Reply to the comments by Referee # 1

We would like to thank the referee for the valuable comments and suggestions. In the following, the comments by the referee are listed in Italic, and our reply is provided for each comment in Roman.

Comment:
. . . Indeed different results are to be expected, but it would be good to think of some criterion to evaluate the performance of the models in an objective way. In which way is the proposed model better than previous ones? External validation of the results would perhaps be a convincing way to promote the approach, and here are a few things that come to mind.

- *The prior distribution on the parameters should be very explicitly described (all the results depend on it). Then, the posterior distribution could be compared to the prior, for instance using overlaid kernel density plots. This would give a visualization of how much information is gathered on the parameters from the data. Perhaps some parameters are easier to estimate than others?*

Reply: We omitted to specify the prior distribution. We appreciate the referee for pointing out that.
In this paper, a uniform distribution is used as the prior distribution of each parameter. If we want to use a different prior distribution, the posterior distribution can be obtained as a product between the prior distribution and the histogram in Figure 1.

Comment:

- *A simulation study on synthetic data generated from the model would also be informative. How precisely can we identify the model parameters using synthetic dataset (using the same number of observations as in the real dataset)?*

Reply: In this paper, we consider a situation where the model is an approximation of the actual process. If we generate a synthetic dataset under given parameters, the spread of the parameters would be very small, which is highly different from the actual situation. We have no idea how we can appropriately generate a synthetic dataset good for a benchmark of our method.

Comment:

- *The model quality could be evaluated based on its predictive performance: for instance, the parameters could be inferred using the first 80% data points, and the remaining 20% data points could be predicted. Certainly other criteria could be envisioned, such as Bayes factors. In fact the statistical literature is quite rich on this topic (see [6]).*

Reply: We tried the estimation without using the last five agemarkers, i.e., we used the first 80% of the age markers and $\delta^{18}O$ data. The followings are the age–depth relationship and the accumulation as a function of age estimated without using the last five agemarkers. The estimate using all the age markers is also indicated with grey lines.
The estimate of the age as a function of depth was very slightly different near the bottom. The variation of the estimated accumulation–age relationship also showed a small time shift between the estimate without using the last five agemarkers and that using all the markers. However, the difference was mostly within the uncertainty between the 10th and 90th percentile. Thus, this difference near the bottom would be acceptable.
In the PMCMC, the likelihood of the parameter vector $\boldsymbol{\theta}$, $p(\boldsymbol{y}_{1:Z}|\boldsymbol{\theta})$, is evaluated at each iteration of the MCMC procedure. In other words, the predictive performance for a given $\boldsymbol{\theta}$ is evaluated at each step of the MCMC. Thus, we think the procedure of the PMCMC would provide a good choice of the parameter in terms of the predictive performance.

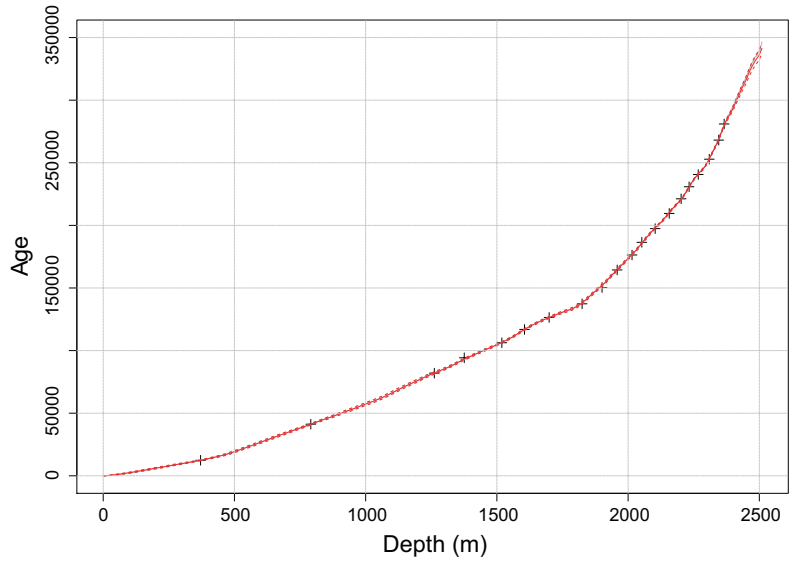

Fig. A. Estimated age without using the last five agemarkers (red lines) and estimate using all the agemarkers (grey). Each solid line indicates the median of the posterior distribution. The 10th and 90th percentiles of the posterior are indicated by dotted lines.

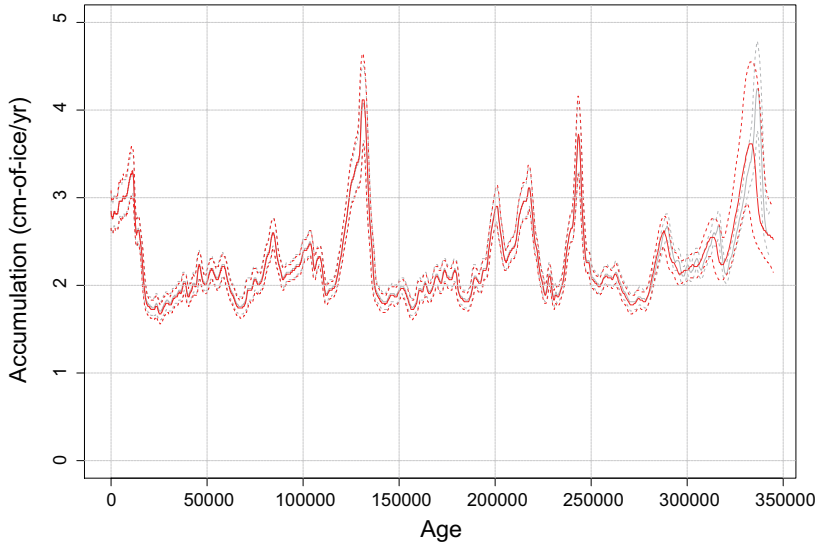

Fig. B. Estimated accumulation rate as a function of age without using the last five agemarkers (red lines) and estimate using all the agemarkers (grey).

Comment:

- *Some aspects of the model seem more arbitrary than others: Gaussian distributions for the noise distributions, the accumulation rate is a random walk process (why not an autoregressive process?). The article could either justify these modelling assumptions in more details, or test various model modifications in practice. The resulting models could be compared, again, using predictive criteria or Bayes factors.*

Reply: It is true we can use various noise distributions and we can consider various models for the accumulation rate. We could select among them by using some metric such as Bayes factors. However, there are a large number of choices, and thus it would take much time to make the selection among those choices. In this revision, we will just add a mention about the fact that we can consider various models and we can select among them by using some metric such as Bayes factors. We will examine the performance of other models in the future works.

Comment:
· The language is clear. The general descriptions of the model and of the methods are fine, but the article should allow readers to reproduce the results; it is not the case here by lack of implementation details (lack of details on the prior distribution), details on the proposal distribution $q(\theta'|\theta)$, etc). Perhaps an appendix could give all the values used in the implementation that are not specified in the main text.

Reply: It is true that we omitted to provide the information on the proposal distribution for the MCMC (Metropolis method) part. In this paper, a zero-mean Gaussian distribution is used as the proposal distribution for each parameter. The variance of the proposal distribution for each parameter is given in Table A. As described above, the prior distributions of the parameters are uniform distributions.

Comment:
· Inconsistent notation: $\delta^{18}\mathrm{O}$ or $\delta^{18}\mathrm{O}_z$ or $\delta^{18}\mathrm{O}(z)$ data.

Reply: We will unify thos expressions into "δ^{18}O data".

Comment:
· State space models are called "sequential Bayesian models" in the article, which is non-standard and a bit misleading, because nothing is really "Bayesian" about them (Bayes formula is just used to obtain the recursion formula for the filtering distributions). "Bayesian" usually refers to inference methods treating parameters as random variables, and does not refer to models. Hence, non-Bayesian approaches could have been applied to the model of the article. Another common term for state space models is "hidden Markov models".

Reply: It is true that the word "sequential Bayesian models" was not appropriate. We refer to it as "state space model".

Comment:
· The model description is split into Section 2 & 3, starting in "continuous time" and with the description of Θ_z (Section 2), and then switches to discrete time and to the description of A_z and of the measurement distributions (Section 3). These sections could perhaps be combined in one section.

Reply: Section 2 is intended to review the glaciological model proposed by the existing study (Parrenin et al., 2007). Section 3 is intended to formulate a state space model based on the glaciological model in Section 2, which might be common in the glaciological community. That is the reason why we divided into two sections.

Comment:

· *page 945: Equation (14) should read*

$$p(\xi_{z+1}|\xi_z, \boldsymbol{\theta}) = \mathcal{N}\left(\xi_z + \frac{1}{A_z\Theta_z},\ \sigma_\nu^2\right), \tag{1}$$

according to Equation (12)...?

Reply: We appreciate the referee for this correction. We will correct Eq. (14).

Comment:

· *page 946: when there are multiple observations δ^{18}O within an interval of one meter, a mean is used (presumably, without modifying the standard deviation σ_w). This seems unfair, as when there are more observations, the uncertainty should be reduced. One simple approach would be to use the mean of the observations at each meter, but with a variance σ_w divided by the number of observations.*

Reply: The time sequence of δ^{18}O contains short-term fluctuations. These short-term fluctuations are regarded as noises, which are difficult to model. However, they have short-term auto-correlation, and δ^{18}O within one meter interval usually takes similar values. Therefore, even if a mean of multiple observations within one meter interval is used, it would not be appropriate to divide σ_w by the number of observations.

Comment:

· *Again, the prior distribution on the parameter θ should absolutely be specified somewhere.*

Reply: As described above, we use a uniform distribution as the prior distribution of each parameter. We will add a mention on the prior distribution in the revised manuscript.

Comment:

· *page 948: why isn't σ_ε included in the parameter θ? More details should be given on this. Does the method fail if this parameter was included in θ? What are the values given to it, in the end?*

Reply: The standard deviation σ_ε is provided together with the age marker data by Kawamura et al. (2007). The following table (Table A) shows the depth and age for each age marker (tie point) as well as σ_ε.

Comment:

In the proposed model, the transition is non-linear (because of the term $1/A_z\Theta_z$) but the noise distributions are Gaussian (if we use the parametrization $x_z = (\xi_z, \log A_z)$ instead of $x_z = (\xi_z, A_z)$). Thus, a "locally optimal" particle filter approach could be implemented, that is, instead of propagating the particles using $p(x_{z+1}|x_z, \theta)$ and weighting using $p(y_{z+1}|x_{z+1}, \theta)$, one could sample from $p(x_{z+1}|x_z, y_{z+1}, \theta)$ and weight the particles using $p(y_{z+1}|x_z, \theta)$; these two distributions are Gaussian. This is called the optimal proposal scheme in [1]; it could reduce the variance of the likelihood estimator.

Reply: As mentioned in Concluding Remarks, we are considering to improve the proposal scheme in the future. However, we are also considering to extend the model and possibly we may use non-Gaussian distribution for the noise distribution. That is the reason why we have not yet tuned the proposal scheme for the SMC part.

Comment:

page 951, line 11: "using the SMC" + algorithm ?

Reply: Right. We meant the SMC algorithm.

Depth	Age	Uncertainty of the age ($2\sigma_\varepsilon$)
371.00	12390	400
791.00	41200	1000
1261.61	81973	2230
1375.67	94240	1410
1518.91	106263	1220
1605.27	116891	1490
1699.17	126469	1660
1824.80	137359	2040
1900.74	150368	2230
1958.31	164412	2550
2015.01	176353	2880
2052.23	186470	2770
2103.14	197394	1370
2156.67	209523	1980
2202.02	221211	890
2232.45	230836	780
2267.28	240633	1230
2309.35	252866	1160
2345.32	268105	1980
2366.01	280993	1600
2389.31	290909	1210
2412.25	301628	880
2438.37	313205	840
2462.36	324774	1110
2505.4	343673	2000

Table A. The depth, the age, and the uncertainty of the age at each tie point.

Comment:
page 952, line 5: perhaps give the formula for the likelihood estimator, since it is quite central in the particle MCMC method?

Reply: We thank the referee for the suggetion. We approximate $p(\boldsymbol{y}_z|\boldsymbol{y}_{1:z-1},\boldsymbol{\theta})$ as follows:

$$
\begin{aligned}
&p(\boldsymbol{y}_z|\boldsymbol{y}_{1:z-1},\boldsymbol{\theta}) \\
&= \int p(\boldsymbol{y}_z|\boldsymbol{x}_z,\boldsymbol{\theta})\, p(\boldsymbol{x}_z|\boldsymbol{y}_{1:z-1},\boldsymbol{\theta})\, d\boldsymbol{x}_z \\
&= \int p(\boldsymbol{y}_z|\boldsymbol{x}_z,\boldsymbol{\theta})\, p(\boldsymbol{x}_{0:z}|\boldsymbol{y}_{1:z-1},\boldsymbol{\theta})\, d\boldsymbol{x}_{0:z} \\
&\approx \frac{1}{N}\sum_{i=1}^{N} \int p(\boldsymbol{y}_z|\boldsymbol{x}_z,\boldsymbol{\theta})\, \delta\left(\boldsymbol{x}_{0:z} - \boldsymbol{x}^{(i)}_{0:z|z-1}\right) d\boldsymbol{x}_{0:z} \\
&= \frac{1}{N}\sum_{i=1}^{N} p(\boldsymbol{y}_z|\boldsymbol{x}^{(i)}_{0:z|z-1},\boldsymbol{\theta}),
\end{aligned}
$$

where we used the assumption introduced in Page 947:

$p(\boldsymbol{y}_z|\boldsymbol{x}_{0:z},\boldsymbol{\theta}) = p(\boldsymbol{y}_z|\boldsymbol{x}_z,\boldsymbol{\theta})$.

We then approximate the logarithm of $p(\boldsymbol{y}_{1:Z}|\boldsymbol{\theta})$:

$$
\log \hat{p}(\boldsymbol{y}_{1:Z}|\boldsymbol{\theta}) = \sum_{z=1}^{Z} \log\left[\frac{1}{N}\sum_{i=1}^{N} p(\boldsymbol{y}_z|\boldsymbol{x}^{(i)}_{0:z|z-1},\boldsymbol{\theta})\right].
$$

Comment:

page 952, not much details is given on the tuning of the proposal distribution $q(\theta'|\theta)$. How is the variance tuned? Using preliminary runs?

Reply: We performed some preliminary runs to find out the landscape of the posterior distribution. Then, the width of $q(\theta'|\theta)$ was taken to be small enough in comparison with the width of the target posterior distribution.

Comment:

page 952, Equation (35) and onwards: it is not very clear that only the likelihood estimator $\hat{p}(y_{1:Z}|\theta^*)$ on the numerator is calculated at each step, and that the one in the denominator is kept fixed. The method would not be valid if both the numerator and the denominator estimators were drawn at each step.

Reply: The denominator is kept fixed at each step. We will modify the desciption.

Comment:

page 953, line 13: "this greatly reduces the computational cost": does this refer to the memory cost instead of the computational cost? Is the memory cost a problem here?

Reply: We agree we should revise the description more concretely. It reduces the computational time because it can skip some processes for handling the whole sequence of 2510 steps ($Z = 2510$ in this paper) for 5000 particles. However, it is also true that the memory cost is also essential. If 5,000 particles for the whole sequence of 2510 steps of two variables are retained for all of the 50,000 MCMC steps, a TB-sized memory would be required.

Comment:

page 953, line 13: "$p(y_{1:k}|\theta)$" should be $p(y_{1:z}|\theta)$?

Reply: We thank the referee for the correction.

Comment:

End of section 4: perhaps mention other particle MCMC methods. In particular, some variations such as particle Gibbs, and particle Gibbs with ancestor sampling (see [3]), would be applicable here and could significantly improve the performance.

Reply: We will add a mention on other particle MCMC methods. We appreciate the referee for the comment.

Comment:

Section 5: some comments could be made on the correlations between the components of the posterior distribution. If they are not close to zero, perhaps some pairwise scatter plots would be informative.

Reply: It is true some of the parameters are closely correlated with each other. For example, the two parameters for Eq. (19), a and b, have an anti-correlation as shown in the following figure (Figure C) We will add some two-dimensional histograms.

Comment:

Section 5: some indication that the Markov chains have mixed would be appreciated, for instance using traceplots instead of histograms. There is no indication in the text that multiple chains, with the same tuning parameters and starting from various points, lead to similar results. By the way, how were the Markov chains initialized? And how long was the burn-in period? Why was a sample kept every fifth iteration and not at every iteration?

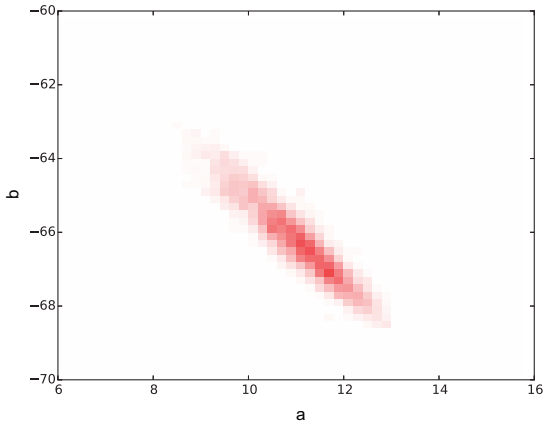

Fig. C. Two dimensional histogram for the joint posterior distribution of a and b.

Reply: As described above, we performed some preliminary runs to find out the landscape of the posterior distribution. The initial point of a Markov chain is determined around the center of the posterior indicated by the preliminary runs.

We kept every fifth iteration in order to reduce the computational time. Actually, since each sample typically has a high correlation with some subsequent MCMC samples, the estimate would not get worse even if four samples are discarded for every five steps.

Comment:
On the dataset: how large is it? Can it be plotted in some way? Can it be downloaded somewhere? It seems that the maximum depth is $Z = 2,500m$, and that there are a few dozen age markers (from Figure 2); it should be described in the text.

Reply: We use 25 age markers as shown above. The δ^{18}O data are published by Watanabe et al. (2003). We will add a plot of the δ^{18}O data. As mentioned above, Z is taken to be 2510 (m) in this paper. It will also be described in the revised manuscript.

Comment:
page 956, on the computational cost: there should be some mentions of parallel computing, which could make 250,000 iterations with 5,000 particles much faster to run than 1,250,000 iterations with 1,000 particles. There is a rich literature on how to implement particle filters on parallel computing hardware.

Reply: It is true that the computation with 5000 particles can be much faster than that with 1000 particles if we use a parallel computer having larger than 1000 processors. We will add a description on this point.

Comment:
Figure 7-9 could be replaced by traceplots of the chains, starting from a few initial points, and plots of the average acceptance rates against number of particles, for a fixed proposal $q(\theta'|\theta)$.

Reply: We think the histograms would be striking to check the convergence. The average acceptance rates were 0.03, 0.12, and 0.17 with 1000, 3000, and 5000 particles, respectively. However, the average acceptance rates will be described.

References

– Watanabe, O., Jouzel, J., Johnsen, S., Parrenin, F., Shoji, H., and Yoshida, N: Homogeneous climate variability across East Antarctica over the past three glacial cycles, Nature, 422, 509–512, 2003.