Journal of Glaciology



Article

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

Received: 19 June 2020
Revised: 25 November 2020
Accepted: 9 December 2020
First published online: 18 January 2021

Keywords:
Basal ice; glacier hydrology; ice-sheet modeling; ice velocity; subglacial processes

Author for correspondence:
Douglas Brinkerhoff,
E-mail: douglas1.brinkerhoff@umontana.edu

© The Author(s), 2021. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

cambridge.org/jog

Constraining subglacial processes from surface velocity observations using surrogate-based Bayesian inference

Douglas Brinkerhoff[1], Andy Aschwanden[2] and Mark Fahnestock[2]

[1]Department of Computer Science, University of Montana, Missoula, MT, USA and [2]Geophysical Institute, University of Alaska Fairbanks, Fairbanks, AK, USA

Abstract

Basal motion is the primary mechanism for ice flux in Greenland, yet a widely applicable model for predicting it remains elusive. This is due to the difficulty in both observing small-scale bed properties and predicting a time-varying water pressure on which basal motion putatively depends. We take a Bayesian approach to these problems by coupling models of ice dynamics and subglacial hydrology and conditioning on observations of surface velocity in southwestern Greenland to infer the posterior probability distributions for eight spatially and temporally constant parameters governing the behavior of both the sliding law and hydrologic model. Because the model is computationally expensive, characterization of these distributions using classical Markov Chain Monte Carlo sampling is intractable. We skirt this issue by training a neural network as a surrogate that approximates the model at a sliver of the computational cost. We find that surface velocity observations establish strong constraints on model parameters relative to a prior distribution and also elucidate correlations, while the model explains 60% of observed variance. However, we also find that several distinct configurations of the hydrologic system and stress regime are consistent with observations, underscoring the need for continued data collection and model development.

Introduction

Glaciers and ice sheets convert potential energy in the form of accumulated ice at high elevations into heat, either by viscous dissipation within the ice itself or by frictional dissipation at the interface between the ice and the underlying bedrock or sediment. This latter process, hereafter referred to as 'sliding', is responsible for >90% of observed surface velocity over much of Greenland, even in regions that are not particularly fast flowing (Maier and others, 2019). Because variations in ice flow dynamics make up >50% of contemporary ice loss in Greenland (Mouginot and others, 2019), correctly modeling sliding is as critical to predicting future Greenland mass loss as having reliable climate models. Ensemble modeling of Greenland's future has shown that uncertainty in ice dynamics accounts for between 26 and 53% of variance in sea level rise projections over the next century (Aschwanden and others, 2019).

Observations (e.g. Iken and Bindschadler, 1986) and theoretical considerations (e.g. Weertman, 1964; Lliboutry, 1968; Fowler, 1979) suggest that basal sliding depends on basal effective pressure. However, explicitly modeling basal effective pressure – and more generally, modeling the subglacial hydrologic system – remains among the most significant open problems in glacier dynamics. The difficulty results from a discrepancy in spatial and temporal scales between the physics driving sliding and water flux versus the scale of glaciers and ice sheets: physics at the bed occur on the order of a few meters with characteristic timescales of minutes, while relevant timescales for ice-sheet evolution occur over kilometers and years. To upscale glacier hydrology to a scale relevant to the overlying ice, a variety of approximations have been proposed, including different physical phenomena thought to be morphologically relevant such as a continuum approximation of linked cavities (Bueler and van Pelt, 2015), a lattice model of conduits or a combination thereof (Werder and others, 2013; De Fleurian and others, 2014; Hoffman and others, 2016; Downs and others, 2018; Sommers and others, 2018). However, validating the models of sliding and hydrology remains elusive, partly due to potential model misspecification, but also due to a lack of sufficient observational constraints on model parameters such as hydraulic conductivity of different components of the subglacial system, characteristic length scales of bedrock asperities and the scaling between effective pressure and basal shear stress.

Previous assimilation of surface velocity observations

The above challenges are not new, and ice-sheet modelers have used geophysical inversion methods (e.g. Parker and Parker, 1994) in glaciological applications to circumvent them for over two decades (e.g. MacAyeal, 1993; Morlighem and others, 2010; Gillet-Chaulet and others, 2012; Favier and others, 2014; Joughin and others, 2014; Cornford and others, 2015). Commonly, a linear relationship between basal shear stress and velocity is adopted,

and then surface velocities are inverted for a spatially varying basal friction field such that the resulting surface velocities are close to observations. This approach lumps all basal processes into one field, a frictional parameter that varies in space while ignoring temporal variability, exchanging the capability of longer-term predictive power for spatial fidelity to observations at an instant.

Several variants on this approach exist. For example Habermann and others (2012) performed the above procedure with a pseudo-plastic power law. Larour and others (2014) assimilated surface altimetry data into re-constructions of transient ice flow. The novelty of their approach was that surface mass balance and basal friction were determined in time as well as space, resulting in adjusted modeled surface heights and time-varying velocities that best fit existing altimetry. Such an approach allows for a better quantification of time-evolving basal and surface processes and a better understanding of the physical processes currently missing in transient ice-flow models. Their study also demonstrated that large spatial and temporal variability is required in model characteristics such as basal friction. However, for prognostic modeling, such approaches cannot be applied because we cannot assimilate future observations. As such, a middle ground between purely empirical and local process modeling must be found.

Several recent studies have taken this approach. Pimentel and Flowers (2011) used a coupled flowband model of glacier dynamics and hydrology to model the propagation of meltwater-induced acceleration across a synthetic Greenland-esque domain, and established that the presence of channels can substantially reduce the sensitivity of the system to fast influxes of meltwater. Hoffman and others (2016) showed that for a 3-D synthetic domain based on West Greenland, a weakly-connected drainage system helps to explain the temporal signal of velocity in the overlying ice. The previous two studies, although not formally assimilating observations, compared their model results to observations in an effort to validate their qualitative results. Minchew and others (2016) directly inverted surface velocities at Hofsjokull Ice Cap for a spatially varying basal shear stress, and in conjunction with a Coulomb friction law, inferred the distribution of effective pressure. Brinkerhoff and others (2016) used a Bayesian approach to condition a 0-D model of glacier hydrology and sliding on surface velocity and terminus flux observations to infer probability distributions over unknown ice dynamics and hydrologic model parameter. Although not coupled to an ice dynamics model, Irarrazaval and others (2019) present a Bayesian inference over the lattice model of Werder and others (2013), constraining the position and development of subglacial channels from observations of water pressure and tracer transit times. Aschwanden and others (2016) demonstrated that outlet glacier flow can be captured using a simple local model of subglacial hydrology, but further improvements are required in the transitional zone with speeds of 20–100 m a^{-1}. This disagreement between observed and simulated speeds most likely arises from inadequacies in parameterizing sliding and subglacial hydrology. Finally and notably, Koziol and Arnold (2018) inverted velocity observations from West Greenland to determine a spatially-varying traction coefficient after attenuation by effective pressure derived from a hydrologic model.

Our approach

In this study, we seek to expand on previous approaches by coupling a state of the art subglacial hydrology model to a 2.5-D (map plane plus an ansatz spectral method in the vertical dimension) model of ice dynamics through a general sliding law (hereafter referred to as the *high-fidelity model*), and to then infer the distribution of practically unobservable model parameters such that the ice surface velocity predicted by the model is statistically consistent with spatially explicit observations over a region in western Greenland. Throughout the study, *we assume spatially and temporally constant parameters in the hydrologic and sliding model* so that spatial and temporal variability in basal shear stress is only attributable to differences in modeled physical processes.

It is likely that there exists substantial non-uniqueness in model parameter solutions. Different controlling factors in the hydrology model may compensate for one another, as may parameters in the sliding law: for, example, the basal traction coefficient could be made lower if sheet conductivity is made higher, leading to a lower mean effective pressure. In order to fully account for these tradeoffs and to honestly assess the amount of information that can be gained by looking solely at surface velocity, we adopt a Bayesian approach (e.g. Tarantola, 2005) in which we characterize the complete joint posterior probability distribution over the parameters, rather than point estimates.

Inferring the joint posterior distribution is not analytically tractable, so we rely on numerical sampling via a Markov Chain Monte Carlo (MCMC) method instead. Similar inference in a coupled hydrology-dynamics model has been done before (Brinkerhoff and others, 2016). However, in the previous study the model was spatially averaged in all dimensions, and thus inference was over a set of coupled ordinary differential equations. Here, we work with a model that remains a spatially explicit and fully coupled system of partial differential equations. As such, the model is too expensive for a naive MCMC treatment. To skirt this issue, we create a so-called *surrogate model*, which acts as a computationally efficient approximation to the expensive coupled high-fidelity model. We note that this idea is not new to glaciology; Tarasov and others (2012) used a similar approach to calibrate parameters of paleoglaciological models based on chronological indicators of deglaciation.

To construct the surrogate, we run a 5000 member ensemble of multiphysics models through time, each with parameters drawn from a prior distribution, to produce samples of the modeled annual average velocity field. This is computationally tractable because each of these model runs is independent, and thus can be trivially parallelized. We reduce the dimensionality of the space of these model outputs through a principal component analysis (PCA) (Shlens, 2014), which identifies the key modes of model variability. We refer to these modes as *eigenglaciers*, and (nearly) any velocity field producible by the high-fidelity model is a linear combination thereof. To make use of this decomposition, we train an artificial neural network (Goodfellow and others, 2016) to control the coefficients of these eigenglaciers as a function of input parameter values, yielding a computationally trivial map from parameter values to a distributed velocity prediction consistent with the high-fidelity model. Unfortunately, neural networks are high variance maps, which is to say that the function is sensitive to the choice of training data. To reduce this variance (and to smooth the relationship between parameters and predictions), we employ a Bayesian bootstrap aggregation approach (Breiman, 1996; Clyde and Lee, 2001) to generate a committee of surrogate models, which are averaged to yield a prediction.

Surrogate in hand, we use the manifold Metropolis-adjusted Langevin algorithm (mMALA; Girolami and Calderhead, 2011) to draw a long sequence of samples from the posterior probability distribution of the model parameters. mMALA utilizes both gradient and Hessian information that are easily computed from the surrogate to efficiently explore the posterior distribution. Because the surrogate model itself is based on a finite sample of a random function, we use a second Bayesian bootstrap procedure to integrate over the surrogate's random predictions, effectively accounting for model error in posterior inference (Huggins and Miller,

https://doi.org/10.1017/jog.2020.112 Published online by Cambridge University Press

2019) induced by using the surrogate (rather than the high-fidelity model) for inference.

We find that high-fidelity model is able to reproduce many of the salient features of the observed annual average surface velocity field for a terrestrially terminating subset of southwestern Greenland, with the model explaining on average ~60% of the variance in observations. As expected, we find significant correlations in the posterior distribution of model parameters. However, we also find that surface velocity observations provide substantial constraints on most model parameters. To ensure that the distribution inferred using the surrogate is still reasonable given the high-fidelity model, we select a handful of samples from the posterior distribution, feed them back into the high-fidelity model, and show that the resulting predictive distribution remains consistent with observations. The process described above is applicable to the broad class of problems in which we would like to perform Bayesian inference over a limited number of parameters given an expensive deterministic model.

Study area

We focus our study on the region of western Greenland centered around Russell Glacier (Fig. 1). The domain runs from the ice margin to the ice divide, covering an area of ~36,000 km^2. This region was selected because it strikes a balance between being simple and being representative: all glacier termini are terrestrial, which means that the effects of calving can be neglected in this study, surface slopes are modest, and surface meltwater runoff rates are neither extreme nor negligible, yet there is still substantial spatial variability in glacier speed even near the margin, from a maximum of 150 m a^{-1} over the deep trench at Isunnguata Sermia, to <30 m a^{-1} just 20 km to the north.

Additionally, this region of Greenland has long been a hotspot for observations due to its proximity to the town of Kangerlussuaq. The bed is well-constrained by Operation IceBridge flightlines, and throughout this study, we use the basal topography of BedMachine V2 (Morlighem and others, 2017). We force the model with surface meltwater runoff computed with HIRHAM (Mottram and others, 2017), averaged by month between 1992 and 2015. As such, our forcing is time-varying but periodic with a period of 1 year. When comparing modeled to observed velocities (henceforth called $\mathbf{u}_{obs}$), we use as our observation the inSAR-derived annual average velocity fields of Joughin and others (2018), further averaged over the years 2014 through 2018.

Numerical models

Ice dynamics

Viscous flow

The flow of the ice sheet over a volume Ω is modeled as a low Reynolds number fluid using a hydrostatic approximation to Stokes' equations (Pattyn, 2003)

$$\nabla \cdot \tau' = \rho_i g \nabla z_s, \tag{1}$$

where

$$\tau' = \begin{bmatrix} 2\tau_{xx} + \tau_{yy} & \tau_{xy} & \tau_{xz} \\ \tau_{xy} & \tau_{xx} + 2\tau_{yy} & \tau_{yz} \end{bmatrix}. \tag{2}$$

z_s is the glacier surface elevation, ρ_i is the ice density, g is the gravitational acceleration and τ_{ij} is a component of the deviatoric stress tensor given by

$$\tau_{ij} = 2\eta\dot{\epsilon}_{ij}, \tag{3}$$

with $\dot{\epsilon}$ the symmetrized strain rate tensor. The viscosity

$$\eta = \frac{A^{-(1/n)}}{2}(\dot{\epsilon}_{II} + \dot{\epsilon}_0)^{1-1/n} \tag{4}$$

is dependent on the second invariant of the strain rate tensor $\dot{\epsilon}_{II}$. Note that we make an isothermal approximation, and take the ice softness parameter A to be a constant. We explicitly note that this assumption may be questionable. However, because models of Greenland thermal conditions frequently do not match borehole observations in the region considered here (e.g. Harrington and others, 2015) and sliding in this region is an order of magnitude greater than deformation (Maier and others, 2019), we choose to avoid the additional computational expense and uncertainty associated with introducing a thermal model. The exponent in Glen's flow law $n = 3$.

Boundary conditions

At the ice surface Γ_{z_s} and terminal margin Γ_T (where the ice thickness is assumed to approximate zero), we impose a no-stress boundary condition

$$\tau' \cdot \mathbf{n} = \mathbf{0}, \tag{5}$$

where $\mathbf{n}$ is the outward pointing normal vector, and $\mathbf{0}$ is the zero vector.

The remaining lateral boundary Γ_L is synthetic in the sense that there are no natural physical boundary conditions that should be applied there. Here, we adopt the boundary condition of Papanastasiou and others (1992), who suggest that the boundary term appearing in the weak form of Eqn (1) (the second term in Eqn (B2)) not be replaced by an arbitrary condition (no stress, e.g.), but rather retained and included as an unknown to be determined as part of the solution procedure. Although this does not lead to a unique solution in the strong form of the differential equation, it does lead to one after discretization with the finite element method. The resulting boundary condition for linear Lagrange finite elements specifies that the curvature of both velocity components vanishes at a point near the boundary which for a sufficiently smooth velocity field outside of the domain approximates a stress free boundary at an infinitely distant location. Griffiths (1997) refers to this as the 'no boundary condition' and show that it is equivalent to solving a reduced order equation in the neighborhood of the boundary, which for the discretization that we describe below reduces to the solution of the shallow ice approximation.

At the basal boundary Γ_{z_b} we impose the sliding law

$$\tau' \cdot \mathbf{n} = -\beta^2 N^p \|\mathbf{u}\|_2^{q-1}\mathbf{u}, \tag{6}$$

with β^2 the basal traction coefficient and $\mathbf{u}$ the ice velocity, and we use $\|\cdot\|_2$ to denote the standard L_2 norm. We note that this sliding law has some theoretical (Fowler, 1987) and empirical (Budd and others, 1979; Bindschadler, 1983) support, but does not satisfy Iken's bound (Iken, 1981). As such there are alternative sliding laws that may be preferable (e.g. Schoof, 2005). However, we defer a detailed comparison of different sliding laws, and condition this study on Eqn (6) being a reasonable (and numerically stable) approximation to the true subglacial process.

https://doi.org/10.1017/jog.2020.112 Published online by Cambridge University Press

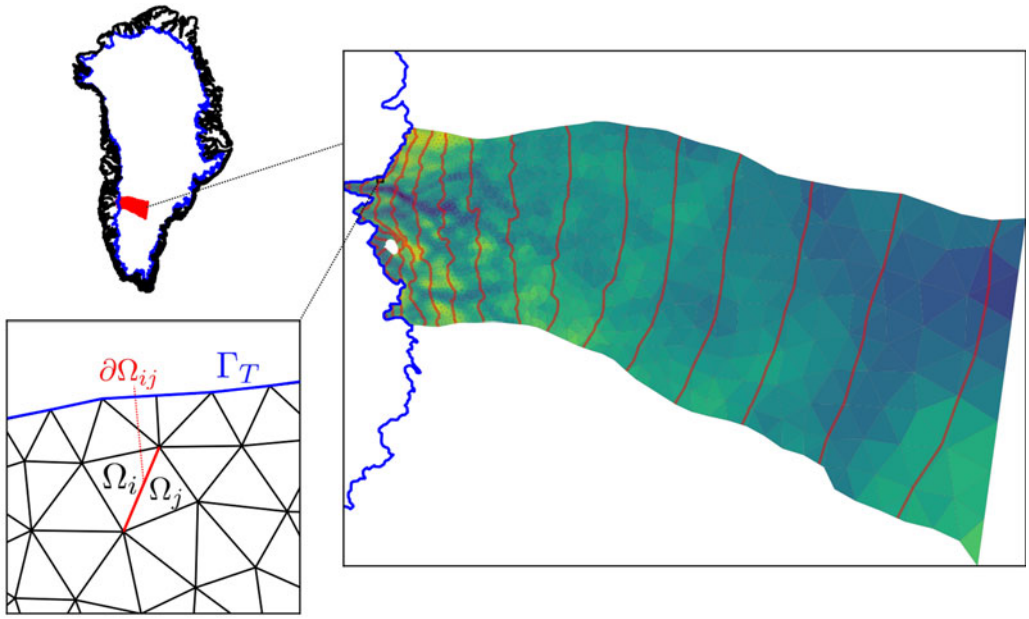


Fig. 1. Study area, with location of domain in Greenland (top left), detailed modeling domain with the computational mesh overlain with bedrock elevation and surface contours (right), and closeup of mesh with domains used in modeling labeled (bottom left, see text). Note that the equilibrium line altitude is at approximately the 1500 m contour. Ω_i represent individual mesh cells, $\delta\Omega_{ij}$ the boundary between them and Γ_T the terminal boundary.

The effective pressure N is given by the ice overburden pressure $P_0 = \rho_i g H$ less the water pressure P_w

$$N = P_0 - P_w.$$

The exponents p and q control the non-linear response of basal shear stress to the effective pressure and velocity (respectively). We note several limiting cases of this sliding law: when $p = q = 1$, we recover the linear Budd law (Budd and others, 1979). When $p = 0$, we get the pressure-independent Weertman law (Weertman, 1957). In the limit $q \to \infty$, we recover a perfectly plastic model of basal stress (e.g. Kamb, 1991).

In practice, we use a re-parameterized version of Eqn (6)

$$\tau_b = \gamma^2 \hat{N}^p \|\hat{\mathbf{u}}\|^{q-1} \hat{\mathbf{u}}, \tag{7}$$

where $N/\text{Scale}(N) = \hat{N}$ is the effective pressure non-dimensionalized by the ice overburden averaged over the model domain, and $\mathbf{u}/\text{Scale}(\mathbf{u}) = \hat{\mathbf{u}}$ is similar, with the characteristic scale of $\mathbf{u}$ taken to be 50 m a^{-1}. Thus, the resulting relationship between γ^2 (which has units of stress) and β^2 is

$$\beta^2 = \frac{\gamma^2}{\text{Scale}(N)^p \text{Scale}(\mathbf{u})^q}. \tag{8}$$

This transformation is helpful because the power law terms on the right-hand side of Eqn (6) can vary by several orders of magnitude, thus requiring that β^2 does the same in order to maintain a given characteristic surface velocity. The γ^2 parameterization circumvents this scale issue. We take γ^2, p and q to be unknown but spatially and temporally constant.

Hydrologic model

In order to predict the effective pressure N on which the sliding law depends, we couple the above ice dynamics model to a hydrologic model that simulates the evolution of the subglacial and englacial storage via fluxes of liquid water through an inefficient linked cavity system and an efficient linked channel system. This model closely follows the model GlaDS (Werder and others, 2013), with some alterations in boundary conditions, discretization and opening rate parameterization.

Over a disjoint subdomain $\bar{\Omega}_i \subset \bar{\Omega}$, $\bigcup_{i\in\mathcal{T}} \bar{\Omega}_i = \bar{\Omega}$, where $\mathcal{T}$ is the set of triangles in the finite element mesh, the hydraulic potential $\phi = P_w + \rho_w g z_b$ (with z_b the bedrock elevation) evolves according to the parabolic equation

$$\frac{e_v}{\rho_w g}\frac{\partial \phi}{\partial t} + \nabla \cdot \mathbf{q} - \mathcal{C} + \mathcal{O} = m, \tag{9}$$

where P_w is the water pressure, ρ_w the density of water, $\mathbf{q}$ the horizontal flux, $\mathcal{C}$ the rate at which the cavity system closes (pushing water into the englacial system), $\mathcal{O}$ the rate at which it opens and m is the recharge rate (either from the surface, basal melt or groundwater). The hydraulic potential is related to the effective pressure by

$$N = \rho_w g z_b + P_0 - \phi. \tag{10}$$

The horizontal flux is given by the Darcy–Weisbach relation

$$\mathbf{q} = -k_s h^{\alpha_s} \|\nabla \phi\|_2^{\beta_s - 2} \nabla \phi, \tag{11}$$

a non-linear function of the hydraulic potential, characteristic cavity height h, bulk conductivity k_s and turbulent exponents α_s and β_s.

The average subglacial cavity height h evolves according to

$$\frac{\partial h}{\partial t} = \mathcal{O} - \mathcal{C}. \tag{12}$$

Here, we model the subgrid-scale glacier bed as self-similar, with bedrock asperity heights modeled with a log-normal distribution:

$$\log h_r \sim \mathcal{N}(\log \bar{h}_r, \sigma_h^2), \tag{13}$$

and a characteristic ratio r of asperity height to spacing. Thus, the

https://doi.org/10.1017/jog.2020.112 Published online by Cambridge University Press

opening rate is given by

$$\mathcal{O} = \int_0^\infty \mathrm{Max}\left(\|\mathbf{u}(\varsigma = 1)\|_2 r(1 - \frac{h}{h_r}), 0 \right) P(h_r)\, \mathrm{d}h_r, \tag{14}$$

where we use $P(\cdot)$ to denote the probability density function. For $\sigma_h^2 = 0$, this expression is equivalent to the standard opening rate used in previous studies (e.g. Werder and others, 2013), albeit reparameterized. However, this implies that once the cavity size reaches h_r, then the opening rate becomes zero: for a glacier moving increasingly quickly due to a high water pressure, there is no mechanism for subglacial storage capacity to increase. For $\sigma_h^2 > 0$, our formulation regularizes the opening rate such that there is 'always a bigger bump', but with a diminishing effect away from the modal bump size. Here, we make the somewhat arbitrary choice that $\sigma_h^2 = 1$ and take $\bar{h}_r$ to be a tunable parameter.

The cavity closing rate is given by

$$\mathcal{C} = \frac{2}{n^n} A h |N|^{n-1} N. \tag{15}$$

Over a domain edge $\partial\Omega_{ij}$ (the edge falling between subdomains $\bar{\Omega}_i$ and $\bar{\Omega}_j$), mass conservation implies that

$$\frac{\partial S}{\partial t} + \frac{\partial Q}{\partial s} = \frac{\Xi - \Pi}{\rho_w L} + m_c, \tag{16}$$

with S the size of a channel occurring along that edge, Ξ the opening rate due to turbulent dissipation, Π the rate of sensible heat changes due to pressure change and m_c the exchange of water with adjacent domains. We adopt the constitutive relations given in Werder and others (2013) for each of these terms. The channel discharge Q is given by another Darcy–Weisbach relation

$$Q = -k_c S_c^\alpha \left\| \frac{\partial \phi}{\partial s} \right\|_2^{\beta-2} \frac{\partial \phi}{\partial s}, \tag{17}$$

where k_c is a bulk conductivity for the efficient channelized system. The channel size evolves according to

$$\frac{\partial S}{\partial t} = \frac{\Xi - \Pi}{\rho_i L} - \mathcal{C}_c, \tag{18}$$

with channel closing rate

$$\mathcal{C}_c = \frac{2}{n^n} A S |N|^{n-1} N. \tag{19}$$

Substitution of Eqn (18) into Eqn (16) leads to an elliptic equation

$$\frac{\partial Q}{\partial s} = \frac{\Xi - \Pi}{L}\left(\frac{1}{\rho_i} - \frac{1}{\rho_w} \right) + m_c. \tag{20}$$

The exchange term with the surrounding sheet is given by

$$[\mathbf{q} \cdot \mathbf{n}]_+ + [\mathbf{q} \cdot \mathbf{n}]_- = m_c, \tag{21}$$

which states that flux into (or out of) a channel is defined implicitly by the flux balance between the two adjacent sheets.

Boundary conditions

We impose a no-flux boundary condition across boundaries $\Gamma_T \cup \Gamma_L$ in both the sheet and conduit model:

$$\mathbf{q} \cdot \mathbf{n} = 0 \tag{22}$$

$$Q = 0 \tag{23}$$

At first glance, this seems to be a strange choice: how then, does water exit the domain? To account for this, we impose the condition that whenever $\phi > \phi_{z_s}$, where $\phi_{z_s} = \rho_w g z_s$ is the surface potential, any excess water immediately runs off. Because the margins are thin, and the flux across the lateral boundary is zero, the hydraulic head there quickly rises above the level of the ice surface, and the excess water runs off. This heuristic is necessary to avoid the numerically challenging case when potential gradients would imply an influx boundary condition. With a free flux boundary, the model would produce an artificial influx of water from outside the domain in order to keep channels filled, which is particularly problematic in steep topography. Most of the time, the chosen inequality condition has the practical effect of setting the hydraulic potential on the terminus to atmospheric pressure. We note that a better solution would be to devise a model that allows for unfilled conduits along with an explicit modeling of the subaerial hydrologic system. However, we defer that development to later work and condition the results of this study on the heuristic described above.

In addition to this condition, we also enforce the condition that channels do not form at the margins (i.e. $S = 0$ on $\Gamma_T \bigcup \Gamma_L$). At the terminus, this ensures that there are no channels with unbounded growth perpendicular to the terminus, and also to ensure that lateral boundaries (where $H > 0$) are not used as preferential flow paths.

Surrogate model

The solution of the coupled model defined above defines a function $\mathcal{F} : \mathbb{R}_+^k \to \mathbb{R}^{n_p}$ that maps from a parameter vector

$$\mathbf{m} = [k_s, k_c, \bar{h}_r, r, \gamma^2, p, q, e_v]^T \tag{24}$$

of length $k = 8$ to a vector of annually-averaged surface speeds defined at each point on the computational mesh

$$\mathcal{F}(\mathbf{m}) = \frac{1}{t_1 - t_0} \int_{t_0}^{t_1} \|\mathbf{u}(t; \mathbf{m})|_{z=z_s}\|_2 \mathrm{d}t,$$

where $t_0 = 15$ years and $t_1 = 20$ years, i.e. the result of running the high-fidelity model with time-varying meltwater forcing for 20 years given parameters $\mathbf{m}_i$, computing the speed at the surface, and taking its average over the last 5 years to ensure that the model has reached dynamic equilibrium. We emphasize that we are dealing in speeds, but that further study could extend the methods presented here to consideration of the complete vector quantity.

The evaluation of $\mathcal{F}$ is computationally expensive and we anticipate needing to evaluate it many times in order to approximate parameter uncertainty through, for example, an MCMC sampling scheme, which cannot be easily parallelized. We therefore seek to create a function $\mathcal{G} : \mathbf{R}_+^k \to \mathbb{R}^{n_p}$ that yields approximately the same map as $\mathcal{F}$, but at a substantially lower cost.

A variety of mechanisms may be used to construct such an approximation, here called the *surrogate model*. To construct the surrogate, we take a machine-learning approach, in which

https://doi.org/10.1017/jog.2020.112 Published online by Cambridge University Press

we create a large (but finite) set of model input and output pairs $D = \{(\mathbf{m}_i, \mathcal{F}(\mathbf{m}_i))\}$. We then use these input–output pairs as training examples over which to optimize the parameters of a highly flexible function approximator, in this case an artificial neural network. We note that each sample is independent, and thus the evaluation of the high-fidelity model for each ensemble member can be performed concurrently.

Large ensemble

In order to construct the training data for $\mathcal{G}$, we must select the values $\mathbf{m}_i$ over which $\mathcal{F}$ should be evaluated. Because all values in $\mathbf{m}$ are positive, yet we do not wish to bias the dataset toward certain regions of the plausible parameter set over others, we choose to draw $\mathbf{m}$ from a log-uniform distribution with lower and upper bounds $\mathbf{b}_\mathrm{L}$ and $\mathbf{b}_\mathrm{U}$:

$$\log_{10}(\mathbf{m}) \sim \mathcal{U}(\mathrm{Bound}_\mathrm{L}, \mathrm{Bound}_\mathrm{U}). \tag{25}$$

We refer to this distribution as $P_\mathrm{em}(\mathbf{m})$. The specific values of the bounds are given in Table 1, but in general, parameters vary a few orders of magnitude in either direction from values commonly found in the literature. Note that this distribution is *not* the prior distribution that we will use for Bayesian inference later on. Rather, it is an extremal bound on what we believe viable parameter values to be. However, the support for the distributions is the same, ensuring that the surrogate model is not allowed to extrapolate.

One viable strategy for obtaining training examples would be to simply draw random samples from $P_\mathrm{em}(\mathbf{m})$, and evaluate the high-fidelity model there. However, because we would like to ensure that there is a sample 'nearby' all locations in the feasible parameter space, we instead generate the samples using the quasi-random Sobol sequence (Sobol and others, 2011), which ensures that the parameter space is optimally filled (the sequence is constructed such that the sum of a function evaluated at these samples converges to the associated integral over the domain as quickly as possible). Although the Sobol sequence is defined over the k-dimensional unit hypercube, we transform it into a quasi-random sequence in the space of $P_\mathrm{em}(\mathbf{m})$ using the percent point function.

With this distribution of parameters in hand, we evaluate $\mathcal{F}$ on each sample $\mathbf{m}_i$. Using 48 cores, this process took ~4 d for 5000 samples. Note that some parameter combinations never converged, in particular cases where γ^2 was too low and the resulting velocity fields were many orders of magnitude higher than observed. We discarded those samples and did not use them in subsequent model training.

Surrogate architecture

Dimensionality reduction

We construct the surrogate model $\mathcal{G}$ in two stages. In the first stage, we perform a PCA (Shlens, 2014) to extract a limited set of basis functions that can be combined in linear combination such that they explain a maximal fraction of the variability in the ensemble. Specifically, we compute the eigendecomposition

$$\mathcal{S} = V\Lambda V^T, \tag{26}$$

where Λ is a diagonal matrix of eigenvalues and the columns of V are the eigenvectors of the empirical covariance matrix

$$\mathcal{S} = \sum_{i=1}^{m} \omega_{d,i}\left[\log_{10}\mathcal{F}(\mathbf{m}_i) - \log_{10}\bar{\mathcal{F}}\right]^2, \tag{27}$$

Table 1. Upper and lower bounds for both the log-uniform distribution used to generate surrogate training examples, as well as the log-beta prior distribution

Parameter	Lower bound	Upper bound
k_s	10^{-4}	10^{0}
k_c	10^{-4}	10^{0}
$\bar{h}_r$	10^{-3}	10^{1}
r	10^{-2}	10^{1}
γ^2	10^{5}	10^{7}
p	10^{-1}	1.2
q	10^{-1}	1.2
e_v	10^{-4}	10^{-2}

with ω_d a vector of weights such that $\sum_{i=1}^{m}\omega_{d,i} = 1$ (defined later in Eqn (39)) and

$$\log_{10}\bar{\mathcal{F}} = \sum_{i=1}^{m}\omega_{d,i}\log_{10}\mathcal{F}(\mathbf{m}_i). \tag{28}$$

We work with log-velocities due to the large variability in the magnitude of fields that are produced by the high-fidelity model.

The columns of V are an optimal basis for describing the variability in the velocities contained in the model ensemble. They represent dominant model modes (Fig. 2) (in the climate literature, these are often called empirical orthogonal functions). We refer to them as 'eigenglaciers' in homage to the equivalently defined 'eigenfaces' often employed in facial recognition problems (Sirovich and Kirby, 1987). The diagonal entries of Λ represent the variance in the data (once again, here these are a large set of model results) explained by each of these eigenglaciers in descending order. As such, we can simplify the representation of the data by assessing the fraction of the variance in the data still unexplained after representing it with j components

$$f(j) = 1 - \frac{\sum_{i=1}^{j}\Lambda_{ii}}{\sum_{i=1}^{m}\Lambda_{ii}}. \tag{29}$$

We find a cutoff threshold c for the number of eigenglaciers to retain by determining $c = \max_j \in \{1, \ldots, m\}: f(j) > s$. We set $s = 10^{-4}$, which is to say that we retain a sufficient number of basis functions such that we can represent 99.99% of the velocity variability in the model ensemble. For the experiments considered here, $c \approx 50$.

Any velocity field that can be produced by the high-fidelity model can be approximately represented as

$$\mathcal{F}(\mathbf{m}) \approx \sum_{j=1}^{c}\lambda_j(\mathbf{m})V_j, \tag{30}$$

where V_j is the j-th eigenglacier, and λ_j is its coefficient. The (row) vector $\lambda(\mathbf{m})$ can thus be thought of as a low-dimensional set of 'knobs' that control the recovered model output.

Artificial neural network

Unfortunately, we do not a priori know the mapping $\lambda(\mathbf{m})$. In the second stage of surrogate creation, we seek to train a function $\lambda(\mathbf{m};\theta)$ with trainable parameters $\theta = \{W_l, b_l, \alpha_l, \beta_l : l = 1, \ldots, L\}$ such that the resulting reconstructed velocity field is as close to the high-fidelity model's output as possible, where L is the number of network blocks (see below). For this task, we use a deep but narrow residual neural network. The architecture of this network is shown in Figure 3. Our choice to use a neural network (as opposed to alternative flexible models like Gaussian process regression and polynomial chaos

https://doi.org/10.1017/jog.2020.112 Published online by Cambridge University Press

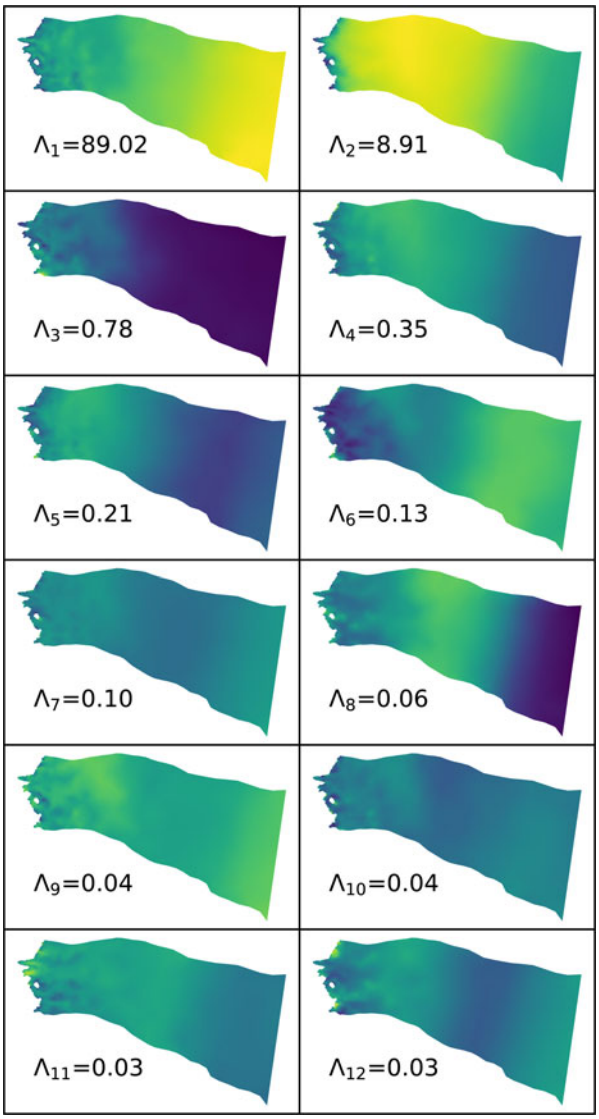


Fig. 2. First 12 basis functions in decreasing order of explained variance for one of 50 bootstrap-sampled ensemble members. The color scale is arbitrary.

expansion) was motivated primarily by the relatively high dimensionality of our predictions, for which Gaussian processes and polynomial chaos expansions are not well suited due to the difficulty of modeling cross-covariance.

As is common for artificial neural networks, we repeatedly apply a four operation block with input h_{l-1} and output h_l. As input to the first block we have our parameter vector, so $h_0 = \mathbf{m}$. In each block, the first operation is a simple linear transformation

$$\hat{\mathbf{a}}_l = \mathbf{h}_{l-1} W_l^T + \mathbf{b}_l, \tag{31}$$

where W_l and $\mathbf{b}_l$ are respectively a learnable weight matrix and bias vector for block l. To improve the training efficiency of the neural network, the linear transformation is followed by so-called layer normalization (Ba and others, 2016), which z-normalizes then rescales the intermediate quantity $\hat{a}_l$

$$\mathbf{a}_l = \alpha_l \frac{\hat{\mathbf{a}}_l - \mu_l}{\sigma_l} + \beta_l, \tag{32}$$

where μ_l and σ_l are the mean and standard deviation of $\hat{\mathbf{a}}_l$, and α_l and β_l are learnable layerwise scaling parameters. Next, in order for the artificial neural network to be able to represent non-linear functions, we apply an activation

$$\hat{\mathbf{z}}_l = \mathrm{ReLU}(\mathbf{a}_l), \tag{33}$$

where

$$\mathrm{ReLU}(x) = \mathrm{Max}(x, 0) \tag{34}$$

is the rectified linear unit (Glorot and others, 2011). Next we apply dropout (Srivastava and others, 2014), which randomly zeros out elements of the activation vector during each epoch of the training phase

$$\mathbf{z}_l = \hat{\mathbf{z}}_l \odot R, \tag{35}$$

where R is a vector of Bernoulli distribution random variables with mean p. After training is complete and we seek to evaluate the model, we set each element in R to p, which implies that the neural network produces deterministic output with each element of the layer weighted by the probability that it was retained during training. Dropout has been shown to effectively reduce overfitting by preventing complex co-adaptation of weights: by never having guaranteed access to a given value during the training phase, the neural network learns to never rely on a single feature in order to make predictions.

Finally, if dimensions allow (which they do for all but the first and last block), the output of the block is produced by adding its input

$$\mathbf{h}_l = \mathbf{z}_l + \mathbf{h}_{l-1}, \tag{36}$$

a so-called residual connection (He and others, 2016) which provides a 'shortcut' for a given block to learn an identity mapping. This mechanism has been shown to facilitate the training of deep neural networks by allowing an unobstructed flow of gradient information from the right end of the neural network (where the data misfit is defined) to any other layer in the network. For each of these intermediate blocks, we utilize $n_h = 64$ nodes.

At the last block as $l = L$, we have that $\lambda(\mathbf{m}) = \mathbf{h}_{L-1} W_L^T + \mathbf{b}_L$. In this study, $L = 5$. $\lambda(\mathbf{m})$ is then mapped to a log-velocity field via V, as described above. The complete surrogate model is thus defined as

$$\mathcal{G}(\mathbf{m}) = 10^{\lambda(\mathbf{m}) V^T}. \tag{37}$$

Surrogate training

To train this model, we minimize the following objective

$$I(\theta) \propto \sum_{i=1}^{m} \sum_{j=1}^{n_p} \omega_{d,i} A_j \left[\log_{10} \mathcal{G}(\mathbf{m}_i; \theta)_j - \log_{10} \mathcal{F}(\mathbf{m}_i)_j\right]^2, \tag{38}$$

where A_j is the fractional area of the j-th gridpoint, and $\omega_{d,i} \in [0, 1]$, $\sum_{i=1}^{m} \omega_{d,i} = 1$ is the weight of the i-th training example model error. The former term is necessary because our computational mesh resolution is variable, and if were to simply compute the integral as a sum over gridpoints, we would bias the estimator toward regions with high spatial resolution.

The model above is implemented in pytorch, which provides access to objective function gradients via automatic differentiation (Paszke and others, 2019). We minimize the objective using the ADAM optimizer (Kingma and Ba, 2014), which is a variant of minibatch stochastic gradient descent. We use a batch size of 64 input–output pairs (i.e. 64 pairs of parameters and their associated high-fidelity model predictions), and begin with a learning rate of $\eta = 10^{-2}$, that is exponentially decayed by one order of

https://doi.org/10.1017/jog.2020.112 Published online by Cambridge University Press

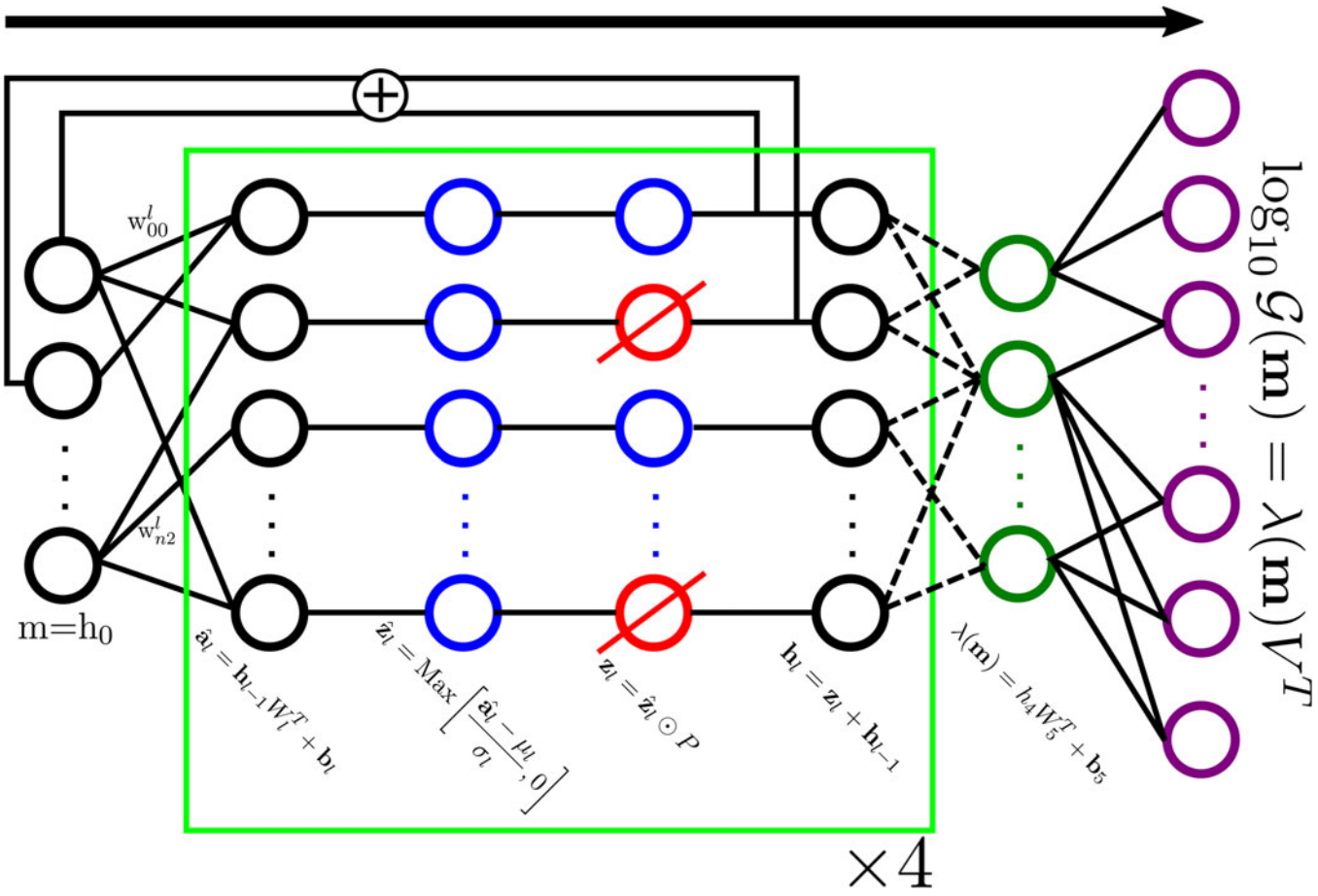


Fig. 3. Architecture of the neural network used as a surrogate model in this study, consisting of four repetitions of linear transformation, layer normalization, dropout and residual connection, followed by projection into the velocity field space through linear combination of basis functions computed via PCA.

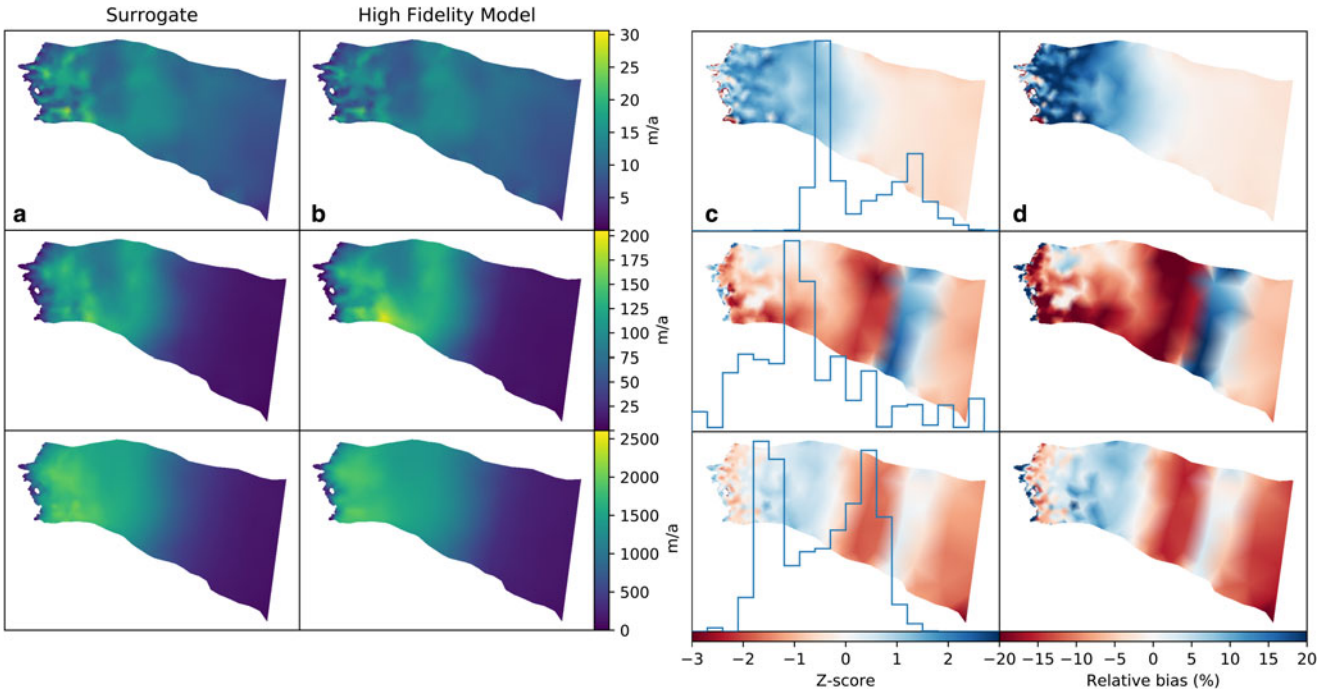


Fig. 4. Comparison between emulated velocity field (a) and modeled velocity field (b) for three random instances of **m**. Note the different velocity scales for each row. These predictions are out of set: the surrogate model was not trained on these examples, and so is not simply memorizing the training data. (c) Difference between high-fidelity and surrogate modeled speeds, normalized by standard deviation of surrogate model ensemble (a *z*-score), with histogram of the same shown by blue line. (d) Difference between high-fidelity and surrogate modeled speeds, normalized by high-fidelity model speeds.

https://doi.org/10.1017/jog.2020.112 Published online by Cambridge University Press

magnitude per 1000 epochs (an epoch being one run through all of the training instances). We run the optimization for 4000 epochs.

The results of the surrogate training are shown in Figure 4. We find that for most instances, the surrogate model produces a velocity field in good agreement with the one produced by the high-fidelity model: in most cases the predicted velocities fall within 20% of the high-fidelity model's predictions. Furthermore, in >99% of instances the nodally defined high-fidelity model predictions fall within three of the ensemble's standard deviations of its own mean (although these residuals are clearly non-normal). The exception to this occurs in instances where the velocity fields are more than three orders of magnitude greater than observations. Since we intend to use the surrogate for inference and such a velocity field implies that the parameters that created it are unlikely to be consistent with observations anyways, this extreme-value misfit will not influence the inference over glacier model parameters.

Bayesian bootstrap aggregation

Neural networks are known to be high-variance models, in the sense that while the high-fidelity model may exhibit a monotonic relationship between input parameters and output velocities, the neural network may exhibit high frequency 'noise', similar to that exhibited when fitting high-order polynomials to noisy data. This noise is problematic in that it tends to yield local minima that prohibit optimization and sampling procedures from full exploration of the parameter space. In order to reduce this variance, we introduce Bayesian bootstrap aggregation (Breiman, 1996; Clyde and Lee, 2001) (so-called bagging), in which we train the surrogate described B times, with the sample weights used in Eqn (38) each time randomly drawn from the Dirichlet distribution

$$\omega_{d,i} \sim \text{Dirichlet}(\mathbf{1}), \tag{39}$$

where $\mathbf{1}$ is a vector of ones with length m, the number of training instances.

This procedure yields B independent instances of $\mathcal{G}$ (with single instances hereafter referred to as $\mathcal{G}_i$), which are combined as a committee. One way to think about this process is that the high-fidelity model is the mean of a distribution, and each ensemble member is a 'data point' (a random function) drawn from that distribution. The optimal estimate of the true mean (once again, the high-fidelity model) is the sample mean of the bootstrap samples

$$\bar{\mathcal{G}}(\mathbf{m}) = \sum_{i=1}^{B} \omega_{e,i}\mathcal{G}_i(\mathbf{m}), \tag{40}$$

with the weights $\omega_{e,i} \in [0, 1]$, $\sum_{i=1}^{B} \omega_{e,i} = 1$. Although this aggregation reduces predictive error (i.e. yields a better approximation to the high-fidelity model) relative to using a single model, uncertainty remains because we are approximating the true mean with the mean based on a finite number of samples. To account for this residual uncertainty in the surrogate model, we can once again employ Bayesian bootstrapping (Rubin, 1981). In principle, we assume that the sample (the computed members of the bagging committee) provide a reasonable approximation to the population (all possible members of the bagging committee) when estimating variability in the mean. In practice, this means that we model $G(\mathbf{m})$ as a random function given by Eqn (40) augmented with Dirichlet distributed weights

$$\omega_{e,i} \sim \text{Dirichlet}(\mathbf{1}). \tag{41}$$

Bayesian inference

We would like to infer the posterior distribution of model parameters $\mathbf{m}$ given observations $\mathbf{d}$, with the added complexity that the random surrogate described above is only an approximation to the high-fidelity model. The former is done using MCMC sampling (Kass and others, 1998), the details of which are described in Appendix C. The latter can be accomplished by marginalizing over the surrogate distribution, or equivalently the bootstrap weights ω_e.

$$P(\mathbf{m}|\mathbf{d}) = \int P(\mathbf{m}, \omega_e|\mathbf{d})\mathrm{d}\omega_e \tag{42}$$

Applying Bayes theorem to the right-hand side, we have that

$$\begin{aligned} P(\mathbf{m}|\mathbf{d}) &\propto \int P(\mathbf{d}|\mathbf{m}, \omega_e)P(\mathbf{m}, \omega_e)\mathrm{d}\omega_e \\ &\propto \int P(\mathbf{d}|\mathbf{m}, \omega_e)P(\mathbf{m})P(\omega_e)\mathrm{d}\omega_e, \end{aligned} \tag{43}$$

where we have used the fact that the bootstrap weights and model parameters are independent. On the left-hand side is the quantity of interest, the posterior distribution of model parameters given observations, while inside the integral, $P(\mathbf{d}|\mathbf{m})$ is the likelihood of observing the data given a set of model parameters, and $P(\mathbf{m})$ is the prior distribution over model parameters.

Likelihood model

Observations of surface velocity are reported as a field, as are the model predictions, and thus we have an infinite-dimensional Bayesian inference problem (Bui-Thanh and others, 2013; Petra and others, 2014) because there are an infinite number of real-valued coordinates at which to evaluate misfit. However, in contrast to previous studies, rather than finite observations with an infinite parameter space, we have the converse, with continuous (infinite) observations and finite-dimensional parameters. To circumvent this difficulty, we propose a relatively simple approximation that can account for observational correlation and a variable grid size. We first assume a log-likelihood of the form

$$\log P(\mathbf{d}|\mathbf{m}, \omega_e) \propto -\frac{1}{2}\int_{\bar{\Omega}}\int_{\bar{\Omega}'} \frac{r(x)r(x')}{\sigma(x, x')}\rho_d^2 \mathrm{d}\bar{\Omega}'\mathrm{d}\bar{\Omega}, \tag{44}$$

where ρ_d is the data density (number of observations per square meter), $\sigma(x, x')$ is a covariance function

$$\sigma(x, x') = \sigma_{obs}^2 + \sigma_{cor}^2\left(1 + \frac{d(x, x')}{2l^2}\right)^{-1} \tag{45}$$

that superimposes white noise with variance σ_{obs}^2 and rational exponential noise with variance σ_{cor}^2 and characteristic length scale l, which we take as four times the local ice thickness. The former term is intended to account for aleatoric observational uncertainty. The latter is a catch-all intended to account for epistemological uncertainty in the flow model and systematic errors in derivation of the velocity fields, with the rational exponential kernel having 'heavy tails' that represent our uncertainty in the true correlation length scale of such errors. Although they represent our best efforts at specifying a sensible likelihood model, we emphasize that they are also somewhat arbitrary and can have significant effects on the resulting analysis. However, in the absence of a more clearly justified model, we assume the one presented here.

$r(x)$ is a residual function given by

$$r(x) = \bar{\mathcal{G}}(x; \mathbf{m}, \omega_e) - \|\mathbf{u}_{obs}\|_2(x), \tag{46}$$

https://doi.org/10.1017/jog.2020.112 Published online by Cambridge University Press

where $\mathbf{u}_{obs}$ is the satellite derived, annually averaged velocity field described in the 'Study area' section, and in which we omit writing the dependence on $\mathbf{m}$ for brevity.

Because solutions are defined over a finite element mesh, we split the integrals in Eqn (44) into a sum over dual mesh elements T in collection $\mathcal{T}$

$$\log P(\mathbf{d}|\mathbf{m},\, \omega_e) \propto -\frac{1}{2}\sum_{T\in\mathcal{T}}\sum_{T'\in\mathcal{T}}\int_T\int_{T'}\frac{r(x)r(x')}{\sigma(x,\, x')}\rho^2 \mathrm{d}T'\mathrm{d}T. \quad (47)$$

Finally, we make the approximation

$$\int_T\int_{T'}\frac{r(x)r(x')}{\sigma(x,\, x')}\rho^2\mathrm{d}T'\mathrm{d}T \approx \frac{r(x_T)r(x_{T'})}{\sigma(x_T,\, x_{T'})}\rho^2 A_{T'}A_T, \quad (48)$$

where x_T are the coordinates of the barycenter of T (the finite element mesh nodes) and A_T its area. Defining

$$\mathbf{r}^T = [r(x_1),\, r(x_2),\, \ldots,\, r(x_N)] \quad (49)$$

and

$$\Sigma^{-1} = K\hat{\Sigma}^{-1}K, \quad (50)$$

where $\hat{\Sigma}_{ij} = \sigma(x_i,\, x_j)$ and $K = \mathrm{Diag}([\rho A_1,\, \rho A_2,\, \ldots,\, \rho A_N])$ yields the finite-dimensional multivariate-normal likelihood

$$P(\mathbf{d}|\mathbf{m}) \propto \exp\left[-\frac{1}{2}\mathbf{r}^T\Sigma^{-1}\mathbf{r}\right]. \quad (51)$$

Prior distribution

In principle, we have very little knowledge about the actual values of the parameters that we hope to infer and thus would like to impose a relatively vague prior during the inference process. However, because the surrogate is ignorant of the model physics, we must avoid allowing it to extrapolate beyond the support of the ensemble. One choice that fulfills both of these objectives is to use as a prior the same log-uniform distribution that we used to generate the surrogate. However, the ensemble distribution was designed to cover as broad a support as possible without biasing the surrogate toward fitting parameter values near some kind of mode and does not represent true prior beliefs about the parameter values. Instead, we adopt for the parameters a scaled log-Beta prior

$$\frac{\log_{10}\mathbf{m} - \mathrm{Bound_L}}{\mathrm{Bound_U} - \mathrm{Bound_L}} \sim \mathrm{Beta}(\alpha = 2,\, \beta = 2) \quad (52)$$

This prior reflects our belief that good parameters values are more likely located in the middle of the ensemble, while also ensuring that regions of parameter space outside the support of the ensemble have zero probability.

Marginalization over ω_e

In order to perform the marginalization over bootstrap weights, we make the Monte Carlo approximation

$$\int P(\mathbf{d}|\mathbf{m},\, \omega_e)P(\mathbf{m})P(\omega_e)\mathrm{d}\omega_e \approx \sum_{i=1}^{N} P(\mathbf{d}|\mathbf{m},\, \omega_{e,i})P(\mathbf{m}), \quad (53)$$

with $\omega_{e,i}$ drawn as in Eqn (53), where N is a number of Monte Carlo samples. The terms in the sum are independent, and may be computed in parallel. However, they are also analytically intractable. Thus, we draw samples from each of the summand distributions (the posterior distribution conditioned on an instance of ω_e) using the MCMC procedure described below, then concatenate the sample to form the posterior distribution approximately marginalized over ω_e. The marginalization of the posterior distribution in this way is similar to BayesBag (Bühlmann, 2014; Huggins and Miller, 2019), but with bootstrap sampling applied over models rather than over observations.

Results

Posterior distribution

The diagonal entries in Figure 5 show the prior and posterior marginal distributions for each of the eight parameters in $\mathbf{m}$. One immediate observation is that the posterior distributions for all parameters exhibit a significantly reduced variance relative to the prior distribution. This implies that surface velocity information alone conveys information not only about the sliding law, but also about the parameters of the hydrologic model.

Hydrology parameters

We find that the hydraulic conductivity has a mean value of $k_s = \sim 10^{-3}\ \mathrm{m}^{1-\alpha_s+\beta_s}$, but with a 95% credibility interval of about an order of magnitude in either direction. Unsurprisingly, this parameter exhibits a strong negative correlation with characteristic bedrock bump height h_r: because flux through the inefficient system is a function that increases with both transmissivity and cavity height, an increase in one term can be compensated for by the other. Interestingly, bedrock bump heights most consistent with observations are on the order of meters. We emphasize that this does not imply that average cavity heights are on the order of meters; in fact, the model typically predicts average cavity thickness on the order of tens of centimeters (see Fig. 8). Rather, this result implies that the model should never reach $h = h_r$, at which point the opening rate begins to decouple from velocity. Nonetheless, this rather large bedrock asperity size introduces the *potential* for very large cavities to form. This tendency is offset by a very low bump aspect ratio r, which tends to be <0.1. Conditioned on the hypothesized physics, the observations indicate an inefficient drainage system formed around large and low-slope bedrock features.

A particularly interesting feature of these results is found in the distribution over channel transmissivity k_c. Of the various parameters governing subglacial hydrology, this one is the most poorly constrained. As shown in Figure 8, there are a number of drainage configurations that are consistent with observations, from essentially negligible to extensive. This insensitivity means that a broad array of channel conductivities are possible, and also implies that more research is needed either to quantify the influence of the efficient system on ice dynamics or to directly observe the channel network in order to constrain this value for prognostic modeling. We note that the null hypothesis that the surface velocity is simply insensitive to k_c is not supported by our results, as k_c exhibits strong correlations with parameters (e.g. the sliding law exponent q) that clearly affect the ice velocity.

The englacial porosity e_v controls the speed at which the hydrologic head changes in response to alterations in flux or forcing. We find that this parameter is relatively poorly constrained by observations relative to prior assumptions. This is not surprising: we would expect the influence of this parameter to primarily manifest itself by controlling the rate of change of water pressure and hence velocity. Since we only consider time-integrated quantities here, this characteristic is not well constrained. Nonetheless, this study suggests a porosity that is on the lower end of the plausible spectrum of values. This indeterminacy also motivates the potential utility for time dependent inversion (see 'Discussion').

https://doi.org/10.1017/jog.2020.112 Published online by Cambridge University Press

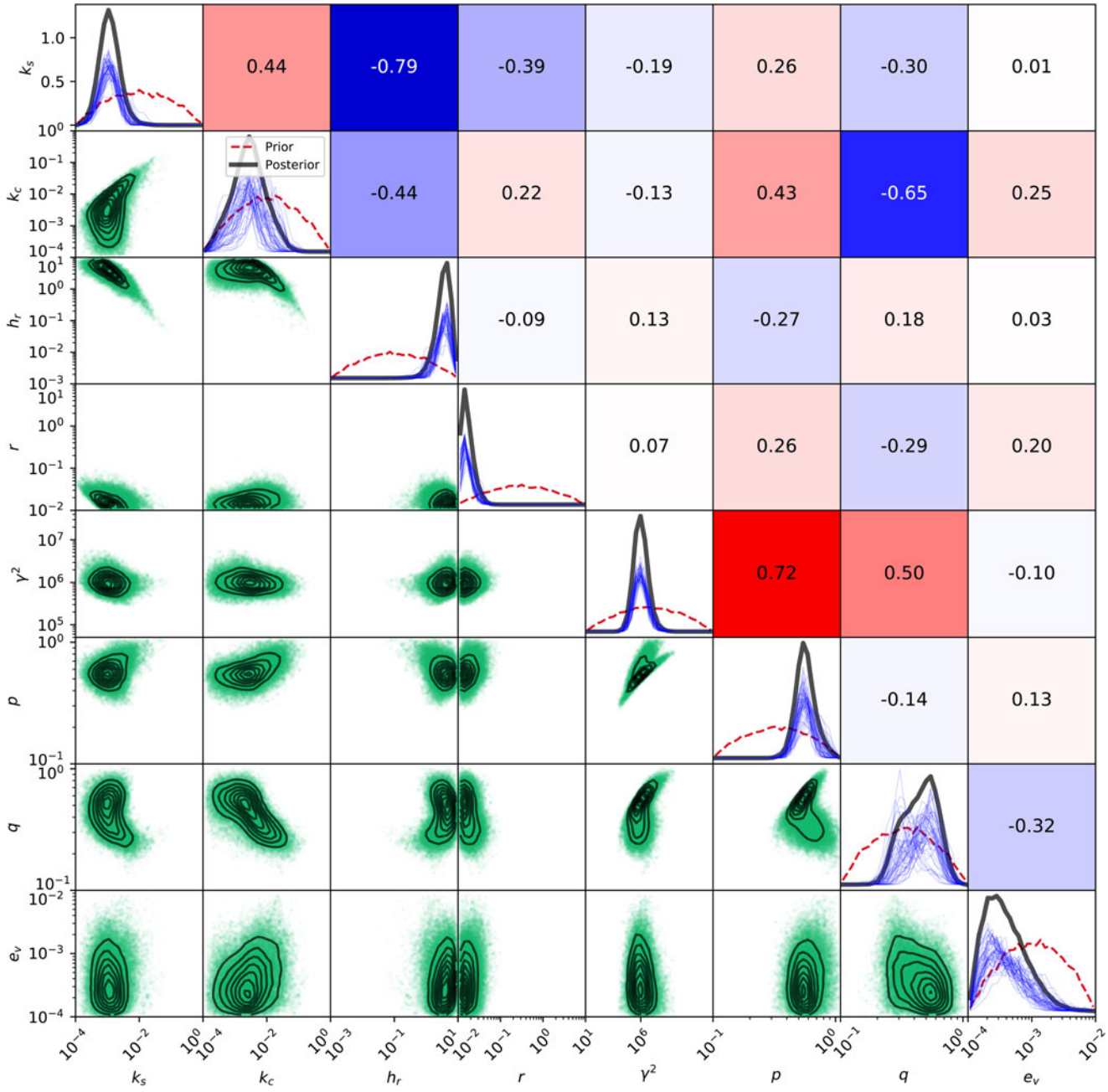


Fig. 5. Posterior distributions. (Diagonal) Marginal distributions for the posterior (black) and prior distribution (red), with BayesBag posteriors in blue (at half scale for clarity). (Below diagonal) Pairwise marginal distributions illustrate correlation structure between parameters. (Above diagonal) Correlation coefficient for each pair of parameters, with red and blue corresponding to positive and negative correlations, respectively.

Sliding law parameters

γ^2 exhibits a strong positive correlation with p. This is simply the result of an increase in p yielding an immediate decrease in the sliding law pressure term (which is typically less than unity), and thus a commensurate increase in γ^2 will yield a similar sliding velocity. This is also true (although to a much lesser extent) of γ^2 and q. γ^2 is strongly constrained by observations, as it sets the scale of glacier velocity, which is directly observable.

The pressure exponent p has a median value of $p = \sim 0.5$, with a relatively small variance. Similarly, the sliding law exponent q also has a median value of $q = \sim 0.5$, but with a significantly larger spread. This spread is distinctly non-Gaussian. Indeed, based on the curvature evident in the joint distributions between q and most other variables, it seems that the distribution over q is the superposition of two overlapping distributions, one associated with a value of q closer to 0.6 (which agrees well with Aschwanden and others (2016), and the other (somewhat less probable) mode $q = \sim 0.2$. This latter secondary mode implies that pseudoplasticity may also be an appropriate bed model. It seems possible that this 'indecision' on the part of the sampler implies that different regions of the glacier might be better fit by different sliding laws, an unsurprising result if some regions are underlain by till and some directly by bedrock. These two modes also lead to different preferred hydrologic parameters: in the pseudoplastic mode, we see a greater transmissivity and a smaller characteristic asperity size (by about an order of magnitude in each case) compared to the less plastic mode.

Posterior predictive distribution

The inference above was performed using a surrogate model, and while the surrogate reproduces predictions from the high-fidelity

https://doi.org/10.1017/jog.2020.112 Published online by Cambridge University Press

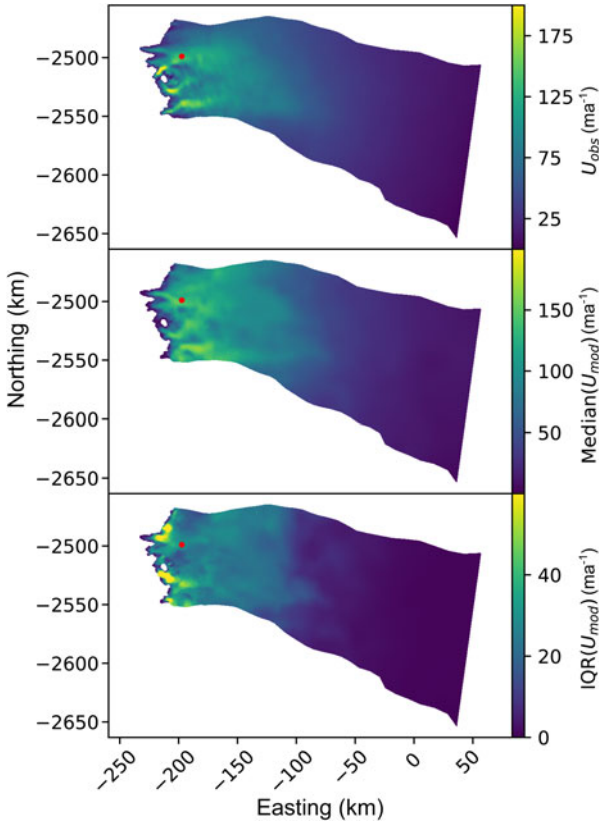


Fig. 6. Posterior predictive distribution. (Top) Observed velocity for study site. (Middle) Median of predicted velocity fields computed by running the high-fidelity model on samples from the posterior distribution from Figure 5. (Bottom) Interquartile range of velocity posterior predictive distribution. The red dot is the location at which a time series is extracted for Figure 9. Note the smaller color scale relative to the top two plots.

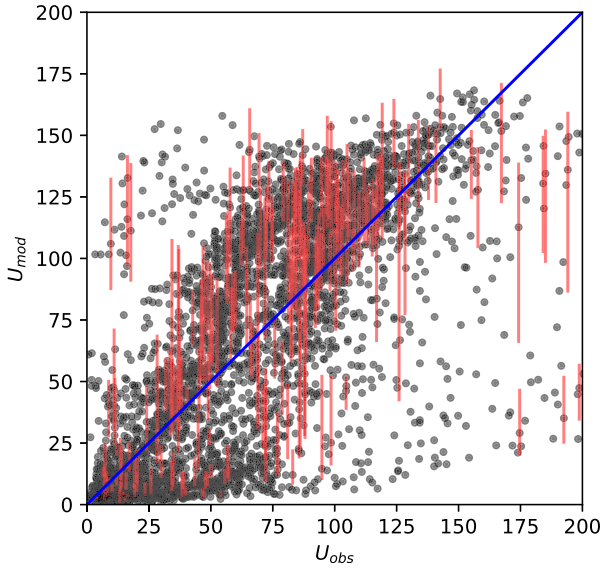


Fig. 7. Observed versus median modeled velocity from 50 ensemble members. The 5th and 95th quantile from the ensemble are given by red lines, plotted for every 20 points. Blue line gives a one-to-one correspondence. Median Bayesian $R^2 = 0.6$.

model in the large majority of circumstances, we have yet to ensure that samples from the posterior distributions inferred using the surrogate produce velocity fields that are consistent with observations when fed back into the high-fidelity model. We note that we do not expect perfect correspondence to observations: the model is necessarily a substantial simplification of a highly complex and heterogeneous physical system. Rather, we seek to verify that samples drawn from the posterior distribution conditioned on the surrogate model lead to velocity predictions by the high-fidelity model that are consistent with observations to the extent that this is possible.

We selected 256 random samples from the posterior distribution shown in Figure 5, and ran the high-fidelity model with these parameter values. Figure 6 shows the mean velocity field as well as the interquartile range, along with the observed velocity. We find that the model fits the observations reasonably well, with an appropriate pattern of fast flow in the outlet glaciers and slow flow in the interior. The transition between these two regimes near the equilibrium line altitude (ELA) is also well-captured by the model. However, the model produces velocity predictions that are somewhat more diffuse than observations, and also fails to match the high-velocities evident in some steep marginal areas. The spread in model predictions is consistent with the imposed observational uncertainty, with an interquartile range (IQR) of between 20 and 30 m a^{-1} over most of the ice sheet below the ELA. Above the ELA, the predicted spread is *lower* than the observational uncertainty in slow flowing regions, indicating that the model is less sensitive to parameter choice in this region than the faster flowing areas downstream. Nonetheless, sliding still makes up ~80% of the modeled (and presumably observed) surface velocity there. Conversely, the model error induced by the surrogate leads to somewhat higher spread in some fast flowing regions near the margin, likely due to these being the places where significant non-linearity in the model (e.g. channelizations, reaching the 'elbow' of the sliding law, etc.) occur, and hence are more challenging to emulate.

It is also useful to establish the degree to which the optimized model explains the observation. Figure 7 shows the velocity observations versus predictions in the form of a scatter plot, as well as the model's predictive spread. Clearly, the model carries substantial predictive power, however there is also substantial variability around the 1 : 1 line. One simple goodness-of-fit metric is the Bayesian R^2 (Gelman and others, 2019), which measures the variance in model predictions relative to the variance of model predictions plus the variance of the residuals. For a model that perfectly models the data, $R^2 = 1$, and for values less than unity R^2 quantifies the fraction of data variance explained by the model. After weighting points by corresponding area, we find a median value of $R^2 = 0.6$, indicating that the model explains 60% of the variance in the observations. Taking this number and the results in Figure 7 together, particularly given the non-Gaussianity of the residuals, we think that the model presented here is underparameterized: a model that allows for some spatial variability in basal conditions would likely fit the data better, and would also be conceptually justifiable, given that different regions of the bed have different geology and sediment cover. However, determining how to parameterize this variability without a wholesale return to the difficulties associated with spatially explicit traction coefficients remains a challenge.

Hydrologic configuration

Although our surrogate model does not provide direct access to the state variables of the hydrologic model, the posterior predictive samples do. In Figure 8, we show the hydraulic potential, channel flux and subglacial cavity size for a weakly, moderately and strongly channelized posterior sample (specifically, posterior samples corresponding to the 16th, 50th and 84th percentile annually integrated flux through the conduit system), all of which produce velocities that are (more or less) equivalently consistent with observations. In the weakly channelized case, large

https://doi.org/10.1017/jog.2020.112 Published online by Cambridge University Press

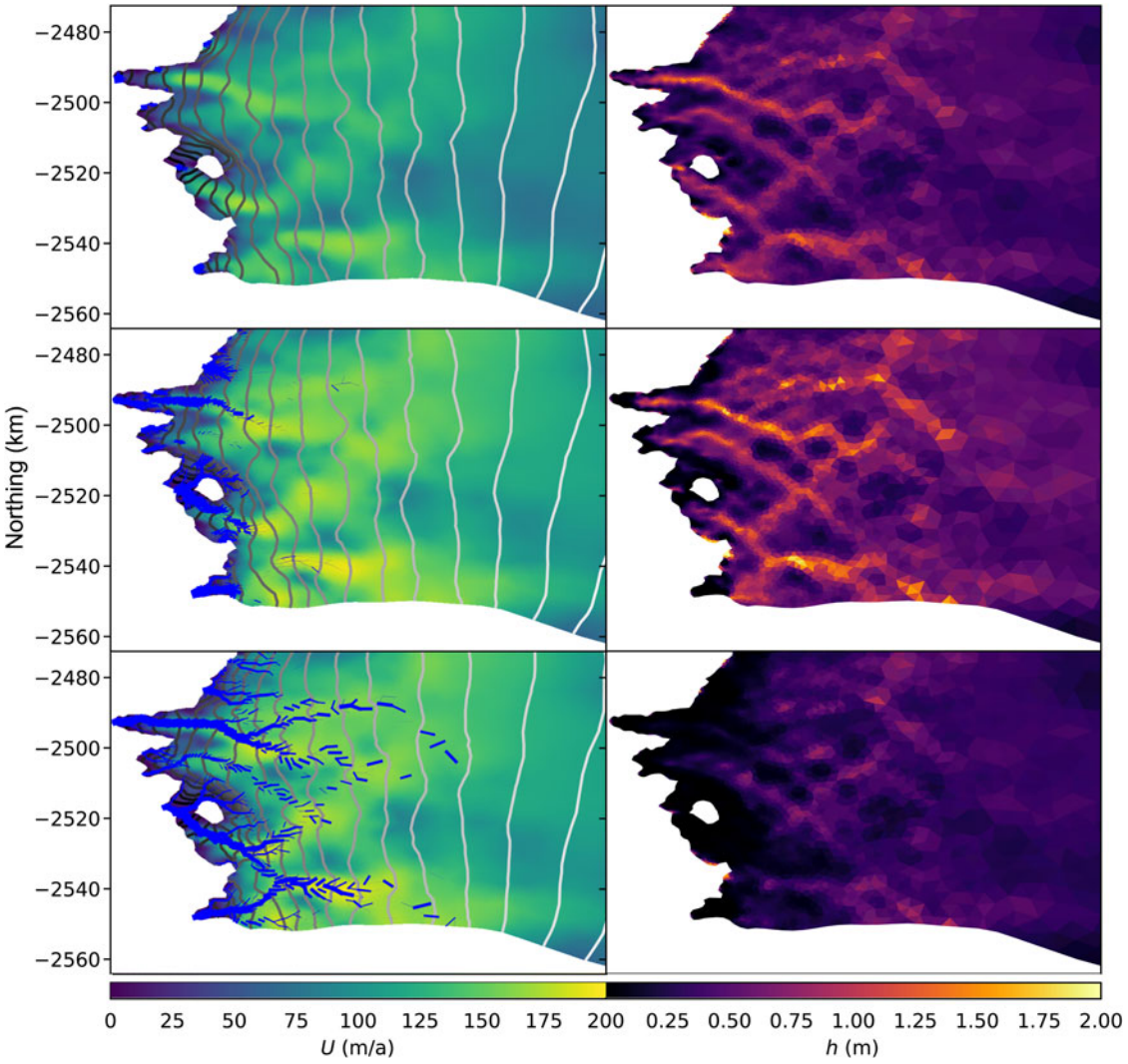


Fig. 8. (Left) Annual average configuration of channels for the simulation according to the 16th (top), 50th (middle) and 84th (bottom) quantile of annually integrated channelized system flux. The widest blue line is ∼300 $m^3\ s^{-1}$ while the smallest visible lines are $10^{-2}\ m^3\ s^{-1}$. Contours show the hydropotential. (Right) Associated distributed water layer thickness fields.

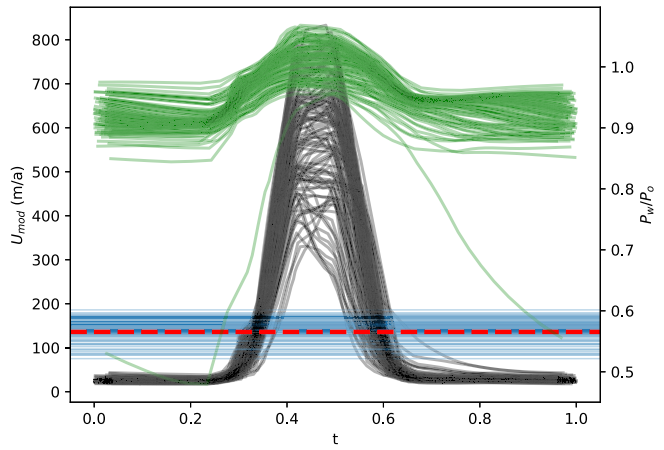


Fig. 9. Time series of velocity (black) over a single year at the red point in Figure 6, modeled annual averages (blue), observed annual average (red) and fraction of overburden (green).

channels occur only near the terminus, where large upstream areas and low overburden pressures allow very large but highly localized channels to form. We note that this low channelization case produces a spacious distributed system, with h frequently reaching 1 m in areas of convergent topography (e.g. the bottom of troughs). A much more well-developed channelized system develops in the moderately channelized sample. However, the inefficient drainage system magnitude remains similar, indicating that despite its greater extent, the channelized system transports relatively little water. Conversely, in the most channelized model run, channels extends nearly all the way to the ELA. The resulting distributed system configuration has much less capacity, with the average cavity size rarely exceeding 0.25 m.

Temporal changes in velocity

Although we constructed the surrogate model and inferred parameters based on time-averaged velocities, the underlying model is still time-dependent and it is of substantial interest to examine the time-dependent behavior of the model. Figure 9 shows the ice sheet's speed and water pressure as a fraction of overburden in the middle of Isunnguata Sermia, coincident with the red dot in Figure 6. Although we find similar qualitative behavior in each simulation, namely an increase in water pressure associated with the onset of meltwater in the spring and a coincident increase in velocity, the peak velocity and speedup duration varies significantly between simulations. This spread in behavior occurs despite annual average velocities that are consistent with observations conditioned on the uncertainty assumptions stated

https://doi.org/10.1017/jog.2020.112 Published online by Cambridge University Press

above. This spread is most acutely driven by uncertainty in the englacial porosity e_v, which plausibly varies by nearly two orders of magnitude, and controls the water pressure rate of change.

In nearly all simulations, water pressure is uniformly high throughout the year, reaching or exceeding overburden pressure during the meltwater season. This uniformly high pressure is consistent with observations for this reason. However, the annual pattern of velocity remains *inconsistent* with the observational record (e.g Andrews and others, 2014; Moon and others, 2014), in particular the lack of a significant winter speed-up. One important future line of inquiry that we are currently undertaking is whether the current model (or any currently proposed hydrologic model) can replicate this time-varying field for *any parameter combination*. If so, then the posterior parameter variance will likely be reduced substantially. However, to answer this question in the negative would call into considerable question the utility of hydrologic models for glaciological modeling.

Discussion

Model selection

To paraphrase Box and others (1987): 'All models are wrong, but some are useful'. Despite the relative robustness of the Bayesian framework here, its ability to quantify parametric uncertainty, and the model's encouraging ability to reproduce many salient features of the velocity observations, we remain skeptical of drawing conclusions that are too certain. This skepticism emerges primarily from the issue of model misspecification: it is almost certainly the case that neither the hydrologic model nor the chosen sliding law (nor even the first-order ice dynamics) are a wholly appropriate approximation of the true physics. This is clearly seen in Figure 7, which indicates that the residuals between the predicted and observed velocities possess systematic (rather than random) biases. As such, the model is wrong, but is it useful? We argue that this study represents a first step toward a defensible mechanism of predicting glacier sliding into the future. However, the physics simulated here are only one possibility, and perhaps not the best possibility. As such, one useful next step toward the goal of a prognostic sliding law would be to repeat the procedure presented here with a variety of candidate models, and to use a formal model selection criterion such as Akaike's information criterion (Akaike, 1998)

$$AIC = 2k - 2\log P(\mathbf{d}|\mathbf{m}), \tag{54}$$

which estimates the relative information loss of a set of candidate models with respect to the true data generating process, to select between them. Indeed, we can do this very simply for the model presented here and, for example, an unregularized inversion of basal traction of the type popularized in MacAyeal (1993). In the above, k is the number of parameters, which in the case of this study is $k = 9$ (including the data variance). In the spatially varying inversion, $k = 4042$, which is the number of gridcells plus one. In the study presented above, the log probability at the a posteriori most probable parameter estimate is (to a constant that cancels when comparing AIC between two models) $\log P(\mathbf{d}|\mathbf{m}) \propto -74$. In the case of the spatially varying inversion, the log-likelihood is effectively zero, representing a nearly perfect fit to the data. Thus, we have $AIC \approx 166$ for the model presented here, and $AIC \approx 8042$ for a spatially varying inversion (although this number will decrease substantially in the presence of regularization, which induces a spatial covariance that decreases the number of effective parameters). Thus, although the model presented here does not fit the data as well, this disadvantage is more than offset by its simplicity with respect to minimizing the loss of information relative to a perfect model of glacier physics.

Nonetheless, it is unlikely that the model presented here is the optimal one. We intend to explore this question systematically in the future by examining both alternative hydrologic and sliding parameterizations, as well as (re-)introducing spatially varying parameters in such a way that a model selection criterion such as AIC is optimized. It is highly likely that an optimal model accounts both for parameters that vary subject to a to-be-determined smoothness constraint coupled with more advanced physical models. The framework suggested here provides a consistent methodology for coupled model optimization that can be applied to any model configuration, without the need for the implementation of time-dependent adjoints, which may be time-consuming and numerically challenging to implement.

Including time-dependent observations

Another important consideration is that we use observations that are averaged over the year, thus likely discarding important information contained in time rates of change and temporal patterns. Fortunately, the procedure presented here is easily amenable to time-dependent inversion. The only substantive difference is in the construction of the surrogate (rather than train a network to predict the coefficients of the eigenglaciers presented in Figure 2, these basis functions must be explicit in time as well) and the likelihood function (which must now include observations at different points in time and also explicitly model spatio-temporal covariance).

Supplementary datasets

In addition to time-varying data, it will also be important to augment velocity observations with other measurements. In particular, including borehole measurements of water pressure would likely yield a much smaller admissible parameter space by constraining the rate of change in pointwise storage in the coupled sub-/englacial hydrologic system. Similarly, radar-derived estimates of channel extent (Livingstone and others, 2017) would provide a statistical target for determining which of the samples presented in Figure 8 is most consistent with reality. The Bayesian framework offers a natural mechanism for incorporating diverse observations into the likelihood model, and the wide availability of such observations represents a major avenue for improvement in parameter estimation for sliding prediction.

Spatial generalization

Finally, it remains to be seen whether the parameter distributions inferred here are transferable to other parts of Greenland, and whether the associated models can exhibit similar fidelity to data. It stands to reason that parameters that likely depend on the underlying geology, such as average asperity height $\bar{h}_r$, the ratio of asperity height to spacing r and the traction coefficient γ^2 should vary across Greenland, while parameters that are more intrinsic to the ice configuration, such as hydraulic conductivities, sliding law exponents and englacial porosity should remain close to constant. At the very least, this study supports the notion that when parameters vary across space, it is possible that they may do so at geologically relevant spatial scales while still maintaining good fidelity to observations.

Conclusions

We developed a coupled model of subglacial hydrology and glacier flow, and used it to infer the posterior probability distribution of eight key model parameters. Because the model is computationally expensive, this inference was non-trivial. We first had to construct a large ensemble of concurrent model runs, with ensemble members determined by sampling from the space of admissible parameter combinations. We then used the resulting samples to train an artificial neural network to act as a surrogate for expensive model

https://doi.org/10.1017/jog.2020.112 Published online by Cambridge University Press

physics. Because the neural network was not a perfect reproduction of model physics, we introduced a double bootstrap aggregation approach to both smooth the surrogate's response to different parameters, and also to robustly account for model error. With the surrogate in hand, we ran an MCMC method to draw samples from the posterior distribution given an observed annual average velocity field. We found that the velocity observation provided substantial information about all of the model parameters relative to a prior distribution, although some were more strongly constrained than others. In particular, we found that both transmissivity of the subglacial conduit network and the englacial porosity remain highly uncertain, and this uncertainty leads to a qualitative variety of solutions that are consistent with observations. Nonetheless, we find that this eight parameter model can account for 60% of variance in the observational dataset, and produces velocity fields that are spatially consistent with observations.

Acknowledgements. We acknowledge Ruth Mottram for providing the HIRHAM surface mass-balance fields. We thank Mauro Werder who provided key insights when reimplementing GlaDS in FEniCS. We acknowledge Scientific Editor Michelle Koutnik and three anonymous reviewers, whose insights and suggestions greatly improved the quality of this manuscript. A.A., M.A.F. and D.J.B. were supported by NASA Cryosphere Grant NNX17AG65G. A Jupyter Notebook in which the ensemble of surrogates is constructed and MCMC sampling performed can be found at https://github.com/douglas-brinkerhoff/glacier-hydrology-emulator-ensemble.

References

Akaike H (1998) Information theory and an extension of the maximum likelihood principle. In Parzen E, Tanabe K and Kitagawa G eds. *Selected Papers of Hirotugu Akaike*. New York, NY: Springer, pp. 199–213.

Andrews LC and 7 others (2014) Direct observations of evolving subglacial drainage beneath the Greenland Ice Sheet. *Nature* **514**(7520), 80–83.

Aschwanden A and 7 others (2019) Contribution of the Greenland Ice Sheet to sea level over the next millennium. *Science Advances* **5**(6), eaav9396.

Aschwanden A, Fahnestock MA and Truffer M (2016) Complex Greenland outlet glacier flow captured. *Nature Communications* **7**, 10524. doi: 10.1038/ncomms10524.

Ba JL, Kiros JR and Hinton GE (2016) Layer normalization. *arXiv preprint arXiv:1607.06450*.

Balay S and 21 others (2017) PETSc users manual revision 3.8. Technical report, Argonne National Lab. (ANL), Argonne, IL (United States).

Bindschadler R (1983) The importance of pressurized subglacial water in separation and sliding at the glacier bed. *Journal of Glaciology* **29**(101), 3–19.

Box GE, Draper NR and others (1987) *Empirical Model-Building and Response Surfaces*, **vol. 424**. New York: Wiley.

Breiman L (1996) Bagging predictors. *Machine Learning* **24**(2), 123–140.

Brinkerhoff D and Johnson J (2015) Dynamics of thermally induced ice streams simulated with a higher-order flow model. *Journal of Geophysical Research: Earth Surface* **120**(9), 1743–1770.

Brinkerhoff DJ, Meyer CR, Bueler E, Truffer M and Bartholomaus TC (2016) Inversion of a glacier hydrology model. *Annals of Glaciology* **57** (72), 84–95.

Budd W, Keage P and Blundy N (1979) Empirical studies of ice sliding. *Journal of Glaciology* **23**(89), 157–170.

Bueler E and van Pelt W (2015) Mass-conserving subglacial hydrology in the Parallel Ice Sheet Model version 0.6. *Geoscientific Model Development* **8**(6), 1613–1635. doi: 10.5194/gmd-8-1613-2015.

Bühlmann P (2014) Discussion of big Bayes stories and BayesBag. *Statistical Science* **29**(1), 91–94.

Bui-Thanh T, Ghattas O, Martin J and Stadler G (2013) A computational framework for infinite-dimensional Bayesian inverse problems Part I: the linearized case, with application to global seismic inversion. *SIAM Journal on Scientific Computing* **35**(6), A2494–A2523.

Butcher JC (2016) *Numerical Methods for Ordinary Differential Equations*. Chichester, UK: John Wiley & Sons.

Clyde M and Lee HJC (2001) Bagging and the Bayesian Bootstrap. In *the Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics*.

Cornford SL and 14 others (2015) Century-scale simulations of the response of the West Antarctic Ice Sheet to a warming climate. *The Cryosphere* **9**(4), 1579–1600. doi: 10.5194/tc-9-1579-2015.

De Fleurian B and 6 others (2014) A double continuum hydrological model for glacier applications. *The Cryosphere* **8**, 137–153.

Downs JZ, Johnson JV, Harper JT, Meierbachtol T and Werder MA (2018) Dynamic hydraulic conductivity reconciles mismatch between modeled and observed winter subglacial water pressure. *Journal of Geophysical Research: Earth Surface* **123**(4), 818–836.

Favier L and 8 others (2014) Retreat of Pine Island Glacier controlled by marine ice-sheet instability. *Nature Climate Change* **5**(2), 1–5. doi: 10.1038/nclimate2094.

Fowler A (1979) A mathematical approach to the theory of glacier sliding. *Journal of Glaciology* **23**(89), 131–141.

Fowler A (1987) Sliding with cavity formation. *Journal of Glaciology* **33**(115), 255–267.

Gelman A, Goodrich B, Gabry J and Vehtari A (2019) R-squared for Bayesian regression models. *The American Statistician* **73**(3), 307–309.

Geuzaine C and Remacle JF (2009) Gmsh: a 3-D finite element mesh generator with built-in pre- and post-processing facilities. *International Journal for Numerical Methods in Engineering* **79**(11), 1309–1331.

Gillet-Chaulet F and 8 others (2012) Greenland ice sheet contribution to sea-level rise from a new-generation ice-sheet model. *The Cryosphere* **6**(6), 1561–1576.

Girolami M and Calderhead B (2011) Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(2), 123–214.

Glorot X, Bordes A and Bengio Y (2011) Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 315–323.

Goodfellow I, Bengio Y and Courville A (2016) *Deep Learning*. Cambridge, MA, USA: MIT Press.

Griffiths DF (1997) The 'no boundary condition' outflow boundary condition. *International Journal for Numerical Methods in Fluids* **24**(4), 393–411.

Habermann M, Maxwell D and Truffer M (2012) Reconstruction of basal properties in ice sheets using iterative inverse methods. *Journal of Glaciology* **58**(210), 795–807. doi: 10.3189/2012JoG11J168.

Harrington JA, Humphrey NF and Harper JT (2015) Temperature distribution and thermal anomalies along a flowline of the Greenland ice sheet. *Annals of Glaciology* **56**(70), 98–104.

He K, Zhang X, Ren S and Sun J (2016) Identity mappings in deep residual networks. In *European Conference on Computer Vision*, Springer, pp. 630–645.

Hoffman MJ and 9 others (2016) Greenland subglacial drainage evolution regulated by weakly connected regions of the bed. *Nature Communications* **7**(1), 1–12.

Huggins JH and Miller JW (2019) Using bagged posteriors for robust inference and model criticism. *arXiv preprint arXiv:1912.07104*.

Iken A (1981) The effect of the subglacial water pressure on the sliding velocity of a glacier in an idealized numerical model. *Journal of Glaciology* **27**(97), 407–421.

Iken A and Bindschadler RA (1986) Combined measurements of subglacial water pressure and surface velocity at Findelengletscher, Switzerland, conclusions about drainage system and sliding mechanism. *Journal of Glaciology* **32**(110), 101–119.

Irarrazaval I and 5 others (2019) Bayesian inference of subglacial channel structures from water pressure and tracer-transit time data: a numerical study based on a 2-D geostatistical modeling approach. *Journal of Geophysical Research: Earth Surface* **124**(6), 1625–1644. doi: 10.1029/2018JF004921.

Joughin I, Smith BE and Howat IM (2018) A complete map of Greenland ice velocity derived from satellite data collected over 20 years. *Journal of Glaciology* **64**(243), 1–11. doi: 10.1017/jog.2017.73.

Joughin I, Smith BE, Shean DE and Floricioiu D (2014) Brief communication: further summer speedup of Jakobshavn Isbræ. *The Cryosphere* **8**(1), 209–214. doi: 10.5194/tc-8-209-2014.

Kamb B (1991) Rheological nonlinearity and flow instability in the deforming bed mechanism of ice stream motion. *Journal of Geophysical Research: Solid Earth* **96**(B10), 16585–16595.

Kass RE, Carlin BP, Gelman A and Neal RM (1998) Markov chain Monte Carlo in practice: a roundtable discussion. *The American Statistician* **52**(2), 93–100.

Kingma DP and Ba J (2014) Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koziol CP and Arnold N (2018) Modelling seasonal meltwater forcing of the velocity of land-terminating margins of the Greenland ice sheet. *The Cryosphere* **12**(3), 971–991. doi: 10.5194/tc-12-971-2018.
Larour E and 8 others (2014) Inferred basal friction and surface mass balance of the Northeast Greenland Ice Stream using data assimilation of ICESat (Ice Cloud and land Elevation Satellite) surface altimetry and ISSM (Ice Sheet System Model). *The Cryosphere* **8**(6), 2335–2351. doi: 10.5194/tc-8-2335-2014.
Livingstone SJ, Chu W, Ely JC and Kingslake J (2017) Paleofluvial and subglacial channel networks beneath Humboldt Glacier, Greenland. *Geology* **45**(6), 551–554.
Lliboutry L (1968) General theory of subglacial cavitation and sliding of temperate glaciers. *Journal of Glaciology* **7**(49), 21–58.
Logg A, Mardal KA and Wells G (2012) *Automated Solution of Differential Equations by the Finite Element Method: The FEniCS Book*, **vol. 84.** Berlin: Springer Science & Business Media.
MacAyeal DR (1993) A tutorial on the use of control methods in ice-sheet modeling. *Journal of Glaciology* **39**(131), 91–98. doi: 10.3189/S0022143000015744.
Maier N, Humphrey N, Harper J and Meierbachtol T (2019) Sliding dominates slow-flowing margin regions, Greenland Ice Sheet. *Science Advances* **5**(7), eaaw5406. doi: 10.1126/sciadv.aaw5406.
Milne-Thomson LM, Abramowitz M and Stegun I (1972) *Handbook of Mathematical Functions*, US Government Printing Office.
Minchew B and 7 others (2016) Plastic bed beneath Hofsjökull Ice Cap, central Iceland, and the sensitivity of ice flow to surface meltwater flux. *Journal of Glaciology* **62**(231), 147–158. doi: 10.1017/jog.2016.26.
Moon T and 6 others (2014) Distinct patterns of seasonal Greenland glacier velocity. *Geophysical Research Letters* **41**(20), 7209–7216.
Morlighem M and 5 others (2010) Spatial patterns of basal drag inferred using control methods from a full-Stokes and simpler models for Pine Island Glacier, West Antarctica. *Geophysical Research Letters* **37**(14).
Morlighem M and 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), 11–051.
Mottram R and 6 others (2017) Surface mass balance of the Greenland ice sheet in the regional climate model HIRHAM5: present state and future prospects. *Low Temperature Science* **75**, 105–115. doi: 10.14943/lowtemsci.75.105.
Mouginot J and 8 others (2019) Forty-six years of Greenland Ice Sheet mass balance from 1972 to 2018. *Proceedings of the National Academy of Sciences* **116**(19), 9239–9244.
Papanastasiou TC, Malamataris N and Ellwood K (1992) A new outflow boundary condition. *International Journal for Numerical Methods in Fluids* **14**(5), 587–608.
Parker RL (1994) *Geophysical Inverse Theory*. **vol. 1.** Princeton, NJ, USA: Princeton University Press.
Paszke A and 20 others (2019) Pytorch: An imperative style, high-performance deep learning library. In Wallach H, Larochelle H, Beygelzimer A, de Alché-Buc F, Fox E and Garnett R eds. *Advances in Neural Information Processing Systems* **32**, 8024–8035.
Pattyn F (2003) A new three-dimensional higher-order thermomechanical ice sheet model: basic sensitivity, ice stream development, and ice flow across subglacial lakes. *Journal of Geophysical Research: Solid Earth* **108**(B8).
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.
Pimentel S and Flowers GE (2011) A numerical study of hydrologically driven glacier dynamics and subglacial flooding. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **467**(2126), 537–558.
Roberts GO, Rosenthal JS and others (2001) Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science* **16**(4), 351–367.
Rubin DB (1981) The Bayesian bootstrap. *The Annals of Statistics* **9**(1), 130–134.
Schoof C (2005) The effect of cavitation on glacier sliding. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **461** (2055), 609–627.
Shlens J (2014) A Tutorial on Principal Component Analysis. *arXiv preprint arXiv:1404:1100*.
Sirovich L and Kirby M (1987) Low-dimensional procedure for the characterization of human faces. *JOSA A* **4**(3), 519–524.
Sobol IM, Asotsky D, Kreinin A and Kucherenko S (2011) Construction and comparison of high-dimensional Sobol' generators. *Wilmott* **2011**(56), 64–79.
Sommers A, Rajaram H and Morlighem M (2018) SHAKTI: subglacial hydrology and kinetic, transient interactions v1. 0. *Geoscientific Model Development* **11**(7), 2955–2974.
Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* **15**(56), 1929–1958.
Tarantola A (2005) *Inverse Problem Theory and Methods for Model Parameter Estimation*, **vol. 89.** Philadelphia: SIAM.
Tarasov L, Dyke AS, Neal RM and Peltier WR (2012) A data-calibrated distribution of deglacial chronologies for the North American ice complex from glaciological modeling. *Earth and Planetary Science Letters* **315**, 30–40.
Weertman J (1957) On the sliding of glaciers. *Journal of Glaciology* **3**(21), 33–38.
Weertman J (1964) The theory of glacier sliding. *Journal of Glaciology* **5**(39), 287–303.
Werder MA, Hewitt IJ, Schoof CG and Flowers GE (2013) Modeling channelized and distributed subglacial drainage in two dimensions. *Journal of Geophysical Research: Earth Surface* **118**(4), 2140–2158.
Zienkiewicz OC, Taylor RL and Zhu JZ (2005) *The Finite Element Method: Its Basis and Fundamentals*. Oxford, UK: Elsevier.

Appendix A Symbol tables

See Tables 2 and 3.

Table 2. Symbols used in defining the high-fidelity model

Symbol	Value	Units	Description
A	10^{-16}	$\mathrm{Pa}^{-n}\ \mathrm{a}^{-1}$	Ice softness
α_s	$\frac{5}{4}$		Sheet thickness flux exponent
α_c	$\frac{5}{4}$		Channel size flow exponent
β_s	$\frac{3}{2}$		Sheet potential flux exponent
β_c	$\frac{3}{2}$		Channel potential flux exponent
β^2		$\mathrm{Pa}^{1-p}\ \mathrm{m}^{-q}\ \mathrm{a}^{-q}$	Traction coefficient
e_v			Englacial porosity
$\dot{\boldsymbol{\epsilon}}$		a^{-1}	Strain rate tensor
$\dot{\epsilon}_0$	10^9	a^{-1}	Strain rate regularization
$\dot{\epsilon}_{\mathrm{II}}$		a^{-1}	Second invariant of strain rate tensor
g	9.81	$\mathrm{m\ s}^{-2}$	Gravitational acceleration
γ		Pa	Scaled traction coefficient
Γ_{z_b}			Basal boundary
Γ_L			Non-terminus lateral boundary
Γ_{z_s}			Surface boundary
Γ_T			Terminal boundary
h		m	Average cavity thickness
$\bar{h}_r$		m	Average bedrock bump size
H		m	Ice thickness
η		Pa a	Ice viscosity
L	3.35×10^5	$\mathrm{J\ kg}^{-1}$	Latent heat of fusion
k_c		$\mathrm{m}^{2-2\alpha_c+\beta_c}\ \mathrm{a}^{-1}\ \mathrm{Pa}^{1-\beta}$	Channel conductivity
k_s		$\mathrm{m}^{1-\alpha_s+\beta_s}\ \mathrm{a}^{-1}\ \mathrm{Pa}^{1-\beta}$	Sheet conductivity
$\dot{m}$		$\mathrm{m\ a}^{-1}$	Specific meltwater
m_c		$\mathrm{m}^2\ \mathrm{a}^{-1}$	Channel-cavity meltwater exchange
n	3		Glen's flow law exponent
n_p			Number of points in FEM mesh
$\mathbf{n}$			Normal vector
N		Pa	Effective pressure
p			Sliding law pressure exponent
P_0		Pa	Ice overburden pressure
P_w		Pa	Water pressure
Ψ			Finite element basis function
q			Sliding law velocity exponent
$\mathbf{q}$		$\mathrm{m}^2\ \mathrm{a}^{-1}$	Cavity flux
Q		$\mathrm{m}^3\ \mathrm{a}^{-1}$	Channel discharge
r			Ratio of asperity height to spacing
ρ_i	917	$\mathrm{kg\ m}^{-3}$	Ice density
ρ_w	1000	$\mathrm{kg\ m}^{-3}$	Freshwater density
S		m^2	Channel size
Scale (N)	10^6	Pa	Effective pressure scale
Scale ($\mathbf{u}$)	50	$\mathrm{m\ a}^{-1}$	Velocity scale
σ_h^2		m	logarithmic std. dev. of bed asperity size
s			Thickness-scaled vertical coordinate

(*Continued*)

Table 2. (*Continued.*)

Symbol	Value	Units	Description
τ'		Pa	Hydrostatic deviatoric stress tensor
$\boldsymbol{\tau}_d$		Pa	Driving stress
$\mathbf{u}$		m a^{-1}	Horizontal velocity vector
$\bar{\mathbf{u}}$		m a^{-1}	Vertically-averaged velocity vector
$\mathbf{u}_d$		m a^{-1}	Shear velocity
ϕ		Pa	Hydraulic potential
ξ			Lagrange basis function
Ξ		J m^{-1} a^{-1}	Dissipative heating
Ψ		J m^{-1} a^{-1}	Pressure heating
z_b		m	Bed elevation
z_s		m	Surface elevation
Ω			3-D ice domain
$\delta\Omega_{ij}$			Boundary between subdomains i and j
$\bar{\Omega}$			Horizontal extent of ice

Table 3. Symbols used in defining the surrogate model and MCMC sampling

Symbol	Description
a	MCMC acceptance probability
$\hat{\mathbf{a}}_l$	Output of linear transform
$\mathbf{a}_l$	Output of layer normalization
α_l	Layer normalization scaling
α	Prior parameter
$\mathbf{b}_l$	Trainable bias vector
β_l	Layer normalization offset
β	Prior parameter
Bound_L	Parameter lower bound
Bound_U	Parameter upper bound
c	Number of retained eigenglaciers
$d(x, x')$	Distance
$\mathbf{d}$	Data vector
Δ	MCMC step size
f	Fraction of explained variance
$\mathcal{F}$	High-fidelity model
$\mathcal{G}$	Surrogate model
$\mathbf{h}_l$	Residual sum
$\hat{H}$	Approximate Hessian
k	Parameter vector length
K	Number of observations per subdomain matrix
l	Length scale of data correlation
L	Number of ANN blocks
$\mathbf{m}$	Vector of model parameters
$P_\text{em}(\mathbf{m})$	Evaluation sampling distribution
$Q(\cdot \mid \cdot)$	MCMC proposal function
$r(x)$	Data residual function
R	Dropout matrix
$\mathbf{r}$	Residual vector
ρ_d	Data density
s	Explained variance threshold
$\sigma(x, x')$	Covariance function
σ_{obs}	Data white noise std.
σ_{cor}	Data correlated noise std.
$\mathcal{S}$	Model empirical covariance
$\hat{\Sigma}$	Data covariance matrix
Σ	Area-scaled data covariance matrix
V	Matrix of ensemble eigenvectors
λ	Eigenglacier coefficients
Λ	Diagonal matrix of ensemble eigenvalues
θ	Surrogate model trainable parameters
W_l	Trainable weight matrix
$\hat{\mathbf{z}}_l$	Output of activation
$\mathbf{z}$	Output of dropout
ω_d	Vector of bootstrap weights for surrogate training
ω_e	Vector of bootstrap weights for aggregation

Appendix B Discretization and numerical solution of the high-fidelity model

Momentum balance

We discretize the momentum equations using a mixed finite element method. Introducing a terrain-following ς-coordinate

$$\varsigma = \frac{z_s - z}{H}, \tag{B1}$$

where z_s is the upper ice surface, H is the ice thickness and z is the vertical coordinate, we decompose the domain as $\Omega = \bar{\Omega} \times [0, 1]$. Introducing a test function $\Psi(x, y, \varsigma)$, multiplying it by Eqn (1), and integrating over the domain, we obtain the following variational formulation: find $\mathbf{u} \in U$, such that

$$\begin{aligned} 0 = &\int_{\bar{\Omega}} \int_1^0 (\bar{\nabla}\Psi + \partial_\varsigma \Psi \bar{\nabla}\varsigma) \cdot \tau' H \, \mathrm{d}\varsigma \, \mathrm{d}\Omega \\ &- \int_{\Gamma_l} \int_1^0 \Psi \cdot \tau' \cdot \mathbf{n}\varsigma \, \mathrm{d}\Gamma \\ &- \int_{\bar{\Omega}} \int_1^0 \Psi \cdot \tau_d \, \mathrm{d}\varsigma \, \mathrm{d}\Omega \\ &+ \int_{\bar{\Omega}} \Psi \cdot \beta^2 N^p \|\mathbf{u}\|_2^{p-1} \mathbf{u} \, \mathrm{d}\Omega|_{\varsigma=1}, \\ &\forall \Psi \in V, \end{aligned} \tag{B2}$$

where $\bar{\nabla}$ is the gradient operator in the two map-plane dimensions and $\tau_\text{d} = \rho g H \bar{\nabla} z_s$ is the gravitational driving stress, and with U, $V \in W^{1,2}(\Omega)$, and where $W^{1,2}$ is a Sobolev space over the model domain Ω. To discretize the weak form, we restrict Ψ to a finite subset of V:

$$\Psi \in \hat{V} \subset V, \tag{B3}$$

where

$$\hat{V} = V_{\bar{\Omega}} \otimes V_{\bar{\Omega}} \otimes V_0 \otimes V_0 \tag{B4}$$

is a tensor product of function spaces defined over $\bar{\Omega}$ and [0, 1], respectively. For $V_{\bar{\Omega}}$, we use the continuous piecewise linear Lagrange basis $\{\xi_i\}_{i=1}^{n_p}$, where n_p is the number of gridpoints in a mesh defined on $\bar{\Omega}$ (Zienkiewicz and others, 2005). For V_0, we utilize the basis set

$$\left\{\psi_1 = 1, \psi_2 = \frac{1}{n+1}[(n+2)\varsigma^{n+1} - 1]\right\}. \tag{B5}$$

Using the standard Galerkin approximation $\hat{U} = \hat{V}$, we introduce the ansatz solution

$$\mathbf{u}(x, y, \varsigma) = \sum_{i \in n} \left[\bar{\mathbf{u}}_i + \mathbf{u}_{\text{d},i} \frac{1}{n+1}[(n+2)\varsigma^{n+1} - 1]\right] \xi_i(x, y), \tag{B6}$$

where $\bar{\mathbf{u}}$ is the vertically averaged velocity, and $\mathbf{u}_\text{d}$ is the deviation from that average induced by vertical shearing. The above expression implies that the solution in the vertical dimension is a linear combination of a constant (i.e. the shallow-shelf approximation) and a polynomial of order $n + 1$, which corresponds to the analytical solution of the isothermal shallow ice approximation. As such, this discretization scheme allows for the exact recovery of both shallow ice and shallow shelf solutions in the appropriate asymptotic regimes, while not requiring the formation of a full 3-D mesh (the ς dimension always has one layer, ranging over $\varsigma \in [0, 1]$). Intercomparison has shown that approximate solutions produced by this method agree well with more

https://doi.org/10.1017/jog.2020.112 Published online by Cambridge University Press

expensive 3-D discretizations of the hydrostatic Stokes' equations (Brinkerhoff and Johnson, 2015), but we emphasize that this method does not 'converge' to the solution of the Blatter–Pattyn equations, as the fixed basis does not allow for either h or p refinement.

Hydrology

We seek to solve Eqn (9) on each subdomain $\bar{\Omega}_j$ and Eqn (20) on each subdomain boundary Γ_{ij}. To discretize, we multiply both by the same test function θ and integrate by parts, leading to the variational problem: find $\phi \in \Phi$ such that

$$0 = \sum_j \int_{\bar{\Omega}_j} \theta \frac{e_\mathrm{v}}{\rho_\mathrm{w} g} \frac{\partial \phi}{\partial t} - \nabla\theta \cdot \mathbf{q} + \theta(\mathcal{C} - \mathcal{O} - m)\, \mathrm{d}\Omega + \sum_j \sum_{i<j} \int_{\Gamma_{ij}} -\frac{\partial \theta}{\partial S} Q + \theta\left(\frac{\Xi - \Pi}{L}\left(\frac{1}{\rho_i} - \frac{1}{\rho_\mathrm{w}}\right) - \mathcal{C}_\mathrm{c}\right) \mathrm{d}\Gamma \quad \forall \theta \in \Theta,$$

where $\Phi,\ \Theta \in W^{1,2}(\bar{\Omega})$. We have used natural boundary conditions, continuity between channel segments, and continuity between the sheet and edges to cancel boundary terms. To discretize this equation, we restrict $\hat{\Phi} \subset \Phi$, $\hat{\Theta} \subset \Theta$ to function spaces defined by the continuous piecewise linear Lagrange basis.

Although Eqn (12) and (18) are ordinary differential equations, it is convenient to put them in a variational form: find $h \in Z$, $S \in \Sigma$ such that

$$0 = \sum_j \int_{\Omega_j} \left[\frac{\partial h}{\partial t} - \mathcal{O} + \mathcal{C}\right] w \mathrm{d}\Omega + \sum_i \sum_{j<i} \int_{\Gamma_{ij}} \left[\frac{\partial S}{\partial t} - \frac{\Xi - \Pi}{\rho_i L} + \mathrm{C}_\mathrm{c}\right] \nu \mathrm{d}\Gamma, \quad \forall w \in Z,\ \nu \in \Sigma, \tag{B7}$$

with $Z \in L^2(\bar{\Omega})$, $\Sigma \in L^2(\Gamma)$. We restrict these functions spaces to a constant basis over each subdomain (i.e. order-zero discontinuous Galerkin over both mesh elements and edges).

Numerical solution

We use finite element software FEniCS (Logg and others, 2012) to compile all of the variational problems described above. We solve the problems over an isotropic computational mesh with variable resolution, ranging from ~250 m diameter elements near the margins to ~1 km near the ice divide. The mesh was created using a Delaunay Triangulation routine in the package gmsh (Geuzaine and Remacle, 2009).

We use the implicit Euler method (Butcher, 2016) to discretize all time steps. Although only accurate to $\mathcal{O}(\Delta t)$, we have found that the implicit Euler method leads to substantially improved stability in the non-linear cavity and conduit equations compared to higher order explicit (e.g. Runge–Kutta) or semi-implicit (e.g. Crank–Nicholson) methods. We deal with the integral in the opening rate $\mathcal{O}$ using Gauss–Legendre numerical quadrature of order seven (Milne-Thomson and others, 1972).

Because the system of equations are non-linear and strongly coupled, we perform Newton's method on a single residual encompassing all seven equations *simultaneously*, using a Jacobian inferred from an automated symbolic computation of the Gâteaux derivative. Note that this implies that we must solve a large non-linear system at each time step. Because of the poor conditioning of the problem, we have found direct solution of the linear system of equations for each Newton update is required. To this end, we use MUMPS, which is implemented in PETSc (Balay and others, 2017).

We employ an adaptive time-stepping procedure that ensures convergence: the time step is slowly increased until Newton's method fails to produce a residual with a specified relative tolerance (10^{-6}) within a certain number of iterations, at which point the time step is reduced by half and the solver tries again until convergence is achieved, after which time-stepping proceeds.

Appendix C Manifold Metropolis-adjusted Langevin algorithm

MCMC methods operate by performing a random walk in parameter space, with candidate for the next position $\hat{\mathbf{m}}_{t+1}$ determined according to a proposal distribution $Q(\,\cdot \mid \cdot\,)$

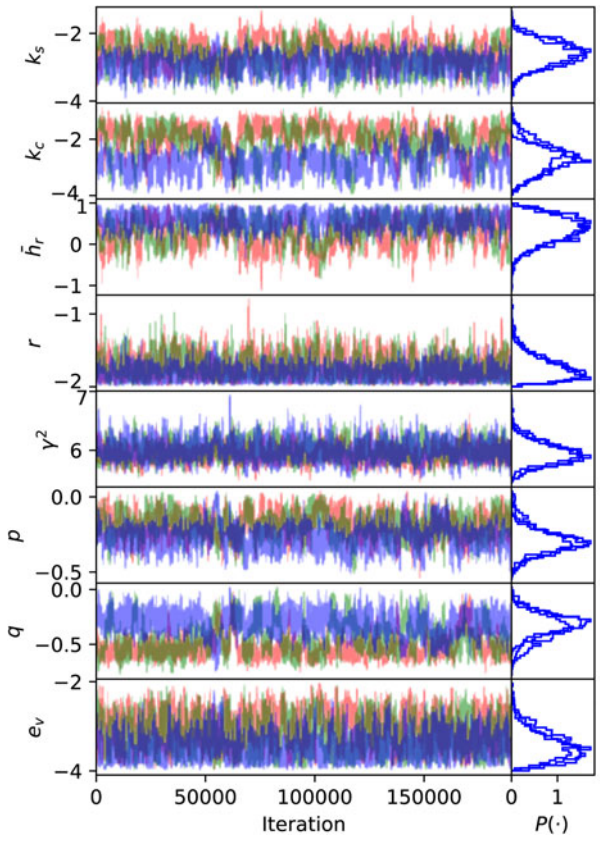


Fig. 10. Three Markov chains over the base-10 logarithm of parameter values (left, RGB), each for a different random value of ω_e. The 'fuzzy caterpillar' pattern indicates good mixing. The right plot shows histogram of the blue sample, after being divided into three disjoint sub-chains. The very similar histograms indicate a converged chain.

$$\hat{\mathbf{m}}_{t+1} \sim Q(\hat{\mathbf{m}}_{t+1} | \mathbf{m}_t). \tag{C1}$$

A given candidate parameter vector is accepted or rejected according to its posterior probability relative to the current position in parameter space:

$$a = \min\left(1,\ \frac{P(\hat{\mathbf{m}}_{t+1}|\mathbf{d}) Q(\mathbf{m}_t | \hat{\mathbf{m}}_{t+1})}{P(\mathbf{m}_t|\mathbf{d}) Q(\hat{\mathbf{m}}_{t+1}|\mathbf{m}_t)}\right), \tag{C2}$$

where a is the probability of acceptance. If a proposal is accepted, then $\mathbf{m}_{t+1} := \hat{\mathbf{m}}_{t+1}$; otherwise, $\mathbf{m}_{t+1} := \mathbf{m}_t$. In the limit as $t \to \infty$ (and under some restrictions on the proposal distribution), the set of samples produced by this procedure converges to the true posterior distribution $P(\mathbf{m}|\mathbf{d})$.

Because of the potential for highly correlated parameters, a simple application of (e.g.) the Metropolis–Hastings algorithm (which utilizes an isotropic Gaussian distribution centered around the current position as a proposal distribution) is unlikely to efficiently explore the space. However, because of the availability of automatic differentiation for the surrogate model we have easy access to the gradient of the log-posterior. This allows for a sampler that can efficiently steer itself toward probable regions of parameter space. Furthermore, because this inference problem is low dimensional, it is straightforward to compute the gradient of the gradient (i.e. the Hessian matrix), which allows for an efficient scaling of the proposal distribution.

One method which allows us to capitalize on this availability of derivatives is the manifold-Metropolis adjusted Langevin algorithm (mMALA, Girolami and Calderhead, 2011). mMALA operates as described above, but with proposal distribution given by

$$Q(\hat{\mathbf{m}}_{t+1}|\mathbf{m}_t) = \mathcal{N}(\mathbf{m}_t - \Delta \hat{H}^{-1} \nabla \log P(\mathbf{d}|\mathbf{m}_t,\, \omega_e),\, 2\Delta \hat{H}^{-1}), \tag{C3}$$

where $\hat{H}$ is a an approximation to the Hessian that is regularized to be positive definite. This method is very similar to the stochastic Newton MCMC

https://doi.org/10.1017/jog.2020.112 Published online by Cambridge University Press

method proposed by Petra and others (2014), but with the use of an analytical (rather than numerically approximated) Hessian and a generalization to step size $\Delta \neq 1$, which we have found to be critical for numerical stability. For each summand in Eqn (53), we initialize the sampler from the maximum a posteriori point, which is computed via Newton's method (again, trivial to implement due to the availability of the Hessian), initialized from a random draw from the prior distribution. We run the sampler for 2×10^5 iterations, with a step size selected by a simple moving average scheme that aims to keep the sampler's acceptance rate at ∼0.56, the theoretically optimal value for mMALA (Roberts and others, 2001). Performing this process for each summand leads to $N = 100$ randomly initialized chains, which helps to minimize the likelihood that any individual chain is stuck in a local minimum. We discard the first 10^4 samples as burn-in. The resulting chains are shown parameter-wise in Figure 10. From a qualitative perspective, the chains exhibit good mixing, as indicated by the 'fuzzy caterpillar' pattern. We ensure that the distributions are approximately stationary by dividing each chain into thirds, and overlaying the resulting histograms; we find that the histograms are very similar, indicating approximate MCMC convergence. Remaining MCMC error is further ameliorated by taking the expectation over independent chains.

https://doi.org/10.1017/jog.2020.112 Published online by Cambridge University Press