Bassetto, Theofanis Karaletsos, and Jakob H. Macke. Likelihood-free inference with emulator networks. In *Proceedings of The 1st Symposium on Advances in Approximate Bayesian Inference*, 2019.

[37] 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.

[38] Simon Lyddon, Chris C. Holmes, and Stephen G. Walker. General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 2017.

[39] Eve Marder and Adam L Taylor. Multiple models to capture the variability in biological neurons and networks. *Nature neuroscience*, 2011.

[40] Takuo Matsubara, Jeremias Knoblauch, François-Xavier Briol, and Chris J Oates. Robust generalised Bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2022.

[41] Kimia Nadjahi, Alain Durmus, Umut Simsekli, and Roland Badeau. Asymptotic guarantees for learning generative models with the sliced-wasserstein distance. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019.

[42] Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli. Statistical and topological properties of sliced probability divergences. In *Advances in Neural Information Processing Systems*, 2020.

[43] Peter Orbanz and Yee Whye Teh. Bayesian nonparametric models. *Encyclopedia of Machine Learning*, 2010.

[44] George Papamakarios and Iain Murray. Fast ϵ-free inference of simulation models with Bayesian conditional density estimation. In *Advances in Neural Information Processing Systems*, 2016.

[45] George Papamakarios, David C. Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *International Conference on Artificial Intelligence and Statistics*, 2018.

[46] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[47] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Found. Trends Mach. Learn.*, 2018.

[48] Georg Pichler, Pierre Colombo, Malik Boudiaf, Günther Koliander, and Pablo Piantanida. A differential entropy estimator for training neural networks. In *International Conference on Machine Learning*, 2022.

[49] John Platt and Alan Barr. Constrained differential optimization. In *Neural Information Processing Systems*, 1987.

[50] Marcel Reginatto, Paul Goldhagen, and Sonja Neumann. Spectrum unfolding, sensitivity analysis and propagation of uncertainties with the maximum entropy deconvolution code MAXED. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 2002.

[51] Herbert E. Robbins. An empirical bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and basic theory*, 1956.

[52] Federico Scala, Dmitry Kobak, Matteo Bernabucci, Yves Bernaerts, Cathryn René Cadwell, Jesus Ramon Castro, Leonard Hartmanis, Xiaolong Jiang, Sophie Laturnus, Elanine Miranda, et al. Phenotypic variation of transcriptomic cell types in mouse motor cortex. *Nature*, 2021.

[53] Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 2007.

[54] Alvaro Tejero-Cantero, Jan Boelts, Michael Deistler, Jan-Matthis Lueckmann, Conor Durkan, Pedro J. Gonçalves, David S. Greenberg, and Jakob H. Macke. sbi: A toolkit for simulation-based inference. *Journal of Open Source Software*, 2020.

[55] Eric Thrane and Colm Talbot. An introduction to Bayesian inference in gravitational-wave astronomy: Parameter estimation, model selection, and hierarchical models. *Publications of the Astronomical Society of Australia*, 2019.

[56] Nicholas Tolley, Pedro LC Rodrigues, Alexandre Gramfort, and Stephanie Jones. Methods and considerations for estimating parameters in biophysically detailed neural models with simulation based inference. *bioRxiv*, 2023.

[57] Pravin M. Vaidya. An O(n log n) algorithm for the all-nearest-neighbors problem. *Discrete & Computational Geometry*, 1989.

[58] Maxime Vandegar, Michael Kagan, Antoine Wehenkel, and Gilles Louppe. Neural empirical Bayes: Source distribution estimation and its applications to simulation-based inference. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[59] Yixin Wang, Andrew C. Miller, and David M. Blei. Comment: Variational Autoencoders as Empirical Bayes. *Statistical Science*, 2019.

[60] Simon N. Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 2010.

[61] Omry Yadan. Hydra - a framework for elegantly configuring complex applications. Github, 2019. URL https://github.com/facebookresearch/hydra.

A Appendix

A.1 Software and data

We use PyTorch [46] for the source distribution estimation and hydra [61] to track all configurations. Code to reproduce results is available at https://github.com/mackelab/sourcerer.

A.2 Simulators and sources

Here we provide a definition of the four benchmark tasks Two Moons (TM), Inverse Kinematics (IK), Simple Likelihood Complex Posterior (SLCP) and Gaussian Mixture (GM), as well as the two high-dimensional simulators, the SIR and Lotka-Volterra model. We also describe the original source distribution used to generate the synthetic observations, and the bounds of the reference uniform distribution on the parameters.

A.2.1 Two moons simulator

Dimensionality	$x \in \mathbb{R}^2, \theta \in \mathbb{R}^2$
Bounded domain	$[-5, 5]^2$
Original source	$\theta \sim \mathcal{U}([-1, 1]^2)$
Simulator	$\boldsymbol{x}\|\theta = \begin{bmatrix} r\cos(\alpha) + 0.25 \\ r\sin(\alpha) \end{bmatrix} + \begin{bmatrix} -\|\theta_1 + \theta_2\|/\sqrt{2} \\ (-\theta_1 + \theta_2)/\sqrt{2} \end{bmatrix}$, where $\alpha \sim U(-\pi/2, \pi/2)$, $r \sim \mathcal{N}(0.1, 0.01^2)$.
References	Vandegar et al. [58], Lueckmann et al. [37]

A.2.2 Inverse Kinematics simulator

Dimensionality	$x \in \mathbb{R}^2, \theta \in \mathbb{R}^4$
Bounded domain	$[-\pi, \pi]^4$
Original source	$\theta \sim \mathcal{N}(0, \text{Diag}(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}))$
Simulator	$x_1 = \theta_1 + l_1 \sin(\theta_2 + \epsilon) + l_2 \sin(\theta_2 + \theta_3 + \epsilon) + l_3 \sin(\theta_2 + \theta_3 + \theta_4 + \epsilon)$, $x_2 = l_1 \cos(\theta_2 + \epsilon) + l_2 \cos(\theta_2 + \theta_3 + \epsilon) + l_3 \cos(\theta_2 + \theta_3 + \theta_4 + \epsilon)$, where $l_1 = l_2 = 0.5$, $l_3 = 1.0$ and $\epsilon \sim \mathcal{N}(0, 0.00017^2)$.
References	Vandegar et al. [58]

A.2.3 SLCP simulator

Dimensionality	$x \in \mathbb{R}^8, \theta \in \mathbb{R}^5$
Bounded domain	$[-5, 5]^5$
Original source	$\theta \sim \mathcal{U}([-3, 3]^5)$
Simulator	$x\|\theta = (x_1, \ldots, x_4)$, $x_i \sim \mathcal{N}(m_\theta, S_\theta)$, where $m_\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$, $S_\theta = \begin{bmatrix} s_1^2 & \rho s_1 s_2 \\ \rho s_1 s_2 & s_2^2 \end{bmatrix}$, $s_1 = \theta_3^2, s_2 = \theta_4^2, \rho = \tanh\theta_5$.
References	Vandegar et al. [58], Lueckmann et al. [37]

A.2.4 Gaussian mixture simulator

Dimensionality	$x \in \mathbb{R}^2, \theta \in \mathbb{R}^2$
Bounded domain	$[-5, 5]^2$
Original source	$\theta \sim \mathcal{U}([0.5, 1]^2)$
Simulator	$x\|\theta \sim 0.5\mathcal{N}(x\|\theta, I) + 0.5\mathcal{N}(x\|\theta, 0.01 \cdot I)$.
References	Sisson et al. [53]

A.2.5 SIR model

Dimensionality	$x \in \mathbb{R}^{50}$, $\theta \in \mathbb{R}^2$
Bounded domain	$[0.001, 3]^2$
Original source	$\beta \sim LogNormal(\log(0.4), 0.5)$ $\gamma \sim LogNormal(\log(0.125), 0.2)$
Simulator	$x\|\theta = (x_1, \ldots, x_{50})$, where $x_i = I_i/N$ equally spaced and I is simulated from $\frac{dS}{dt} = -\beta\frac{SI}{N}$, $\frac{dI}{dt} = \beta\frac{SI}{N} - \gamma I$, $\frac{dR}{dt} = \gamma I$ with initial values $S = N - 1$, $I = 1$, $R = 0$ and $N = 10^6$.
References	Lueckmann et al. [37]

A.2.6 Lotka-Volterra model

Dimensionality	$x \in \mathbb{R}^{100}$, $\theta \in \mathbb{R}^4$
Bounded domain	$[0.1, 3]^4$
Original source	$\theta' \sim \mathcal{N}(0, 0.5^2)^4$, pushed through $\theta = f(\theta') = \exp(\sigma(\theta'))$, where σ is the sigmoid function.
Simulator	$x\|\theta = (x_1^X, \ldots, x_{50}^X, x_1^Y, \ldots, x_{50}^Y)$, where $x_i^X \sim \mathcal{N}(X, 0.05^2)$, $x_i^Y \sim \mathcal{N}(Y, 0.05^2)$ equally spaced, and X, Y are simulated from $\frac{dX}{dt} = \alpha X - \beta XY$, $\frac{dY}{dt} = -\gamma Y + \delta XY$ with initial values $X = Y = 1$.
References	Glöckler et al. [17]

A.3 Pseudocode and details on source estimation for benchmark tasks

Pseudocode for Sourcerer is provided in Algorithm 1.

For both the benchmark tasks and high dimensional simulators, sources were estimated from 10000 synthetic observations that were generated by simulating samples from an original previously defined source.

For the benchmark tasks, we used $T = 500$ linear decay steps from $\lambda_{t=0}$ to $\lambda_{t=T} = \lambda$ and optimized the source model using the Adam optimizer with a learning rate of 10^{-4} and weight decay of 10^{-5}. The two high dimensional simulators were optimized with a higher learning rate of 10^{-3} and $T = 50$ linear decay steps. In both cases, early stopping was performed when the overall loss in Eq. (4) did not improve over a set number of training iterations.

As a baseline, we compare to Neural Empirical Bayes (NEB) as described in Vandegar et al. [58]. Specifically, we use the biased estimator with 1024 samples per observation ($\mathcal{L}_{1024}$), which are used to compute the Monte Carlo integral. Unlike our Sliced-Wasserstein-based approach, NEB does not operate on the whole dataset of observations directly but attempts to maximize the marginal likelihood per observation and thus uses part of the observations as a validation set. To ensure a fair comparison, we increased the number of observations to 11112 for all NEB experiments, which results in a training dataset of 10000 observations when using 10% as a validation set. For training, we again used the Adam optimizer (learning rate 10^{-4}, weight decay 10^{-5}, training batch size 128).

A.4 Source model

Throughout all our experiments, we use neural samplers as the source models [58]. The sampler architecture is a three-layer multi-layer perceptron with dimension of 100, ReLU activations and batch normalization as our source model. Samples are generated by drawing a sample $s \sim \mathcal{N}(0, I)$ from the standard multivariate Gaussian and then (non-linearly) transforming s with the neural network.

A.5 Surrogates for the benchmark tasks

We follow Vandegar et al. [58] and train RealNVP flows [13] as surrogates for the four benchmark tasks. For all benchmark tasks, the RealNVP surrogates have a flow length of 8 layers with a hidden dimension of 50.

Surrogates for the benchmark tasks were trained using the Adam optimizer [25] on 15000 samples and simulator evaluations from the uniform distribution over the bounded domain (learning rate 10^{-4}, weight decay $5 \cdot 10^{-5}$, training batch size 256). In addition, 20% of the data was used for validation.

Algorithm 1: Sourcerer

Inputs: Source model q_ϕ constrained on the bounded domain B_Θ, observed dataset $\mathcal{D} = \{x_1, ..., x_n\} \sim p_o(x)$, differentiable model $p(x|\theta)$ to draw samples from (simulator or surrogate), number of samples m to estimate entropy, regularization schedule $\lambda_{t=1}, ..., \lambda_{t=T}$.
Outputs: Trained source model $q_\phi(\theta)$.

$t \leftarrow 0$;
while *not converged* **do**
  $\theta_1, \ldots, \theta_n \sim q_\phi(\theta)$; `# sample parameters for pushforward`
  $x'_i \sim p(x|\theta_i)$; `# sample pushforward`
  $\theta'_1, \ldots, \theta'_m \sim q_\phi(\theta)$; `# sample parameters for entropy estimation`
  $\lambda \leftarrow \lambda_{t=t}$ **if** $t \leq T$ **else** $\lambda_{t=T}$; `# schedule lambda`
  $\mathcal{L} \leftarrow \lambda H(\{\theta'_1, \ldots, \theta'_m\}) + (1-\lambda) D(\{x_1, \ldots x_n\}, \{x'_1, \ldots, x'_n\})$; `# compute loss`
  $\phi \leftarrow \phi - \text{Adam}(\nabla_\phi \mathcal{L})$; `# update source model`
  $t \leftarrow t + 1$
return q_ϕ

To train surrogate models for the SIR and Lotka-Volterra model, we first reduce the simulator dimension in observation space to 25 in both cases. Additionally, we add a small amount of independent Gaussian noise ($\mathcal{N}(X, 0.01^2)$) to the output of the SIR simulator to avoid training the normalizing flow surrogate with simulations from a deterministic likelihood. We then use 10^6 simulations to train and validate (20% validation set) both surrogate models, again using the Adam optimizer (learning rate $5 \cdot 10^{-4}$, weight decay $5 \cdot 10^{-5}$, training batch size 256).

A.6 Kozachenko-Leonenko entropy estimator

Our use of neural samplers requires us to use a sample-based estimate of (differential) entropy, since no tractable likelihood is available (see Sec. 2.5).

We use the Kozachenko-Leonenko estimator [28, 3] for a set of samples $\{\theta_i\}_{i=1}^n$ from a distribution $p(\theta) \in P(\Theta)$, given by

$$H(q_\phi) \approx \frac{d}{m} \left[\sum_{i=1}^{n} \log(d_i) \right] - g(k) + g(n) + \log(V_d), \tag{8}$$

where d_i is the distance of θ_i from its k-th nearest neighbor in $\{\theta_j\}_{j \neq i}$, d is the dimensionality of Θ, m is the number of non-zero values of d_i, g is the digamma function, and V_d is the volume of the unit ball using the same distance measure as used to compute the distances d_i.

The Kozachenko-Leonenko estimator is differentiable and can be used for gradient-based optimization. The all-pairs nearest neighbor problem can be efficiently solved in $\mathcal{O}(n \log n)$ [57]. In practice, we find all nearest neighbors by computing all pairwise distances on a fixed number of samples. Throughout all experiments, 512 source distribution samples were used to estimate the entropy during training.

A.7 Uniqueness of maximum entropy source distribution

Here, we prove the uniqueness of the maximum entropy source distribution (Proposition 2.1). First, however, we demonstrate for a simple example that the source distribution without the maximum entropy condition is not unique.

Example of non-uniqueness Consider the (deterministic) simulator $x = f(\theta) = |\theta|$. Further assume that our observed distribution is the uniform distribution $p(x) = \mathcal{U}(x; a, b)$, where $0 < a < b$. Due the symmetry of f, the source distribution $p(\theta)$ for the observed distribution $p(x)$ is not unique. Any convex combination of form $\alpha u_1(\theta) + (1-\alpha) u_2$, where $u_1(\theta) = \mathcal{U}(\theta; -b, -a)$ and $u_2(\theta) = \mathcal{U}(\theta; a, b)$ and $\alpha \in [0, 1]$ provides a source distribution. The maximum entropy source distribution is unique and is attained if both distributions are weighted equally with $\alpha = 0.5$.

Proof of Proposition 2.1 First, let us state Proposition 2.1 in full:

Let $\Theta \subset \mathbb{R}^{d_\Theta}$ and $\mathcal{X} \subset \mathbb{R}^{d_\mathcal{X}}$ be the parameter and observation spaces, respectively. Suppose that Θ is compact. Let $\mathcal{P}(\Theta) \subset L^1(\Theta)$ and $\mathcal{P}(\mathcal{X}) \subset L^1(\mathcal{X})$ be the set of probability measures on Θ and $\mathcal{X}$ respectively. Let $Q = \{q | q^\# = p_o \text{ almost everywhere } \} \subset \mathcal{P}(\Theta)$ be the set of source distributions for a given likelihood $p(x|\theta)$ and data distribution $p_o \in \mathcal{P}(\mathcal{X})$. Suppose that Q is non-empty and compact (in the L^1 norm topology). Then $q^* = \arg\max_{q \in Q} H(q)$ exists and is unique.

First, by the compactness assumption on Θ, the (differential) entropy of all $q \in P(\Theta)$ is bounded above (by the entropy of the uniform distribution on Θ), and so in particular it is finite. By the compactness assumption on Q, the entropy achieves its supremum of Q, that is, there exists a q^* such that $H(q^*) = \arg\max_{q \in Q} H(q)$. To show that q^* is unique (up to L^1-null sets), it is sufficient to show two results: (1) that the set Q is a convex set, and (2) that entropy is strictly concave. In this case, if we have two distinct suprema q_1^* and q_2^*, then any convex combination of q_1^*, q_2^* is a valid source distribution with higher entropy, causing a contradiction. For the remainder of this proof, we let q_1 and q_2 be two distinct source distributions. Their convex combination $q = \alpha q_1 + (1-\alpha) q_2$, $\alpha \in [0, 1]$ is a valid probability distribution supported on both of the supports of q_1 and q_2.

(1) *Sources distributions are closed under convex combination*: q is also a source distribution, since

$$\begin{aligned} q^\#(x) &= \int p(x|\theta) \cdot (\alpha q_1(\theta) + (1-\alpha) q_2(\theta)) d\theta \\ &= \alpha \int p(x|\theta) q_1(\theta) d\theta + (1-\alpha) \int p(x|\theta) q_2(\theta) d\theta \\ &= \alpha p_o(x) + (1-\alpha) p_o(x) = p_o(x). \end{aligned} \tag{9}$$

(2) *Entropy is (strictly) concave*: the entropy of q satisfies

$$\begin{aligned} H(q) &= -\int (\alpha q_1(\theta) + (1-\alpha) q_2(\theta)) \cdot \log(\alpha q_1(\theta) + (1-\alpha) q_2(\theta)) d\theta \\ &\geq -\int [\alpha q_1(\theta) \log(q_1(\theta)) + (1-\alpha) q_2(\theta) \log(q_2(\theta))] d\theta \\ &= \alpha H(q_1) + (1-\alpha) H(q_2), \end{aligned} \tag{10}$$

where we used the fact that the function $f(x) = x \log x$ is convex on $[0, \infty)$, and hence $-f$ is concave. Furthermore, $f(x)$ is strictly convex on $[0, \infty)$, so for any $\theta \in \Theta$, the equality of the integrands

$$\alpha q_1(\theta) + (1-\alpha) q_2(\theta)) \log(\alpha q_1(\theta) + (1-\alpha) q_2(\theta)) = \alpha q_1(\theta) \log(q_1(\theta) + (1-\alpha) q_2(\theta) \log(q_2(\theta) \tag{11}$$

holds if and only if $\alpha \in \{0, 1\}$ or $q_1(\theta) = q_2(\theta)$. Since q_1 and q_2 are assumed distinct, that is, it holds $q_1(\theta) \neq q_2(\theta)$ on a positive measure set, the integral equality in Eq. (10) only holds if $\alpha \in \{0, 1\}$, and thus entropy is strictly concave, which concludes our proof.

□

Regularized regression as an approximation to constrained optimization In practice, we approximate the optimization problem in Eq. (2) with the regularized regression objective in Eq. (3). As a result, we cannot use the result of Proposition 2.1 to guarantee the uniqueness of our solution. However, the dynamic schedule approach to λ we use in our work (see Appendix A.3) is similar to the penalty method of approximating solutions to constrained optimization tasks [16, 8]. Future work could use this connection to apply theoretical knowledge of constrained optimization in the source distribution estimation setting.

A.8 Examples related to the average posterior distribution

In general, the average posterior distribution is not a source distribution. The average posterior distribution is defined in Eq. (7). The infinite data limit is given by $G_n(\theta) \xrightarrow{n \to \infty} G(\theta) = \int p(\theta|x) p_o(x) dx$.

Here, we provide two examples, one based on coin flips, and one based on a Gaussian bimodal likelihood to illustrate this point.

Coin-flip example Consider the classical coin flip example, where the probability of heads (H) follows a Bernoulli distribution with parameter θ. The source distribution estimation problem for this setting would consist of the outcomes of flipping n distinct coins, with potentially different values θ_i.

Proposition A.1. *Suppose we have a Beta prior distribution on the Bernoulli parameter* $\theta \sim Beta(\alpha, \beta)$ *with parameters* $\alpha = \beta = 1$*, and that the empirical measurements consist of* 70% *heads, i.e.:*

$$p_o(x) = \begin{cases} 0.7 & x = H \\ 0.3 & x = T \end{cases}$$

Then the average posterior $G(\theta) = \int p(\theta|x)p_o(x)dx$ *is* not *a source distribution for* $p_o(x)$.

Proof: Since the Beta distribution is the conjugate prior for the Bernoulli likelihood, the single-observation posteriors are known to be $p(\theta|x = \text{H}) = Beta(2, 1)$ and $p(\theta|x = \text{T}) = Beta(1, 2)$. Hence, the average posterior is

$$G(\theta) = 0.3 \cdot Beta(1, 2) + 0.7 \cdot Beta(2, 1). \tag{12}$$

However, the ratio of heads observed when pushing this distribution through the Bernoulli simulator is

$$\begin{aligned} G^{\#}(x = \text{H}) &= \int_0^1 \theta[0.3 \cdot Beta(\theta; 1, 2) + 0.7 \cdot Beta(\theta; 2, 1)]d\theta \\ &= \int_0^1 \theta \left[0.3\frac{1-\theta}{B(1,2)} + 0.7\frac{\theta}{B(2,1)}\right] d\theta \\ &= 2\int_0^1 [0.3\theta(1-\theta) + 0.7\theta^2]d\theta \\ &= 0.3\theta^2 + \frac{2}{3}0.4\theta^3\Big|_0^1 \approx 0.567 \neq 0.7, \end{aligned} \tag{13}$$

where we have used the fact that the Beta function takes the values $B(1, 2) = B(2, 1) = 1/2$. Therefore, the pushforward of the average posterior distribution does not recover the correct ratio of heads, and so it is not a source distribution.

Gaussian bimodal example As another illustrative example to show the differences between average posterior and estimated source, we consider a one-dimensional, bimodal Gaussian likelihood given by $x|\theta \sim 0.5\mathcal{N}(x|\theta - 1, 0.3^2) + 0.5\mathcal{N}(x|\theta + 1, 0.3^2)$ and the source $\mathcal{N}(\theta|0, 0.25^2)$. We use the `sbi` package [54] and perform neural posterior estimation with the uniform prior $\theta \sim \mathcal{U}([-5, 5])$ to obtain the average posterior and compare it to the source estimated with our approach.

While the estimated source matches the original source closely, the average posterior is visibly different and substantially broader (Fig. A1). As expected, this difference persists when sampling from the average posterior and estimated source to simulate from the likelihood. The pushforward distributions in data space of the original and estimated source match, while the one of the average posterior is again substantially different (Fig. A1).

Additional average posteriors (in comparison to original and estimated source distributions) for the Two Moons and Gaussian mixture are shown in Fig. A6.

A.9 Details on source estimation for the single-compartment Hodgkin-Huxley model

We use the simulators as described in Bernaerts et al. [2] for our source estimation. This work provides a uniform prior over a specified box domain, which we use as the reference distribution for source estimation. Since the simulator parameters live on different orders of magnitude, we transform the original m-dimensional box domain to the $[-1, 1]^m$ cube. Note that this transformation does not affect the maximum entropy source distribution. This is because this scaling results in a constant term added to the (differential) entropy. More specifically, for a random variable X (associated with

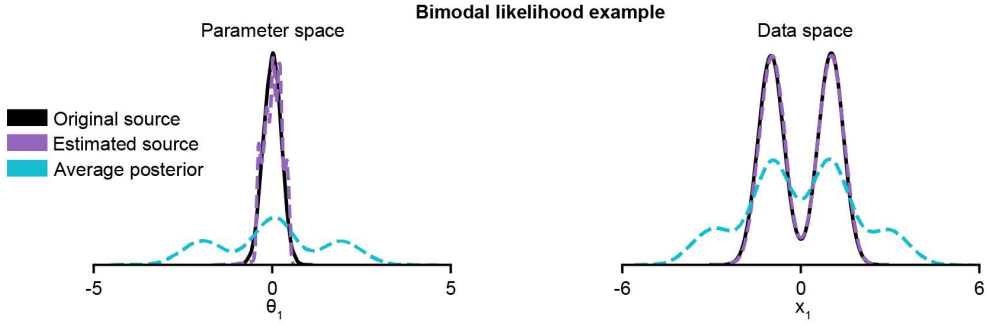

Figure A1: Failure of the average posterior as a source distribution for the bimodal likelihood example. Each of the individual posteriors is bimodal, resulting in an average posterior with 3 modes (left), the secondary modes produce observations which are not observed in the data distribution when pushed through the likelihood (right), and should not be part of the source distribution.

its probability density $p(x)$), the (differential) entropy of X scaled by a (diagonal) scaling matrix D and shifted by a vector c is given by

$$H(DX + c) = H(X) + \log(\det D). \tag{14}$$

The surrogate is trained on 10^6 parameter-simulation pairs produced by sampling parameters from the uniform distribution and simulating with the sampled parameters. We do not use the simulated traces directly, but instead compute 5 commonly used summary statistics [2, 18]. These are the number of spikes k transformed by a $\log(k + 3)$ transformation (ensuring it is defined in the case of $k = 0$), the mean of the resting potential, and the first three moments (mean, variance, and skewness) of the voltage during the stimulation.

As our surrogate, we choose a deterministic multi-layer perceptron, because we found that the internal noise has almost no noticeable effect on the summary statistics, so that the likelihood $p(x|\theta)$ is essentially a point function. We are able to make this choice because the sample based nature of our source distribution estimation approach is less sensitive to sharp likelihood functions, whereas likelihood-based approaches could struggle with such problems.

The multi-layer perceptron (MLP) surrogate has 3 layers with a hidden dimension of 256. ReLU activations and batch normalization were used. Training of the MLP was done with Adam (learning rate $5 \cdot 10^{-4}$, weight decay 10^{-5}, training batch size 4096). Again, 20% of the data were used for validation.

A.10 Computational Resources

All numerical experiments reported in this work were performed on GPU using an NVIDIA A100 GPU. A single source estimation run for a benchmark task using the Sourcerer approach (for one value of λ) took approx. 30 seconds. In comparison, learning the source using NEB for the same task took approx. 2 minutes (see Table A1). A source estimation run for Sourcerer on the high-dimensional tasks took approx. 10 min. When the observations are high-dimensional, training a surrogate (if required) makes up the majority of the computational cost. For the Hodgkin-Huxley task, training a surrogate took approx. 20 minutes, after which estimating the source distribution with Sourcerer took approx. 30 seconds.

Table A1: **Wall-clock runtime comparison between Sourcerer and NEB.** Time in seconds measured on an Nvidia A100 GPU. Average and standard deviation are shown over 5 runs. For all three settings (Sourcerer with and without entropy regularization, NEB), surrogate models for the benchmark simulators were used. Sourcerer converges noticeably faster than the NEB baseline.

Method	Sur. (w/o reg.)	Sur. (with reg.)	NEB
TM	29.4 (8.5)	63.9 (10.1)	145.2 (13.9)
IK	28.5 (6.9)	66.7 (10.0)	116.8 (22.6)
SLCP	71.7 (12.8)	53.1 (12.2)	91.6 (9.9)
GM	26.6 (5.4)	46.2 (9.2)	98.5 (15.5)

A.11 Supplementary figures

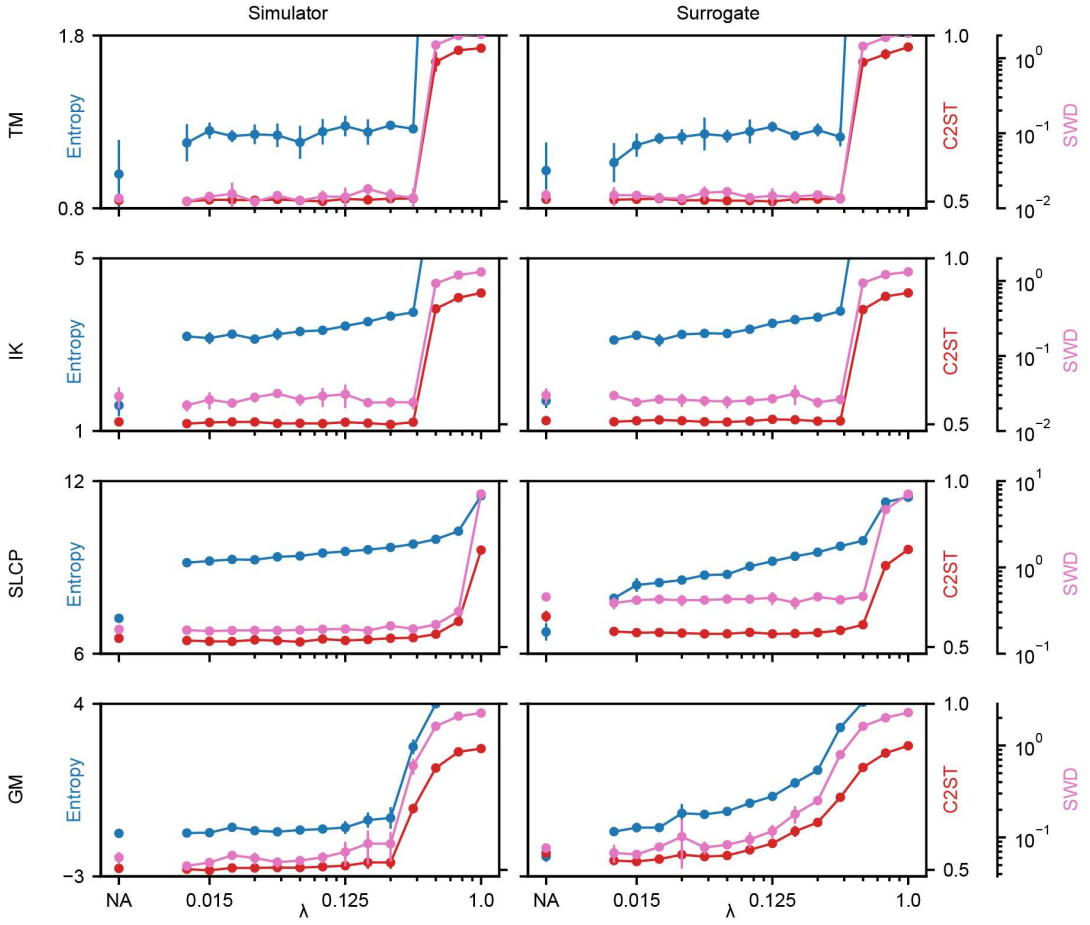

Figure A2: Extended results for source distribution estimation on the benchmark tasks (Fig. 3) for different choices of λ. In addition to the C2ST accuracy and entropy, here the Sliced-Wasserstein distance (SWD) between the observations and the pushforward distribution of the estimated source is shown. Mean and standard deviation were computed over five runs.

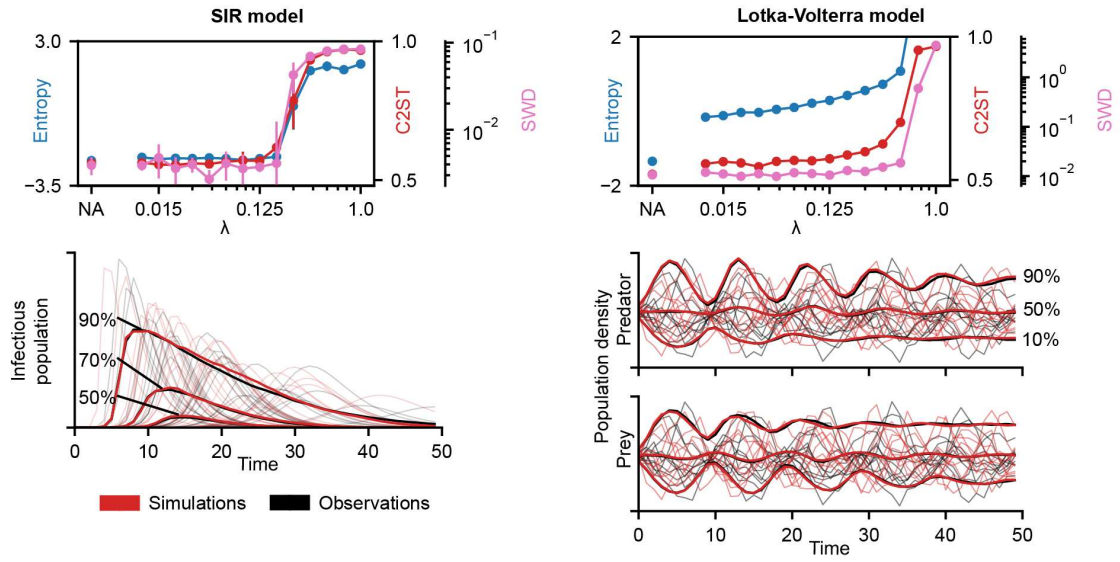

Figure A3: Extended results for source distribution estimation on the differentiable SIR and Lotka-Volterra models (Fig. 4). In addition to the Sliced-Wasserstein distance (SWD), the C2ST accuracy between the observations and the pushforward distribution of the the estimated source is shown. Despite the high-dimensional data space of the simulators (50 and 100 dimensions), the estimated sources achieve a good C2ST accuracy (below 60%) for various choices of λ. Mean and standard deviation were computed over five runs. Additionally, percentile values of all samples computed per time point between simulations (simulated with parameters from the estimated source) and observations (simulated with parameters from the original source) closely match.

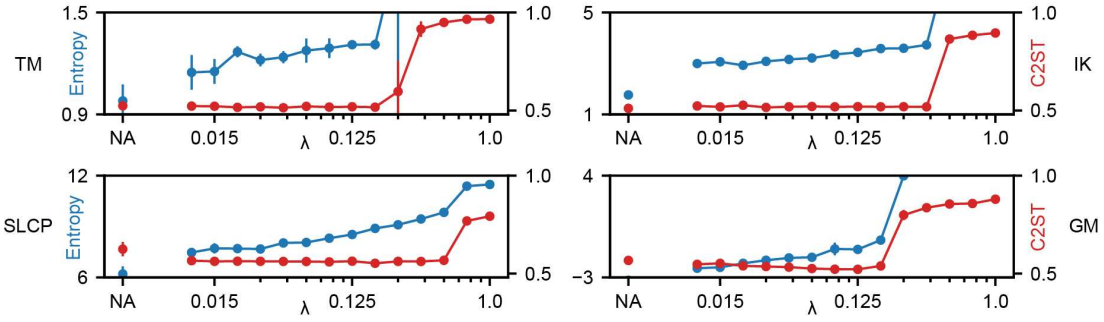

Figure A4: Sourcerer with Maximum Mean Discrepancy (MMD) as the differentiable, sample-based distance. We use MMD with an RBF kernel and the median distance heuristic for selecting the kernel length scale. Source estimation is performed without (NA) and with entropy regularization for different choices of λ. For these tasks, MMD produces similar results to the previously used SWD (Fig. 3b). These results show that Sourcerer is compatible with other sample-based, differentiable distances other than the SWD. For all cases, mean C2ST accuracy between observations and simulations (lower is better) as well as the mean entropy of estimated sources (higher is better) over five runs are shown together with the standard deviation.

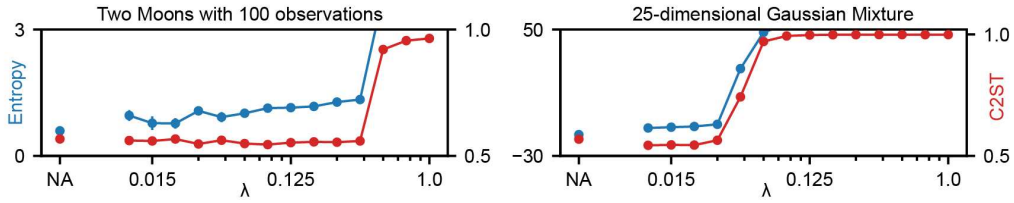

Figure A5: Experiments with less observations and higher-dimensional sources. Source estimation without (NA) and with entropy regularization for different choices of λ. For the Two Moons task, the number of observations was reduced from 10000 to 100. For the Gaussian Mixture task, the dimensionality was increased from 2 to 25. These results show that Sourcerer is robust to small datasets of observations, and can estimate high-dimensional source distributions. For all cases, mean C2ST accuracy between observations and simulations (lower is better) as well as the mean entropy of estimated sources (higher is better) over five runs are shown together with the standard deviation.

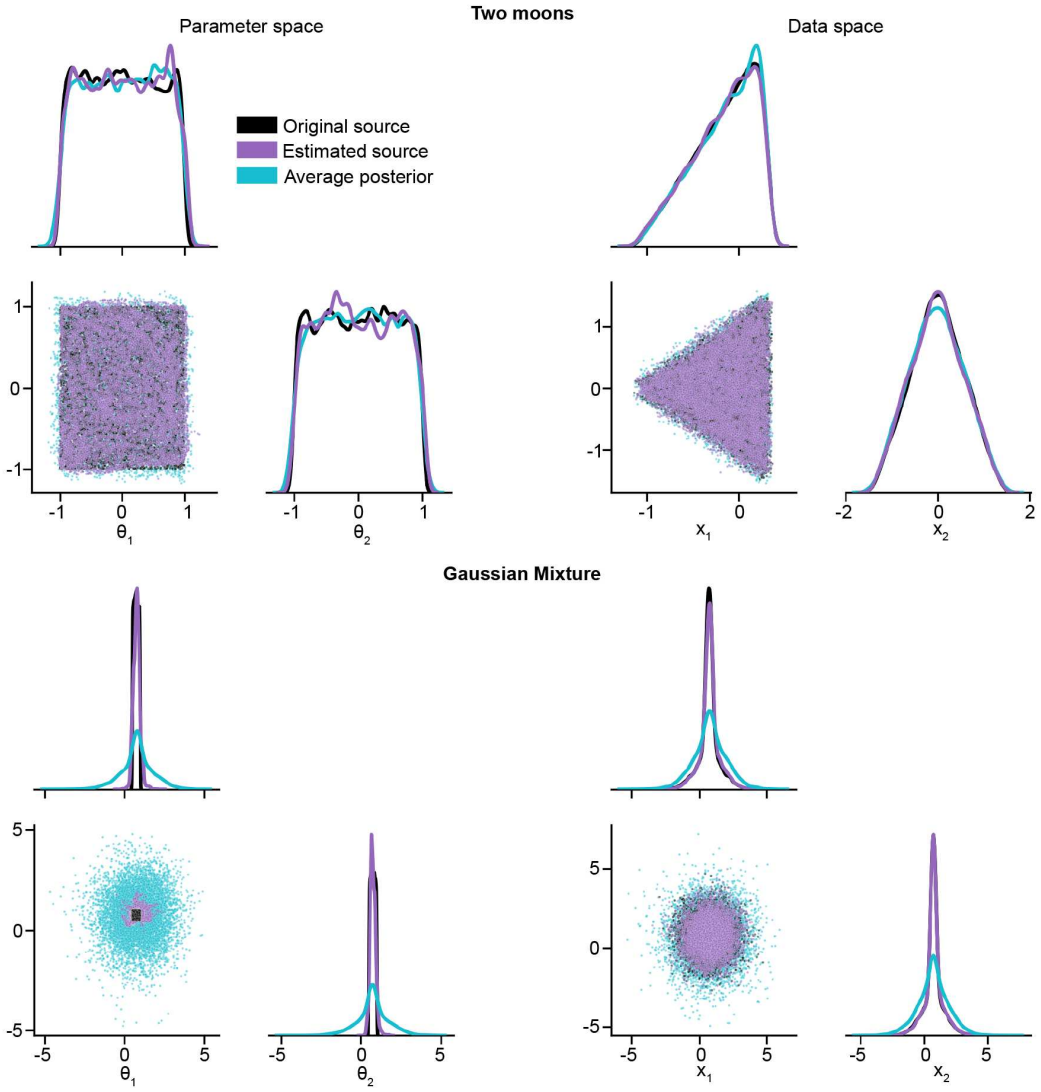

Figure A6: Original and estimated sources distributions as well as average posterior distribution for Two Moons and Gaussian Mixture simulator with uniform prior $\theta \sim \mathcal{U}([-5, 5]^2)$. For simulators for which the likelihood is unimodal and narrow, such as the Two Moons simulator, the average posterior can be a good approximation of a source distribution. However, for simulators where the likelihood is broader, such as the Gaussian Mixture simulator, the average posterior is too broad, and does not reproduce the data distribution p_o well, when compared to estimates of source distributions.

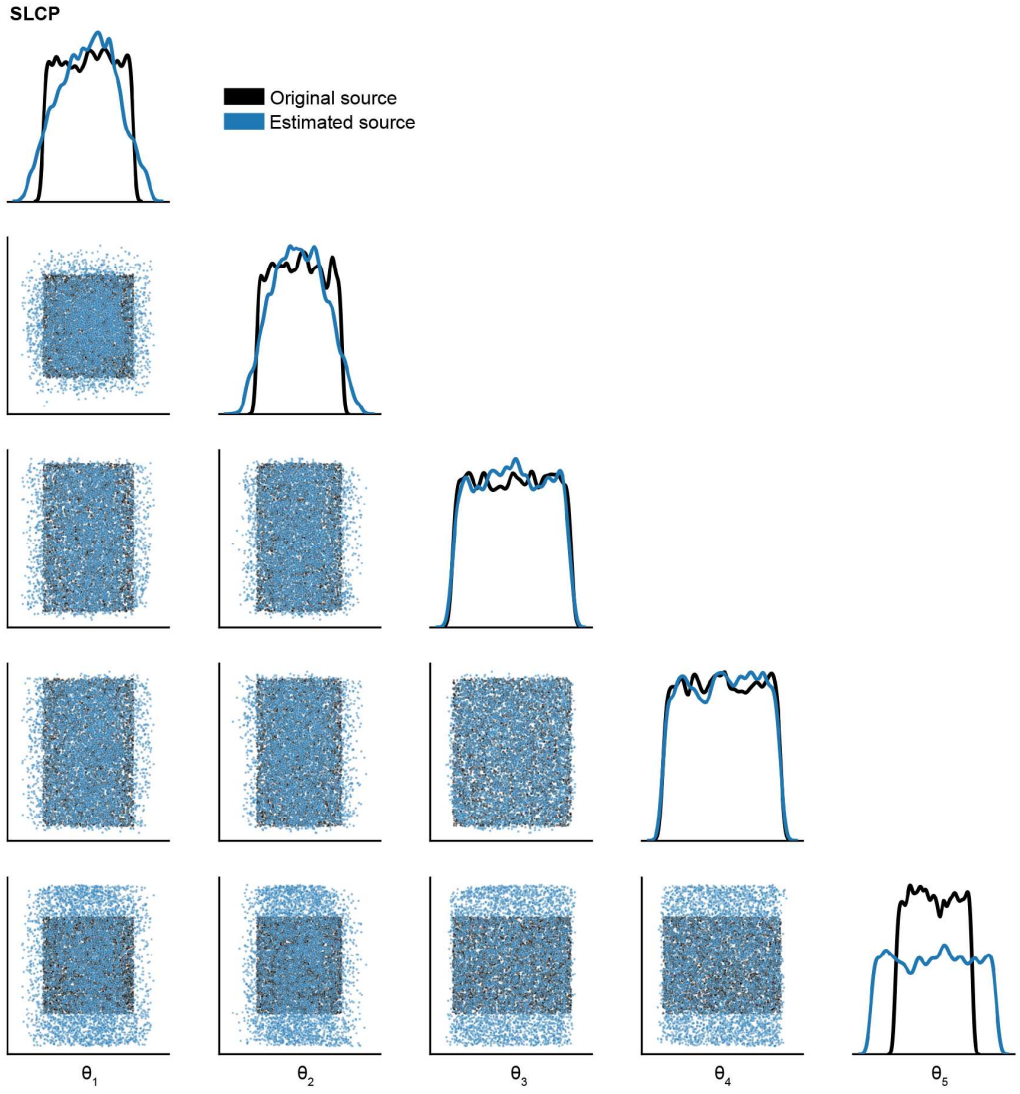

Figure A7: Original and estimated source distributions for the benchmark SLCP simulator. The estimated source has higher entropy than the original source.

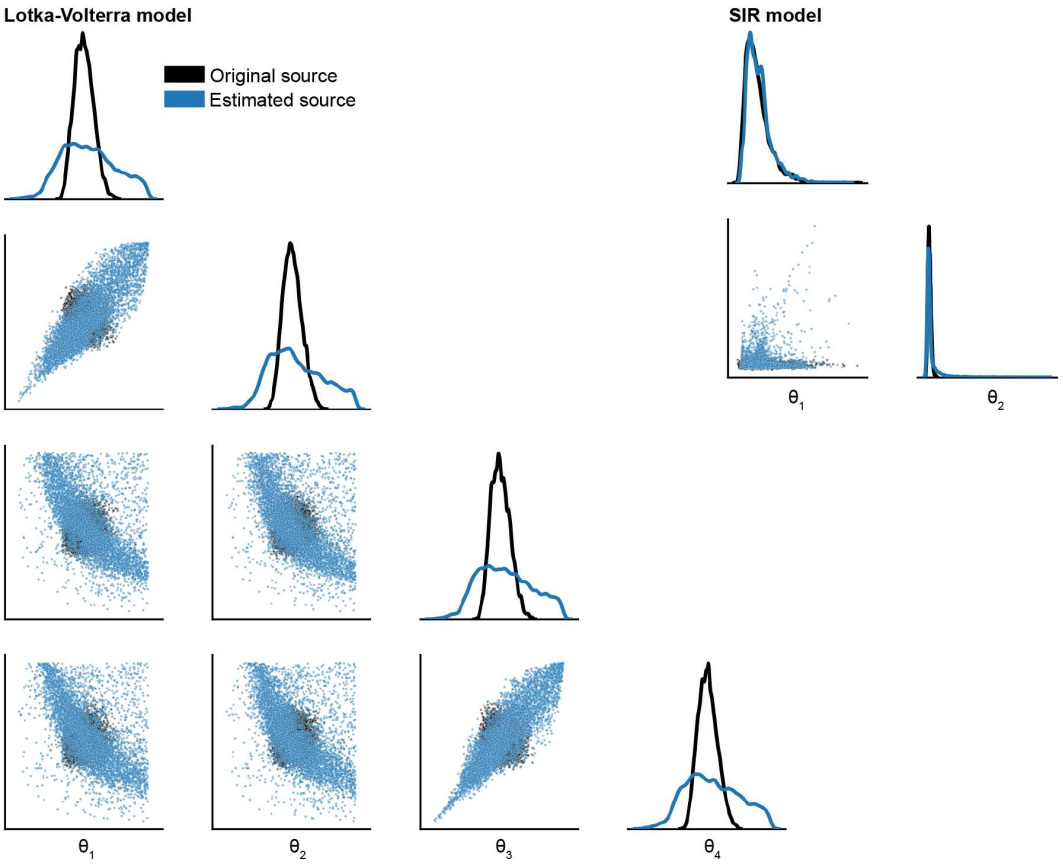

Figure A8: Original and estimated source distributions for the SIR and Lotka-Volterra model. For the Lotka-Volterra model, the estimated source has higher entropy than the original source.

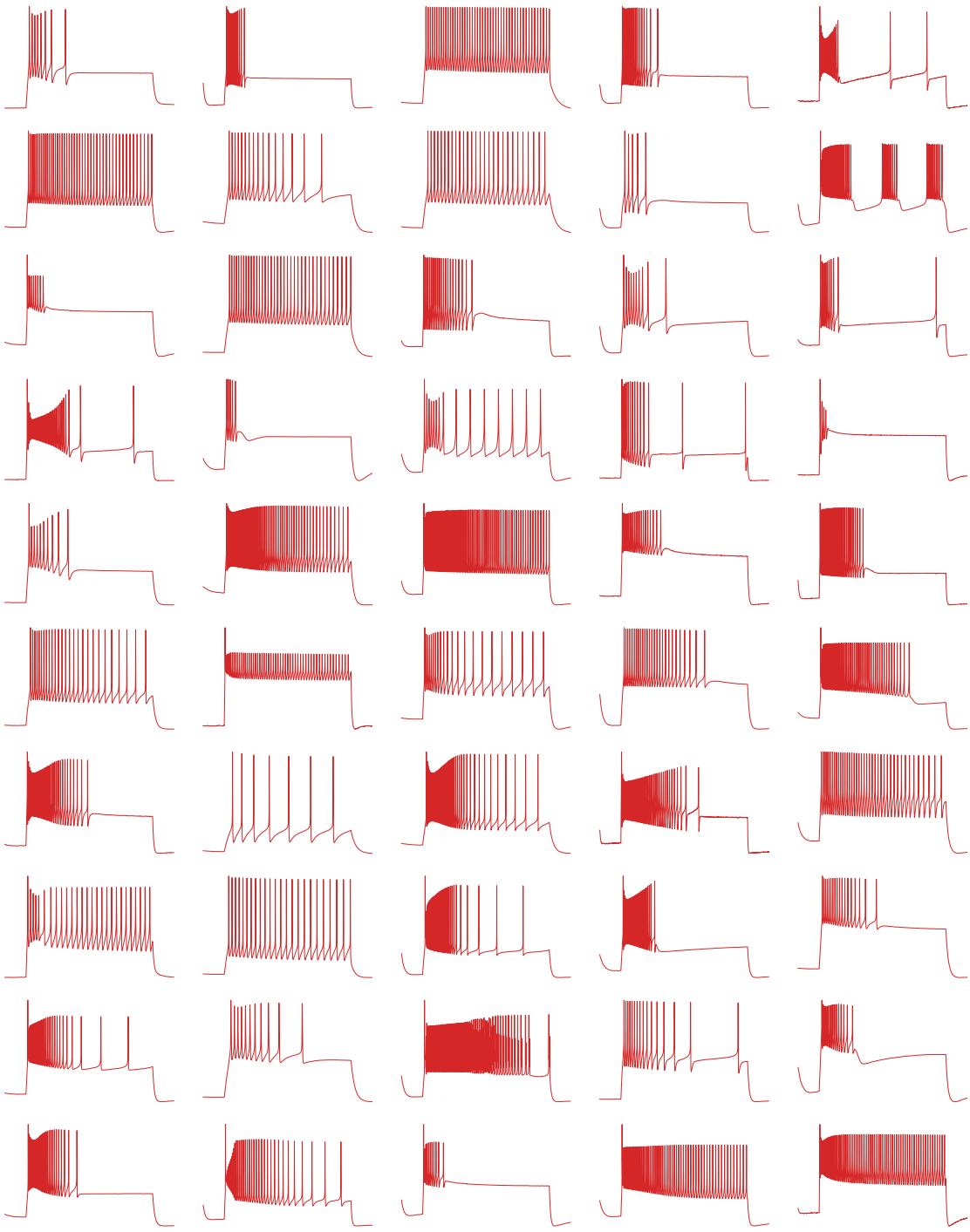

Figure A9: 50 random example traces produced by sampling from the estimated source and simulating with the Hodgkin-Huxley model.

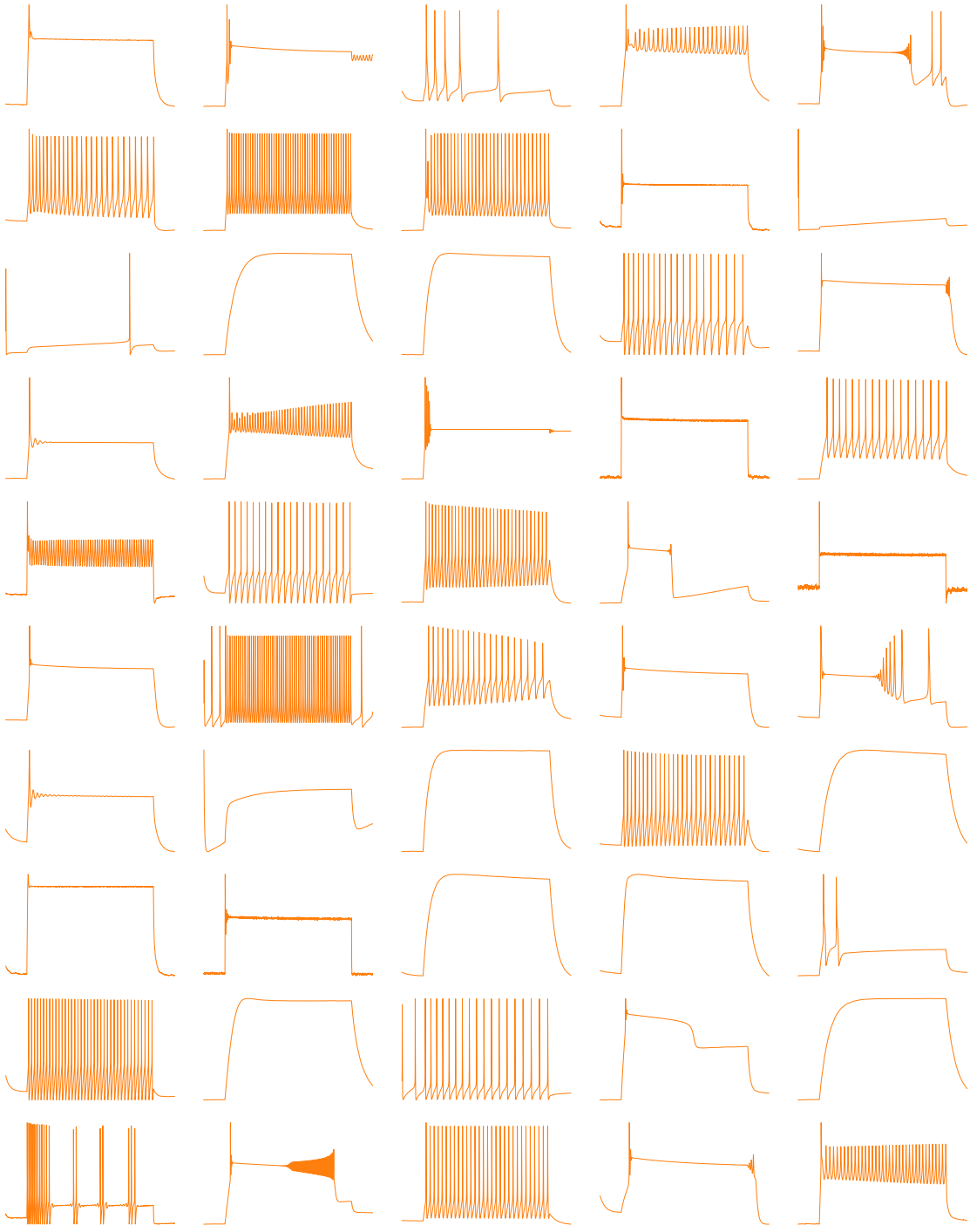

Figure A10: 50 random example traces produced by sampling from the uniform distribution over the box domain and simulating with the Hodgkin-Huxley model.

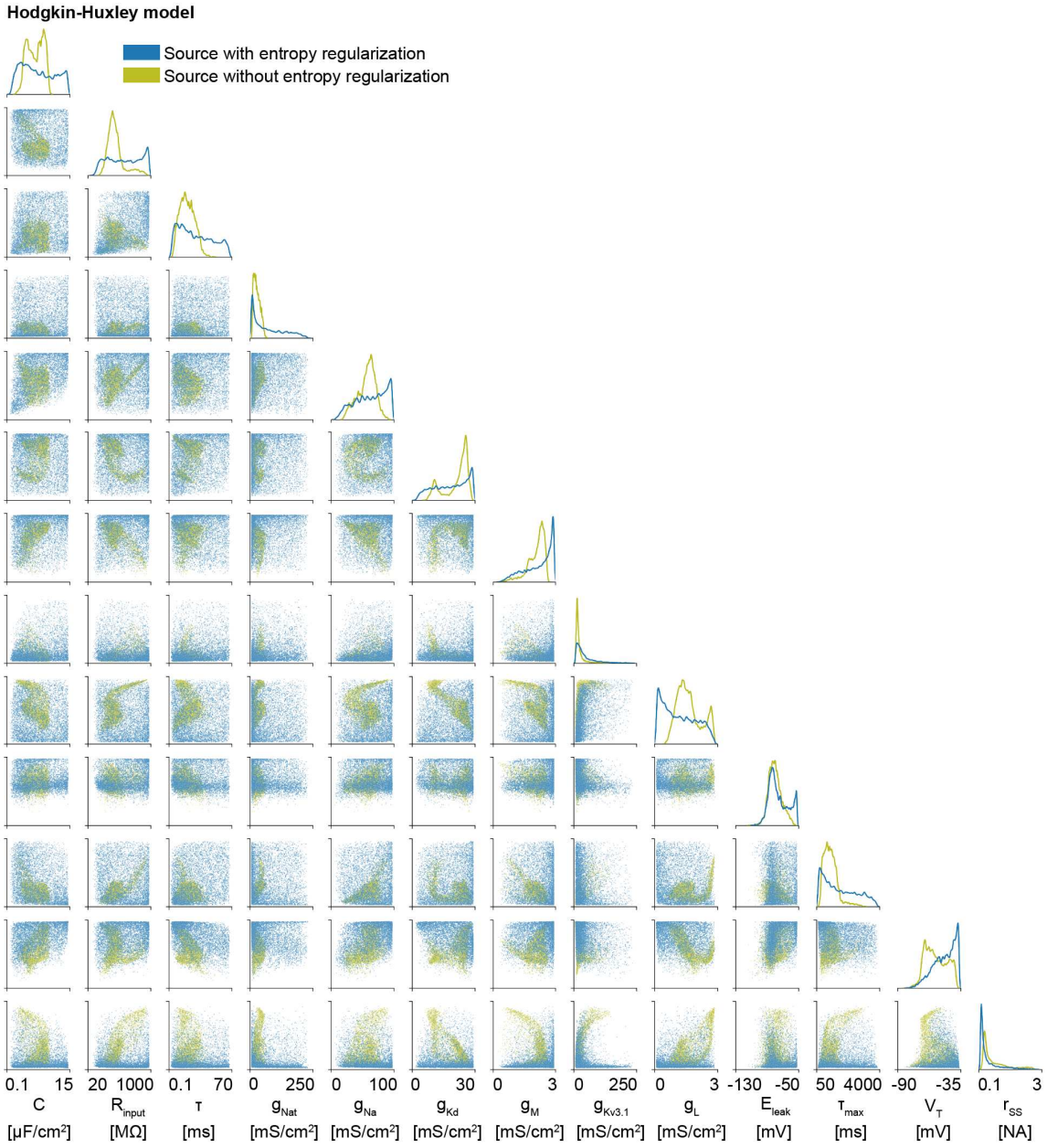

Figure A11: Estimated sources using for Hodgkin-Huxley task with the entropy regularization ($\lambda = 0.25$) and without the entropy regularization. Without, many viable parameter settings are missed, which would have significant downstream effects if the learned source distribution is used as a prior distribution for inference tasks.

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 demonstrate in Table 1 our claim that we achieve source distributions with higher entropy than a state-of-the-art comparison, and show results in Fig. 4 and Fig. 5 that our method recovers source distributions on high dimensional tasks and the electrophysiological data, respectively.

   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 clearly mark the limitations discussion in Sec. 5.

   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: Proposition 2.1 is stated with a full set of assumptions and a complete proof in Appendix A.7.

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 pseudocode of our method in Algorithm 1. We provide full details of the architecture of the source model and surrogates in Appendices A.4 and A.5, respectively.

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 use public data from existing work which we reference for the electrophysiological dataset. The code necessary to reproduce our results is available at https://github.com/mackelab/sourcerer.

   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 provide full details on training the source model in Appendix A.3, A.4, A.5, A.6 and A.9.

   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: The numerical results in Table 1 are reported with estimated standard deviations, and the figures include error bars showing the standard deviation over an independent set of runs with different random seeds.

   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 specify the computational resources used in our numerical experiments in Appendix A.10. We provide a breakdown of the approximate computation time for each of the experiments performed in this work.

   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 conform with all aspects of the NeurIPS Code of Ethics.

   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 is fundamental in that we develop a new approach to solving the source distribution estimation problem. We do not develop new classes of models, nor do we apply our approach to problems with societal implications. We do not foresee any direct or indirect misuse of this work.

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: This work does not involve models that have a high 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 use a dataset of electrophysiological recordings from Scala et al. [52], which we cite in the main text.

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: The public repository contains the code to reproduce our results, along with necessary documentation. It is licensed under the MIT license.

    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 involve crowdsourcing nor research with human subjects.

    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: This work does not involve crowdsourcing nor research with 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.