Exploring the use of transformation group priors and the method of maximum relative entropy for Bayesian glaciological inversions

Robert J. ARTHERN

Natural Environment Research Council, British Antarctic Survey, Cambridge, UK
Correspondence: Robert J. Arthern <rart@bas.ac.uk>

ABSTRACT. Ice-sheet models can be used to forecast ice losses from Antarctica and Greenland, but to fully quantify the risks associated with sea-level rise, probabilistic forecasts are needed. These require estimates of the probability density function (PDF) for various model parameters (e.g. the basal drag coefficient and ice viscosity). To infer such parameters from satellite observations it is common to use inverse methods. Two related approaches are in use: (1) minimization of a cost function that describes the misfit to the observations, often accompanied by explicit or implicit regularization, or (2) use of Bayes' theorem to update prior assumptions about the probability of parameters. Both approaches have much in common and questions of regularization often map onto implicit choices of prior probabilities that are made explicit in the Bayesian framework. In both approaches questions can arise that seem to demand subjective input. One way to specify prior PDFs more objectively is by deriving transformation group priors that are invariant to symmetries of the problem, and then maximizing relative entropy, subject to any additional constraints. Here we investigate the application of these methods to the derivation of priors for a Bayesian approach to an idealized glaciological inverse problem.

KEYWORDS: ice-sheet modelling

1. INTRODUCTION

One of the tasks presently facing glaciologists is to advise how the Greenland and Antarctic ice sheets might contribute to sea-level rise under the range of different climatic conditions that could occur in the future. To be genuinely useful to policymakers, planners of coastal infrastructure and other investors that are sensitive to future sea level, these glaciological forecasts will need to deliver information about the probability of the various possible outcomes that could be realized by the ice sheets. This makes it important to characterize the probability density function (PDF) of the sea-level contributions from Greenland and Antarctica.

Some estimates of how the climate of the atmosphere and oceans might evolve are available from general circulation models. These climate projections can be used to force dynamical models of the flow of an ice sheet, giving a forecast of the future contribution to sea level (e.g. Joughin and others, 2014; Cornford and others, 2015). At present these glaciological simulations are well adapted to investigating the sensitivity of the forecast to various perturbations in forcing, model parameters or initial conditions (e.g. Bindschadler and others, 2013). However, unless we can obtain reliable estimates of the probability of any particular perturbation actually occurring, the models cannot be used to evaluate the PDF of the contributions that ice from Greenland and Antarctica will make to sea level.

Glaciological forecasts over the next century or two will only be accurate if the models used to simulate the future begin in a state for which the geometry and flow speed are closely representative of the present-day ice sheets in Greenland and Antarctica. As in weather forecasting, the selection of the initial conditions for the simulation is an important component of the forecast. The procedure for setting up a model in a realistic starting state is known as initialization. One of the best ways to initialize large ice-sheet models is to use inverse methods (e.g. MacAyeal, 1992). These optimize the basal drag coefficient, viscosity or similar model parameters, to ensure that the model state from which the forecast proceeds agrees closely with a wide variety of measurements from satellites, aircraft and field campaigns. In this way the model starts from a state where the shape and flow speed of the ice accurately reflect what is happening now.

Ultimately, we are seeking to determine the complete PDF for the sea-level rise contribution from Greenland and Antarctica at times in the future that are relevant to planning decisions. Any errors in the initial conditions will propagate to give uncertainty in the simulation of future behavior. To quantify this uncertainty it is important to first characterize the uncertainty in the initial conditions in probabilistic terms. This makes a Bayesian approach to model initialization attractive, since it offers a probabilistic interpretation, while allowing information from satellites and other observational data to influence the joint PDF for viscosity, drag coefficient and other parameter values, as well as the initial values of state variables. Once this joint PDF has been obtained it can either be used to design ensembles for Monte Carlo experiments that evolve multiple simulations with a variety of initial conditions and parameter values, or as input to more formal probabilistic calculations that evaluate how uncertainty in the present state will propagate into the simulation used to forecast the future of the ice sheet. In this study we adopt a Bayesian approach to the problem of model initialization, and to the inversion of model parameters, such as the basal drag coefficient and viscosity.

A number of Bayesian inversions have been described previously by glaciologists (e.g. Berliner and others, 2008; Gudmundsson and Raymond, 2008; Raymond and Gudmundsson, 2009; Tarasov and others, 2012; Petra and others, 2014; Zammit-Mangion and others, 2014). A key requirement in applying Bayesian methods is the definition

of a prior probability distribution for the parameters that we wish to identify. In this paper, our particular focus will be on how we can specify prior information for model parameters that we have very little useful information about. A good example is the basal drag coefficient. This can vary enormously, depending on details of the subglacial environment that are completely unknown to us. The drag coefficient can be effectively zero for ice floating on water, but effectively infinite for ice frozen motionless to the bed. In many places in Greenland and Antarctica we do not know which condition applies. Furthermore, we do not know whether there are narrow water-filled channels, large water-filled cavities or broad sheets of water under the ice, so we cannot specify the length scale on which the basal drag coefficient might vary with any certainty. This makes it difficult to specify the prior PDF that is needed for any Bayesian inversion of this parameter.

One of our goals in this study is to reduce the subjectivity attached to glaciological forecasts. The general approach of defining the initial state of an ice-sheet model using inverse methods and then running the model forward in time to produce a forecast of the future might seem to provide a strategy for prediction that is physically based, mechanistic and largely free of subjectivity. By free of subjectivity we mean that different scientists should provide the same forecasts of the future behaviour of the ice sheet, assuming: (1) they are given the same set of observations; (2) they make the same rheological assumptions about the deformation of ice or sediment; and (3) they use the same conservation equations in the physical model that represents forces and mass fluxes within the ice sheet. However, even with observations, rheological assumptions and conservation equations in common, there is scope for making subjective decisions in the application of inverse methods that are used to identify parameters in the model, or the initial values for state variables. This applies particularly to the specification of the prior PDF for those parameters.

Subjective decisions made in defining the prior PDF will influence the initial state, and this, in turn, will affect the forecast of the ice sheet. The rate of ice flow into the ocean is sensitive to the basal drag coefficient and the ice viscosity (Schoof, 2007). Furthermore, the forecast of the ice sheet is typically obtained by solving a nonlinear system of equations, and it may be quite sensitive to small changes in initial conditions or parameter values. Models that specify different prior PDFs for the spatial variations in viscosity and basal drag could potentially produce quite different projections of sea level.

The subjectivity attached to glaciological inverse methods is not usually emphasized, and not much consideration has been given to whether it is important or not, so we consider it in some detail here. We do not claim to have a recipe to eliminate all subjective decisions from glaciological forecasts, nor is it our intention to criticize glaciological inversions that have relied upon them. There will always be decisions about which model to use, which datasets to include, which parameters to invert for and which methods to use to regularize the inversion. In common with many previous studies, our work has involved a variety of such decisions in mapping spatial patterns of basal drag and estimating the flow speeds within the Antarctic ice sheet (Arthern and others, 2015). The motivation for the present study is to explore whether this approach can be improved upon by working within a probabilistic framework. Tasked with providing probabilistic estimates of the contribution of the ice sheets to sea level, our goal is that those forecasts should be made as objectively as currently available techniques allow.

2. BAYESIAN INFERENCE OF MODEL PARAMETERS USING OBSERVATIONS

Suppose we are trying to estimate a vector, $\boldsymbol{\theta}$, comprised of N parameters, $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_N]^{\mathrm{T}}$, which may include the basal drag coefficient, β, at many different locations and viscosity, η, at many different points within the ice sheet. Bayes' theorem provides a recipe for modifying a prior PDF for these parameters, $p(\boldsymbol{\theta})$, to include the information provided by new data, $\boldsymbol{x} = [x_1, x_2, \ldots, x_M]^{\mathrm{T}}$, which may include observations from satellites, aircraft, field parties or laboratory experiments. The result is the posterior PDF,

$$p_{\mathrm{p}}(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p_{\mathrm{l}}(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p_{\mathrm{n}}(\boldsymbol{x})}. \tag{1}$$

The prior $p(\boldsymbol{\theta})$ is a PDF, defined such that $p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$ is the probability that the parameters lie within a vanishingly small 'volume' element $\mathrm{d}\boldsymbol{\theta} = \mathrm{d}\theta_1 \mathrm{d}\theta_2 \ldots \mathrm{d}\theta_{\mathrm{N}}$, located at $\boldsymbol{\theta}$, within an N-dimensional parameter space, Θ, that includes all possible values of the parameters. The term prior reflects that this is the PDF before we have taken account of the information provided by the data. The information provided by the data, $\boldsymbol{x}$, is encoded in the likelihood function, $p_{\mathrm{l}}(\boldsymbol{x}|\boldsymbol{\theta})$. The likelihood function can be assumed known, provided two conditions are met. First, our physical model must be capable of estimating the measured quantities, $\boldsymbol{x}$, if supplied with parameter values, $\boldsymbol{\theta}$. Second, we must be able to estimate the PDF of residuals between these model-based estimates and the data, $\boldsymbol{x}$ (e.g. if model deficiencies can be neglected, this amounts to knowing the distribution of the observational errors). The likelihood function, $p_{\mathrm{l}}(\boldsymbol{x}|\boldsymbol{\theta})$, is then proportional to the PDF for observing the data, $\boldsymbol{x}$, given that the parameters take particular values, $\boldsymbol{\theta}$. The denominator, $p_{\mathrm{n}}(\boldsymbol{x})$, is defined as $p_{\mathrm{n}}(\boldsymbol{x}) = \int_{\Theta} p_{\mathrm{l}}(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$, and can simply be viewed as a normalizing constant, defined so the posterior PDF gives a total probability of $\int_{\Theta} p_{\mathrm{p}}(\boldsymbol{\theta}|\boldsymbol{x})\,\mathrm{d}\boldsymbol{\theta} = 1$, when integrated over all possible values within the parameter space, Θ. To avoid ambiguity we will use subscripts to identify various different posterior PDFs (p_{p1}, p_{p2}, etc.), likelihoods (p_{l1}, p_{l2}, etc.), priors (p_1, p_2, etc.) and normalizing constants (p_{n1}, p_{n2}, etc.).

The notation for conditional probabilities, $P(A|B)$, denotes probability of event A given that B is true. The posterior, $p_{\mathrm{p}}(\boldsymbol{\theta}|\boldsymbol{x})$, is the PDF for the parameters, $\boldsymbol{\theta}$, given that the data, $\boldsymbol{x}$, take the particular values observed. This means that, after we have taken account of all the information provided by the data, $\boldsymbol{x}$, the posterior PDF, $p_{\mathrm{p}}(\boldsymbol{\theta}|x)\,\mathrm{d}\boldsymbol{\theta}$, gives the updated probability that the parameters lie within a small volume, $\mathrm{d}\boldsymbol{\theta}$, of parameter space located at $\boldsymbol{\theta}$. Selecting the values of $\boldsymbol{\theta}$ that maximize $p_{\mathrm{p}}(\boldsymbol{\theta}|x)$ provides a Bayesian estimate for the most likely value of the parameters.

The likelihood, function, $p_{\mathrm{l}}(\boldsymbol{x}|\boldsymbol{\theta})$, sometimes written $L(\boldsymbol{\theta};\boldsymbol{x})$ or $L(\boldsymbol{\theta})$, can be considered as a function of $\boldsymbol{\theta}$ for the observed values of the data, $\boldsymbol{x}$. It is sometimes the case when applying Bayes' rule that the likelihood, $L(\boldsymbol{\theta})$, is negligible except within a narrowly confined region of parameter space, while the prior, $p(\boldsymbol{\theta})$, describes a much broader distribution. This situation would indicate great

prior uncertainty in parameter values, $\boldsymbol{\theta}$, but much less uncertainty once the information from the data is incorporated using Bayes' rule. In such cases, the information provided by the likelihood function, $L(\boldsymbol{\theta})$, overwhelms the information provided by the prior, $p(\boldsymbol{\theta})$. Specifying the prior accurately in such circumstances is perhaps not so important, since any sufficiently smooth function much broader than the likelihood function would produce a similar posterior PDF. However, we should not be complacent just because there are some circumstances in which it is not very important to specify the prior PDF accurately. There is no guarantee that this situation will correspond to glaciological inversions of the type that we are considering. Many aspects of the subglacial environment are barely constrained by data, so it is in our interests to specify the prior PDF carefully.

In this paper we will apply two principles advocated by Jaynes (2003) to constrain the choice of prior PDF: (1) we will exploit symmetries of the ice-sheet model, by requiring that the prior PDF is invariant to a group of transformations that do not alter the mathematical specification of the inverse problem, and (2) using this invariant prior as a reference function, we will include additional constraints by seeking the PDF that maximizes the relative entropy subject to those constraints. Both approaches are described in detail by Jaynes (2003), and we will only make brief introductory remarks about them (Sections 5 and 6). Our intention is to guide, and as far as possible eliminate, the subjective decisions made during the inverse problem that defines the initial state of the model from the observations, particularly with respect to the choice of a prior PDF for the parameters. Although we concentrate here on methods advocated by Jaynes (2003), reviews by Kass and Wasserman (1996) and Berger (2006) provide a broader perspective and include additional background on the use of formal rules for the selection of prior PDFs.

3. THE CLOSE RELATIONSHIP BETWEEN THE PRIOR PDF AND THE REGULARIZATION OF INVERSE METHODS

Although we will use Bayesian methods, our investigation is relevant to a wide variety of glaciological inverse methods, many of which have been described in the literature without mention of Bayes' theorem.

In particular, our approach is related to a broad class of inverse methods that minimize a cost function, J_{misfit}, that quantifies the misfit between the model and data. A common example is choosing model parameters that minimize the mismatch between model velocities, $\boldsymbol{u}$, and observations of the ice velocity, $\boldsymbol{u}^*$, from satellites, so that the cost function is the unweighted sum of the squares of the misfit, e.g. $J_{\text{misfit}} = \frac{1}{2}\sum_i \left(\boldsymbol{u}_i - \boldsymbol{u}_i^*\right)^2$, or some similar function weighted by estimates of the observational error covariance, $\boldsymbol{R}$, e.g. $J_{\text{misfit}} = \frac{1}{2}\sum_{ij}\left(\boldsymbol{u}_i - \boldsymbol{u}_i^*\right)R_{ij}^{-1}\left(\boldsymbol{u}_j - \boldsymbol{u}_j^*\right)$. Other cost functions to characterize the misfit between the model and the data have been proposed (Arthern and Gudmundsson, 2010; Morlighem and others, 2010) and the choice of cost function is therefore one way that subjective decisions can influence the inversion.

There are other aspects of the inversion that require subjective decisions. Generally speaking, simply minimizing the mismatch with observations does not uniquely define the spatially varying fields of basal drag and viscosity. This is because many different combinations of parameters allow the model to agree equally well with the available observational data. This is often dealt with by constraining parameters, using some kind of explicit or implicit regularization.

Regularization introduces additional information about parameters (e.g. requiring that they be close to some estimated value, or that they are small in magnitude, or that they vary smoothly in space). Before proceeding we will describe how regularization of inverse methods as commonly applied in glaciology relates to our Bayesian inversion.

The purpose of regularization is either to turn an ill-posed problem into a well-posed problem, or an ill-conditioned problem into a well-conditioned problem. As defined by Hadamard (1902), an ill-posed problem either has no solution, more than one solution, or a solution that varies discontinuously when small perturbations are made to any quantitative information provided (i.e. data). In any of these three cases it becomes impossible to precisely define a solution to the problem. On a practical level, especially when performing calculations numerically, we may come across problems that are not exactly ill-posed in the above sense, but are ill-conditioned. This means that a unique solution exists, but small changes to the data from measurement errors or numerical roundoff can result in large changes to that solution. If the resulting loss of precision is too great, we may be willing to constrain the solution in some other way, by regularizing the problem.

To be more concrete, we will give some simple examples of how a glaciological inversion can be ill-posed or ill-conditioned, beginning with an example of a problem that does not have a unique solution. Suppose we would like to find the initial state for a model of ice of constant thickness flowing down a featureless inclined plane. Furthermore, suppose we know the ice thickness, the surface elevation and the flow speed at the surface (i.e. the data). Suppose now that we have no information about the drag coefficient at the base of the ice, or the ice viscosity, but wish to determine these using inverse methods.

For a slab of the given thickness, the ice speed at the surface could be matched either by a rigid slab that is sliding at its base, or by a slab that is fixed at the base, but deforming through internal shearing (Fig. 1). None of the data provide information about the subsurface flow speed so we cannot distinguish between these two possibilities, or between these and some intermediate solution that is any combination of sliding and internal shearing that matches the specified surface velocity.

In practical applications, to avoid such non-uniqueness, it is rare that viscosity and basal drag are solved for simultaneously. Rather, it is commonly assumed that one or other of these quantities is known perfectly. On floating ice shelves the viscosity is usually solved for on the assumption that basal drag is zero. By contrast, on grounded ice, the basal drag is usually solved for on the assumption that the viscosity perfectly obeys some rheological flow law with known parameters, so that the viscosity can be considered known.

The assumption that either the basal drag or the viscosity is perfectly known regularizes what would otherwise be an ill-posed problem, by avoiding a multiplicity of non-unique solutions. However, there are difficulties with this approach. First, for ice shelves where the bathymetry is poorly mapped, it may be difficult to be certain there is no basal

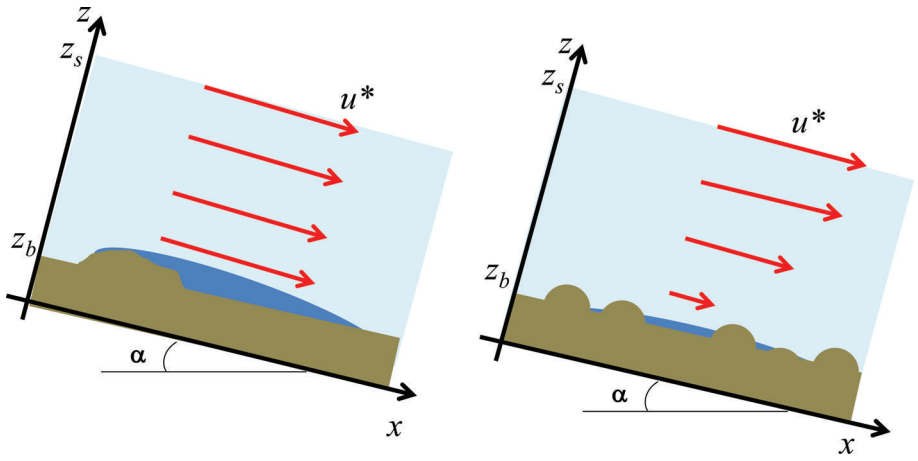

Fig. 1. Simultaneous inversion for basal drag coefficient, β, and viscosity, η, is not well posed. Any observed surface velocity could be produced either by a well-lubricated base with high viscosity (left), or by a slab with high basal drag and low viscosity (right). Prior information about basal drag and/or viscosity is needed to determine which situation is more likely. The inversion may also be ill-conditioned if features at the bed are too small to affect the shape or flow speed of the upper surface. The usual remedy for non-uniqueness or ill-conditioning is to regularize the problem, and this can be interpreted in Bayesian terms as specifying prior probabilities for basal drag and viscosity. The coordinate axes used for the simple slab model are shown.

drag from some unidentified contact with the sea floor (Fürst and others, 2015). Second, on grounded ice, many factors that are not included in commonly used flow laws can affect the viscosity. These include such poorly known factors as impurity content, anisotropy, grain size, geothermal heating and damage from crevassing. This makes it problematic to assume we have perfect knowledge of the viscosity. Another consideration is that we have assumed in the problem specified above that the ice thickness is known perfectly. However the thickness is poorly known in many places, and it might be useful to invert for this field, rather than assume it is known perfectly (Raymond and Gudmundsson, 2009). Clearly, the problem of non-uniqueness would become even more acute if the observations of ice thickness were unavailable and we tried to solve for ice thickness as well as the basal drag and viscosity.

In the above example there are three parameters that we would like to invert for: ice thickness, ice viscosity and basal drag coefficient. Rather than assume we have perfect knowledge of two of these, it would be more realistic to acknowledge that there is considerable uncertainty in each, and to seek a compromise solution that jointly reflects these uncertainties.

For problems with many uncertain parameters a Bayesian approach is attractive. Rather than one set of parameters that minimize the cost function, J_{misfit}, Bayesian inversion seeks the posterior joint PDF, $p_{\text{p}}(\boldsymbol{\theta}|\boldsymbol{x})$, for the parameters. This means that the combined uncertainty in basal drag coefficient, viscosity and thickness can be evaluated. If we wish, we can later seek the values of the parameters that maximize the joint PDF, allowing us to solve simultaneously for the most likely values of all three quantities. To perform such Bayesian inversion we will need to define a prior PDF, $p(\boldsymbol{\theta})$, for the parameters. It is this aspect that we concentrate on in this paper.

As an example of ill-conditioning, suppose the slipperiness at the base of the slab is not uniform as assumed above, but has fluctuations on some scale. As the characteristic size of these fluctuations decreases, their effect on the flow at the surface will diminish, until they become too small to have any significant effect (Bahr and others, 1994; Gudmundsson, 2003). At the smallest scales their effect on the surface elevation and flow speed will be smaller than the accuracy with which these data are measured. Any inverse method that seeks to recover the fluctuations in basal drag on such a fine scale will be corrupted by the errors in surface elevation and surface velocity. In extreme cases, wild and unrealistic variations in basal drag might be introduced in an attempt to match the flow speed in the model to noise in the observations. This is known as overfitting. The usual remedy is to apply some form of regularization.

There are various different ways of regularizing inversions of basal drag to avoid overfitting, but a common approach is to enforce smoothness of the recovered pattern of basal drag. Many of the iterative algorithms that are used to minimize the cost function have a convenient property: they introduce features in basal drag on coarse scales in the first few iterations, then add progressively finer scales in later iterations (Maxwell and others, 2008; Habermann and others, 2012). Simply stopping after some number of iterations can prevent unrealistic fine-scale features being added. Deciding when to stop is a more vexing question, but there are criteria that can serve as a guide (Maxwell and others, 2008; Habermann and others, 2012). One remaining issue is that the regularized solution depends upon the initial guess for parameters used to start the very first iteration. Again, this is an opportunity for different people to make different choices.

A different form of regularization that is often used is Tikhonov regularization (e.g. Jay-Allemand and others, 2011; Gillet-Chaulet and others, 2012; Sergienko and Hindmarsh, 2013). Here the data–model misfit cost function, J_{misfit}, is replaced by $J_{\text{total}} = J_{\text{misfit}} + J_{\text{reg}}$, where J_{reg} is a term that penalizes solutions for the basal drag coefficient, β, that are not smooth and promotes those that are. A common choice is $J_{\text{reg}} = \lambda_{\text{reg}} \int |\nabla\beta|^2 \, \text{dS}$, for some constant λ_{reg}, which adds a term proportional to the area integral of the square of the magnitude of the horizontal gradient in basal drag coefficient (e.g. Sergienko and Hindmarsh, 2013). Adding this term to the data–model misfit cost function

before minimization favors solutions for basal drag that have small gradients, hence the wildly fluctuating high-frequency oscillations that might otherwise be introduced by overfitting are reduced.

When Tikhonov regularization is used, the value of λ_{reg} can be varied to increase or decrease the strength of the regularization. It can be difficult to know what value to use for this parameter. Some heuristic conventions exist for selecting λ_{reg}, among them plotting the L-curve (e.g. Jay-Allemand and others, 2011), or making use of a discrepancy principle (Maxwell and others, 2008), but in real applications these do not always provide an obvious choice (Vogel, 1996).

It can also be difficult to know whether to penalize gradients in the drag parameter or its logarithm, i.e. $J_{\mathrm{reg}} = \lambda_{\mathrm{reg}} \int |\nabla \ln \beta|^2 \, \mathrm{dS}$. Other options include the square of basal drag, $J_{\mathrm{reg}} = \lambda_{\mathrm{reg}} \int |\beta|^2 \, \mathrm{dS}$, or its logarithm, $J_{\mathrm{reg}} = \lambda_{\mathrm{reg}} \int |\ln \beta|^2 \, \mathrm{dS}$, but it is not always obvious why one should use one form rather than another, or even some combination. It is clear there is scope for many different choices in applying Tikhonov regularization, and we have not even mentioned all of them.

Regularization requires the introduction of information that does not come from the observational data, $\boldsymbol{x}$, that we have available from satellites, aircraft, field observations or laboratory experiments. This extra information must come from somewhere. The source of much of the subjectivity that we refer to in this paper is that the practitioners of the inverse methods often simply decide what seems reasonable. It is here that many of the subjective decisions that we would prefer to avoid can arise.

How smooth should the field of basal drag be? What should be the starting guess for iterative minimization of the cost function? What form of Tikhonov regularization should be used? How much can the viscosity vary from some prescribed approximation of the ice rheology, such as Glen's flow law? Viewed from the Bayesian perspective, all of these decisions amount to the selection of priors for basal drag and viscosity.

One of the attractions of Bayes' theorem is that it can provide the joint PDF for the parameters, given some observations with known error distribution. Crucially, the theorem cannot be applied without a prior for the parameters. This requirement to define a prior PDF for the parameters brings into the open many of the subjective decisions that are often made in an ad hoc fashion in the process of regularizing inversions.

As noted in many studies using Bayesian methods (e.g. Gudmundsson and Raymond, 2008; Petra and others, 2014), the link between regularization and specification of the prior can often be made explicit by taking the negative of the logarithm of Eqn (1),

$$-\ln p_{\mathrm{p}}(\boldsymbol{\theta}|\boldsymbol{x}) = -\ln p_{\mathrm{l}}(\boldsymbol{x}|\boldsymbol{\theta}) - \ln p(\boldsymbol{\theta}) + \ln p_{\mathrm{n}}(\boldsymbol{x}). \qquad (2)$$

Now, we identify a misfit function, $J_{\mathrm{misfit}} = -\ln p_{\mathrm{l}}(\boldsymbol{x}|\boldsymbol{\theta})$, defined as the negative of the log-likelihood, and a regularization term, $J_{\mathrm{reg}} = -\ln p(\boldsymbol{\theta})$, that is the negative of the logarithm of the prior, and $J_0 = \ln p_{\mathrm{n}}(\boldsymbol{x})$, which is just a constant offset for any given set of observations. Then it is clear that choosing parameters, $\boldsymbol{\theta}$, that maximize the posterior PDF is the same as choosing them to minimize a misfit function, $J_{\mathrm{total}} = J_{\mathrm{misfit}} + J_{\mathrm{reg}} + J_0 = -\ln p_{\mathrm{p}}(\boldsymbol{\theta}|\boldsymbol{x})$. The relationship, $J_{\mathrm{reg}} = -\ln p(\boldsymbol{\theta})$, means for instance that quadratic regularization terms, such as $J_{\mathrm{reg}} = \lambda_{\mathrm{reg}} \int |\nabla \beta|^2 \, \mathrm{dS}$, correspond to specifying a Gaussian density function for $p(\boldsymbol{\theta})$, such as $\exp(-\lambda_{\mathrm{reg}} \int |\nabla \beta|^2 \, \mathrm{dS})$, and vice versa. From a Bayesian perspective the various options for Tikhonov regularization described above are just different ways of specifying a prior PDF for the parameters.

Working in the Bayesian framework provides some clarity to the definition of the cost function, J_{misfit}, since it suggests that if we want the most likely parameters we should use the negative log-likelihood function, $-\ln p_{\mathrm{l}}(\boldsymbol{x}|\boldsymbol{\theta})$, to characterize the misfit with data, rather than unweighted least-squares or some other choice. It also clarifies the process of regularization, since it requires that the information to be added is explicitly formulated in terms of a prior PDF for the parameters. However, simply adopting the Bayesian approach does not tell us what the prior, $p(\boldsymbol{\theta})$, should be. So how should priors be defined for parameters that we have so little information about?

4. SUBJECTIVE PRIORS

One possible way to define a prior, $p(\boldsymbol{\theta})$, is to leave this up to the individual scientist performing the inversion. In the case of inverting for the basal drag under an ice sheet this seems a questionable choice. The posterior PDF, $p_{\mathrm{p}}(\boldsymbol{\theta}|\boldsymbol{x})$, will be used to define the parameters for the model, and these parameters are an important component of the sea-level forecast. The glaciological forecast usually requires us to solve a nonlinear system of equations, which may be sensitive to small changes in parameter values or initial conditions, and we know that flow of ice into the ocean is sensitive to the basal drag coefficient and the ice viscosity (Schoof, 2007). This suggests that models that specify different prior PDFs for the spatial variations in basal drag could produce quite different projections of sea level. At present it is difficult to know how important this effect could be, but as more forecasts are produced, each with different models and different inversion methods, it will become easier to evaluate the degree of spread among projections.

Often a great deal of effort and cost is expended in developing the physical model, collecting the observations, $\boldsymbol{x}$, and characterizing the error covariance of those observations. It seems questionable to apply such dedication to deriving the likelihood, $p_{\mathrm{l}}(\boldsymbol{x}|\boldsymbol{\theta})$, but then multiply this function by a prior that is left up to individual choice, either through explicit definition of a subjective prior, or implicitly through choices of regularization strategy. This could be justified if the scientist performing the inversion has some real insight into the range of variation and the length scale on which basal drag varies. As mentioned above, there is great uncertainty regarding the subglacial environment and it is difficult to know how the insight needed to define the prior would be obtained. We emphasize that the prior, $p(\boldsymbol{\theta})$, is logically independent of the observations, $\boldsymbol{x}$, that will be used in the inversion, so these observations cannot be used to provide insight into what the prior should be.

Another way of defining the prior, $p(\boldsymbol{\theta})$, would be to delegate the task to experts on the subglacial environment, asking them to define the prior for the basal drag, viscosity, etc. The justification for this would be that there are people who (without regard of the observations that we will use in the inversion) can provide us with useful information on the range of values that the viscosity and basal drag coefficient

can take, and the length scales they will vary over. If such experts exist then their views should be taken into account in definition of the prior, $p(\boldsymbol{\theta})$. However, it may be difficult to find anyone with such comprehensive information about the details of the subglacial environments of Greenland and Antarctica.

In the end, the main justifications for using subjective priors, or indeed heuristic approaches to regularization, may be (1) that they are easy to implement, (2) that it can plausibly be assumed, or checked after the fact, that the main results of the forecast are not too sensitive to the details of how this regularization is performed and (3) that it can be difficult to imagine what else could be done. The first point is certainly true, and should not be downplayed, since it allows large-scale calculations to be performed that could not be otherwise (e.g. Gillet-Chaulet and others, 2012; Morlighem and others, 2013; Joughin and others, 2014; Petra and others, 2014; Arthern and others, 2015; Cornford and others, 2015). The second point may well be true also, but seems to require that we address the third. After all, without first considering what else we might do to regularize the problem it is hard to argue that it won't make much difference. In the following sections we outline two principles that have been advanced by Jaynes (2003) as a way of defining prior PDFs for parameters when minimal information about them is available.

5. TRANSFORMATION GROUP PRIORS

Transformation group priors use symmetries of the problem to constrain the function, $p(\boldsymbol{\theta})$, that is used as a prior PDF. In many mathematical problems knowledge of some particular symmetry can be extremely valuable, because it allows us to rule out a wide range of possible solutions that do not exhibit that symmetry. For instance, if there is some prior information available to us that can be written as mathematical expressions involving $\boldsymbol{\theta}$ and if there are transformations that can be applied to these expressions that do not alter them in any way, then Jaynes (2003) argues that those transformations should also leave the prior, $p(\boldsymbol{\theta})$, unchanged. The motivation is to ensure consistency, so that for two problems where we have the same prior information we assign the same prior probabilities (Jaynes, 2003). Based on this, Jaynes (2003) argues that we should select priors that are invariant to a group of transformations that do not alter the specified problem. Surprisingly, in some cases, identifying the symmetries in the form of a group of transformations that leave the problem unchanged and then requiring that the function $p(\boldsymbol{\theta})$ is invariant to those transformations can completely determine which function to use as a prior. The value of using transformation group priors is perhaps best appreciated by imagining that we use a prior that does not respect the symmetries of the specified problem. Then we would, in effect, be claiming access to additional information that is not inherent in the problem specification, and, if called upon, we should be able to provide a reasoned explanation of where that information has come from.

6. MAXIMIZING RELATIVE ENTROPY

In addition to the symmetries of the problem, we may have other information that is relevant to specification of the prior. Sometimes this information can be expressed in the form of constraints that the PDF must satisfy. One common class of constraints are expectations of the form,

$$\int_\Theta p(\boldsymbol{\theta}) f_i(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} = F_i. \tag{3}$$

For instance, if we have reason to believe that the expected value for the vector of parameters $\boldsymbol{\theta}$ is $\overline{\boldsymbol{\theta}}$, we would apply a constraint with $f_i = \boldsymbol{\theta}$, $F_i = \overline{\boldsymbol{\theta}}$. A similar constraint with $f_i = F_i = 1$ requires that the PDF, $p(\boldsymbol{\theta})$, is normalized such that it integrates to one. Jaynes (2003) provides a recipe for incorporating such constraints, arguing that we should favor the PDF that maximizes the relative entropy subject to whatever constraints are imposed. The relative entropy of a PDF, $p(\boldsymbol{\theta})$, is a functional, $H(p)$, defined with respect to a reference PDF, $\pi(\boldsymbol{\theta})$, as

$$H = -\int_\Theta p(\boldsymbol{\theta}) \ln\left[\frac{p(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right] \mathrm{d}\boldsymbol{\theta}. \tag{4}$$

Multiple constraints of the form given by Eqn (3) can be imposed using Lagrange multipliers $\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_Q\}$, by seeking stationary points of the functional,

$$H_1(p, \boldsymbol{\lambda}) = -\int_\Theta p(\boldsymbol{\theta}) \ln\left[\frac{p(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right] \mathrm{d}\boldsymbol{\theta} + \sum_{i=1}^{Q} \lambda_i \left[\int_\Theta p(\boldsymbol{\theta}) f_i(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} - F_i\right]. \tag{5}$$

As described by Jaynes (2003), when the normalization constraint is enforced and other constraints, $i = 1, 2, \ldots, Q$, are also imposed, stationary points of $H_1(p, \boldsymbol{\lambda})$ are provided by PDFs of the form

$$p(\boldsymbol{\theta}) = \frac{\pi(\boldsymbol{\theta}) \exp\left[\sum_{i=1}^{Q} \lambda_i f_i(\boldsymbol{\theta})\right]}{Z}, \tag{6}$$

$$Z(\boldsymbol{\lambda}) = \int_\Theta \pi(\boldsymbol{\theta}) \exp\left[\sum_{i=1}^{Q} \lambda_i f_i(\boldsymbol{\theta})\right] \mathrm{d}\boldsymbol{\theta}, \tag{7}$$

$$\frac{\partial \ln Z}{\partial \lambda_i} = \int_\Theta p(\boldsymbol{\theta}) f_i(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} = F_i. \tag{8}$$

Solving Eqn (8) often provides a convenient way of identifying values for the Lagrange multipliers, $\boldsymbol{\lambda}$, such that H_1 is stationary and the constraints are enforced. Once these values of $\boldsymbol{\lambda}$ have been obtained, Eqn (6) provides the PDF and $Z(\boldsymbol{\lambda})$ plays the role of a normalizing constant. If there are no constraints other than normalization, then finding stationary points of H_1 results in $p(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$. This means that $\pi(\boldsymbol{\theta})$ can be viewed as a preliminary prior that will be modified, such that any additional constraints on $p(\boldsymbol{\theta})$ are satisfied. Jaynes (2003) argues that the PDF, $p(\boldsymbol{\theta})$, that maximizes H, subject to whatever constraints are imposed has many attractive features. Roughly speaking, H can be viewed as quantifying the 'uninformativeness' of the PDF, $p(\boldsymbol{\theta})$. Maximizing H is therefore a way of guarding against selecting a prior that is too prescriptive about which values of $\boldsymbol{\theta}$ are likely. This provides a safeguard against ruling things out that could possibly happen. A prior probability obtained in this way is guaranteed to satisfy the constraints, but is otherwise as uninformative as possible.

We would obviously prefer to have a very informative prior, since then we would know exactly which parameter values to use in our model. It may then seem strange that we are selecting the least informative prior possible, subject to the information introduced by the reference distribution and the constraints. The point is that we should only use an informative prior if we actually have the information to back

it up. Here, once we have defined a reference distribution, the extra information is being introduced in the form of constraints, or through the data that we will introduce later via Bayes' theorem. For each constraint that we impose, the prior will become more informative, relative to the original PDF, $\pi(\boldsymbol{\theta})$. If we were to subjectively choose a prior more informative than demanded by the constraints we would be guilty of subjectively introducing information into the inversion without good reason, and this is exactly what we are hoping to avoid, as far as possible. A prior that is too prescriptive about which parameter values are possible will only lead to overconfidence in the accuracy of our forecasts and to surprises when the forecast fails to deliver such accuracy.

It may seem that the problem has now simply changed from finding $p(\boldsymbol{\theta})$ to finding the preliminary prior, $\pi(\boldsymbol{\theta})$. This is where a combination of the two approaches outlined above can be used. Jaynes (2003) suggests that invariance to a transformation group defining the symmetry of the specified problem should be used to define $\pi(\boldsymbol{\theta})$. Having obtained $\pi(\boldsymbol{\theta})$, any additional constraints can then be imposed by maximizing the relative entropy, H, subject to those constraints. This is the procedure that we will adopt in the rest of this paper.

7. APPLICATION TO A SIMPLE GLACIOLOGICAL PROBLEM

To introduce the methods outlined above, we will consider the simple problem of estimating the viscosity and basal drag coefficient for a slab of uniform thickness flowing down a plane. Although this is a highly simplified problem compared with the initialization of large-scale models of the Greenland and Antarctic ice sheets, it will turn out to contain many of the essential features of the more difficult three-dimensional problem, and therefore serves as a useful starting point to illustrate the methods.

We define a coordinate system in which the x- and y-axes are parallel to the planar bed of the ice sheet, which slopes downwards at an angle α below horizontal in the direction of increasing x, with no slope in the direction of increasing y (Fig. 1). The z-axis is taken to be normal to the bed, positive upwards, with $z = z_\mathrm{b}$ defining the bed and $z = z_\mathrm{s}$ defining the surface. The thickness, $h = z_\mathrm{s} - z_\mathrm{b}$, is assumed uniform, and velocity in the x-direction, $u(z)$, is a function of depth. Any vertical shearing within the slab leads to a shear stress σ_{xz}. The ice density, ρ, is assumed constant. The basal drag coefficient is β, and the viscosity within the slab is η, which is assumed constant with depth in this simplified problem. Within the slab, body forces from gravity are balanced by gradients in stress. At the lower boundary a linear sliding law is assumed, so the shear stress $\sigma_{xz} = \beta u$. At the upper boundary the shear stress vanishes, so $\sigma_{xz} = 0$. For now, we will assume nothing about β or η, other than that they are positive constants.

For this simple system, conservation of momentum in the x-direction gives the system of equations:

$$\begin{aligned} \eta\partial_z u &= \sigma_{xz}, && \text{in } z_\mathrm{b} < z < z_\mathrm{s},\\ \partial_z \sigma_{xz} &= -\rho g \sin\alpha, && \text{in } z_\mathrm{b} < z < z_\mathrm{s},\\ \sigma_{xz} &= 0, && \text{on } z = z_\mathrm{s},\\ -\sigma_{xz} + \beta u &= 0, && \text{on } z = z_\mathrm{b}, \end{aligned} \tag{9}$$

where g is acceleration due to gravity. Generally, before performing an inversion using the model we will already have available a discrete version of the momentum equations. For illustrative purposes we can consider a simple finite-difference discretization of the above system on a uniform grid that has n velocity levels at grid spacing $\Delta = h/(n-1)$ and velocities in the x-direction of u_1, u_2, etc. These are defined on each of the different levels, so u_1 is the sliding velocity at the base and u_n is the flow velocity at the upper surface. We define

$$\boldsymbol{A} = \boldsymbol{X}^\mathrm{T}\boldsymbol{D}\boldsymbol{X}, \tag{10}$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & & & & & & \\ -1 & 1 & & & & & \\ & \ddots & \ddots & & & & \\ & & \ddots & \ddots & & & \\ & & & -1 & 1 & & \\ & & & & -1 & 1 \end{bmatrix}, \quad \boldsymbol{f} = \begin{bmatrix} 0 \\ \rho g \sin\alpha \\ \vdots \\ \vdots \\ \rho g \sin\alpha \\ 0 \end{bmatrix},$$

$$\boldsymbol{D} = \begin{bmatrix} \frac{\beta}{\Delta} & & & & & \\ & \frac{\eta}{\Delta^2} & & & & \\ & & \ddots & & & \\ & & & \ddots & & \\ & & & & \frac{\eta}{\Delta^2} & \\ & & & & & \frac{\eta}{\Delta^2} \end{bmatrix}, \quad \boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_{n-2} \\ u_{n-1} \\ u_n \end{bmatrix}. \tag{11}$$

The discretized system corresponding to Eqn (9) is then

$$\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f}. \tag{12}$$

This is obviously a highly simplified model. We have only introduced it here so we have a very simple discrete system that we can use to illustrate how the methods can be applied in more general circumstances. More sophisticated ice-sheet models can be written in a very similar form, with $\boldsymbol{A}$ a symmetric positive-definite matrix, $\boldsymbol{u}$ a vector of velocities and $\boldsymbol{f}$ a vector comprised of body forces and forcing terms from boundary conditions.

If $\boldsymbol{A}$ and $\boldsymbol{f}$ are known, solving the system defined by Eqn (12) provides an estimate, $\boldsymbol{A}^{-1}\boldsymbol{f}$, for the velocities, $\boldsymbol{u}$. In this paper we are not particularly interested in such a straightforward application of the model. Instead we would like to consider very general inferences about the velocity field, $\boldsymbol{u}$, the forces, $\boldsymbol{f}$, and the matrix, $\boldsymbol{A}$, remembering that this matrix depends on the parameters, β and η, that we are trying to identify. We will also consider the possibility that the model is not perfect, so Eqn (12) is satisfied only approximately.

Note that in the above example, if we can estimate the matrix $\boldsymbol{A}$, we can later derive the parameters β and η by computing $\boldsymbol{D} = \boldsymbol{X}^{-\mathrm{T}}\boldsymbol{A}\boldsymbol{X}^{-1}$. To keep the following discussion as generally applicable as possible, we will not yet assume any particular form for the system matrix, $\boldsymbol{A}$, except that it is positive-definite and symmetric. Later, we will return to the problem with the particular $\boldsymbol{A}$ defined by Eqn (10).

We will include in our vector of parameters, $\boldsymbol{\theta}$, all of the quantities that we might perhaps want to estimate. These will include velocities, $\boldsymbol{u}$, forces, $\boldsymbol{f}$, the upper triangular (including leading diagonal) elements, $\boldsymbol{A}^\mathrm{u}$, of the symmetric matrix, $\boldsymbol{A}$, and the upper triangular (including leading diagonal) elements, $\boldsymbol{C}^\mathrm{u}$, of the model error covariance, $\boldsymbol{C}$. Some of these would not usually be regarded as

'parameters', but we will continue to use this terminology for the unknown quantities that we would like to be able to estimate. We have only included the upper triangular elements of symmetric matrices in our list of parameters because the complete matrix, e.g. $\boldsymbol{A}(\boldsymbol{A}^{\mathrm{u}})$, can always be recovered from these if needed.

For now, our only concern is to find a function that we can use as a prior for velocities, $\boldsymbol{u}$, forces, $\boldsymbol{f}$, and the upper triangular elements of $\boldsymbol{A}$ and $\boldsymbol{C}$, based on what we know about the relationships between them.

It is important to recognize that we will not introduce any observations of any of the quantities until we have obtained the prior and are ready to include those observations using Bayes' rule. Equally, once we have obtained a very general prior, we can later impose additional constraints on the form of the matrix, $\boldsymbol{A}$. If expert knowledge or independent estimates of parameter values from laboratory experiments are available these can also be introduced later using Bayes' theorem.

8. DERIVING A GENERAL PRIOR FOR DISCRETE SYMMETRIC POSITIVE-DEFINITE SYSTEMS

On the assumption that a symmetric positive-definite matrix, $\boldsymbol{A}$, exists that relates velocities, $\boldsymbol{u}$, and body forces, $\boldsymbol{f}$, with finite error covariance and finite bias, we have the following prior information:

$$\begin{gathered}\boldsymbol{u}-\boldsymbol{A}^{-1}\boldsymbol{f}=\boldsymbol{\epsilon},\\ \mathrm{E}_{\boldsymbol{u},\boldsymbol{f}|\boldsymbol{A},\boldsymbol{C}}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}\right]=\boldsymbol{C},\\ \boldsymbol{u},\boldsymbol{f}\in\mathbb{R}^{n}\quad \boldsymbol{A},\boldsymbol{C}\in\mathcal{P}(n).\end{gathered}\tag{13}$$

Since we are assuming that the model has already been discretized using some particular set of basis functions, the velocities, $\boldsymbol{u}$, and body forces, $\boldsymbol{f}$, belong to the set $\mathbb{R}^n$ of real vectors of known dimension n. The matrices $\boldsymbol{A}$ and $\boldsymbol{C}$ belong to the set $\mathcal{P}(n)$ of $n\times n$ real symmetric positive-definite matrices. The conditional expectation $\mathrm{E}_{\boldsymbol{u},\boldsymbol{f}|\boldsymbol{A},\boldsymbol{C}}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}\right]$, is the average of $\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}$ over the velocity, $\boldsymbol{u}$, and body forces, $\boldsymbol{f}$, for particular values of $\boldsymbol{A}$ and $\boldsymbol{C}$. We will refer to the matrix $\boldsymbol{C}$ as the model error covariance. It is possible that the model is biased, so to be strict we should perhaps refer to this as the mean-squared error. It is defined as $\boldsymbol{C}=\boldsymbol{Cov}(\boldsymbol{\epsilon})+\boldsymbol{b}\boldsymbol{b}^{\mathrm{T}}$, where $\boldsymbol{b}=\mathrm{E}_{\boldsymbol{u},\boldsymbol{f}|\boldsymbol{A},\boldsymbol{C}}[\boldsymbol{\epsilon}]$ is the expected bias of the model velocities represented by $\boldsymbol{A}^{-1}\boldsymbol{f}$, averaged over all possible forcings, $\boldsymbol{f}$, and $\boldsymbol{Cov}(\boldsymbol{\epsilon})=\mathrm{E}_{\boldsymbol{u},\boldsymbol{f}|\boldsymbol{A},\boldsymbol{C}}\left[(\boldsymbol{\epsilon}-\boldsymbol{b})(\boldsymbol{\epsilon}-\boldsymbol{b})^{\mathrm{T}}\right]$ is the error covariance of the model, averaged over all possible forcings.

Before looking at any data, we do not know anything about $\boldsymbol{u}$, $\boldsymbol{f}$, $\boldsymbol{A}$ or $\boldsymbol{C}$, except for the information contained in Eqn (13). Before we can use Bayes' theorem to make inferences about $\boldsymbol{u}$, $\boldsymbol{f}$, $\boldsymbol{A}$ and $\boldsymbol{C}$ from data, we need a prior PDF so that $p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$ is the prior probability that the parameters lie within a small interval, $\mathrm{d}\boldsymbol{\theta}$, of parameter space located at $\boldsymbol{\theta}$. For our particular choice of parameters, this takes the form

$$\begin{aligned}&p(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^{\mathrm{u}},\boldsymbol{C}^{\mathrm{u}})\,\mathrm{d}\boldsymbol{u}\,\mathrm{d}\boldsymbol{f}\,\mathrm{d}\boldsymbol{A}^{\mathrm{u}}\,\mathrm{d}\boldsymbol{C}^{\mathrm{u}}\equiv\\ &\quad\text{probability parameters lie within volume element}\\ &\quad\{\mathrm{d}\boldsymbol{u}\,\mathrm{d}\boldsymbol{f}\,\mathrm{d}\boldsymbol{A}^{\mathrm{u}}\,\mathrm{d}\boldsymbol{C}^{\mathrm{u}}\}\text{ at }\{\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^{\mathrm{u}},\boldsymbol{C}^{\mathrm{u}}\},\end{aligned}\tag{14}$$

where $\mathrm{d}\boldsymbol{u}=\prod_i \mathrm{d}u_i$, $\mathrm{d}\boldsymbol{f}=\prod_i \mathrm{d}f_i$, $\mathrm{d}\boldsymbol{A}^{\mathrm{u}}=\prod_{i,j\geq i}\mathrm{d}A_{ij}$ and $\mathrm{d}\boldsymbol{C}^{\mathrm{u}}=\prod_{i,j\geq i}\mathrm{d}C_{ij}$ are the standard (i.e. Lebesgue) measures for integration over elements of $\boldsymbol{u}$, $\boldsymbol{f}$, $\boldsymbol{A}^{\mathrm{u}}$ and $\boldsymbol{C}^{\mathrm{u}}$, the latter being the $n(n+1)/2$ upper triangular elements of $\boldsymbol{A}$ and $\boldsymbol{C}$, respectively. To label the domains of integration for $\boldsymbol{u}$, $\boldsymbol{f}$, $\boldsymbol{A}^{\mathrm{u}}$ and $\boldsymbol{C}^{\mathrm{u}}$, we write $\mathcal{U}=\{\boldsymbol{u}\in\mathbb{R}^n\}$, $\mathcal{F}=\{\boldsymbol{f}\in\mathbb{R}^n\}$, $\mathcal{A}^+=\left\{\boldsymbol{A}^{\mathrm{u}}\in\mathbb{R}^{n(n+1)/2}\text{ such that }\boldsymbol{A}(\boldsymbol{A}^{\mathrm{u}})\in\mathcal{P}(n)\right\}$, and $\mathcal{C}^+=\left\{\boldsymbol{C}^{\mathrm{u}}\in\mathbb{R}^{n(n+1)/2}\text{ such that }\boldsymbol{C}(\boldsymbol{C}^{\mathrm{u}})\in\mathcal{P}(n)\right\}$, and $\Theta=\mathcal{U}\times\mathcal{F}\times\mathcal{A}^+\times\mathcal{C}^+$ for the parameter space that combines all of $\mathcal{U}$, $\mathcal{F}$, $\mathcal{A}^+$ and $\mathcal{C}^+$.

The information defined by the prior information (Eqn (13)) is invariant under several transformations:

T1($\boldsymbol{\Phi}$): Orthogonal transformations, with $\boldsymbol{\Phi}$ an $n\times n$ orthogonal matrix, such that $\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}=\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}=\boldsymbol{I}$,

$$\boldsymbol{u}\mapsto\boldsymbol{\Phi}\boldsymbol{u},\quad \boldsymbol{f}\mapsto\boldsymbol{\Phi}\boldsymbol{f},\quad \boldsymbol{A}\mapsto\boldsymbol{\Phi}\boldsymbol{A}\boldsymbol{\Phi}^{\mathrm{T}},\quad \boldsymbol{C}\mapsto\boldsymbol{\Phi}\boldsymbol{C}\boldsymbol{\Phi}^{\mathrm{T}}.\tag{15}$$

T2(a,b): Change of units. Rescaling by $a>0$ and $b>0$,

$$\boldsymbol{u}\mapsto a\boldsymbol{u},\quad \boldsymbol{f}\mapsto ab\boldsymbol{f},\quad \boldsymbol{A}\mapsto b\boldsymbol{A},\quad \boldsymbol{C}\mapsto a^2\boldsymbol{C}.\tag{16}$$

T3($\boldsymbol{r}$): Superposition of solutions,

$$\boldsymbol{u}\mapsto\boldsymbol{u}+\boldsymbol{r},\quad \boldsymbol{f}\mapsto\boldsymbol{f}+\boldsymbol{A}\boldsymbol{r},\ \boldsymbol{A}\mapsto\boldsymbol{A},\quad \boldsymbol{C}\mapsto\boldsymbol{C}.\tag{17}$$

T4(q): Switch velocities for forces and model for inverse. Repeated q times, with $q=1$ or $q=2$,

$$\boldsymbol{u}\mapsto\boldsymbol{f},\quad \boldsymbol{f}\mapsto\boldsymbol{u},\quad \boldsymbol{A}\mapsto\boldsymbol{A}^{-1},\quad \boldsymbol{C}\mapsto\boldsymbol{A}\boldsymbol{C}\boldsymbol{A}.\tag{18}$$

According to Jaynes (2003) it is important to specify the transformations as a mathematical group. Mathematically, a *group* is a set of elements (e.g. A, B, C, etc.) that also has an operation, $\times$, that takes two elements A and B of the set and relates them to a third element P. To be a group the set and the operation must together satisfy certain conditions: (1) the operation must be *closed* so that the product, $P=A\times B$, always lies within the set; (2) the operation must be *associative*, so that for any A, B and C in the set, $(A\times B)\times C=A\times(B\times C)$; (3) the set must contain an *identity* element, I, such that $A\times I=A$ for any element A; (4) each element of the set must have an *inverse*, A^{-1}, such that $A\times A^{-1}=I$. In Appendix A we show that the transformations T1 to T4 satisfy these conditions individually and that their direct product defines a transformation group.

As noted by Jaynes (2003), the key role played by the preliminary reference prior, $\pi(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^{\mathrm{u}},\boldsymbol{C}^{\mathrm{u}})$, is to define a volume measure for the parameter space, Θ. To qualify as a transformation group prior, this volume measure must be invariant to the group of transformations that do not alter the mathematical specification of the problem, but it need not correspond to the Lebesgue volume measure, $\mathrm{dV}=\mathrm{d}\boldsymbol{u}\,\mathrm{d}\boldsymbol{f}\,\mathrm{d}\boldsymbol{A}^{\mathrm{u}}\,\mathrm{d}\boldsymbol{C}^{\mathrm{u}}$, that would apply if the parameter space had a standard Euclidean geometry. The following distance metric is invariant under the group of transformations defined above:

$$\begin{aligned}\mathrm{ds}^2=&\,\mathrm{Tr}\left[\boldsymbol{A}^{-1}\,\mathrm{d}\boldsymbol{A}\,\boldsymbol{A}^{-1}\,\mathrm{d}\boldsymbol{A}\right]+\mathrm{Tr}\left[\boldsymbol{C}^{-1}\,\mathrm{d}\boldsymbol{C}\,\boldsymbol{C}^{-1}\,\mathrm{d}\boldsymbol{C}\right]\\ &+\mathrm{d}\boldsymbol{u}^{\mathrm{T}}\,\boldsymbol{C}^{-1}\,\mathrm{d}\boldsymbol{u}+\mathrm{d}\boldsymbol{f}^{\mathrm{T}}\,(\boldsymbol{A}\boldsymbol{C}\boldsymbol{A})^{-1}\,\mathrm{d}\boldsymbol{f},\end{aligned}\tag{19}$$

where Tr indicates the trace, obtained by summing elements on the main diagonal. We have written $\mathrm{d}\boldsymbol{A}$, $\mathrm{d}\boldsymbol{C}$, $\mathrm{d}\boldsymbol{u}$ and $\mathrm{d}\boldsymbol{f}$ to represent infinitesimal changes to $\boldsymbol{A}$, $\boldsymbol{C}$, $\boldsymbol{u}$ and $\boldsymbol{f}$. The invariance of ds^2 to the transformation group can be shown by applying the transformations to Eqn (19), and using the invariance of $\mathrm{Tr}\left[\boldsymbol{A}^{-1}\,\mathrm{d}\boldsymbol{A}\,\boldsymbol{A}^{-1}\,\mathrm{d}\boldsymbol{A}\right]$ to inversion $\boldsymbol{A}\mapsto\boldsymbol{A}^{-1}$, scale changes $\boldsymbol{A}\mapsto b\boldsymbol{A}$, and to congruent transformations of the form $\boldsymbol{A}\mapsto\boldsymbol{B}\boldsymbol{A}\boldsymbol{B}^{\mathrm{T}}$, where $\boldsymbol{B}$ is an invertible $n\times n$ matrix (Moakher and Zéraï, 2011).

From expressions given by Moakher and Zéraï (2011), the volume element that corresponds to the distance metric, ds^2, is

$$dV = 2^{n(n-1)/2}|\boldsymbol{CA}|^{-(n+3)/2}\,d\boldsymbol{u}\,d\boldsymbol{f}\,d\boldsymbol{A}^u\,d\boldsymbol{C}^u, \tag{20}$$

where $d\boldsymbol{u}$, $d\boldsymbol{f}$, $d\boldsymbol{A}^u$ and $d\boldsymbol{C}^u$ are the usual Lebesgue measures. The following transformation group prior is therefore a suitable preliminary prior for derivation of a maximum-entropy PDF,

$$\begin{aligned}\pi(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^u,\boldsymbol{C}^u)\,d\boldsymbol{u}\,d\boldsymbol{f}\,d\boldsymbol{A}^u\,d\boldsymbol{C}^u = \\ Z_0^{-1}\,2^{n(n-1)/2}\,|\boldsymbol{CA}|^{-(n+3)/2}\,d\boldsymbol{u}\,d\boldsymbol{f}\,d\boldsymbol{A}^u\,d\boldsymbol{C}^u,\end{aligned} \tag{21}$$

where Z_0 is a non-dimensional constant that will be determined by normalization.

Because the matrices $\boldsymbol{A}$ and $\boldsymbol{C}$ are defined to be positive-definite, the function $\pi(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^u,\boldsymbol{C}^u)$ is finite everywhere within the parameter space, Θ, provided that Z_0 is finite. However, $\pi(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^u,\boldsymbol{C}^u)$ is an 'improper' prior, as described by Jaynes (2003). This means that if we attempt to integrate Eqn (20) over the parameter space, Θ, we find that this integral does not exist. Therefore a finite Z_0 cannot be defined such that the prior, $\pi(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^u,\boldsymbol{C}^u)$, is a normalized PDF over the entire parameter space, Θ. To allow us to interpret the function $\pi(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^u,\boldsymbol{C}^u)$ as a PDF, we will consider a restricted section of parameter space, Θ_ϵ, for which the necessary integral exists, and for which there is a well-defined limiting process $\Theta_\epsilon \to \Theta$ as $\epsilon \to 0^+$. For example, rather than an improper uniform prior for a single parameter over the interval $(-\infty,\infty)$, a well-defined uniform prior can be defined over a range $[-l/\epsilon, l/\epsilon]$, with l a constant that is finite and positive. Restricted domains $\mathcal{U}_\epsilon$, $\mathcal{F}_\epsilon$, $\mathcal{A}_\epsilon^+$ and $\mathcal{C}_\epsilon^+$ that are subsets of $\mathcal{U}$, $\mathcal{F}$, $\mathcal{A}^+$ and $\mathcal{C}^+$, respectively, are defined in Appendix B. The restricted parameter space is then derived from the Cartesian product, $\Theta_\epsilon = \mathcal{U}_\epsilon \times \mathcal{F}_\epsilon \times \mathcal{A}_\epsilon^+ \times \mathcal{C}_\epsilon^+$. Later, having derived a posterior PDF that depends upon ϵ, we can investigate its behavior in the limit $\epsilon \to 0^+$. In Appendix B, we also define a separate non-dimensional parameter, γ, that controls the smallest diagonal entries of positive-definite matrices.

Having defined the restricted parameter space, Θ_ϵ, we can seek the PDF, $p(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^u,\boldsymbol{C}^u)$, that maximizes the relative entropy, H,

$$H(p) \equiv -\int_{\Theta_\epsilon} p(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^u,\boldsymbol{C}^u)\ln\left[\frac{p(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^u,\boldsymbol{C}^u)}{\pi(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^u,\boldsymbol{C}^u)}\right]d\boldsymbol{u}\,d\boldsymbol{f}\,d\boldsymbol{A}^u\,d\boldsymbol{C}^u, \tag{22}$$

subject to whatever constraints are imposed. In our case the constraints are that the PDF must be normalized, so that

$$\int_{\Theta_\epsilon} p(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^u,\boldsymbol{C}^u)\,d\boldsymbol{u}\,d\boldsymbol{f}\,d\boldsymbol{A}^u\,d\boldsymbol{C}^u = 1, \tag{23}$$

and that the conditional expectation of the error covariance, $\mathrm{E}_{\boldsymbol{u},\boldsymbol{f}|\boldsymbol{A},\boldsymbol{C}}\left[(\boldsymbol{u}-\boldsymbol{A}^{-1}\boldsymbol{f})(\boldsymbol{u}-\boldsymbol{A}^{-1}\boldsymbol{f})^{\mathsf{T}}\right]$ is equal to $\boldsymbol{C}$. Using the product rule, $P(A|B)P(B) = P(A,B)$, for conditional probabilities of events A and B gives

$$\begin{aligned}p(\boldsymbol{u},\boldsymbol{f}\,|\,\boldsymbol{A}^u,\boldsymbol{C}^u) &= p(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^u,\boldsymbol{C}^u)/p_{n2}(\boldsymbol{A}^u,\boldsymbol{C}^u),\\ p_{n2}(\boldsymbol{A}^u,\boldsymbol{C}^u) &= \int_{\mathcal{U}_\epsilon\mathcal{F}_\epsilon} p(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^u,\boldsymbol{C}^u)\,d\boldsymbol{u}\,d\boldsymbol{f}.\end{aligned} \tag{24}$$

The constraint, $\mathrm{E}_{\boldsymbol{u},\boldsymbol{f}|\boldsymbol{A},\boldsymbol{C}}\left[(\boldsymbol{u}-\boldsymbol{A}^{-1}\boldsymbol{f})(\boldsymbol{u}-\boldsymbol{A}^{-1}\boldsymbol{f})^{\mathsf{T}}\right] = \boldsymbol{C}$, takes the form

$$\int_{\mathcal{U}_\epsilon\mathcal{F}_\epsilon} p(\boldsymbol{u},\boldsymbol{f}\,|\,\boldsymbol{A}^u,\boldsymbol{C}^u)\left[(\boldsymbol{u}-\boldsymbol{A}^{-1}\boldsymbol{f})(\boldsymbol{u}-\boldsymbol{A}^{-1}\boldsymbol{f})^{\mathsf{T}}-\boldsymbol{C}\right]d\boldsymbol{u}\,d\boldsymbol{f} = 0. \tag{25}$$

The constraints can be imposed using Lagrange multipliers, λ_1, a scalar, and $\boldsymbol{\Lambda}_2(\boldsymbol{A}^u,\boldsymbol{C}^u)$, a positive-definite symmetric matrix. We seek stationary points for the following quantity:

$$\begin{aligned}&H_2(p,\lambda_1,\boldsymbol{\Lambda}_2) \equiv\\ &-\int_{\Theta_\epsilon} p(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^u,\boldsymbol{C}^u)\ln\left[\frac{p(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^u,\boldsymbol{C}^u)}{\pi(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^u,\boldsymbol{C}^u)}\right]d\boldsymbol{u}\,d\boldsymbol{f}\,d\boldsymbol{A}^u\,d\boldsymbol{C}^u\\ &+\lambda_1\left(\int_{\Theta_\epsilon} p(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^u,\boldsymbol{C}^u)\,d\boldsymbol{u}\,d\boldsymbol{f}\,d\boldsymbol{A}^u\,d\boldsymbol{C}^u - 1\right)\\ &-\int_{\mathcal{A}_\epsilon^+\mathcal{C}_\epsilon^+}\mathrm{Tr}\left(\boldsymbol{\Lambda}_2\left[\int_{\mathcal{U}_\epsilon\mathcal{F}_\epsilon} p(\boldsymbol{u},\boldsymbol{f}|\boldsymbol{A}^u,\boldsymbol{C}^u)\right.\right.\\ &\qquad\left.\left.(\boldsymbol{u}-\boldsymbol{A}^{-1}\boldsymbol{f})(\boldsymbol{u}-\boldsymbol{A}^{-1}\boldsymbol{f})^{\mathsf{T}}\,d\boldsymbol{u}\,d\boldsymbol{f}-\boldsymbol{C}\right]\right)d\boldsymbol{A}^u\,d\boldsymbol{C}^u.\end{aligned} \tag{26}$$

Wherever $p_{n2}(\boldsymbol{A}^u,\boldsymbol{C}^u)$ does not vanish, we can define

$$\begin{aligned}Z_1 &= \exp[1-\lambda_1],\\ \boldsymbol{\Psi}(\boldsymbol{A}^u,\boldsymbol{C}^u) &= \boldsymbol{\Lambda}_2(\boldsymbol{A}^u,\boldsymbol{C}^u)/p_{n2}(\boldsymbol{A}^u,\boldsymbol{C}^u).\end{aligned} \tag{27}$$

Then H_2 is stationary for PDFs of the form

$$p(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^u,\boldsymbol{C}^u) = Z_1^{-1}\pi(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^u,\boldsymbol{C}^u)\mathrm{e}^{-\mathrm{Tr}\left\{\boldsymbol{\Psi}\left[(\boldsymbol{u}-\boldsymbol{A}^{-1}\boldsymbol{f})(\boldsymbol{u}-\boldsymbol{A}^{-1}\boldsymbol{f})^{\mathsf{T}}-\boldsymbol{C}\right]\right\}}. \tag{28}$$

The normalization constraint is satisfied for

$$\begin{aligned}&Z_1(\boldsymbol{\Psi}) =\\ &\int_{\Theta_\epsilon}\pi(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^u,\boldsymbol{C}^u)\mathrm{e}^{-\mathrm{Tr}\left\{\boldsymbol{\Psi}\left[(\boldsymbol{u}-\boldsymbol{A}^{-1}\boldsymbol{f})(\boldsymbol{u}-\boldsymbol{A}^{-1}\boldsymbol{f})^{\mathsf{T}}-\boldsymbol{C}\right]\right\}}\,d\boldsymbol{u}\,d\boldsymbol{f}\,d\boldsymbol{A}^u\,d\boldsymbol{C}^u,\end{aligned} \tag{29}$$

in which Z_1 is regarded as a functional of $\boldsymbol{\Psi}(\boldsymbol{A}^u,\boldsymbol{C}^u)$. To evaluate the function $\boldsymbol{\Psi}$ we require that the first variation of $Z_1(\boldsymbol{\Psi})$ with respect to $\boldsymbol{\Psi}$ is zero. This is analogous to identifying values for Lagrange multipliers by solving Eqn (8). Since the preliminary prior, $\pi(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^u,\boldsymbol{C}^u)$, is independent of $\boldsymbol{u}$, carrying out an integration over $\mathcal{U}$ provides the approximation

$$\begin{aligned}&Z_1(\boldsymbol{\Psi}) =\\ &\pi^{\frac{n}{2}}\int_{\mathcal{F}_\epsilon\mathcal{A}_\epsilon^+\mathcal{C}_\epsilon^+}\pi(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^u,\boldsymbol{C}^u)|\boldsymbol{\Psi}|^{-\frac{1}{2}}\mathrm{e}^{\mathrm{Tr}[\boldsymbol{\Psi C}]}\,d\boldsymbol{f}\,d\boldsymbol{A}^u\,d\boldsymbol{C}^u + \text{E.E.},\end{aligned} \tag{30}$$

where π written without arguments represents the constant, and E.E. represents edge effects arising because we have integrated over $\mathcal{U}$ rather than $\mathcal{U}_\epsilon$. Here we neglect these edge effects, in anticipation that they become unimportant in the limit $\epsilon \to 0^+$, whereupon $\mathcal{U}_\epsilon \to \mathcal{U}$. Then, requiring that the first variation of $Z_1(\boldsymbol{\Psi})$ with respect to $\boldsymbol{\Psi}$ is zero provides

$$\begin{aligned}&\delta Z_1(\boldsymbol{\Psi}) =\\ &\pi^{\frac{n}{2}}\int_{\mathcal{F}_\epsilon\mathcal{A}_\epsilon^+\mathcal{C}_\epsilon^+}\pi(\boldsymbol{u},\boldsymbol{f},\boldsymbol{A}^u,\boldsymbol{C}^u)|\boldsymbol{\Psi}|^{-\frac{1}{2}}\mathrm{e}^{\mathrm{Tr}[\boldsymbol{\Psi C}]}\delta\boldsymbol{\Psi}\left(\boldsymbol{C}-\frac{1}{2}\boldsymbol{\Psi}^{-1}\right)d\boldsymbol{f}\,d\boldsymbol{A}^u d\boldsymbol{C}^u\\ &= 0.\end{aligned} \tag{31}$$

Since this must be true for any $\delta\boldsymbol{\Psi}$, and the quantities preceding $\delta\boldsymbol{\Psi}$ in the integrand are all positive, applying the

fundamental lemma of the calculus of variations provides

$$\boldsymbol{\Psi} = \frac{1}{2}\,\boldsymbol{C}^{-1}. \tag{32}$$

We therefore arrive at the following expression for the prior:

$$p(\boldsymbol{u}, \boldsymbol{f}, \boldsymbol{A}^{\mathrm{u}}, \boldsymbol{C}^{\mathrm{u}}) =$$
$$Z_2^{-1}\pi(\boldsymbol{u}, \boldsymbol{f}, \boldsymbol{A}^{\mathrm{u}}, \boldsymbol{C}^{\mathrm{u}}) \exp\left[-\frac{1}{2}\,(\boldsymbol{u} - \boldsymbol{A}^{-1}\boldsymbol{f})^{\mathsf{T}}\boldsymbol{C}^{-1}(\boldsymbol{u} - \boldsymbol{A}^{-1}\boldsymbol{f})\right],$$
$$Z_2 = \int_{\Theta_\epsilon} \pi(\boldsymbol{u}, \boldsymbol{f}, \boldsymbol{A}^{\mathrm{u}}, \boldsymbol{C}^{\mathrm{u}}) \exp\left[-\frac{1}{2}\,(\boldsymbol{u} - \boldsymbol{A}^{-1}\boldsymbol{f})^{\mathsf{T}}\boldsymbol{C}^{-1}(\boldsymbol{u} - \boldsymbol{A}^{-1}\boldsymbol{f})\right] \mathrm{d}\boldsymbol{u}\,\mathrm{d}\boldsymbol{f}\,\mathrm{d}\boldsymbol{A}^{\mathrm{u}}\,\mathrm{d}\boldsymbol{C}^{\mathrm{u}}. \tag{33}$$

9. INTRODUCING ADDITIONAL INFORMATION USING BAYES' THEOREM

Deriving the prior PDF is only the first step of our inversion. We still have information available to us that we have not used. In particular, we have not yet introduced any of the observational data, $\boldsymbol{x}$. We will assume that these data provide an estimate for the velocity at the upper surface, $\boldsymbol{u}^*$, and that it also allows us to estimate the forces, $\boldsymbol{f}^*$. Having obtained a prior density function and a likelihood function, we can write the posterior PDF (Eqn (1)) using Bayes' rule as

$$p_{\mathrm{p}}(\boldsymbol{u}, \boldsymbol{f}, \boldsymbol{A}^{\mathrm{u}}, \boldsymbol{C}^{\mathrm{u}} | \boldsymbol{u}^*, \boldsymbol{f}^*) = \frac{p_{\mathrm{l}}(\boldsymbol{u}^*, \boldsymbol{f}^* | \boldsymbol{u}, \boldsymbol{f}, \boldsymbol{A}^{\mathrm{u}}, \boldsymbol{C}^{\mathrm{u}}) p(\boldsymbol{u}, \boldsymbol{f}, \boldsymbol{A}^{\mathrm{u}}, \boldsymbol{C}^{\mathrm{u}})}{p_{\mathrm{n}}(\boldsymbol{u}^*, \boldsymbol{f}^*)}, \tag{34}$$

with

$$p_{\mathrm{n}}(\boldsymbol{u}^*, \boldsymbol{f}^*) = \int_{\Theta_\epsilon} p_{\mathrm{l}}(\boldsymbol{u}^*, \boldsymbol{f}^* | \boldsymbol{u}, \boldsymbol{f}, \boldsymbol{A}^{\mathrm{u}}, \boldsymbol{C}^{\mathrm{u}}) p(\boldsymbol{u}, \boldsymbol{f}, \boldsymbol{A}^{\mathrm{u}}, \boldsymbol{C}^{\mathrm{u}})\,\mathrm{d}\boldsymbol{u}\,\mathrm{d}\boldsymbol{f}\,\mathrm{d}\boldsymbol{A}^{\mathrm{u}}\,\mathrm{d}\boldsymbol{C}^{\mathrm{u}}, \tag{35}$$

where $p_{\mathrm{l}}(\boldsymbol{u}^*, \boldsymbol{f}^* | \boldsymbol{u}, \boldsymbol{f}, \boldsymbol{A}^{\mathrm{u}}, \boldsymbol{C}^{\mathrm{u}})$ is the likelihood function and $p_{\mathrm{n}}(\boldsymbol{u}^*, \boldsymbol{f}^*)$ is the normalizing constant. The posterior PDF should then approach a well-defined limit as $\epsilon \to 0^+$ and $\gamma \to 0^+$, so that $\Theta_\epsilon \to \Theta$. If it does not, this may be a sign that the problem remains ill-posed and additional information is needed. Other information could also be introduced using Bayes' theorem. For instance, we may have data from laboratory experiments that can help to constrain the ice viscosity, or there may be additional information about basal drag from experts in subglacial processes. We have tried to define a very general prior that does not rely on such expert knowledge, but if credible information from experts can be obtained there is no reason it could not be introduced later using Bayes' theorem.

10. RETURNING TO THE SIMPLE SLAB PROBLEM

We now return to our simple slab problem. To give a practical illustration of how a Bayesian estimation might proceed, we will derive a posterior PDF for the basal drag coefficient, β, and the viscosity, η, for our slab. For this example we will make numerous simplifying assumptions, some of which would not be applicable in real situations. Nevertheless, many of the methods that we will use would apply to more general problems.

Specifying the likelihood requires knowledge of the accuracy of the various observations that are to be introduced, and how the errors in those observations are correlated. In a real inversion the estimation of the likelihood function might be quite involved, but to illustrate the methods we will assume Gaussian distributed errors in the estimation of surface velocity, $\boldsymbol{u}^*$, and body forces, $\boldsymbol{f}^*$. We then have the likelihood function

$$p_{\mathrm{l}}(\boldsymbol{u}^*, \boldsymbol{f}^* | \boldsymbol{u}, \boldsymbol{f}, \boldsymbol{A}^{\mathrm{u}}, \boldsymbol{C}^{\mathrm{u}}) =$$
$$\exp\left[-\frac{1}{2}\,(\boldsymbol{H}\boldsymbol{u} - \boldsymbol{u}^*)^{\mathsf{T}}\boldsymbol{R}_{\mathrm{u}}^{-1}(\boldsymbol{H}\boldsymbol{u} - \boldsymbol{u}^*) - \frac{1}{2}\,(\boldsymbol{f} - \boldsymbol{f}^*)^{\mathsf{T}}\boldsymbol{R}_{\mathrm{f}}^{-1}(\boldsymbol{f} - \boldsymbol{f}^*)\right], \tag{36}$$

where $\boldsymbol{R}_{\mathrm{u}}$ and $\boldsymbol{R}_{\mathrm{f}}$ are the error covariances for the observations $\boldsymbol{u}^*$ and $\boldsymbol{f}^*$, respectively. Note that these are distinct from the model error, $\boldsymbol{C}$, that was introduced previously. For our simple slab model we only have one observation of surface velocity, so $\boldsymbol{u}^*$ and $\boldsymbol{R}_{\mathrm{u}}$ are scalars, and $\boldsymbol{H} = [0, 0, 0, \ldots, 0, 0, 1]$ simply selects the surface velocity, so that $\boldsymbol{H}\boldsymbol{u} = u_{\mathrm{n}}$.

The observational data, $\boldsymbol{u}^*$ and f^*, are not the only information that we need to include. We also know that the form of the system matrix is given by $\boldsymbol{A} = \boldsymbol{X}^{\mathsf{T}}\boldsymbol{D}\boldsymbol{X}$, where $\boldsymbol{D}$ is diagonal and the matrix $\boldsymbol{X}$ is known. To begin with, we assume no information about the basal drag coefficient, β, or viscosity, η, except that these are positive, so we treat the matrix $\boldsymbol{D}$ as an unknown diagonal matrix with positive elements $\boldsymbol{D}^{\mathrm{d}} = [D_{11}, D_{22}, \ldots, D_{nn}]$ on its diagonal. To introduce this information we apply a one-to-one coordinate transformation, defined in Appendix B, from $\boldsymbol{A}^{\mathrm{u}}$ to $\boldsymbol{Y} = \left[\boldsymbol{D}^{\mathrm{d}}, \boldsymbol{M}^{\mathrm{sl}}\right]$ and we recognize that our system matrix is a special case of Eqns (B1) and (B2), where $\boldsymbol{M}$ is the identity matrix. This corresponds to $\boldsymbol{M}^{\mathrm{sl}} \to 0$, or, more precisely, all elements of $\boldsymbol{M}^{\mathrm{sl}}$ bounded within the interval $[-\delta M, \delta M]$, defined by some vanishingly small quantity, δM. To transform the PDF to the new coordinates we need to know the Jacobian of the coordinate transform from $\boldsymbol{A}^{\mathrm{u}}$ to $\boldsymbol{Y}$. This can be found by explicitly writing out the dependence of A_{ij} on M_{ij}, D_{ii} and X_{ij} and differentiating, then reordering the variables so that the Jacobian of this transformation is a triangular matrix (e.g. Magnus and Neudecker, 1980; Mathai, 1997). The determinant of this triangular matrix is then given by the product of its diagonal elements, which results in

$$|\mathcal{J}_1| = \left|\frac{\partial \boldsymbol{A}^{\mathrm{u}}}{\partial \boldsymbol{Y}}\right| = |\boldsymbol{X}|^{n+1}|\boldsymbol{D}|^{\frac{n-1}{2}}. \tag{37}$$

This can be used to define the transformed PDF

$$p_{\mathrm{p2}}(\boldsymbol{u}, \boldsymbol{f}, \boldsymbol{D}^{\mathrm{d}}, \boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{C}^{\mathrm{u}} | \boldsymbol{u}^*, \boldsymbol{f}^*) = p_{\mathrm{p}}(\boldsymbol{u}, \boldsymbol{f}, \boldsymbol{A}^{\mathrm{u}}(\boldsymbol{Y}), \boldsymbol{C}^{\mathrm{u}} | \boldsymbol{u}^*, \boldsymbol{f}^*)|\mathcal{J}_1|. \tag{38}$$

Using the product rule, $P(A|B)P(B) = P(A, B)$, for conditional probabilities of events A and B gives

$$p_{\mathrm{p3}}(\boldsymbol{u}, \boldsymbol{f}, \boldsymbol{D}^{\mathrm{d}}, \boldsymbol{C}^{\mathrm{u}} | \boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{u}^*, \boldsymbol{f}^*) = \frac{p_{\mathrm{p2}}(\boldsymbol{u}, \boldsymbol{f}, \boldsymbol{D}^{\mathrm{d}}, \boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{C}^{\mathrm{u}} | \boldsymbol{u}^*, \boldsymbol{f}^*)}{p_{\mathrm{n3}}(\boldsymbol{M}^{\mathrm{sl}} | \boldsymbol{u}^*, \boldsymbol{f}^*)}, \tag{39}$$

which provides the posterior PDF for unknown quantities, given values for the data, $\boldsymbol{u}^*$, $\boldsymbol{f}^*$, and the known matrix elements, $\boldsymbol{M}^{\mathrm{sl}}$. As usual, the denominator,

$$p_{\mathrm{n3}}(\boldsymbol{M}^{\mathrm{sl}} | \boldsymbol{u}^*, \boldsymbol{f}^*) =$$
$$\int_{\mathcal{U}_\epsilon \mathcal{F}_\epsilon \mathcal{D}_\epsilon^+ \mathcal{C}_\epsilon^+} p_{\mathrm{p2}}(\boldsymbol{u}, \boldsymbol{f}, \boldsymbol{D}^{\mathrm{d}}, \boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{C}^{\mathrm{u}} | \boldsymbol{u}^*, \boldsymbol{f}^*)\,\mathrm{d}\boldsymbol{u}\,\mathrm{d}\boldsymbol{f}\,\mathrm{d}\boldsymbol{D}^{\mathrm{d}}\,\mathrm{d}\boldsymbol{C}^{\mathrm{u}}, \tag{40}$$

can be viewed as a normalizing constant. The notation $\mathcal{D}_\epsilon^+$ refers to the restricted domain $l_{\mathrm{D}}\gamma < D_i^{\mathrm{d}} < l_{\mathrm{D}}/\epsilon$, with l_{D}

positive and finite, and the standard (Lebesgue) measure $\mathrm{d}\boldsymbol{D}^{\mathrm{d}} = \prod_{\mathrm{i}} \mathrm{dD}^{\mathrm{d}}_{\mathrm{i}}$ is used.

Many similar manipulations can be considered. For instance, if we want to assume some particular model error, $\boldsymbol{C}^{\mathrm{u}}$, or, more precisely, that all elements of $\boldsymbol{C}^{\mathrm{u}}$ lie within some small interval, $[-\delta C, \delta C]$, of such an estimate for vanishingly small δC, we could modify this to

$$p_{\mathrm{p4}}(\boldsymbol{u}, \boldsymbol{f}, \boldsymbol{D}^{\mathrm{d}} | \boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{C}^{\mathrm{u}}, \boldsymbol{u}^*, \boldsymbol{f}^*) = \frac{p_{\mathrm{p2}}(\boldsymbol{u}, \boldsymbol{f}, \boldsymbol{D}^{\mathrm{d}}, \boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{C}^{\mathrm{u}} | \boldsymbol{u}^*, \boldsymbol{f}^*)}{p_{\mathrm{n4}}(\boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{C}^{\mathrm{u}} | \boldsymbol{u}^*, \boldsymbol{f}^*)}, \quad (41)$$

where the normalizing constant is then

$$p_{\mathrm{n4}}(\boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{C}^{\mathrm{u}} | \boldsymbol{u}^*, \boldsymbol{f}^*) = \int_{\mathcal{U}_\epsilon \mathcal{F}_\epsilon \mathcal{D}^+_\epsilon} p_{\mathrm{p2}}(\boldsymbol{u}, \boldsymbol{f}, \boldsymbol{D}^{\mathrm{d}}, \boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{C}^{\mathrm{u}} | \boldsymbol{u}^*, \boldsymbol{f}^*)\, \mathrm{d}\boldsymbol{u}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{D}^{\mathrm{d}}. \quad (42)$$

If we are more interested in estimating the parameters $\boldsymbol{D}^{\mathrm{d}}$ than the forces and velocities within the slab, we can integrate over $\boldsymbol{u}$ and $\boldsymbol{f}$ to compute the marginalized distribution,

$$p_{\mathrm{p5}}(\boldsymbol{D}^{\mathrm{d}} | \boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{C}^{\mathrm{u}}, \boldsymbol{u}^*, \boldsymbol{f}^*) = \int_{\mathcal{U}_\epsilon \mathcal{F}_\epsilon} p_{p4}(\boldsymbol{u}, \boldsymbol{f}, \boldsymbol{D}^{\mathrm{d}} | \boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{C}^{\mathrm{u}}, \boldsymbol{u}^*, \boldsymbol{f}^*)\, \mathrm{d}\boldsymbol{u}\, \mathrm{d}\boldsymbol{f}. \quad (43)$$

If we assume constant viscosity in the slab, we also know that $D_{ii} = D_{22}$ for $i > 2$. To use this information we make a second transformation of coordinates of the form $\widetilde{\boldsymbol{D}}^{\mathrm{d}} = \{D_\beta, D_\eta, \boldsymbol{\zeta}\}$, with $D_\beta = \beta\Delta^{-1} = D_{11}$, $D_\eta = \eta\Delta^{-2} = D_{22}$, and a set of residuals that will be assumed zero, given by $\boldsymbol{\zeta} = \{(D^{-1}_{33} - D^{-1}_{22}), (D^{-1}_{44} - D^{-1}_{22}), \ldots, (D^{-1}_{nn} - D^{-1}_{22})\}$. There is a subtlety in the choice of $\boldsymbol{\zeta}$: later we will explain why we have taken residuals of inverses of diagonal elements, rather than diagonal elements themselves. We use the Jacobian

$$|\mathcal{J}_2| = \left| \frac{\partial \boldsymbol{D}^{\mathrm{d}}}{\partial \widetilde{\boldsymbol{D}}^{\mathrm{d}}} \right| = \left| \prod_{i=1}^{n-2} \left(\zeta_i + D_\eta^{-1} \right)^{-2} \right|, \quad (44)$$

to give

$$p_{\mathrm{p6}}(D_\beta, D_\eta, \boldsymbol{\zeta} | \boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{C}^{\mathrm{u}}, \boldsymbol{u}^*, \boldsymbol{f}^*) = p_{\mathrm{p5}}(\boldsymbol{D}^{\mathrm{d}}(D_\beta, D_\eta, \boldsymbol{\zeta}) | \boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{C}^{\mathrm{u}}, \boldsymbol{u}^*, \boldsymbol{f}^*) |\mathcal{J}_2|. \quad (45)$$

Then the posterior PDF for D_β and D_η, assuming known $\boldsymbol{\zeta}$, $\boldsymbol{M}^{\mathrm{sl}}$, $\boldsymbol{C}^{\mathrm{u}}$, $\boldsymbol{u}^*$ and $\boldsymbol{f}^*$, is

$$p_{\mathrm{p7}}(D_\beta, D_\eta | \boldsymbol{\zeta}, \boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{C}^{\mathrm{u}}, \boldsymbol{u}^*, \boldsymbol{f}^*) = \frac{p_{\mathrm{p6}}(D_\beta, D_\eta, \boldsymbol{\zeta} | \boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{C}^{\mathrm{u}}, \boldsymbol{u}^*, \boldsymbol{f}^*)}{p_{\mathrm{n7}}(\boldsymbol{\zeta} | \boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{C}^{\mathrm{u}}, \boldsymbol{u}^*, \boldsymbol{f}^*)}, \quad (46)$$

with

$$p_{\mathrm{n7}}(\boldsymbol{\zeta} | \boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{C}^{\mathrm{u}}, \boldsymbol{u}^*, \boldsymbol{f}^*) = \int_{l_{\mathrm{D}}\gamma}^{l_{\mathrm{D}}/\epsilon} \int_{l_{\mathrm{D}}\gamma}^{l_{\mathrm{D}}/\epsilon} p_{\mathrm{p6}}(D_\beta, D_\eta, \boldsymbol{\zeta} | \boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{C}^{\mathrm{u}}, \boldsymbol{u}^*, \boldsymbol{f}^*) \mathrm{dD}_\beta\, \mathrm{dD}_\eta. \quad (47)$$

Making substitutions from the equations above, we have

$$p_{\mathrm{p7}}(D_\beta, D_\eta | \boldsymbol{\zeta}, \boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{C}^{\mathrm{u}}, \boldsymbol{u}^*, \boldsymbol{f}^*) = \frac{\int_{\mathcal{U}_\epsilon \mathcal{F}_\epsilon} |\mathcal{J}_1| |\mathcal{J}_2| p_{\mathrm{l}}(\boldsymbol{u}^*, \boldsymbol{f}^* | \boldsymbol{u}, \boldsymbol{f}, \boldsymbol{A}^{\mathrm{u}}, \boldsymbol{C}^{\mathrm{u}}) p(\boldsymbol{u}, \boldsymbol{f}, \boldsymbol{A}^{\mathrm{u}}, \boldsymbol{C}^{\mathrm{u}})\, \mathrm{d}\boldsymbol{u}\, \mathrm{d}\boldsymbol{f}}{p_{\mathrm{n7}}(\boldsymbol{\zeta} | \boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{C}^{\mathrm{u}}, \boldsymbol{u}^*, \boldsymbol{f}^*) p_{\mathrm{n4}}(\boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{C}^{\mathrm{u}} | \boldsymbol{u}^*, \boldsymbol{f}^*) p_{\mathrm{n}}(\boldsymbol{u}^*, \boldsymbol{f}^*)}. \quad (48)$$

Reassuringly, this can be interpreted as an application of Bayes' rule (Eqn (1)), with parameters $\boldsymbol{\theta} = \{D_\beta, D_\eta, \boldsymbol{u}, \boldsymbol{f}\}$ and data $\boldsymbol{x} = \left\{\boldsymbol{\zeta}, \boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{C}^{\mathrm{u}}, \boldsymbol{u}^*, \boldsymbol{f}^*\right\}$, followed by marginalization over $\boldsymbol{u}$ and $\boldsymbol{f}$. The Jacobians, $|\mathcal{J}_1|$ and $|\mathcal{J}_2|$, just apply the changes of coordinates mentioned above, and the chain rule, $P(A, B, C) = P(A|B, C)P(B|C)P(C)$, can be used to rewrite the denominator in the more usual form, $p_{\mathrm{n}}(\boldsymbol{x})$.

To evaluate the prior, $p(\boldsymbol{u}, \boldsymbol{f}, \boldsymbol{A}^{\mathrm{u}}, \boldsymbol{C}^{\mathrm{u}})$, we need the determinant $|\boldsymbol{A}|$. For our simple problem $\boldsymbol{X}$ is triangular and has a determinant given by the product of elements on the leading diagonal. This gives $|\boldsymbol{X}| = 1$, so $|\boldsymbol{A}|$ is given by $|\boldsymbol{A}| = |\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}| = |\boldsymbol{D}||\boldsymbol{X}|^2 = |\boldsymbol{D}|$. To impose that viscosity is constant in the slab, we require that all elements of $\boldsymbol{\zeta}$ are bounded within some small interval, $[-\delta\zeta, \delta\zeta]$, for vanishingly small $\delta\zeta$. Then $|\mathcal{J}_1|$, $|\mathcal{J}_2|$ and $|\boldsymbol{A}|$ are approximated by

$$\begin{aligned} |\mathcal{J}_1| &= |\boldsymbol{X}|^{n+1} |\boldsymbol{D}|^{\frac{n-1}{2}} = \left|D_\beta\right|^{\frac{(n-1)}{2}} \left|D_\eta\right|^{\frac{(n-1)(n-1)}{2}}, \\ |\mathcal{J}_2| &= \left|D_\eta\right|^{2(n-1)}, \\ |\boldsymbol{A}| &= \left| \prod_i D_{ii} \right| = D_\beta D_\eta^{(n-1)}. \end{aligned} \quad (49)$$

If we take γ to be a constant, independent of ϵ, then taking the limit $\epsilon \to 0^+$ in Eqn (48), produces the posterior PDF over a restricted section of parameter space for which all D_{ii} are greater than some constant, $l_{\mathrm{D}}\gamma$. Collecting all factors that do not depend on D_β or D_η into one constant of normalization, Z_8, then gives

$$\begin{aligned} p_{\mathrm{p8}}(D_\beta, D_\eta | \boldsymbol{\zeta}, \boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{C}^{\mathrm{u}}, \boldsymbol{u}^*, \boldsymbol{f}^*) &= \lim_{\epsilon \to 0^+} p_{\mathrm{p7}} \\ &= Z_8^{-1} D_\beta^{-2} D_\eta^{-2} \mathcal{I}(D_\beta, D_\eta), \end{aligned} \quad (50)$$

$$Z_8 = \int_{l_{\mathrm{D}}\gamma}^{\infty} \int_{l_{\mathrm{D}}\gamma}^{\infty} D_\beta^{-2} D_\eta^{-2} \mathcal{I}(D_\beta, D_\eta)\, \mathrm{dD}_\beta\, \mathrm{dD}_\eta, \quad (51)$$

where the integral, $\mathcal{I}(D_\beta, D_\eta)$, is

$$\mathcal{I}(D_\beta, D_\eta) = \int_{\mathcal{UF}} \mathrm{e}^{\left[-\frac{1}{2}(\boldsymbol{H}\boldsymbol{u}-\boldsymbol{u}^*)^{\mathrm{T}}\boldsymbol{R}_u^{-1}(\boldsymbol{H}\boldsymbol{u}-\boldsymbol{u}^*) - \frac{1}{2}(\boldsymbol{f}-\boldsymbol{f}^*)^{\mathrm{T}}\boldsymbol{R}_f^{-1}(\boldsymbol{f}-\boldsymbol{f}^*) - \frac{1}{2}(\boldsymbol{u}-\boldsymbol{A}^{-1}\boldsymbol{f})^{\mathrm{T}}\boldsymbol{C}^{-1}(\boldsymbol{u}-\boldsymbol{A}^{-1}\boldsymbol{f})\right]} \mathrm{d}\boldsymbol{u}\, \mathrm{d}\boldsymbol{f}. \quad (52)$$

This evaluates to

$$\mathcal{I}(D_\beta, D_\eta) = \sqrt{\frac{(2\pi)^{2n} |\boldsymbol{R}_{\mathrm{u}}| |\boldsymbol{R}_{\mathrm{f}}| |\boldsymbol{C}|}{|\boldsymbol{C}_2|}} \exp\left[-\frac{1}{2} \left(\boldsymbol{u}^* - \boldsymbol{H}\boldsymbol{A}^{-1}\boldsymbol{f}^*\right)^{\mathrm{T}} \boldsymbol{C}_2^{-1} \left(\boldsymbol{u}^* - \boldsymbol{H}\boldsymbol{A}^{-1}\boldsymbol{f}^*\right) \right], \quad (53)$$

where π written without arguments represents the constant, and

$$\boldsymbol{C}_2 = \left[\boldsymbol{R}_{\mathrm{u}} + \boldsymbol{H}\boldsymbol{C}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{H}\boldsymbol{A}^{-1}\boldsymbol{R}_{\mathrm{f}}\boldsymbol{A}^{-1}\boldsymbol{H}^{\mathrm{T}}\right]. \quad (54)$$

For our simple problem, the dependence of the matrix, A^{-1}, on D_β and D_η can be computed explicitly as

$$\boldsymbol{A}^{-1} = \boldsymbol{X}^{-1}\boldsymbol{D}^{-1}\boldsymbol{X}^{-\mathrm{T}},$$

$$\boldsymbol{X}^{-1} = \begin{bmatrix} 1 & & & & & \\ 1 & 1 & & & & \\ 1 & 1 & 1 & & & \\ & & & \ddots & & \\ 1 & 1 & \cdots & & 1 & \\ 1 & 1 & \cdots & & 1 & 1 \end{bmatrix}, \quad \boldsymbol{D}^{-1} = \begin{bmatrix} D_\beta^{-1} & & & & \\ & D_\eta^{-1} & & & \\ & & \ddots & & \\ & & & D_\eta^{-1} & \\ & & & & D_\eta^{-1} \end{bmatrix}, \quad (55)$$

so Eqns (50) and (51), with substitutions from Eqns (53–55), provide the posterior PDF for D_β and D_η.

For this simple example we have assumed that viscosity in the slab is constant, so elements of $\boldsymbol{\zeta}$ are bounded close to zero, and that $\boldsymbol{A} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}$, with $\boldsymbol{X}$ and $\boldsymbol{D}$ given by Eqn (11). We have also assumed that the model error, $\boldsymbol{C}$, is known, that the thickness, h, is known and that observational data for velocity, $\boldsymbol{u}^*$, and forces, $\boldsymbol{f}^*$, along with their error covariances, $\boldsymbol{R}_{\mathrm{u}}$ and $\boldsymbol{R}_{\mathrm{f}}$, are available. In the limit of vanishing model error and vanishing error in estimation of forces, such that $\boldsymbol{H}\boldsymbol{C}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{H}\boldsymbol{A}^{-1}\boldsymbol{R}_{\mathrm{f}}\boldsymbol{A}^{-1}\boldsymbol{H}^{\mathrm{T}} = 0$, we can make the further simplifying assumption, $\boldsymbol{C}_2 = \boldsymbol{R}_{\mathrm{u}}$. Then the integral defined by Eqn (51) can be evaluated in the limit $\gamma \to 0^+$. In that case, the posterior PDF defined by $p_{\mathrm{p9}} = \lim_{\gamma\to 0^+} p_{\mathrm{p8}}$ can be normalized, resulting in

$$p_{\mathrm{p9}}(D_\beta, D_\eta|\boldsymbol{\zeta}, \boldsymbol{M}^{\mathrm{sl}}, \boldsymbol{C}^{\mathrm{u}}, \boldsymbol{u}^*, \boldsymbol{f}^*) = \lim_{\gamma\to 0^+} p_{\mathrm{p8}}$$
$$= Z_9^{-1} D_\beta^{-2} D_\eta^{-2} \exp\left[-\frac{1}{2}\left(\boldsymbol{u}^* - \boldsymbol{H}\boldsymbol{A}^{-1}\boldsymbol{f}^*\right)^{\mathrm{T}} \boldsymbol{R}_{\mathrm{u}}^{-1}\left(\boldsymbol{u}^* - \boldsymbol{H}\boldsymbol{A}^{-1}\boldsymbol{f}^*\right)\right], \tag{56}$$

with

$$Z_9 = \lim_{\gamma\to 0^+} \int_{l_{\mathrm{D}}\gamma}^{\infty}\int_{l_{\mathrm{D}}\gamma}^{\infty} D_\beta^{-2} D_\eta^{-2} \exp\left[-\frac{1}{2}\left(\boldsymbol{u}^* - \boldsymbol{H}\boldsymbol{A}^{-1}\boldsymbol{f}^*\right)^{\mathrm{T}} \boldsymbol{R}_{\mathrm{u}}^{-1}\left(\boldsymbol{u}^* - \boldsymbol{H}\boldsymbol{A}^{-1}\boldsymbol{f}^*\right)\right] \mathrm{d}D_\beta\, \mathrm{d}D_\eta, \tag{57}$$

$$= \frac{2\boldsymbol{R}_{\mathrm{u}}}{(n-1)(n-2)^2(\rho g \sin\alpha)^2} \left\{\frac{\sqrt{\pi}\boldsymbol{u}^*}{\sqrt{2\boldsymbol{R}_{\mathrm{u}}}}\left[\mathrm{erf}\left(\frac{\boldsymbol{u}^*}{\sqrt{2\boldsymbol{R}_{\mathrm{u}}}}\right) + 1\right] + \exp\left[-\frac{1}{2}\boldsymbol{u}^*\boldsymbol{R}_{\mathrm{u}}^{-1}\boldsymbol{u}^*\right]\right\}. \tag{58}$$

Figure 2 shows this posterior PDF for basal drag coefficient, β, and viscosity, η, calculated according to Eqn (56). The plots show results for non-dimensional quantities, $\widetilde{\boldsymbol{u}} = \boldsymbol{u}/u^*$, $\widetilde{\boldsymbol{f}} = \boldsymbol{f}/(\rho g \sin\alpha)$, $\widetilde{z} = z/h$, $\widetilde{\beta} = \beta u^*/(\rho g h \sin\alpha) = D_\beta u^*/(\rho g(n-1)\sin\alpha)$, $\widetilde{\eta} = \eta u^*/(\rho g h^2 \sin\alpha) = D_\eta u^*/[\rho g(n-1)^2 \sin\alpha]$. Colors show the PDF normalized by the maximum value. In this example we set $n = 1000$ and $\boldsymbol{C}_2 = \boldsymbol{R}_u = (0.05u^*)^2$, which corresponds to 5% error in velocity, perfect knowledge of forces, $\boldsymbol{f}^*$, and negligible model error.

Figure 3 shows multiple profiles of non-dimensional velocity, $\widetilde{\boldsymbol{u}}$, as a function of non-dimensional depth, $\widetilde{z}$, overlain on the same plot. Different curves are plotted for each combination of non-dimensional basal drag coefficient, $\widetilde{\beta}$, and viscosity, $\widetilde{\eta}$. Each curve is colored according to the posterior probability for the particular combination of non-dimensional basal drag coefficient, $\widetilde{\beta}$, and viscosity, $\widetilde{\eta}$, that it represents. Profiles for higher probabilities are plotted on top of those with lower probabilities. The profile corresponding to values of $\widetilde{\beta}$ and $\widetilde{\eta}$ that maximize the posterior PDF is shown as a white dashed curve.

Interestingly, even though we did not introduce any information about the viscosity of the slab or the basal drag coefficient, the posterior PDF shown in Figure 2 has a well-defined maximum. Figure 3 shows that the parameter values that maximize the posterior PDF correspond to a particular velocity profile through the slab. As discussed by Jaynes (2003), the maximum entropy distribution can be interpreted as the distribution that can be achieved in the greatest

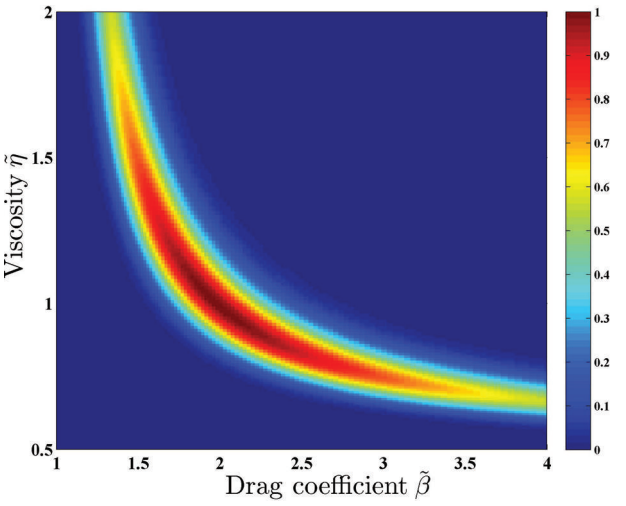

Fig. 2. The posterior PDF for different values of non-dimensional basal drag coefficient, $\widetilde{\beta}$, and viscosity, $\widetilde{\eta}$.

number of ways, subject to the imposed constraints, so we might perhaps expect this velocity profile to be realized with highest probability in a natural system, if viscosity is constant in the slab and the symmetries encoded by the transformation group are respected by the physical system, but there are otherwise no constraints on the values taken by the viscosity and basal drag coefficient. However, in a more realistic inversion, we would probably want to introduce information about the viscosity derived from laboratory experiments. If the uncertainty in viscosity can be estimated, so that a likelihood function can be derived, the information could be introduced using Bayes' theorem. Figures 4 and 5 show the effect of weighting the posterior PDF shown in Figure 2 by a Gaussian likelihood function, $\exp\left[-\frac{1}{2}(\widetilde{\eta} - 5)^2/0.25\right]$, which corresponds to an estimate $\widetilde{\eta} = 5 \pm 0.5$. Again, there is a preferred velocity profile, but now the most likely parameter values correspond to a velocity profile that is

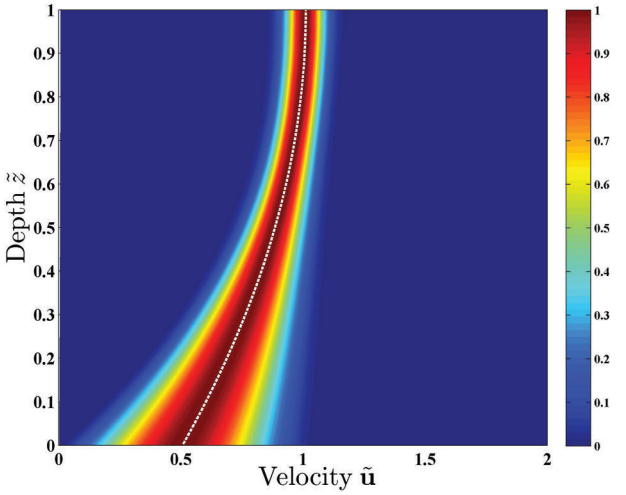

Fig. 3. Multiple profiles of non-dimensional velocity, $\widetilde{u}$, overlain on the same plot, each colored according to the posterior probability for the particular combination of non-dimensional basal drag coefficient, $\widetilde{\beta}$, and viscosity, $\widetilde{\eta}$, that it represents. The profile corresponding to values of $\widetilde{\beta}$ and $\widetilde{\eta}$ that maximize the posterior PDF is shown as a white dashed curve.

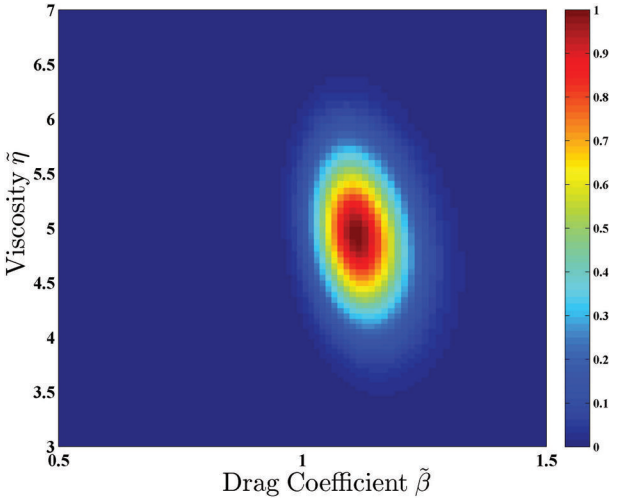

Fig. 4. As Figure 2, but with a constraint on non-dimensional viscosity, $\tilde{\eta} = 5 \pm 0.5$.

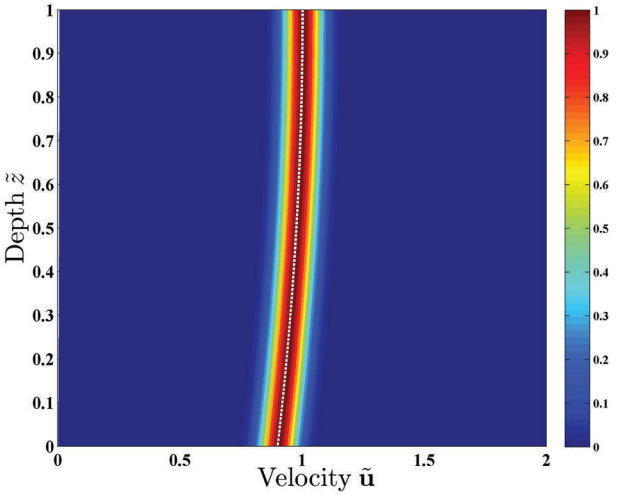

Fig. 5. As Figure 3, but with a constraint on non-dimensional viscosity, $\tilde{\eta} = 5 \pm 0.5$.

almost uniform with depth. An alternative approach would be to require that the expected value of the viscosity is equal to the laboratory-derived value, using a constraint of the form described by Eqn (3). Although we have concentrated here on a simple linear rheology, similar constraints could be applied using values of viscosity predicted by Glen's flow law, or some other rheology.

11. DISCUSSION

More realistic situations than a simple slab of uniform viscosity can obviously be imagined, but in this paper we wanted to illustrate the main features of the methods without introducing too many complications into the model. We have illustrated how manipulations of the PDF can be made using coordinate transformations, the product rule, marginalization and Bayes' theorem. For more general problems, some of the details would be different, but many aspects of the approach outlined above could still be applied.

To make our example problem as simple as possible we have made a number of assumptions that could be questioned. In addition to the observational constraints, we have imposed constraints on $\boldsymbol{\zeta}$, to make the viscosity uniform within the slab, and on $\boldsymbol{M}^{\mathrm{sl}}$, to impose the particular structure of our very simple finite-difference model. We have also assumed that $\boldsymbol{C}^{\mathrm{u}}$ and $\boldsymbol{R}_{\mathrm{f}}$ can be made arbitrarily small, to impose the assumption of negligible model error and negligible errors in the estimation of forces. Of course, in a more realistic situation, the viscosity would not be constant, the model structure would be more complicated and it would be difficult to justify either the perfect model assumption or perfect knowledge of forces. As an alternative to neglecting model error, we could perhaps marginalize over model error, $\boldsymbol{C}$, treating it as a nuisance parameter, or use some other estimate for model error, based on the approximate magnitude of neglected terms in an asymptotic expansion describing the model. It would also be worth trying to relax the unrealistic assumption that we make negligible error in estimation of forces, $\boldsymbol{f}^{*}$.

An important consideration in our example is that we simplified the PDF to a function of only two parameters so that it could be plotted. To do this, we imposed the constraints $\boldsymbol{\zeta} \to \boldsymbol{0}$, $\boldsymbol{M}^{\mathrm{sl}} \to \boldsymbol{0}$ and $\boldsymbol{C} \to \boldsymbol{0}^{+}$. However, these constraints do not provide sufficient information to provide a well-defined posterior PDF unless we also describe how these limits are approached. This is an example of the Borel–Kolmogorov paradox (e.g. Jaynes, 2003). The important point to consider is that conditional probabilities, e.g. $p(A|B)$, can only be defined unambiguously with respect to an event B that has non-zero probability, so we must consider a limiting process, such as the requirement specified above that all elements of $\boldsymbol{\zeta} = \{(D_{33}^{-1} - D_{22}^{-1}), (D_{44}^{-1} - D_{22}^{-1}), \ldots, (D_{nn}^{-1} - D_{22}^{-1})\}$ are bounded within some small interval, $[-\delta\zeta, \delta\zeta]$, for vanishingly small $\delta\zeta$. Perhaps surprisingly, there are many different ways to represent uniform viscosity in the slab, and these can result in different posterior PDFs. Instead of using $\boldsymbol{\zeta}$, we could have represented constant viscosity by requiring that all elements of $\widetilde{\boldsymbol{\zeta}} = \{(D_{33} - D_{22}), (D_{44} - D_{22}), \ldots, (D_{nn} - D_{22})\}$ are bounded within some small interval $\left[-\delta\widetilde{\zeta}, \delta\widetilde{\zeta}\right]$ for vanishingly small $\delta\widetilde{\zeta}$. Then we would have obtained a different Jacobian, $(|\mathcal{J}_2| = 1)$, instead of Eqn (44), and hence a different posterior PDF. Similarly, a different coordinate transformation could have been used in place of Eqns (B1) and (B2).

In situations where different coordinates, such as $\boldsymbol{\zeta}$ or $\widetilde{\boldsymbol{\zeta}}$, can be used it can be difficult to know which option should apply. In our example there are additional desirable symmetries that can perhaps help. Allowing for a factor of $|\boldsymbol{CA}|^{-1}$ in Eqn (21) that originates from the final two terms in Eqn (19) and provides scale-invariance, our use of Eqns (21), (37), (B1) and (B2) is consistent with treating diagonal elements of $\boldsymbol{D}$, which are known to be positive, as *scale* parameters, and elements of $\boldsymbol{M}^{\mathrm{sl}}$, which are known to be bounded within $[-1, 1]$, as *location* parameters in the terminology of Jeffreys (1961) and Jaynes (2003). Then, the advantage of using $\boldsymbol{\zeta}$ over $\widetilde{\boldsymbol{\zeta}}$ is that we obtain a PDF that does not depend upon n in the limit $n \to \infty$. This means the PDF we obtain converges to a well-defined function as we increase the resolution of the finite-difference model. It still seems possible that some other choice of variables might also satisfy this symmetry, so for now we make no claim that the particular posterior PDF that we have obtained is unique in

satisfying all of the symmetries of the problem that we have identified. If it turns out not to be unique, the question of whether there are other symmetries that we have yet to take into account would arise. Each symmetry that we can identify places constraints on the prior, but there is no guarantee that it can be specified uniquely.

Although we advocate the Bayesian approach to inversion, there may be problems for which it is too expensive to derive the posterior PDF in full. To consider how maximizing the posterior PDF relates to minimization of a cost function, we can take the negative of the logarithm of Eqn (56). We obtain

$$J_{\text{total}} = \frac{1}{2}\left(\boldsymbol{u}^* - \boldsymbol{H}\boldsymbol{A}^{-1}\boldsymbol{f}^*\right)^{\mathrm{T}}\boldsymbol{R}_{\mathrm{u}}^{-1}\left(\boldsymbol{u}^* - \boldsymbol{H}\boldsymbol{A}^{-1}\boldsymbol{f}^*\right) + 2\ln\beta + 2\ln\eta + J_0, \tag{59}$$

where J_0 is a constant offset. Since $\boldsymbol{H}\boldsymbol{A}^{-1}\boldsymbol{f}^*$ is a model-based estimate of the surface velocity, this is a quadratic misfit function, defined by the negative of the log-likelihood function, with an additional term, $J_{\text{reg}} = 2\ln\beta + 2\ln\eta$. This term appears in place of the more conventional Tikhonov regularization terms, J_{reg}, mentioned in Section 3. In this particular example, for which we have assumed we can neglect model error, there are no arbitrary coefficients λ_{reg}, so there is no requirement for an L-curve analysis to fix the degree of regularization applied. The general case, where model error cannot be neglected, would not be so straightforward.

The terms $2\ln\beta$ and $2\ln\eta$ penalize large values of basal drag, β, and viscosity, η. This provides qualitative support to the algorithms that are commonly used in glaciological inversions (e.g. Morlighem and others, 2013; Joughin and others, 2014; Petra and others, 2014; Arthern and others, 2015). However, most previous approaches have not exploited the symmetries of the model in the specification of their regularization, or in the characterization of the prior PDF, and therefore are effectively introducing additional information about the subglacial environment into the inversion. In itself this does not represent a problem, if some reasoned explanation can be provided for where this information has come from, but our poor knowledge of the subglacial environment perhaps makes it difficult to furnish a very convincing explanation. Comparing many diverse approaches to the inversion (e.g. Favier and others, 2014) and assessing the consequences for the simulation of the future behavior of the ice sheet would offer one way to explore the consequences for predictions of sea level.

It will be interesting to explore how these methods apply in models with a greater number of spatial dimensions than the simple slab problem. The prior determined by Eqn (33) is very general, and could be used for any discretized system of equations with a symmetric positive-definite matrix. As an example, a more complicated three-dimensional model of Antarctica can also be written in the form $\boldsymbol{A} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}$, with $\boldsymbol{X}$ known and $\boldsymbol{D}$ a diagonal matrix that depends upon viscosity and basal drag parameters (Arthern and others, 2015). In that case, the matrices $\boldsymbol{X}$ and $\boldsymbol{D}$ are different from the simple model considered above, and a different coordinate transformation would be needed to separate out the known and unknown parts of the matrix. Nevertheless, it seems likely that very similar methods could be used to provide a posterior PDF for the viscosity and basal drag in a realistic geographical setting.

Some of the mathematical aspects of our approach could be developed further. We have considered a discretized model from the outset, but the equivalent problem for continuous linear differential operators could also be investigated. Other avenues open to exploration include the application of similar methods to a nonlinear ice rheology, further characterization of group orbits produced as the transformation group acts on the parameter space, mathematical consideration of the actual length scales of basal drag that are important in the ice-sheet prediction problem, further investigation of sensitivity to the limiting process that is used to define the restricted parameter space in more general cases than considered here, and investigations that relate the prior defined here to other approaches (e.g. Jeffreys, 1961; Kass and Wasserman, 1996; Berger, 2006).

12. CONCLUSIONS

In this paper, we have described an exploratory study to investigate whether transformation group priors and the maximization of relative entropy might have a role to play in glaciological inversions for viscosity and basal drag. These inversions are used to initialize forecasts of the ice sheets, and their formulation in Bayesian terms is an essential prerequisite to probabilistic forecasts of the sea-level contribution from Greenland and Antarctica.

Our initial findings are that adopting a Bayesian approach that uses transformation group priors and the maximization of relative entropy does add complexity to the problem. Nevertheless, having investigated a very general problem with a model that is based upon a symmetric positive-definite matrix, and having applied this to a highly simplified problem for a slab of ice flowing down an inclined plane, it does seem that these methods could be used to initialize ice-sheet models. Rather than an ad hoc and subjective approach to regularization of the glaciological inverse problem, this would provide a more formulaic approach to the definition of priors for the parameters.

The great advantage of the Bayesian approach is that it allows the complete probability distribution of the model parameters to be evaluated. This could be of considerable value, either in setting up ensembles for Monte Carlo simulations of the future behavior of an ice sheet, or in more formal calculations of the probability of various possible contributions to sea level that might come from the ice sheets of Greenland and Antarctica.

ACKNOWLEDGEMENTS

The research was funded by the UK Natural Environment Research Council, including NERC grant NE/L005212/1. I am extremely grateful to the editor and two anonymous reviewers for very insightful and useful comments that significantly improved the manuscript. Richard Hindmarsh, Hilmar Gudmundsson, Jan De Rydt, Jonathan Kingslake and Dan Goldberg all provided useful comments on an earlier draft that have led to improvements.

REFERENCES

Arthern RJ and Gudmundsson GH (2010) Initialization of ice-sheet forecasts viewed as an inverse Robin problem. *J. Glaciol.*, **56**(197), 527–533 (doi: 10.3189/002214310792447699)

Arthern RJ, Hindmarsh RCA and Williams CR (2015) Flow speed within the Antarctic ice sheet and its controls inferred from satellite observations. *J. Geophys. Res. Earth Surf.*, **120**, 1171–1188 (doi: 10.1002/2014JF003239)

Bahr DB, Pfeffer WT and Meier MF (1994) Theoretical limitations to englacial velocity calculations. *J. Glaciol.*, **40**(135), 509–518

Berger J (2006) The case for objective Bayesian analysis. *Bayesian Anal.*, **1**(3), 385–402 (doi: 10.1214/06-BA115)

Berliner LM, Jezek K, Cressie N, Kim Y, Lam CQ and Van der Veen Q (2008) Modeling dynamic controls on ice streams: a Bayesian statistical approach. *J. Glaciol.*, **54**(187), 705–714 (doi: 10.3189/002214308786570917)

Bindschadler RA and 27 others (2013) Ice-sheet model sensitivities to environmental forcing and their use in projecting future sea level (the SeaRISE project). *J. Glaciol.*, **59**(214), 195–224 (doi: 10.3189/2013JoG12J125)

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

Favier L and 8 others (2014) Retreat of Pine Island Glacier controlled by marine ice-sheet instability. *Nature Climate Change*, **4**(2), 117–121

Fürst JJ and 7 others (2015) Assimilation of Antarctic velocity observations provides evidence for uncharted pinning points. *Cryosphere*, **9**, 1427–1443 (doi: 10.5194/tc-9-1427-2015)

Gillet-Chaulet F and 8 others (2012) Greenland ice sheet contribution to sea-level rise from a new-generation ice-sheet model. *Cryosphere*, **6**(6), 1561–1576 (doi: 10.1038/nclimate2094)

Gudmundsson GH (2003) Transmission of basal variability to a glacier surface. *J. Geophys. Res.: Solid Earth*, **108**(B5), 2253 (doi: 10.1029/2002JB002107)

Gudmundsson GH and Raymond M (2008) On the limit to resolution and information on basal properties obtainable from surface data on ice streams. *Cryosphere*, **2**, 167–178 (doi: 10.5194/tc-2-167-2008)

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

Hadamard J (1902) Sur les problèmes aux dérivés partielles et leur signification physique. *Princeton Univ. Bull.*, **13**, 49–52

Jay-Allemand M, Gillet-Chaulet F, Gagliardini O and Nodet M (2011) Investigating changes in basal conditions of Variegated Glacier prior to and during its 1982–1983 surge. *Cryosphere*, **5**(3), 659–672 (doi: 10.5194/tc-5-659-2011)

Jaynes ET (2003) *Probability theory: the logic of science*. Cambridge University Press, Cambridge

Jeffreys H (1961) *Theory of probability*, 3rd edn. Oxford University Press, Oxford

Joughin I, Smith BE and Medley B (2014) Marine ice sheet collapse potentially underway for the Thwaites Glacier basin, West Antarctica. *Science*, **344**(6185), 735–738 (doi: 10.1126/science.1249055)

Kass RE and Wasserman L (1996) The selection of prior distributions by formal rules. *J. Am. Stat. Assoc.*, **91**(435), 1343–1370 (doi: 10.2307/2291752)

MacAyeal DR (1992) The basal stress-distribution of Ice Stream-E, Antarctica, inferred by control methods. *J. Geophys. Res. Solid Earth*, **97**(B1), 595–603 (doi: 10.1029/91JB02454)

Magnus JR and Neudecker H (1980) The elimination matrix: some lemmas and applications. *SIAM J. Algebraic Discrete Meth.*, **1**(4), 422–449 (doi: 10.1137/0601049)

Mathai AM (1997) *Jacobians of matrix transformations and functions of matrix argument*. World Scientific Publishing, New York

Maxwell D, Truffer M, Avdonin S and Stuefer M (2008) An iterative scheme for determining glacier velocities and stresses. *J. Glaciol.*, **54**(188), 888–898 (doi: 10.3189/002214308787779889)

Moakher M and Zéraï M (2011) The Riemannian geometry of the space of positive-definite matrices and its application to the regularization of positive-definite matrix-valued data. *J. Math. Imaging Vis.*, **40**(2), 171–187 (doi: 10.1007/s10851-010-0255-x)

Morlighem M, Rignot E, Seroussi H, Larour E, Ben Dhia H and Aubry D (2010) Spatial patterns of basal drag inferred using control methods from a full-Stokes and simpler models for Pine Island Glacier, West Antarctica. *Geophys. Res. Lett.*, **37**, L14502 (doi: 10.1029/2010GL043853)

Morlighem M, Seroussi H, Larour E and Rignot E (2013) Inversion of basal friction in Antarctica using exact and incomplete adjoints of a higher-order model. *J. Geophys. Res. Earth Surf.*, **118**, 1746–1753 (doi: 10.1002/jgrf.20125)

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 J. Sci. Comput.*, **36**(4), A1525–A1555 (doi: 10.1137/130934805)

Raymond MJ and Gudmundsson GH (2009) Estimating basal properties of ice streams from surface measurements: a non-linear Bayesian inverse approach applied to synthetic data. *Cryosphere*, **3**(2), 265–278 (doi: 10.5194/tc-3-265-2009)

Schoof C (2007) Marine ice-sheet dynamics. Part 1. The case of rapid sliding. *J. Fluid Mech.*, **573**, 27–55 (doi: 10.1017/S0022112006003570)

Sergienko OV and Hindmarsh RCA (2013) Regular patterns in frictional resistance of ice-stream beds seen by surface data inversion. *Science*, **342**(6162), 1086–1089 (doi: 10.1126/science.1243903)

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 Planet. Sci. Lett.*, **315**(SI), 30–40 (doi: 10.1016/j.epsl.2011.09.010)

Vogel CR (1996) Non-convergence of the L-curve regularization parameter selection method. *Inverse Probl.*, **12**(4), 535 (doi: 10.1088/0266-5611/12/4/013)

Zammit-Mangion A, Rougier J, Bamber J and Schoen N (2014) Resolving the Antarctic contribution to sea-level rise: a hierarchical modelling framework. *Environmetrics*, **25**(4, SI), 245–264 (doi: 10.1002/env.2247)

APPENDIX A: THE TRANSFORMATION GROUP

Mathematically, a *group* is a set of elements (e.g. A, B, C, etc.) that also has an operation $\times$ that takes two elements, A and B, of the set and relates them to a third element, P. To be a group the set and the operation must together satisfy certain conditions: (1) the operation must be *closed*, so that the product $P = A \times B$ always lies within the set; (2) the operation must be *associative*, so that for any A, B and C in the set $(A \times B) \times C = A \times (B \times C)$; (3) the set must contain an *identity* element, I, such that $A \times I = A$ for any element A; (4) each element of the set must have an *inverse* A^{-1}, such that $A \times A^{-1} = I$.

Varying the parameter $\boldsymbol{\Phi}$ within the set of orthogonal matrices defines the set of possible transformations, $\mathrm{T1}(\boldsymbol{\Phi})$, that could be applied. For transformation groups the operation $\times$ represents successive application of two transformations from the set, so that $\mathrm{T1}(\boldsymbol{\Phi}_2) \times \mathrm{T1}(\boldsymbol{\Phi}_1)$ represents the transformation $\mathrm{T1}(\boldsymbol{\Phi}_1)$ followed by the transformation $\mathrm{T1}(\boldsymbol{\Phi}_2)$. It is then straightforward to show that $\mathrm{T1}(\boldsymbol{\Phi})$, together with this operation, defines a group. Closure follows, since

$$\mathrm{T1}(\boldsymbol{\Phi}_2) \times \mathrm{T1}(\boldsymbol{\Phi}_1) = \mathrm{T1}(\boldsymbol{\Phi}_2\boldsymbol{\Phi}_1), \tag{A1}$$

and $\boldsymbol{\Phi}_2\boldsymbol{\Phi}_1$ is orthogonal. Associativity follows from associativity of matrix multiplication:

$$\begin{aligned}\mathrm{T1}(\boldsymbol{\Phi}_3) \times (\mathrm{T1}(\boldsymbol{\Phi}_2) \times \mathrm{T1}(\boldsymbol{\Phi}_1)) = \mathrm{T1}(\boldsymbol{\Phi}_3(\boldsymbol{\Phi}_2\boldsymbol{\Phi}_1)) = \\ \mathrm{T1}((\boldsymbol{\Phi}_3\boldsymbol{\Phi}_2)\boldsymbol{\Phi}_1) = (\mathrm{T1}(\boldsymbol{\Phi}_3) \times \mathrm{T1}(\boldsymbol{\Phi}_2)) \times \mathrm{T1}(\boldsymbol{\Phi}_1).\end{aligned} \tag{A2}$$

The identity is $I_{T1} = T1(\boldsymbol{I})$, where $\boldsymbol{I}$ is the identity matrix. The inverse is $T1(\boldsymbol{\Phi}^T)$, since $T1(\boldsymbol{\Phi}^T) \times T1(\boldsymbol{\Phi}) = I_{T1}$.

Similar considerations can be used to show that when $\times$ represents successive application of the transformations, (T2(a,b), with $\times$), (T3($\boldsymbol{r}$), with $\times$) and (T4(q), with $\times$) also define transformation groups. For the transformation group defined by T2(a,b), closure follows from $T2(a_1,b_1) \times T2(a_2,b_2) = T2(a_1a_2, b_1b_2)$, associativity follows from that of multiplication, the identity is $I_{T2} = T2(1,1)$ and the inverse is $T2(1/a, 1/b)$. For T3($\boldsymbol{r}$), closure follows from $T3(\boldsymbol{r}_1) \times T3(\boldsymbol{r}_2) = T2(\boldsymbol{r}_1 + \boldsymbol{r}_2)$, associativity follows from that of addition, the identity is $I_{T3} = T3(\boldsymbol{0})$ and the inverse is T3(–$\boldsymbol{r}$). The transformation T4(1) is its own inverse, so T4(2) = T4(1) × T4(1) is the identity, I_{T4}. The set composed of T4(1) and T4(2), along with $\times$, defines a cyclic transformation group. For T4(q), closure and associativity $(A \times B) \times C = A \times (B \times C)$ can be verified trivially for all possible combinations of A, B and C that are either T4(1) or the identity, I_{T4} (i.e. T4(2)).

The four groups (T1($\boldsymbol{\Phi}$), with $\times$), (T2(a,b), with $\times$), (T3($\boldsymbol{r}$), with $\times$) and (T4(q), with $\times$) can be combined by taking their direct product. The direct product of groups is defined so that the sets of transformations are combined using their Cartesian product, and operations are applied component by component: a simple example of a direct product for two groups ($\{A_1, B_1\}$ with $\times_1$) and ($\{C_2, D_2\}$ with $\times_2$) is the group composed of the set of ordered pairs $\{(A_1, C_2), (A_1, D_2), (B_1, C_2), (B_1, D_2)\}$, with the operation $(A_1, C_2) \times (B_1, D_2) = (A_1 \times_1 B_1, C_2 \times_2 D_2)$.

APPENDIX B: THE RESTRICTED PARAMETER SPACE

We define the restricted domains $\mathcal{U}_\epsilon = \{\boldsymbol{u} \in [-l_u/\epsilon, l_u/\epsilon]^n\}$, $\mathcal{F}_\epsilon = \{\boldsymbol{f} \in [-l_f/\epsilon, l_f/\epsilon]^n\}$ from the Cartesian product of n finite intervals, with l_u and l_f being positive constants. As $\epsilon \to 0^+$, $\mathcal{U}_\epsilon$ and $\mathcal{F}_\epsilon$ approach the domains $\mathcal{U}$ and $\mathcal{F}$, respectively. To introduce a similar restricted domain, $\mathcal{A}_\epsilon^+$, for the matrix $\boldsymbol{A}$ it is useful to first consider a transformation of coordinates from $\boldsymbol{A}^u$. If $\boldsymbol{X}$ is an arbitrary invertible square matrix, such as the one defined by Eqn (11), any symmetric positive-definite matrix, $\boldsymbol{A}$, can be written as

$$\boldsymbol{A} = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}, \tag{B1}$$

with $\boldsymbol{W}$ symmetric positive-definite. Let $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{W})$ be the diagonal matrix composed of the diagonal elements of $\boldsymbol{W}$. All diagonal elements $\boldsymbol{D}^d = \{D_{11}, D_{22}, \ldots, D_{nn}\}$ are positive, since $\boldsymbol{W}$ is symmetric positive-definite, so $\boldsymbol{W}$ can be further decomposed as

$$\boldsymbol{W} = \boldsymbol{D}^{\frac{1}{2}} \boldsymbol{M} \boldsymbol{D}^{\frac{1}{2}}, \tag{B2}$$

where $\boldsymbol{M}$ has diagonal elements $M_{ii} = 1$. Our simple example system, defined by Eqn (11), corresponds to the special case where $\boldsymbol{M}$ is the identity. More generally, since $\boldsymbol{W}$ is symmetric positive-definite, $\boldsymbol{M}$ must also be symmetric positive-definite. A necessary (but not sufficient) condition for this is that the off-diagonal elements lie in the range $-1 < M_{i,j\neq i} < 1$. This implies that $\boldsymbol{M}^{sl}$, the elements of $\boldsymbol{M}$ that are strictly below the leading diagonal, lie within

$$\mathcal{M} = \{\boldsymbol{M}^{sl} \in (-1,1)^{n(n-1)/2}, \text{ such that } \boldsymbol{M}(\boldsymbol{M}^{sl}) \in \mathcal{P}(n)\}, \tag{B3}$$

where $\boldsymbol{M}(\boldsymbol{M}^{sl})$ has diagonal elements $M_{ii} = 1$, elements $\boldsymbol{M}^{sl}$ strictly below the leading diagonal and elements above the leading diagonal determined by the fact that $\boldsymbol{M}$ is symmetric.

For any particular invertible matrix, $\boldsymbol{X}$, there is a one-to-one coordinate transform from $\boldsymbol{A}^u$ to $\boldsymbol{Y} = \left[\boldsymbol{D}^d, \boldsymbol{M}^{sl}\right]$. In our restricted parameter space we require that diagonal elements of $\boldsymbol{D}$ lie in the range $\mathcal{D}_\epsilon^+ = \{\boldsymbol{D}^d \in [l_D\gamma, l_D/\epsilon]^n\}$ with l_D finite and positive. For now, we will leave γ as a general parameter that controls the smallest values of D_{ii}. Our restricted domain is then

$$\mathcal{A}_\epsilon^+ = \Big\{\boldsymbol{A}^u \in \mathbb{R}^{n(n+1)/2}, \text{ such that } \boldsymbol{A}(\boldsymbol{A}^u) = \boldsymbol{X}^T \boldsymbol{D}^{\frac{1}{2}} \boldsymbol{M} \boldsymbol{D}^{\frac{1}{2}} \boldsymbol{X}, \text{ with } \boldsymbol{M}^{sl} \in \mathcal{M} \text{ and } \boldsymbol{D}^d \in \mathcal{D}_\epsilon^+\}. \tag{B4}$$

We can derive a restricted domain, $\mathcal{C}_\epsilon^+$ for $\boldsymbol{C}^u$, in a similar way as

$$\mathcal{C}_\epsilon^+ = \Big\{\boldsymbol{C}^u \in \mathbb{R}^{n(n+1)/2}, \text{ such that } \boldsymbol{C}(\boldsymbol{C}^u) = \boldsymbol{V}^{\frac{1}{2}} \boldsymbol{P} \boldsymbol{V}^{\frac{1}{2}}, \text{ with } \boldsymbol{P}^{sl} \in \mathcal{M}_p \text{ and } \boldsymbol{V}^d \in \mathcal{V}_\epsilon^+\}, \tag{B5}$$

where $\boldsymbol{V}(\boldsymbol{V}^d)$ is a diagonal matrix of variances $\boldsymbol{V}^d$, restricted to the domain $\mathcal{V}_\epsilon^+ = \{\boldsymbol{V}^d \in [l_V\gamma, l_V/\epsilon]^n\}$ with l_V finite and positive. The symmetric positive-definite correlation matrix, $\boldsymbol{P}(\boldsymbol{P}^{sl})$, has diagonal elements $P_{ii} = 1$ and elements $\boldsymbol{P}^{sl}$ strictly below the diagonal. The set $\mathcal{M}_p = \mathcal{M}$ is defined in the same way as Eqn (B3), but for $\boldsymbol{P}^{sl}$ rather than $\boldsymbol{M}^{sl}$. The restricted parameter space is derived from the Cartesian product $\Theta_\epsilon = \mathcal{U}_\epsilon \times \mathcal{F}_\epsilon \times \mathcal{A}_\epsilon^+ \times \mathcal{C}_\epsilon^+$.

MS received 23 March 2015 and accepted in revised form 25 June 2015