Dear Dr. Whitehouse,

We would like to thank you and both reviewers for your attention to our manuscript. We appreciate the interest of the reviewers in our study, their constructive comments and their help in improving the quality of the paper. We have provided detailed individual responses to each reviewer and we first highlight here the major modifications in the revised manuscript with respect to the first version:

The definition of our research theme

We clarify that this study is focused on the methodology of implementing a Bayesian calibration method for firn models. Furthermore, we aim at demonstrating how our findings of uncertainty related to firn model parameterisation affect firn model output. In this way, we try to show more specifically the direct impact of our results for the broader community of firn model users. We also exploit more the Bayesian framework of our study in order to illustrate how this method can be of practical use to other modellers of the glaciology community. We emphasize that this study is not an estimation of uncertainty in ice sheet wide mass balance estimates from altimetry.

For these purposes, we made the following modifications:

a) We added a section discussing both intra- and inter-model spread on firn model output. The spread in results are computed by exploiting posterior ensembles of parameters. The firn model outputs investigated are compaction rates and the age of firn at the firn-ice transition because both these variables are of important interest to firn model users.

b) We removed the paragraph approximating the impact of our results in terms of compaction rates if they were to be upscaled at the scale of the Greenland ice sheet. These calculations were too speculative, and a thorough uncertainty analysis would require taking many other aspects into account, which is beyond the scope of this study. We subsequently modified the abstract to remove any mention of ice sheet wide estimations.

The boundary forcing

As highlighted by the reviewers, it was important to account for uncertain forcing of firn models: the climatic input and the surface density conditions. In this revised version of our study, we allow realistic magnitudes of these uncertainties to propagate into the model calibration process and thus, to influence the parameter estimates and their credible intervals. This was implemented by adding random perturbations in these fields. The climatic random perturbations are based on typical values of RACMO2 errors and biases with respect to weather station measurements. We provide all the details of the implementation of the random perturbations in the updated manuscript and its Supplementary Information. We note however that this study is not a complete sensitivity analysis to climatic conditions and/or to fresh snow density. Our goal here is to let reasonable estimates of errors in those fields to be accounted for in the calibration process.

Interpretability of our results

As mentioned above, we discuss more thoroughly intra- and inter-model spread for variables of interest to the firn science community. We believe that the coefficients of variation, expressed in %, provide a straightforward overview to the reader of the uncertainty one can expect in firn model output. In general, we focus more on the implications of our findings in terms of uncertainty than on the specific comparison between performances of the MAP and original models. We also facilitated the understanding of the multi-dimensionality of the calibration method for the reader. As suggested by Reviewer 2, Figure 3 has been changed to explicitly demonstrate correlations between different parameters.

Many additional modifications have been integrated to the manuscript and all of these are detailed and explained in our responses. Responses to the reviewers and a marked-up version of the updated manuscript are provided below.

Best regards,
Vincent Verjans, on behalf of authors

Response to Reviewer 1

We thank the reviewer for their effort in reading and evaluating our study. Their comments were accounted for and we believe that our subsequent modifications have improved our study. Our response to their comments is provided below. The original text from the referee is in black italic and our responses are in blue. Throughout the response, we refer the reviewer to the revised manuscript for evaluating the modifications to the study. A marked-up copy of the revised manuscript is attached below (p.9 and below).

The authors point out that "Results of the calibration would depend on the particular climate model used for forcing" (Line 4, Page 4). This important point is mentioned in Data and Methods, but not in Discussion and Conclusions. From reading the latter two sections, I gained the impression that the authors suggest replacing the original parameters with the MAP parameters (with the exception of the LZ model). However, could the difference in parameter values also result from different model forcing? Which parameters would the Bayesian calibration provide as output if you would use, for example, the climate conditions assumed by Herron and Langway (1980)? I understand that this is difficult to quantify, but I suggest at least to highlight this in Discussion and Conclusions or to test the stability of the calibration under different forcing.

The close link between firn model parameters and climatic forcing is a challenge for firn modelling and the dependence of the parameterisation to the forcing is unavoidable. In order to try and account for this, we introduce uncertainty in the climatic forcing in the calibration process by applying random perturbations to both the temperature and accumulation fields of RACMO2. These perturbations are based on published errors and biases of RACMO2 with respect to weather station data. We explain this in Section 2.2 (p.11 l.24) and provide all the technical details of the implementation in the Supplementary Information (Section S2). We hope that this addresses the issue of *"stability of the calibration"*. This development leads to larger uncertainty ranges of the parameters, as we expected. The intervals now incorporate ranges that one can expect to reach using any realistic climatic forcing. We believe that this demonstrates even further the benefits of the Bayesian approach: the interesting outcome of a calibration is not only the best-fit parameter set (i.e. the MAP), but also the robust uncertainty ranges. Accounting for this uncertainty is crucial for any firn model application. As suggested by the reviewer, we add an extra sentence in the Conclusion (p.18 l.5) to raise the awareness of the readers to this issue. Please note that we introduce random perturbations in surface density also, since this is another uncertain boundary condition for firn densification models.

I appreciate that the authors investigate the impact of the new calibration on a Greenland scale. Nevertheless, I feel the comparison could be improved by showing the numbers in the context of total Greenland mass change. Furthermore, the three models are designed to simulate dry firn compaction, while the sensitivity analysis extends to the entire accumulation area. A very substantial part of the Greenland accumulation area is subject to melt and refreezing. How does this influence the informative value of your sensitivity analysis?

In hindsight, we agree with the reviewer that this section was a little speculative. A thorough sensitivity analysis would require taking many other aspects into account, including the *"melt and refreezing"*. Such ice sheet scale sensitivity analysis is beyond the scope of our study and we have removed this section of the Discussion. Instead, we show in a simple example how the Bayesian uncertainty assessment translates into uncertainty in model outputs (p.16 l.49). We use two metrics that are of interest to the broad firn science community: the compaction rate and the age of firn at pore close-off depth. This replacement renders our Discussion section less speculative, more embedded in the Bayesian framework and clearer about the implications of our findings for the wider glaciology community.

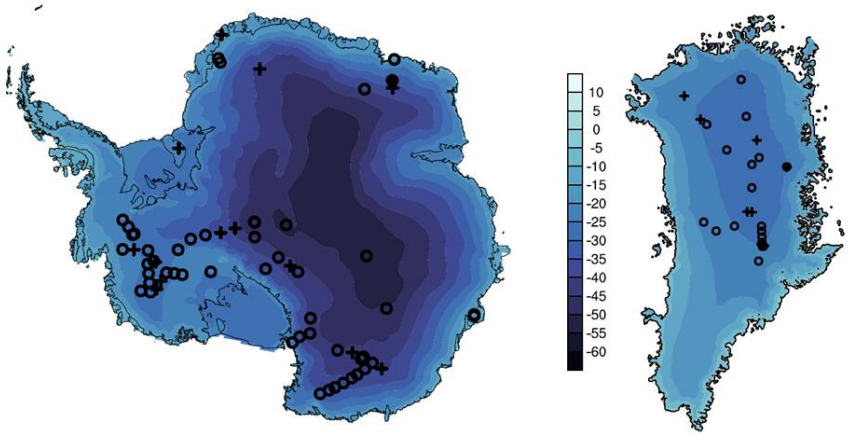
Figures: I appreciate the good quality of the figures. My only suggestion is to use the same colour scale for Greenland and Antarctica in Figure 1. As it is now, and being fully aware that the two colour bars are different, it is difficult to anticipate the differences in climate at the core locations. For both maps, why does the colour bar represent a temperature range that exceed the actual range in climate conditions?

We thank the review for his comment on the quality of our figures. Following their suggestions, we have tried many possible alternatives for Figure 1. Using the same colour scale for both ice sheets resulted in a very red-looking Greenland. Thus, we tried using only a blue colour scale to avoid this problem. The result was not satisfactory (see the

figure at the end of the document). We have thus decided to keep the original map. However, the colour bars are modified as the reviewer suggested, and we have added a statement "Note the different colour bars" in the caption.

References: I had only a brief look at the references, but noticed that the bibliography for Shepherd et al., Science, 2012, might contain some errors. Looking up the article, I found for example a different DOI ("10.1126/science.1228102" instead of "5b0143 [pii]").

We reviewed our list of references. Please note that the citation to the study of Shepherd et al. (2012) has been replaced by a more recent altimetry-based assessment of Antarctic mass balance: Shepherd et al. (2019).



Response to Reviewer 2

We thank the reviewer for the comments provided and for their effort in bringing forward suggestions for improving our study. Based on their comments, we proceeded to several adjustments and we believe that these improve the quality and robustness of our methodology. Our response to their comments is provided below. The original text from the referee is in black italic and our responses are in blue. Throughout the response, we refer the reviewer to the revised manuscript for evaluating the modifications to the study. A marked-up copy of the revised manuscript is attached below (p.9 and below).

The paper aims to evaluate and improve the parametrisations of 3 firn densification models, for the dry snow part. It is written in a concise and clear matter, but I find that the scope of the paper is too narrow. The authors use a sound method, and find weakly different results from the original publications for 2 of the 3 models. They make a good effort in discussing the implications of their findings in terms of densification physics, but still, to my mind, they consider only part of the problem, and it is difficult to make use of their findings without the rest of the picture.

We would like to emphasize that this manuscript is mainly concerned with presenting an advanced methodology which we hope will contribute to a more rigorous use of models in firn science and glaciology in general. Following the advice of the reviewer, we aimed at clarifying the scope of our study in the manuscript. We have also tried to better demonstrate how our findings, and a Bayesian approach to model calibration in general, can contribute to the research of the broader glaciology community. Throughout this response, we try to demonstrate how the updated version of the manuscript fulfils these aims.

Away from meteorological stations, surface temperature and accumulation are not well known. Even if RACMO is as good as it gets, there are biases that can be several °C in temperature. Simply assuming that RACMO is right is not acceptable. A range of scenarios, based on the known biases of RACMO, or uncertainty derived from also using MAR, and different reanalyses is necessary.

We agree that the first version of the study was not sufficiently addressing the problem of the dependency of parameter calibration to climatic forcing. In order to better take this into account, we modify our calibration process. We introduce random perturbations to both the temperature and accumulation fields from RACMO2. These perturbations are based on published errors and biases of RACMO2 with respect to weather station data. We explain this in Section 2.2 (p.11 l.24) and provide all the technical details of the implementation in the Supplementary Information (Section S2). With these random perturbations, the credible intervals of our parameter values have become larger as we expected. The intervals now incorporate ranges that one can expect to reach using any realistic climatic forcing. We believe that this demonstrates even further the benefits of the Bayesian approach: the interesting outcome of a calibration is not only the best-fit parameter set (i.e. the MAP), but also the uncertainty ranges.
We decided to follow the first suggestion of the reviewer and not use other forcing products such as MAR or reanalyses products. Our point of view is that any study making use of firn models should be aware of their empirical nature. Ideally, a model calibration to the specific forcing used should be performed before any application of firn models. The goal of this study is not to provide a specific "best fit" parameter combination for each one of the numerous climatic products that are commonly used (RACMO, MAR, HIRHAM, CESM, NHM-SMAP, ERA, NCEP, MERRA, etc.). Therefore, we have decided to follow the approach of introducing the observation-based random climatic perturbations. We consider that the credible intervals obtained are representative of the range of parameter values that would be applicable using any of these forcing products.
As the reviewer points out, one cannot *"assume that RACMO is right"*. But firn models are generally used coupled to climatic models, and using these climatic models for calibrating firn model parameters is thus sensible. The addition of random noise aims at making the parameterisation less model-specific. We add an extra sentence in the Conclusion (p.18 l.5) in order to clarify this.

In many cases, you can find a mean accumulation derived from the ice core the density was measured on, why not use that?

Our main reason for not using accumulation rates derived from the ice core is that we use a dynamic climatic forcing. The advantage of accumulation rates produced by a climate model is that we have time-varying values and not only a

long-term mean or annual values. In several applications of firn models (e.g. correction of altimetry measurements and surface mass balance modelling), it is crucial that the model outputs remain as correct as possible throughout the year. However, monthly accumulation rates can be strongly different from long-term or annual means. A second reason for our approach is mentioned in the response above: firn models are mostly used forced with climate model products. Thus, calibrating the parameters with these products is a reasonable and pragmatic approach, although climate models include unavoidable errors.

1.2 surface density. I support the idea to use measured density, but measuring surface density is not easy. It comes with an uncertainty of 20% at least. You need to do a sensitivity study of your whole process with randomy different surface density within a reasonable uncertainty range. Surface density is actually a pretty sensitive parameter.

We followed the advice of the reviewer to introduce random noise in the surface density boundary condition. This is also added in Section 2.2 (p.11 l.35) and explained in greater details in the Supplementary Information (section S2). Our point of view is similar to that with respect to the climatic boundary condition: the goal of this study is not to proceed to a thorough sensitivity analysis of firn model output to surface density. But the reviewer is correct that uncertainty in surface density must propagate in the calibration process and must ultimately be translated into uncertainty in firn model output. We believe that the addition of random noise fulfils this aim.

Assumptions that go into the model, like the steady state assumption, and the 1D assumption. In some areas, we know these assumptions are not right. In your outliers to the models, when it is an outlier to all models, did you consider that maybe some of these assumptions were not right?

We agree with the reviewer that many assumptions inherent to current firn models are not physically realistic. Our goal is to investigate the parameterisation of existing, and commonly used, firn models. We have not developed a new firn model. We hope that future developments to firn models will make them closer to the physical processes underlying firn densification. Please note that we highlight these structural shortcomings of firn models in the conclusion (p.18 l.7).

the physical formulation : you discussed this thing well, but did not combine all your model output to give the global uncertainty, or compare the within-model to the acrossmodels uncertainty. For instance, it would be interesting to see a fouth row on figure 5 with the outputs of the 3 models on the same figure, and comment why the uncertainty ranges don't overlap. I am surprised by that, and it makes me think that your 95% confidence interval is underestimated.

In order to better evaluate both intra-model and inter-model uncertainties, we have added an analysis of variability in firn model output in the Discussion (p.16 l.49). We focus on compaction anomalies and firn age at the pore close-off depth because these are firn model outputs of common interest. The parameter-related (i.e. intra-model) uncertainties can be directly compared with the combined parameter- and model-related (i.e. inter-model) uncertainty. This analysis is based on the standard deviations obtained from the posterior ensembles of parameter combinations and thus further exploits the benefits of our Bayesian approach. It highlights that uncertainties are larger when accounting for inter-model spread, especially for compaction rates.
Concerning Figure 5, it is noticeable that the uncertainty ranges have expanded due to the addition of climatic noise. The uncertainty ranges of the three models now clearly overlap. We think that the reader can compare the ranges of the three models because the observation profile, the x-axis and the y-axis are the same for the three models at a given site.

My major problem with the method you used is that you assumed that your parameters were independent, when in reality they are not at all, as shown the covariance matrix on Fig S4 (e.g. k1 and E1 having a covariance of 0.94). You don't show your prior error covariance matrix, so it's difficult to assess exactly what you did. A better description in the prior assumptions, including a comparison of the prior and posterior probability distributions (for instance adding the prior to Fig 3) is needed. From what I read, you included no covariance in your parameters in your prior, and that is not right. For instance, in fig 3, showing k1 and E1 independently makes no sense, because you could always compensate any error in E1 by a change in k1*, it would make more sense to show these either in a 2D plot, or for a fixed E1 in the case of k1 and vice versa.*

Following the comments of the reviewer, we have modified the prior distributions of pairs of parameters for which we have an a priori knowledge of some correlation structure. The pairs of parameters are (k_0^*, E_0), (k_1^*, E_1), (k_0^{Ar}, E_g) and (k_1^{Ar}, E_g). We provide this information in Section 2.4 (p.12 l.40) and in the caption of Table 1. We provide all the technical details of the estimation of the prior correlations in the Supplementary Information (section S3).
It is important to keep in mind that prior independence does not imply posterior independence. If any correlations are plausible given the data, posterior distributions can show, potentially very strong, correlations. This is illustrated in the posterior correlations (Figure 3) and covariance matrices (Figure S4 and Table S2). Please note also that we inform the reader about this in the main manuscript in Section 2.4 (p.13 l.2) and that the updated Figure 3 will make this clearer. Although we now mention the importance of posterior correlations in the main manuscript (p.14 l.29 and p.16 l.15), we have preferred to discuss the details of the posterior correlations only in the Supplementary Information (section S7). We think that this discussion is of interest to firn model developers but probably not to the majority of the firn model users. We hope that discussing more our prior distributions at (p.12 l.40), at (p.13 l.19), at (p.14 l.29) in the caption of Table 1 and in the Supplementary Information (section S3) will make our approach more understandable for the reader.

As suggested, we have changed Figure 3 to two-dimensional graphs. We believe that this will help the reader to visualize the multi-dimensionality of the model calibration. Note also that we now explicitly state that the posterior distributions are characterised by some correlation features (p.14 l.29). Figure 3 still provides marginal posterior distributions in parameters but we removed the limits of the 95% credible intervals to avoid overloading the graph (these are still available in Table 1 and can be compared with the weakly informative prior distributions also given in Table 1).

5.1 Data. Here, I'm with you, I don't want to go through checking each dataset again, but you should at least give some information about these data, citing the spencer paper, and maybe a few other, to illustrate what is the uncertainty in the data , given different methods (weighting a full core is not quite the same as doing gamma, or CT..).

We modify the manuscript in order to refer the reader to our Data Availability section and to our Supplementary Information as soon as we introduce the dataset we use (p.10 l.32). We also add a paragraph dedicated to measurement uncertainty for firn depth-density profiles in Section 2.1 (p.11 l.9). In this paragraph, we refer to several studies that investigate the magnitude of measurement uncertainty.
It is important to highlight that we selected these 91 cores carefully. It was a meticulous effort to look at every profile individually and ruling it as acceptable or not. Unfortunately, even the most recent density measurements do not come with any uncertainty estimation. Thus, the selection relied on our appreciation of the quality of the data. The SUMup dataset has more than 1000 cores and simply taking all these cores was not possible (even though it would have been an easy solution). We also rejected more than half of the Spencer dataset, and we explored other data sources. We know that the selection criteria "accepted by the authors" is not ideal. However, we believe that this was the best solution given the firn core data available.
In our approach, the variances of the likelihood function can mitigate consequences of measurement errors on the calibration process. Also, using DIP as a metric reduces the impact of single measurement errors. Since DIP quantifies a smoothed depth-density profile, individual errors in point measurements of density in depth should (1) have a minor impact on DIP and (2) average out. Also, the use of 69 cores in the calibration should mitigate the effect of a bias in anyone of the 69 cores. Finally, the percentages of DIP used for the calculations of variances are consistent (and even conservative) with respect to the uncertainty estimates of the studies referenced at (p.11 l.10).

5.2 Metric. Is DIP really the best metric? Is it faithful? You are the first to use this metric for the calibration of firn air models. Everyone else was using rho(z) directly. Does it give the same answer? You should demonstrate that. Is it a good metric also for other applications of firn modelign, such as the close-off depth estimation? A paragraph demonstrating the usefullness and validity of this metric would be nice.

We have added a paragraph to discuss our choice of the DIP metric in Section 2.1 (p.10 l.49). In this paragraph, we highlight that DIP is a metric commonly used by firn modellers and by the firn science community in general. Using all individual $\rho(z)$ measurements has several drawbacks: a much stronger sensitivity to individual errors and the fact that the number of $\rho(z)$ values per core is very variable and dependent on the measurement technique. The

approach of Herron and Langway (1980) had been considered. They asserted that the slope of $ln\left(\frac{\rho}{\rho_i-\rho}\right)$ is linear in depth for both stages and thus tuned the modelled slopes to the observed slopes. However, by looking at these slopes for the firn cores, we estimated that this is a very crude approximation. Other authors have calibrated with respect to the depth at which the firn reaches 550 and 830 kg m^{-3} density. This has the drawback that it does not capture the shape of the depth-density profile. Another solution could have been to smooth the $\rho(z)$ profiles (which is the approach of Spencer et al. (2001) and Morris (2018)). By integrating the porosity in depth, DIP is a metric comparable to this approach. We hope that these points and the paragraph added to the manuscript address the concerns of the reviewer with respect to the DIP metric.

The general objective of the paper is, if I quote the abstract to "demonstrate how modeland parameter-related uncertainties potentially affect ice sheet mass balance assessments". It's a great idea, and such a thorough assessment of firn models has not been done. Lundin et al. 2017 highlighted some model deficiencies, but did nothing to remedy those. I support a future version of this paper to be published with a quantitative answer to this question, but we are not there yet.

The reviewer pointed out in this remark that some statements in the first version of the paper were inadequate with respect to its general objective. This study is focused on the methodology of a Bayesian calibration process applied to firn models. Furthermore, we show how uncertainties derived from this method have potential impacts for the broader firn science community. In order to clarify this, we remove the statement "how model- and parameter-related uncertainties potentially affect ice sheet mass balance assessments" from the abstract. For the same reasons, we also removed the section about a Greenland-wide estimate of firn compaction uncertainty from the Discussion. We deemed it too speculative given that many other factors must be taken into account for such an estimation (most notably the impact of melting outside the dry snow zone, climatic gradients in topographically complex areas, geographical patterns of surface density and satellite measurements accuracy). While fully evaluating uncertainty in ice sheet mass balance assessments is beyond the scope of our study, we strongly believe that our work is an important step towards this objective. We are currently working in this direction and we hope to submit an ice-sheet wide uncertainty study in the future, however we agree with the reviewer that "we are not there yet".

As mentioned above, the section removed from the Discussion has been replaced by another section (p.16 l.49). The latter aims at demonstrating the usefulness of Bayesian posterior uncertainty evaluations in terms of (1) compaction rates and (2) age of firn at pore close-off.

If I dont want to run the community firn model (CFM) 30,000 times but just a few, to get a gist of the uncertainty in my specific application, should I use a wide range of (T, accum) scenarios (step 1)? or rather use different model physics (step 3)? Or one of the models, but with a range of parameters (step 4)? This is a practical question that would be really useful to future users of the community firn model.

The modifications to the manuscript offer two different ways to address the practical problem of running 30 000 simulations for assessments of parameter-related and model-related uncertainty (steps 3 and 4).
The first, and more straightforward, approach is to use the coefficients of variation we give (p.17 l.10, p.17 l.13 and Table 3) when discussing uncertainty in computed compaction rates and ages of firn at pore-close off. These figures provide an approximation of the typical uncertainty one can expect due to uncertainty in parameter values and in choice of model. We believe that the coefficients of variations, expressed in%, provide a straightforward overview to the reader of the uncertainty one can expect in firn model output.
The second approach requires slightly more work and more computational effort, but it is more rigorous. We introduce the notion of the normal approximation to the posterior densities in Section 2.4 (p.14 l.9). The normal approximation to the posterior is a practical and commonly used solution in such circumstances (see Gelman et al. (2013)). Any firn-model user can sample a high number of parameter combinations from these normal distributions. Firn model output based on this sample can subsequently be used to build uncertainty intervals for any application. All the details and information required for generating samples of parameter combinations are provided in the Supplementary Information (section S6). The question of how large the generated sample should be cannot be answered unequivocally. Statistical theory tells that the uncertainty intervals (e.g. at 95% precision level or any other) will converge to the true intervals as the sample size increases. We used 500 samples in the study but it was clear that at most sites, the uncertainty intervals were very close to those reached with much fewer samples. One can start

computing uncertainty intervals with a certain number of samples (e.g. 50) and evaluate how intervals change with larger number of samples. A good rule of thumb is that an optimal sample sized is reached once uncertainty intervals remain stable with a further growth in the sample size.
Concerning the climate-related uncertainty (step 1), we believe that this is beyond the scope of this study. We have integrated climatic noise, which allows propagation of climatic uncertainty in the estimation of uncertainty intervals of parameter values. As such, the range of parameter values used in computing the coefficients of variations and characterising the posterior distributions incorporates parameter values influenced by climatic noise. Estimating the plausible range of climatic forcing at any given location in Greenland or Antarctica must be addressed by studies dedicated to regional climate models. Interestingly, many intercomparisons of climate models have been submitted and/or published recently.

And finally, when you have done that, it would be great to go back and recalculate the uncertainty in mass balance from altimetry, using the above mentionned decomposition. This is the great paper I'd like to read. It's marginally more work from what you have done, running a few more simulations on the same framework, but I think it would be really worth it.

We discuss this aspect in our response above. We believe that this study is a step towards this direction and our long-term objective is indeed to "*recalculate the uncertainty in mass balance from altimetry*". However, a thorough calculation requires taking many other aspects into account and goes much further than the methodology we present in this study. We work thoroughly on this long-term objective and we welcome the interest of the reviewer in this subject.

Method: Why did you go for Monte-Carlo, rather than a form of generalized least squares, which would have converged in <10 runs very likely, and could have easily dealt with covariance in model parameters? I agree that this problem is non linear, and underdetermined, but it is also monotonic, so least squares are applicable, and converge much faster. That being said, your method is valid.

The reviewer is correct that multivariate least squares regression would have required less model runs and would probably have reached close "best fit" parameter estimates. We have decided to follow the Bayesian MCMC approach for two reasons. First and foremost, it is much more powerful to assess uncertainty in parameter values given the available data. This point is of major importance to our study. Secondly, it provides a robust framework for future developments. In the future, it will be desirable to integrate other type of data to firn model calibration (e.g. strain rates and radar reflection if such data become publicly available) and to apply such a calibration method to models of greater complexity which exhibit strongly non-linear behaviour (e.g. ice sheet models, climate models or even firn models that become more complex). Bayesian MCMC are a powerful tool in such circumstances. Finally, there is scope to reduce computational costs as recent algorithmic developments (such as Hamiltonian Monte Carlo) allow faster posterior convergence.

choice of models. Why did you choose the Arthern model as one of your 2 models, rather than the IMAU versions (ligtenberg or kuipers)? We already know that the physical formulation of Arthern is not right (Lundin 2017). You later discuss the IMAU version. I think it would make more sense to publish an optimisation for these rather than the Arthern, which shows no sensitivity to the accumulation rate, something we know is wrong.

We discussed the choice of models in length before carrying out this study. Finally, we favoured to focus on the Arthern model because it is the original formulation of the Ligtenberg and Kuipers Munneke versions, and also in order to demonstrate that its inherent flaws can be corrected through our calibration method. Model sensitivities can be evaluated by derivative calculations. We have added a paragraph in the Discussion section (p.16 l.19) where we show that the over-sensitivity of the Arthern model to accumulation rates is appropriately rectified.

Bayesian calibration of firn densification models

Vincent Verjans[1], Amber A. Leeson[1], Christopher Nemeth[2], C. Max Stevens[3], Peter Kuipers Munneke[4], Brice Noël[4] and Jan Melchior van Wessem[4]

[1] Lancaster Environment Centre, Lancaster University, Lancaster, LA1 4YW, UK.
[2] Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, UK
[3] Department of Earth and Space Sciences, University of Washington, Seattle, WA, USA
[4] Institute for Marine and Atmospheric research Utrecht, Utrecht University, Utrecht, the Netherlands

Correspondence to: Vincent Verjans (v.verjans@lancaster.ac.uk)

Abstract.
Firn densification modelling is key to understanding ice sheet mass balance, ice sheet surface elevation change, and the age difference between ice and the air in enclosed air bubbles. This has resulted in the development of many firn models, all relying to a certain degree on parameter calibration against observed data. We present a novel Bayesian calibration method for these parameters, and apply it to three existing firn models. Using an extensive dataset of firn cores from Greenland and Antarctica, we reach optimal parameter estimates applicable to both ice sheets. We then use these to simulate firn density and evaluate against independent observations. Our simulations show a significant decrease (~~25~~ 24 and ~~55~~56%) in observation-model discrepancy for two models and a smaller increase (~~11~~15%) for the third. As opposed to current methods, the Bayesian framework allows for robust uncertainty analysis related to parameter values. Based on our results, we review some inherent model assumptions and demonstrate how ~~model~~ firn model choice and ~~parameter-related~~ uncertainties in parameter values ~~potentially affect ice sheet mass balance assessments~~cause spread in key model outputs.

1 Introduction

On the Antarctic and Greenland ice sheets (AIS and GrIS), snow falling at the surface progressively compacts into ice, passing through an intermediary stage called firn. The process of firn densification depends on local conditions, primarily the temperature, the melt rate and the snow accumulation rate, and accurate modelling of densification is key to several applications in glaciology. Firstly, variability in firn densification affects altimetry measurements of ice sheet surface elevation changes. Consequently, it is a large contributor of uncertainty in mass balance estimates that rely on a correct conversion from measured volume changes to mass changes (Li and Zwally, 2011; ~~Shepherd et al., 2012;~~ McMillan et al., 2016; Shepherd et al., 2019). Errors in the firn related correction can lead to over- or underestimation of mass changes related to surface processes, and to misinterpreting elevation change signals as changes in mass balance and in ice flow dynamics. Secondly, firn models are used to estimate the partitioning of surface meltwater into runoff off the ice sheet, and refreezing within the firn column, which strongly influences mass loss rates (van den Broeke et al., 2016). Model estimates of current and future surface mass balance of the AIS and GrIS would thus benefit from an improved knowledge of the sensitivity of the densification process to climatic conditions. And finally, the densification rate determines the firn age at which air bubbles ~~become~~ are trapped in the ice matrix. Knowing this age is crucial for precisely linking samples of past atmospheric composition, which are preserved in these bubbles, to paleo-temperature indicators, which come from the water isotopes in the ice (Buizert et al., 2014).

Firn densification has been the subject of numerous modelling studies over the last decades (e.g. Herron and Langway, 1980; Goujon et al., 2003; Helsen et al., 2008; Arthern et al., 2010; Ligtenberg et al., 2011; Simonsen et al., 2013; Morris and Wingham, 2014; Kuipers Munneke et al., 2015). However, there is no consensus on the precise formulation that such models should use. Most models adopt a two-stage densification process with the first stage characterising faster densification for firn with density less than a critical value, and then slower densification in the second stage. The firn-model intercomparison of Lundin et al. (2017) demonstrated that, even for idealised simulations, inter-model disagreements are large in both stages. Firn compaction is driven by the pressure exerted by the overlying firn layers. Dry firn densification depends on numerous microphysical mechanisms acting at the scale of individual grains, such as grain-boundary sliding, vapour transport, dislocation creep and lattice diffusion (Maeno and Ebinuma, 1983; Alley, 1987; Wilkinson, 1988). Deriving formulations closely describing the densification of firn at

the macroscale as a function of these mechanisms is challenging. Consequently, most models rely on simplified governing formulations that are calibrated to ~~match~~ agree with observations. The final model formulations have usually been tuned to data either from AIS (Helsen et al., 2008; Arthern et al., 2010; Ligtenberg et al., 2011) or from GrIS (Simonsen et al., 2013; Morris and Wingham, 2014; Kuipers Munneke et al., 2015), consisting of drilled firn cores from which depth-density profiles are measured. However, the calibration of firn densification rates to firn depth-density profiles requires the assumption of a firn layer in steady state. To overcome this limitation, some models have been calibrated against other type of data such as strain rate measurements (Arthern et al., 2010; Morris and Wingham, 2014) or annual layering detected by radar reflection (Simonsen et al., 2013), but such measurements remain scarce and do not extend to firn at great depths below the surface. Ultimately, firn model calibration is an inverse problem that relies on using observational data to infer parameter values.

In this study, we adopt a Bayesian approach in order to address firn model calibration. This provides a rigorous mathematical framework for estimating distributions of the model parameters (Aster et al., 2005; Berliner et al., 2008). Bayesian inversion has been applied in several glaciological studies, and it has been demonstrated that this methodology improves our ability to constrain poorly known factors such as basal topography (Gudmundsson, 2006; Raymond and Gudmundsson, 2009; Brinkerhoff et al., 2016a), basal friction coefficients (Gudmundsson, 2006; Berliner et al., 2008; Raymond and Gudmundsson, 2009), ice viscosity (Berliner et al., 2008) and the role of the subglacial hydrology systems on ice dynamics (Brinkerhoff et al., 2016b). In the Bayesian framework, model parameters are considered as random variables for which we seek an *a posteriori* probability distribution that captures the probability density over the entire parameter space. This distribution allows not only to identify the most likely parameter combination, but also allows us to set confidence limits on the range of values in each parameter that is statistically reasonable. This enables us to quantify uncertainty in model results, to challenge the assumptions inherent to the model itself and to assess correlation between different parameters. Calculations rely on Bayes' theorem (see Sect. 2.4 and Eq. (7)), but because of the high-dimensional parameter space and the non-linearity of firn models, solutions cannot be computed in closed form. As such, we apply rigorously designed Monte Carlo methods to approximate the target probability distributions efficiently. By exploiting the complementarity between the Bayesian framework and Monte Carlo techniques, we recalibrate three benchmark firn models and improve our understanding of their associated uncertainty.

2 Data and Methods

2.1 Firn densification data

In order to calibrate three firn densification models, we use observations of firn depth-density profiles from 91 firn cores (see Data Availability and Supplementary Information) located in different climatic conditions on both the GrIS (27 cores) and the AIS (64 cores) (Fig. 1). Using cores from both ice sheets is important since we seek parameter sets that are generally-applicable and not location-specific. We only consider dry densification since meltwater refreezing is poorly represented in firn models and wet-firn compaction is absent (Verjans et al., 2019). As such, we select cores from areas with low mean annual melt (<0.006 m w.e. yr^{-1}) but spanning a broad range of annual average temperatures (-55 to -20°C) and accumulation rates (0.02 to 1.06 m w.e. yr^{-1}). For each core, we use the depth-integrated porosity (DIP), also called firn air content. We calculate DIP until 15 m depth ($DIP15$, Eq. (1)). For sufficiently deep measurements, we also calculate $DIPpc$, Eq. (2), taken below 15 m and until pore close-off depth (z_{pc}, where a density of 830 kg m^{-3} is reached). These are the observed quantitative values used for the calibration:

$$DIP15 = \int_0^{15} \frac{\rho_i - \rho}{\rho_i} dz \quad (1)$$

$$DIPpc = \int_{15}^{z_{pc}} \frac{\rho_i - \rho}{\rho_i} dz \quad (2)$$

where z (m) increases downwards, ρ is the density of firn (kg m^{-3}) and ρ_i is the density of ice (917 kg m^{-3}). In Eq. (2), we consider porosity only below 15 m to avoid dependency between $DIP15$ and $DIPpc$. We choose to use both $DIP15$ and $DIPpc$ in order to account for first- and second-stage densification. One of the cores has only a single density measurement above 15 m depth and thus its $DIP15$ value is discarded. We note that 48 cores are too shallow to reach z_{pc} and so cores which do reach this depth provide a stronger constraint to the Bayesian inference method. This is sensible because these deep cores carry information about both stages of the densification process. We use DIP as the evaluation metric for the models because of the crucial role of this variable in both surface mass balance modelling and altimetry-based ice sheet mass balance assessments (Ligtenberg et al., 2014). We note that it is

commonly used in firn model intercomparison exercises (Lundin et al., 2017; Stevens et al., 2020) and is a quantity of interest for field measurements (Vandecrux et al., 2018). Due to its formulation (Eq. (1) and (2)), DIP represents the mean depth-density profile and thus is robust to the presence of individual errors and outliers in density measurements.

We separate the dataset into calibration data (69 cores) and independent evaluation data (22 cores). The latter is selected semi-randomly; we ensure that it includes a representative ratio of GrIS-AIS cores and that it covers all climatic conditions, including an outlier of the dataset with high accumulation and temperature (see Supplementary Information). The resulting evaluation data has 8 GrIS and 14 AIS cores; 11 of the 22 cores extend to z_{pc}.

Observed firn density can be prone to measurement uncertainty, which previous studies point out is about 10%, though it is variable in depth and between measurement techniques employed (Hawley et al., 2008; Conger and McClung, 2009; Proksch et al., 2016). We outline our procedure to account for measurement uncertainty in Sect. 2.4.

2.2 Climate model forcing

At the location of each core, we simulate firn densification under climatic forcing provided by the RACMO2.3p2 regional climate model (RACMO2 hereafter) at 5.5 km horizontal resolution for GrIS (Noël et al., 2019) and 27 km for AIS (van Wessem et al., 2018). ~~Results of the calibration would depend on the particular climate model used for forcing.~~ Each firn model simulation consists of a spin-up by repeating a reference climate until reaching a firn column in equilibrium, which is followed by a transient period until the core-specific date of drilling. The reference climate is taken as the first 20-year period of RACMO2 forcing data (1960-1979 and 1979-1998 for GrIS and AIS respectively). The number of iterations over the reference period depends on the site-specific accumulation rate and mass of the firn column (mass from surface down to z_{pc}). We ensure that the entire firn column is refreshed during the spin-up but fix the minimum and maximum number of iterations to 10 (200 years spin-up) and 50 (1000 years spin-up). We note that at 33 sites, the core was drilled before the last year of the reference climate and so the transient period is effectively a partial iteration of the spin-up period.

Results of the calibration would depend on the particular climate model used for forcing. We thus propagate uncertainty in modelled climatic conditions into our calibration of firn model parameters by perturbing the temperature and accumulation rates of RACMO2 with normally distributed random noise. Standard deviations of the random perturbations are based on reported errors of RACMO2 (Noël et al., 2019; van Wessem et al., 2018 – see more details in the Supplementary Information). By introducing these perturbations, uncertainty intervals on our parameter values encompass the range of values that would result from using other model-based or observational climatic input.

In addition to the climatic forcing, another surface boundary condition is the fresh snow density, ρ_0. It is taken as a constant site-specific value. Each value is taken in agreement with the shallow densities measured in the corresponding core of the dataset. We prefer this approach to the use of available parameterisations of ρ_0 (Helsen et al., 2008; Kuipers Munneke et al., 2015) to avoid any error in the fresh snow parameterisation to affect the calibration process. Fresh snow density is a poorly constrained boundary condition (e.g. Fausto et al., 2018). We account for uncertainty in this parameter by adding normally distributed random noise with standard deviation 25 kg m^{-3} to ρ_0 at every model time step.

2.3 Firn densification models

We use the Community Firn Model (Stevens et al., ~~2018~~2020) as the framework of our study because it incorporates the formulations of all three densification models investigated: HL (Herron and Langway, 1980), Ar (Arthern et al., 2010) and LZ (Li and Zwally, 2011). The Robin hypothesis (Robin, 1958) constitutes the fundamental assumption of HL, Ar and LZ. It states that any fractional decrease of the firn porosity, $\frac{\rho_i-\rho}{\rho_i}$, ~~$\rho/(\rho-\rho_i)$,~~ is proportional to an increment in overburden stress. This translates into densification rates depending on a rate coefficient c, assumed different for stage-1 and stage-2 densification:

$$\begin{cases} \frac{d\rho}{dt} = c_0\,(\rho_i - \rho), & \rho \le 550\ kg\ m^{-3} \\ \frac{d\rho}{dt} = c_1\,(\rho_i - \rho), & \rho > 550\ kg\ m^{-3} \end{cases} \quad (3)$$

The formulations of the rate coefficients rely on calibration and thus differ between the three models investigated:

HL

$$\begin{cases} c_0 = \dot{b}^a k_0^* \exp\left(\frac{-E_0}{RT}\right) \\ c_1 = \dot{b}^b k_1^* \exp\left(\frac{-E_1}{RT}\right) \end{cases} \quad (4)$$

Ar

$$\begin{cases} c_0 = \rho_w \dot{b}^\alpha k_0^{Ar} g \exp\left(\frac{-E_c}{RT} + \frac{E_g}{RT_{av}}\right) \\ c_1 = \rho_w \dot{b}^\beta k_1^{Ar} g \exp\left(\frac{-E_c}{RT} + \frac{E_g}{RT_{av}}\right) \end{cases} \quad (5)$$

LZ

$$\begin{cases} c_0 = \beta_0 lz_a (273.15 - T)^{lz_b} \dot{b} \\ c_1 = \beta_1 lz_a (273.15 - T)^{lz_b} \dot{b} \end{cases} \quad (6)$$

$$with \begin{cases} \beta_0 = lz_{11} + lz_{12}\dot{b} + lz_{13} T_{av} \\ \beta_1 = \beta_0 \left(lz_{21} + lz_{22}\dot{b} + lz_{23} T_{av}\right)^{-1} \end{cases}$$

where $\dot{b}$ is the accumulation rate (m w.e. yr^{-1}), T the temperature (K), T_{av} the annual mean temperature, R the gas constant, g gravity and ρ_w the water density (1000 kg m^{-3}). All remaining terms are model-specific tuning parameters. For $\dot{b}$, we use the mean accumulation rate over the lifetime of each specific firn layer because it better approximates the overburden stress than the annual mean (Li and Zwally, 2011). HL and Ar use Arrhenius relationships with activation energies (E terms) capturing temperature sensitivity and exponents characterising the exponential proportionality of the rate coefficients to the accumulation rate. Originally, Herron and Langway (1980) inferred all values from calibration based on 17 firn cores, from which they inferred the values for the six free parameters (Table 1) of HL. In contrast, Arthern et al. (2010) fixed the accumulation exponents in advance ($\alpha = \beta = 1$) and took activation energies (E_c, E_g) from measurements of microscale mechanisms: Nabarro-Herring creep for E_c and grain-growth for E_g. Still, they noted a mismatch with the activation energy fitting their data best. The k_0^{Ar} and k_1^{Ar} parameters were tuned to three measured time series of strain rates collected in relatively warm and high accumulation locations of AIS. Here, we consider all five $\alpha, \beta, k_0^{Ar}, k_1^{Ar}, E_g$ as free parameters (Table 1) but keep E_c fixed because of its strong correlation with E_g; our use of monthly model time steps and depth-density profiles as calibration data is not suitable for differentiating effects of $\frac{E_g}{RT_{av}}$ and $\frac{E_c}{RT}$. Equation (6) shows that LZ has eight free parameters (Table 1), all denoted by lz in this paper. In contrast to our approach to Ar, we do not add additional accumulation rate exponents to $\dot{b}$ in Eq. (6) because the dependence of c_0 and c_1 on $\dot{b}$ also involves the coefficients lz_{12} and lz_{22} in the definition of β_0 and β_1. Li and Zwally (2011) performed their calibration of Eq. (6) against firn cores only from the GrIS. Later, Li and Zwally (2015) proposed a new parameterisation for β_0 and β_1, but calibrated for Antarctic firn. Since one of the goals of this study is to find a densification formulation applicable to firn in both GrIS and AIS, we choose to apply our calibration method only to Eq. (6) specified in Li and Zwally (2011). However, in our results' analysis (Sect. 3), we also consider the performance of the Li and Zwally (2015) model on the AIS cores of our dataset.

2.4 Bayesian calibration

In our approach, the free parameters of the firn models are identified as the quantities of interest and we define this parameter set as θ. Hereafter, 'original model values' refers to the values originally attributed by Herron and Langway (1980), Arthern et al. (2010) and Li and Zwally (2011) to their respective sets of free parameters θ. The calibration process relies on Bayes' theorem (Eq. (7)) which allows to update a prior probability distribution $P(\theta)$ for θ based on observed data Y.

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)} \quad (7)$$

We use normal and weakly informative priors centred about the original model values so that the constraint of the prior on $P(\theta|Y)$ is minor (Table 1). As indicated by Morris and Wingham (2014), in HL and Ar, the values of the Arrhenius pre-exponential factors (k_0^*, k_1^*, k_0^{Ar} and k_1^{Ar}) are correlated with their corresponding activation energies (E_0, E_1 and E_g). At a given temperature, a change of the value in the pre-exponential factor can be compensated by adjusting the activation energy to keep the densification rates constant. We express our *a priori* knowledge of these

correlations in the prior distributions (see Supplementary Information). No other pair of parameters in HL, Ar or LZ are clearly correlated *a priori*, but the calibration process captures *a posteriori* correlations by confronting the models to data. The data Y consists of the observed $DIP15$ and $DIPpc$ values of the calibration data. The marginal likelihood, $P(Y)$, is a constant term independent of θ and does not influence the calibration. We use a normal likelihood function $P(Y|\theta)$, which quantifies the match of the modelled DIP values with the observed:

$$P(Y|\theta) \propto \exp\left[\frac{-1}{2}(X_{15}-Y_{15})^T\Sigma_{15}^{-1}(X_{15}-Y_{15}) - \frac{1}{2}\left(X_{pc}-Y_{pc}\right)^T\Sigma_{pc}^{-1}\left(X_{pc}-Y_{pc}\right)\right] \quad (8)$$

where X_{15} and Y_{15} are vectors containing all modelled and observed values for the calibration data of $DIP15$ respectively, and similarly for X_{pc} and Y_{pc}. We use diagonal covariance matrices Σ_{15} and Σ_{pc} with site-specific variances. The variances determine the spread allowed for the model outputs compared to the observed values and ~~we calculate them~~are calculated by taking ~~5~~10% and ~~10~~20% margins around $DIP15$ and $DIPpc$ measurements respectively. Allowing for such spread is necessary because multiple causes may lead to model-observation discrepancy such as firn model errors, measurement uncertainties and~~,~~ ~~potential errors in the climatic forcing~~discrepancies induced by the random perturbations applied to RACMO2 forcing and to ~~ and approximations in fixing~~ ρ_0. This particular form of the likelihood function assumes independence between model errors in $DIP15$ and in $DIPpc$, which is ensured by our calculation of $DIPpc$ only from 15 m depth to z_{pc} (Eq. (2)). It also assumes normally distributed model errors with respect to the observed values. Both these aspects were verified with preliminary assessments, along with our calculations for the covariance matrices Σ_{15} and Σ_{pc}, as discussed in the Supplementary Information. The posterior distribution $P(\theta|Y)$ gives a probability distribution over the parameter space of a given model conditioned on the calibration data. In our case, with weakly informative priors (Table 1), the distribution $P(\theta|Y)$ is essentially governed by the likelihood function (Eq. (8)). We note here that extreme parameter combinations in the LZ model can lead to negative densification rates. In such cases, we set the modelled DIP values to 0, which leads to extremely low values for the likelihood and for the posterior probability of such parameter sets.

There is no analytical form of $P(\theta|Y)$ and we must investigate the parameter space to generate an ensemble of θ_i approximating $P(\theta|Y)$. Such an investigation is achieved efficiently using Markov Chain Monte Carlo methods. We apply the well-known Random Walk Metropolis (RWM) algorithm (Hastings, 1970) and summarize it in Fig. 2, on which we base the brief following description. A given model starts with a certain parameter set θ_i and simulates firn profiles at all the calibration sites. Its $DIP15$ and $DIPpc$ results are compared with observations and the general performance of the model using θ_i is quantified by the likelihood. From there and with the prior distributions assumed, the posterior probability $P(\theta_i|Y)$ is computed following Eq. (7). At this point, the state of the chain is θ_i (Fig. 2a). The RWM then proposes a new θ_i^* from a proposal distribution (Fig. 2b). For the latter, we use the symmetric multivariate normal (MVN) distribution which is centred about θ_i. This implies that the random choice of θ_i^* depends only on the current state θ_i and on the proposal covariance in the MVN, Σ_{prop}, which is discussed below. Using the parameter combination θ_i^*, the model simulates profiles at all calibration sites again (Fig. 2c) and $P(\theta_i^*|Y)$ is computed (Fig. 2d). From there, we either accept or reject the proposed θ_i^* in the ensemble approximating $P(\theta|Y)$. The probability of accepting θ_i^* depends on the ratio $P(\theta_i^*|Y)/P(\theta_i|Y)$ (Fig. 2e). The set saved in the ensemble (Fig. 2g) is θ_i^* if accepted or θ_i if θ_i^* was rejected. The saved set becomes the updated current status for the next iteration θ_{i+1} (Fig. 2a) and the algorithm iterates this process. The RWM has the property that the chain will ultimately converge to a stationary distribution that represents the posterior $P(\theta|Y)$. Thus, after a sufficiently high number of iterations of the algorithm, the ensemble of parameter sets is representative of $P(\theta|Y)$. We verify adequate convergence using a number of tests, which are shown in the Supplementary Information. The proposal variance Σ_{prop} must account for dependence between the different components of θ, i.e. the value of one free parameter can influence the value of another free parameter for the model to reach a good match with the observed data. Σ_{prop} can capture this dependence between parameters and, for optimality, it is updated every given number of iterations (100 in our study) using Eq. (9) (Rosenthal, 2010):

$$\Sigma_{prop} = \frac{2.38^2}{p}\,\Sigma_{cov} \quad (9)$$

where Σ_{cov} is the covariance matrix between the free parameters of the model at this stage of the iterative chain and p is the number of free parameters.

From the posterior probability distributions, we can infer the Maximum a Posteriori (MAP) estimates of each model (MAP_{HL}, MAP_{Ar}, MAP_{LZ}). These are the modes of the multi-dimensional distributions over the space of free

parameters and have been identified as the most likely sets by the RWM. The MAP estimates can be compared to the corresponding original model values of the parameters. The posterior distributions additionally incorporate the uncertainty in the parameter values. By performing posterior predictive simulations on the evaluation data, we can assess this remaining uncertainty (Gelman et al., 2013). More specifically, we can assume that a large (500) random sample of the ensemble of accepted θ is representative of the posterior distribution. As such, model results computed with all sets of this sample inform about model performance accounting for uncertainty. Intuitively, a large spread in results from a 500 random sample would indicate a large range of possible sets for the free parameters and thus a high uncertainty in parameter values.
Since there is no analytical form of our posterior distributions, and to facilitate future firn model uncertainty assessments, we can approximate the posterior distributions with MVN distributions whose means and covariances are set to the posterior means and posterior covariance matrices of the calibration. This allows straightforward sampling of random parameter sets instead of relying on posterior samples of the MCMC. We provide information about the normal approximations and assess their validity in the Supplementary Information. Such normal approximations are asymptotically exact and are commonly applied to analytically intractable Bayesian posterior distributions (Gelman et al., 2013).

3 Results

We present the results of the calibration process after ~~30000~~ 15000 algorithm iterations and compare the MAP and original models' performances against the 22 evaluation cores. We also evaluate the uncertainty of the posterior distributions and compare performances between the different MAP models. All the evaluation simulations are performed without climatic and surface density noise in order to make the evaluation fully deterministic.

For ~~both~~ HL and even more so for Ar, the posterior distributions for the parameters demonstrate some strong disagreements with the original values (Figs. 3a, 3b). The 95% credible intervals for each parameter (Table 1) incorporate 95% of the marginal probability density in the posterior. Two original parameter values of HL (a, b) and three of Ar (E_g, α, β) lie in the tails of the posterior distributions~~outside these intervals, and one of HL (k_0^*) is at the lower edge of the interval~~ (Figs. 3a, 3b) and even outside these intervals in the case of b, E_g, α and β. This indicates that our analysis provides strong evidence against these original values. The strongest disagreements relate to the accumulation exponents of both models (a, b, α, β). In contrast, the original LZ values agree better with the posterior distribution and all lie within the 95% credible intervals (Table 1 and Fig. 3c). The posterior distributions show some strong correlation between certain pairs of parameters (Fig. 3). Notable examples are the pre-exponential factors and their corresponding activation energy in HL and Ar, for which the posterior correlations are even stronger than in the prior distributions. The complete correlation matrices and a detailed analysis of all posterior correlation features are provided in the Supplementary Information.

We use the original models and the MAP estimates to simulate firn profiles at the evaluation sites and we compare DIP results with the observed values. This is an effective way to assess possible improvements in parameter estimates reached through our method since the evaluation sites were not used in the calibration process. The match between observations and the model is improved for MAP_{HL} (Fig. 4a) and even more for MAP_{Ar} (Fig. 4b), with the original Ar strongly underestimating DIP values. These improvements translate into significantly reduced root mean squared errors (RMSE) in modelled values of both $DIP15$ (-~~26~~24% for HL and -45% for Ar) and $DIPpc$ (-22% and -~~60~~61%) (Table 2).
For LZ, ~~the evaluation against the test set is inconclusive, with a worse~~the relative performance of the MAP_{LZ} model for both $DIP15$ and $DIPpc$ is worse (+2% and +~~22~~24% in RMSE) but differences are of smaller magnitude ~~and a slight improvement for $DIP15$ (-1 %)~~ (Table 2 and Fig. 4c). ~~The relative agreement in parameter values between~~Parameter values of MAP_{LZ} and the original LZ are closer, which explains more moderate differences in RMSE compared to HL and Ar. Comparing modelled and observed depth-density profiles of evaluation data illustrates the differences in performance visually (e.g. Fig. 5). Profiles of the original models of HL and Ar frequently lie outside the credible intervals of their respective MAP models. In contrast, profiles of MAP_{LZ} and of the original LZ tend to be close together. At the climatic outlier of our evaluation data (DML in Fig. 5), ~~MAP_{LZ} performs slightly worse than the original LZ (Fig. 5i) but~~ improvements are reached for the three MAP models (Figs. 5c, 5f, 5i). ~~MAP_{HL} and MAP_{Ar} (Figs. 5c, 5f).~~ This demonstrates benefits of this method even at the limits of the calibration range.

Compared to the original HL, MAP_{HL} reaches improvements in $DIP15$ for 12 of the 22 evaluation cores and in $DIPpc$ for 5 of the 11 evaluation cores (Fig. 6a). Generally, MAP_{HL} performs better at AIS sites and worse at GrIS sites. An analysis of the improvement of MAP_{HL} as a function of climatic variables (Fig. 6a) shows that the original HL gives better results in a narrow range of T_{av}: from -30 to -25 °C. As such, the better performance at the GrIS evaluation sites is likely due to the original HL being better suited for the particular temperature range corresponding to the conditions of the latter sites. In contrast, MAP_{HL} seems more appropriate for covering a wider range of climatic conditions. For Ar, the original model shows better performance than MAP_{Ar} at few evaluation sites (~~5~~ 6 for $DIP15$ and 2 for $DIPpc$) which are only in AIS and confined to low-accumulation conditions (Fig. 6b). This is counterintuitive given that Arthern et al. (2010) tuned the original Ar to measurements from high accumulation sites of AIS. Finally, the original LZ performs better than MAP_{LZ} at most GrIS sites (Fig. 6c), which is unsurprising given that its original calibration was GrIS-specific. Again, this seems related to the original LZ performing significantly better in the same narrow range of temperatures as for HL. In total, MAP_{LZ} performs better for ~~11~~ 10 of the 22 $DIP15$ and 4 of the 11 $DIPpc$ evaluation measurements.

As explained in Sect. 2.3, the original LZ model was developed for GrIS firn only (Li and Zwally, 2011) and later complemented by an AIS-specific model (Li and Zwally, 2015). Using both of these on the evaluation sites of their respective calibration ice sheet, we construct an LZ dual model, which thus really consists of two different models. The RMSE for $DIP15$ of LZ dual is slightly ~~lower~~ larger (+8 ~~9~~ %) than that of MAP_{LZ} ~~but~~ and significantly larger (+3~~8~~7 %) for $DIPpc$. We note that the higher RMSE values of LZ dual are strongly affected by its densification scheme performing very poorly at the climatic outlier of the evaluation data, with conditions that are outside of the calibration range of Li and Zwally (2015).

We also compare MAP results with the IMAU firn densification model (IMAU-FDM), which has been used frequently in recent mass balance assessments from altimetry (Pritchard et al., 2012; Babonis et al., 2016; McMillan et al., 2016; Shepherd et al., ~~2018~~2019). IMAU-FDM was developed by adding two tuning parameters to both densification stages of Ar. All four extra-parameters are different for AIS (Ligtenberg et al., 2011) and GrIS (Kuipers Munneke et al., 2015), thus also resulting in two separate models. On the evaluation data, its performance for $DIP15$ is slightly better than MAP_{Ar} and MAP_{LZ} but worse than MAP_{HL}, and its performance for $DIPpc$ is significantly worse than all three MAP models (Table 2).

To assess the uncertainty captured by the Bayesian posterior distributions, we compute results on the evaluation data with the 500 parameter sets randomly selected from each of the three posterior ensembles. For all three models, the average performance of their random sample is similar to the corresponding MAP performance, with a maximum RMSE change of 6% (Table 2). This demonstrates a low uncertainty in the optimal parameter combinations identified by calibration. Furthermore, the best performing 95th percentile of the random selection allows the construction of the uncertainty intervals shown in Figs 4, 5. Of the original models, LZ reaches the lowest RMSE values. Of all models, MAP_{HL} performs best in $DIP15$ and MAP_{Ar} in $DIPpc$ (Table 2). MAP_{LZ} performs worse than the other MAP models even when accounting for uncertainty by using the 500-samples random selections (Table 2). ~~MAP_{HL} performs better in both $DIP15$ and $DIPpc$ than any model tested (Table 2). Comparing the performances of MAP models, MAP_{HL} is followed by MAP_{Ar} and then MAP_{LZ}. This order is still valid for the 500-samples random selections, which account for uncertainty (Table 2).~~

4 Discussion ~~and Implications~~

This calibration method is potentially applicable to models of similar complexity in a broad range of research fields. We exploit it here to investigate the parameter space of HL, Ar and LZ, and to re-estimate optimal parameter values conditioned on observed calibration data; no further complexity is introduced since the number of empirical parameters remains the same. We treat the accumulation exponents of Ar (α, β) as free parameters whereas Arthern et al. (2010) decided to fix their values to 1. Analogous to a and b in HL, these exponents capture the mathematical relationship between densification rates and the accumulation rate, used as a proxy for load increase on any specific firn layer. No physical argument favours a linear proportionality between densification and load increase and any prescribed value for these exponents is a choice of the model designer. Unlike Arthern et al. (2010), Herron and

Langway (1980) previously inferred $a = 1$ and $b = 0.5$. Our calibration data shows strong evidence against both these pairs of values; all four are in the extreme tails of the posterior distributions ~~outside the posterior 95% credible intervals~~ (Fig. 3a, 3b). Our results of stage-1 exponents (a, α) smaller than 1 indicate a weaker increase in densification rates with pressure. In firn, the load is supported at the contact area between the grains, which increases on average due to grain rearrangement (in stage-1) and grain growth. As such, firn strengthens in time and the actual stress on ice grains increases slower than the total load (Anderson and Benson, 1963). Morris and Wingham (2014) incorporated this by including a temperature-history function, causing slower densification of firn previously exposed to higher temperatures. This is consistent with both grain rearrangement and grain growth because these processes are enhanced at higher temperatures (Alley, 1987; Gow et al., 2004). Lower values of the stage-2 exponents (b, β) illustrate the larger strength of high-density firn with larger contact areas between grains. The same can be applied to the LZ model by investigating the posterior correlation between its free parameters. It shows a positive correlation coefficient (0.~~6~~74) between the accumulation-related parameters of both stages; lz_{12} and lz_{22}. High values of lz_{12} make β_0 more sensitive to $\dot{b}$ (Eq. (6)). However, β_0 appears in the numerator of the β_1 calculation (Eq. (6)) and higher values of lz_{22} thus moderate the sensitivity of stage-2 densification to $\dot{b}$. As such, positively correlated lz_{12} and lz_{22} provide further evidence that stage-1 densification rates are more sensitive to accumulation rates. This example demonstrates how posterior correlations provide insights into model behaviour. The posterior correlations of all three models are further discussed in the Supplementary Information.

In the IMAU model introduced in Sect. 3, tuning parameters have been added to Ar in order to reduce its sensitivity to accumulation rates (Ligtenberg et al., 2011; Kuipers Munneke et al., 2015). The calibration method presented in this study detects and adjusts for this over-sensitivity without the need for more tuning parameters in the governing densification equations. The sensitivity of stage-1 densification to $\dot{b}$ can be computed from the derivative of the rate coefficient:

$$\frac{\partial c_0}{\partial \dot{b}} = \rho_w \, k_0^{Ar} \, g \, \exp\left(\frac{-E_c}{RT} + \frac{E_g}{RT_{av}}\right) \alpha \, \dot{b}^{\alpha-1} \quad (10)$$

Similarly, the derivative $\frac{\partial c_1}{\partial \dot{b}}$ is obtained by replacing k_0^{Ar} and α with k_1^{Ar} and β. Our calibration process strongly favours smaller values of α, β and E_g with respect to the original values (Fig. 3b). We can compare the magnitudes of the derivatives under the original Ar parameterisation and under the MAP parameterisation. The magnitudes vary for particular combinations of T_{av} and $\dot{b}$. Under all the annual mean climatic regimes of our dataset, the MAP parameters result in a decreased sensitivity of both stage-1 and stage-2 densification rates to $\dot{b}$.

HL, Ar and LZ only use temperature and accumulation rates as input variables. Other models use additional variables hypothesised to affect densification rates. These include the temperature-history mentioned above (Morris and Wingham, 2014), firn grain size (Arthern et al., 2010), impurity content (Freitag et al., 2013) and a transition region between stage-1 and stage-2 densification (Morris, 2018). Other models are explicitly based on micro-scale deformation mechanisms (Alley, 1987; Arthern and Wingham, 1998; Arnaud et al., 2000). These efforts undoubtedly contribute to progressing towards physically based models. A potential problem with such approaches is overfitting calibration data by adding parameters to model formulations while detailed firn data remain scarce. As long as more firn data is not available to appropriately constrain the role of each variable in model formulations, we favour the use of parsimonious models relying on few input variables. It is noteworthy that MAP_{LZ}, which relies on eight free parameters, performs worse on the evaluation data than MAP_{HL} and MAP_{Ar} with two fewer free parameters. This highlights that gains in model accuracy should rely not only on better calibration of parameters but also on a reconsideration of the governing densification equations. Additionally, firn core data invokes the assumption of a steady-state depth-density profile. As such, parameter calibration poorly captures seasonal climatic effects on densification. Comprehensive datasets of depth-density profiles (Koenig and Montgomery, 2019) are very valuable to model development. Efforts in collecting and publishing strain rate measurements from the field (Hawley and Waddington, 2011; Medley et al., 2015; Morris et al., 2017), and possibly from laboratory experiments (Schleef and Löwe, 2013), can further benefit model calibration and the progress towards more representative equations.

In order to quantify the consequences of our calibration, we investigate two aspects for which firn models are of common use: calculating firn compaction rates and predicting the age of firn at z_{pc} depth, age_{pc} (yr). At every site i of our dataset, we compute the 2000-2017 total compaction anomaly, $cmp_{an,i}$ (m), and the $age_{pc,i}$ value with each of the 500 parameter sets randomly drawn from the posterior ensembles of the three different models (HL, Ar, LZ). This

allows evaluation of both parameter-related and model-related uncertainty. Total compaction anomaly (cmp_{an}) – calculated as the cumulative anomaly in surface elevation change due only to firn compaction changes during the 2000-2017 period with respect to the climatic reference period – is given by:

$$cmp_{an,i} = cmp_{tot,i}^{00-17} - 17cmp_{ref,i}^{yr} \quad (11)$$

where cmp_{tot}^{00-17} (m) is the total firn compaction over 2000-2017 and cmp_{ref}^{yr} (m yr^{-1}) is the annual mean compaction over the reference period. At all sites, we compute the coefficients of variation (CV) for both cmp_{an} and age_{pc} from the 500 simulations with each model, and we average the CVs across all sites. CV is the ratio of the standard deviation to the mean and provides an effective assessment of relative dispersion of model results. Because low mean values of cmp_{an} can inflate its CV, we consider only half of the sites at which the mean computed cmp_{an} is highest. For all three models, the CV values for both cmp_{an} and age_{pc} lie between 5.5 and 7.5% (Table 3). These values give typical uncertainty in firn model output related to uncertain parameter values. Proceeding to the same calculations but using all three models, i.e. an inter-model ensemble of 1500 simulations at each site, gives an overview of the combined parameter- and model-related uncertainty. The CVs are 19.5% for cmp_{an} and 7.5% for age_{pc}, demonstrating larger inter-model disagreement on cmp_{an} calculations (Table 3). In the dry snow zone of GrIS, simulated compaction anomalies are typically around 20 cm over 2000-2017, and thus come with an uncertainty of the order of ±4 cm. Since pore close-off age here is around 250 years, a reasonable uncertainty range on this value is ±19 years. In contrast, on the drier AIS, pore close-off age is about 1000 years thus this range increases to ±75 years. Compaction anomalies hover around 0 cm on most of the dry zone of AIS because it has not experienced the strong recent surface warming of GrIS. Absolute uncertainty is thus reduced but still critical given the large area of AIS over which it must be integrated when mass balance trends are evaluated. Such numbers provide an order of magnitude of errors in firn model outputs that must be accounted for in altimetry-based mass balance assessments and in ice core studies, respectively.

~~As an example of consequences of our calibration, we investigate its effects on GrIS firn thickness change under the 2000-2017 climate. At all 27 GrIS sites of our dataset, we compute the cumulative anomaly in thickness change due to firn compaction during the latter period with respect to the reference 1960-1980 period (see Supplementary Information). Altimetry-based mass balance surveys could interpret any change in firn thickness not captured by firn models as a net gain or loss of ice. The root mean square difference in compaction anomaly between MAP_{HL} and the original HL is 1.5 cm over the 2000-2017 period. If we extrapolate this model-discrepancy to the entire accumulation area of GrIS and convert it to mass, the disagreement between both models corresponds to about 20.6 Gt of ice cumulated over 2000-2017. The same process applied to the original Ar and LZ and their respective MAP yields discrepancies of about 76.6 and 27.4 Gt respectively. Between all six different models, the largest disagreement corresponds to 83.4 Gt, which decreases to 72.1 Gt when considering only MAP models. For reference, we note that the absolute compaction anomaly over the period is equivalent to 870 Gt if averaged across all six models. These discrepancy figures are approximate, since different climatic sensitivities of models and variability in climatic changes will cause compensating effects in different areas. Still, they provide an order of magnitude for potential errors in mass balance assessments that are related to choices of model and of parameter values if firn densification is exposed to a realistic climatic shift.~~

~~In addition to different cumulative responses,~~We further investigate the ~~six models show~~ different sensitivities in terms of monthly values of compaction anomalies over the 2000-2017 period of the original and MAP models of HL, Ar and LZ (Fig. 7). Ar shows the strongest sensitivity to climatic conditions diverging from these of the reference period; compaction responds strongly to the general increases on GrIS in temperature and accumulation rate, especially in late summer. Due to its lower values for α, β and E_g, MAP_{Ar} exhibits less extreme compaction anomalies than the original Ar and thus less seasonal variability. In sharp contrast to Ar, HL-computed compaction rates remain relatively stable, due to low activation energy values that smooth out the seasonal variability. Firn core observations provide little information and constraints on seasonal patterns of densification. However, it is noteworthy that MAP_{Ar} and MAP_{LZ} tend to show comparable short-scale sensitivities (insets in Fig. 7), despite structural differences in the models' governing equations. This might indicate that these models fare relatively well in capturing seasonal fluctuations of densification rates and their sensitivity to climate shifts.

5 Conclusion

We have implemented a Bayesian calibration method to estimate optimal parameter combinations applicable to GrIS and AIS firn for three benchmark firn densification models (HL, Ar, LZ). An extensive dataset of 91 firn cores was separated into calibration and independent evaluation data. Two optimised models (MAP_{HL}, MAP_{Ar}) showed

significant improvement against the evaluation data, while MAP_{LZ} reached results close to, but slightly worse, ~~to~~ than its original version and inferior to MAP_{HL} and MAP_{Ar}. When compared to other models of greater complexity, the MAP models showed comparable or even improved performances~~, especially MAP_{HL}~~. Furthermore, the Bayesian approach provides a robust way to evaluate the uncertainty related to parameter value choice, which is a major deficiency of current models. By introducing realistic climatic perturbations in the calibration process, the uncertainty intervals obtained account for the effects of an uncertain climatic forcing. However, at most sites where we evaluated, all three models' uncertainty intervals do not cover observed *DIP* values. As such, although model results can be improved by re-calibration methods, model tuning alone is insufficient to reach exact fidelity of firn densification models. The formulation of models' governing equations impacts the remaining errors with respect to observations, which highlights deficiencies in our understanding of dry firn densification. Developing a well-constrained physically detailed model is challenging given the number of mechanisms affecting densification rates and their dependency on microstructural properties of firn, which are difficult to observe. Our study demonstrates that, despite these observational limitations, thorough calibration methods relying only on climatic variables can substantially improve firn model accuracy, and constrain uncertainties.

Author contributions. VV, AL and CN conceived this study. VV performed the development of the calibration method, performed the model experiments and led writing of the manuscript. AL and CN supervised the work. MS developed the Community Firn Model. PKM provided firn core data. BN and JMVW provided the RACMO2 forcing data. All authors provided comments and suggested edits to the manuscript.

Data availability. 41 of the 91 firn cores are from the SUMup dataset (2019 release), which is publicly available from the Arctic Data Center (doi: 10.18739/A26D5PB2S). 41 of the 91 firn cores are from the dataset compiled by Matt Spencer (Spencer et al., 2001), which is publicly available via the NASA Global Change Master Directory as 'LSSU_PSU_Firn_data' (https://drive.google.com/drive/folders/0B_IQfVYZbcWFMDc4M2ZjOTEtNWNhOS00NjdmLTkxMjctYWZlNmZkMDg2OGFh?hl=en_US). 5 of the 91 firn cores were provided by Joe McConnell and Ellen Mosley-Thompson and are available on request through PKM. 2 of the 91 cores are available via the PANGEA website (https://doi.pangaea.de/10.1594/PANGAEA.227732) and (https://doi.pangaea.de/10.1594/PANGAEA.615238). 1 of the 91 cores is available via the NOAA website (ftp://ftp.ncdc.noaa.gov/pub/data/paleo/icecore/antarctica/newall/). 1 of the 91 cores is available via the USAP website (http://www.usap-dc.org/view/dataset/609215). All Antarctic RACMO2.3p2 climate data used are available on request through JMVW and yearly climate variables are free to download from the IMAU website (https://www.projects.science.uu.nl/iceclimate/publications/data/2018/index.php). All Greenland RACMO2.3p2 climate data used are available on request through BN and yearly SMB and components are free to download (https://doi.pangaea.de/10.1594/PANGAEA.904428). The Community Firn Model is available for download on GitHub (https://github.com/UWGlaciology/CommunityFirnModel).

Acknowledgements. We thank Lora Koenig and Lynn Montgomery for making the SUMup dataset of firn cores available and easily accessible (Koenig and Montgomery, 2019). Matt Spencer is also acknowledged for publishing a separate dataset of firn cores (Spenceret al., 2001). We thank Joe McConnell and Ellen Mosley-Thompson, supported by the NSF-NASA PARCA Project, for providing additional firn core data (Bales et al., 2009; Banta & McConnell, 2007; McConnell et al. 2002; Mcconnell et al., 2000; Mosley-Thompson et al., 2001). We thank Malcolm McMillan for his interest in the study and for providing insight into the subject of ice sheet mass balance assessments. VV thanks Elizabeth Morris for pointing out errors in geographical coordinates of some of the firn cores and for her endless interest in firn densification. The Centre for Polar Observation and Modelling is acknowledged for supporting VV in his research. AL and CN research is supported by EPSRC, *A Data Science for the Natural Environment*, EP/R01860X/1. PKM acknowledges support from NESSC (Netherlands Earth System Science Centre). BN was funded by the NWO (Netherlands Organisation for Scientific Research) VENI grant VI.Veni.192.019. We thank all contributors to the development of the CFM who are not authors of this study. All authors thank two anonymous referees for their time and effort in reviewing the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

References

Alley, R. B.: Firn Densification By Grain-Boundary Sliding : a First Model, Le J. Phys. Colloq., 48(C1), C1-249-C1-256, doi:10.1051/jphyscol:1987135, 1987.

Anderson, D. L. and Benson, C. S.: The densification and diagenesis of snow, in W. D. Kingery, ed. Ice and snow: properties, processes, and applications, pp. 391–411, MIT Press, Cambridge, MA., 1963.

Arnaud, L., Gay, M., Barnola, J.-M. J. M. and Duval, P.: Physical modeling of the densification of snow / firn and ice in the upper part of polar ice sheets, in Physics of Ice Core Records, edited by T. Hondoh, pp. 285–305, Hokkaido University Press, Sapporo, Japan., 2000.

Arthern, R. J. and Wingham, D. J.: The Natural Fluctuations of Firn Densification and Their Effect on the Geodetic Determination of Ice Sheet Mass Balance, Clim. Change, 40(3–4), 605–624, doi:10.1023/A:1005320713306, 1998.

Arthern, R. J., Vaughan, D. G., Rankin, A. M., Mulvaney, R. and Thomas, E. R.: In situ measurements of Antarctic snow compaction compared with predictions of models, J. Geophys. Res. Earth Surf., 115, 1–12, doi:10.1029/2009JF001306, 2010.

Aster, R. C., Borchers, B. and Clifford, H. T.: Parameter estimation and inverse problems, Elsevier, Amsterdam., 2005.

Babonis, G. S., Csatho, B. and Schenk, T.: Mass balance changes and ice dynamics of Greenland and Antarctic ice sheets from laser altimetry, Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci., 41(July), 481–487, doi:10.5194/isprsarchives-XLI-B8-481-2016, 2016.

Berliner, L. M., Jezek, K., Cressie, N., Kim, Y., Lam, C. Q. and Van Der Veen, C. J.: Modeling dynamic controls on ice streams: A Bayesian statistical approach, J. Glaciol., 54(187), 705–714, 2008.

Brinkerhoff, D. J., Aschwanden, A. and Truffer, M.: Bayesian Inference of Subglacial Topography Using Mass Conservation, Front. Earth Sci., 4(February), 1–15, doi:10.3389/feart.2016.00008, 2016a.

Brinkerhoff, D. J., Meyer, C. R., Bueler, E., Truffer, M. and Bartholomaus, T. C.: Inversion of a glacier hydrology model, Ann. Glaciol., 57(72), 84–95, doi:10.1017/aog.2016.3, 2016b.

van den Broeke, M. R., Enderlin, E. M., Howat, I. M., Kuipers Munneke, P., Noël, B. P. Y., van de Berg, W. J., van Meijgaard, E. and Wouters, B.: On the recent contribution of the Greenland ice sheet to sea level change, The Cryosphere, 10, 1933–1946, doi:10.5194/tc-10-1933-2016, 2016.

Buizert, C., Gkinis, V., Severinghaus, J. P., He, F., Lecavalier, B. S., Kindler, P., Leuenberger, M., Carlson, A. E., Vinther, B., Masson-Delmotte, V., White, J. W. C., Liu, Z., Otto-Bliesner, B. and Brook, E. J.: Greenland temperature response to climate forcing during the last deglaciation, Science (80-.)., 345(6201), 1177–1180, doi:10.1126/science.1254961, 2014.

Conger, S. M. and McClung, D.: Instruments and Methods: Comparison of density cutters for snow profile observations, J. Glaciol., 55, 163–169, doi:10.3189/002214309788609038, 2009.

Fausto, R. S., Box, J. E., Vandecrux, B., van As, D., Steffen, K., MacFerrin, M., Machguth H., Colgan W., Koenig L. S., McGrath D., Charalampidis, C., and Braithwaite, R. J.: A Snow Density Dataset for Improving Surface Boundary Conditions in Greenland Ice Sheet Firn Modeling, Front. Earth Sci., 6, 51 pp., doi:10.3389/feart.2018.00051, 2018.

Freitag, J., Kipfstuhl, S., Laepple, T. and Wilhelms, F.: Impurity-controlled densification: a new model for stratified polar firn, J. Glaciol., 59(218), 1163–1169, doi:10.3189/2013jog13j042, 2013.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D.: Bayesian Data Analysis, Third Edit., CRC Press Taylor & Francis Group, Boca Raton., 2013.

Goujon, C., Barnola, J.-M. and Ritz, C.: Modeling the densification of polar firn including heat diffusion: Application to close-off characteristics and gas isotopic fractionation for Antarctica and Greenland sites, J. Geophys. Res. Atmos., 108, n/a-n/a, doi:10.1029/2002JD003319, 2003.

Gow, A. J., Meese, D. A. and Bialas, R. W.: Accumulation variability, density profiles and crystal growth trends in ITASE firn and ice cores from West Antarctica, Ann. Glaciol., 39, 101–109, doi:10.3189/172756404781814690, 2004.

Gudmundsson, G. H.: Estimating basal properties of glaciers from surface measurements, in Knight, P.G., ed. Glacier science and environmental change. Oxford, Blackwell, pp. 415–417., 2006.

Hastings, W. K.: Monte Carlo sampling methods using Markov chains and their applications, Biometrika, 57(1), 97–109, 1970.

Hawley, R., Brandt, O., Morris, E., Kohler, J., Shepherd, A. and Wingham, D.: Techniques for measuring high-resolution firn density profiles: Case study from Kongsvegen, Svalbard. J. Glaciol., 54(186), 463-468, doi:10.3189/002214308785837020, 2008.

Hawley, R. L. and Waddington, E. D.: Instruments and Methods in situ measurements of firn compaction profiles using borehole optical stratigraphy, J. Glaciol., 57(202), 289–294, doi:10.3189/002214311796405889, 2011.

Helsen, M. M., van den Broeke, M. R., van de Wal, R. S. W., van de Berg, W. J., van Meijgaard, E., Davis, C. H., Li,

Y. and Goodwin, I.: Elevation changes in antarctica mainly determined by accumulation variability, Science (80-.)., 320, 1626–1629, doi:10.1126/science.1153894, 2008.
Herron, M. and Langway, C.: Firn densification: an empirical model, J. Glaciol., 25(93), 373–385, https://doi.org/10.3189/S0022143000015239, 1980.
Koenig, L. and Montgomery, L.: Surface mass balance and snow depth on sea ice working group (SUMup) snow density subdataset, Greenland and Antarctica, 1950-2018, , doi:10.18739/A26D5PB2S, 2019.
Kuipers Munneke, P., Ligtenberg, S. R. M., Noël, B. P. Y., Howat, I. M., Box, J. E., Mosley-Thompson, E., McConnell, J. R., Steffen, K., Harper, J. T., Das, S. B. and van den Broeke, M. R.: Elevation change of the Greenland ice sheet due to surface mass balance and firn processes, 1960-2013, The Cryosphere, 9, 3541–3580, doi:10.5194/tcd-9-3541-2015, 2015.
Li, J. and Zwally, H. J.: Modeling of firn compaction for estimating ice-sheet mass change from observed ice-sheet elevation change, Ann. Glaciol., 52, 1–7, doi:10.3189/172756411799096321, 2011.
Li, J. and Zwally, H. J.: Response times of ice-sheet surface heights to changes in the rate of Antarctic firn compaction caused by accumulation and temperature variations, J. Glaciol., 61, 1037–1047, doi:10.3189/2015JoG14J182, 2015.
Ligtenberg, S. R. M., Helsen, M. M. and Van Den Broeke, M. R.: An improved semi-empirical model for the densification of Antarctic firn, The Cryosphere, 5, 809–819, doi:10.5194/tc-5-809-2011, 2011.
Ligtenberg, S. R. M., Kuipers Munneke, P., and van den Broeke, M. R.: Present and future variations in Antarctic firn air content, The Cryosphere, 8, 1711–1723, doi:10.5194/tc-8-1711-2014, 2014.
Lundin, J. M. D., Stevens, C. M., Arthern, R., Buizert, C., Orsi, A., Ligtenberg, S. R. M., Simonsen, S. B., Cummings, E., Essery, R., Leahy, W., Harris, P., Helsen, M. M. and Waddington, E. D.: Firn Model Intercomparison Experiment (FirnMICE), J. Glaciol., 63, 401–422, doi:10.1017/jog.2016.114, 2017.
Maeno, N. and Ebinuma, T.: Pressure sintering of ice and its implication to the densification of snow at polar glaciers and ice sheets, J. Phys. Chem., 87(21), 4103–4110, doi:10.1021/j100244a023, 1983.
McMillan, M., Leeson, A., Shepherd, A., Briggs, K., Armitage, T., Hogg, A., Kuipers Munneke, P., van den Broeke, M., Noël, B., van de Berg, W. J., Ligtenberg, S., Horwath, M., Groh, A., Muir, A. and Gilbert, L.: A high-resolution record of Greenland mass balance, Geophys. Res. Lett., 43, 7002–7010, doi:10.1002/2016GL069666, 2016.
Medley, B., Ligtenberg, S. R. M., Joughin, I., Van Den Broeke, M. R., Gogineni, S. and Nowicki, S.: Antarctic firn compaction rates from repeat-track airborne radar data: I. Methods, Ann. Glaciol., 56(70), 155–166, doi:10.3189/2015AoG70A203, 2015.
Morris, E. M.: Modeling Dry-Snow Densification without Abrupt Transition, Geosciences, 8(12), 464, doi:10.3390/geosciences8120464, 2018.
Morris, E. M. and Wingham, D. J.: Densification of polar snow: Measurements, modeling, and implications for altimetry, J. Geophys. Res. Earth Surf., 119, 349–365, doi:10.1002/2013JF002898, 2014.
Morris, E. M., Mulvaney, R., Arthern, R. J., Davies, D., Gurney, R. J., Lambert, P., De Rydt, J., Smith, A. M., Tuckwell, R. J. and Winstrup, M.: Snow Densification and Recent Accumulation Along the iSTAR Traverse, Pine Island Glacier, Antarctica, J. Geophys. Res. Earth Surf., 122, 2284–2301, doi:10.1002/2017JF004357, 2017.
Noël, B., van de Berg, W. J., van Wessem, J. M., van Meijgaard, E., van As, D., Lenaerts, J. T. M., Lhermitte, S., Kuipers Munneke, P., Smeets, C. J. P. P., van Ulft, L. H., van de Wal, R. S. W., and van den Broeke, M. R.: Modelling the climate and surface mass balance of polar ice sheets using RACMO2 –Part 1: Greenland (1958–2016), The Cryosphere, 12, 811–831, https://doi.org/10.5194/tc-12-811-2018, 2018.
Noël, B., van de Berg, W. J., Lhermitte, S. and van den Broeke, M. R.: Rapid ablation zone expansion amplifies north Greenland mass loss, Sci. Adv., 5(9), eaaw0123, doi:10.1126/sciadv.aaw0123, 2019.
Pritchard, H. D., Ligtenberg, S. R. M., Fricker, H. A., Vaughan, D. G., Van Den Broeke, M. R. and Padman, L.: Antarctic ice-sheet loss driven by basal melting of ice shelves, Nature, 484(7395), 502–505, doi:10.1038/nature10968, 2012.
Proksch, M., Rutter, N., Fierz, C., and Schneebeli, M.: Intercomparison of snow density measurements: bias, precision, and vertical resolution, The Cryosphere, 10, 371–384, doi:10.5194/tc10-371-2016, 2016.
Raymond, M. J. and Gudmundsson, G. H.: Estimating basal properties of ice streams from surface measurements: A non-linear Bayesian inverse approach applied to synthetic data, The Cryosphere, 3(2), 265–278, doi:10.5194/tc-3-265-2009, 2009.
Robin, G. d. Q.: Glaciology III: Seismic shooting and related investigations, in Norwegian-British-Swedish Antarctic Expedition, 1949–52, Scientific Results, vol. 5, Norsk Polarinstit, Oslo., 1958.
Rosenthal, J.: Optimal Proposal Distributions and Adaptive MCMC, (1), 1–25, doi:10.1201/b10905-5, 2010.
Schleef, S. and Löwe, H.: X-ray microtomography analysis of isothermal densification of new snow under external mechanical stress, J. Glaciol., 59(214), 233–243, doi:10.3189/2013JoG12J076, 2013.
Shepherd, A., Gilbert, L., Muir, A. S., Konrad, H., McMillan, M., Slater, T., Briggs, K., Sundal, V., Hogg, A. and

Engdahl, E.: Trends in Antarctic Ice Sheet elevation and mass. Geophysical Research Letters, 46, 8174–8183, doi:10.1029/2019GL082182, 2019.

Simonsen, S. B., Stenseng, L., Adalgeirsdóttir, G., Fausto, R. S., Hvidberg, C. S. and Lucas-Picher, P.: Assessing a multilayered dynamic firn-compaction model for Greenland with ASIRAS radar measurements, J. Glaciol., 59, 545–558, doi:10.3189/2013JoG12J158, 2013.

Spencer, M. K., Alley, R. B. and Creyts, T. T.: Preliminary firn-densification model with 38-site dataset, J. Glaciol., 47, 671–676, https://doi.org/10.3189/172756501781831765, 2001.

Stevens, C. M., Verjans, V., Lundin, J. M. D., Kahle, E. C., Horlings, A. N., Horlings, B. I., and Waddington, E. D.: The Community Firn Model (CFM) v1.0, Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2019-361, in review, 2020.

Verjans, V., Leeson, A. A., Max Stevens, C., MacFerrin, M., Noël, B. and Van Den Broeke, M. R.: Development of physically based liquid water schemes for Greenland firn-densification models, The Cryosphere, 13(7), 1819–1842, doi:10.5194/tc-13-1819-2019, 2019.

van den Broeke, M. R., Enderlin, E. M., Howat, I. M., Kuipers Munneke, P., Noël, B. P. Y., van de Berg, W. J., van Meijgaard, E. and Wouters, B.: On the recent contribution of the Greenland ice sheet to sea level change, The Cryosphere, 10, 1933–1946, doi:10.5194/tc-10-1933-2016, 2016.

van Wessem, J. M., Jan Van De Berg, W., Noël, B. P. Y., Van Meijgaard, E., Amory, C., Birnbaum, G., Jakobs, C. L., Krüger, K., Lenaerts, J. T. M., Lhermitte, S., Ligtenberg, S. R. M., Medley, B., Reijmer, C. H., Van Tricht, K., Trusel, L. D., Van Ulft, L. H., Wouters, B., Wuite, J. and Van Den Broeke, M. R.: Modelling the climate and surface mass balance of polar ice sheets using RACMO2 - Part 2: Antarctica (1979-2016), The Cryosphere, 12(4), 1479–1498, doi:10.5194/tc-12-1479-2018, 2018.

Vandecrux, B., MacFerrin, M., Machguth, H., Colgan,W. T., van As, D., Heilig, A., Stevens, C. M., Charalampidis, C., Fausto, R. S., Morris, E. M., Mosley-Thompson, E., Koenig, L., Montgomery, L. N., Miège, C., Simonsen, S. B., Ingeman-Nielsen, T., and Box, J. E.: Firn data compilation reveals widespread decrease of firn air content in western Greenland, The Cryosphere, 13, 845–859, doi:10.5194/tc-13-845-2019, 2019.

Wilkinson, D.: A pressure-sintering model for the densification of polar firn and glacier ice, J. Glaciol., 34(116), 40–45, doi:10.3189/S0022143000009047, 1988.

Parameter	Value in original model	Prior distribution	MAP	95 % Credible interval
k_0^* [m w.e.$^{-a}$]	11	$N(11, 100)$	17.4	7.58; 28.4
k_1^* [m w.e.$^{-b}$]	575	$N(575, 9\ 10^4)$	524	260; 1060
E_0 [J mol^{-1}]	10 160	$N(10160, 4\ 10^6)$	10 840	9 000; 12 290
E_1 [J mol^{-1}]	21 400	$N(21400, 4\ 10^6)$	20 800	18 900; 22 300
a [/]	1	$N(1, 0.4)$	0.91	0.74; 1.02
b [/]	0.5	$N(0.5, 0.4)$	0.63	0.54; 0.78
k_0^{Ar} [m w.e.$^{-\alpha}$]	0.07	$N(0.07, 4.9\ 10^{-3})$	0.077	0.046; 0.137
k_1^{Ar}[m w.e.$^{-\beta}$]	0.03	$N(0.03, 9\ 10^{-4})$	0.025	0.015; 0.048
E_c [J mol^{-1}]	60 000	Fixed: 60000	/	/
E_g [J mol^{-1}]	42 400	$N(42400, 16\ 10^6)$	40 900	39 700; 42 000
α [/]	1	$N(1, 0.4)$	0.80	0.66; 0.89
β [/]	1	$N(1, 0.4)$	0.68	0.59; 0.81
lz_a	8.36	$N(8.36, 36)$	7.31	3.93; 12.82
lz_b	-2.061	$N(-2.061, 2)$	-2.124	-2.319; -1.896
lz_{11}	-9.788	$N(-9.788, 36)$	-14.710	-20.839; -5.469
lz_{12}	8.996	$N(8.996, 36)$	7.269	2.680; 17.724
lz_{13}	-0.6165	$N(-0.6165, 1)$	-1.019	-1.389; -0.509
lz_{21}	-2.0178	$N(-2.0178, 2)$	-1.513	-2.970; -0.258
lz_{22}	8.4043	$N(8.4043, 36)$	6.0203	4.911; 12.942
lz_{23}	-0.0932	$N(-0.0932, 0.25)$	-0.0913	-0.133; -0.0460

Table 1. Information for the free parameters of HL (top), Ar (middle) and LZ (low). $N(x, y)$ designates a normal distribution of mean x and variance y. The variances in the prior distributions are taken to generate weakly informative distributions. Some prior correlation is prescribed for the pairs (k_0^*, E_0), (k_1^*, E_1), (k_0^{Ar}, E_g), (k_1^{Ar}, E_g) and (k_0^{Ar}, k_1^{Ar}) (see Supplementary Information). MAP estimates and ~~Credible~~ credible intervals are results from the calibration process.

Model	RMSE ($DIP15$) [m]	RMSE ($DIPpc$) [m]
HL original	0.503	2.395
HL MAP	0.382	1.862
HL 500 random sample	0.396	1.899
Ar original	0.772	4.566
Ar MAP	0.426	1.780
Ar 500 random sample	0.448	1.889
LZ original	0.452	1.812
LZ dual	0.505	3.883
LZ MAP	0.463	2.392
LZ 500 random sample	0.486	2.296
IMAU-FDM	0.418	2.681

Table 2. Model results on the evaluation data. The errors are calculated with respect to the observations of depth integrated porosity until 15 m depth and until pore close-off.

Coefficient of Variation	HL	Ar	LZ	Combined (HL, Ar, LZ)
cmp_{an}	5.8%	5.8%	6.5%	19.5%
age_{pc}	6.5%	5.8%	7.5%	7.5%

Table 3. Coefficients of variation for the 2000-2017 cumulative compaction anomaly (cmp_{an}) and firn age at pore close-off depth (age_{pc}). Values are computed from results of 500 randomly selected parameter combinations from the posterior ensembles of each model (HL, Ar, LZ). Coefficients of variation are averaged across all sites of the dataset.

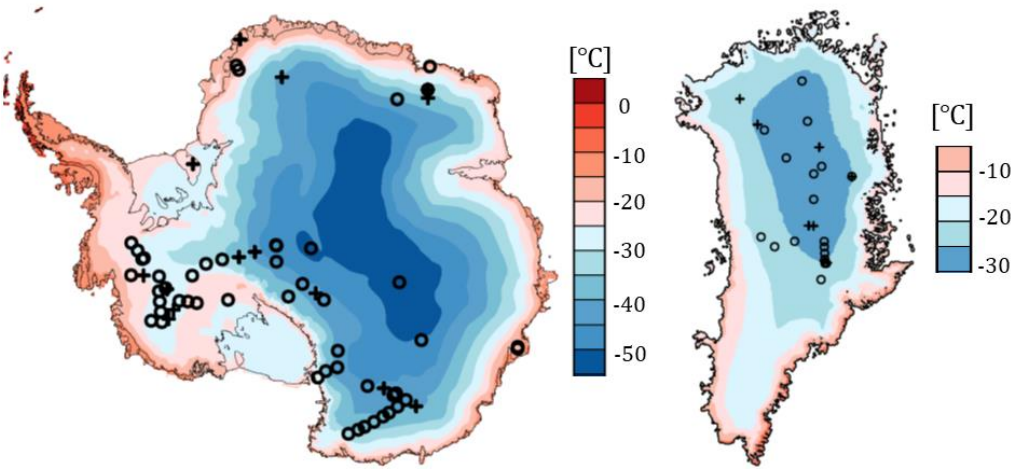


Figure 1. Maps of Antarctic (left) and Greenland (right) ice sheets. Background is mean annual air temperature as modelled by RACMO2. Note the different colour scales.

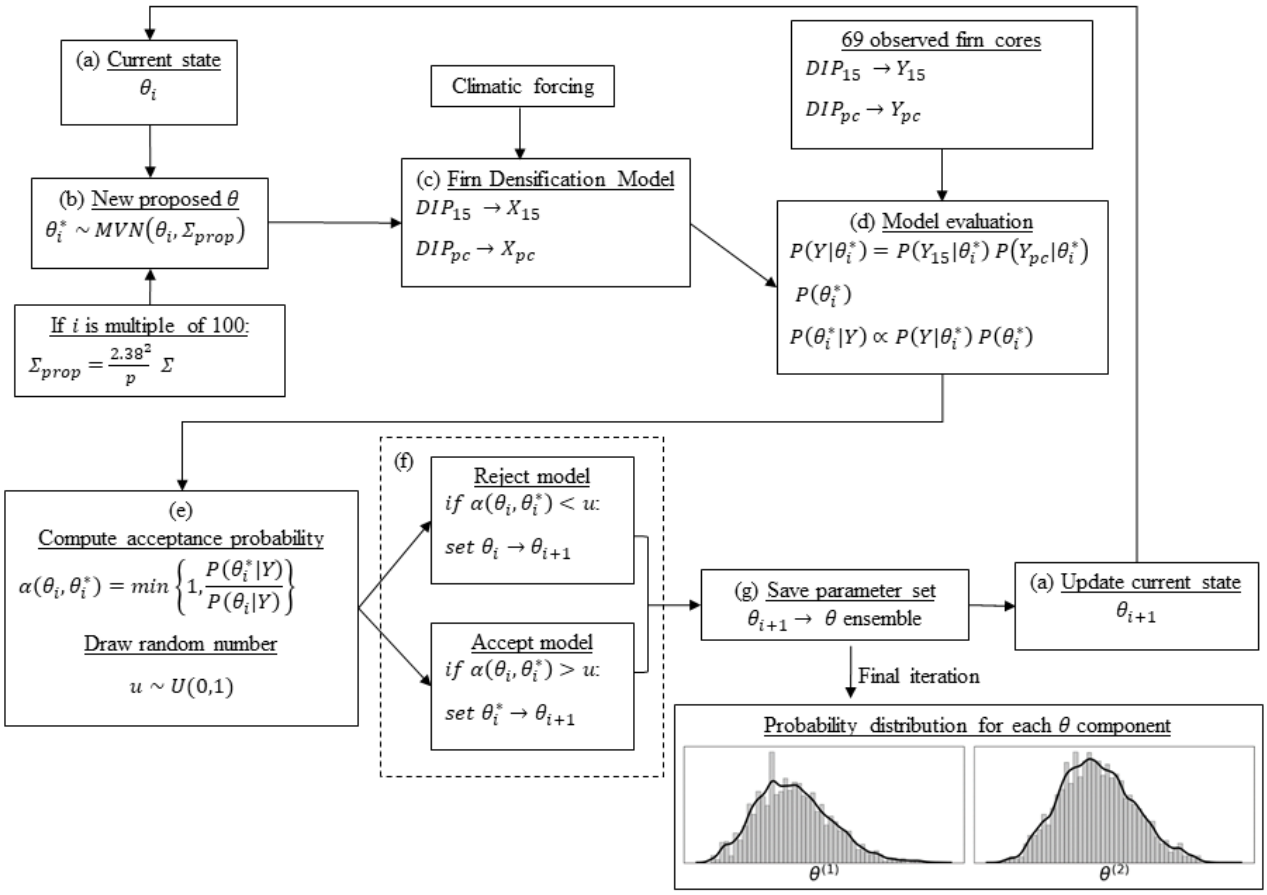


Figure 2. Implementation of the Random Walk Metropolis algorithm. θ represents a parameter combination of any given firn densification model investigated.

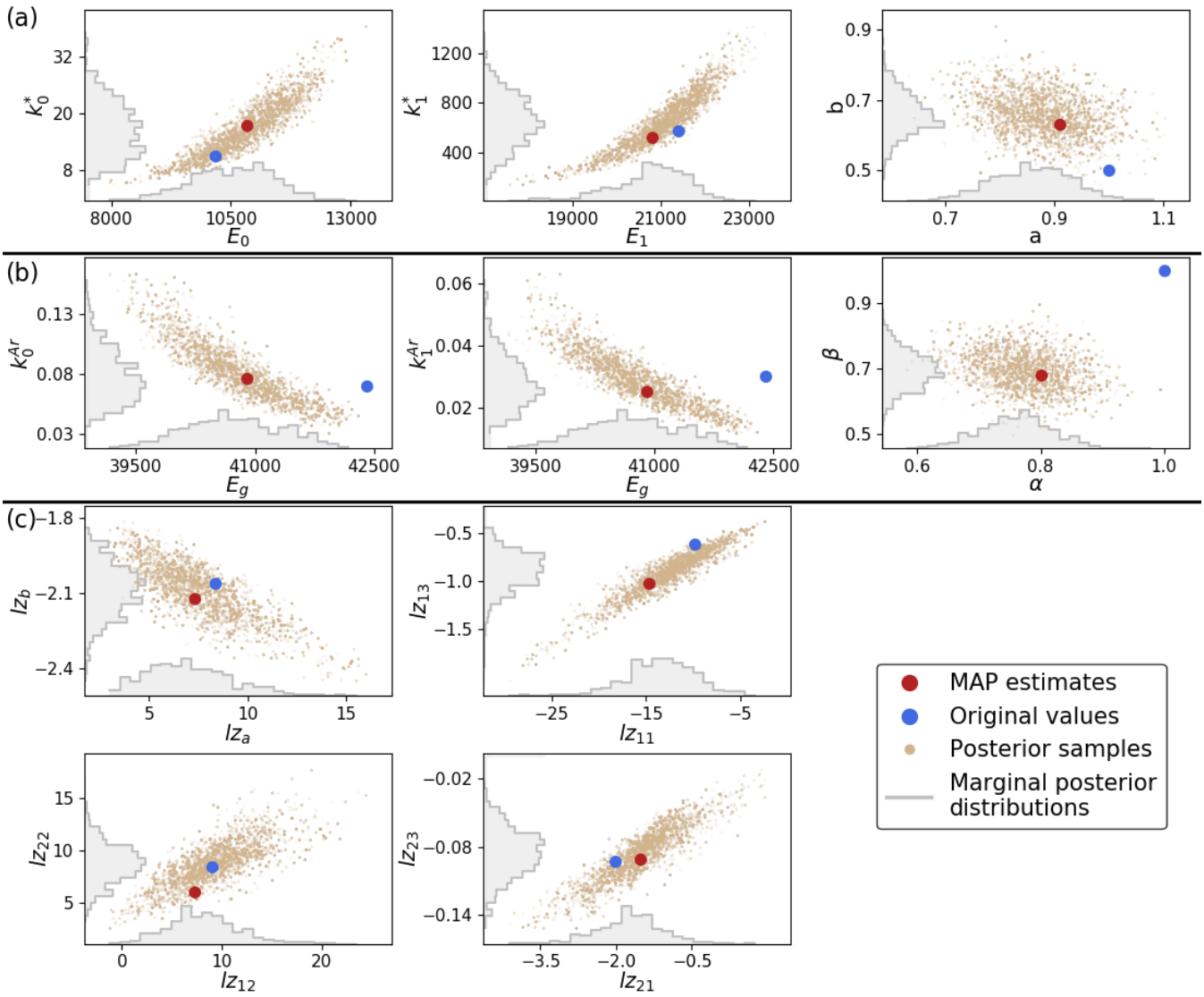


Figure 3. Posterior probability distributions, shown for pairs of parameters, for (a) HL, (b) Ar, (c) LZ. Where possible, correlated parameters share the same graph (see Supplementary Information for full correlation matrices).

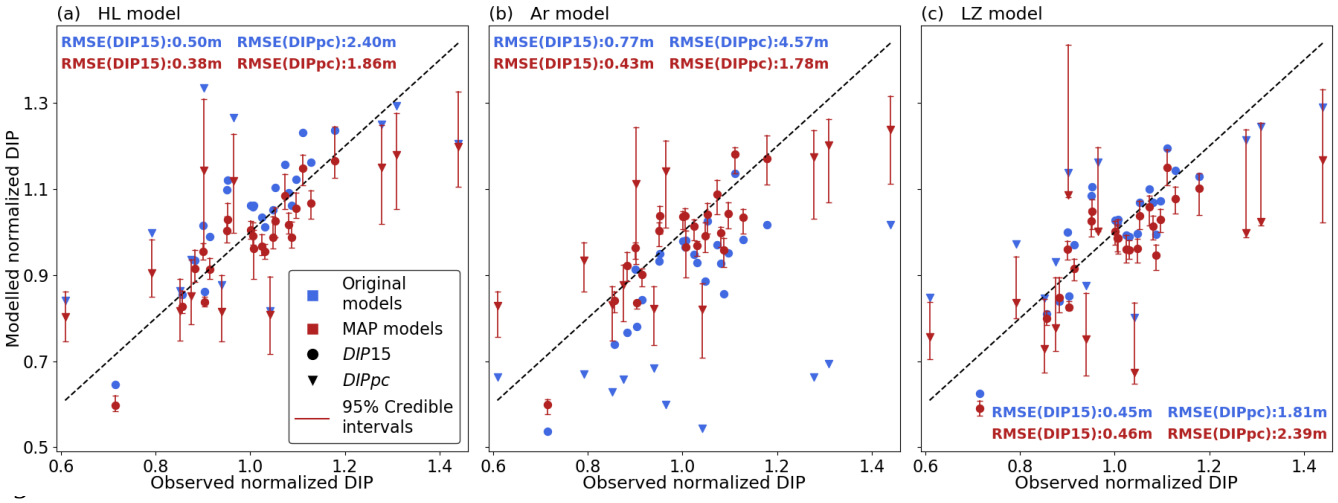


Figure 4. Comparison of evaluation data *DIP* with model results

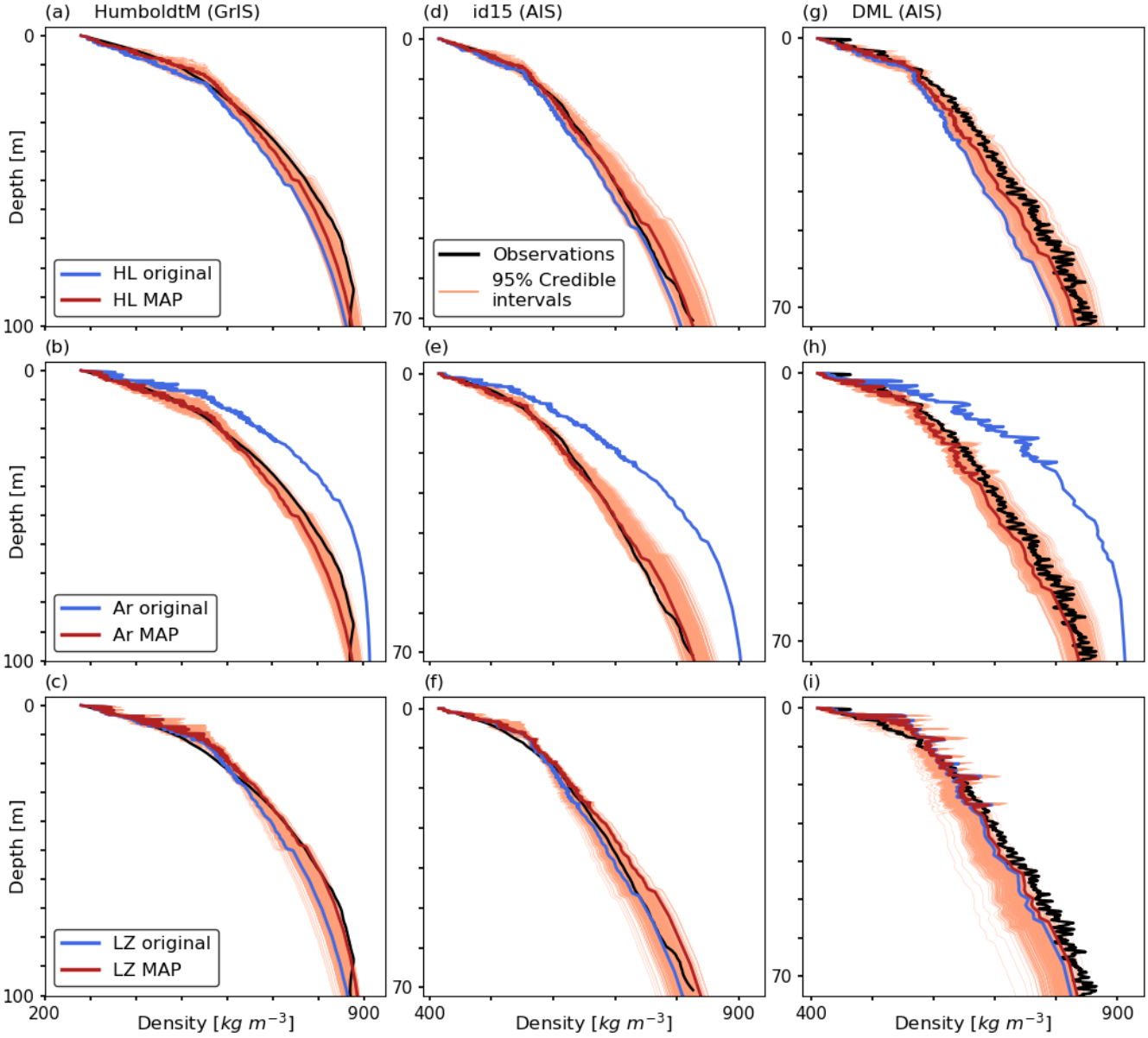


Figure 5. Depth-density profiles at three evaluation sites. ~~DMZ~~ DML is a climatic outlier of our dataset with particularly high temperatures and accumulation rates.

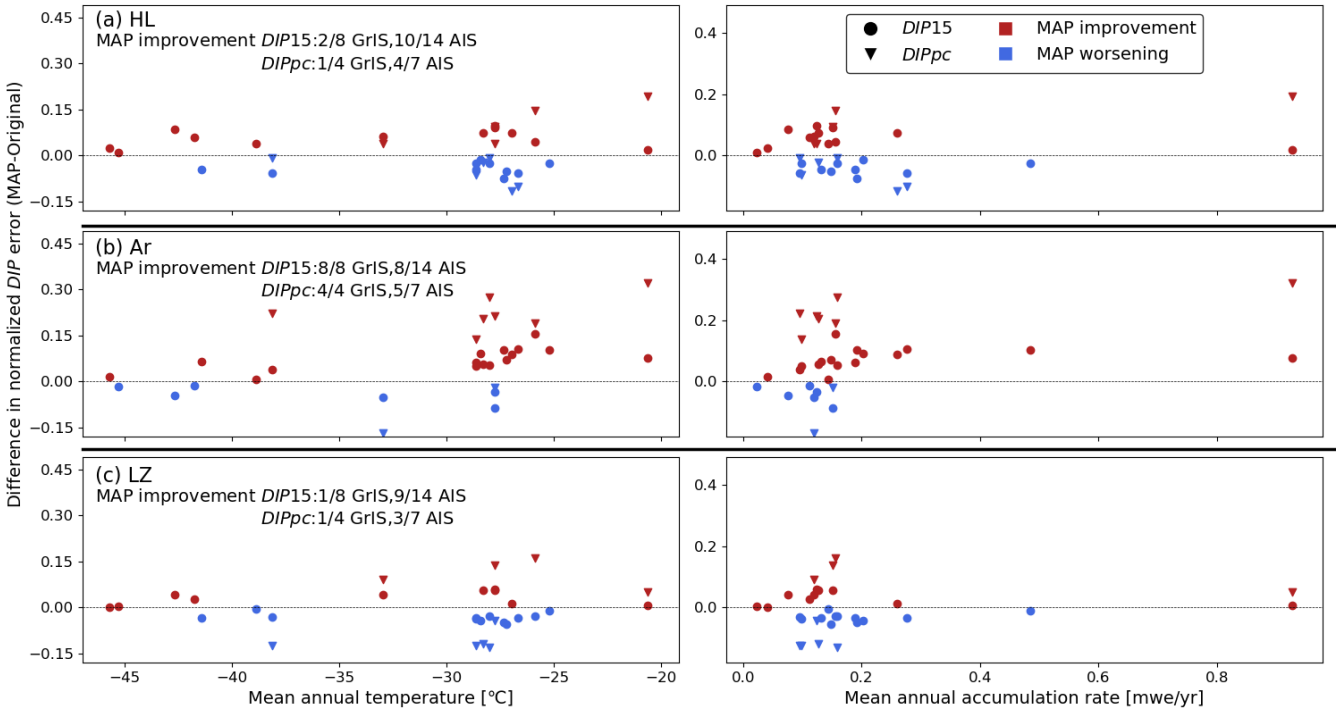


Figure 6. Improvements of the MAP models with respect to the original models for the evaluation data. The ratios indicate the ratios of cores for which an improvement is achieved by the corresponding MAP.

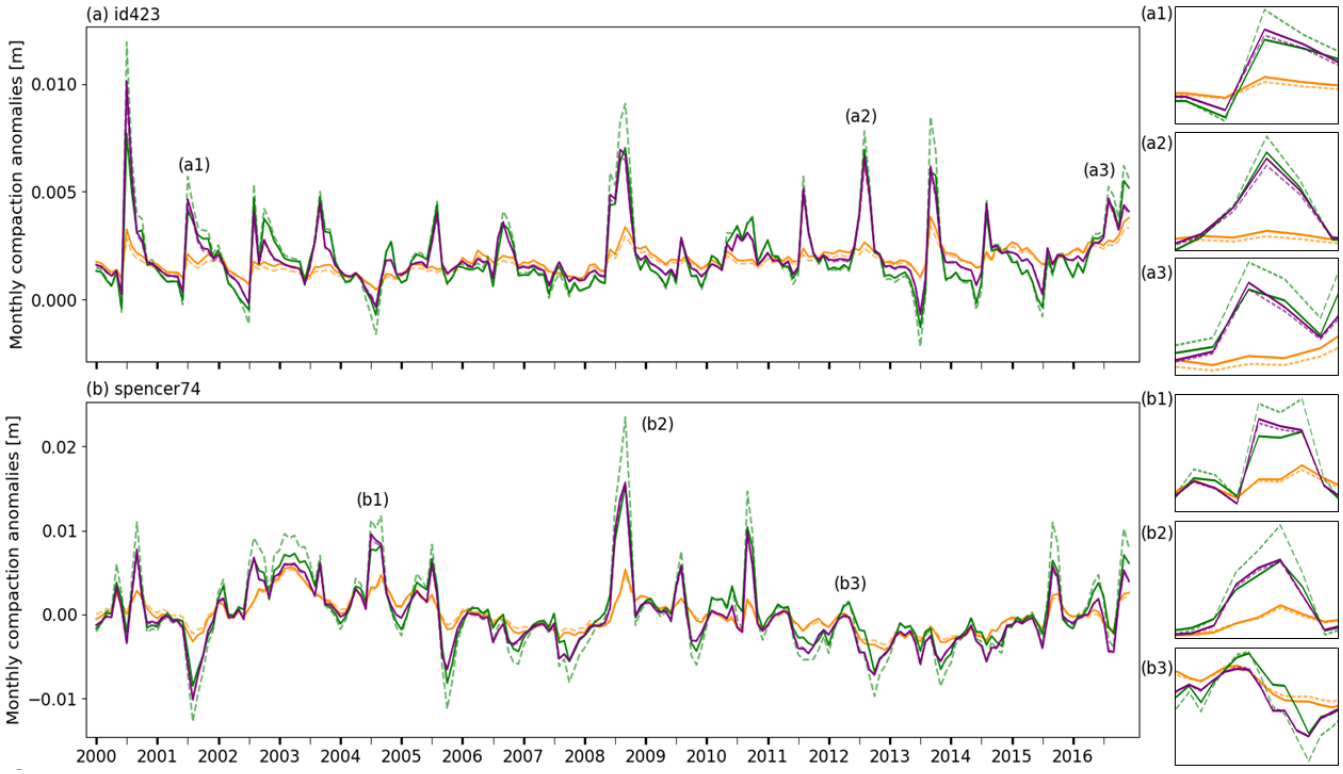


Figure 7. Monthly time-series of compaction anomalies at two sites on the GrIS. Insets show details for particular intervals of the time-series.

S1 Separation between the calibration and the evaluation data

The 91 sites of the dataset span a broad range of temperature and accumulation rate conditions (Table S1 and Fig. S1). As explained in the main text, our objective is to select the evaluation data (22 cores) randomly but still making it representative of (i) all climatic conditions and (ii) the ratio of GrIS to AIS sites of the dataset. We separate the 91 observed cores in three tiers of lowest, middle and highest T_{av} and we select randomly 7 cores in each tier for the evaluation data. We repeat this random selection until 5 to 10 out of the 21 cores are from GrIS, with the remainder from AIS. Finally, our dataset includes two sites that are climatic outliers with respect to the others (DML and spencer4 in Table S1) with high T_{av} and $\dot{b}$ values (Figure S1). We select randomly one of these for the evaluation data. Proceeding to the selection based on $\dot{b}$ rather than T_{av} would be similar given the strong correlation between both variables.

S2 Application of random noise in the boundary conditions

In order to let uncertainty in RACMO2 output affect the calibration process, we perturb the temperature and accumulation time series that serve as climatic forcing for the firn models. At each iteration (a round of simulations with a given parameter set at all the calibration sites) and for each individual calibration site, we randomly draw an individual climatic perturbation value c_p from a standard Normal distribution (Eq. (S1)). As such, every calibration site has a specific c_p value, which changes at each iteration. We use observed statistics of RACMO2 errors in temperature and Surface Mass Balance to determine the perturbation.

For GrIS, Noël et al. (2019) report RMSE values with respect to field observations for temperature and surface mass balance flux of 2.1 K and 69 m w.e. yr^{-1} respectively (in their Supplementary Material).

Each monthly value of the RACMO2 time series is therefore perturbed by the corresponding RMSE value scaled by c_p (Eq. (S2), (S3), (S4)).

We favour this approach rather than drawing a different random perturbation at each time step. The latter method would cause perturbations of opposite signs to occur on a very short timescale, which would result in unrealistic short term climatic variability (e.g. a very warm perturbation could be immediately followed by a very cold perturbation in the next month). Also, using the same c_p value to quantify the magnitude of the perturbation for temperature and accumulation preserves the strong correlation between these variables. Warm (cold) temperature perturbations coincide with high (low) accumulation perturbations, which keeps our random perturbations physically plausible.

The part of the total accumulation perturbation attributed to each monthly time step is weighted by the actual accumulation at that time step. This attributes larger absolute noise in accumulation to high-accumulation months and lower absolute noise to low-accumulation months (Eq. (S3), (S4)).

Our approach is summarized in Eq. (S1), (S2), (S3) and (S4). These equations are applied at all iterations of the calibration process.

$$c_p \sim N(0,1) \quad \text{at all calibration sites} \tag{S1}$$

$$T_t^* = T_t + c_p \sigma_T \quad \text{at all } t \tag{S2}$$

$$\dot{b}_{tot}^* = n_{yr} c_p \sigma_{SMB} \tag{S3}$$

$$\dot{b}_t^* = \dot{b}_t + \dot{b}_{tot}^* \frac{\dot{b}_t}{\sum_t \dot{b}_t} \tag{S4}$$

where T_t and $\dot{b}_t$ are temperature and accumulation rate as computed by RACMO2 at time step t and the $*$ superscript denotes the perturbed quantity. n_{yr} is the total number of years in a given simulation, $\dot{b}_{tot}^*$ is the total accumulation perturbation applied for that simulation and σ_T and σ_{SMB} are the temperature and surface mass balance flux RMSE values (as mentioned above, σ_T = 2.1 K and σ_{SMB} = 69 m w.e. yr^{-1} for GrIS). Note that by using a RMSE value on the surface mass balance flux, we overestimate uncertainty because the observed RMSE is mostly driven by errors in melt amounts which do not apply at the sites of our dataset, all from the dry snow zone area. For AIS, we apply the exact

same process for perturbing the temperature variables. We use the RMSE value reported by van Wessem et al. (2018) and set σ_T = 1.3 K. The accumulation conditions of AIS forces the use of a slightly different method for perturbing the accumulation rate. In terms of magnitude, RACMO2 errors are much larger in coastal areas, where accumulation rates are high. In contrast, in the dry interior of the ice sheet where most of the cores of our dataset come from, the magnitude of RACMO2 errors is small due to low accumulation rates. As such, applying noise based on the ice sheet wide RMSE value would result in noise signals larger than actual accumulation values at most of our dry sites. We thus use the average RACMO surface mass balance bias of 5% (van Wessem et al., 2018) as a proxy for one standard deviation. For AIS, Eq. (S3) and (S4) are replaced by Eq. (S5).

$$\dot{b}_t^* = \dot{b}_t + 0.05\, c_n\, \dot{b}_t \quad \text{(S5)}$$

As explained in Sect. 2.2, we also let uncertainty in fresh snow density, ρ_0, affect the calibration process by applying random perturbations to each ρ_0^t. In contrast to the climatic perturbation, the perturbation in ρ_0 must not be iteration specific but can be specific to each single time step t. Indeed, it is not unrealistic that a month with anomalously low fresh snow density is immediately followed by a month of anomalously high fresh snow density for example. We determine surface density values at each site from the firn cores of our dataset, ρ_0^{core}, and we perturb these values based on a standard deviation of 25 kg m^{-3}. As such, adding noise to ρ_0 simplifies to Eq. (S6).

$$\rho_{0,t}^* \sim N(\rho_0^{core}, 25) \quad \text{(S6)}$$

We emphasize that the aim of this study is not to conduct a complete sensitivity analysis of firn densification to climatic forcing and to fresh snow density. The objective of the perturbations is to let reasonable estimates of errors in those fields to be accounted for in the calibration process.

S3 Prior correlations in HL and Ar

The Arrhenius form of HL and Ar (Eq. (4) and (5)) allows us to include some correlation in the prior distributions over the parameters of these models. The values of the Arrhenius pre-exponential factors (k_0^*, k_1^*, k_0^{Ar} and k_1^{Ar}) are correlated with their corresponding activation energies (E_0, E_1 and E_g). For any given constant temperature, modelled densification rates, $\frac{d\rho}{dt}$, can be kept constant despite a change in the pre-exponential factor if the corresponding activation energy is changed accordingly and vice versa. As such, changes in these parameters can potentially compensate in the calculation of DIP values and of the likelihood function (Eq. (8)).

By enforcing constant $\frac{d\rho}{dt}$, exact compensation is ensured by the following equalities:

$$k_{0,mv}^* = k_{0,HL}^* \exp\left(\frac{E_{0,mv} - E_{0,HL}}{R\,T}\right) \quad \text{(S7)}$$

$$k_{1,mv}^* = k_{1,HL}^* \exp\left(\frac{E_{1,mv} - E_{1,HL}}{R\,T}\right) \quad \text{(S8)}$$

$$k_{0,mv}^{Ar} = k_{0,Ar}^{Ar} \exp\left(\frac{E_{g,Ar} - E_{g,mv}}{R\,T}\right) \quad \text{(S9)}$$

$$k_{1,mv}^{Ar} = k_{1,Ar}^{Ar} \exp\left(\frac{E_{g,Ar} - E_{g,mv}}{R\,T}\right) \quad \text{(S10)}$$

where HL and Ar subscripts denote the original values in HL and Ar, and the mv subscript denotes a modified value of the parameter. Firstly, we generate 10000 random values of temperature T in the range of annual mean temperatures covered by our dataset. Secondly, for each random temperature, we generate random values of $E_{0,mv}$, $E_{1,mv}$ and $E_{g,mv}$ in an interval of ±500 J mol^{-1} around the original values. Thirdly, we calculate the corresponding values in the pre-exponential factors from Eq. (S7), (S8), (S9) and (S10). This results in 10000 pairs of $(k_{0,mv}^*, E_{0,mv})$, $(k_{1,mv}^*, E_{1,mv})$, $(k_{0,mv}^{Ar}, E_{g,mv})$ and $(k_{1,mv}^{Ar}, E_{g,mv})$, from which we calculate correlation coefficients. The absolute values of all four correlation coefficients lie in the interval [0.75; 0.78]. We decide to fix all prior correlation coefficients to -0.75 (HL parameters, negatively correlated) and 0.75 (Ar parameters, positively correlated). The process described necessarily results in perfectly correlated $k_{0,mv}^{Ar}$ and $k_{1,mv}^{Ar}$. We also set the prior correlation between these parameters to 0.75.

We emphasize here that any other pair of *a priori* uncorrelated parameters can certainly be correlated *a posteriori* if the calibration process identifies such quantitative behaviour when the observed data is considered. This is highlighted and further discussed in Sect. S7.

S4 The likelihood function, Eq. (8)

The covariance matrices Σ_{15} and Σ_{pc} that appear in Eq. (8) are diagonal matrices with the site-specific variances on the diagonal. At each site, we take ~~5~~10% of the observed $DIP15$ and ~~10~~20% of the observed $DIPpc$ as the standard deviation, and the variance value is the square of the standard deviation. We take the higher value of ~~10~~20% for $DIPpc$ because density errors propagate in firn models. Equation (3) shows that densification rates depend on the density value itself, resulting in error propagation through time. As such, if a model shows an offset compared to observations at 15 m depth, it is likely to show an even stronger offset at z_{pc}. Taking a higher variance alleviates the strength of this effect on the likelihood calculations by allowing a larger spread of model results compared to observed $DIPpc$ values.

The form of Eq. (8) corresponds to a normal likelihood function. This assumes that model DIP results are normally distributed around the observed values. To support this assumption, we conducted a preliminary verification of errors in $DIP15$ $(X_{15} - Y_{15})$ and $DIPpc$ $\left(X_{pc} - Y_{pc}\right)$ computed with the three original models (HL, Ar, LZ) on the entire dataset ~~and . we~~ We compute a basic Kolmogorov-Smirnov test for both sets of errors: residuals in $DIP15$ and in $DIPpc$. The resulting p-values are very large: 0.94 and 0.86 respectively. The distributions of these errors are thus in line with a normal distribution. We show the Quantiles-Quantiles plots for both sets of residuals in Figure S2. ~~all six sets of errors. The resulting p-values are all above 0.05. The distribution of these errors are thus in line with a normal distribution, and this is despite the presence of some outliers in the errors in $DIP15$ that heavily disfavour the assumption of normality. We show the Quantiles-Quantiles plots for all six sets of errors in Figure S2.~~ As explained in the main text, the form of Eq. (8) also assumes independence between errors in $DIP15$ and $DIPpc$, which is the reason why $DIPpc$ is calculated only from depths below 15 m. As such, observations-model discrepancies are essentially governed by parameter values of stage-1 densification for $DIP15$ and by parameter values of stage-2 densification for $DIPpc$, with little interaction between both. The same preliminary verification as mentioned above allows us to evaluate the correlation between $DIP15$ and $DIPpc$ errors for all three original models on the entire dataset. This yields correlation coefficients of 0.13, 0.60 and -0.01 for the original models HL, Ar and LZ respectively. ~~This yields a correlation coefficient of 0.40. The three correlation coefficients computed for each of the original models individually are all below 0.3.~~

S5 Convergence diagnostics

For convergence of the RWM algorithm, the chain must traverse between the peaks of the target posterior distribution multiple times. Simply examining the trace of the RWM algorithm for each parameter provides an effective way to verify this criterion. The trace is the history of accepted parameter values over the entire chain. We show this sampling history in Fig. S3. The fuzzy appearance for each parameter of each model indicates an efficient exploration of the parameter space as the samples from RWM algorithm oscillate around the posterior mode.

In addition to this, we compute the Gelman-Rubin statistic, which provides a numerical test for convergence (Gelman et al., 2013). The motivation behind this test is that if each chain (run independently) converges to the same posterior distribution, then the variances within each chain should be approximately the same. For each model (HL, Ar, LZ), we launch three different chains from different initial parameter values. For each parameter of each model, we calculate the mean within sample variance W:

$$W = \frac{s_1^2 + s_2^2 + s_3^2}{3} \quad \text{(S1)}$$

where s^2 denotes the variance of an individual chain. We then calculate the between sample variance:

$$B = \frac{n}{3-1}\sum_{i=1}^{3}\left(\bar{\theta}_i - \bar{\bar{\theta}}\right)^2 \quad \text{(S2)}$$

where n denotes the number of iterations within each chain, $\bar{\theta}_i$ the mean parameter value within each chain and $\bar{\bar{\theta}}$ is the mean of $(\bar{\theta}_1, \bar{\theta}_2, \bar{\theta}_3)$. From there, the estimate of the variance of the posterior distribution is given by:

$$\hat{\sigma}^2 = \frac{n-1}{n}W + \frac{1}{n}B \quad \text{(S3)}$$

And the Gelman-Rubin statistic is defined as:

$$R = \sqrt{\frac{\hat{\sigma}^2}{W}} \quad \text{(S4)}$$

Large values of R indicate that estimates of θ values between the different chains are significantly different. With more iterations, the chains progressively converge to the same stationary distributions and the estimates of θ become similar, resulting in values of R close to 1. We reach $R < 1.1$ for all parameters, which proves adequate convergence (Gelman et al., 2013). Two parameters of the LZ model needed a larger number of iterations to reach $R < 1.1$.

S6 Normal approximation to the posterior

The ensembles of parameter combinations obtained for each model provide large samples, representative of the posterior probability distributions over their respective parameter space. The most efficient way to assess parameter-related uncertainty is to run a model with a high number of random parameter combinations from these ensembles, which is demonstrated in Sect. 3. However, this means that for any firn modelling study, access must be easy to such posterior ensembles or an MCMC algorithm must be re-executed. To circumvent these practical difficulties, it is approximately correct to sample random parameter combinations from a multivariate normal distribution centred about the mean of the posterior ensemble and with covariance matrix set to the posterior ensemble covariance matrix. This is commonly referred to as a normal approximation to the posterior (Gelman et al., 2013). Table S2 provides both the posterior mean and posterior covariance for the HL, Ar and LZ models.

We assess how random samples from the normal approximations compare to samples from the posterior ensembles in Fig. S4. Posterior samples and the normal approximations are very similar, with correlations only slightly less well captured in the tails of the distributions. It results in a slight overestimation of uncertainty and thus conservative estimates of uncertainty. This has been confirmed by additional model simulations with values sampled from the normal approximations (not shown).

~~S4 Compaction anomaly calculation~~

~~For each model (HL, Ar, LZ, MAP_{HL}, MAP_{Ar}, MAP_{LZ}), the calculation of compaction anomalies is the same. First, we compute at each site i the mean annual compaction rate (specific to each model) during the reference period 1960-1980, $cmp^{yr}_{ref,i}$. During this period, the firn column, and thus also compaction rates, are in steady-state. Secondly, we compute the total compaction over the transient 2000-2017 period, $cmp^{00-17}_{tot,i}$. Finally, we calculate the total compaction anomaly over 2000-2017:~~

$$cmp^{00-17}_{an,i} = cmp^{00-17}_{tot,i} - 17 cmp^{yr}_{ref,i} \quad \text{(S5)}$$

~~We compute the root mean squared difference of the compaction anomalies $cmp^{00-17}_{an,i}$ at all sites between any pair of model j and k, $rmsd(j,k)$. We take this as an approximate average discrepancy between models j and k in compaction anomalies over GrIS, to which corresponds a discrepancy in estimated mass of ice. We take an approximate GrIS accumulation area of $1.5\ 10^{12}$ m^2 (Noël et al., 2019). Using an ice density ρ_i of 917 kg m^{-3}, we reach the total ice-mass equivalent discrepancy $\|m^{tot}_{an}\|(j,k)$ corresponding to their disagreement in cmp^{00-17}_{an} values.~~

$$\|m^{tot}_{an}\|(j,k) = 1.5\ 10^{12}\ rmsd(j,k)\ \rho_i \quad \text{(S6)}$$

~~The figures of disagreement between models given in Section 4 are the inter-model differences of $\|m^{tot}_{an}\|$. The model-specific values of cmp^{00-17}_{an} at the 27 GrIS sites are given in Table S2.~~

S7 Posterior correlation between parameters

The joint posterior distributions for the parameters of each model also allow us to analyse the models' internal structure, i.e. the correlation between their different parameters. The full correlation matrices are given in Fig. S4. In HL, the strongest correlation coefficients r are unsurprisingly found for the pairs of pre-exponential factor and activation energy governing densification in stage-1 (k_0^* and E_0) and in stage-2 (k_1^* and E_1) with r of 0.91 and 0.92 respectively~~0.94 in both cases~~. Higher activation energies (E_0 and E_1) imply stronger thermal barriers to the densification process and thus slower densification, and the pre-exponential factors (k_0^* and k_1^*) counter-balance the effect to still match observed DIP values. In the same way, the activation energies are negatively correlated with their respective accumulation rate exponent (a and b), but more moderately (r values of approximately ~~-0.4~~-0.5). The negative correlation of ~~-0.44~~-0.28 between a and b themselves might be linked to the density at 15 m being the lower boundary and the upper boundary condition for the calculation of $DIP15$ and $DIPpc$ respectively. Higher values of a tend to cause lighter firn at 15 m depth. Lower E_0 values compensate for this effect on $DIP15$ because the shallow firn densifies faster due to its greater sensitivity to temperature variations. The lighter 15 m depth density also affects $DIPpc$, and lower values of b compensate for this by enhancing the densification rate, which explains the negative correlation between a and b. In Ar, the interpretation is more challenging due to the use of a same activation energy in both stages. There is a strong correlation between the activation energy E_g and both pre-exponential factors k_0^{Ar} (r = -

~~0.94~~ 0.89) and k_1^{Ar} (r = ~~-0.95~~ -0.90), for the same reason as in HL. As such, this induces a strong positive correlation between the latter parameters (r = ~~0.88~~0.76). The negative correlation between α and k_1^{Ar} (r = ~~-0.37~~ -0.41) is more surprising because these parameters apply to different stages, but it reveals an interesting pattern. Higher temperatures raise densification rates at warmer sites, where accumulation rates are also higher thus further amplifying this effect. Higher accumulation rates nevertheless cause light recently deposited firn to be buried rapidly, which may cause lower density firn governed by the fast stage-1 densification to extend below 15 m. To avoid underestimation of $DIPpc$ at such sites, stage-1 densification rates must remain low enough but there is no possibility for adjusting a stage-1 specific activation energy. Lower α values generate this effect while only marginally affecting densification at colder low-accumulation sites. Thus, high k_1^{Ar} without a complementary lower α would cause $DIPpc$ underestimation at warm and high accumulation sites. We note here that, through the calibration process, the data enhanced the prior correlations we assigned in the HL and Ar models. Analysis of correlation coefficients in LZ is less straightforward because its governing equations, Eq. (6), are less interpretable than the plain Arrhenius relationship of HL and Ar. Still, we highlight some correlated pairs of parameters. As could be expected from Eq. (6), lz_a and lz_b are negatively correlated (r = ~~-0.73~~ -0.80). Also, the independent term of stage-1 densification lz_{11} is strongly correlated with the corresponding temperature-related parameter (lz_{13}, r = ~~0.9~~0.94). The same is valid for stage-2 densification between lz_{21} and lz_{23} (r = ~~0.93~~0.90). The positive correlation between lz_{12} and lz_{22} (r = ~~0.6~~0.74) is discussed in the main text.

Tables

Site	Lat	Lon	Core depth [m]	Year	Mean $\dot{b}$ [m w.e. yr^{-1}]	Mean T [°C]	ρ_0 [kg/m^3]	$DIP15$ [m]	Var $DIP15$ [m^2]	$DIPpc$ [m]	Var $DIPpc$ [m^2]
EGRIP	75.63	-35.98	20.1	2017	0.113	-29.0	285	7.816	0.611	/	/
Summit *	72.58	-38.47	22.1	2017	0.205	-28.4	330	7.500	0.562	/	/
id359	73.94	-37.63	102.4	1993	0.124	-28.8	240	6.708	0.450	11.456	5.250
id369	75.00	-30.00	19.9	1997	0.135	-27.6	335	7.454	0.556	/	/
id373	75.25	-37.62	100.8	1993	0.106	-29.5	275	7.826	0.612	12.372	6.123
id385	76.00	-43.49	109.8	1995	0.124	-29.3	315	7.857	0.617	13.186	6.955
id423 *	76.62	-36.40	143.2	1993	0.093	-29.1	310	7.716	0.595	10.666	4.550
id514	77.25	-49.22	119.6	1995	0.163	-28.3	300	7.575	0.574	13.217	6.987
id531 *	77.45	-51.06	75.0	2009	0.198	-27.4	320	7.434	0.553	/	/
id534	80.00	-41.14	96.0	1994	0.105	-28.4	335	7.811	0.610	11.345	5.148
Basin8	69.80	-36.49	29.8	2003	0.350	-25.6	300	7.396	0.547	/	/
D2	71.80	-46.34	101.3	2003	0.421	-23.4	370	7.051	0.497	14.097	7.949
D4	71.39	-43.94	143.9	2003	0.390	-24.6	300	7.394	0.547	12.770	6.523
HumboldtM *	78.47	-56.98	141.9	1995	0.384	-24.8	280	8.062	0.650	10.947	4.794
NASAE1 *	74.98	-29.97	19.9	1997	0.135	-27.6	340	7.394	0.547	/	/
spencer6 *	72.57	-37.62	82.3	1994	0.176	-29.0	360	4.889	0.239	/	/
spencer16	71.75	-40.75	15.0	1954	0.289	-27.0	340	7.216	0.521	/	/
spencer17	77.95	-39.18	60.0	1973	0.080	-29.3	300	5.002	0.250	7.781	2.421
spencer66 *	70.75	-35.96	109.0	1987	0.247	-27.3	300	7.340	0.539	14.852	8.823
spencer67	70.63	-35.83	128.6	1988	0.262	-27.0	325	7.098	0.504	14.114	7.968
spencer68 *	70.65	-37.48	105.6	1988	0.263	-26.9	325	7.172	0.514	14.505	8.416
spencer69	70.67	-38.79	24.8	1988	0.252	-27.1	305	7.184	0.516	/	/
spencer70	70.64	-39.62	100.1	1988	0.262	-27.0	290	6.772	0.459	14.026	7.869
spencer71	71.76	-35.87	77.8	1988	0.203	-28.2	275	7.043	0.496	13.094	6.858
spencer72	71.48	-35.88	25.7	1988	0.207	-28.0	330	7.223	0.522	/	/
spencer73	71.15	-35.85	70.8	1988	0.214	-27.7	340	7.230	0.523	/	/
spencer74	70.85	-35.85	26.2	1988	0.264	-26.9	330	7.087	0.502	/	/
SouthPole	-90.00	0.00	122.9	2001	0.055	-47.8	325	7.613	0.580	22.312	19.913
Newall	-77.58	162.50	111.1	1989	0.043	-31.2	305	7.160	0.513	4.132	0.683
Berkner *	-79.61	-45.72	178.2	1995	0.124	-28.3	345	6.255	0.391	9.658	3.731
DML *	-71.41	-9.92	78.2	2007	0.902	-20.6	410	6.037	0.364	10.228	4.185

id9	-76.77	-101.74	111.6	2012	0.313	-24.7	470	6.194	0.384	12.119	5.875
id10	-76.95	-121.22	62.0	2011	0.213	-28.4	355	6.947	0.483	/	/
id11	-77.06	-89.14	114.5	2001	0.346	-26.5	415	5.879	0.346	11.201	5.019
id12	-77.61	-92.25	67.8	2001	0.301	-27.8	350	6.019	0.362	/	/
id13	-77.68	-124.00	59.3	2000	0.155	-28.2	350	6.411	0.411	/	/
id14	-77.76	153.38	97.1	2006	0.048	-44.6	360	6.833	0.467	17.516	12.272
id15 *	-77.84	-102.91	70.7	2001	0.486	-25.1	415	5.853	0.343	/	/
id17	-77.88	158.46	98.5	2006	0.058	-41.1	350	6.419	0.412	11.687	5.464
id18	-77.96	-95.96	57.4	2010	0.354	-28.0	335	6.752	0.456	/	/
id19	-78.08	-120.08	57.8	2000	0.171	-27.7	315	6.253	0.391	/	/
id20	-78.12	-95.65	70.5	2001	0.324	-27.7	385	6.265	0.393	/	/
id22 *	-78.33	-124.48	59.9	2000	0.152	-27.7	285	6.509	0.424	8.989	3.232
id24	-78.43	-115.92	59.8	2000	0.318	-27.8	390	6.295	0.396	/	/
id26	-78.73	-111.50	60.7	2000	0.329	-27.8	350	6.427	0.413	/	/
id28	-79.04	149.68	100.1	2006	0.040	-44.6	405	6.703	0.449	15.584	9.714
id29 *	-79.13	-122.27	63.1	2000	0.127	-27.8	300	6.507	0.423	9.926	3.941
id30	-79.16	-104.97	72.7	2001	0.306	-28.7	400	5.921	0.351	/	/
id33	-79.38	-111.24	104.8	2000	0.239	-28.2	370	6.159	0.379	12.943	6.701
id35 *	-79.48	-112.09	160.0	2011	0.162	-28.0	460	6.181	0.382	11.824	5.592
id39	-80.62	-122.63	57.5	1999	0.094	-25.9	370	6.253	0.391	/	/
id43	-81.20	-126.17	48.3	1999	0.070	-24.5	325	6.268	0.393	4.975	0.990
id46	-82.00	-110.01	62.2	2002	0.180	-27.8	340	6.161	0.380	/	/
id48	-83.50	-104.99	61.7	2002	0.220	-31.0	360	6.098	0.372	/	/
id49 *	-84.40	140.63	50.1	2007	0.023	-45.4	340	6.886	0.474	/	/
id50	-85.00	-105.00	44.9	2002	0.157	-36.3	360	6.422	0.412	/	/
id51	-85.78	145.72	41.7	2007	0.033	-46.1	310	6.767	0.458	/	/
id52 *	-86.50	-107.99	71.6	2002	0.147	-38.8	340	6.882	0.474	/	/
id53	-86.84	95.31	20.8	2003	0.042	-53.3	355	6.535	0.427	/	/
id54 *	-88.00	-107.98	54.1	2002	0.133	-41.4	355	7.009	0.491	/	/
id55	-88.51	178.53	99.3	2007	0.081	-48.2	320	6.880	0.473	/	/
id56	-89.93	144.39	139.5	2002	0.080	-48.6	345	6.319	0.399	25.046	25.092
spencer1	-80.00	-120.00	307.0	1968	0.120	-27.2	350	6.987	0.488	10.314	4.255
spencer4	-66.72	113.18	200.9	1989	1.060	-22.0	380	7.848	0.616	12.847	6.602
spencer5	-74.50	123.17	49.5	1980	0.037	-51.8	345	8.262	0.683	/	/
spencer7	-85.25	166.50	79.9	19997	0.028	-39.7	305	7.003	0.490	8.202	2.691
spencer8	-66.77	112.80	180.0	1997	0.488	-22.7	385	7.385	0.545	10.640	4.528

spencer22	-73.60	-12.43	25.5	1996	0.220	-22.5	380	3.920	0.154	/	/
spencer25	-74.02	-12.02	26.5	1996	0.171	-30.7	390	5.412	0.293	/	/
spencer29 *	-75.00	2.00	20.6	1996	0.072	-42.9	320	7.602	0.578	/	/
spencer33	-70.68	44.32	123.5	1978	0.114	-33.1	385	6.385	0.408	7.022	1.972
spencer34 *	-70.68	44.32	109.0	1978	0.114	-33.1	375	6.161	0.380	6.909	1.909
spencer61	-73.10	39.75	99.7	1978	0.069	-42.3	360	7.005	0.491	16.245	10.556
spencer62 *	-71.18	45.97	100.2	1997	0.091	-38.2	395	7.049	0.497	16.344	10.686
spencer76	-90.00	0.00	122.1	1997	0.055	-47.8	360	4.906	0.241	25.586	26.185
spencer77	-75.00	147.00	15.8	1961	0.042	-46.1	385	7.184	0.516	/	/
spencer78 *	-74.00	143.00	16.0	1961	0.043	-45.5	375	7.205	0.519	/	/
spencer79	-73.00	142.00	15.7	1961	0.057	-44.0	325	7.148	0.511	/	/
spencer80	-73.00	141.00	16.0	1961	0.057	-44.0	355	6.876	0.473	/	/
spencer81	-72.00	140.00	16.9	1961	0.080	-42.7	335	6.936	0.481	/	/
spencer82 *	-71.00	139.00	15.6	1961	0.120	-41.6	375	6.848	0.469	/	/
spencer83	-72.00	143.00	15.7	1961	0.087	-41.3	405	6.796	0.462	/	/
spencer84	-72.00	146.00	16.2	1961	0.086	-40.9	410	6.876	0.473	/	/
spencer85	-72.00	148.00	15.9	1961	0.096	-40.2	360	6.745	0.455	/	/
spencer86	-72.00	151.00	15.8	1961	0.103	-39.7	400	6.963	0.485	/	/
spencer87	-72.00	154.00	15.9	1961	0.130	-38.0	355	6.430	0.414	/	/
spencer88	-72.00	156.00	15.7	1961	0.130	-37.6	395	7.050	0.497	/	/
spencer89	-72.00	159.00	15.7	1961	0.115	-35.7	370	6.665	0.444	/	/
spencer90	-83.47	138.80	340.5	1994	0.020	-45.2	420	/	/	10.046	4.037
spencer91	-83.47	-138.80	47.0	1987	0.058	-27.1	295	7.037	0.495	3.530	0.499
spencer92	-78.47	106.80	179.3	1996	0.022	-54.6	360	8.790	0.773	20.368	16.594

Table S1. The 91 firn core dataset used in this study. * symbols indicate the core is part of the evaluation data. Lat and Lon designate latitude and longitude respectively. Year indicates the year of drilling of the core. $\dot{b}$ is the accumulation rate. T is the temperature. ρ_0 is the surface density boundary condition that was derived individually for each core by extrapolating density measurements until the surface (random noise is added to ρ_0 as discussed in Sect. S2). Var designates the site-specific variance used for the terms of Σ_{15} and Σ_{pc} (see Text ~~S2~~ S4 for their calculation). The core spencer90 has only a single density measurement above 15 m depth and its $DIP15$ is discarded.

	Parameters	Posterior mean	Posterior covariance matrix
HL	$k_0^*, k_1^*, E_0, E_1, a, b$	$[16.7, 649, 10760, 21000, 0.88, 0.66]$	$\begin{bmatrix} 34.4 & 40.2 & 4500 & 324 & -0.0685 & -0.0195 \\ 40.2 & 44000 & 618 & 161000 & 1.087 & -3.670 \\ 4502 & 618 & 710000 & 7080 & -29.95 & 1.94 \\ 324 & 1610000 & 7080 & 694000 & 7.86 & -27.51 \\ -0.0685 & 1.087 & -29.95 & 7.86 & 0.0051 & -0.0012 \\ -0.0195 & -3.670 & 1.94 & -27.51 & -0.0012 & 0.0036 \end{bmatrix}$
Ar	$k_0^{Ar}, k_1^{Ar}, E_g, \alpha, \beta$	$[0.080, 0.028, 40900, 0.78, 0.69]$	$\begin{bmatrix} 5.62\ 10^{-4} & 1.55\ 10^{-4} & -12.66 & 9.65\ 10^{-5} & -3.23\ 10^{-4} \\ 1.55\ 10^{-4} & 7.41\ 10^{-5} & -4.64 & -2.04\ 10^{-4} & 1.05\ 10^{-4} \\ -12.66 & -4.64 & 360000 & 11.0 & 4.67 \\ 9.65\ 10^{-5} & -2.04\ 10^{-4} & 11.0 & 3.30\ 10^{-3} & -1.01\ 10^{-3} \\ -3.23\ 10^{-4} & 1.05\ 10^{-4} & 4.67 & -1.01\ 10^{-3} & 3.12\ 10^{-3} \end{bmatrix}$
LZ	$lz_a, lz_b, lz_{11}, lz_{12}, lz_{13}, lz_{21}, lz_{22}, lz_{23}$	$[7.56, -2.091, -14.71, 7.269, -1.019, -1.513, 6.0203, -0.09127]$	$\begin{bmatrix} 5.27 & -0.198 & -1.20 & -1.68 & -0.0239 & 5.53\ 10^{-3} & -0.0606 & 4.13\ 10^{-3} \\ -0.198 & 0.0116 & 0.218 & -0.0612 & 0.0134 & -0.0158 & -2.29\ 10^{-3} & -7.37\ 10^{-4} \\ -1.20 & 0.218 & 14.6 & -3.96 & 0.801 & 0.368 & 0.354 & 0.0129 \\ -1.68 & -0.0612 & -3.96 & 13.3 & -0.309 & -0.0850 & 5.40 & 0.0166 \\ -0.0239 & 0.0134 & 0.801 & -0.309 & -0.0502 & -0.0173 & 0.0252 & -4.42\ 10^{-4} \\ 5.53\ 10^{-3} & -0.0158 & 0.368 & -0.0850 & -0.0173 & 0.446 & -0.429 & 0.0131 \\ -0.0606 & -2.29\ 10^{-3} & 0.354 & 5.40 & 0.0252 & -0.429 & 3.94 & -2.59\ 10^{-4} \\ 4.13\ 10^{-3} & -7.37\ 10^{-4} & 0.0129 & 0.0166 & -4.42\ 10^{-4} & 0.0131 & -2.59\ 10^{-4} & 4.80\ 10^{-4} \end{bmatrix}$

Table S2. The posterior means and covariance matrices for the free parameters of HL, Ar and LZ. These statistics can be used to generate random parameter combinations following a normal approximation.

~~cmp_{ant}^{00-17} [m]~~	~~HL~~	~~MAP_{HL}~~	~~Ar~~	~~MAP_{Ar}~~	~~LZ~~	~~MAP_{LZ}~~
~~EGRIP~~	~~0.626~~	~~0.648~~	~~0.735~~	~~0.656~~	~~0.675~~	~~0.695~~
~~Summit~~	~~0.464~~	~~0.461~~	~~0.442~~	~~0.442~~	~~0.481~~	~~0.475~~
~~id359~~	~~0.748~~	~~0.763~~	~~0.866~~	~~0.781~~	~~0.815~~	~~0.837~~
~~id369~~	~~0.609~~	~~0.587~~	~~0.597~~	~~0.574~~	~~0.623~~	~~0.617~~
~~id373~~	~~0.613~~	~~0.637~~	~~0.735~~	~~0.649~~	~~0.664~~	~~0.693~~
~~id385~~	~~0.319~~	~~0.333~~	~~0.407~~	~~0.356~~	~~0.373~~	~~0.393~~
~~id423~~	~~0.528~~	~~0.559~~	~~0.649~~	~~0.566~~	~~0.565~~	~~0.600~~
~~id514~~	~~0.358~~	~~0.367~~	~~0.409~~	~~0.378~~	~~0.409~~	~~0.423~~
~~id531~~	~~0.336~~	~~0.333~~	~~0.304~~	~~0.328~~	~~0.373~~	~~0.371~~
~~id534~~	~~0.216~~	~~0.214~~	~~0.189~~	~~0.205~~	~~0.229~~	~~0.237~~
~~Basin8~~	~~0.764~~	~~0.760~~	~~0.728~~	~~0.671~~	~~0.641~~	~~0.611~~
~~D2~~	~~0.244~~	~~0.255~~	~~0.222~~	~~0.164~~	~~0.158~~	~~0.133~~
~~D4~~	~~0.560~~	~~0.553~~	~~0.550~~	~~0.465~~	~~0.500~~	~~0.479~~
~~HumboldtM~~	~~0.630~~	~~0.616~~	~~0.606~~	~~0.534~~	~~0.562~~	~~0.545~~
~~NASAE1~~	~~0.591~~	~~0.568~~	~~0.578~~	~~0.555~~	~~0.605~~	~~0.596~~
~~spencer6~~	~~0.380~~	~~0.386~~	~~0.391~~	~~0.367~~	~~0.393~~	~~0.399~~
~~spencer16~~	~~0.499~~	~~0.489~~	~~0.500~~	~~0.428~~	~~0.462~~	~~0.448~~
~~spencer17~~	~~0.503~~	~~0.534~~	~~0.611~~	~~0.538~~	~~0.536~~	~~0.566~~
~~spencer66~~	~~0.963~~	~~0.944~~	~~0.955~~	~~0.912~~	~~0.926~~	~~0.889~~
~~spencer67~~	~~0.832~~	~~0.812~~	~~0.821~~	~~0.762~~	~~0.761~~	~~0.744~~
~~spencer68~~	~~0.895~~	~~0.875~~	~~0.895~~	~~0.831~~	~~0.838~~	~~0.811~~
~~spencer69~~	~~0.935~~	~~0.915~~	~~0.920~~	~~0.880~~	~~0.895~~	~~0.857~~
~~spencer70~~	~~1.078~~	~~1.056~~	~~1.060~~	~~1.017~~	~~1.025~~	~~0.982~~
~~spencer71~~	~~1.094~~	~~1.066~~	~~1.126~~	~~1.042~~	~~1.107~~	~~1.093~~
~~spencer72~~	~~0.809~~	~~0.806~~	~~0.852~~	~~0.780~~	~~0.812~~	~~0.799~~
~~spencer73~~	~~0.793~~	~~0.786~~	~~0.835~~	~~0.756~~	~~0.782~~	~~0.763~~
~~spencer74~~	~~0.856~~	~~0.842~~	~~0.861~~	~~0.785~~	~~0.786~~	~~0.771~~

~~**Table S2.** Total compaction anomaly over 2000-2017 (cmp_{ant}^{00-17}) at the 27 GrIS sites. See text S3 for calculation details.~~

Figures

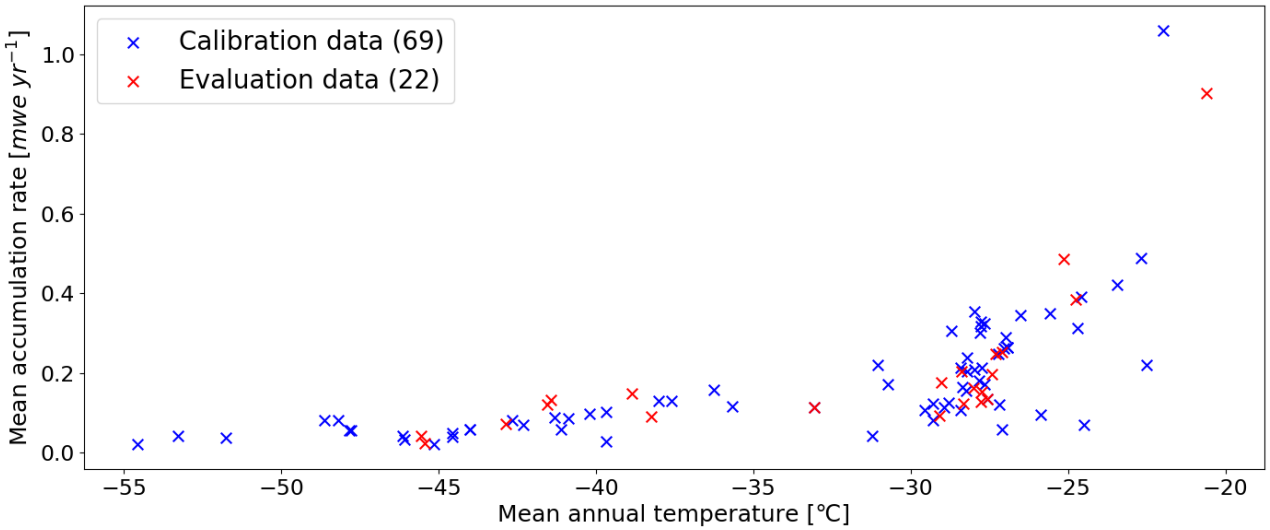


Figure S1. Climatic conditions at the 91 sites of the dataset

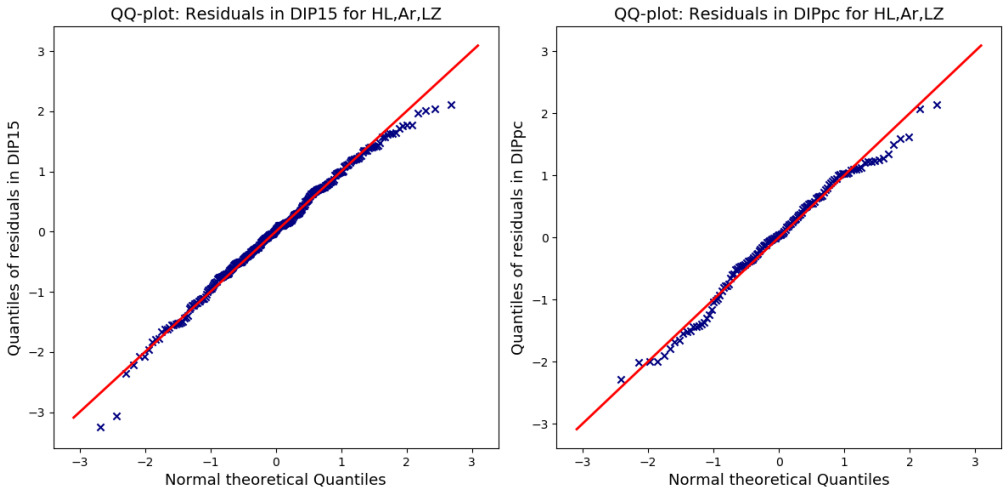


Figure S2. Quantiles-Quantiles plots for the errors of the three original models (HL, Ar, LZ) computed on the entire dataset. The alignment of the points along the red line informs about the fit to a normal distribution.

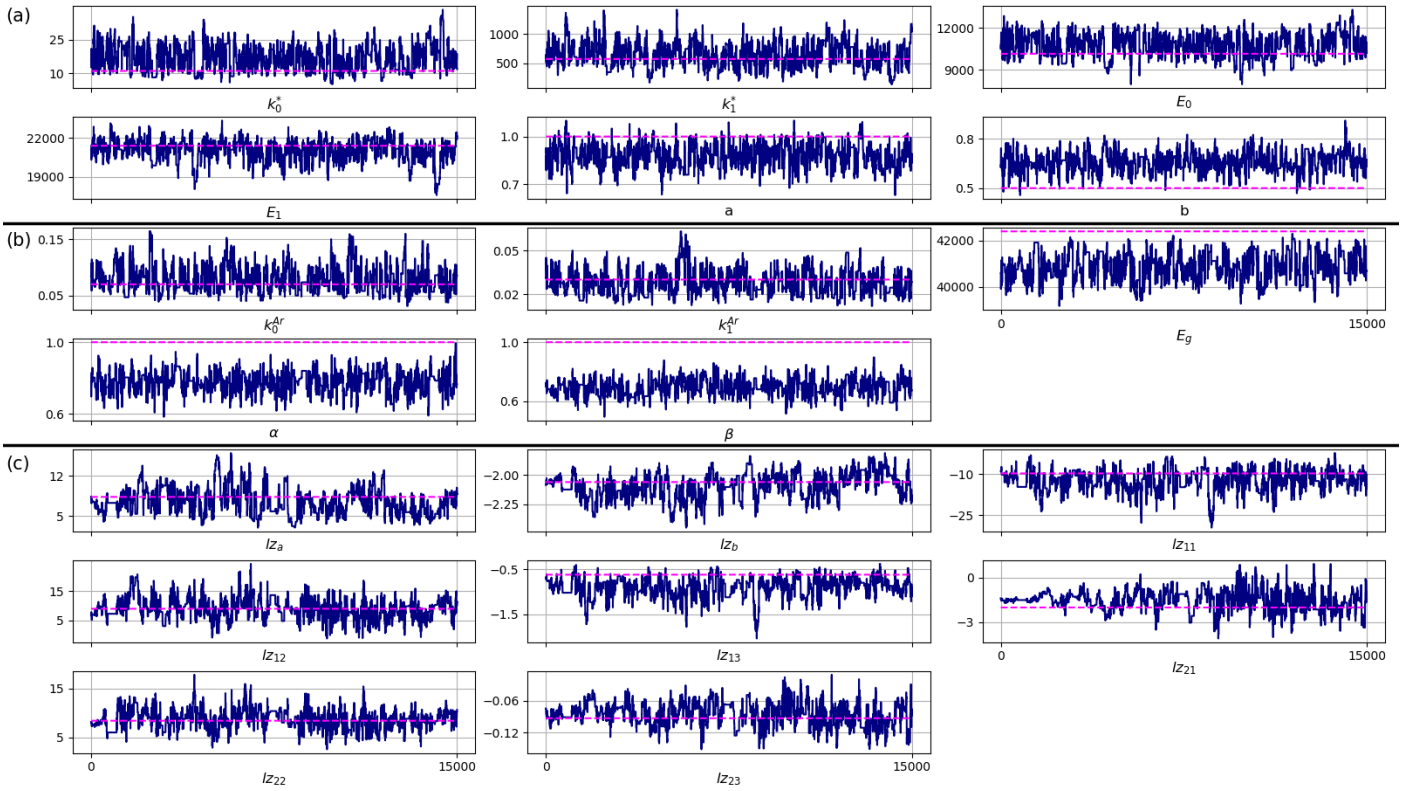


Figure S3. Sampling chains of each parameter for (a) HL, (b) Ar, (c) LZ. The x-axis displays the iteration number, the y-axis displays the parameter value. The dashed pink line shows the value of the original model, which is also the starting point of each chain.

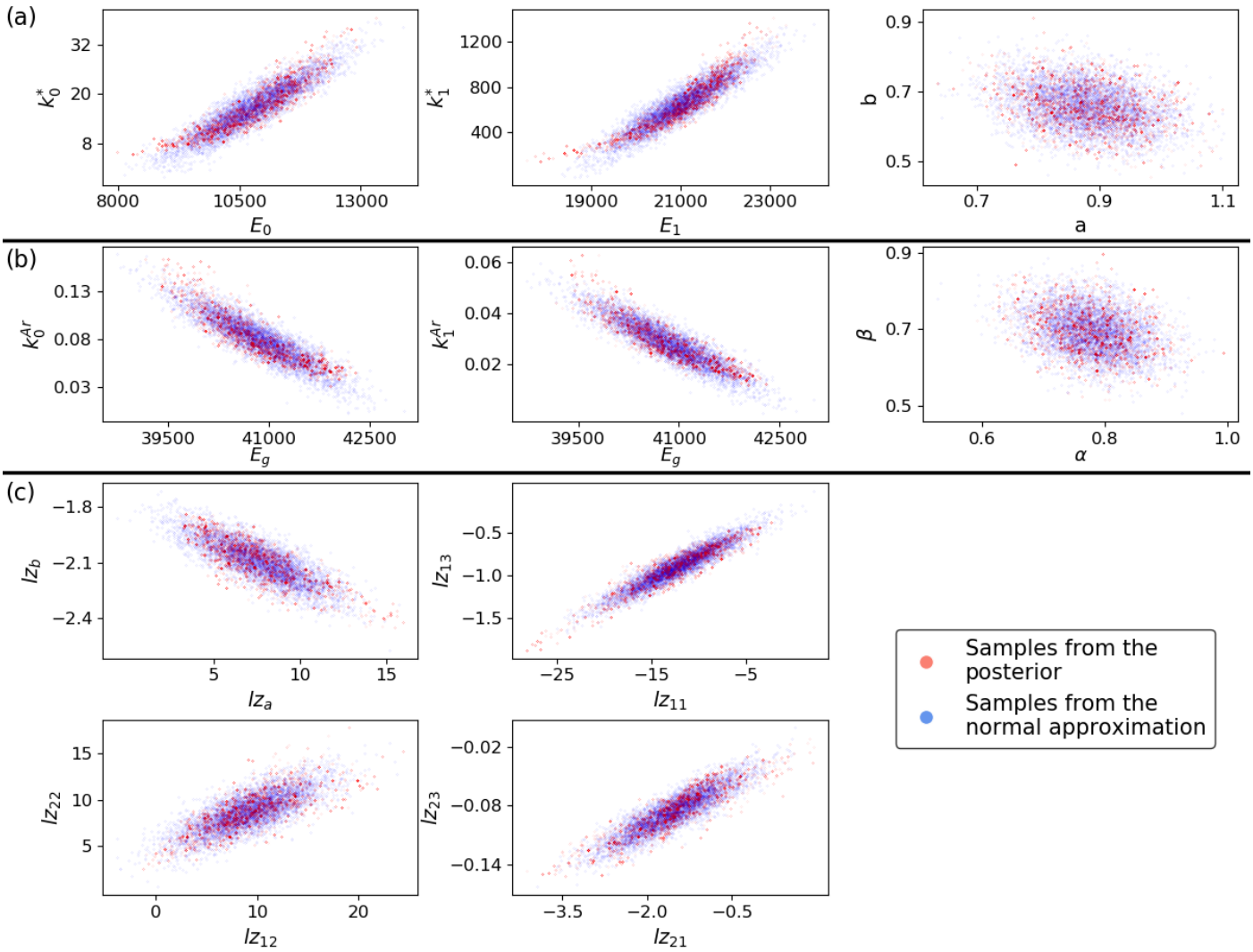


Figure S4. Evaluation of the normal approximations to the posterior distributions for (a) HL, (b) Ar, (c) LZ. Where possible, correlated parameters share a same graph.

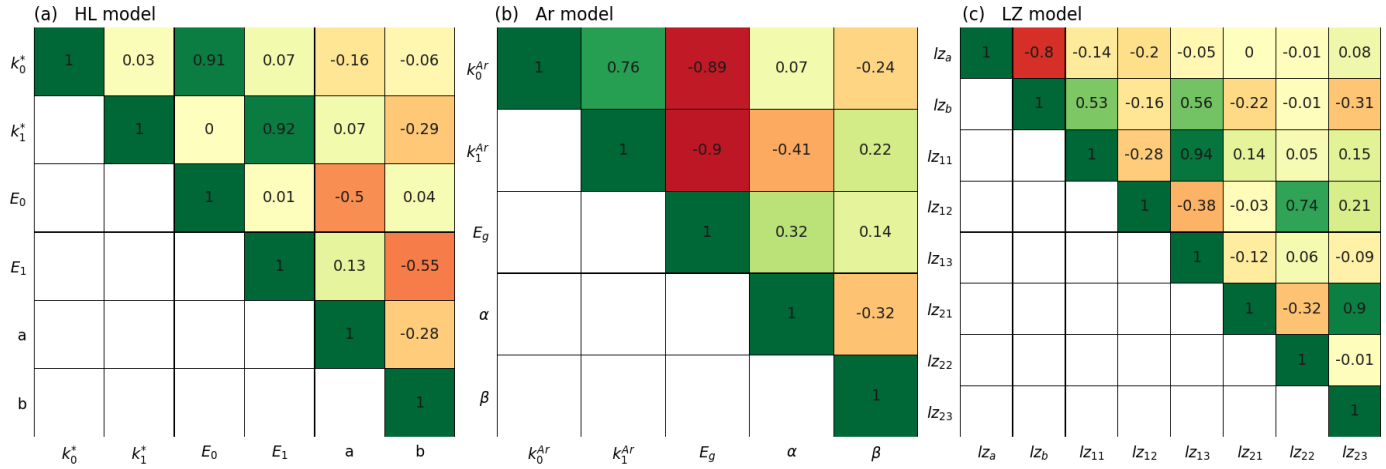


Figure S5. Posterior correlation matrices.