We thank you a lot for the helpful comments and ideas. As suggested, we added the glacier d'Arolla from the ISMIP-HOM experiments in a time-dependent formulation. We also summarized time stepping methods for full-Stokes models. We modified the title to reflect the good performance of the Picard iteration with the approximately exact step sizes.

All upcoming line and page references refer to the new manuscript.

Summary: The paper implements a Newton method with step-size control for solving a full Stokes model describing the dynamics of glacial ice. In idealised diagnostic numerical experiments two types of step size control are compared with each other and to Picard iterations with and without step-size control.

General comments:

The chosen topic is important and the paper is well structured and seemingly without any obvious technical mistakes. My main concern is about the novelty of this paper, for the following reasons:

1. The Newton method is faster than Picard is expected and Newtons method is already used in ice sheet models extensively, even with step size control (e.g. in Elmer/Ice). It is true that Newton iterations do not converge for all cases especially without running a few initial Picard iterations, but for the numerical experiments chosen in this paper my guess is that Newton works in most existing codes already. Perhaps exact step size is new but I think in that case a brief summary of the most common step size control methods in the most used ice sheet models should be provided. Furthermore, a significant superiority of exact step-sizes compared to Armijo is not clear from this paper.

   We added a summary of the step size controls and methods (Newton/Picard) used in full-Stokes models (lines 23-31). As recommended later, we added time-dependent experiments (pages 14-23).

2. In comparison to the previous manuscript which is referred to in this paper: Schmidt, 2013 https://arxiv.org/pdf/2307.02930.pdf In Schmidt, 2013 more mathematical details and proves are given, but the method and experiments seems identical. Is Schmidt, 2013 indented to be published?

   We intend to publish this manuscript. In fact, a new version with some corrections is in the second review round at *Numerical Algorithms*. A preprint is on *ResearchSquare* at https://www.researchsquare.com/article/rs-3354498/v1 . Due to your feedback, we added time-dependent experiments (pages 14-23) and references to glaciology models (lines 23-31) to this manuscript. These changes make this manuscript more different to the mathematical manuscript.

Nevertheless, I think the paper can be published after major revisions. My suggestions for addressing my above concerns are:

1. If Schmidt, 2013 https://arxiv.org/pdf/2307.02930.pdf is not already submitted somewhere, I think the mathematical proofs here are novel enough for publication. I am not sure if the Cryosphere is the best journal for this, but possibly the mathematical proofs could be explained and summarized in the main text, and the proofs themselves be kept in the appendix.

2. Otherwise, a solution which probably fits the Cryosphere better, is to extend the numerical experiments and literature study so that the importance to glaciology is clearer. If the authors can show that it is possible to use Newtons method in realistic simulations in an effective way without initial Picard iterations, I think it is a major contribution to the community.

We follow this approach.

Specifically the following points should be included:

(a) In the introduction, go through the most important ice sheet models (especially those using the full Stokes model) and review what method they are using to resolve the non-linearity of the problem

We summarized in the introduction, how *ElmerIce*, *ISSM*, *COMSOL multipyhsics*, and *FELIX-S* solve the full-Stokes equations (lines 23-31).

(b) Extend the numerical experiments. If the aim of this paper is to show that some method is better than another and could be useful for an ice sheet modeller, it must be for a relevant example. Most importantly, a time-dependancy (free surface movement) should be introduced. This is important since the first time-step (the only one studied here) usually behaves quite differently in terms of number of iterations, as compared to later time-steps when free surface has relaxed. In a real application, there is also a contact problem between the ice and bedrock, which introduces an extra non-linearity. The more realistic example (3D, variable bedrock) the better. Also, I think CPU-times should be measured rather than only number of iterations.

We added the two-dimensional glacier d'Arolla with variable surface and also with a sliding-part as in the ISMIP-HOM experiments $E1$ and $E2$. We measured the computation times for the time-dependent experiments and the three-dimensional experiment.

If the original paper Schmidt, 2013 https://arxiv.org/pdf/2307.02930.pdf is kept as a separate paper, I think still some of the mathematical details must be added to this paper. I am lacking some mathematical justification of some of the key points of the paper, for instance the justification of the extra diffusion term in the equations, and explanation behind Algorithm 3.

We rearranged the section about step size controls (pages 5-6). We introduced the Armijo step size (lines 122-129) first. Then, we motivate the exact step sizes (lines 130-142). Additionally, we explained the algorithm for the exact step sizes a bit more (lines 132-134).

We stated the reasons for adding the diffusion term (lines 81-82) and explained this in the appendix (A1 The variational formulation).

Specific comments:

The abstract and introduction could flow better. The sentences could be longer or better connected.

We changed the abstract due to additional content and tried to connect the sentences better. We also changed the introduction a bit to get a better flow.

L 4 and 5: I think these sentences can be removed from the abstract "For the step size control, we need a minimization problem. Minimizing a specific convex function is equivalent to solving the full-Stokes equations."

We removed these sentences.

L 18: nonlinear → non-linearly?

We corrected the grammar.

L 18: make it clearer that the stationary variation of these equations is the standard one

We added, how the new shape of glaciers is calculated. This should make clearer why the stationary equations are important (lines 18-20 and in detail in section 6.2).

L 56: These spaces are for linear Stokes, see e.g. the work by Belenki et al: https://www.jstor.org/stable/41582741 for the appropriate spaces

Thanks for the reference. We added an explanation in the appendix (A1 The variational formulation) with reference to this literature (lines 359-360) and explain that the additional diffusion term μ_0 allows us to use the simpler spaces for linear stokes.

L 60: Why do you add a diffusion term? This is non-standard and not included in other ice sheet models, which might make the reader wonder if these results are applicable to their model. I suggest trying to get rid of this term.

We added a short explanation for the diffusion term in the appendix (A1 The variational formulation). Additionally, we did all experiments again with $\mu_0 = 0$ (and added this information in line 85).

However, we partly overwrote our reference solution by our methods that we compared. This led to small mistakes in Fig. 4 and 7 for Newton's method and the Picard iteration with exact step sizes. Now, both methods are a bit better for the later iterations. Additionally, Fig. 3 was slightly wrong for the exact step sizes. Now, the Picard iteration with exact step sizes has a slightly higher difference at the surface to the reference solution and Newton's method with exact step sizes a slightly lower.

Regarding the diffusion term in theory: Without the diffusion term, we have to formulate the problem in a subspace of $W^{1,1+1/n}(\Omega)^N$. For functions $f \in W^{1,1+1/n}(\Omega)^N$, we know

$$\int_\Omega |f|^{1+1/n}\,dx < \infty \quad \text{and} \quad \int_\Omega |\nabla f|^{1+1/n}\,dx < \infty.$$

This is not enough for the well-posedness of the first summand in the Picard iteration and the directional derivative of the variational formulation Eq. (6). The additional diffusion term allows us to have $f \in H^1(\Omega)^N = W^{1,2}(\Omega)^N$. Then, we know

$$\int_\Omega |f|^2\,dx < \infty \quad \text{and} \quad \int_\Omega |\nabla f|^2\,dx < \infty.$$

Now, we can apply Hölder's inequality to show $\int_\Omega |D\boldsymbol{v_{k+1}} : \nabla\boldsymbol{\phi}| < \infty$.

L 71: I would not say that Picard is standard. Include some references for ice sheet models that use Picard rather than Newton. Most probably use a combination of the two.

We added ice models, which use the Picard iteration and replaced "standard" by "common".

L 92-93: It's true that in Hirn μ_0 is 0 (and even delta=0 !) . This should be discussed - why is it not possible to set them to zero in this work?

We changed the simulations to $\mu_0 = 0$. Only for the infinite-dimensional theory $\mu_0 > 0$ is necessary. For the first few iterations, the solution is nearly identical. Some small changes are visible for more iterations.

Algorithm 3: what is the stopping criteria, i.e. how many for-loop iterations do you do?

We used 25 iterations and added this in the text (lines 140-142). One could probably optimize the number of used iterations. However, in terms of the computation time, this is not important.

L 104: For the reader of the cryosphere, I think you should introduce the two step-size control methods with a bit more words, discuss pros and cons etc, rather than just stating the algorithm.

We motivated Armijo and exact step sizes more. Additionally, we introduced the Armijo step sizes first (lines 122-129) and used them as a motivation for the exact step sizes (lines 130-142).

Equation 10-11: Does the fact that the terms are smaller really mean that one can surely say that the solution is not impacted significantly, especially in a time-dependent simulation it is not clear to me.

We set $\mu_0 = 0$ for the simulations. The term $\delta > 0$ is used in ice models. Thus, we did not set $\delta = 0$. However, as we have for typical ranges of $|D\boldsymbol{v}|$ in the computer the relation $D\boldsymbol{v} = D\boldsymbol{v} + \delta$ due to the machine precision, the δ term should not influence this region. It will have an influence on really small values of $|D\boldsymbol{v}|$, but there are also other inaccuracies due to, e.g., grid resolution, unknown Ω, B, or N. Thus, we think that a small value of δ does not change the solution too much compared to the other sources of inaccuracies.

Figure 2: I think it is misrepresentative to compare to models that are not full stokes, please only include full stokes models for the dashed lines.

We specified in section 5.3 that we only compared with full-Stokes models. (*aas2*, *cma1*, *jvj1*, *mmr1*, *rhi3*, and *ssu1*). The displayed standard deviation in Fig. 2 seems to us quite similar to Fig. 6 in [Pattyn2008], just with another scaling on the y-axis. However, we forgot the models *aas1*, *oga1*, and *rhi1* and added them calculating for the dashed lines.

About the experiments: I strongly recommend that you add a more complex example, something with more complex geometry and time-dependency. A time-dependent simulation of the arolla glacier maybe.

We added the experiments ISMIP-HOM $E1$ and $E2$ in a time-dependent formulation. For the time-dependent simulation, we had to add

$$-\int_\Omega p \mathrm{div}\boldsymbol{v}\, dx$$

to the functional in Eq. (7) because for the time-dependent problem our initial guesses are not necessary divergence-free.

About the experiments: Add plots with relative error vs CPU-time, to show if the step-size control increases the work per iteration significantly or not

We did not do a relative error vs computation time plot for two main reasons: The computation times for all methods are quite similar for the time-dependent experiments (Table 3 and Table 5). For the three-dimensional example, Newton's method itself takes more time than the Picard iteration (Table 1). However, the step size control is not computationally relevant. Thus, we think there would be too many details in the plots without giving important information. Instead, we added the tables with computation times to give this information. Secondly, one could reduce the computation of the step size control by parallelizing the calculation of the integrals. The *MUMPS* solver automatically uses more processors, if available. Thus, we have a comparision between one optimized algorithm and one proof of concept algorithm. Additionally, one could try if a lower number of bisections for the exact step sizes produces results of similar quality. Therefore, the computational effort for calculating the step sizes could be reduced compared to the complete iteration.

The outlook: This is a little bit informally written.

We rewrote the outlook by adding some possible research directions and reformulating the text a bit more formally.

Additionally, we made minor corrections due to mistakes in the reference in Schmidt, 2023: We removed the statement that the function G would be continuously differentiable and the theoretical statements about the convergence speed. The function G is only Gâteaux differentiable.