Improved convergence and stability properties in a three-dimensional higher-order ice sheet model

Reply to List of Comments

by J.J. Fürst, O. Rybak, H. Goelzer, B. De Smedt, P. de Groen and P. Huybrechts

First of all we want to thank all three reviewers for the critical and therefore useful comments they gave on the presented document. All comments are considered and helped to improve the quality of our work. In the following the responses to the reviewers comments are denoted in italic.

Review 3:

General comments

This manuscript presents a finite difference stencil with a different choice of flux points than used by Pattyn (2003) and compares the performance of a certain linear and nonlinear solver choice at a single grid resolution for a few model problems. Although the issues of discretization properties and solver convergence are mixed in the manuscript, but I will discuss them separately here.

The described stencil is a classical method for discretizing nonlinear elliptic terms, therefore it's hard to call "novel". The choice of locations at which to evaluate such intermediate quantities is a topic discussed in most books on finite difference methods, Ferziger and Perić (1996) or LeVeque (2007) for example. More recent work regarding choice of flux points has used the mimetic finite difference formalism to maintain local conservation and add support for non-smooth and unstructured meshes, see Hyman et al. (2002) for an introduction. In any case, the present formulation is a classical choice of flux points (though not usually regarded as the best—face midpoints would typically be preferred for conservation reasons), not locally conservative, and still only second order accurate. It may be better than the unconventional choice used by Pattyn (2003), but it doesn't qualify as "novel". (The approach used by "DIR" is typically dismissed by the numerical methods community due to undesirable accuracy, stability, and sparsity properties.)

If the goal is to compare the performance of two stencils, the accuracy, stability, and conservation properties should be compared in a setting independent of the iterative solver. A sequence of successively refined grids should be constructed and errors evaluated. A problem with a manufactured solution can be used to provide an exact solution or a very high resolution solution computed with a known robust method can be used as a reference. If the authors would like to establish that their choice of stencil is more practical, it would be worth giving special attention to the performance in regions with discontinuous bed slope since basal topography is rough in practice (at the spatial scale of feasible meshes). The method should be evaluated on its ability to produce non-oscillatory solutions, to locally conserve momentum (integrate stress around a control volume, the current formulation appears to be non-conservative due to the transformed coordinates, but I could be wrong), and accuracy (e.g. as assessed by grid refinement).

The iterative solver convergence results appear to be strongly influenced by artifacts of the solution algorithm and what appears to be a flawed methodology for solver tolerance. See my comments for lines 14.23, 15.12, and 19.6 for details. The linear solver especially is not robust for this problem and without further analysis, no significance should be ascribed to its erratic convergence behavior. I am a developer of the PETSc solvers package and we answer questions about similar solver issues every day via our mailing lists and support email. Such solver-related issues are documented in various books including Saad (2003); Kelley (1995); Smith et al.

(1996). While the different convergence behavior for Bi-CGSTAB with each finite difference stencil is perhaps an interesting observation, it does not warrant a paper. The present method uses no preconditioning and I am surprised that the authors were patient enough to wait for more than one million linear iterations. It is extremely common for iterative solvers, especially for nonsymmetric problems, to not converge at all without preconditioning. Regardless, the required iteration count for second order elliptic operators necessarily increases proportional to (Δx)−1 under grid refinement. This can be rectified by preconditioning, which is usually mandatory and, depending on the method, can solve the problem in a constant number of iterations independent of grid size. Additionally, the slow convergence of the Picard nonlinear iteration can be overcome by switching to a Newton iteration with grid sequencing for globalization. The nonlinear equations for every model configuration in this paper can be solved to a relative tolerance of 10−10 in at most 10 linear iterations independent of grid resolution using a Newton-Krylov-Multigrid method. A demonstration of this using a Q1 finite element discretization has been available as a tutorial in PETSc (Balay et al., 2011) since early 2010, see src/snes/examples/tutorials/ex48.c. I mention this not to cry “you didn’t cite my (unpublished) work,” but rather to illustrate constructively that the iterative solution methods in this paper cannot be considered robust or efficient by any reasonable standard, for example, compared to what is achievable by applying established methods in a straightforward manner.

First of all we want to thank especially this reviewer for the amount and quality of the provided comments on numerical concepts and methods in scientific computing. A lot of insight was gained and the manuscript essentially profited from a mathematically more precise vocabulary. It has never been the purpose to claim that the stencil we use for the discretisation is novel but if at all we thought this about the reformulation of the force balance in the two operators. However, the three places in the text, where we refer to our variant of a discretisation as novel, are rewritten.

We also agree that the presented work does not represent a state of the art method in mathematical standards but rather should be seen as ‘small’ step forward in improving an existing way of discretising the higher-order LMLa equations. A large community applies existing code that uses a discretisation that resembles the DIR scheme which exhibits the described problems. From this background, the presented work is rather a practical approach to ice sheet modelling than a mathematically thorough study on characteristics of a specific way of discretisation. The suggested discretisation could provide groups using a similar discretisation with a quickly implementable method that solves some numerical issues without replacing their established solving method or their preconditioning. The revised version of the manuscript now also gives additional references to state of the art mathematical discretisation techniques, preconditioning and solving methods for nonsymmetric systems of linear equations. The reader is therefore directly guided to relevant, general literature but also to specific advances presented in SIAM publications.

The suggested ideas on testing our spatial discretisation using successive grid refinements and detailed analysis of performance at discontinuities are suitable tools for a mathematically rigorous study. Since our manuscript represents a model development of an existing ice sheet model and since STAG is not state of the art for discretising the Stokes equation, we think it is a pure academic exercise to mathematically vindicate our approach further. Concerning preconditioning the used variant of BiCG is based on a simple diagonal inverse, which is certainly not the most sophisticated choice but proved its robustness for our ice sheet model applications.

By now, we have certainly a broader view on ongoing research in numerical method for solving nonsymmetric systems of linear equations. In the future we hope that we can pick up some of the ideas for implementation in our ice sheet models.

The paper needs a moderate amount of editing for English grammar.

During the revision process, English grammar was further adapted within the internal correspondences.

Specific comments

2.8 ISMIP-HOM offers relatively non-discerning test cases and did not establish the correctness of reported results. For the test cases relevant to the model, the range is reported for the results is as large as the mean, indicating that there were implementation errors in those submissions, casting further doubt on the significance of statistics like the mean. This sentence seems to suggest an overemphasis on ISMIP-HOM as a "benchmark" and even as sufficient demonstration that the claimed equations are being solved.

> *This comment is consistent with the latter remark on verification and validation. We agree with the reviewer that the ISMIP-HOM model intercomparison does not provide neither a verification nor a validation of the solution provided by any numerical implementation. The presented experiments have no analytic solution nor is a reliable solution provided with a very high accuracy. In consequence it is not a strict "mathematical benchmark" but as indicated an intercomparison opportunity for glaciologicists to test the feasibility of their models. The intercomparison was also meant to reveal dynamic differences between higher-order models and the more comprehensive FS models. For these reasons, we avoided the word 'benchmark' and sticked to 'model intercomparison study' throughout the text. The most prominent adjustments are listed:*
>
> ***Correction** in the abstract: 'The reprise of the ISMIP-HOM experiments indicates that both discretisations are capable of reproducing the higher-order model inter-comparison results.'*
>
> ***Correction** in the introduction: 'Thereafter results from repeating the ISMIP-HOM experiments (Pattyn et al.,2008) are presented for the two discretisations.'*
>
> ***Correction** in the conclusion: 'In general both discretisations reproduce the results presented in the ISMIP-HOM model intercomparison study.'*

3.23 Use of "entirely" is not really appropriate here. They solve the Stokes problem without further changes to the continuum model.

> ***Corrected** as follows: 'The most comprehensive Full-Stokes (FS) models solve the force balance equation without further simplifications to the underlying equations of continuum mechanics.'*

4.1 While SIA may be "feasible", this does not imply that it is "suitable". In particular, the long-term evolution is significantly influenced by flow in regions where SIA is a very poor approximation. That SIA has an a priori assumption of minimal sliding should be mentioned more explicitly in this part.

> ***Corrected** as suggested along with the comment from reviewer 2.*

4.29 These equations do not only involve horizontal derivatives, so "horizontal PDEs" is an odd choice of terms. The PDEs are genuinely 3D PDEs, but only solve for two components of velocity.

> ***Corrected** accordingly: '... remaining two partial differential equations (PDE) for the horizontal velocity components.'*

5.1 This paragraph is confusing since it mixes stability and consistency of a discretization with convergence of an algebraic solver for the discrete equations. There is no citation for "smoothing algorithms", but it is usually the responsibility of an algebraic solver to solve the algebraic equations up to a tolerance at least as small as spatial discretization error.

> ***Corrected** the paragraph. Now we distinguish between stability, consistency of the discretisation and the convergence properties of a numerical solver.*

5.13 This sentence is just not true. A discretization may produce algebraic equations for which a certain iterative method performs well, but the spatial discretization in no way provides "direct control" of the convergence properties of the iterative solver.

> ***Corrected as follows:*** *'... , an adjustment of the discretisation of the underlying dynamic equations gives direct control on the consistency of their numerical representation. '*

5.20 Mattheij et al is an odd choice of reference for "proposing" a staggered grid finite difference method. Harlow and Welch (1965), for example, would be a more appropriate reference for staggered grid FD for incompressible flow with free surface.

> *We thank the reviewer for the provided reference and it is added. Following the general comments additional reference was made to LeVeque (2007) and Ferziger and Peric (1997). While the first provides a profound basis on numerical concepts the latter specialises on the discretisation of differential equations.*
>
> ***Corrected.***

5.27 The model is not "validated" or even "verified" in this work, it is only compared to some other models that have not been verified or validated either. "Verification" and "validation" have precise meanings when applied to computation. See Babuska and Oden (2004) or Roache (1997) for details and further discussion.

> *As already mentioned under point 2.8. the reviewer correctly states that the ISMIP-HOM did not claim the correctness of the reported results. Therefore we follow the suggestion of the reviewer and avoided the terms 'validated' and 'verified'.*
>
> ***Corrected.***

7.16 Remove ambiguity by writing an equation for "horizontal gradients of the vertically directed shearing field".

> ***Corrected*** *by adding: '($\partial_x \tau_{xz}, \partial_y \tau_{yz} << \partial_z \tau_{zz}$)'.*

9.1 In most cases, the actual solutions have only one weak derivative and certainly not two classical derivatives. This is a somewhat pedantic matter because the finite difference method is algebraically equivalent to a Petrov-Galerkin method in which only one derivative is used.

> *We agree with the reviewer and want to stress that the conversion of the force balance equation to a weak formulation also allows for a uniform treatment of the bulk equation and the upper and lower boundary conditions. Consequently it prevents inconsistencies in the discretisation of the force balance of the boundary layers and the intermediate ice layers.*
>
> ***Not adjusted.*** *The weak formulation is commonly used in finite element (FE) implementations of the force balance in ice dynamics. For details we refer the interested reader to Zwinger et al., 2007 and Pattyn et al., 2008, which provides additional references to further FE ice dynamic models and standard literature on the subject.*

10.11 This evaluation of viscosity at staggered points has little to do with the Arakawa B grid which is was originally formulated to address the divergence relation in the shallow water equations. The classification of Arakawa and Lamb (1977) applies more directly to the choice of location for ice thickness nodes relative to velocity nodes. Although the elliptic equations in the LMLa model are of different character, there is still a useful analogy in the divergence operator applied to stress, however compatible (locally conservative) discretizations for this would be based on a C-type grid instead of the B grid chosen here. It would be useful if the authors could justify their rationale for choosing the present formulation instead of the more commonly used C grid.

> *We admit that the reference to the Arakawa B grid was meant to give an additional illustration of the centred computation of viscosities. Indeed we admit that this gives rise for confusion since in the original form it was meant to describe computational nodes for velocities and pressure (while the latter one translates into ice thickness). To avoid any confusion we omit the reference to the Arakawa grid in the context of viscosities.*

> ***Corrected** by removing the reference to Arakawa and Lamb (1977).*

13.9 A negative diagonal and positive off-diagonal is not a "CFL criterion". I think the authors are trying to invoke the notion of "positivity" as described in Chapter 4 of Wesseling (2001) which notes the discrete maximum principle under the sign criterion above combined with a sum of zero (needed anyway for consistency). A discrete maximum principle is not needed for stability (indeed, it does not exist for high order approximations[1]), but it is a desirable property for a discretizations, especially when solutions are not smooth. This section should use "discrete maximum principle" instead of "stability" and should not use the term "CFL criterion" which refers to an entirely different issue.

> ***Corrected** along the lines of one comment from the first reviewer. We rewrote the chapter and show now that the STAG discretisation gives a matrix in the class of M-matrices which in this particular case is equivalent to the maximum principle. The chapter has been restructured while the previously suggested condition, which now serves for invertibility, remains the same. The title was change to 'Operator Invertibility'.*

13.15 This sentence does not make sense. The compact stencil improves matrix sparsity which reduces the cost per iteration. Additionally, the larger diagonal relative to off-diagonal entries makes local smoothers more effective. This property is made more precise by the technique of local Fourier analysis which quantifies h-ellipticity, a necessary and sufficient condition for the existence of a pointwise smoother, an important property for efficient multigrid schemes, see Trottenberg et al. (2001) for details. It appears that a preconditioner is not used in the present work, but popular inexpensive preconditioners such as incomplete factorization, Jacobi, and Gauss-Seidel are more effective when when the coefficient of h-ellipticity is larger. Analyzing the present method within this context may explain the observed convergence rates.

> *We agree with the reviewer that leaving the compact stencil for the discretisation does not imply an increased error propagation. We correct this mistake by reformulating: 'Leaving the compact stencil can give rise to enhanced, spurious oscillations in the solution close to the lateral boundaries.'*
>
> ***Corrected.***
>
> *Concerning the preconditioning of the BiCG solver. It was not explicitely mentioned in the manuscript that our implementation uses a basic diagonal preconditioning. In the manuscript we provide now the information on the preconditioner. We admit that this might not be the most favourable choice and therefore we give the reference to Manguoglu et al. (2009) in the outlook. Our aim remains however to show the influence of the discretisation on the convergence behaviour.*
>
> ***Corrected** by adding: 'One prominent solver for such systems is the bi-conjugate gradient (BiCG) method (Press et al., 2003) with a diagonal preconditioner.'*

14.23 The BiCG method is not a monotone method for nonsymmetric matrices, therefore the residual can grow between successive iterations. In fact, the monotonicity of BiCG is notorious and motivated the development (and success) of the stabilized variant (Van der Vorst, 1992) which is also not a monotone method, but is less erratic. Even with a monotone method such as GMRES (Saad and Schultz, 1986) (in exact arithmetic), there is no guarantee that successive

[1] *Wesseling (2001) introduces the topic in the context of advection, for which linear nonoscillatory schemes are at most first order accurate (Godunov's Theorem). All practical discretizations for advection of higher than first order accuracy are nonlinear. The situation is different for elliptic systems, which typically have far fewer problems with oscillations, for which second order linear nonoscillatory schemes are readily available, and for which higher order linear methods are frequently practical.*

iterates should move significantly between iterations until the residual is small. Indeed, it is common for iterates to remain essentially constant for several iterations before converging rapidly. If the convergence test is actually comparing successive iterates as indicated in the text, then it is a fundamental misuse of the method. If instead the norm of the residual ||Ax − b|| is being used as a convergence test, this should be stated. A combination of relative and absolute tolerances are typically used to define convergence.

> *It is indeed necessary to specify the used convergence test of the linear iterations more in detail. A build in criteria is chosen as described in Press et al. (2003) which is accessed by itol=4. This criteria uses the maximum norm to define a relative tolerance for successive solutions based on the preconditioner and the BiCG search direction vector.*
>
> ***Corrected** as suggested: 'A build in criteria is applied that successively computes the ratio of two maximum norms. The numerator shows the precondition matrix applied on the residual while the denominator holds the current correction vector (see Press et al., 2003). A relative tolerance   defines the threshold for which the iteration is aborted.'*
>
> *Concerning BiCG and other available solver method see answer on 19.6.*

15.12 Although the term "error" is used, I am once again left with the impression that the norm of the difference between successive iterations is used to define convergence. Although the Picard iteration can be shown to be contractive under modest conditions, there is no guarantee that a small difference between successive iterations implies that the current iterate is anywhere near the solution. Indeed, I have observed this near-stagnation for the same continuum equations as modeled here, but with nonlinear basal sliding. Instead of comparing successive iterates, the residual of the nonlinear equations should be used as a convergence test.

> The convergence test in the non-linear iterations is precisely the one the reviewer states. Obviously our description is ambiguous and therefore we follow the formulation of the reviewer to specify the criteria. Moreover we avoid now the term 'error' in our manuscript and use 'accuracy' instead.
>
> ***Corrected** as suggested in section 'Non-linear Iteration': 'Instead, true Picard iterations are used until the ratio of the Euclidean norm of the residual and the norm of the solution between successive iterations falls below a certain threshold.'*
>
> ***Corrected** as suggested in section 'Residual Decrease': 'The norm of the residual between successive solutions for the velocity field during iteration is referred to as the iteration accuracy.'*

19.6 As mentioned before, BiCG is well-known to exhibit erratic convergence behavior. Divergence of this iteration says little to nothing about the spatial discretization, instead it is an indication that the operator should either use a different Krylov method (e.g. BiCGSTAB, GMRES, or many others) or use a more effective preconditioner. If the authors wish to assert that the DIR scheme is not stable, they can compute the eigenvalues with smallest real part. (This can be done efficiently for large sparse matrices using software packages such as SLEPc Hernandez et al. (2005).)

It is also worth noting that since the continuum equations are self-adjoint and uniformly elliptic, spatial discretizations can produce a positive definite matrix for which the conjugate gradient method can be used, thus also guaranteeing monotonic convergence (in exact arithmetic). This property is "automatic" for Galerkin methods such as the finite element method. Finite difference methods often do not lose symmetry due to boundary condition implementation and the handling of non-uniform grids. This is not fundamental, however, and the mimetic approach (e.g. Hyman et al., 2002) produces finite difference schemes that do preserve symmetry.

> *The reviewer surely is right that there are mathematically strict criteria for the stability of discretisations of any operator. An extensive eigenvalue analysis however exceeds the scope of a model development study. The aim of the study is to improve a discretisation that is widely used in glaciology. We still think that using the same preconditioned solver for two*

discretisations still allows a direct comparison. In our case we note that the previous DIR scheme actually can cause divergence of the solution. For the same experiments STAG does not diverge. Admittedly BiCG exhibits erratic convergence behaviour and therefore divergence might arise from this fact. But Fig. 4 clearly shows the positive influence of the discretisation on the successive accuracies. Therefore we do not agree that divergence says little to nothing about the spatial discretisation. The more regular decay in the STAG accuracy (see Fig. 4) reassures us on the influence of the STAG discretisation on the solvers convergence behaviour.

Not adjusted.

We appreciate this comment together with 14.23 since they point out insufficiencies of the BiCG method but they also provide references to mathematical improvements of this solving method and others approaches. For completeness we provide now reference to recent development of mathematical solvers as well as preconditioning and finite difference discretisations.

***Correction** in summary & outlook: 'To decrease computational costs significantly we suggest the application of other mathematical solvers as since BiCG is known to exhibit erratic convergence behaviour. A stabilised version of Bi-CG (BiCGstab) was suggested by Van der Vorst (1992) and its convergence rate was improved with ML(K)-BiCGstab applying a Lanczos process using multiple starting left Lanczos vectors (Yeung et al., 1999). But there are also other techniques to solve nonsymmetric systems of linear equations as the generalized minimal residual (GMRES in Saad and Schultz, 1986), multigrid approaches (see Wesseling and Sonneveld, 1980; Trottenberg et al., 2001), the induced dimension reduction (IDR) method (Sonneveld and van Gijzen, 2008; van Gijzen and Sonneveld, submitted) or a combination of IDR and BiCGstab called IDRstab (Sleijpen and van Gijzen, 2010). In line with replacing the solver is the use of a more appropriate preconditioner for matrix inversion for incompressible fluid dynamics (see Manguoglu et al., 2009).'*

25.15 The second term on the right hand side should use $f_{k-1/2}$.

***Corrected** but in fact it was the first term that bore the mistake.*