Journal of Glaciology

Article

Cite this article: Riel B, Minchew B (2023). Variational inference of ice shelf rheology with physics-informed machine learning. *Journal of Glaciology* 1–20. https://doi.org/10.1017/jog.2023.8

Received: 6 September 2022
Revised: 22 January 2023
Accepted: 30 January 2023

Keywords:
Glacial rheology; glacier modeling; ice dynamics; ice rheology; ice shelves

Author for correspondence:
Bryan Riel, E-mail: briel@zju.edu.cn

Variational inference of ice shelf rheology with physics-informed machine learning

Bryan Riel[1,2] and Brent Minchew[2]

[1]School of Earth Sciences, Zhejiang University, 310027 Hangzhou, China and [2]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

Abstract

Floating ice shelves that fringe the coast of Antarctica resist the flow of grounded ice into the ocean. One of the key factors governing the amount of flow resistance an ice shelf provides is the rigidity (related to viscosity) of the ice that constitutes it. Ice rigidity is highly heterogeneous and must be calibrated from spatially continuous surface observations assimilated into an ice-flow model. Realistic uncertainties in calibrated rigidity values are needed to quantify uncertainties in ice sheet and sea-level forecasts. Here, we present a physics-informed machine learning framework for inferring the full probability distribution of rigidity values for a given ice shelf, conditioned on ice surface velocity and thickness fields derived from remote-sensing data. We employ variational inference to jointly train neural networks and a variational Gaussian Process to reconstruct surface observations, rigidity values and uncertainties. Applying the framework to synthetic and large ice shelves in Antarctica demonstrates that rigidity is well-constrained where ice deformation is measurable within the noise level of the observations. Further reduction in uncertainties can be achieved by complementing variational inference with conventional inversion methods. Our results demonstrate a path forward for continuously updated calibrations of ice flow parameters from remote-sensing observations.

1. Introduction

Viscous flow of ice in glaciers and ice sheets is governed by gravitational driving forces and resisting tractions at ice-rock boundaries, as well as internal stresses resulting from stretching and compression. For laterally confined ice shelves that flow within embayments, flow is resisted by shear stresses at the margins where faster-flowing ice is in contact with rock or immobile ice. Basal shear stresses can further resist flow where ice is locally grounded at ice rises or pinning points. The total resistance, or buttressing, provided by ice shelves to upstream grounded ice is a key modulator for potential changes in flow speed of the grounded ice to changes in atmospheric or oceanic conditions. However, accurate quantification of buttressing stresses and modeling of ice shelf flow depends on well-calibrated estimates of ice rheological parameters throughout the modeling domain.

Observations of ice flow, whether in an experimental or natural setting, are the only means by which we can infer mechanical properties such as ice rheology. Specifically, spatially continuous flow velocity measurements permit robust strain rate estimates, which can have considerable spatial variability due to differing flow regimes, ice rheology, ice geometry and other factors. These strain rates are linked to stresses within the ice through an appropriate constitutive relation, where the most commonly used relation is Glen's Flow Law:

$$\begin{aligned}\tau_{ij} &= 2\eta\dot{\epsilon}_{ij} \\ &= B\dot{\epsilon}_e^{\frac{1-n}{n}}\dot{\epsilon}_{ij},\end{aligned} \tag{1}$$

where τ_{ij} is the deviatoric stress tensor, η is the effective dynamic viscosity, B is the ice rigidity, n is the stress exponent, $\dot{\epsilon}_{ij}$ is the strain rate tensor and $\dot{\epsilon}_e = \sqrt{\dot{\epsilon}_{ij}\dot{\epsilon}_{ij}/2}$ (where we apply the summation convention for repeated indices) is the effective strain rate computed as the square root of the second invariant of the strain rate tensor (Glen, 1958). Note that a prefactor defined as $A = B^{-n}$ is also commonly used in Glen's Flow Law. All of the terms in Eqn (1) vary spatially with different intrinsic lengthscales. The stress exponent, n, is set by the dominant mechanisms of creep that drive the deformation of ice and is dependent on the stress regime, grain size, ice temperature and crystallographic fabric (Goldsby and Kohlstedt, 2001). The prefactor, B, which we refer to as the ice rigidity, shares the same dependencies as the exponent, in addition to interstitial water content, impurities and damage (Cuffey and Paterson, 2010). Thus, both B and n are lumped parameters in Glen's Flow Law that represent a combination of factors and mechanisms which generally cannot be observed continuously at the scale of ice shelves and ice sheets. Rather, B and n must be inferred from ice surface velocity and thickness observations for each area of interest.

To construct a tractable inverse problem, Glen's Flow Law is first injected into an appropriate dynamical framework (i.e. governing equations for ice flow) to obtain a non-linear mapping from parameters (B and n) to observables (ice velocity) over the entire modeling domain. This mapping, or forward problem, can be used in an optimization framework to

https://doi.org/10.1017/jog.2023.8 Published online by Cambridge University Press

estimate parametric values that optimally reconstruct the surface observations (MacAyeal, 1989, 1993). The outcome of the inverse problem, a 2D map of B and n, can then be used in Glen's Flow Law to compute stresses within the ice, which allows for further prognostic simulations to project the evolution of ice flow for a given study area in response to changing climatic conditions. However, for static datasets, i.e. snapshots of velocity and thickness at a given time epoch, B and n cannot be uniquely determined, and independent constraints on one of the parameters is required to reduce the non-uniqueness. In this work, we focus only on inference of a spatially varying rigidity B, noting that recent work has demonstrated that n may be estimated in Greenland and Antarctica independently under certain flow conditions (e.g. Bons and others, 2018; Millstein and others, 2022), leading to a value of $n \approx 4$ which is consistent with experimental analysis of ice deformation under realistic pressure environments and strain rates (Qi and Goldsby, 2021).

Still, estimation of the optimal rigidity field is equivalent to drawing only a single sample of B from the total statistical *distribution* of fields that could explain the observations nearly as well as the optimal one. This distribution is influenced by observational uncertainties as well as modeling uncertainties. For the latter, modeling uncertainties can stem from factors such as model resolution (sensitivity of the forward model to variations in parameter values) and model misspecification where the model fails to capture relevant physics or makes improper assumptions about certain aspects of the physics. Overall, quantification of the distribution of parameter values is of equal importance to estimating the optimal values, and it is ultimately necessary for obtaining a realistic distribution of future ice states conditioned on current-day observations (Aschwanden and others, 2021).

In this work, we aim to develop a framework for estimating the distribution of ice rigidity for large study areas that combines information extracted from relevant surface observations with information obtained from prior theories, experimental/observational studies, etc. While such a framework has a long history in Bayesian inference, our primary consideration in this work is a matter of scalability to large datasets as well as to a large number of effective model parameters. To that end, we build upon recent developments in variational inference and physics-informed machine learning to address the problem of scalability. We use a combination of neural networks for modeling continuous surface observations with variational Gaussian Processes for modeling ice rigidity probability distributions. The mapping between surface observations and rigidity is provided by partial differential equations (PDEs) describing ice flow, which ultimately allow us to include a physics-informed loss function to the training objective for the machine learning models. Both classes of models allow for training with stochastic gradient descent, which is critical for scaling the inference method to large datasets. We target select ice shelves in Antarctica for demonstrating the proposed methods as they provide a number of favorable modeling simplifications while maintaining adequate complexity and large spatial extents suitable for examining the advantages and disadvantages of the proposed methods.

2. Methodology

In this section, we will introduce the governing equations for ice flow that link spatial variations in our parameter of interest, ice rigidity, to observations of ice shelf velocity and thickness. We then introduce a probabilistic physics-informed machine learning framework designed to produce a probability distribution of rigidity fields consistent with the surface observations. We discuss how we utilize variational Bayesian techniques to perform inference at the scale of large ice shelves, observed with large datasets.

2.1. Ice flow momentum balance forward model

Given a spatial domain with spatial coordinates specified by $\mathbf{x}$, where for two dimensions $\mathbf{x} = [x,\ y]$, our goal is to estimate the most likely spatial field of ice rigidity, $B = B(\mathbf{x})$, conditional on observations of the flow of ice shelves and their geometry. To that end, we utilize a momentum balance method to estimate B that computes resistive stresses that optimally balance gravitational driving stresses. Within ice shelves, resistive (vertical) shear stresses at the base are negligible due to contact with relatively inviscid seawater. Ice is a thin film with small thicknesses relative to the horizontal dimensions (aerial extent). Thus, we are justified in employing the widely used shallow-shelf approximation (SSA), which assumes negligible vertical shearing in a thin film and vertically integrates viscosity and stresses in the ice column to obtain a simplified 2D framework for the governing equations of flow in ice shelves. We write the SSA momentum balance as:

$$r_x = \frac{\partial}{\partial x}\left(2\eta h\left(2\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right)\right) + \frac{\partial}{\partial y}\left(\eta h\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right)\right) - \rho_{\mathrm{i}} g h \frac{\partial s}{\partial x}, \tag{2}$$

$$r_y = \frac{\partial}{\partial y}\left(2\eta h\left(2\frac{\partial v}{\partial y} + \frac{\partial u}{\partial x}\right)\right) + \frac{\partial}{\partial x}\left(\eta h\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right)\right) - \rho_{\mathrm{i}} g h \frac{\partial s}{\partial y}, \tag{3}$$

where u and v are the horizontal velocity components of the velocity vector, $\boldsymbol{u}$, along the x- and y-directions, respectively, and taken to be constant with depth; h is the ice thickness; s is the ice surface elevation; $2\eta = B\dot{\epsilon}_{\mathrm{e}}^{1-n/n}$ is the effective dynamic viscosity of ice (Eqn 1); ρ_{i} is the mass density of ice (assumed to be constant at $\rho_{\mathrm{i}} = 917\ \mathrm{kg/m^3}$); and g is the gravitational acceleration. In the above formulation, r_x and r_y on the left-hand side represent *residual* terms in the momentum balance. For the general SSA as applied to flowing ice where ice is grounded, such as at ice streams, these residual terms are non-zero and correspond to basal drag. However, since drag at the base of ice shelves is assumed to be negligible because of the negligible viscosity of seawater, r_x and r_y are nominally zero. Thus, we seek to construct the field $B(\mathbf{x})$ that provides an optimal balance between the membrane stresses (first two terms on the right-hand side) and the driving stress (last term on the right-hand side) such that r_x and r_y are close to zero.

2.2. Reparameterization for ice rigidity

Before proceeding, we first introduce a commonly used reparameterization of the rigidity B:

$$B(\mathbf{x}) = B_0(\mathbf{x})\,\mathrm{e}^{\theta(\mathbf{x})}, \tag{4}$$

where B_0 ($\mathbf{x}$) represents some reference field of the rigidity and θ ($\mathbf{x}$) corresponds to logarithmic rigidity variations about B_0 ($\mathbf{x}$). Since B is a strictly positive quantity, this reparameterization allows for the transformation of an inequality-constrained inference problem on B to an unconstrained inference problem on θ while also compressing the dynamic range of rigidity variations to log-space (Shapero and others, 2021; Brinkerhoff, 2022). Analysis of inferred variations in B can thus be performed by inserting any inferred variations in θ into the above equation. Proper choice of the reference rigidity B_0 depends on the prior information available, which we discuss in detail in later sections.

https://doi.org/10.1017/jog.2023.8 Published online by Cambridge University Press

2.3. Probabilistic inference of ice rigidity from remote-sensing data

Observations of horizontal ice velocity over ice-sheet margins have been widely available for the past decade thanks to the prevalence of remote-sensing platforms and efficient data processing methodologies (Joughin and others, 2010; Mouginot and others, 2017; Gardner and others, 2019). At the same time, improved integration of ice-penetrating radar and surface velocities using mass conservation techniques have allowed for more accurate and higher resolution maps of ice thickness and bathymetry (Morlighem and others, 2017). Specifically, over ice shelves, it is common practice to convert observations of surface elevation, which are well constrained, to ice thickness by assuming hydrostatic equilibrium and applying corrections for firn layers derived from in situ thickness data (Morlighem and others, 2020). Thus, we can estimate spatial gradients and compute the SSA momentum balance directly for a given rigidity field when velocity and thickness observations are spatially continuous over an ice shelf. However, observation noise and data gaps generally degrade estimates of observation gradients, resulting in non-physical estimates of SSA terms that require an additional gradient operation. Therefore, we seek to formulate a method for approximating the continuous functions underlying the surface velocity, ice thickness and ice rigidity fields that balance the reconstruction accuracy of the observed data while resulting in minimal SSA residuals, **r**. Furthermore, such a method should allow for rigorous quantification of the uncertainties associated with θ, which will be driven by observation noise, varying sensitivities of the SSA equations to spatial variables, and prior knowledge on the range of values for θ.

To that end, we utilize Bayes' Theorem to construct the joint posterior probability distribution for rigidity and the reconstructed surface velocity and ice thickness, conditioned on a set of velocity and thickness observations assimilated into the SSA momentum balance. Let us first define the observation vector $\mathbf{d} = [u,\ v,\ h]$, reconstructed observation vector $\hat{\mathbf{d}} = [\hat{u},\ \hat{v},\ \hat{h}]$, and SSA residual vector $\mathbf{r} = [r_x,\ r_y]$. In the following, we consider that all three vectors vary spatially over an ice shelf, i.e. $\mathbf{d} = \mathbf{d}(\mathbf{x})$, $\hat{\mathbf{d}} = \hat{\mathbf{d}}(\mathbf{x})$ and $\mathbf{r} = \mathbf{r}(\mathbf{x})$. We can then write the joint posterior distribution for θ and $\hat{\mathbf{d}}$ as (Appendix A):

$$p(\theta,\ \hat{\mathbf{d}}|\mathbf{d},\ \mathbf{r}) \ \propto\ p(\mathbf{d}|\hat{\mathbf{d}})p(\mathbf{r}|\theta,\ \hat{\mathbf{d}})p(\theta). \tag{5}$$

The first two terms on the right-hand side of the equation are the data likelihood distributions, where $p(\mathbf{d}|\hat{\mathbf{d}})$ measures the probability of observing **d** for a given set of predictions $\hat{\mathbf{d}}$ while $p(\mathbf{r}|\theta,\ \hat{\mathbf{d}})$ measures the probability of computing **r** (which is nominally [0, 0] for ice shelves) through evaluation of the SSA momentum balance using $\hat{\mathbf{d}}$ and θ. The last term on the right is the prior distribution $p(\theta)$, which encodes our prior knowledge on θ values without having seen any observations.

2.4. Variational inference of ice rigidity

Due to non-linearities in the SSA momentum balance and potentially non-Gaussian data likelihoods, the posterior for θ and $\hat{\mathbf{d}}$ must be approximated, either by drawing random samples from $p(\theta,\ \hat{\mathbf{d}}|\mathbf{d},\ \mathbf{r})$ or by constructing a suitable approximating distribution. The former strategy is based on the general class of Markov Chain Monte Carlo (MCMC) approaches and tends to be suitable for a low or moderate number of model dimensions. However, for spatial domains with sizes typical of Antarctic ice shelves, the number of model dimensions will be quite high, regardless of the model used to represent the spatial fields. Therefore, we instead employ a variational inference framework wherein we aim to construct an approximating distribution, $q(\theta,\ \hat{\mathbf{d}})$, for θ and $\hat{\mathbf{d}}$ that is minimally divergent from the true posterior $p(\theta,\ \hat{\mathbf{d}}|\mathbf{d},\ \mathbf{r})$. The choice of approximating distribution generally involves a trade-off between computational complexity (e.g. number of trainable parameters) and accuracy in capturing relevant statistics and correlations in the target posterior (Blei and others, 2017). We discuss the choice of approximating distribution in the next section.

At this stage, let us also enforce that $q(\theta,\ \hat{\mathbf{d}})$ is the product of two independent variational distributions, $q(\theta,\ \hat{\mathbf{d}}) = q(\theta)q(\hat{\mathbf{d}})$. This independence will allow us to use separate machine learning models to reconstruct observations and predict rigidity while using a joint objective function to tune their parameters simultaneously. Moreover, we restrict $q(\hat{\mathbf{d}})$ to produce only the mean reconstructed velocity and thickness, assuming that information about formal observation uncertainties can be obtained. Later, we will discuss how this assumption on $q(\hat{\mathbf{d}})$ will affect the posterior variance estimates for θ.

As is commonly done in variational inference, we utilize the Kullback–Leibler (KL) divergence for measuring the difference between two probability distributions:

$$\mathcal{KL}\big[q(\theta,\ \hat{\mathbf{d}})\|p(\theta,\ \hat{\mathbf{d}}|\mathbf{d},\ \mathbf{r})\big] = \int q(\theta,\ \hat{\mathbf{d}}) \log \frac{q(\theta,\ \hat{\mathbf{d}})}{p(\theta,\ \hat{\mathbf{d}}|\mathbf{d},\ \mathbf{r})}\, d\theta\, d\hat{\mathbf{d}}. \tag{6}$$

By minimizing the KL-divergence, we are tuning the variational distribution to be close to the target posterior distribution from an informational perspective. However, the existence of the posterior distribution in the denominator of the log term generally makes computing the KL-divergence intractable. As shown in Appendix A, we can instead maximize a stochastic variational lower bound, referred to as the Evidence Lower Bound (ELBO):

$$\begin{aligned}\mathrm{ELBO} = {} & E_{\theta\sim q(\theta)}\big[\log p(\mathbf{r}|\theta,\ \hat{\mathbf{d}})\big] + \log p(\mathbf{d}|\hat{\mathbf{d}}) \\ & - \mathcal{KL}\big[q(\theta)\|p(\theta)\big].\end{aligned} \tag{7}$$

The first term for the ELBO is the expected value of the log-likelihood for the SSA residual vector, **r**, for a given set of predictions $\hat{\mathbf{d}}$ and integrated over the variational distribution for θ (see Appendix A for estimation of this integral over ice shelves). The second t erm for the ELBO is the data log-likelihood for the observations **d** given predictions $\hat{\mathbf{d}}$. The last term is the KL-divergence between the variational distribution $q(\theta)$ and prior $p(\theta)$ for the rigidity, which can be computed analytically for certain pairs of distributions, e.g. two multivariate normal distributions. Overall, the ELBO provides an objective function for learning an approximating distribution for the spatial field of rigidity, as well as the mean ice velocity and thickness fields. In the next section, we discuss the choice of approximating distribution $q(\theta)$ and the parameterization of spatial fields such that the ELBO can be computed efficiently for large datasets and spatial domains.

2.5. Parameterization of spatial fields and physics-informed machine learning

We seek representations of the mean spatial fields $\hat{\mathbf{d}}(\mathbf{x})$ and an approximating distribution $q(\theta)$ that permit efficient evaluation of the ELBO in Eqn (7). For ice-sheet modeling in general, spatial fields are typically discretized into an irregular mesh of triangular finite elements. However, this approach can lead to a large number of nodal parameters for a given spatial field, particularly for fields with high spatial variability with length scales on the order of a few ice thicknesses. Large numbers of nodes can

https://doi.org/10.1017/jog.2023.8 Published online by Cambridge University Press

severely restrict the choice of approximating distributions that are computationally feasible. For example, for a mesh with N nodes, an approximating multivariate normal distribution with a mean vector of N elements and a covariance matrix of N^2 elements would require excessive computer memory when N is greater than a few thousand nodes.

Our strategy in this work is to adopt mesh-free parameterizations of spatial fields using two different machine learning models: (1) a neural network for the mean fields $\hat{\mathbf{d}}(\mathbf{x})$; and (2) a variational Gaussian Process (VGP) for the approximating distribution for the rigidity field $\theta(\mathbf{x})$. Both models fall under the broad classification of function approximators for a given set of data, while the VGP (and Gaussian Processes in general) also approximates the probability distribution of functions that generate the given dataset. For the mean fields $\hat{\mathbf{d}}(\mathbf{x})$, a dense, feedforward neural network, $f_{\boldsymbol{\psi}}$, is used to represent the surface observations on a point-by-point basis:

$$\hat{\mathbf{d}}_i = f_{\boldsymbol{\psi}}(\mathbf{x}_i), \tag{8}$$

where $\hat{\mathbf{d}}_i$ is the vector of neural network predictions at the i-th coordinate $\mathbf{x}_i$, and $\boldsymbol{\psi}$ represents the total set of weights and biases of the hidden layers. The choice of feedforward networks in this work is motivated by the ability to generate predictions of $\hat{\mathbf{d}}_i$ at arbitrary coordinates, which facilitates mini-batch training by stochastic gradient descent. For the ice shelf models investigated here, we found that four-layer feedforward networks with 50–100 nodes per layer provide a suitable balance between expressivity and computational efficiency. Additionally, we use tanh activation functions as they are continuously differentiable and result in spatial gradients that are spatially smooth (Riel and others, 2021).

For the rigidity variational distribution, we exploit the ability of Gaussian Processes (GPs) to construct an approximating multivariate normal distribution through computation of a mean vector for any set of coordinates and a covariance matrix for any pairwise set of coordinates (Rasmussen, 2003). Multivariate normal distributions permit modeling of correlations between variables and have been shown to agree well with MCMC-sampled posteriors for ice dynamics inverse problems (Petra and others, 2014). The mean and covariance both require the evaluation of a kernel function that describes the expected similarity between data points for a given coordinate pair. Here, we use a squared exponential kernel:

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{(\mathbf{x}-\mathbf{x}')^T(\mathbf{x}-\mathbf{x}')}{2L^2}\right), \tag{9}$$

where σ^2 is the kernel variance and L is a covariance length scale that describes the characteristic distance up to which function values at $\mathbf{x}$ and $\mathbf{x}'$ are correlated. However, as described in Appendix B, the use of standard GPs to infer an approximating distribution for a given dataset will be computationally prohibitive for large datasets, which stems from the need to evaluate the kernel at N coordinates in order to construct and invert a covariance matrix of size $N \times N$. For our study areas, N is usually on the order of tens of thousands of points. Variational GPs are a modification of standard GPs that utilize the concept of sparse *inducing index points* for overcoming the prohibitive $\mathcal{O}(N^2)$ memory and $\mathcal{O}(N^3)$ time requirements of large sets of index points for standard GPs (Titsias, 2009). Rather than evaluating the kernel at N coordinates, we evaluate the kernel at M inducing index coordinates, $\mathbf{z}$, where $M \ll N$. The inducing index points are sometimes referred to as support points and act as a low-dimensional latent representation of the data (Quinonero-Candela and Rasmussen, 2005). As explained in Appendix B, computation of a posterior mean and covariance matrix at an arbitrary number of prediction points involves evaluation of the kernel at the prediction and inducing index points and combining the kernel evaluations with a full-rank covariance matrix at the inducing points. The number of inducing points, M, is a prescribed parameter that provides additional control on the complexity of the predicted variable field. For the ice shelves studied here, we find that M between 300 and 1000 provides sufficient expressiveness for the rigidity fields. Overall, the total set of trainable variables, $\boldsymbol{\phi}$, for the VGP consists of the inducing point locations $\mathbf{z}$, the inducing point mean values, the covariance matrix values at the inducing points and the hyperparameters of the kernel (σ^2 and L). In the following, we denote the approximating multivariate normal distribution as $q_{\boldsymbol{\phi}}(\theta)$.

While GPs and VGPs are usually trained in a supervised fashion for a training dataset of N examples, we do not have a set of ground truth rigidity values for θ. We instead insert $q_{\boldsymbol{\phi}}(\theta)$ into the ELBO in Eqn (7) in order to train $f_{\boldsymbol{\psi}}$ and $q_{\boldsymbol{\phi}}(\theta)$ simultaneously. This style of training is consistent with recent developments in physics-informed machine learning where available physics knowledge (in our case, the SSA momentum balance) is used to augment observables with pseudo-observables, which can be useful in domains where direct observations are impossible or difficult to obtain (Raissi and others, 2019; Riel and others, 2021). Here, ice surface velocity and thickness are observable, and their reconstructed values (generated by $f_{\boldsymbol{\psi}}$) are used in conjunction with rigidity samples drawn from $q_{\boldsymbol{\phi}}(\theta)$ in order to compute the SSA residual pseudo-observables, which should nominally be zero. We evaluate the SSA residuals at a random set of C 'collocation' coordinates, $\mathbf{x}_c$, not included in the training data for $f_{\boldsymbol{\psi}}$ in order to maximize the information extracted from the pseudo-observables (Raissi and others, 2019).

The ELBO in Eqn (7) can now be written in terms of the neural network parameters $\boldsymbol{\psi}$ and the VGP parameters $\boldsymbol{\phi}$:

$$\begin{aligned}\mathrm{ELBO}(\boldsymbol{\psi}, \boldsymbol{\varphi}) = {} & E_{\theta_c \sim q_{\varphi}(\theta_c)}\left[\log p(\mathbf{r}|\theta_c, f_{\psi}(\mathbf{x}_c))\right] \\ & + \log p(\mathbf{d}|f_{\psi}(\mathbf{x})) - \mathcal{KL}\left[q_{\varphi}(\theta_z)\|p(\theta_z)\right],\end{aligned} \tag{10}$$

where we use the shorthand notation $\theta(\mathbf{x}_c) = \theta_c$ and $\theta(\mathbf{z}) = \theta_z$ to indicate the rigidity field evaluated at the collocation points $\mathbf{x}_c$ and the inducing index points $\mathbf{z}$, respectively. The neural network $f_{\boldsymbol{\psi}}$ is evaluated at the surface observation coordinates $\mathbf{x}$. In this work, we minimize the negative of the ELBO using the Adam optimizer (Kingma and Ba, 2014) with a learning rate between $1\mathrm{e}^{-4}$ and $5\mathrm{e}^{-4}$ and batch sizes varying from 256 to 1024 examples. Additionally, we often found it beneficial to pre-train the network $f_{\boldsymbol{\psi}}$ using only the middle term of the ELBO (i.e. fitting the observations without the rigidity field). The pre-training results in a $f_{\boldsymbol{\psi}}$ that can then be fine-tuned with the full ELBO in order to improve overall training convergence with the VGP. Typically, we make a qualitative assessment of convergence by monitoring the loss values for training and testing datasets (using a train-test split of 85 and 15% of the data, respectively) and stop training when reductions in the loss value are negligible (Fig. S1). A summary of the neural network and variational inference training framework is described in Figure 1.

2.6. Selection of data likelihood and prior distributions

To concretize the ELBO training objective for the neural network $f_{\boldsymbol{\psi}}$ and VGP $q_{\boldsymbol{\phi}}(\theta)$, we now discuss specification of the data likelihoods for the surface observations and SSA residuals and the prior distribution for the rigidity.

https://doi.org/10.1017/jog.2023.8 Published online by Cambridge University Press

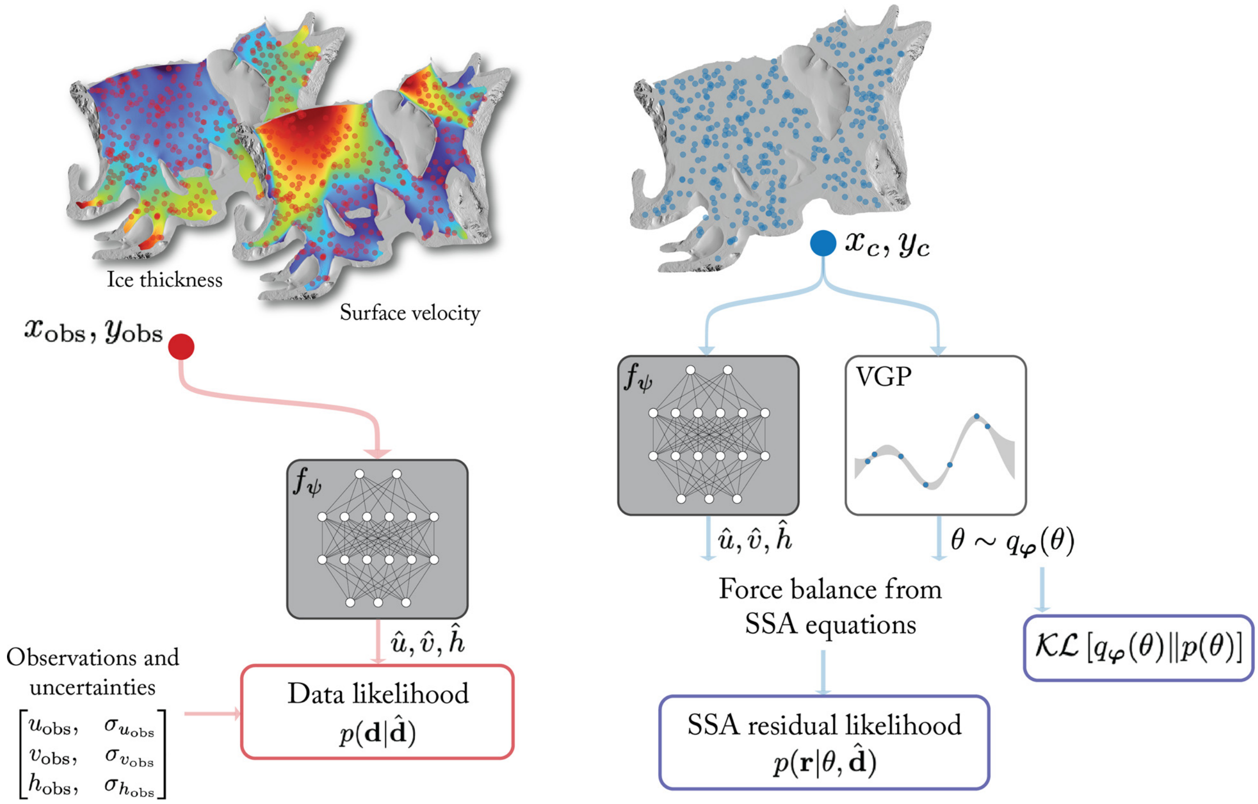
Ice thickness
Surface velocity
$x_{\text{obs}}, y_{\text{obs}}$
f_ψ
$\hat{u}, \hat{v}, \hat{h}$
Observations and uncertainties
$\begin{bmatrix} u_{\text{obs}}, & \sigma_{u_{\text{obs}}} \\ v_{\text{obs}}, & \sigma_{v_{\text{obs}}} \\ h_{\text{obs}}, & \sigma_{h_{\text{obs}}} \end{bmatrix}$
Data likelihood $p(\mathbf{d}|\hat{\mathbf{d}})$
x_c, y_c
f_ψ
VGP
$\hat{u}, \hat{v}, \hat{h}$
$\theta \sim q_\varphi(\theta)$
Force balance from SSA equations
$\mathcal{KL}\left[q_\varphi(\theta) \| p(\theta)\right]$
SSA residual likelihood $p(\mathbf{r}|\theta, \hat{\mathbf{d}})$

Fig. 1. Illustration of physics-informed variational inference framework. The two main components of the framework are the neural network f_ψ tasked with reconstructing surface observations (grey box) and the variational Gaussian Process (VGP) tasked with generating the distribution $q(\theta)$ (white box), which is a variational approximation to the posterior distribution of the normalized ice rigidity parameter θ. The training loss function is the sum of: (1) a data likelihood loss measuring the consistency between the reconstructed (predicted) and observed observations; (2) the expected value of the SSA residual likelihood given predicted rigidity values and observations; and (3) the KL-divergence between the variational and prior distributions for rigidity. For the data likelihood, training data and corresponding uncertainties and spatial coordinates are sampled from remote-sensing observations (red dots). For the SSA residual likelihood, an independent set of collocation coordinates (x_c and y_c, blue dots) are sampled from the model domain, which are input to f_ψ and the VGP in order to evaluate the SSA momentum balance at those coordinates.

2.6.1. Data likelihood specification

For the likelihood $p(\mathbf{d}|\hat{\mathbf{d}})$, we use independent normal distributions for the velocity components and ice thickness:

$$\begin{aligned} p(\mathbf{d}|\hat{\mathbf{d}}) &= p(u|\hat{u})p(v|\hat{v})p(h|\hat{h}), \\ p(u|\hat{u}) &= \mathcal{N}(u; \hat{u}, \sigma_u^2), \\ p(v|\hat{v}) &= \mathcal{N}(v; \hat{v}, \sigma_v^2), \\ p(h|\hat{h}) &= \mathcal{N}(h; \hat{h}, \sigma_h^2), \end{aligned} \tag{11}$$

where the different σ^2 variables correspond to the variances of each observable. The observation variances may be prescribed or learned, and in this work, we prescribe the variances to be scaled values of formal observation uncertainties provided with the datasets. We found that a scaling factor between 10 and 30 prevented the neural network f_ψ from overfitting the velocity data in order to generate velocities that are more consistent with the SSA momentum balance (e.g. mitigating high spatial frequency velocity variations that are not well-modeled by the SSA approximation). We further note that we apply a limited amount of spatial smoothing to the velocity and thickness data (using a Gaussian filter with a window size of roughly half of the mean ice thickness of the shelf) in order to mitigate high-frequency observation noise and improve training convergence.

For the likelihood $p(\mathbf{r}|\theta, \hat{\mathbf{d}})$, we again use a normal distribution:

$$p(\mathbf{r}|\theta, \hat{\mathbf{d}}) = \mathcal{N}(\mathbf{r}(\theta, \hat{\mathbf{d}}); \mathbf{0}, \sigma_{\mathbf{r}}^2), \tag{12}$$

where the mean of zero encodes that we expect $\mathbf{r} = \mathbf{r}(\theta, \hat{\mathbf{d}})$ to be zero on ice shelves, and $\sigma_{\mathbf{r}}^2$ is the expected variance of the residuals. Prescribing a proper value of $\sigma_{\mathbf{r}}^2$ for the likelihood distribution involves a careful consideration of the observation uncertainties and the expected uncertainties in θ. In the general case, one could perform a Monte Carlo estimate of $\sigma_{\mathbf{r}}^2$ by inputting random realizations of the velocity data and samples of θ from the prior into the SSA equations to get an expected range of values for $\mathbf{r}$. Using this approach, we find that $\sigma_{\mathbf{r}} \approx 1$ kPa works well for most cases and can be decreased for lower observation noise. Note that this approach does not explicitly consider situations where the SSA (with zero basal drag for ice shelves) is expected to perform poorly, such as pinning points where ice shelves become locally grounded over bathymetric highs. These *epistemic* uncertainties should correspond to an increase in $\sigma_{\mathbf{r}}^2$. An additional improvement to the methods presented here is to use a full variational distribution for $q(\hat{\mathbf{d}})$, rather than using only the mean values $\hat{\mathbf{d}}$. Our restriction to the mean values for $\hat{\mathbf{d}}$ likely reduces the estimated posterior variance for θ, which can be partially compensated by inflating $\sigma_{\mathbf{r}}^2$. Overall, the likelihood formulation can be improved by modeling the full probability distribution for the surface observations, explicit handling of epistemic uncertainties, incorporation of spatially correlated uncertainties (e.g. Brinkerhoff, 2022) and using a non-Gaussian probability distribution. We leave these improvements for future exploration.

2.6.2. Prior distribution

As discussed earlier, the VGP uses a squared exponential kernel function for posterior inference of rigidity (Eqn 9). Likewise, we use the squared exponential kernel to construct the GP prior $p(\theta)$ in order to encourage spatial coherence between predictions of θ at different coordinates:

$$k_{\text{prior}}(\mathbf{x}, \mathbf{x}') = \sigma_\theta^2 \exp\left(-\frac{(\mathbf{x}-\mathbf{x}')^T(\mathbf{x}-\mathbf{x}')}{2L_\theta^2}\right), \tag{13}$$

where the prior variance and length scale, σ_θ^2 and L_θ, are prescribed and not tuned during minimization of the negative ELBO. For ice dynamics well-described by the SSA approximation, L_θ can usually be chosen to be some finite multiple of the mean ice shelf thickness, which is commensurate with the longitudinal stress coupling length scale (Cuffey and Paterson, 2010). The complete prior distribution is written as:

$$p(\theta(\mathbf{x})) = \mathcal{N}(\mathbf{0}, k_{\text{prior}}(\mathbf{x}, \mathbf{x})). \tag{14}$$

Since we enforce a prior mean of zero, the prior variance σ_θ^2 will depend on the value of B_0 used for the reparameterization of rigidity. In this work, we investigate two different strategies: (i) assume that B_0 is uniform; or (ii) estimating a spatial field $B_0(\mathbf{x})$ independently through an inversion using traditional control methods. For the second strategy, recall that traditional control methods generally use an optimization objective based on the misfit between observed and predicted ice velocities, which is different from the momentum balance optimization objective used here. Therefore, the velocity misfit-based objective will implicitly have different spatially varying sensitives to the parameter field than the momentum balance objective, which provides an opportunity to combine the two objectives in a complementary manner. For the two different strategies for selecting B_0, we also correspondingly adjust the prior variance. For a uniform B_0, we set $\sigma_\theta^2 = 1$ to allow for a relatively large variation in θ over the ice shelf. For a B_0 obtained from a control method inversion, we reduce σ_θ^2 to 0.2, which encodes our belief that the values of B from the inversion are relatively well-constrained, and θ thus represents smaller deviations of B dictated by the momentum balance optimization objective.

2.7. Generating shelf-wide samples of the ice rigidity

While the VGP utilizes inducing index points to allow for per-batch prediction of θ and the corresponding posterior covariance matrices, we require an algorithm for generating a random sample of θ with a given spatial resolution over the entire modeling domain. Even with the use of inducing points, assembling global posterior mean and covariance matrices for the entire domain would require a very large amount of memory for uniform grids with sizes exceeding tens of thousands of grid points ($\mathcal{O}(NM)$ memory requirements for N grid points and M inducing points). We therefore utilize an MCMC-based approach where a random sample of θ is generated at a limited set of coordinates within the ice shelf and used as a seed to grow a full chain over the entire shelf, which is equivalent to block-sampling approaches for sampling Gaussian Markov Random Fields (e.g. Rue, 2001). Note that this form of MCMC utilizes the posterior statistics learned from variational inference and does not involve evaluation of a physics-based forward model. Specifically, we apply Gibbs sampling on a block-by-block basis where a block is defined as a small subset of the uniform grid. For each block, the mean $\boldsymbol{\mu}_\theta$ and covariance matrix $\boldsymbol{\Sigma}_\theta$ are computed using the trained VGP. In this way, we can grow a random chain through successive evaluation of smaller multivariate normal distributions rather using a multivariate normal with a very large global covariance matrix. As an example, assume that the first two blocks are combined in the following partition:

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{bmatrix}, \quad \boldsymbol{\mu}_\theta = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \\ \boldsymbol{\Sigma}_\theta = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \tag{15}$$

Then, initialization of a Gibbs chain would use the following identities:

$$\boldsymbol{\mu}_\theta, \ \boldsymbol{\Sigma}_\theta \leftarrow q_\varphi(\theta(\mathbf{x})), \tag{16a}$$

$$\boldsymbol{\theta}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}), \tag{16b}$$

$$\boldsymbol{\theta}_2 \sim \mathcal{N}\big(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{\theta}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\big). \tag{16c}$$

In words, we use the trained VGP to generate a mean vector and covariance matrix for two blocks of coordinates, which are then partitioned according to Eqn (15). A sample $\theta_1 = \theta(\mathbf{x}_1)$ is drawn directly from the mean and covariance of the first block. Then, a sample $\theta_2 = \theta(\mathbf{x}_2)$ is drawn from a multivariate normal with a mean and covariance that are computed using the Schur complement. We then set $\boldsymbol{\mu}_1 \leftarrow \boldsymbol{\mu}_2$, $\boldsymbol{\theta}_1 \leftarrow \boldsymbol{\theta}_2$, $\boldsymbol{\Sigma}_{11} \leftarrow \boldsymbol{\Sigma}_{22}$, and use the VGP to predict the mean and covariance for the next set of blocks. The process is repeated for all blocks in the modeling domain. With this approach, each sample is independent of the previous sample and is accepted with a probability of 1. In our experiments, we found that if the block size was chosen to be too small, the variance of the final sample was artificially large, likely due to excessive truncation of the covariance matrix (depending on the resolution of the uniform grid relative to the length scale of the posterior variations in rigidity). Here, we found a block size of 1000 grid points works well for grids with cell sizes equal to roughly half or a quarter of the prior covariance lengthscale. Overall, we can validate the samples derived from the Gibbs sampler by viewing traceplots (plots of samples at independent coordinates) and cross-comparing the standard deviation fields between the sample standard deviation and the standard deviation predicted directly by the VGP (Fig. S6).

2.8. Related work

Bayesian inference has long been applied to geophysical inverse problems, and as computational resources and inference algorithms improve, the complexity and size of the physical models investigated has increased. Within glaciological inverse problems, Bayesian formulations of the posterior distributions have been used as cost functions for obtaining point estimates of basal topography and friction for grounded ice streams (Pralong and Gudmundsson, 2011). For fully Bayesian inference, Petra and others (2014) developed an MCMC method for estimating the posterior distribution for ice-sheet models with a large number of parameters, utilizing low-rank approximations of data likelihood Hessian matrices in order to reduce computational complexity while improving sample efficiency. Similarly, Gopalan and others (2021) used a Gibbs sampler in order to sample for ice stream model parameters for a simpler model applicable to slower-flowing ice. While MCMC methods generally serve as 'gold standards' for Bayesian inference, they do not scale well to large problem sizes. MCMC methods that invoke simpler proposal distributions usually require many more samples in order to sufficiently sample the posterior, whereas methods that can utilize the problem structure to improve

https://doi.org/10.1017/jog.2023.8 Published online by Cambridge University Press

sample efficiency require more computational resources (Petra and others, 2014).

Methods that approximate the posterior distribution, rather than sample from it, provide appealing alternatives to MCMC. Both Isaac and others (2015) and Babaniyi and others (2021) utilize a Gaussian approximation of the posterior centered on the *maximum a posteriori* (MAP) point (i.e. a Laplace approximation) in order to infer basal drag parameters for ice sheets. While Laplace approximations subvert the need for generating posterior samples (and the forward model evaluations associated with each sample), they can lead to posterior approximations that fail to capture much of the probability mass when the posterior is sufficiently non-Gaussian or multi-modal (Penny and others, 2007). In contrast, variational methods that utilize the KL-divergence as an optimization criterion (as done here) tend to favor approximating distributions that match the moments of the target distribution (e.g. mean and variance), which tends to capture more probability mass. Sufficient capturing of probability mass can be especially important for posterior predictive modeling where non-linearities can lead to a large spread of predictions (e.g. see the section on ice shelf buttressing).

To that end, Brinkerhoff (2022) introduced a variational inference method to jointly infer basal drag and ice rheology at a catchment-scale for glaciers. Importantly, the KL-divergence was used to estimate an optimal approximating distribution that also uses a Gaussian process prior, similar to the approximating distributions used in our work. A finite number of eigenvectors of the prior covariance are used to construct a linear model that permits inference at a lower dimension, similar to the sparse inducing points used in the VGP. The construction of the eigenvectors utilizes a coarse grid in the Fourier domain to model signals with a discrete set of spatial frequencies. Thus, the method of Brinkerhoff (2022) shares many of the same features proposed here, with two main differences. Firstly, we take the ice surface velocity and thickness as given (subject to smoothing) and use the momentum balance based on the SSA as our forward model in order to compute $\mathbf{r}$. On the other hand, all of the previous approaches strictly enforce $\mathbf{r} = 0$ (i.e. they take satisfaction of the SSA momentum balance as a constraint) and use predicted velocities as the forward model. The difference between the two approaches may be interpreted as taking two different pathways to the same solution, where the difference in forward models will lead to different sensitivities to the ice rheology parameters (see Discussion section). Furthermore, our parameterization of the spatial fields with neural networks and VGPs permits independent evaluation of the spatial gradients in the SSA momentum balance using automatic differentiation. This independence allows for batch-based stochastic gradient descent, which is advantageous for large datasets.

3. Application to simulated ice shelves

We now apply the physics-informed variational framework to simulated ice shelves in order to evaluate the recovery of ice rigidity under varying degrees of model complexity and uncertainty and data noise. Furthermore, the simulated ice shelves allow us to isolate which mechanical factors control the inferred rigidity uncertainties, which will aid in building intuition for application of the framework to natural settings.

3.1. 1D ice shelf

We first simulate a laterally confined ice shelf using 1D SSA equations where lateral drag is parameterized assuming a rectangular bed with width w in the across-flow direction (Nick and others, 2010). The momentum equation in the along-flow direction reduces to:

$$\frac{\partial}{\partial x}\left(4\eta h \frac{\partial u}{\partial x}\right) - \frac{2Bh}{w}\left(\frac{5u}{w}\right)^{1/n} = \rho_i g h \frac{\partial s}{\partial x}. \qquad (17)$$

As there are no time derivatives in the momentum balance for ice flow, the above equation is solved for u for a given thickness profile h. The thickness is then evolved using mass conservation:

$$\frac{\partial h}{\partial t} = a - \frac{\partial q}{\partial x}, \qquad (18)$$

where a is the accumulation rate (set to 0 for this example) and $q = uh$ is the horizontal ice flux through a vertical column of ice.

3.1.1. Simulation setup

We simulate an ice shelf with a width of 30 km and a length of 100 km, which is comparable to ice shelves of several ice streams in West Antarctica, such as Rutford Ice Stream. We prescribe a spatially varying B profile that is periodic in the along-flow direction while setting the flow law exponent to be uniform at $n = 3$. After simulating the shelf for 400 years to an approximate steady-state, we extract 200 random velocity and thickness values over the model domain to use as training data. We add spatially correlated noise by convolving a 1D field of independent Gaussian noise with a Gaussian kernel with a lengthscale of 5 km. We train a feedforward neural network with four hidden layers of 50 nodes each in order to reconstruct the velocity and thickness and a VGP with 15 inducing index points in order to predict the log rigidity θ. We use an a priori value of the prefactor $B_0 = 400\ \mathrm{yr}^{1/3}$ kPa. For the prior distribution for θ, we describe a prior standard deviation of $\sigma_\theta = 1$ and a correlation lengthscale of $L_\theta = 15$ km. For the data likelihood parameterizing the residual SSA terms, we use an independent normal distribution with a mean of zero and a standard deviation of $\sigma_r = 2.0$.

3.1.2. Evaluation of variational inference

As is commonly done in studies investigating variational inference techniques to approximate a target posterior distribution, we compare the estimated variational distribution with direct samples from the posterior using MCMC. Here, we utilize a No U-Turn Sampler scheme implemented in the NumPyro Python package (Phan and others, 2019), which uses automatic differentiation to efficiently generate sample trajectories for moderately high numbers of model parameters. We use the velocity and thickness predictions from the neural network as the observations such that MCMC only samples the rigidity posterior distribution, which allows for a more direct comparison between the variational and sampled posterior.

We find that both MCMC and variational inference recover a posterior mean profile for θ that is close to the true values for areas >20 km upstream from the ice front (Fig. 2). Close to the ice front, both methods predict uncertainties that are substantially larger due to thinner ice, which reduces the sensitivity of the longitudinal and lateral membrane stresses (left-hand terms in Eqn (17)) to rigidity variations. These uncertainties near the ice front can likely be reduced by augmenting the SSA residual loss with a dynamic boundary condition at the ice front that balances longitudinal stresses normal to the ice front with the difference in hydrostatic pressure between the ice and ocean water (Larour and others, 2005). Pair plots of marginal distributions of θ at different locations along the ice shelf show that the variational approach is able to recover strong covariances between θ samples for locations that are relatively close to each other while ensuring samples

https://doi.org/10.1017/jog.2023.8 Published online by Cambridge University Press

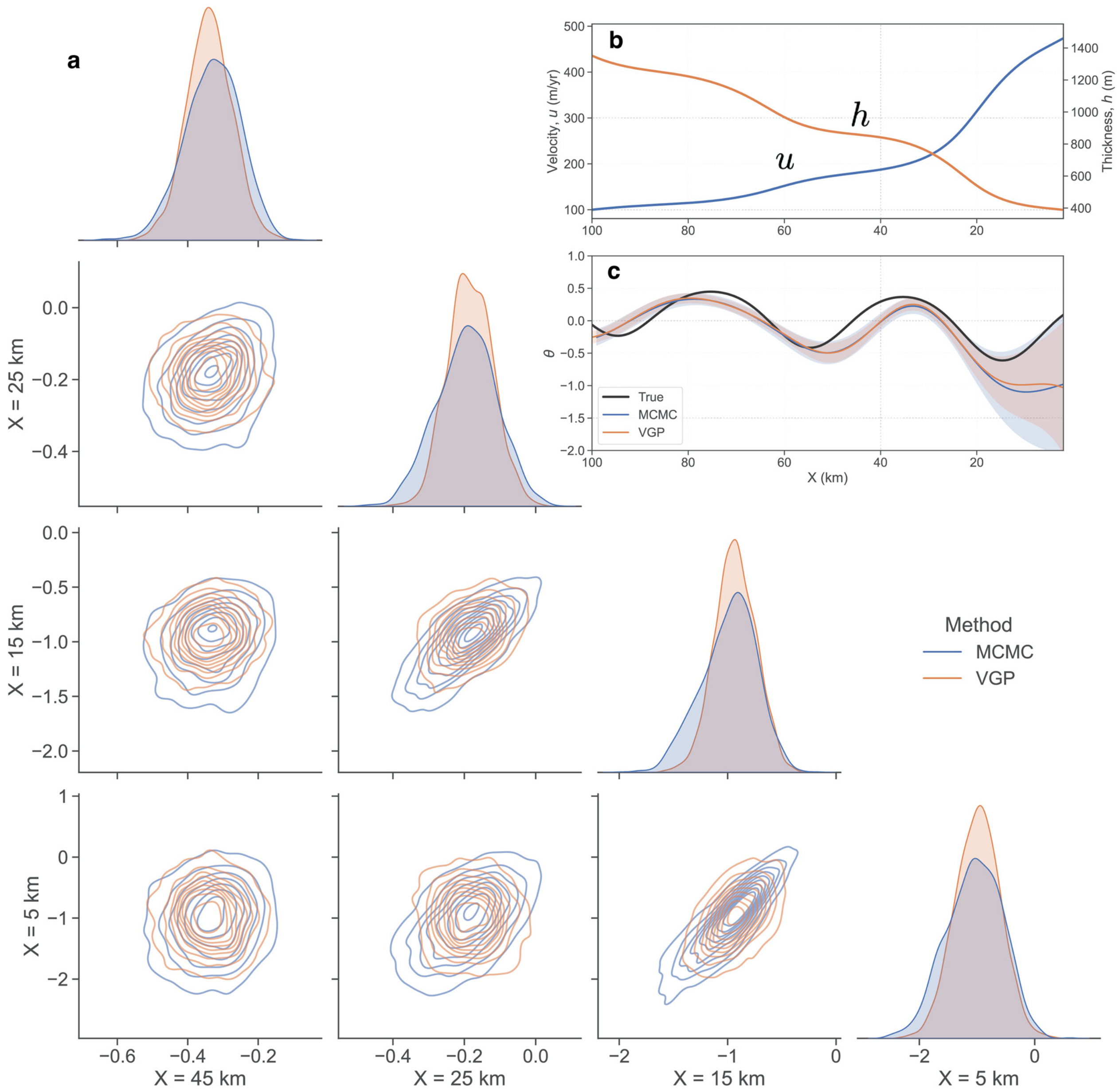

Fig. 2. Posterior inference of ice rigidity for simulated 1D ice shelf. (a) Posterior marginal distributions for θ at different locations along the ice shelf. The grounding line is at $X = 0$ km and the ice front is at $X = 100$ km. Diagonal plots show the 1D marginals computed from posterior samples generated with MCMC (blue) and the variational Gaussian process (VGP; orange). Off-diagonal plots show 2D covariance plots for the same sample set. All marginals have been smoothed using a Gaussian kernel density estimator. (b) Velocity (blue) and ice thickness (orange) of ice shelf used for posterior inference. (c) Comparison of the true $\theta(X)$ against the mean $\theta(X)$ computed from the MCMC samples (blue) and the VGP (orange). The shaded regions correspond to 2σ posterior uncertainties. Overall, the posterior distributions for MCMC and VGP are very similar. The largest deviations occur near the ice front where the marginals exhibit stronger non-Gaussian behavior, which cannot be modeled by the VGP.

are uncorrelated for larger pair-wise distances. In general, the strength of the posterior covariance will be modulated by the physical model as well as the prior correlation lengthscale. Closer to the ice front, the marginal distributions derived from MCMC indicate a slight deviation from Gaussian behavior, which is again likely due to the lower ice thicknesses limiting ice stress sensitivity to rigidity variations. Since the variational distribution is constrained to be a multivariate normal, it is unable to recover the skewed, non-Gaussian behavior in the marginals in these areas. For applications where it is desirable to place more probability mass in the longer-tails of the posterior distribution, one could simply increase relevant variances (σ_θ, $\sigma_\mathbf{r}$) to inflate the variational posterior variances or use a more flexible variational approximation not restricted to multivariate normals (Rezende and Mohamed, 2015).

3.2. 2D ice shelf

We now simulate a 2D ice shelf using the icepack ice flow modeling software (Shapero and others, 2021). Similar to the synthetic ice shelf presented in Shapero and others (2021), we prescribe a semi-circular shelf geometry with four inlet glaciers of varying widths (Fig. 3). Additionally, we prescribe a bed topography that results in a few pinning points where the flotation height is positive, i.e. the ice is actually grounded at these locations. Under the shallow-stream approximation, we prescribe a basal drag friction coefficient proportional to the flotation height such that friction is only non-zero for grounded ice. Such pinning points in the form of ice rumples are common in ice shelves in Antarctica. However, assuming a fully floating ice shelf during inversion for rheological parameters will introduce errors into

https://doi.org/10.1017/jog.2023.8 Published online by Cambridge University Press

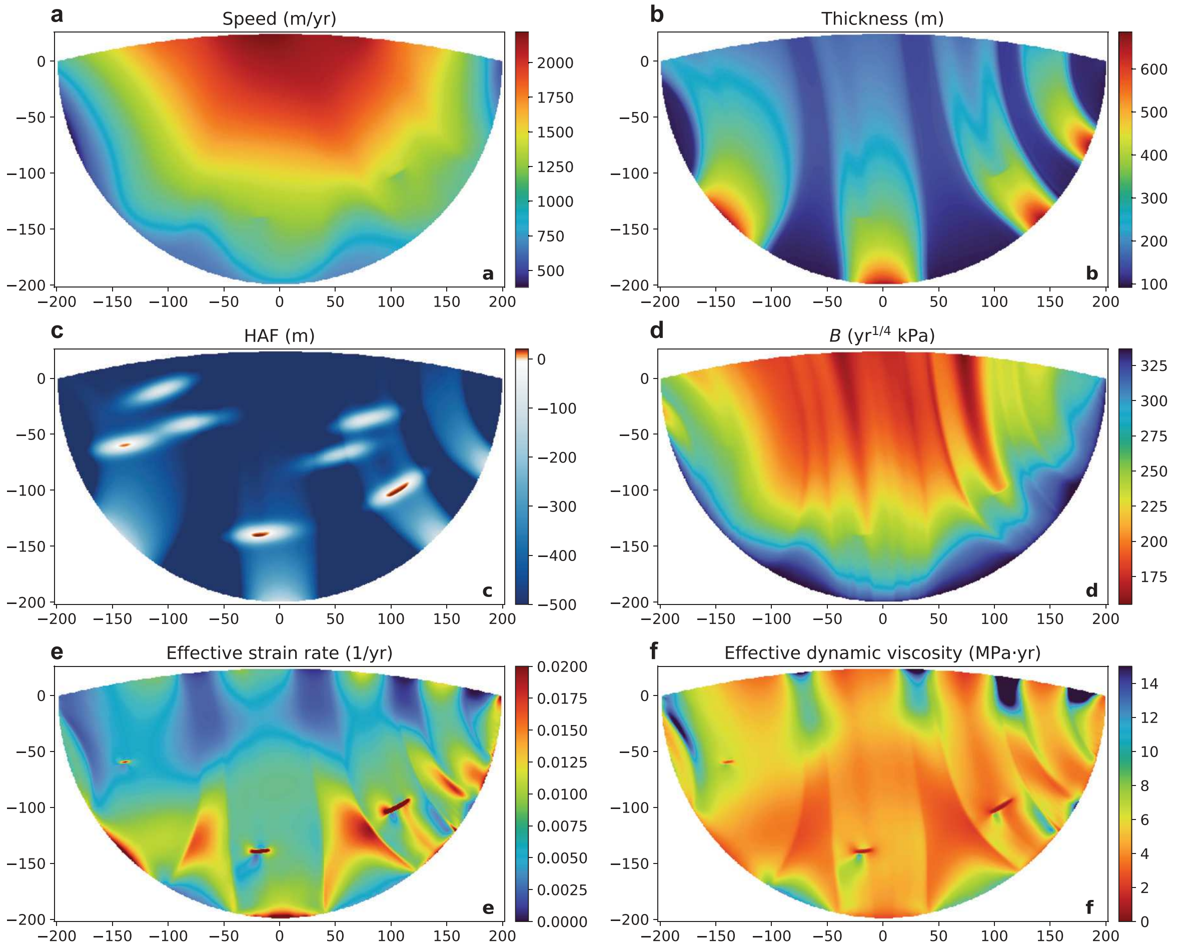

Fig. 3. 2D ice shelf with simulated damage evolution and pinning points. The simulation outputs shown are computed from 700 years of spinup in order to achieve steady state. The ice shelf is fed by four inlet ice streams, as evidenced by the flow speed (a) and ice thickness (b). Height-above-flotation (HAF) in (c) shows the location and orientation of the prescribed pinning points. The steady-state ice rigidity B (d) reflects damage accumulation due to shear margin weakening and ice thinning due to large strain-rates over the pinning points. The effective strain rate (e) and effective dynamic viscosity (f) are approximately inversely related and show strong shearing in the ice between the inlet flow, as well as over the pinning points. Strain rates are lower closer to the ice front. The effective viscosity exhibits a mix of long-wavelength variations within flow units and short-wavelength variations near the shear margins.

the inferred parameter field due to model mismatch. Therefore, by purposefully injecting modeling errors into the estimation procedure, we can assess how the two different cost functions and the estimated parameter uncertainties respond to such errors.

We use icepack to first simulate the evolution of shelf velocity and thickness for roughly 500 years with a constant ice rigidity, B_0, corresponding to an ice temperature of $-5\,^{\circ}$C, a net surface mass balance of 0, and a stress exponent of $n = 4$. Here, we choose $n = 4$ in order to evaluate the sensitivity of the rigidity inference to 2D ice stress variations that are more likely to be found in natural environments of fast-flowing ice (Bons and others, 2018; Millstein and others, 2022). After the first simulation stage, we apply a continuum damage mechanics model that modulates the rigidity field with an evolving damage factor, D, such that $B_D = (1 - D)B_0$ (Borstad and others, 2013; Albrecht and Levermann, 2014). This approach provides a physically realistic means to obtain a spatially varying prefactor field with rheology-modifying processes such as shear weakening. We run the damage-enhanced model for an additional 200 years to achieve approximate steady-state. At the end of the simulation, we can observe substantial spatial variation in damage, where ice is nearly undamaged at the grounding line (due to a zero-damage boundary condition) and highly damaged near the ice front, at shear margins and downstream of the pinning points. The dynamic effective viscosity field shows concentrated low viscosities near the pinning points and higher viscosities between the inlet ice streams where deformation rates are lower. Overall, the viscosities exhibit a mix of short- and long-wavelength features, which are mirrored in the effective strain rate field.

3.2.1. Variational inference setup

For recovery of the rigidity field, as we discussed during selection of the prior variance, we explore both the conventional control method-based inversion and the variational inference approach based on the momentum balance objective, as well as a combination of the two where we use the inversion to set B_0 for the prior. For all approaches, we use the simulated ice surface elevation to compute ice thickness by assuming hydrostatic equilibrium (buoyancy). Over floating ice, the thickness values derived from buoyancy are identical to the simulated thickness, but over the pinning points, the actual thickness values are lower, which results in an overestimation of the driving stress variations using the buoyancy conversion (Fig. S2). Furthermore, assuming flotation for the entire ice shelf will neglect the basal drag provided by the pinning points. The combined data and modeling errors will impact recovery of the prefactor field, which we explore shortly.

The control method inversion is again performed with icepack, using a Gauss–Newton solver to minimize a joint objective function that combines a velocity prediction error function and a regularization function based on the first-derivative of the log rigidity field, θ. The inversion includes Dirichlet boundary

https://doi.org/10.1017/jog.2023.8 Published online by Cambridge University Press

conditions for velocity values at the grounding line and Neumann (dynamic) boundary conditions at the ice-ocean interface. For the variational inference problem, we select 20 000 uniformly random locations on the ice shelf to extract velocity and thickness values to use as training data for the network f_ψ (feedforward network of four layers of 100 nodes each), which is only tasked with reconstructing the surface observations. We select an additional, independent set of 20 000 random locations, or collocation points, for training the VGP (with 750 inducing index points), which is tasked with predicting the parameters of the variational distribution $q(\theta)$. Both the number of observations and collocation points will influence the effective spatial resolution of θ and can be interpreted as a mesh-free analog of the mesh element size. For all priors, we prescribe a lengthscale of 15 km, and for the prior with a uniform B_0, we use a value of $B_0 = 260$ yr$^{1/4}$ kPa. After training, we evaluate training performance by reconstructing the surface observations over the entire model domain (using f_ψ), as well as the predicted SSA residuals (using f_ψ and mean rigidity as predicted by the VGP). For the variational inference predictions, the observation misfits and drag residual are minimal over most of the modeling domain but are higher over the two largest pinning points (Fig. S4). The higher errors are a function of oversmoothing of the observations and model mismatch, which amounts to assuming ice is floating over the grounded pinning points. As a consistency check, we use the posterior samples of θ to generate stochastic predictions of velocity using the standard forward model and find that velocity errors are generally <5% of the flow speed, with higher error values localized to the pinning points (Fig. S5). We note that the velocity errors are commensurate with those from the conventional inversion.

3.2.2. Evaluation of rigidity reconstruction errors

A more detailed comparison of the recovery error for B (recovered using $B = B_0\,e^{\theta}$) between the control method inversion and variational inference reveals that the two methods are complementary. The control method inversion has the lowest overall error bias, but the areas where the errors are largest are systematically upstream of the pinning points (Fig. 4). Since we assume all ice is floating for the forward model, the missing resistive stress provided by drag at grounded ice is compensated by artificially making the ice stiffer upstream of the pinning points, which acts to slow the ice down in a manner that allows the predicted velocities to match the observed velocities. In contrast, the B recovered by variational inference (which uses the SSA momentum balance as a forward model) shows larger errors directly over the pinning points, as well as in areas where ice is stagnant (low strain rate). Over the pinning points, the true rheology sharply varies from about 250 to 200 prior yr$^{1/4}$ kPa (Fig. 3d). However, the prior lengthscale of 15 km encourages spatially smoother fields of ice rigidity, which limits the dynamic range of ice stresses that can be modeled in order to satisfy the SSA momentum balance. Since the driving stress variations over the pinning points are overestimated due to the buoyancy assumption (Fig. S2), the preferred solution is to smooth out all stress variations over the grounded ice in order to minimize the residual SSA terms. Upstream of the pinning points and closer to the grounding line, the recovery errors are actually lower using variational inference as compared to the control method inversion. The spatial patterns in the recovery errors are similar to the patterns of residual SSA components (Fig. S4). Finally, by using the control method inversion as the reference B_0 for variational inference, we can minimize much of the recovery errors closer to the ice front and in areas where strain rates are lower but flow speeds are still high, i.e. areas where the inversion has greater sensitivity and where the dynamic boundary condition at the ice-ocean interface provides additional constraints on the rigidity (Fig. 4c).

3.2.3. Evaluation of rigidity uncertainties

The predicted uncertainties for θ are consistent with the reconstruction errors: uncertainties are higher closer to the ice front where ice thicknesses are lower (as observed in the 1D case), as well as in more stagnant ice where strain rates are lower (Figs 4d, 5). Uncertainties near the ice front are reduced when using B_0 from the control method inversion (Fig. S3), which is again likely due to the incorporation of the dynamic boundary condition at the ice-ocean interface which also reduced reconstruction errors. In areas where ice is thinner but strain rates are higher (e.g. higher shear strain rate in the areas between the fast-flowing ice), the balance between extensional stresses and lateral drag also provides sufficient signal for reducing uncertainties. In a few isolated patches, even when effective strain rates are low and ice is relatively thin, slightly positive lateral forces that act as a 'pull' on the ice can also reduce uncertainties (Fig. 5). Directly over the pinning points, uncertainties are low due to the high

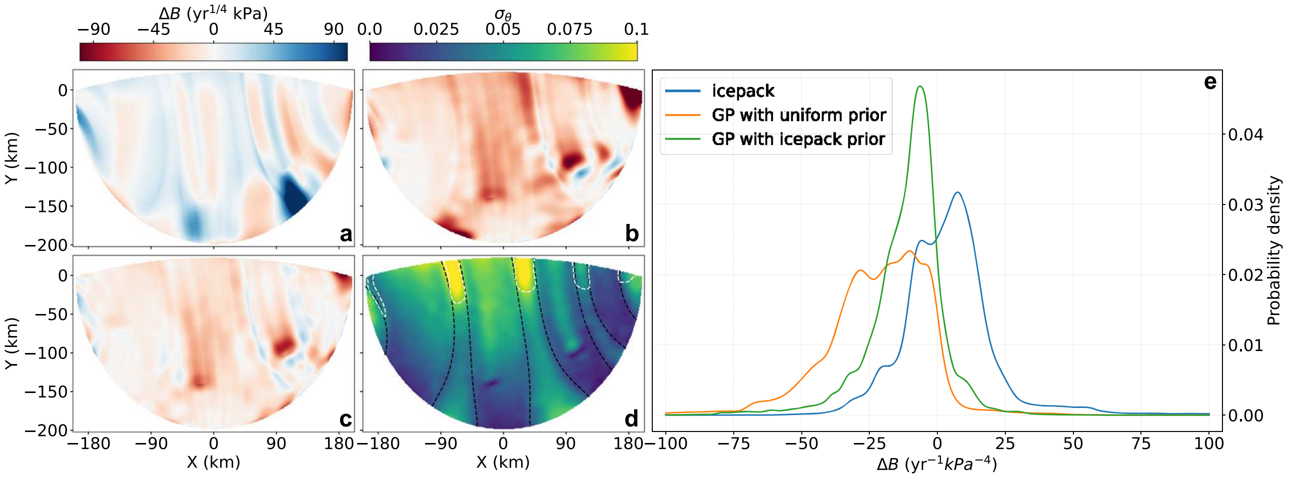

Fig. 4. Comparison of reconstruction errors for mean inferred ice rigidity between control method inversion and the proposed variational inference method. (a) Error ($B - B_{\text{true}}$) for control method inversion using icepack. (b) Error for variational inference with a uniform B_0 field. (c) Error for variational inference using the control method inversion for B_0. (d) Inferred uncertainty for normalized rigidity parameter θ using the control method inversion for B_0. Black dashed lines correspond to a thickness contour of 150 m while the white dashed lines correspond to an effective strain rate contour of $10^{-2.6}$ a^{-1}. (e) Histograms of errors for different methods. Higher reconstruction errors and uncertainties are mostly concentrated in thinner ice and areas with lower effective strain rates.

https://doi.org/10.1017/jog.2023.8 Published online by Cambridge University Press

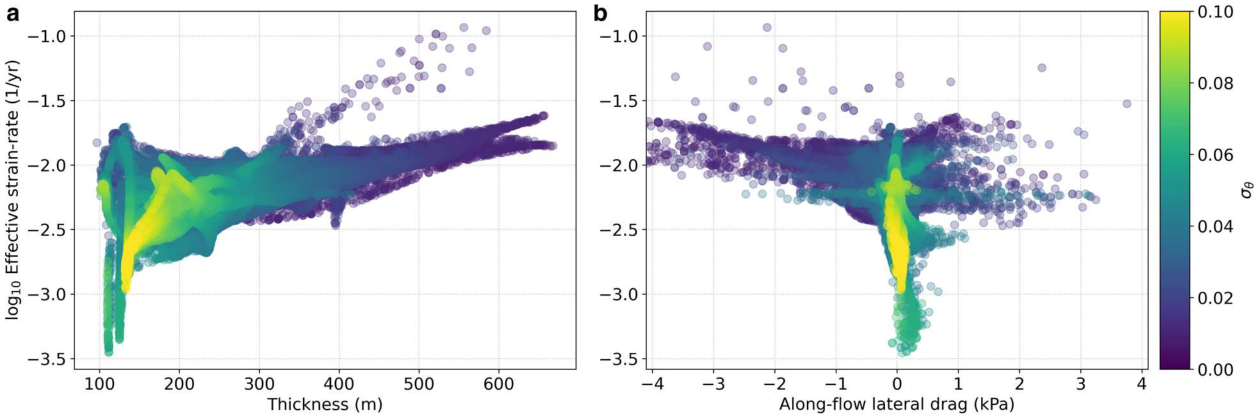

Fig. 5. Uncertainty for normalized rigidity θ vs ice thickness, along-flow lateral drag and effective strain rate for simulated 2D ice shelf. (a) Effective strain rate vs ice thickness with colors corresponding to uncertainty in θ. (b) Same as (a) but for effective strain rate vs along-flow lateral drag. Here, lateral drag is computed as the transverse gradients of the SSA momentum balance projected to the along-flow direction, where negative values denote flow resistance. While ice thickness is the first-order control on rigidity uncertainties, higher strain rates can reduce uncertainties in thinner ice. Positive lateral forces can also reduce uncertainties where effective strain rates are low.

strain rates there. This mismatch between low uncertainties and high SSA residuals over the pinning points may be compensated by inflating the variance σ_τ^2 over areas where measured flotation height is near zero (see Discussion section). Downstream of the pinning points, we observe higher uncertainties due to downstream thinning of the ice. Overall, the uncertainty maps for θ indicate that areas with thicker ice and higher strain rates are better constrained, and targeted inverse modeling (e.g. estimating B_0 from a control method inversion) can be an effective tool for further reducing uncertainties (Fig. S3).

4. West Antarctica ice shelves

We now apply our methods to select large ice shelves in West Antarctica, specifically the Larsen C Ice Shelf (LCIS), Filcher-Ronne Ice Shelf (FRIS), Ross Ice Shelf (RIS) and the combined Brunt Ice Shelf with Stancombe-Wills Ice Tongue and Riiser-Larsen Ice Shelf (B-SW-RL) (Fig. 6). These ice shelves are fairly representative of shelf environments on the Antarctic coast and serve as a robust testing suite for several reasons. Firstly, they encompass a large area (48, 380, 440 and 68×10^3 km^3 for LCIS, FRIS, RIS and B-SW-RL, respectively), corresponding to a large number of effective modeling parameters in order to test the inference capacity of the VGP. Secondly, the ice shelves are subject to different flow and buttressing environments. Large ice rises in Larsen C have favored the formation of large rifts, the evolution of which are complicated by the presence of mechanically weak suture zones that likely contain large proportions of mechanically weak marine ice (Jansen and others, 2013; Kulessa and others, 2014; Borstad and others, 2017). Within Ross Ice Shelf (the largest ice shelf in Antarctica), a mix of ice rises, ice rumples and large islands serve to create a heterogeneous flow environment involving localized grounding, rift formation and shear margin weakening. Many of these pinning points lie in the western portion of the shelf off the Siple Coast, which drains much of the West Antarctic Ice Sheet through fast-flowing ice streams. Filchner-Ronne is also fed by several fast-flowing ice streams with large ice thicknesses, leading to larger driving stresses over the ice shelf with the highest overall flow speeds of the ice shelves examined here. The Brunt-Stancomb-Wills-Riiser-Larsen shelf complex (B-SW-RL) is subject to lower buttressing than Larsen C or Ronne-Filchner due to lack of embayments. However, within the Riiser-Larsen shelf are a few prominent pinning points that do provide limited buttressing but also serve as potential areas of model mismatch, similar to the synthetic ice shelf we previously investigated. Additionally, much of the ice in the Stancomb-Wills ice tongue is more loosely packed, leading to large surface gradients at the edges of individual ice units that are not well-matched to velocity variations.

4.1. Remote-sensing data and pre-processing strategy

For RIS and B-SW-RL, we use the MEaSUREs velocity mosaic (Rignot and others, 2011; Mouginot and others, 2012), which combines speckle tracking of SAR images from various satellite platforms with feature tracking of Landsat 8 images and has a nominal temporal coverage between 2009 and 2016. For LCIS and FRIS, we use a 2020 annual velocity mosaic provided by ITS_LIVE, which is derived from feature tracking of Landsat 7 and 8 images over Antarctica (Gardner and others, 2019). From a visual inspection, we found that the ITS_LIVE mosaic exhibited fewer velocity artifacts for LCIS and FRIS, whereas the MEaSUREs mosaic exhibited fewer artifacts over B-SW-RL and provided full coverage over RIS. Ice thickness and elevation data are derived from BedMachine V2 (Morlighem and others, 2020), which combines radar-estimated thickness profiles with mass conservation constraints and firn corrections in order to obtain continuous thickness maps. While the nominal year for the thickness data is 2015, the correspondence between the velocity and thickness data are sufficient for the spatial resolution of our analysis (assuming an upper bound of ≈5 km of motion for feature advection). For all velocity and thickness rasters, we first perform a void-filling operation that uses a spring-based PDE constraint to fill in missing data (D'Errico, 2012). The rasters are then filtered to ≈10–15 times the average ice thickness using a Savitzky–Golay filter in order to remove high-frequency components not resolvable by the SSA momentum balance.

4.2. Variational inference setup

As with the simulated ice shelf, we first invert for B using icepack in order to optimize a cost function combining a velocity misfit term (weighted by the formal uncertainties for the velocity estimates) and a regularization term based on first-order spatial gradients to encourage smoother solutions. All ice shelves are discretized into 2D triangular finite elements with an average size of 5 km. A penalty parameter controlling the relative contribution of the regularization term is selected with a standard L-curve analysis, independently for each ice shelf. For each ice shelf, we use feedforward neural networks with four layers of 100 nodes each and VGPs with 600–900 inducing index points.

https://doi.org/10.1017/jog.2023.8 Published online by Cambridge University Press

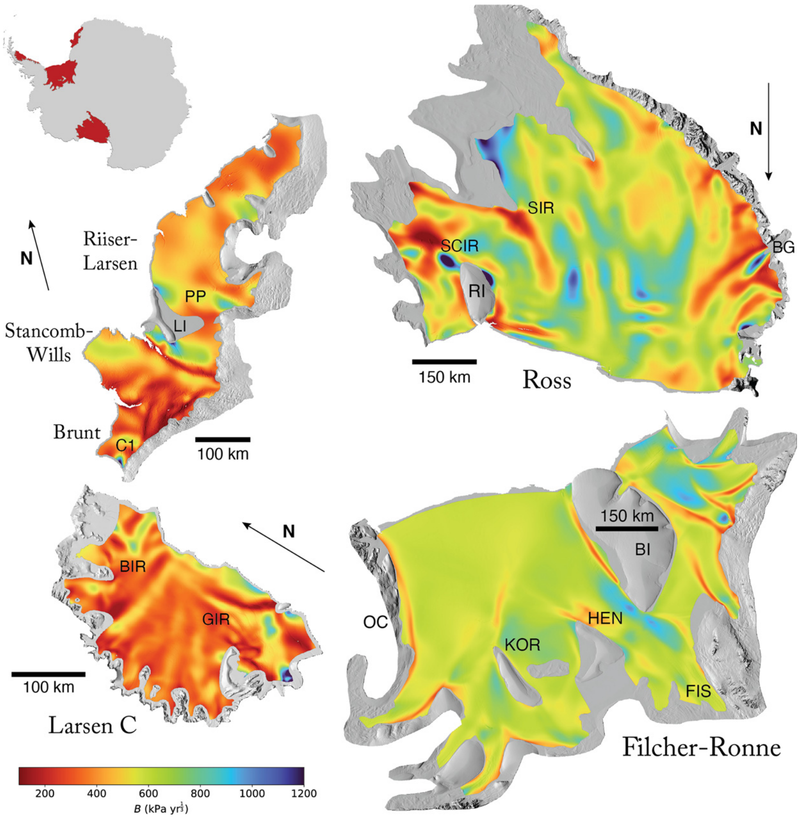

Fig. 6. Estimated mean ice rigidity B for West Antarctic ice shelves. Specific features in Ross Ice Shelf are Shirase Coast Ice Rumples (SCIR), Steershead Ice Rise (SIR), Roosevelt Island (RI) and Byrd Glacier (BG). At Filcher-Ronne Ice Shelf are Korff (KOR) and Henry (HEN) ice rises, Berkner Island (BI), Foundation Ice Stream (FIS), and Orville Coast (OC). At Larsen C are Bawden (BIR) and Gipps (GIR) Ice Rises. At Brunt-Stancomb-Wills-Riiser-Larsen is Chasm 1 (C1), Lyddan Island (LI), and an unnamed pinning point (PP). Inset at the top left shows the location of the ice shelves in Antarctica. Overall, areas of soft ice are inferred at shear margins and large surface crevasses, while areas of stiffer ice are associated with thick ice in compressional zones.

The training data for the feedforward neural networks and the collocation points for the VGP are independently constructed with a point density of 2 points per 10 square kilometers, which is commensurate with the size of the mesh elements used for the icepack inversion. This point density for the VGP is sufficient to model a moderate-resolution (~20 km) θ field, which when combined with the higher-resolution B_0 from the control method inversion results in rigidity samples that resolve important variations near shear margins and rifts. The neural network training observations are randomly sampled directly from the raster grids, whereas the collocation points are randomly generated using a latin hypercube sampling scheme. Finally, we use the estimated B field as the reference field B_0 for parameterization of θ, setting the prior variance for θ to 0.2^2.

4.3. Analysis of rigidity mean and uncertainty fields

To a first-order approximation, ice is inferred to be stiffer for FRIS and RIS than for LCIS and BWSRL, and average rigidity values for LCIS are the lowest of the four (Fig. 6). These first-order trends are well-matched by modeled ice shelf surface temperatures where temperatures for FRIS are generally around −25 to −30 °C, whereas for LCIS they range from −15 to −10 °C (Fig. S8). However, all ice shelves exhibit significant spatial variability in inferred ice rigidity beyond surface temperature variations in order to minimize SSA residuals (Figs S9, S10). For FRIS, the estimated mean B field is broadly consistent with results from prior studies (e.g. MacAyeal and others, 1998; Larour and others, 2005). Ice is inferred to be substantially softer in the shear margins where strong lateral shearing leads to viscous dissipation and elevated ice temperatures. These shear margins are prominent in the Ronne Ice Shelf where fast-flowing floating ice is in contact with rock (along the Orville coast and Berkner Island) or stagnant ice, as is the case downstream of the Korff Ice Rise. As discussed in Larour and others (2005), larger basal melt rates on the northern tip of the Henry Ice Rise are coincident with softer ice. Within the Filchner Ice Shelf, lower overall values of B indicate softer ice, again in the shear margins where ice streams flow onto the shelf and are in contact with stagnant ice. A large lateral surface crevasse close to the ice front is also associated with higher strain rates and softer ice. We can also observe localized regions of substantially stiffer ice, such as downstream of the Foundation Ice Stream and upstream of the Korff and Henry Ice Rises. These regions are associated with larger driving stresses (Fig. S7) such that ice is inferred to be stiffer in order to provide enough resistive stresses to balance those driving stresses. Ice is also inferred to be stiffer closer to the grounding line where colder ice is advected by the ice streams. For all ice shelves, rigidity uncertainties are mostly lower where ice is thicker and strain rates are larger, similar to what was observed for the simulated 1D and 2D ice shelves (Fig. 7).

Similar to FRIS, the ice in the central portions of RIS are inferred to be more rigid, likely due to relatively cold surface temperatures of −20 °C. However, we can also observe zones of softer ice near shear margins and localized areas of grounding. At the inlet of the Byrd Glacier to the west, prominent shear margins separating the fast-flowing inlet ice from more stagnant shelf ice are coherent for more than 300 km downstream of the grounding line (Fig. S7), which results in substantial shear weakening. In the central trunk of the Byrd Glacier inlet, the reduction in flow speed as the ice flows onto RIS leads to enhanced compressional stress and thickening of the ice, leading to inferred higher B values. On the east side of RIS, the Shirase Coast Ice Rumples (SCIR) at the outlet of the MacAyeal and Bindschadler Ice Streams significantly modify the flow field and ice thickness due to grounding of the ice, consistent with the simulated pinning points for the synthetic ice shelf. Thinning of ice downstream of SCIR and diversion of the shear margins toward Roosevelt Island (RI) are both dynamical effects that modify the buttressing capability of ice in this region (Still and others, 2019; Still and Hulbe, 2021). In our inferred B field, the ice covering the rumples is inferred to be

https://doi.org/10.1017/jog.2023.8 Published online by Cambridge University Press

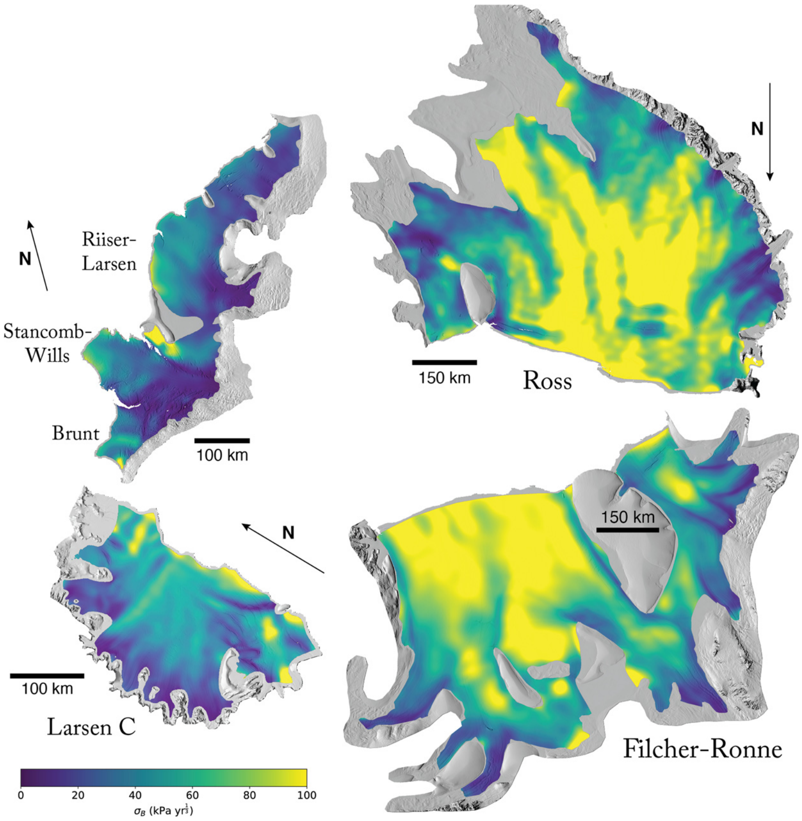

Fig. 7. Estimated 1-σ B uncertainties for West Antarctic ice shelves. Uncertainties are generally larger for higher B values (scale-dependence) and for areas with thinner ice and lower driving stresses. Uncertainties tend to be lower closer to the grounding line.

softer while the downstream ice connected to RI is inferred to be stiffer. Alternatively, the ice upstream of Steershead Ice Rise (SIR) is near-stagnant, leading to very high inferred values for B. Downstream of SIR is a streakline of thin ice coincident with the shear margin of the inlet of MacAyeal and Bindschadler Ice Streams, leading to a narrow zone of soft ice that persists nearly all the way to the ice front.

At LCIS, the softest ice is inferred within highly localized areas corresponding to surface crevassing, including the large rift originating from the Gipps Ice Rise (Khazendar and others, 2011; Larour and others, 2021). It is likely that some fraction of the inferred softness is due to not explicitly including rifts (geometrically and dynamically) within the ice flow model, which can reproduce a significant proportion of the observed strain rates with active opening/closing of rifts (Larour and others, 2021). As is the case with FRIS, stiffer ice is inferred near the grounding line where colder and thicker ice is advected downstream by the inlet ice streams. Within the ice shelf, areas in between faster flowing ice correspond to thinner ice and higher strain rates, resulting in softer ice. Unlike FRIS, the proximity of the fast flowing inlet ice streams with one another limits the areal extent of stagnant ice over Larsen C. High effective strain rates between ice streams are aligned with the initiation of suture zones where mechanically weak marine ice (sourced from warmer ocean water) has been observed to accumulate at the base of LCIS (Kulessa and others, 2014). The initial portion of the suture zones within approximately 20–30 km downstream of promontories and peninsulas are associated with inferred softer ice. Upstream of the Bawden Ice Rise (BIR), strain rates are substantially lower and correspond to larger inferred B values. Here, the correspondence between large fractures and a simulated confluence of meltwater plumes is hypothesized to stimulate abundant accretion of marine ice, which can actually lead to ice stiffening (Khazendar and others, 2011).

Finally, for B-SW-RL, ice is inferred to be substantially softer in the mélange area that separates the Brunt Ice Shelf from the Stancomb-Wills Ice Tongue, as well as in the mélange that separates the latter from the Riiser-Larsen Ice Shelf. These areas, which contain a heterogeneous mixture of marine ice, sea ice and ice shelf debris, have previously been inferred to exhibit lower rigidity values (within a continuum mechanics model) and act to bind large ice fragments to the coast (Khazendar and others, 2009). Since the mélange is less coherent than meteoric ice advected from the ice streams, it deforms readily and corresponds to high strain rates. Additionally, prominent surface crevasses throughout B-SW are also associated with softer ice, including several transverse rifts close to the grounding line of Brunt Ice Shelf and a frontal rift separating the northeastern corner of Brunt Ice Shelf from the Stancomb-Wills Ice Tongue. Since the nominal temporal coverage of the MEaSUREs velocity data is 2009–2016, the Halloween Crack has not yet initiated (De Rydt and others, 2019). At the southern edge of Brunt Ice Shelf at the base of Chasm 1, ice is actually inferred to have high mean B, but since uncertainties are large here (Fig. 7), we consider this to be a smoothing artifact stemming from larger thickness errors near the large rifts. Upstream of the prominent pinning point on the Riiser-Larsen Ice Shelf (PP in Fig. 6), ice is inferred to be stiffer, similar to what we observed with the pinning points for the simulated ice shelf as a compensation for unmodeled basal drag. The thinner ice downstream of the pinning point is correspondingly inferred to be softer. We do note that the orientation of the flow field relative to the pinning point is more oblique than that of our simulated shelf, which likely is the source of the more complex strain rate pattern adjacent to the pinning point (Figs S7, S11). Finally, upstream of Lyddan Island in the mélange at the eastern edge of Stancomb-Wills, ice is inferred to have high rigidity, but as this area corresponds to both low strain rates and low driving stress, the uncertainty in rigidity is very large.

4.4. Posterior predictive distributions and ice shelf buttressing

After obtaining the variational distribution that best approximates the posterior distribution for the ice rigidity, we can compute a

https://doi.org/10.1017/jog.2023.8 Published online by Cambridge University Press

posterior predictive distribution for any quantity or forward model that depends on the rigidity. The most straightforward way to accomplish this is to generate random samples from the variational distribution and pass each sample through the forward model of interest, i.e. Monte Carlo approximation. For example, one could perform a dynamic perturbation analysis on specific ice shelves by applying some form of stress perturbation at the ice front (calving event, gain/loss of buttressing sea ice, etc.) and running prognostic simulations for different realizations of the rigidity, sampled from the posterior distribution. This type of analysis has been performed in many studies to assess sensitivity of ice shelves to changing climate conditions (e.g. Schlegel and others, 2018; Nias and others, 2019), but usually the rigidity field is varied by choosing some uniform upper and lower bound guided by expected temperature variations or other a priori knowledge on creep mechanisms. By instead using the posterior distribution to draw samples of the rigidity, we automatically incorporate information derived from surface observations while also allowing known physical laws (e.g. SSA equations) to induce realistic covariances between values of the rigidity over finite length scales. In other words, the combined information from data and flow equations results in more realistic samples of physical parameters consistent with all available knowledge.

Since one of the most important physical implications of ice shelf rheology is the amount of buttressing applied to inland grounded ice, we use the variational distribution for B to compute the distribution of maximum buttressing factors following Fürst and others (2016). The normal buttressing number is defined as (Gudmundsson, 2013):

$$K_n = 1 - \frac{\hat{\mathbf{n}} \cdot \boldsymbol{\tau}\hat{\mathbf{n}}}{N_0}, \tag{19}$$

where $\boldsymbol{\tau}$ is the deviatoric stress tensor, the quantity $N_0 = \frac{1}{2}\rho_i(1 - \rho_i/\rho_w)gh$ is the vertically integrated pressure exerted by the ocean (with density ρ_w) on the ice shelf, and $\hat{\mathbf{n}}$ is a normal vector selected to be aligned with the second principal stress, following Fürst and others (2016). By performing systematic calving simulations where ice is removed from an ice shelf up to different buttressing factor isolines, the increase in ice flux across the ice front or grounding line can be predicted for various buttressing factors (Fürst and others, 2016). The buttressing factor above which ice flux is projected to rapidly increase then serves as a buttressing threshold for a given ice shelf. The isoline corresponding to the threshold can then delineate regions of ‘passive’ shelf ice (PSI), defined as ice that can be removed without significantly altering the flow dynamics of the adjacent ice. As the normal force in the buttressing factor is computed from the ice stress tensor, which itself depends on the rigidity B to estimate the stress components, the buttressing factor will be subject to random variations consistent with the posterior samples of B. We can therefore estimate the expected variation in PSI consistent with the surface observations. To estimate a more realistic estimate of PSI area specific to calving, we only include buttressing factor isolines that form polygons that intersect the ice front, meaning we exclude areas of isolated PSI closer to the grounding line.

The buttressing thresholds originally presented by Fürst and others (2016) corresponded to flux increases across the ice front, leading to threshold values of 0.3–0.4 for the ice shelves investigated here. Alternatively, thresholds defined for increased ice flux across the grounding line are found to be a better predictor for ice shelf stability in response to instantaneous calving events (Reese and others, 2018; Mitcham and others, 2022). These buttressing values tend to range from 0.8 to 0.9. For the purposes of comparison with the result of Fürst and others (2016), we use a lower threshold of 0.4 roughly corresponding to a step increase in flux across the ice front. Due to slight biases between our inferred mean B fields and the fields estimated by Fürst and others (2016), our threshold value of 0.4 is slightly higher than that used by Fürst and others (2016) in order to roughly match the PSI regions in that study.

We observe variations in PSI that lie roughly within the bounds computed from ±10% variation of the mean B, following Fürst and others (2016) (Fig. 8). However, we can observe additional spatial and statistical patterns beyond the simple ±10% variations. For the ice shelves that are laterally confined by embayments, there are a significant number of samples of the PSI boundary that exceed the upper and lower bounds. Over Larsen C, the PSI boundary samples are slightly skewed toward lower PSI areas. However, several posterior samples of B actually connect passive ice centered on the rift originating from GIR to passive ice at the ice front, which increases total PSI area and slightly reduces the vulnerability of Larsen C to ice loss. Over FRIS and Ross, the PSI distribution is more symmetrical, although the former has a long tail of lower PSI areas, which correspond to a slight increase in vulnerability of those shelves to ice loss. Finally, over B-SW-R, the distribution of PSI is near-symmetric and lies well within the ±10% bounds. However, the difference in spatial extent between the ±10% bounds is larger than for the other ice shelves, particularly for the Stancomb-Wills ice tongue, which indicates a greater sensitivity to variations and uncertainties in inferred ice rigidity. This sensitivity is likely reflective of the lack of lateral confinement and drag and highlights the importance of embayment geometry on ice shelf buttressing force. Overall, these results demonstrate that calibration of ice shelf rigidity and associated uncertainties using surface data can both inflate/deflate predictive uncertainties and needs to be performed on a shelf-by-shelf basis.

5. Discussion

We demonstrated our proposed physics-informed variational inference framework by estimating the posterior distribution of ice rigidity for synthetic and large-scale ice shelves in Antarctica. The variational inference scheme produces posterior distributions of rigidity that agree well with those estimated by MCMC methods while providing a scalable approach for exploring uncertainties in parameter fields and forward predictions. We now briefly discuss potential avenues for further exploration of ice rheological parameters using distributions of B, as well as future algorithmic and computational improvements.

5.1. Uncertainties in ice rigidity propagated to flow law parameters

In this work, we focused on estimating the variational distribution for ice rigidity, B, and demonstrate how the the inferred uncertainties can be used to form predictive distributions on a derived buttressing factor. However, B was defined using the form of Glen’s Flow Law in Eqn (1), which aggregates multiple physical factors into a single prefactor. The prefactor can be disaggregated using an Arrhenius-type relation with the following form (using the convention that $B = A^{-1/n}$) (Cuffey and Paterson, 2010):

$$A = EA_0 \exp\left\{\frac{-Q_c}{R}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right\}, \tag{20}$$

where A_0 is a reference prefactor value, R is the ideal gas

https://doi.org/10.1017/jog.2023.8 Published online by Cambridge University Press

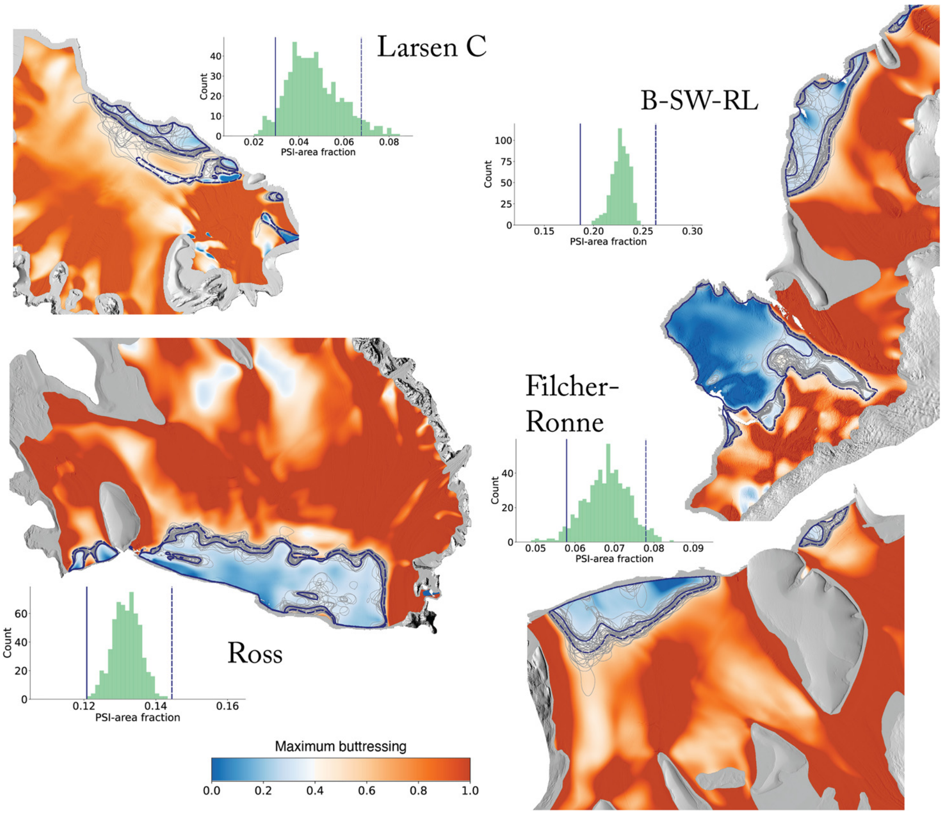

Fig. 8. Stochastic analysis of maximum buttressing factor for West Antarctic ice shelves, following Fürst and others (2016). Background 2D buttressing fields are computed from the mean *B* inferred from variational inference for each ice shelf. The colormap is constructed to highlight a threshold buttressing value of 0.4, which roughly corresponds to a step increase in ice flux across the ice front for removal of ice up to the 0.4 buttressing isoline. Thus, blue areas correspond to 'passive' ice. The thick solid and dashed dark blue lines correspond to the 0.4 isoline for a ±10% variation of *B* about the mean, respectively. Thin gray lines correspond to the 0.4 isoline for *B* samples from the variational posterior distribution. For each ice shelf, a histogram is shown of the passive ice shelf area estimated from samples from the posterior, along with the same ±10% lower and upper bounds shown in the maps.

constant, T is temperature, $T_0 \approx -10\,^\circ\mathrm{C}$ is a transition temperature corresponding to a switch in the activation energy for creep, Q_c, and E is an enhancement factor that depends on the ice crystallographic fabric, grain size, damage, and water and impurity content. Therefore, it is possible to decompose the inferred distribution of B into probability distributions for the unknown parameters in the above relation (all parameters except R, the ideal gas constant, a universal constant whose value is well-constrained) (Ranganathan and Minchew, 2022). However, such a decomposition is highly ill-posed and only possible if relatively strong prior constraints are available for the parameters. For example, ice temperatures can be measured at select locations and modeled independently with an appropriate thermomechanical model. The spatial variations in E are likely to be highly correlated with the deformation mode (e.g. simple shear vs extension), which can be well-approximated from surface strain-rates. On the other hand, the activation energy Q_c, which is temperature dependent through T_0, is likely to be relatively uniform within the two separate temperature regimes partitioned by T_0. The differences in expected spatial variation can thus be used as prior constraints when forming the joint posterior distribution of the parameters in Eqn (20).

5.2. Influence of modeling errors

Models of complex physical systems are generally incomplete and do not fully represent all physical processes found in natural settings. Modeling errors will therefore affect inference of parameter values and associated posterior distributions (Kennedy and O'Hagan, 2001). In the case of ice shelves, we have represented ice flow in a continuum mechanics framework with a momentum balance based on the SSA, which assumes that the vertical profile of ice rigidity for an ice column can be represented by its depth-averaged value and that all ice is floating within the ice shelf. The former assumption likely results in inconsequential prediction errors since ocean water provides minimal drag to the base of ice shelves. The assumption of floating ice is violated in areas where ice is locally grounded, which in the 2D synthetic shelf we observed can cause a localized bias in inferred rigidity values around and upstream of the grounded area. These biases arise from the uniform uncertainties, $\sigma_{\mathbf{r}}$, we prescribed in the likelihood model in Eqn (12). In reality, these uncertainties should be scaled according to expected variations in SSA residuals, which when corresponding to un-modeled basal drag can be informed by estimates of flotation height. A simple scaling of the uncertainties follows from consideration of the sensitivity of the forward model to the SSA residuals r_x and r_y, which are nominally zero over ice shelves. Since the forward model used here directly uses the SSA momentum balance, the sensitivity matrix for each SSA residual component is identity, and the total prediction uncertainty is proportional to the uncertainties in the nominal values for the residuals (Duputel and others, 2014). This approach is appropriate when the primary objective is physical interpretation of the distribution of rigidity values (as discussed in the previous section). However, if the primary goal is to use the posterior distribution of rigidity to construct ensembles of ice flow model runs (e.g. to estimate range of probable contributions to sea level rise), then a bias in the distribution for rigidity is acceptable since an increase in ice rigidity will compensate for the missing basal drag for grounded ice.

Another source of modeling uncertainty comes from our use of a conventional inverse method to pre-compute a B_0 field to be used as a prior mean. This strategy nominally reduces uncertainties in ice rigidity near the ice front (Fig. S3). However, the conventional inversion requires specification of a dynamic boundary condition at the ice front based on the hydrostatic pressure provided by the ocean water. In areas where considerable sea ice has formed at the ice front, uncompensated buttressing stress provided by the sea ice will lead to biased estimates of B_0, which can be considered as an additional source of modeling uncertainty. One strategy to account for such uncertainties is to treat

https://doi.org/10.1017/jog.2023.8 Published online by Cambridge University Press

B_0 as a hyperparameter in order to formulate a hyperprior for the rigidity. However, this approach would require a reformulation of the variational lower bound. A simpler approach would be to pre-construct a spatial map of uncertainties in B_0 due to uncertainties in the dynamic boundary condition by repeating the control method inversion for different values of buttressing stress at the ice front. The map of B_0 uncertainties can then be used to inflate the prior variance σ_θ^2 in areas where B_0 is more uncertain. A more direct quantitative approach would be to incorporate the dynamic boundary conditions in an additional loss function for the VGP-generated rigidity. The hydrostatic pressure provided by the ocean water would be treated as a random variable with some uncertainty, which would allow for the construction of a likelihood function that measured the misfit between the probabilistic longitudinal stress and the probabilistic hydrostatic pressure, both of which can be readily computed by the neural network and VGP using automatic differentiation.

5.3. Integration with numerical ice flow models

The SSA momentum balance is the basis of the forward model for our method, which differs from the forward model of predicted ice velocities used in traditional control-method-based inversions and previous studies investigating Bayesian methods for parameter estimation for ice dynamics (see section on related work). If the surface data are noise-free and the boundary conditions at the grounding line and ice front are known perfectly, the two different forward models would result in identical point estimates of ice rigidity. However, even in this ideal scenario, posterior inference with the two different forward models would lead to uncertainties with different spatial variations due to different model sensitivities. The velocity-based forward model is most sensitive to rigidity variations where velocities are higher, usually closer to the ice front. On the other hand, the momentum-based forward model is most sensitive to rigidity variations closer to the grounding line where driving stresses are higher, as well as in high strain-rate regions where driving stresses are lower (Fig. 5).

The different sensitivities will also lead to different systematic errors in the inferred rigidity. As shown with the synthetic 2D ice shelf, inference with the velocity-based forward model can lead to erroneously high rigidity values upstream of pinning points in order to compensate for omission of the basal drag of the pinning points in the ice shelf momentum balance. Alternatively, for the momentum-based forward model, when observed strain rates are relatively large in magnitude and have spatial length scales smaller than the prior length scale, the inferred rigidity will tend to be *lower* in order to minimize stresses in the momentum balance. This situation is more likely when the noise level in the velocity data is relatively high compared to the signal or if the velocity data contain high-frequency signals below the resolution of the momentum balance. In these cases, one possible strategy is to use the predicted velocities from the control method inversion as the ‘observed’ velocities for the variational inference framework. The predicted velocities from the former would have essentially been pre-filtered according to the momentum balance, leading to manageably low initial values of SSA residuals $\mathbf{r}$ for the momentum-based variational inference.

One possible avenue for future work is to integrate the variational inference scheme of Brinkerhoff (2022), which uses a velocity-based forward model, with the methods presented here in order to provide complementary model sensitivities. Furthermore, recent advances in deep learning-based surrogate modeling could significantly improve the computational efficiency of velocity-based forward models by replacing expensive forward solves with much cheaper neural network predictions (Jouvet and others, 2021).

5.4. Uncertainty quantification in physics-informed machine learning and computational considerations

In this work, we focused on estimating a variational distribution for the ice rigidity, conditional on observations of ice surface velocity and thickness and the SSA governing equations for ice flow. The methods presented here are also applicable to similar classes of physics-informed machine learning problems where parameter fields govern the evolution of observable spatiotemporal fields (e.g. Raissi and others, 2020). For example, the variational inference framework could be used to infer a spatially varying thermal diffusivity for a model governed by the heat equation. Time-dependent observations of temperature profiles would be reconstructed by a neural network, $T(x,\ y) = f_{\boldsymbol{\psi}}(x,\ y,\ t)$, and a VGP would be trained to generate samples of the thermal diffusivity at arbitrary spatial coordinates. Furthermore, if temporal gradients are computable, then the variational inference can be extended to time-dependent PDEs. Overall, as long as the governing equations of the physical system can be evaluated at arbitrary coordinates, the methods presented here are transferable. In some cases, the use of the feedforward neural network $f_{\boldsymbol{\psi}}$ is not necessary when high-quality observations (and their necessary gradients) are available or can be pre-processed. Then, one would only have to estimate the parameters of the VGP for estimating the posterior distribution of the parameter field of interest.

From a computational efficiency standpoint, the VGP used for the variational distribution is a marked improvement from standard GPs, but the need to learn a full-rank Gaussian distribution at the inducing points still prevents the VGP from being applicable to very large spatial domains (or, equivalently, model domains where high-spatial resolution of parameter fields is desired). As the number of inducing points exceeds ≈1000, computational and memory requirements become excessive and training efficiency drops dramatically. While training efficiency of VGPs could be improved through the use of second-order optimizers (e.g. Newton- or quasi-Newton-based optimizers), joint training of VGP and neural network parameters would become intractable since neural networks tend to have a significantly larger number of parameters. Therefore, future work must involve the development of alternative models for variational distributions that are suitable for large effective model dimensions (e.g. Brinkerhoff, 2022).

6. Conclusions

In this work, we present a framework for inferring the posterior distribution of ice rheology for large ice shelves in West Antarctica. Motivated by recent advances in physics-informed machine learning and variational inference, the framework utilizes neural networks to reconstruct spatially dense observations of ice surface velocity and thickness, which allows for mesh-free evaluation of surface variable values and associated spatial gradients. At the same time, we task a variational Gaussian Process to predict the mean and covariance of ice rheological parameters for arbitrary spatial coordinates. By using the momentum balance for ice-flow appropriate for ice shelves, we formulate a mapping from parameters (rheology) to observables (residual momentum) that is inherently parallelizable and allows for joint training of the neural networks and variational Gaussian Process using stochastic gradient descent. The training objective utilizes a variational approximation to Bayesian inference, which provides an explicit way to encode prior rheology information in the form of spatial lengthscales (to modulate smoothing of the inferred rheology

https://doi.org/10.1017/jog.2023.8 Published online by Cambridge University Press

field) and range of variation relative to a reference field. For the latter, we show that using a conventional inversion method to estimate a prior mean field can reduce reconstruction errors, which demonstrates a potentially favorable approach to exploring uncertainties in large-scale ice flow models without injecting them into computationally expensive MCMC samplers.

Using these methods, we demonstrate posterior inference of ice rheology for synthetic 1D and 2D ice shelves. We find that rheological uncertainties are lowest where driving stresses and strain rates are higher, corresponding to larger components of the momentum balance and higher levels of ice deformation, which implies more information about ice rheology. Using the synthetic 2D ice shelves, we also demonstrate how the momentum balance-based forward model can help reduce biases in inferred ice rigidity near areas of localized grounding where the shallow-shelf approximation of the momentum balance is violated. Inference of the distribution of ice rigidity values for select West Antarctic ice shelves reveal a wide range of spatial patterns consistent with highly heterogeneous flow environments. Generally, we find softer inferred ice in shear margins, near pinning points, and around visible surface crevasses. Conversely, ice is inferred to be stiffer where bulk ice temperatures are lower and where compressional stresses result in thickening of ice, such as upstream of certain ice rises or where fast-flowing ice streams flow onto slower shelf ice. Finally, using the posterior covariances of rigidity, we generate stochastic predictions of buttressing factors for the West Antarctic ice shelves and show how different flow environments can result in different ranges of passive shelf ice areas, as well as different levels of non-Gaussian behavior. These results demonstrate the utility of site-specific posterior inference for predictive modeling as opposed to assuming uniform lower and upper bounds for an entire ice sheet.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/jog.2023.8.

Data. ITS_LIVE velocity data can be downloaded at https://nsidc.org/apps/itslive/. MEaSUREs velocity data can be downloaded at https://nsidc.org/data/NSIDC-0484/versions/2. BedMachine V2 thickness data can be downloaded at https://nsidc.org/data/NSIDC-0756. Example scripts for variational inference for 1D and 2D ice shelves can be found at https://github.com/bryanvriel/physics-informed-vi. Neural network and variational Gaussian Process models were written using TensorFlow (https://www.tensorflow.org) and TensorFlow Probability (https://www.tensorflow.org/probability).

Acknowledgements. The authors thank two anonymous reviewers and the Scientific Editor Doug Brinkerhoff for their constructive feedback for improving the quality of this work. This research was partially supported by NSFGEO-NERC grant 1853918, the John W. Jarve (1978) Seed Fund for Science Innovation, and the Earl A Killian III (1978) and Waidy Lee Fund.

References

Albergo MS and 7 others (2021) Introduction to normalizing flows for lattice field theory. ArXiv preprint arXiv:2101.08176.

Albrecht T and Levermann A (2014) Fracture-induced softening for large-scale ice dynamics. *The Cryosphere* **8**(2), 587–605. doi:10.5194/tc-8-587-2014

Aschwanden A, Bartholomaus TC, Brinkerhoff DJ and Truffer M (2021) Brief communication: a roadmap towards credible projections of ice sheet contribution to sea level. *The Cryosphere* **15**(12), 5705–5715. doi:10.5194/tc-15-5705-2021

Babaniyi O, Nicholson R, Villa U and Petra N (2021) Inferring the basal sliding coefficient field for the stokes ice sheet model under rheological uncertainty. *The Cryosphere* **15**(4), 1731–1750. doi:10.5194/tc-15-1731-2021

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

Bons PD and 6 others (2018) Greenland ice sheet: higher nonlinearity of ice flow significantly reduces estimated basal motion. *Geophysical Research Letters* **45**(13), 6542–6548. doi:10.1029/2018GL078356

Borstad C, McGrath D and Pope A (2017) Fracture propagation and stability of ice shelves governed by ice shelf heterogeneity. *Geophysical Research Letters* **44**(9), 4186–4194. doi:10.1002/2017GL072648

Borstad CP, Rignot E, Mouginot J and Schodlok MP (2013) Creep deformation and buttressing capacity of damaged ice shelves: theory and application to Larsen C ice shelf. *The Cryosphere* **7**(6), 1931–1947. doi:10.5194/tc-7-1931-2013

Brinkerhoff DJ (2022) Variational inference at glacier scale. *Journal of Computational Physics* **459**, 111095. doi:10.1016/j.jcp.2022.111095

Cuffey KM and Paterson WSB (2010) *The Physics of Glaciers*. Burlington, Massachusetts: Academic Press.

D'Errico J (2012) inpaint_nans, matlab central file exchange. Accessed June 2021 at https://www.mathworks.com/matlabcentral/fileexchange/4551-inpaint_nans.

De Rydt J, Gudmundsson GH, Nagler T and Wuite J (2019) Calving cycle of the Brunt Ice Shelf, Antarctica, driven by changes in ice shelf geometry. *The Cryosphere* **13**(10), 2771–2787. doi:10.5194/tc-13-2771-2019

Dillon JV and 9 others (2017) Tensorflow distributions. ArXiv preprint arXiv:1711.10604.

Duputel Z, Agram PS, Simons M, Minson SE and Beck JL (2014) Accounting for prediction uncertainty when inferring subsurface fault slip. *Geophysical Journal International* **197**(1), 464–482. doi:10.1093/gji/ggt517

Fürst JJ and 6 others (2016) The safety band of Antarctic ice shelves. *Nature Climate Change* **6**(5), 479–482. doi:10.1038/nclimate2912

Gardner A, Fahnestock M and Scambos T (2019) ITS_LIVE regional glacier and ice sheet surface velocities. Data archived at National Snow and Ice Data Center, doi:10.5067/6II6VW8LLWJ7.

Glen J (1958) The flow law of ice: a discussion of the assumptions made in glacier theory, their experimental foundations and consequences. *IASH Publications* **47**(171), e183.

Goldsby D and Kohlstedt DL (2001) Superplastic deformation of ice: experimental observations. *Journal of Geophysical Research: Solid Earth* **106**(B6), 11017–11030. doi:10.1029/2000JB900336

Gopalan G, Hrafnkelsson B, Aḥhalgeirsdóttir G and Pálsson F (2021) Bayesian inference of ice softness and basal sliding parameters at Langjökull. *Frontiers in Earth Science* **9**, 610069. doi:10.3389/feart.2021.610069

Gudmundsson G (2013) Ice-shelf buttressing and the stability of marine ice sheets. *The Cryosphere* **7**(2), 647–655. doi:10.5194/tc-7-647-2013

Harris CR and 25 others (2020) Array programming with NumPy. *Nature* **585**(7825), 357–362. doi:10.1038/s41586-020-2649-2

Hensman J, Fusi N and Lawrence ND (2013) Gaussian processes for big data. ArXiv preprint arXiv:1309.6835, doi:10.48550/arXiv.1309.6835.

Huszár F (2015) How (not) to train your generative model: scheduled sampling, likelihood, adversary?. ArXiv preprint arXiv:1511.05101.

Isaac T, Petra N, Stadler G and Ghattas O (2015) Scalable and efficient algorithms for the propagation of uncertainty from data through inference to prediction for large-scale problems, with application to flow of the Antarctic ice sheet. *Journal of Computational Physics* **296**, 348–368. doi:10.1016/j.jcp.2015.04.047

Jansen D, Luckman A, Kulessa B, Holland PR and King EC (2013) Marine ice formation in a suture zone on the Larsen C ice shelf and its influence on ice shelf dynamics. *Journal of Geophysical Research: Earth Surface* **118**(3), 1628–1640. doi:10.1002/jgrf.20120

Joughin I, Smith BE, Howat IM, Scambos T and Moon T (2010) Greenland flow variability from ice-sheet-wide velocity mapping. *Journal of Glaciology* **56**(197), 415–430. doi:10.3189/002214310792447734

Jouvet G and 5 others (2021) Deep learning speeds up ice flow modelling by several orders of magnitude. *Journal of Glaciology* **68**(270), 651–664. doi:10.1017/jog.2021.120)

Kennedy MC and O'Hagan A (2001) Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**(3), 425–464. doi:10.1111/1467-9868.00294

Khazendar A, Rignot E and Larour E (2009) Roles of marine ice, rheology, and fracture in the flow and stability of the Brunt/Stancomb-Wills Ice Shelf. *Journal of Geophysical Research: Earth Surface* **114**(F04007). doi:10.1029/2008JF001124

Khazendar A, Rignot E and Larour E (2011) Acceleration and spatial rheology of Larsen C ice shelf, Antarctic Peninsula. *Geophysical Research Letters* **38**(L09502). doi:10.1029/2011GL046775

Kingma DP and Ba J (2014) Adam: a method for stochastic optimization, doi:10.48550/ARXIV.1412.6980.

Kulessa B, Jansen D, Luckman AJ, King EC and Sammonds PR (2014) Marine ice regulates the future stability of a large Antarctic ice shelf. *Nature Communications* **5**(1), 1–7. doi:10.1038/ncomms4707

Larour E, Rignot E, Joughin I and Aubry D (2005) Rheology of the Ronne Ice Shelf, Antarctica, inferred from satellite radar interferometry data using an inverse control method. *Geophysical Research Letters* **32**(L05503). doi:10.1029/2004GL021693

Larour E, Rignot E, Poinelli M and Scheuchl B (2021) Physical processes controlling the rifting of Larsen C Ice Shelf, Antarctica, prior to the calving of iceberg A68. *Proceedings of the National Academy of Sciences* **118**(40). doi:10.1073/pnas.2105080118

MacAyeal DR (1989) Large-scale ice flow over a viscous basal sediment: theory and application to Ice Stream B, Antarctica. *Journal of Geophysical Research: Solid Earth* **94**(B4), 4071–4087. doi:10.1029/JB094iB04p04071

MacAyeal DR (1993) A tutorial on the use of control methods in ice-sheet modeling. *Journal of Glaciology* **39**(131), 91–98.

MacAyeal DR, Rignot E and Hulbe CL (1998) Ice-shelf dynamics near the front of the Filchner-Ronne Ice Shelf, Antarctica, revealed by SAR interferometry: model/interferogram comparison. *Journal of Glaciology* **44**(147), 419–428.

Matthews AGdG, Hensman J, Turner R and Ghahramani Z (2016) On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. In *Artificial Intelligence and Statistics*. Cadiz, Spain: PMLR, pp. 231–239.

Millstein JD, Minchew BM and Pegler SS (2022) Ice viscosity is more sensitive to stress than commonly assumed. *Communications Earth & Environment* **3**(1), 1–7. doi:10.1038/s43247-022-00385-x

Mitcham T, Gudmundsson GH and Bamber JL (2022) The instantaneous impact of calving and thinning on the Larsen C Ice Shelf. *The Cryosphere* **16**(3), 883–901. doi:10.5194/tc-16-883-2022

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**(21), 11051–11061. doi:10.1002/2017GL074954

Morlighem M and 9 others (2020) Deep glacial troughs and stabilizing ridges unveiled beneath the margins of the Antarctic ice sheet. *Nature Geoscience* **13**(2), 132–137.

Mouginot J, Rignot E, Scheuchl B and Millan R (2017) Comprehensive annual ice sheet velocity mapping using Landsat-8, Sentinel-1, and RADARSAT-2 data. *Remote Sensing* **9**(4), 364. doi:10.3390/rs9040364

Mouginot J, Scheuchl B and Rignot E (2012) Mapping of ice motion in Antarctica using synthetic-aperture radar data. *Remote Sensing* **4**(9), 2753–2767. ISSN 2072-4292. doi:10.3390/rs4092753

Nias IJ, Cornford SL, Edwards TL, Gourmelen N and Payne AJ (2019) Assessing uncertainty in the dynamical ice response to ocean warming in the Amundsen Sea Embayment, West Antarctica. *Geophysical Research Letters* **46**(20), 11253–11260. doi:10.1029/2019GL084941

Nick FM, Van der Veen CJ, Vieli A and Benn DI (2010) A physically based calving model applied to marine outlet glaciers and implications for the glacier dynamics. *Journal of Glaciology* **56**(199), 781–794. doi:10.3189/002214310794457344

Penny W, Kiebel S and Friston K (2007) Chapter 24 – variational bayes. In K Friston, J Ashburner, S Kiebel, T Nichols and W Penny (eds), *Statistical Parametric Mapping*. London: Academic Press, pp. 303–312, ISBN 978-0-12-372560-8. doi:10.1016/B978-012372560-8/50024-3.

Petra N, Martin J, Stadler G and Ghattas O (2014) A computational framework for infinite-dimensional Bayesian inverse problems, part II: stochastic Newton MCMC with application to ice sheet flow inverse problems. *SIAM Journal on Scientific Computing* **36**(4), A1525–A1555. doi:10.1137/130934805

Phan D, Pradhan N and Jankowiak M (2019) Composable effects for flexible and accelerated probabilistic programming in NumPyro. ArXiv preprint arXiv:1912.11554.

Pralong MR and Gudmundsson GH (2011) Bayesian estimation of basal conditions on Rutford Ice Stream, West Antarctica, from surface data. *Journal of Glaciology* **57**(202), 315–324. doi:10.3189/002214311796406004

Qi C and Goldsby DL (2021) An experimental investigation of the effect of grain size on 'dislocation creep' of ice. *Journal of Geophysical Research: Solid Earth* **126**(9), e2021JB021824. doi:10.1029/2021JB021824.

Quinonero-Candela J and Rasmussen CE (2005) A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research* **6**, 1939–1959. doi:10.5555/1046920.1194909

Raissi M, Perdikaris P and Karniadakis GE (2019) Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics* **378**, 686–707. doi:10.1016/j.jcp.2018.10.045

Raissi M, Yazdani A and Karniadakis GE (2020) Hidden fluid mechanics: learning velocity and pressure fields from flow visualizations. *Science* **367**(6481), 1026–1030. doi:10.1126/science.aaw4741

Ranganathan M and Minchew B (2022) The effect of uncertainties in creep activation energies on modeling ice flow and deformation. EarthArXiv.

Rasmussen CE (2003) Gaussian processes in machine learning. In *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer, pp. 63–71.

Reese R, Gudmundsson GH, Levermann A and Winkelmann R (2018) The far reach of ice-shelf thinning in Antarctica. *Nature Climate Change* **8**(1), 53–57. doi:10.1038/s41558-017-0020-x

Rezende D and Mohamed S (2015) Variational inference with normalizing flows. In *International Conference on Machine Learning*. Lille, France: PMLR, pp. 1530–1538.

Riel B, Minchew B and Bischoff T (2021) 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), e2021MS002621. doi:10.1029/2021MS002621.

Rignot E, Mouginot J and Scheuchl B (2011) Ice flow of the Antarctic Ice Sheet. *Science* **333**(6048), 1427–1430. doi:10.1126/science.1208336

Rue H (2001) Fast sampling of Gaussian markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**(2), 325–338. doi:10.1111/1467-9868.00288

Schlegel NJ and 8 others (2018) Exploration of Antarctic Ice Sheet 100-year contribution to sea level rise and associated model uncertainties using the ISSM framework. *The Cryosphere* **12**(11), 3511–3534. doi:10.5194/tc-12-3511-2018

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**(7), 4593–4616. doi:10.5194/gmd-14-4593-2021

Still H, Campbell A and Hulbe C (2019) Mechanical analysis of pinning points in the Ross ice shelf, Antarctica. *Annals of Glaciology* **60**(78), 32–41. doi:10.1017/aog.2018.31

Still H and Hulbe C (2021) Mechanics and dynamics of pinning points on the Shirase Coast, West Antarctica. *The Cryosphere* **15**(6), 2647–2665. doi:10.5194/tc-15-2647-2021

Titsias M (2009) Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*. Clearwater, Florida: PMLR, pp. 567–574.

Appendix A. Probabilistic formulation of physics-informed ice rigidity inference

We aim to formulate an objective function for estimating a probability distribution for ice rigidity fields $\theta(\mathbf{x})$ on ice shelves, conditioned on noisy observations of surface velocity and ice thickness (or thickness derived from surface elevations using a hydrostatic assumption). Additionally, we aim to construct mean fields (i.e. smoothed fields) for ice surface velocity and thickness that are consistent with the observations and their uncertainties while also minimizing the SSA momentum balance residuals for random samples of θ. To that end, let us first define an augmented parameter vector for a given ice shelf, $\mathbf{m} = [\theta, \hat{\mathbf{d}}]$, where $\hat{\mathbf{d}} = [\hat{u}, \hat{v}, \hat{h}]$ is the vector of reconstructed surface observations. As stated in the main text, $\theta = \theta(\mathbf{x})$ and $\hat{\mathbf{d}} = \hat{\mathbf{d}}(\mathbf{x})$ are implied, i.e. all scalar and vector quantities in this discussion are spatial fields spanning the ice shelf. Furthermore, we make no assumption on the parameterization of θ and $\hat{\mathbf{d}}$ as this section is focused on formulating the probabilistic relationships between modeled and observed fields. For our observations, we have $\mathbf{d} = [u, v, h]$, as well as a vector of pseudo-observations corresponding to the momentum balance residuals in Eqns (2 and 3), $\mathbf{r} = [r_x, r_y]$. Recall that in our physics-informed machine learning framework, $\mathbf{r} = [0, 0]$ as we seek to learn θ that minimizes the SSA momentum balance residuals for a given set of surface predictions $\hat{\mathbf{d}}$.

We use Bayes' Theorem to obtain an expression for the posterior distribution of $\mathbf{m}$ given our combined observations:

$$p(\mathbf{m}|\mathbf{d}, \mathbf{r}) \propto p(\mathbf{d}|\mathbf{r}, \mathbf{m})p(\mathbf{r}|\mathbf{m})p(\mathbf{m}). \tag{A1}$$

The first two terms on the right-hand side correspond to the joint data likelihood, which encodes the probability of having 'observed' $\mathbf{d}$ and $\mathbf{r}$ for a given

$\mathbf{m}$ within the SSA equations. The third term corresponds to the prior distribution which encodes our prior knowledge on $\mathbf{m}$ values without having seen any observations. We can expand and simplify the above posterior distribution by utilizing what we know about the dependencies between the augmented parameters and observations in the physics-informed machine learning framework. Firstly, the surface observations $\mathbf{d}$ are only being reconstructed by a regression model (e.g. the neural network $f_{\boldsymbol{\psi}}$) that does not explicitly depend on θ; thus, $\mathbf{d}$ is only dependent on $\hat{\mathbf{d}}$. Similarly, the prior can be expressed as $p(\mathbf{m}) = p(\theta)p(\hat{\mathbf{d}})$ since the predictions $\hat{\mathbf{d}}$ are again generated by an independent regression model and are independent of the prior values for θ. Furthermore, we can prescribe a uniform prior for $\hat{\mathbf{d}}$ as no explicit regularization is imposed on its values. We can therefore write the posterior distribution for θ and $\hat{\mathbf{d}}$ as:

$$p(\theta,\ \hat{\mathbf{d}}|\mathbf{d},\ \mathbf{r}) \ \propto\ p(\mathbf{d}|\hat{\mathbf{d}})p(\mathbf{r}|\theta,\ \hat{\mathbf{d}})p(\theta), \tag{A2}$$

which matches Eqn (5) in the main text.

A.1. Variational lower bound to posterior distribution

As discussed in the main text, we use a variational inference framework wherein we aim to construct an approximating, variational distribution $q(\theta,\ \hat{\mathbf{d}})$ that is minimally divergent from the true posterior $p(\theta,\ \hat{\mathbf{d}}|\mathbf{d},\ \mathbf{r})$. Generally, the variational distribution is a parametric distribution that is easy to sample from and has a likelihood that can be easily evaluated. Therefore, choosing an appropriate variational distribution will depend on the desired level of computational complexity (e.g. number of trainable parameters) and the expressiveness of the distribution for capturing the relevant features and statistics of the target posterior. Expected features such as skewness, multiple modes, correlations between parameters, etc., will impact the choice of variational distribution. Posterior distributions with highly complex probability topologies tend to be ill-suited for variational inference and will likely require MCMC-based methods. Even for less complex posteriors, variational distributions that are overly simple, such as an independent normal for each parameter, will fail to capture parameter correlations and can lead to significantly underestimated posterior variances (Blei and others, 2017). In this work, as discussed in Appendix B, we choose to use multivariate normal distributions in which the elements of the mean vector and covariance matrix are the learnable parameters.

A commonly used metric for quantifying the similarity between variational and target distributions is the Kullback–Leibler (KL) divergence, which for our problem is defined as:

$$\mathcal{KL}\left[q(\theta,\ \hat{\mathbf{d}})\|p(\theta,\ \hat{\mathbf{d}}|\mathbf{d},\ \mathbf{r})\right] = \int q(\theta,\ \hat{\mathbf{d}}) \log \frac{q(\theta,\ \hat{\mathbf{d}})}{p(\theta,\ \hat{\mathbf{d}}|\mathbf{d},\ \mathbf{r})}\ d\theta d\hat{\mathbf{d}}. \tag{A3}$$

It is important to note that the above equation is referred to as the *reverse*-KL-divergence, denoted as $\mathcal{KL}[q\|p]$ for brevity, as opposed to the *forward*-KL-divergence, $\mathcal{KL}[p\|q]$. As documented in previous studies (e.g. Huszár, 2015; Albergo and others, 2021), training the variational distribution with a reverse-KL divergence loss encourages mode-seeking behavior (but does not enforce it like the Laplace approximation), i.e. the distribution will be centered closer to regions with the highest posterior probabilities. A forward-KL divergence tends to encourage mean-seeking behavior. For inference of rigidity with posterior distributions that are generally unimodal and well-approximated by Gaussian distributions, the two KL-divergences should lead to similar variational distributions. However, the reverse-KL formulation leads to a more computationally efficient variational lower bound, which we now turn to.

The posterior $p(\theta,\ \hat{\mathbf{d}}|\mathbf{d},\ \mathbf{r})$ in the denominator of the second term in Eqn (A3) causes the integration over all θ to be intractable. In this case, it can be shown (e.g. Titsias, 2009; Hensman and others, 2013; Matthews and others, 2016; Blei and others, 2017) that minimization of the KL-divergence can be replaced by maximization of a stochastic variational lower bound, often referred to as the Evidence Lower Bound (ELBO):

$$\begin{aligned}\mathrm{ELBO}(\theta,\ \hat{\mathbf{d}}) &= E_{\theta,\hat{\mathbf{d}}\sim q(\theta,\hat{\mathbf{d}})}\left[\log p(\mathbf{d}|\hat{\mathbf{d}}) + \log p(\mathbf{r}|\theta,\ \hat{\mathbf{d}})\right]\\ &\quad - \mathcal{KL}\left[q(\theta,\ \hat{\mathbf{d}})\|p(\theta,\ \hat{\mathbf{d}})\right].\end{aligned} \tag{A4}$$

The first term of the ELBO is the expected value of the joint data log-likelihood, integrated over θ and $\hat{\mathbf{d}}$. The second term is the KL-divergence between the variational distribution and the prior distribution. Thus, the ELBO can be interpreted as an optimization objective that encourages the variational distribution $q(\theta,\ \hat{\mathbf{d}})$ to produce samples of θ and $\hat{\mathbf{d}}$ that best fit the surface observations and minimize the SSA momentum balance residuals while also maintaining consistency with the prior distribution. We can further decompose the variational distribution to be $q(\theta,\ \hat{\mathbf{d}}) = q(\theta)q(\hat{\mathbf{d}})$, i.e. the rigidity field and surface predictions are modeled independent of each other. The independence allows us to use different machine learning models for predicting the surface variables and rigidity field. Here, we use a neural network $f_{\boldsymbol{\psi}}$ for the former and a variational Gaussian Process for the latter. Moreover, as the network $f_{\boldsymbol{\psi}}$ only predicts the mean $\hat{\mathbf{d}}$ values, $q(\hat{\mathbf{d}})$ is equivalent to a δ-distribution, e.g. $\int x\delta(x-\hat{x})\,\mathrm{d}x = \hat{x}$. Additionally, we prescribe a uniform prior for $\hat{\mathbf{d}}$ such that $p(\hat{d}_i) = \frac{1}{b_i - a_i}$ where $a_i \le \hat{d}_i \le b_i$ for the i-th component of $\hat{\mathbf{d}}$. These formulations allows the KL-divergence in Eqn (A.4) to be simplified in the following manner:

$$\begin{aligned}\mathcal{KL}\left[q(\theta,\ \hat{\mathbf{d}})\|p(\theta,\ \hat{\mathbf{d}})\right] &= \mathcal{KL}\left[q(\theta)q(\hat{\mathbf{d}})\|p(\theta)p(\hat{\mathbf{d}})\right]\\ &= \int q(\theta)q(\hat{\mathbf{d}}) \log\left(\frac{q(\theta)q(\hat{\mathbf{d}})}{p(\theta)p(\hat{\mathbf{d}})}\right)\mathrm{d}\theta\,\mathrm{d}\hat{\mathbf{d}}\\ &= \mathcal{C}\cdot\int q(\theta)\log\left(\frac{q(\theta)}{p(\theta)}\right)\mathrm{d}\theta\\ &= \mathcal{C}\cdot\mathcal{KL}\left[q(\theta)\|p(\theta)\right],\end{aligned}$$

where $\mathcal{C}$ is a constant resulting from the uniform prior $p(\hat{\mathbf{d}})$. Altogether, the ELBO can be written as:

$$\mathrm{ELBO}(\theta,\ \hat{\mathbf{d}}) = E_{\theta\sim q(\theta)}\left[\log p(\mathbf{r}|\theta,\ \hat{\mathbf{d}})\right] + \log p(\mathbf{d}|\hat{\mathbf{d}}) - \mathcal{KL}\left[q(\theta)\|p(\theta)\right], \tag{A5}$$

which is Eqn (7) in the main text. Note that we exclude the constant $\mathcal{C}$ from the above equation, but generally, one can include a multiplicative factor on the remaining KL-divergence to tune the regularization effect of the prior on the posterior inference.

The expectation operator on the log-likelihood of $p(\mathbf{r}|\theta,\ \hat{\mathbf{d}})$ can be numerically estimated using either a Monte Carlo integration or a quadrature-based integration. For high-dimensional integrals, Monte Carlo methods are favorable over quadrature methods, which require an exponential amount of quadrature points in the number of dimensions. However, the SSA residual vector $\mathbf{r}$ can be evaluated at each point on the ice shelf independently, provided that gradients of θ and $\hat{\mathbf{d}}$ are available in order to compute relevant stress gradients. Both the neural network and variational Gaussian Process models permit the computation of spatial gradients on a coordinate-by-coordinate basis by utilizing automatic differentiation. Therefore, the expectation of $\log p(\mathbf{r}|\theta,\ \hat{\mathbf{d}})$ can be decomposed into a series of one-dimensional integrals. We follow the approach of Dillon and others (2017) and evaluate each 1D integral with a Gauss–Hermite quadrature with 3 quadrature points, which is exact for 1D Gaussian log-likelihoods:

$$E_{\theta\sim q(\theta)}\left[\log p(\mathbf{r}|\theta,\ \hat{\mathbf{d}})\right] \approx \frac{1}{\sqrt{\pi}}\sum_{i=1}^{n} w_i \log p(\mathbf{r}|\sqrt{(}2)\hat{\sigma}_\theta g_i + \hat{\mu}_\theta,\ \hat{\mathbf{d}}), \tag{A6}$$

where n is the number of quadrature points, w_i are the quadrature weights for the quadrature grid points g_i, $\hat{\sigma}_\theta$ is the predicted posterior standard deviation for θ, and $\hat{\mu}_\theta$ is the predicted posterior mean for θ. Here, we compute w_i and g_i using the NumPy function `np.polynomial.hermite.hermgauss` (Harris and others, 2020).

Appendix B. Variational Gaussian processes

In this work, we use a variational Gaussian Process (VGP) to model the variational distribution $q(\theta)$ introduced in Appendix A. Gaussian Processes (GPs) are a class of machine learning models that describe a set of random variables, $\{f(\mathbf{x})|\mathbf{x} \in \mathcal{X}\}$, where $\mathbf{x}$ is a finite set of index points (i.e. spatial or temporal coordinates) from $\mathcal{X}$. GPs assume that every finite collection of random variables follows a multivariate normal distribution (Rasmussen, 2003). Thus, to describe a GP, one needs to specify a mean function, $\mu(\mathbf{x})$, and a covariance, or kernel, function $k(\mathbf{x},\ \mathbf{x}')$ that describes the pairwise similarity between index points $\mathbf{x}$ and $\mathbf{x}'$. In this work, we use a squared exponential kernel function for both posterior inference (Eqn 9) and for constructing the prior distribution (Eqn 13). The difference in usage between the two is that for posterior inference, σ and L are tunable hyperparameters, whereas for the prior, σ_θ and L_θ are fixed values.

https://doi.org/10.1017/jog.2023.8 Published online by Cambridge University Press

Given a vector of training data, $\mathbf{y} = f(\mathbf{x}_t) + \boldsymbol{\epsilon}$, where $\mathbf{x}_t$ represent the training index points and $\boldsymbol{\epsilon}$ represents independent Gaussian observation noise, the mean and covariance functions can be inferred via Bayesian inference in order to to estimate a GP posterior. For predictive points $\mathbf{x}$, i.e. index points not used for training, the closed-form expressions for the GP posterior mean, $\mu_p(\mathbf{x})$, and covariance, $k_p(\mathbf{x},\ \mathbf{x}')$, functions are:

$$\begin{aligned}\mu_p(\mathbf{x}) &= K_{\mathbf{x}\mathbf{x}^t}(\boldsymbol{\epsilon} + K_{\mathbf{x}^t\mathbf{x}^t})^{-1}\mathbf{y},\\ k_p(\mathbf{x},\mathbf{x}') &= k(\mathbf{x},\mathbf{x}') - K_{\mathbf{x}\mathbf{x}^t}(\boldsymbol{\epsilon} + K_{\mathbf{x}^t\mathbf{x}^t})^{-1}K_{\mathbf{x}^t\mathbf{x}'},\end{aligned} \tag{B7}$$

where we adopt the shorthand notation $K_{\mathbf{x}\mathbf{x}'} = k(\mathbf{x},\ \mathbf{x}')$. The above formulations therefore allow for generation of samples of $f(\mathbf{x})$ at any arbitrary index points $\mathbf{x}$ by computing a mean and covariance matrix for a multivariate normal distribution.

For a training set of size N, the time complexity of GP inference scales as $\mathcal{O}(N^3)$ with a storage complexity of $\mathcal{O}(N^2)$, primarily due to the need for computing and inverting a covariance matrix of size $N \times N$ in Eqn (B7). Therefore, when N exceeds several thousand data points, the computational requirements become excessive, even for modern computing platforms. VGPs address this limitation by building a low-dimensional representation of the dataset by introducing a set of *inducing variables*, which are indexed by a set of M coordinates, $\mathbf{z}$, which live in the same space as $\mathcal{X}$ (Quinonero-Candela and Rasmussen, 2005; Titsias, 2009). Importantly, $M \ll N$, such that the inducing variables serve as a much smaller set of latent observations that are used in the above GP inference equations rather than the full training data. In the formulation presented by Hensman and others (2013) which we use here, the locations of $\mathbf{z}$ are unknown and must be estimated during training of the VGP. Two key approximations are used for the inducing variables: (1) function values at non-inducing points, $\mathbf{x}$, are mutually independent given function values at the inducing index points, $\mathbf{z}$; and (2) the posterior distribution $p(f(\mathbf{z})|\mathbf{y})$, which can be expensive to compute for large N, can be approximated with a multivariate normal distribution with mean $\mathbf{m} \in \mathbb{R}^{M\times 1}$ and covariance $\mathbf{S} \in \mathbb{R}^{M\times M}$. The VGP posterior predictive mean, $\mu_q(\mathbf{x})$, and covariance, $k_q(\mathbf{x},\ \mathbf{x}')$, functions then take the form:

$$\begin{aligned}\mu_q(\mathbf{x}) &= K_{\mathbf{xz}}K_{\mathbf{zz}}^{-1}\mathbf{m},\\ k_q(\mathbf{x},\mathbf{x}') &= k(\mathbf{x},\mathbf{x}') - K_{\mathbf{xz}}K_{\mathbf{zz}}^{-1}K_{\mathbf{zx}'} + K_{\mathbf{xz}}\mathbf{B}K_{\mathbf{zx}'},\end{aligned} \tag{B8}$$

where $\mathbf{B} = K_{\mathbf{zz}}^{-1}\mathbf{S}K_{\mathbf{zz}}^{-1}$. The mean and covariance are then used to form the variational distribution for the ice rigidity, $q(\theta(\mathbf{x})) = \mathcal{N}(\mu_q(\mathbf{x}),\ k_q(\mathbf{x},\ \mathbf{x}))$. The above equations now have a time complexity of $\mathcal{O}(NM^2)$ and a storage complexity of $\mathcal{O}(NM)$. Note that one still cannot evaluate the variational kernel function for exceedingly large $\mathbf{x}$, which may be the case for large ice shelves in Antarctica. For this reason, we predict covariance matrices for finite blocks of the modeling domain in order to seed a Gibbs sampler, as explained in the Methodology section of the main text Eqns (16a–16c).

The total set of trainable variables for the VGP, $\boldsymbol{\phi}$, consists of the inducing point locations, $\mathbf{z}$, the inducing point mean values, $\mathbf{m}$, the full-rank inducing point covariance matrix, $\mathbf{S}$, and the hyperparameters of the base kernel function (σ and L of the squared exponential kernel). As we only need to estimate the lower-triangular portion of $\mathbf{S}$, the number of covariance parameters becomes $\frac{M(M+1)}{2}$. Furthermore, since the diagonal values of $\mathbf{S}$ must strictly be positive, we apply a sotfplus transformation followed by small shift of $1\mathrm{e}^{-5}$ in order to avoid numerical issues (see the FillScaleTriL bijector implemented in TensorFlow Probability). The number of inducing points, M, is a prescribed parameter that depends on the complexity of the signal being inferred. For example, for the ice shelves studied in this work, we find that M between 300 and 1000 provided a reasonable trade-off between computational complexity and resolution of the inferred $\theta(\mathbf{x})$. The optimal value of M will also depend on the choice of the length scale L_θ for the prior distribution. Smaller L_θ will lead to $\theta(\mathbf{x})$ with more spatial variability, which will require a larger M.

Optimization of the variational parameters is achieved by inserting the variational distribution $q(\theta(\mathbf{x})) = \mathcal{N}(\mu_q(\mathbf{x}),\ k_q(\mathbf{x},\ \mathbf{x}))$ into the ELBO in Eqn (10). Note that the last term of the ELBO is the KL-divergence of $q(\theta)$ from the prior evaluated at the inducing points, $\mathbf{z}$, which is consistent with the VGPs formulated by Titsias (2009) and Hensman and others (2013). With this formulation, q can be constructed using the variational mean and covariance directly, $q(\theta(\mathbf{z})) = \mathcal{N}(\mathbf{m},\ \mathbf{S})$ (note how Eqn (B8) simplifies to $\mathbf{m}$ and $\mathbf{S}$ when $K_{\mathbf{xz}} = K_{\mathbf{zz}}$). Empirically, we found that $\mathcal{KL}[q_{\boldsymbol{\varphi}}(\theta)\|p(\theta)]$ can also be well-approximated by using Eqn (B8) to construct $q(\theta(\mathbf{x}_c))$ and $p(\theta(\mathbf{x}_c))$ at a mini-batch of collocation coordinates $\mathbf{x}_c$, assuming the mini-batch of coordinates are roughly uniformly distributed throughout the modeling domain. This approach can improve training time when the number of inducing points is significantly larger than the batch size.

https://doi.org/10.1017/jog.2023.8 Published online by Cambridge University Press