Author response to the review of D. Brinkerhoff

Johannes Landmann and co-authors

June 2021

Dear Douglas Brinkerhoff,
Dear Editor,
We thank Douglas Brinkerhoff very much for the valuable comments regarding our manuscript. We really appreciate the thoughtful input and the detailed recommendations to improve the document. We address the raised requests in the form of point-by-point responses and make suggestions for how to update the final manuscript.
Best regards,
Johannes Landmann and co-authors

RC: *Title The observations are not of mass balance, but of surface elevation (specifically in the negative direction). I suggest changing the title to be more precise.*
AR: Thanks for this comment. We agree that the observations we make are in "surface elevation space" and that these are then transformed into a mass change using the density of the medium (snow/ice). However, the reviewer will agree that (i) the surface elevation change of a glacier at a given point is the result of both surface mass balance and a component due to ice emergence (see Cogley et al., 2011, p. 38), and (ii) the changes that we measure at our stakes are only due to the former. Stated differently: what we see is the surface elevation change due to glacier mass balance, but not the total ice thickness change. To avoid possible confusion, we propose to change the title into: "Assimilating near real-time mass balance stake readings into a model ensemble using a particle filter".

RC: *L10 The reader does not yet know what 'model probability' is in the abstract, nor is the abstract notion of 'custom resampling' useful here.*
AR: We understand the point and suggest the following, simplified text: "These observations are assimilated into an ensemble of three temperature index (TI) and one simplified energy-balance mass balance models using a particle filter. By using state augmentation, we assign temporally-varying weights to individual models. We analyse model performance over the observation period, and find that the model probability is highest for [...]".

RC: *L39 Of the three points (first, second, third) made after this line, only one logically follows this statement.*
AR: We agree that the logics might not have been apparent because the enumerations were stretched very far out into the subsequent text. We propose to restructure the text as follows: "In many cases, mass balance analyses are available twice a year, i.e. they are based on seasonal in situ observations (Cogley et al., 2011). This relatively low mass balance analysis frequency is mainly due to the fact that in situ observations are not often made, because they are expensive in terms of both time and manpower. Only recently have low-cost and high-frequency monitoring approaches emerged (Hulth, 2010, Fausto et al., 2012, Keeler and Brugger, 2012, Biron and Rabatel, 2019, Carturan et al., 2019, Gugerli et al., 2019, Netto and Arigony-Neto, 2019). However, even if higher observation frequencies are available with these new approaches, it is not straightforward to calculate analyses at higher frequencies. This is because near real-time estimates are often based on ensemble modelling, in order to enable a correct quantification of uncertainties. Ensemble modelling is used in glaciology in the context of model intercomparison projects (Hock et al., 2019), future projections for ice sheets and mountain glaciers (e.g. Ritz et al., 2015, Shannon et al., 2019, Golledge, 2020, Marzeion et al., 2020, Seroussi et al., 2020), and also to determine the initial conditions for modelling (Eis et al., 2019). However, ensembles are currently not prominent in the calculation of seasonal or daily glacier mass balances. Another reason why calculating higher-frequent glacier mass balance analyses is not straightforward is that there is often a lack of knowledge about the exact short-term parameters in mass balance models. This poses a problem, since e.g. temperature

index (TI) models are parametrizations of the full energy balance equation and deliver inaccurate results when applied with inapt parameters for a specific location [...]".

RC: *L49 List of references should have an e.g. in front of it. There are many other examples of ensemble modelling for ice sheet projection.*
AR: Thanks! We will insert "e.g.".

RC: *L55 'discussed how'→'not clear whether'(?)*
AR: We will exchange the wording as proposed.

RC: *L63 surface point mass balance→surface point ablation. You don't measure massbalance, you measure volume change in one direction.*
AR: Besides the answer given in reply to the comment related to the title (see our first answer), note that we do not only measure ablation but also accumulation. Together, this is the mass balance of the surface.

RC: *L80 as above.*
AR: Idem.

RC: *L103 'cumulative surface height change' is (mathematically) equivalent to 'surface height'. I suggest the latter for brevity.*
AR: We agree in principle. However, the formulation "observations of ice surface height between two time steps" (as the sentence would read then in l. 103) sounds unintuitive to us. We thus suggest the following wording as a compromise: "observations of surface height change since a given point in time (in our case the time at which the camera is set up)".

RC: *Eq. 1 This equation is only valid for bare ice. This is briefly touched on elsewhere, but should be reiterated here. In fact, it might be better to state that the operation relates h(t,z) to* a_{sfc}.
AR: We suggest to circumvent this issue by replacing the density of ice "ρ_{ice}" with the bulk density of snow and/or ice "ρ_{bulk}". We think that this should clarify that this equation is generally valid. We will explain in the text that "ρ_{bulk} is the temporally weighted average of the snow and ice densities ($\mathrm{kg\,m^{-3}}$) at the camera location.".

RC: *L111 'Short snow events ...'. We never see this notion of assigning a high uncertainty SWE estimate again. Is this actually done, and specifically how?*
AR: Yes, we actually assign higher uncertainties to snowfalls. We propose to clarify the "how" as follows: "Short snow events during the melt seasons are assigned a density of 150 $\mathrm{kg\,m^{-3}}$. The calculated snow water equivalent is assigned an uncertainty of 2-3 cm w.e.."

RC: *L133 I'm confused by the lapse rate thing. Why don't you continue to be a Bayesian and just use the probability distribution over the lapse rate inferred from the data without injecting questionable notions of 'significance'? This could then be propagated into downstream analysis.*
AR: We did not use significance testing in the strict statistical sense, but rather as a simple procedure to take background information (in the data assimilation sense) into account on days where the lapse rate cannot be estimated from the meteorological data directly. However, we will implement the suggested way to treat the lapse rates of temperature and precipitation, which we understand as follows: we compute the posterior distribution of the lapse rate for each day (using, e.g., a g-prior of Zellner (Zellner, 1986)) and then use an independent draw from this posterior for each particle in the particle filter's predict step. Note that the information required to describe the procedure will most likely result in an increase in the paper's length.

RC: *L151 In what sense is an outline a surface? I don't understand this line.*
AR: The term "reference surface" is used to highlight that the extent doesn't change over time. We suggest to clarify this in the following way: "[...] mass balances in this study are calculated over a glacier surface area that does not change over time (Elsberg et al., 2001, Huss et al., 2012)."

RC: *L158 'Values of glacier-wide mass balance ...' I don't understand what 'partly harmonized' means in this context?*
AR: It means that some of the mass balance data used for model calibration (i.e. data provided by GLAMOS), are already consistent with geodetic mass balances, thus matching long-term mass changes

(the procedure of ensuring consistency is often referred to as "homogenization", (Huss et al., 2015)). "Partly homogenized" (or "harmonized") means that this procedure has not yet happened for the most recent mass balance data, since no geodetic mass balances are available yet. We will better explain that in the revised text at l.158 and exchange "harmonized" with "homogenized" to avoid confusion: "Glacier-wide mass balances are obtained by extrapolating the in-situ observations, and making the extrapolated values consistent with long-term mass changes. The latter procedure is often referred to as "homogenization" (e.g. Bauder et al., 2007, Huss et al., 2015). For the recent years this homogenization has not yet been applied, since no geodetic mass balances are available yet."

RC: *Eq. 2 Perhaps it's standard notation, but having cprec mean an entirely different thing (with different units) than ccfc is really confusing.*
AR: We will replace c_{prec} with $prcp_{\text{scale}}$ in the entire manuscript to avoid such confusion.

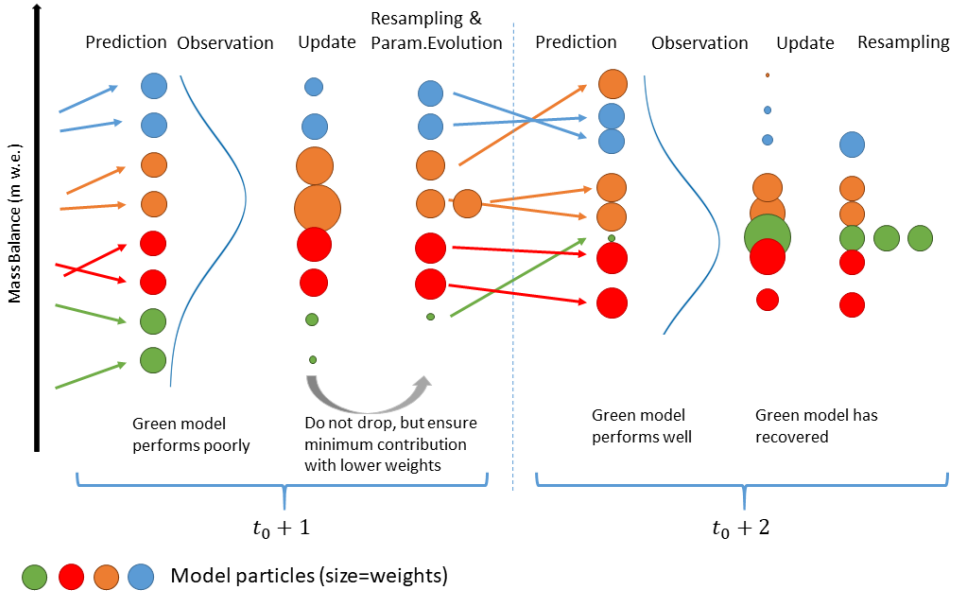
RC: *L214 It would be useful to make a note that G is a function of I.*
AR: We agree. We will use the following notation in the entire manuscript: $G(I_{\text{pot}}, t, z)$

RC: *Eq. 8 Suggest using Δ t rather than dt, as the latter is usually reserved for infinitesimals.*
AR: We will change this as suggested.

RC: *Eq. 10 The 'general framework' also has inside epsilon of H(x), although it doesn't appear that way in this work.*
AR: We see the point. Our notation, which has ϵ outside of $\mathcal{H}(x)$, is meant to signalize that we only consider additive errors in our study. This is a common assumption, and it is not obvious to us why considering other error types would be beneficial in our case. However, we agree that some confusion arises due to the different notations in Eqs. 1 and 10. We thus suggest to move ϵ_t outside of the brackets to match Eq. 10:

$$h(t,z) = \mathcal{H}(b_{\text{sfc}}(t,z)) + \epsilon_{t,z} = \frac{b_{\text{sfc}}(t,z) \cdot \rho_w}{\rho_{\text{bulk}}} + \epsilon_{t,z}$$

RC: *Sec. 3.3 I find it confusing that the parameter update process appears at the end, even though Figure 6 indicates that it happens at the same time as the state prediction.*
AR: The reason why the two processes are merged in fig. 6 is that we couldn't think of a symbol representing the updating of parameters. In general, the order "resampling - parameter update - prediction" is due to the observations (and so the weights) only depending on the physical state ξ, and not on the parameters θ. This is why the parameters can evolve after the resampling, in which it is decided to which models the resampled particles belong. In the example figure, the two orange particles with the same physical state ξ_t can thus obtain different parameters. We suggest to visually separate the parameter evolution from the prediction step by moving the annotation to the resampling side.



RC: *L289 It might be clearer to state explicitly that a particle is always associated with only 1 model over its "lifetime".*
AR: Thanks for this comment. We suggest to add one more sentence to explain what "a particle" means: "[...], which means that, when following a given particle backwards in time, the entire dynamics of the particle is governed by one single model over its "lifetime". In the other direction, a particle can change model during the resampling step. In this case, both the model index $m_{t,k}$ and the entire past trajectory is changed to the new model.".

RC: *Eq. 16 I strongly disagree with the choice of setting $\beta_t = 0$. This is because this is tantamount to the assumption that the model predictions are perfect, which is certainly not the case. In reality, two models are only different in their reliability to the extent that their predictions differ by more than their internal uncertainty and one fits the data better than the other. Setting this model error to zero artificially accentuates the differences in the likelihoods computed for different model and encourages the mode collapse (what you call 'model dominance') exhibited in Figure 11.*
AR: Our choice of setting β_t=0 originates from the fact that for our case, the model error is dominated by the uncertain model inputs, i.e. by the meteorological input and the parameters. In principle, we have thus splitted the value of β into the meteorological input errors η, which we can specify correctly. We suggest to better clarify this point with the following explanation (l. 306): "By introducing both the meteorological input uncertainty η and the parameter uncertainties, we shift the majority of the uncertainty contained in β_t to these variables. Since the remaining uncertainty for β_t is small and hard to quantify, we set β_t=0 for simplicity."

RC: *Table 2 Caption By covariance, do you mean standard deviation?*
AR: Yes, sorry, that's a typo.

RC: *L320 the standard symbol for variance would be σ^2.*
AR: For clarity (one superscript less), we will change "variance" to "standard deviation" and leave the symbol σ_ϵ untouched.

RC: *Eq. 18 An implicit assumption made throughout is that a single model's probability is marginally uniform, or alternatively that $P(m_t) = Dirichlet(1)$, to wit that one model being dominant is just as probable as all four models contributing equally. This is a weird assumption for a time dependent problem, because it means that physical reality is subject to sudden switches between governing principles. Again, this leads to the mode collapse seen in Fig. 11. Predictions might be made substantially more robust by putting a prior on P(mt) such that the more probable case is an averaging of the four models, and deviation from that has to be the result of significant evidence.*
AR: We do not make the assumption that the probability of each single model is marginally uniform, neither explicitly nor implicitly. At the initial time t_0, all four models have the probability 1/4. As the model index never changes during predict steps, the same is true at all later times t if we do not condition on the observations. If we condition on the observations, instead, the model probabilities change at each time a new observation becomes available (this happens during the update step). In other words, we have

$$P(m_t = j \mid y_{1:t-1}) = P(m_{t-1} = j \mid y_{1:t-1}),$$

This means that a model that has high probability at time $t-1$ is favored by the prior also at time t, which means that a sudden switch of the preferred model is unlikely. Such a switch only happens if the evidence for it is strong, the evidence being given by the likelihood ratio and the likelihood of model j being

$$p(y_t \mid m_t = j, y_{1:t-1}) \propto \frac{p(m_t = j \mid y_{1:t})}{p(m_t = j \mid y_{1:t-1})},$$

or, by Equations (17) and (18),

$$p(y_t \mid m_t = j, y_{1:t-1}) \propto \frac{\sum_k p(y_t \mid x_{t,k}) w_{t-1,k} \delta(m_{t,k} - j)}{\sum_k w_{t-1,k} \delta(m_{t-1,k} - j)}.$$

The preferred model can thus only switch if the particles belonging to one model have a much better fit for the new observations than all the other models. This is what happens with our data. It indicates that presumably the forecast values $x_{t,k}$ are overconfident and that all four models have non-negligible model errors. However, in order to specify model errors, we would need physical knowledge of their

order of magnitude and their dependence on meteorological inputs. As a way to mirror this explanation in the manuscript, we suggest to show a figure containing the prediction $b_{sfc}(z,t)_k$ and the likelihood $p(y_t \mid b_{sfc}(z,t))$ for a time t where the model probability changes quickly.

RC: *L338 It's not that there's no stochasticity, it's that mt for a given particle doesn't evolve at all!*
AR: Thanks for raising this point. We think that the statement depends on how the system is viewed and thus how a given particle is defined. When taking the Eulerian view, i.e. when the particle index k is fixed, the model index $m_{t,k}$ is free to change during resampling. When taking the Lagrangian view, i.e. when not fixing k but following a specific particle backwards in time, its model index does not change (see also our response to the comment on line 289). In this sense, we believe that our statement "There is no stochasticity in the evolution of m_t" is correct. To clarify this, we suggest to rephrase l. 338 in the following way: "Because there is no stochasticity in the evolution of m_t though, when the particle index k is fixed,..."

RC: *Sec. 3.4 This section is essentially incomprehensible, with the section on proper scoring reading like it was pasted from a statistical methods paper. This being the Cryosphere, it's important to try to help your reader with some intuition as to what the CRPS actually means, and why its potential impropriety matters. A figure describing the metric might be useful, or perhaps a simple example de-scribing circumstances where the value is high or low. While the rest of the paper is still accessible not understanding CRPS, the analysis breaks down to 'big number bad, low number good,' which is unfortunate given that there is probably much more insight to be gained from the following sections.*
AR: We suggest to simplify the paragraph and take some statistics jargon out of it. We will also add the suggested example (but prefer not to add the figure for not increasing the manuscript's length further): "[...]It takes into account both the deviation of the median forecast from the actual observation (forecast reliability) and the spread of the forecast distribution (forecast resolution). This means that a forecast close to the observation median can receive a poor Continuous Ranked Probability Score (CRPS) if the forecast distribution spread is high, and the other way around. The CRPS is defined as (Hersbach, 2000): [...]Lower values of CRPS correspond to better forecasts. The minimum value is zero, corresponding to a deterministic, perfect forecast." We will introduce further edits to the explanation of non-proper and proper CRPS in the revised manuscript.

RC: *Sec. 4.2.1 This section is quite unclear, specifically what the differences are that these include relative to the 'full' forecast.*
AR: We want to avoid the wording "full" in this paragraph and suggest to change the sentences into: "We have run experiments where the particle filter is limited to using mean parameters and/or single models instead of parameter distributions and the model ensemble. In more than half of the experiments, the resulting CRPS values are higher than the highest CRPS obtained with the ensemble setting and time-variant parameters."

RC: *L435 Perhaps I missed it, but I can't find anything describing what the number in brackets means.*
AR: It is explained in l. 371, but indeed far from the first number occurring in this format. We will add another hint in the Results section (see also comment from Anonymous Referee #2 on this line): "(proper CRPS outside, non-proper CRPS inside the square brackets)".

RC: *Sec. 4.2.2 This section on cross-validation is very clear and good. Maybe it would be useful to comment on the temporal pattern evident in Figure 9, with CRPS in-creasing through time, but at different rates between different cross-validation folds.*
AR: This is indeed a very interesting feature. We suggest to add the following sentences: "The temporal pattern evident in Figure 9 includes an increasing CRPS through time, but at different rates between the individual cross-validation folds. It originates from (1) how representative camera stations are for the elevation band they are located in, (2) how the stations are combined in the cross-validation folds, and (3) the cumulative error characteristics, since we observe cumulative mass balance over time. To mention an example, station RHO 3 can generally be modeled with low errors compared to other stations. This is because the station is located in reasonably flat terrain with only little crevasses. The other stations are instead either in the vicinity of crevasses (RHO 4) or influenced by shadows from the surrounding terrain, dark glacier surface or steep ice (RHO 1 and RHO 2). RHO 1 and RHO 2 show that also neighboring stations can exhibit different melt, leading to a different reproducibility in the cross-validation."

RC: *L481 I don't understand where the '45 distinct model runs' come from. Also, what is a 'random coupling'?*

AR: We suggest to clarify this in the text in the following way: "The 45 model runs come from the 45 parameter sets we could gain from the calibration described in section 3.2. The random coupling is a random connection between the 45 model runs and the 10000 particle trajectories that initiate when the particle filter run starts."

RC: *L495–496 I don't understand this sentence, nor why conditioning initial conditions on observations leads to poorer results.*
AR: This is something that we discussed extensively as well. The only reasons we can think of for why the conditioning can also lead to worse overall results are the following: 1) the mass balance stakes are several meters to hundreds of meters away from the camera installations, and are thus not "true" observations at the camera locations. This might result in the initial conditions being conditioned on biased observations with respect to the camera locations. 2) Either the CRAMPON and/or the GLAMOS uncertainties might be too small or too large. This can cause either of the analyses to be over-confident. 3) The GLAMOS glacier-wide annual mass balances are interpolated from the stake readings using a model, and CRAMPON uses simplified geometries to calculate mass balances. If combined unluckily, this might lead to a stronger deviation from the GLAMOS annual mass balance when conditioned on point observations. Concretely, this means that because it was only possible to mount the cameras in our study on the lower 30% of the glacier surface area, they have a spatial bias and thus might not be representative for the vertical mass balance gradient. The three points are mentioned in the manuscript already, but we would like to elaborate more on point (2) and (3): "Overall differences to the GLAMOS analyses can be explained by [...], (2) the use of only 1-4 point observations located in the ablation zone and covering max. 30% of the glacier surface , compared to the complete network of 5-14 stakes over the entire elevation range used in the GLAMOS analyses, (3) lack of representativeness of the camera observations for the accumulation zone of the glaciers, i.e. biased vertical mass balance gradients, [...]"

RC: *Figure 11 To emphasize earlier comments again, this pattern of mode collapse is strongly indicative of an over-confident likelihood operating in an M-open framework. It's well known that Bayesian inference only 'works' when the models are correctly specified. For Bayesian model averaging (which is what the particle filter is doing in a time dependent way), this still holds: because the true physics are not contained in the set of equations that the filter has available to pick from, yet this additional uncertainty is not explicitly specified, the filter hops between the model that fits the observations in the moment. While I don't expect any additional analysis, I think it would be appropriate to make this assumption explicit in the text, and to perhaps reference it when describing the fast switching between dominant models.*
AR: This is really a valuable inspiration for future work. We agree with the comment but would add that, in our view, it is not the likelihood being overconfident, but rather the prediction and thus the prior following from that. This is because we have chosen the observational error σ_ϵ conservatively. We could experiment with likely values for β_t, which would also depend on the meteorological input uncertainties, until the Particle Filter slowly converges towards one model. However, we believe that this is critical, since values for β_t are volatile and tuning them would be subject to manual intervention. When implementing the procedure suggested for the lapse rates as our reply to the comment on l. 133, we will take another source of uncertainty into account, which counteracts the "hopping" between models. In order to include an explicit error term β_t, we would have to specify a distribution also depending on the meteorological inputs. This would be a major undertaking. Regarding the request to clarify this in the text, we propose the following addition (together with the edits according to the comment on Eq. 16):
L.307 "By introducing both the meteorological input uncertainty η and the parameter uncertainties, we shift the majority of the uncertainty contained in β_t to these variables. Since the remaining uncertainties for β_t are small and hard to quantify, we set β_t=0 for simplicity. With this assumption we neglect some additional uncertainty contained in β_t, which might lead to jumps in the temporal evolution of the model probability."
L. 515 "This model dominance, and especially the fast switches between dominant models, is describing a mode collapse. This might indicate an overconfident likelihood and/or a prior operating in an M-open framework (Bernardo and Smith, 2009) where the "true" model is not a choice amongst the available models. In our case, we believe that the ensemble prior is overconfident, since we have chosen the observational error conservatively. The filter thus switches back and forth quickly between individual models that describe the observations best. We accept the fast switching model dominance as a sign that the overall ensemble performance is improved."

RC: *Figure ?? Two things that are missing from the paper are time series' of state and parameter distributions. It would be very interesting to see the evolution of un-certainty in the predictions away from observations, and also to see how quickly parameters change or revert to the mean.*

AR: We agree that this is very interesting. We will add a figure and a small paragraph discussing this aspect in the new manuscript.

New references

- Bernardo, J. M., & Smith, A. F. (2009). Bayesian theory (Vol. 405). John Wiley & Sons.
- Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. Bayesian inference and decision techniques.