The paper is for the most part well written and I recommend it for publication after minor revisions. My major concern is the big discrepancy (up to 50% relative error) between the adjoint-based and finite-difference based sensitivities shown in Table 1. This seems large even considering numerical errors or high nonlinearities. The finite-difference based sensitivity depends heavily on the increment ε used, especially for highly nonlinear forward maps. Maybe the authors could consider trying different increments.

Detailed review:

Eq (3) In general x is not a scalar, so the definition provided here is not well defined. I would change it to something like $\delta_\varepsilon \mathcal{J} = \frac{\mathcal{J}(x+\varepsilon\delta x)-\mathcal{J}(x-\varepsilon\delta x)}{2\varepsilon}$ This is also the proper finite difference approximation of $\delta\mathcal{J}$ introduced in eq (6).

Eq (4) It is not clear why $\mathcal{L}_0$ has argument x (the control) while $\mathcal{L}_1, \mathcal{L}_2, \ldots$ seem to have the state u as argument. Is it assumed that the initial state u_0 is the control x? If so please mention this. This would be a simplified case because in the following you use the precipitation, surface and basal temperature (which are not an initial states) as a control.

Page 9, Line 14 Note that also the MALI model (Hoffmann et al.) and the FEIS model (Brinkerhoff et al.) use AD, through Trilinos and FEniCS softwares respectively.

Fig 1 I think this figure is hard to understand. What is dy? Why is it set to zero at some point in the reverse sweep? Can you please make the figure and its caption clearer?

Page 11, Line 14 This calving law can create issues with computing sensitivities, because the thickness becomes discontinuous in time.

Table 1 I think that the notation $\frac{\delta\mathcal{J}}{\delta\text{variable}}$ is misleading. In my understanding here you are showing $\delta\mathcal{J}$, computed as in (6), where x is the variable and $\delta X = \varepsilon e_i$, e_i being the unit vector that's equal to 1 at point i (for the cases 1,2,3) and ε is 5% of the initial variable value. Similarly the notation $\frac{\Delta J}{\Delta\text{variable}}$ is misleading.

Page 15, Line 21 The definition of ΔV_{adj} is a bit problematic. Instead of $\frac{\delta\mathcal{J}}{\delta X}$ it should be $\frac{\partial\mathcal{J}}{\partial x}$. Moreover the integral makes sense only in the continuous case. I think it is better to use the notation of eq. (6) and write $\Delta V_{adj} = < \frac{\partial J}{\partial X}, \delta X >$, where $\delta X = \varepsilon\, \chi_\Omega$ and χ_Ω is the characteristic function of Ω, i.e. it's 1 on Ω and 0 otherwise, and ε is 5% of the variable value. In the discrete case $< \cdot, \cdot >$ is simply the l_2 inner product, i.e. a sum over all the nodes of the grid. Does this correspond to how you computed ΔV_{adj}?

Page 15, Line 22 In the definition of ΔV_{fd}, at the denominator there should be only 2, not $2\delta X$.

Appendix While automatic differentiation of functions having corners (e.g. the absolute value or max function) makes sense, the differentiation of discontinuous functions does not makes sense mathematically (one would have to deal with the Dirac delta). For this reason I'm concerned to see the AD implementation of discontinuous functions like floor or ceiling. Why are those function needed in an ice sheet code?