REFEREE REPORT FOR THE PAPER "OF THE GRADIENT ACCURACY IN FULL-STOKES ICE FLOW MODEL: BASAL SLIPPERINESS INFERENCE" BY N. MARTIN AND J. MONNIER

The paper presents a study of the accuracy of an adjoint-based gradient in the context of inverse problems as well as the effect of two approximations of the adjoint (hence the gradient) on the reconstruction and performance of the inverse method. The inverse problem under study is the reconstruction of the basal friction coefficient in an ice sheet inverse (model) problem governed by the nonlinear Stokes equation with nonlinear (Weertman-type) basal boundary conditions. In addition to the gradient study, the authors also study the limitations of the basal friction reconstruction using the correct and approximate adjoints/gradients.

I find the topic of the paper important to be brought to the glaciology/cryosphere community. However, the presentation and writing needs some work. The most important comment/concern I have is that the authors rely heavily on a black-box type adjoint derivation, and the discussion and flow of the paper are strongly influenced by this. Throughout the paper, the study (and conclusions) are very much depending on the AD tool and it is not applicable in a more generic setup (for instance, the nonlinear solver which influences the way the adjoint is solved and consequently the computational cost). I strongly recommend writing down the adjoint equations, and point out concretely the terms dropped to obtain the approximate adjoint/gradient the authors focus on. This will be applicable for any inverse method where approximations are made to the adjoint/gradient. Below I list several issues which I hope the authors will be willing to address.

Here are my comments:

1. **Title:**
   (a) Is it "Off" or "On" the gradient accuracy ... ?
   (b) I don't think "Full" needs to start with caps F; the hyphen is also not necessary.
   (c) The title suggests that the gradient accuracy is studied for the nonlinear Stokes (e.g., for a Newton solver) rather than for the inverse problem. I suggest reformulating it, e.g. "On the gradient accuracy for ice flow inverse problems governed by full Stokes: basal slipperiness inference" or something along these lines.
2. **Abstract:**
   (a) Consistency issue in the abstract and throughout the paper for the parameter in the basal sliding law: "basal slipperiness" versus "friction coefficient". Is it friction or sliding? Since the parameter is preceding the tangential velocity (see eq. 4), I think the correct terminology should be friction coefficient, and friction law (see for instance the ISMIP paper by Pattyn et al.).
   (b) At the first read, it is extremely difficult to understand the difference

between the three variants/methods the authors are referring to: (a) complete or exact or full (!) adjoint, (b) self-adjoint, and (c) incomplete adjoint (especially the later two). Please make sure these terms are clear especially from the abstract and in the intro and be consistent with the chosen terminology throughout the paper.

(c) It is not clear to me where was the accuracy of the adjoint-based gradient studied (or assessed) in the perspective of "variational data assimilation". This never comes up later in the paper (explicitly). Why not use "inverse problems in glaciology"?

(d) Pag 3855 bottom: The comparison between the results reported by Petra et al. and Gudmundsson should be more cautious; Gudmundsson's results are infinitesimal based on perturbation theory, while Petra's results are based on a finite amplitude perturbation; it is not clear how to cleanly / fairly compare these two.

3. **Sections:**

Section 1.

(a) The first two paragraphs should be combined (the second paragraph is in fact a sentence ...). Also, please give a reference for the sensitivity of the ice dynamics to the bedrock.

Section 2.

(a) I suggest adding a reference to the forward model, e.g. Hutter (1983), Paterson (1994).

(b) The discretization details are not necessary so early in the paper. I would move the discussion on the choice of the finite elements (and any numerics related particular details) to the numerics section. (Also, finite elements should not be hyphenated.)

(c) Define $\boldsymbol{Id}$, $\boldsymbol{n}$ and $\boldsymbol{t}$. Also, what is the nt subscript represent?

(d) The reference "Martin and Monnier, 2013" is not published or available. Is there an archive version that the readers could check out? If not, please add the necessary details (in an appendix) so that the reader can follow the discussion.

(e) Pag. 3857 bottom: I am surprised that the authors chose to solve the nonlinear forward problem with a fixed point iterative method when all the ingredients for a fast Newton solver are available.

(f) It would be very useful to have an illustration of the problem setup and geometry.

(g) I might have missed it, what was the Glen's flow law exponent parameter n in the computations (for the nonlinear Stokes)? Also, how is this n related to the n present in equations (6), (19), (26), etc.?

(h) The $\boldsymbol{k}$ variable may be confusing and not necessary. Why not work directly with β? Also, why is k a vector; and later β is not? (I assume by making it bold the authors wanted to suggest that it's a vector.)

(i) Tykhonov should be Tikhonov.

(j) It is common to define both the misfit and the regularization terms with $\frac{1}{2}$ so that when one takes variations with respect to the state and control variables, respectively the 2 cancels.

(k) Vogel is indeed appropriate/classic reference for the discussion on choosing the regularization parameter. I find it strange that the authors are referring to the Martin 2013 PhD thesis in this context.

(l) The difinition of the cost should not change (e.g., $j(\beta)$ versus $j(\beta, \gamma)$), or please redefine.

(m) Pag. 3859 (lines 5): Is it the final cost or the sqrt of the final cost? In fact, shouldn't be the misfit compared to the noise?

(n) Pag. 3859 (lines 10-15): Where is the "twin experiments" terminology coming from? This seems to be just the synthetic data case with no noise. Without adding noise, these tests validate the optimization but don't allow to say much about the inverse problem; too many inverse crimes going on.

(o) Pag. 3859 (lines 25): Not sure if rounding errors are taken care of in AD.

(p) Section 2.5: I personally don't think it is necessary to describe the gradient or "scalar product" tests. However, given the audience of this journal, it might be helpful. I would still combine subsections 2.5.1 and 2.5.2 since the latter contains only one sentence. I suggest recommending additional text books for finite difference checks as refs, such as "Perspectives in Flow Control and Optimization" by Max D. Gunzburger, or the classical finite difference text book by J. Strikwerda.

Section 3.

(a) Pag. 3863 (lines 5): It is not clear what "current model" are the authors referring to.

(b) Pag. 3863 (lines 10-20): Why are there iterations for the adjoint system, which is linear? Is this just an artifact that is inherited from the fixed point solve of the forward problem?

(c) For a paper, in which the main result is the discussion of the importance of deriving and using the correct gradient, I think it is important to write down the equations actually solved by the two approximations (correct/exact versus incorrect gradient), not only discuss how their implementation in an AD framework changes. I would like to see the actual terms that are being ignored for the "self-adjoint" variant.

(d) Eq. 18: Cost has a factor $\frac{1}{2}$ here, different from before (see eq. 10). Why? Also, how is u^{obs} related to u_s^{obs}?

(e) How are the "backward" and "adjoint" problems related?

(f) Pag. 3865 (lines 5): Why not add noise to the observations u^{obs} rather than use an inexact solution of the forward problem?

(g) Pag. 3865 (lines 20-25): The authors state that: "a computational cost well below the one of the exact adjoint". It must be made clear that this computational saving is only due to way things are setup here. In fact, the adjoint problem is linear for both linear and nonlinear Stokes. The cost of computing the gradient is one nonlinear forward solve (= number of linearizations x linear(ized) Stokes - if we are talking about Newton, this number of linearizations is below 10) and **one single linear adjoint (Stokes) solve**.

(h) Pag. 3866: It is not clear why the authors have to do reverse accumulation. This sounds very much like a debugging/study of the particular AD tool rather than a scientific paper with generally useful information in it.

(i) What do the authors mean by "direct solver"? (pag. 3866 - lines 5).

(j) Pag. 3868 (lines 10): It is difficult to say / debatable if 1% noise is realistic.

Sections 4-6

(a) Eqs. 19-23: Is it possible to write these functions in a more compact/general form? (same with eqs. 27-30) Also, it would be very useful to visualize these various frequency functions (or refer to an appropriate figure from the results sections).

(b) Pag. 3869 (lines 10-15): Do i and n really belong to an interval, or should be natural numbers?

(c) Pag. 3870 (line 15): Do the authors have any intuition for why for smaller noise the difference between the exact and self-adjoint methods is larger then for high noise case?

(d) Pag. 3870 (lines 15): Define/specify n_t and L in eq. (25). Is it necessary to repeat the expression for the cost?

(e) Pag. 3870 (lines 25): The following is confusing: "approximation is a priori not valid.", Please reword/clarify.

(f) Pag. 3873 (lines 20): It is not clear what "quality identification" means.

(g) Pag. 3874 (lines 20): What/where is eq. (2.1)?

(h) Pag. 3875 (lines 20): It's not over fitting the cost, rather over fitting the noisy data.

(i) The one sentence "paragraph" in lines 20 should be connected to the next paragraph. Also, "academic situation" sounds a bit weird. I recommend using "model problem" or "synthetic test or model problem", etc. In the same sentence, it is not clear who "they" refer to.

(j) Pag. 3876: Again, the authors make it sound as if the "self-adjoint" gradient would be much easier to compute; this might be the case within their AD framework, where–for whatever reason—iterations for the linear adjoint equations are required. However, I think that is only an artifact of their implementation and not a general issue. In general, the computational cost of solving the exact adjoint and solving the self-adjoint adjoint equation should be very similar.

(k) Pag. 3879: are eqs. 32-35 again necessary? For instance, some of them have been previously defined.

(l) On the same page, I recommend replacing "very partially" with maybe "poorly" or just "partially".

(m) I find it surprising that the CPU time is brought up only in the Conclusions section. Again, I encourage the authors to acknowledge that the much favorable CPU time for the "self-adjoint" method is due to the way the forward and adjoint problems are solved.

4. **References**
   (a) Third reference, why is "Flows" capitalized?
   (b) Martin (2013) - typo: using.

5. **Language issues throughout the paper: grammar, typos, consistency, etc.** *consistency issues:* basal sliding versus friction versus signal β (for the latter, see table 3); *grammar:* pag. 3859 (lines 10-15) compute should be computes; pag. 3862 (lines 5): aim should be aims; pag. 3878 (lines 20): simulations **have** been; same sentence, it should be "one measures", also "measurement points" or "observation points" as opposed to "measure points".