RESPONSE TO REVIEWER # 2

COMMENT # 1.1

> *I am very enthusiastic about this paper, having experienced PBS ensemble collapse in assimilation of frequent and over-confident observations myself, and it has potential applications beyond cryospheric data assimilation. My comments are mostly more in the nature of discussion than requirements for revisions.*

Reply:

We thank Richard Essery for providing this helpful review with encouraging feedback and constructive comments on our manuscript. Please find below a point-by-point response to all comments from the reviewer, including changes made. These revisions have undoubtedly helped to improve the manuscript.

COMMENT # 1.2

> *The broad review of assimilation methods relevant to the subject is interesting, but could possibly be shortened, and it is a little repetitive in places (e.g. MCMC is referred to as "gold standard" 15 times). There is much less in general terms on the nature of filter collapse that is central to this paper, but the simple illustration in Figure 2 could provide insight. I think that the PBS mean is suboptimal because the necessarily broad prior does not efficiently sample the narrow posterior, and the degeneracy is because of the prohibitive cost in Equation 4 for all except the nearest particle.*

Reply:

We thank the reviewer for these suggestions. We agree that the repeated use of the term gold-standard to describe the MCMC samples was redundant. We have now cut this down from 15 to only 5 uses of this particular term to avoid repetition.

As for the review of assimilation methods, we have decided to retain this section to maintain the necessary background material. While we acknowledge that it is extensive, it synthesizes disparate concepts from the wider computational Bayesian literature into a single more accessible overview tailored toward cryospheric data assimilation. Because this is a 'Development and technical paper' in GMD focused specifically on advancing DA methodology, we believe this technical background is both appropriate and necessary for readers less familiar with DA. It should be especially helpful for the less standard cryospheric DA algorithms that we utilize in the paper, i.e., iterative ensemble Kalman, adaptive particle, and MCMC methods.

The reviewer's understanding of the origin of degeneracy (i.e., collapse) when performing importance sampling as visualized in Figure 2 is spot-on. Depending on the expectations being estimated, the optimal proposal is generally much closer to the posterior than the prior [MacKay, 2003, Rainforth et al., 2020]. It is thus unfortunate that, for simplicity, the vast majority if not all particle methods used to date for cryospheric DA [e.g. Charrois et al., 2016, Oberrauch et al., 2024], including the PBS [Margulis et al., 2015], are variations on the seminal bootstrap particle filter [Gordon et al., 1993] in the sense that the prior is used (generally implicitly) as a proposal distribution. We have now extended our discussion of this issue by foregrounding the needle and haystack analogy, based on a similar example in MacKay [2003], placing it already in the introduction, placing it in the context of particle degeneracy, and including a key reference by Morzfeld et al. [2017]:

Changes:

However, issues with their practical implementation can make these methods problematic in certain settings. It is well known that if the proposal distribution, typically the prior, differs significantly from the target ~~distribution then the probability tends to collapse to~~ posterior distribution [MacKay, 2003, Rainforth et al., 2020], then all probability mass risks collapsing to a few or even a single particle, a problem known as particle degeneracy or ensemble collapse ~~[van Leeuwen et al., 2019, Murphy, 2023]~~ [Snyder et al., 2008, Morzfeld et al., 2017, Murphy, 2023]. Degeneracy results in suboptimal approximate Bayesian inference by greatly degrading uncertainty quantification in the posterior simulations. This issue is aggravated by incorporating highly informative (i.e., numerous and/or very precise) observations~~and~~ , as well as when working in higher dimensional state and parameter spaces [van Leeuwen et al., 2019]. Drawing on MacKay [2003], a useful analogy is looking for a needle (target posterior) in a haystack (prior proposal): a larger haystack (broader, higher-dimensionality) or a smaller needle (bigger data, more informative observations) exponentially reduces the chance of hitting the target needle when probing the haystack. While this needle in a haystack effect is a general problem, particle methods are especially vulnerable to it due to high Monte Carlo variance induced by the mismatch between the proposal and target [MacKay, 2003, Rainforth et al., 2020], challenges with localization [Farchi and Bocquet, 2018], and generally poor scaling with high effective dimensions [Snyder et al., 2008, Morzfeld et al., 2017].

Reply:

We have also added a new section 4.3 providing recommendations for applying and extending AdaPBS, where we also discuss collapse in the context of model complexity in our experiments as follows:

Changes:

Experiments 1 and 2 demonstrate the role of problem complexity. For the simpler temperature index model with two parameters, AdaPBS exceeded the desired threshold of $\tau = 0.3$ (i.e., $N_{eff} = 30$ with $N_e = 100$) with $N_{eff} = 84$ in 5 iterations, providing a strong approximation that nearly converged already at 4 iterations (Figure **??**). In Experiment 2, the intermediate complexity snow model FSM2, a higher dimensional parameter space, and numerous observations conspired to make inference challenging [Snyder et al., 2008, Morzfeld et al., 2017]. Following the 'needle in a haystack' analogy for inferring the posterior: increased model complexity may result in an oddly shaped needle, more informative data makes the needle smaller, and higher dimensionality makes the haystack bigger. Thus, AdaPBS required, on average, nearly twice as many iterations (8) to converge with the same $\tau = 0.3$ as Experiment 1. For both experiments, the number of iterations could have been reduced by lowering the bar on diversity to say $\tau = 0.1$, which can suffice in operational settings focused on prediction using point estimates rather than uncertainty quantification [e.g. Oberrauch et al., 2024] .

COMMENT # 1.3

Isn't the mode of the lognormal distribution for the precipitation multiplier in 3.2 equal to 0.86, i.e. not centred on being unbiased? It won't make much difference, but why is a logit-normal distribution chosen instead for perturbing other strictly positive forcing variables in 3.3 (it would make sense for relative humidity, but that is not listed as a perturbed variable)?

Reply:

We thank the reviewer for pointing out this inconsistency. It is true that strictly speaking the mode of a lognormal with associated normal parameters $\mu = 0.1$ and $\sigma = 0.1$ is 0.86, while the mean is 1.25 and the median is 1.10. So the mode itself does not correspond exactly to an unbiased multiplier (i.e. a multiplier of 1), nor do the other two location parameters. They do however al straddle 1 as was our intention. To clarify this and other aspects of the prior we have modified the paragraph as follows:

Changes:

For simplicity, and without loss of generality for the DA, we assume *a priori* parameter independence, such that the joint prior factorizes into the product of its marginal prior distributions. This assumption can be relaxed using additional background knowledge on parameter dependence [e.g., Pirk et al., 2022, Alonso-González et al., 2023] . Yet, even with prior independence, correlation between parameters can be inferred in the posterior via the likelihood [Cao et al., 2025]. The prior probability ~~distributions used to generate the perturbation parameters are~~ distribution for the additive air temperature bias is a normal distribution with mean $\mu_b = 0$ and standard deviation $\sigma_b = 1$~~for the air temperature bias and~~. For the prior on the multiplicative precipitation scaling, we use a strictly positive lognormal distribution with an associated normal mean $\mu_c = 0.1$ (so the actual median perturbation in model space is ~~$\exp(\mu_c) = 1.10$~~$\exp(\mu_c) = 1.11$) and standard deviation $\sigma_c = 0.5$~~for precipitation scaling~~. Note that we carry out inference directly on the unbounded space of the log-transformed precipitation perturbation parameter and then apply an exponential transform to map this back to the model parameter c following common practice [Gelman et al., 2013, Alonso-González et al., 2022]. The ~~modes~~ location parameters of these prior distributions were chosen in a conservative ~~(i.e.centered on being unbiased) way, imitating a real case where in practice it is not possible to know which prior mode might be close to the true forcing bias~~ manner, i.e., near unbiased, with a mean temperature bias of 0 and a precipitation scaling centered close to 1. This corresponds to quite a general case where, without additional background knowledge or data, the orientation of the forcing bias is not known *a priori*. The relatively large prior spread parameters (σ_c and σ_b) encode considerable initial epistemic uncertainty ~~. The air temperature perturbation parameter is an additive bias term, whereas the precipitation perturbation parameter becomes multiplicative after transforming back to the model space using the exponential function~~in the two forcing perturbation parameters.

COMMENT # 1.4

The banana-shaped posterior distribution in Figure 2 is the equifinality which is familiar (because easily visualized) in calibration of temperature index models. A hydrologist seeking to calibrate a 2-parameter model would not consider this challenging at all; a better solution than the degenerate PBS in Figure 2 could be found efficiently by any number of standard methods.

Reply:

We thank the reviewer for this comment and providing the possibility to discuss more traditional calibration approaches from hydrology. This is indeed the classical equifinality result for temperature index models, building off the equifinality thesis of Beven [2006] which we now cite. We have now added the following to Section 4.1 to put the PBS results in the wider context of hydrological model calibration.

Changes:

It may seem surprising that importance sampling via PBS struggles to accurately capture the reference posterior in this apparently simple experiment with a two-dimensional parameter space. An alternative would be gradient-based optimization techniques, such as gradient descent or Newton's method, used in both machine learning [Murphy, 2023] and variational DA [Bannister, 2017], to identify the mode with relatively few iterations and find a Laplace approximation of the posterior [MacKay, 2003]. However, as is often the case in practical DA in the absence of automatic differentiation [Murphy, 2023] , we do not have access to gradients and are focused on the harder problem of 'black-box' (i.e., gradient-free) inference that is naturally handled by ensemble-based methods [Sanz-Alonso et al., 2023]. While similar low-dimensional calibration problems are routinely encountered in hydrology and can be tackled via methods such as generalized likelihood uncertainty estimation [GLUE; Beven and Binley, 1992], SCEM-UA optimization [Vrugt et al., 2003], or DREAM MCMC sampling [Vrugt et al., 2008], Nonetheless, these methods usually involve orders of magnitude more forward model evaluations than the $N_e = 100$ particles used by PBS in this experiment. For example, Clark et al. [2006] demonstrated the need for a computationally intensive SODA approach, using SCEM-UA in an outer loop to optimize parameters, together with an EnKF inner loop for state estimation, to rigorously calibrate a simple two-parameter snow model. Furthermore, the widely used hydrological calibration technique GLUE [Beven and Binley, 199 , tailored to identify equifinality [Beven, 2006], is equivalent to PBS when using a Gaussian likelihood. As such, GLUE could only outperform PBS by increasing the ensemble size or modifying the likelihood, which are equally applicable to PBS. Both black-box optimization and inference are challenging problems where particle methods have proven to be highly competitive [Duan et al., 2023, Chopin and Papaspiliopoulos, 2020] . In fact, these particle methods can be seen as formalizing highly successful meta-heuristic evolutionary algorithms that are widely used for optimization [Del Moral, 2004]. The takeaway here is that particle methods can be improved either by increasing the ensemble size, at considerable computational expense given typically poor scaling [Snyder et al., 2008], or by adapting them for example using various iterative methods [Bugallo et al., 2017, Chopin and Papaspiliopoulos, 2020].

COMMENT # 1.5

There are only a few snow depth measurements in Figure 2, so I guess that the ensemble collapse is due to the observations having high confidence. The assumed observation error should be stated in the text and shown on the figures.

Reply:

The reviewer is correct that precise snow depth observations (i.e. relatively low observation error variance) helps contribute to ensemble collapse here. We also thank the reviewer for pointing out the need to specify the observation error variance. This is now included in the text as follows for Experiment 1

Changes:

These drone-acquired data and forcing datasets have previously been used in data assimilation studies [Alonso-González et al., 2022, Alonso-González et al., 2023] ~~.~~ ~~The~~ Following these studies and the findings of Revuelto et al. [2021], we assume an observation error standard deviation of $\sigma_y = 0.2$ m for the drone-based snow depth retrievals. The corresponding observation error variance σ_y^2 is used to construct the diagonal observation error covariance matrix **R** in the likelihood (**??**). The

Reply:

and Experiment 2:

Changes:

In the second set of experiments, we conducted several site-level simulations using a physics-based snow model of intermediate complexity, the Flexible Snow Model ~~[Essery et al., 2025, FMSM2;]~~[Essery et al., 2025, FSM2;]. In these experiments, we assimilated hourly observations of snow depth at six different geographical locations using both ~~the~~ AdaPBS and ES-MDA for comparison. For simplicity, we assume the same observation error standard deviation of $\sigma_y = 0.2$ m as in Experiment 1. While this is likely to be an overestimate of individual measurement errors at these sites, inflating independent observation errors in our model helps implicitly account for the fact that the actual errors in these data are correlated in time, which reduces their information content [Evensen et al., 2022].

COMMENT # 1.6

Is the KDE in Figure 6 misleading? The PBS posterior approximation appears to be too wide, but it should actually be too narrow.

Reply:

We thank the reviewer for spotting this. It is indeed an artifact caused by the KDE routine `scipy.stats.gaussian_kde` and the default 'scott' bandwidth selection method and exposes the risk of applying default methods. Using the (nearly completely) degenerate PBS weights with $N_{eff} = 1.26$, corresponding to two surviving yet unevenly weighted particles (the big and small green stars in Figure 2) , this routine picked a wide bandwidth factor of $1.26^{-1/5} = 0.95$ such that the mixture of Gaussians at the two isolated peaks, one lower than the other, merges into a single Gaussian. To better visualize the actual degeneracy, we have now applied the KDE to resmapled particles from the PBS, resulting in many identical copies of the two surviving particles, such that after resampling $N_{\text{eff}} = N_e = 10^2$ by definition. Recall that this only superficially solves the weight degeneracy issue, not the underlying deeper path degeneracy problem [Murphy, 2023]. For the KDE this effectively lowers the bandwidth factor to $100^{-1/5} = 0.40$ which at least makes the two surviving particles more visible and thus helps to better visualize the degeneracy. Figure 6 has been modified accordingly, and resampling was also applied to AdaPBS before passing it to the KDE for consistency. The caption of Figure 6 has also been modified to provide more detail on the KDE:

Changes:

Comparison of Gaussian kernel density estimates (KDEs) obtained via `scipy.stats.gaussian_kde` [Scott's rule bandwidth; Virtanen et al., 2020] of the marginal prior and marginal approximate posterior distributions of the parameters from the respective algorithms in the transformed (unbounded) space. The MCMC posterior (orange) from RAM is considered the gold-standard.

Comment # 1.7

In situ meteorological data are available for all of the ESM-SnowMIP sites; is ERA5 used instead for generality? ERA5 grid elevations without downscaling can be expected to be considerably lower than most of the site elevations (Menard at al. 2019, Figure 9). The final version of Essery et al. (2024) is https://doi.org/10.5194/gmd-18-3583-2025. Of the parameters listed in Table 1, rgr0 and rhow will have no influence on snow depth simulations; this could be apparent from the posterior parameters. Can MuSA diagnose parameter importance?

Reply:

We thank the reviewer for highlighting these various points. Unlike in the first experiment, in experiment two we did not apply any topographic downscaling to the ERA5 forcing data for the selected ESM-SnowMIP sites. Moreover, as suspected by the reviewer, we used ERA5 data as forcing rather than the in-situ meteorological data for generality. By not applying a topographic correction, especially for temperature, we also made the DA task more challenging at these sites which are likely at higher elevations than the nearest $0.25°$ ERA5 cell centroid based on a comparison to coarser $0.5°$ GSWP3 cells in Ménard et al. [2019]. Nonetheless, the majority ($\simeq 95\%$) of the prior probability on the additive temperature bias perturbation parameter with $\sigma_b = 1$ K is in the range $[-2, 2]$ K which would accommodate an elevation difference of roughly 300 m according with a standard lapse rate 6.5 K km^{-1}. According to Figure 9 in Ménard et al. [2019], this 300 m elevation difference encompasses most of the sites that we considered except RME which seems to have the largest difference with GSWP3. The AdaPBS still performed quite well at RME (compared to say SNB), so it may be that compared to GSWP3 the elevation differences are smaller between ERA5 and these sites. We now explicitly mention the reasons for this choice of forcing and lack of topographic correction explicitly in the experimental setup (Section 3.3) thanks to the reviewer's comment as follows

Changes:

The meteorological forcing was obtained directly from the nearest grid cell of the ERA5 reanalysis [Hersbach et al., 2020] ~~. Thus,~~ without applying topographic corrections. While site-level meteorological forcing was available from Ménard et al. [2019], we used the more uncertain, unadjusted global reanalysis data to provide a more challenging experimental setup that is generally applicable to any site with snow depth data. Specifically, each water year~~a~~, we assimilated a large batch of $24 \times 365 = 8760$ hourly snow depth observations~~are being assimilated in these experiments~~.

Reply:

We have now also updated the FSM2 reference to the published version [Essery et al., 2025] throughout.

We thank the reviewer for pointing out that these two specific parameters should have no influence on snow depth simulations and for raising the possibility of evaluating parameter importance. It is possible to diagnose parameter importance in MuSA using any of the DA techniques contained therein, including AdaPBS, as long as they provide a sufficiently robust (i.e., non degenerate) approximation of the posterior. Such an evaluation can be performed by looking at the information gain to see how much the entropy is reduced when passing from the marginal prior to the marginal posterior for each parameter [van Hove et al., 2026]; a greater entropy reduction corresponds to greater information gain. Note that this differs from classical sensitivity analysis, which focuses on how much varying a parameter affects some simulated output; the information gain instead measures how much is learned about a parameter by conditioning on actual data. While the possibility of performing such an information gain analysis is an intriguing opportunity for future studies, we consider it beyond the scope of this paper, which is focused on presenting AdaPBS, especially as we are primarily concerned with performance in state space in terms of predicted snow depth in Experiment 2. To help clarify this, we have also added the following paragraph to Section 3.3:

Changes:

Unlike Experiment 1 with the temperature index model, we did not run MCMC in this more complex experiment due to the high computational cost and marginal additional benefits to the analysis. As such, the KLD metric could not be used without an MCMC-based reference distribution. Instead, this experiment focused on comparing the predictive performance in state space of AdaPBS to that of ES-MDA, an advanced and competitive baseline for DA with more complex snow models [Alonso-González et al., 2022, Alonso-González et al., 2026]. In this higher-dimensional experiment, the entire 19-dimensional parameter space is too large for comprehensive and instructive visualization. An alternative could be to investigate the importance of each parameter by estimating the information gain [van Hove et al., 2026], but such parameter analysis strays beyond the focus of this experiment. Except for a handful of forcing parameters partly analyzed in Experiment 1, the majority of the 19 parameters are so-called nuisance parameters [Jaynes, 2003]. They are not of primary interest beyond their effect on the state and their role in making the DA problem more challenging. We thus focus on the performance of AdaPBS and ES-MDA for estimating the snow depth state observed at the selected ESM-SnowMIP sites.

COMMENT # 1.8

It would be easy to add PBS RMSE and bias to Table 3, allowing the reader to judge for themselves how badly it is struggling. There will be enormous redundancy in hourly snow depth measurements because of autocorrelation; would decimated or averaged observations be less challenging?

Reply:

We thank the reviewer for this suggestion and have now added the corresponding metrics for PBS to Table 3 and discussed these new results in the revised text as follows in Section 4.2:

Changes:

The results in Table 3 ~~show is what they do not show: the performance of~~ also show that AdaPBS consistently outperforms PBS in this experiment for all metrics across all sites, with the exception of SOD where the low bias differs by only 0.01. Overall, when considering metrics averaged over all sites, use of AdaPBS leads to a 25%, 41%, and 60% reduction in RMSE, CRPS, and bias, respectively, compared to PBS. A likely reason for the large gains from adaptation is that the ~~non-iterative PBS and the ES. The reason for this is that in this case the PBS would have a tendency to struggle with capturing~~ non-iterative PBS struggles to capture the dense observations, as demonstrated by the average of eight iterations needed in AdaPBS~~, and~~

~~very likely degenerate~~ . As a result, PBS often degenerates completely to a sub-optimal solution~~whereas the~~ . Although not explicitly evaluated here, a single ES update ~~would suffer due to~~ is expected to also lead to a suboptimal solution due to the combined effects of nonlinearity and the highly informative nature of ~~these observations. The expected performance of these two non-iterative schemes is based on~~ the observations. This expectation is supported by Experiment 1 as well as previous experiments comparing the ES-MDA to ~~these non-iterative schemes~~ the ES in snow data assimilation [Aalstad et al., 2018, Alonso-González et al., 2022] and ~~beyond~~ other applications [Pirk et al., 2022, 2024, Keetz et al., 2025].

Reply:

The reviewer is correct that there is likely some redundancy in hourly snow depth measurements especially considering that observation errors are actually autocorrelated. The idea here was to construct a case that was particularly challenging for the PBS by ending up with a very narrow needle in terms of numerous and highly informative observations. Averaging would not necessarily make the assimilation less challenging, since with this additive Gaussian error model the observation error standard deviation σ_y would need to be scaled by $1/\sqrt{N_{av}}$ where N_{av} is the number of observations that are averaged over, to account for the averaging. On the other hand, accounting for auto-correlated errors in the observations would reduce the information content of the snow depth observations and thus make contraction of the posterior less dramatic and easier to capture. While it is possible to add this to the likelihood related following the work of Smith et al. [2010] as cited in the manuscript, this is beyond the scope of our study where we assume uncorrelated observation errors as is typical in DA [Carrassi et al., 2018]. The use of decimated (sub-sampled) observations would certainly make the exercise easier and reduce degeneracy, but this would also cause a loss of information assuming uncorrelated errors. The related technique of data tempering (also known as mini-batching) following Chopin [2002] which is now discussed in the new Section 4.3 provides a similar way to make the problem less challenging while not throwing away information.

Minor errors:

COMMENT # 1.9

p3 "Ensembe Kalman"

Reply:

Fixed to ‘Ensemble Kalman’, thanks for spotting this.

COMMENT # 1.10

p4-5 “comprehensive texts . . . for a comprehensive”

Reply:

Thanks for spotting this unfortunate wording, we have now change the second comprehensive to "thorough"

COMMENT # 1.11

p12 “that we can easy generate”

Reply:

Fixed to ‘that we can easily generate’, thanks.

COMMENT # 1.12

p13 “samples form the target”

Reply:

Fixed typo to ‘from’, thanks for pointing this out.

COMMENT # 1.13

p15 “each iterations proposals density” (missing apostrophe)

Reply:

Thanks for spotting this, we now added the missing apostrophe:

Changes:

The particles $\theta_i^{(k)}$ for each iteration k have been drawn independently from their respective sampling proposal distributions $q^{(k)}$.

COMMENT # 1.14

p19 No reference in the text to Figure 1 (could be related to the steps in 2.6.1).

Reply:

Thanks for spotting this, we have now added the following paragraph just above Figure 1:

Changes:

The overall workflow of the AdaPBS algorithm is visualized in Figure **??**. As with the standard PBS, AdaPBS is initialized by sampling a particle ensemble from the prior, running this ensemble through the forward model, and computing self-normalized importance weights via (**??**). Ensemble diversity is then diagnosed using the effective sample size N_{eff} in (**??**). If N_{eff} exceeds a predefined threshold, the algorithm stops successfully, outputting this diverse ensemble as the posterior approximation. Otherwise, it adapts a new sampling proposal using clipping (**??**) and ensemble statistics via (**??**) and (**??**). A new particle ensemble is drawn from this adapted sampling proposal, run through the forward model, and weights are obtained using (**??**), now incorporating the entire particle history via a mixture proposal (**??**). Adaptation continues until the ESS-based diversity threshold is met or a predefined maximum number of iterations is reached.

COMMENT # 1.15

p20 "cryospheric sciene"

Reply:

Fixed 'science', thanks.

COMMENT # 1.16

p22 "The idea being ..." is an incomplete sentence as written.

Reply:

Thanks for spotting this, we changed the sentence to:

Changes:

The idea ~~being~~ is to accelerate MCMC by ~~initializing at a location~~ starting the chain within or close to the typical set of the target posterior distribution.

COMMENT # 1.17

p23 "FMSM2".

Reply:

Fixed the typo.

COMMENT # 1.18

p23 "7 perturbation parameters to correct the forcing" – only six are listed.

Reply:

TODO add specific humidity or reduce to six forcing parameters? If we only use 6 we need to update total parameter count to 18 not 19

COMMENT # 1.19

p26 The MCMC distribution in Figure 2 should have a colour scale.

Reply:

We thank the reviewer for pointing out the lack of scale for the MCMC heatmap. We have now visualized all the reference posterior MCMC samples in blue and removed the heatmap as the point density colormap does not add much information.

COMMENT # 1.20

p29 Figure 5 would be clearer if ESS in the legend was changed to N_{eff} as in the caption.

Reply:

We thank the reviewer for spotting this inconsistency regarding effective sample size (ESS) notation by using the ESS acronym in the legend and the N_{eff} symbol in the caption. We have updated Figure 5 (and Figure 1) to instead just use N_{eff} for consistency and clarity.

COMMENT # 1.21

> *p29 Rather than ending up selecting a threshold value, you must have started out by doing so if it is preselected (p16).*

Reply:

We thank the reviewer for spotting this unfortunate turn of phrase, we have now changed this to the following which aligns with our original intent:

Changes:

In this case, we ~~ended up selecting~~ used a relatively high ~~N_{eff} diversity threshold value~~ diversity threshold of $\tau = 0.3$ ~~(30%)~~for N_{eff}/N_e, which in practice resulted in the same number of iterations as ES-MDA.

We would like to thank Richard Essery for his constructive and detailed review. We are especially grateful to receive this positive feedback given the key role of FSM2 [Essery et al., 2025] and ESM-SnowMIP [Ménard et al., 2019] in our study.

On behalf of all coauthors,
Kristoffer Aalstad and Esteban Alonso-González

REFERENCES

K. Aalstad, S. Westermann, T. V. Schuler, J. Boike, and L. Bertino. Ensemble-based assimilation of fractional snow-covered area satellite retrievals to estimate the snow distribution at Arctic sites. *The Cryosphere*, 12:247–270, 2018. doi: 10.5194/tc-12-247-2018.

E. Alonso-González, K. Aalstad, M. W. Baba, J. Revuelto, J. I. López-Moreno, J. Fiddes, R. Essery, and S. Gascoin. The Multiple Snow Data Assimilation System (MuSA v1.0). *Geosci. Model Dev.*, 2022. doi: 10.5194/gmd-15-9127-2022.

E. Alonso-González, K. Aalstad, N. Pirk, M. Mazzolini, D. Treichler, P. Leclercq, S. Westermann, J. I. López-Moreno, and S. Gascoin. Spatio-temporal information propagation using sparse observations in hyper-resolution ensemble-based snow data assimilation. *Hydrol. Earth Syst. Sci.*, 27:4637–4659, 2023. doi: 10.5194/hess-27-4637-2023.

E. Alonso-González, A. Harpold, J. D. Lundquist, C. Piske, L. Sourp, K. Aalstad, and S. Gascoin. Ensemble-based data assimilation improves hyperresolution snowpack simulations in forests. *The Cryosphere*, 20:209–225, 2026. doi: 10.5194/tc-20-209-2026.

R. N. Bannister. A review of operational methods of variational and ensemble-variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 143:607–633, 2017. doi: 10.1002/qj.2982.

K. Beven. A manifesto for the equifinality thesis. *Journal of Hydrology*, 320:18–36, 2006. doi: 10.1016/j.jhydrol.2005.07.007.

K. Beven and A. Binley. The future of distributed models: Model calibration and uncertainty prediction. *Hydrol. Process.*, 6:279–298, 1992. doi: 10.1002/hyp.3360060305.

Monica F. Bugallo, Victor Elvira, Luca Martino, David Luengo, Joaquin Miguez, and Petar M. Djuric. Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Processing Magazine*, 34(4):60–79, 2017. doi: 10.1109/MSP.2017.2699226.

W. Cao, K. Aalstad, L.S. Schmidt, S. Westermann, and T.V. Schuler. Glacier data assimilation on an Arctic glacier: Learning from large ensemble twin experiments. *Journal of Glaciology*, 2025. doi: 10.48550/arXiv.2502.09314. Preprint in review.

A. Carrassi, M. Bocquet, L. Bertino, and G. Evensen. Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *WIREs Climate Change*, 9:e535, 2018. doi: 10.1002/wcc.535.

L. Charrois, E. Cosme, M. Dumont, M. Lafaysse, S. Morin, Q. Libois, and G. Picard. On the assimilation of optical reflectances and snow depth observations into a detailed snowpack model. *The Cryosphere*, 10:1021–1038, 2016. doi: 10.5194/tc-10-1021-2016.

N. Chopin. A sequential particle filter method for static models. *Biometrika*, 89:539–552, 2002. doi: 10.1093/biomet/89.3.539.

N. Chopin and O. Papaspiliopoulos. *An Introduction to Sequential Monte Carlo*. Springer, 2020. doi: 10.1007/978-3-030-47845-2.

M. P. Clark, A. G. Slater, A. P. Barrett, L. E. Hay, G. J. McCabe, B. Rajagopalan, and G. H. Leavesley. Assimilation of snow covered area information into hydrologic and land-surface models. *Advances in Water Resources*, 29:1209–1221, 2006. doi: 10.1016/j.advwatres.2005.10.001.

P. Del Moral. *Feynman-Kac Formulae*. Springer, 2004. doi: 10.1007/978-1-4684-9393-1.

Jin-Chuan Duan, Shuping Li, and Yaxian Xu. Sequential monte carlo optimization and statistical inference. *WIREs Computational Statistics*, 15(3):e1598, 2023. doi: 10.1002/wics.1598.

R. Essery, G. Mazzotti, S. Barr, T. Jonas, T. Quaife, and N. Rutter. A flexible snow model (fsm 2.1.1) including a forest canopy. *GMD*, 18:3583–3605, 2025. doi: 10.5194/gmd-18-3583-2025.

G. Evensen, F. C. Vossepoel, and P. J. van Leeuwen. *Data Assimilation Fundamentals*. Springer, 2022. doi: 10.1007/978-3-030-96709-3.

A. Farchi and M. Bocquet. Review article: Comparison of local particle filters and new implementations. *Nonlinear Processes in Geophysics*, 25:765–807, 2018. doi: 10.5194/npg-25-765-2018.

A. Gelman, J.B. Carlin, H. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis*. CRC Press, 3 edition, 2013.

N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F Radar and Signal Processing*, 140:107, 1993. doi: 10.1049/ip-f-2.1993.0015.

H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146: 1999–2049, 2020. doi: 10.1002/qj.3803.

E.T. Jaynes. *Probability Theory*. Cambridge, 2003. doi: 10.1017/CBO9780511790423.

L. T. Keetz, K. Aalstad, R. A. Fisher, C. Poppe Terán, B. Naz, N. Pirk, Y. A. Yilmaz, and O. Skarpaas. Inferring Parameters in a Complex Land Surface Model by Combining Data Assimilation and Machine Learning. *Journal of Advances in Modeling Earth Systems*, 17(6):e2024MS004542, 2025. doi: 10.1029/2024MS004542.

D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge, 2003.

S. A. Margulis, M. Girotto, G. Cortés, and M. Durand. A particle batch smoother approach to snow water equivalent estimation. *Journal of Hydrometeorology*, 16: 1752–1772, 2015. doi: 10.1175/JHM-D-14-0177.1.

C. B. Ménard, R. Essery, A. Barr, P. Bartlett, J. Derry, M. Dumont, C. Fierz, H. Kim, A. Kontu, Y. Lejeune, D. Marks, M. Niwano, M. Raleigh, L. Wang, and N. Wever. Meteorological and evaluation datasets for snow modelling at 10 reference sites: description of in situ and bias-corrected reanalysis data. *Earth Sys. Sci. Data*, 11(2): 865–880, 2019. doi: 10.5194/essd-11-865-2019. URL https://essd.copernicus.org/articles/11/865/2019/.

M. Morzfeld, D. Hodyss, and C. Snyder. What the collapse of the ensemble Kalman filter tells us about particle filters. *Tellus A: Dynamic Meteorology and Oceanography*, 69:1–14, 2017. doi: 10.1080/16000870.2017.1283809.

K.P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT, 2023. URL https://probml.github.io/book2. (last access: 18.12.2023).

Moritz Oberrauch, Bertrand Cluzet, Jan Magnusson, and Tobias Jonas. Improving fully distributed snowpack simulations by mapping perturbations of meteorological forcings inferred from particle filter assimilation of snow monitoring data. *Water Resources Research*, 60(12):e2023WR036994, 2024. doi: 10.1029/2023WR036994.

N. Pirk, K. Aalstad, S. Westermann, A. Vatne, A. van Hove, L.M. Tallaksen, M. Cassiani, and G. Katul. Inferring surface energy fluxes using drone data assimilation in large eddy simulations. *Atmospheric Measurement Techniques*, 15:7293–7314, 2022. doi: 10.5194/amt-15-7293-2022.

Norbert Pirk, Kristoffer Aalstad, Erik Schytt Mannerfelt, François Clayer, Heleen de Wit, Casper T. Christiansen, Inge Althuizen, Hanna Lee, and Sebastian Westermann. Disaggregating the carbon exchange of degrading permafrost peatlands using bayesian deep learning. *Geophysical Research Letters*, 51(10):e2024GL109283, 2024. doi: https://doi.org/10.1029/2024GL109283. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2024GL109283. e2024GL109283 2024GL109283.

T. Rainforth, A. Golinski, F. Wood, and S. Zaidi. Target–Aware Bayesian Inference: How to Beat Optimal Conventional Estimators. *JMLR*, 21:1–54, 2020.

J. Revuelto, E. Alonso-Gonzalez, I. Vidaller-Gayan, E. Lacroix, E. Izagirre, G. Rodríguez-López, and J.I. López-Moreno. Intercomparison of UAV platforms for mapping snow depth distribution in complex alpine terrain. *Cold Regions Science and Technology*, 190:103344, 2021. doi: 10.1016/j.coldregions.2021.103344.

D. Sanz-Alonso, A. M. Stuart, and A. Taeb. *Inverse problems and data assimilation*, volume 107. Cambridge, 2023. doi: 10.1017/9781009414319.

Tyler Smith, Ashish Sharma, Lucy Marshall, Raj Mehrotra, and Scott Sisson. Development of a formal likelihood function for improved bayesian inference of ephemeral catchments. *Water Resources Research*, 46(12), 2010. doi: 10.1029/2010WR009514.

Chris Snyder, Thomas Bengtsson, Peter Bickel, and Jeff Anderson. Obstacles to high-dimensional particle filtering. *Monthly Weather Review*, 136(12):4629 – 4640, 2008. doi: 10.1175/2008MWR2529.1.

Alouette van Hove, Kristoffer Aalstad, and Norbert Pirk. Actively inferring methane sources with drones. *Environmental Data Science*, 5:e2, 2026. doi: 10.1017/eds.2026.10029.

P. J. van Leeuwen, H. R. Künsch, L. Nerger, R. Potthast, and S. Reich. Particle filters for high-dimensional geoscience applications: A review. *Quarterly Journal of the Royal Meteorological Society*, 145:2335–2365, 2019. doi: 10.1002/qj.3551.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, Ilhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272, February 2020. ISSN 1548-7105. doi: 10.1038/s41592-019-0686-2. URL http://dx.doi.org/10.1038/s41592-019-0686-2.

Jasper A. Vrugt, Hoshin V. Gupta, Willem Bouten, and Soroosh Sorooshian. A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resources Research*, 39(8), 2003. doi: 10.1029/2002WR001642.

Jasper A. Vrugt, Cajo J. F. ter Braak, Martyn P. Clark, James M. Hyman, and Bruce A. Robinson. Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with markov chain monte carlo simulation. *Water Resources Research*, 44(12):W00B09, 2008. doi: 10.1029/2007WR006720.