Peer-Review of "Short Communicaton: Motivation for standardizing and normalizing inter-model comparison of computational landscape evolution models"

Anonymous referee

July 6, 2023

The manuscript "Short Communicaton: Motivation for standardizing and normalizing inter-model comparison of computational landscape evolution models" presents a call to the geomorphologic community to devise a concept for the benchmarking of numerical landscape evolution models. The authors give a short overview make why these efforts are necessary and provide a inter-model comparison of potential numerical artefacts connected to timestep length that can affect steady state computations of landscapes between three model softwares that implement a simple detachment-limited fluvial erosion law.

After reading the manuscript I am not quite sure what the authors want to accomplish overall. The manuscript appears a bit unfocused and seems to change scope midway. The introduction makes a good case for the need of benchmarking of LEM models (which I will not dispute, although my idea of what should be benchmarked seems to differ significantly), but then offers a partial introduction into the basics of numerical methods and a demonstration of issues than arise mainly by the choice of numerical schemes and by neglecting the stability criterion.
I do understand that the authors try to span a bridge here between model users and model developers by including a detailed introduction of basic numerical modelling theory. The authors comment on the need of teaching beginners the correct handling of model, but I do not think that this is a topic of scientific research and better suited for presentation in a book, where the theoretical background can be sufficiently covered and a detailed demonstration of known numerical issues can be better appreciated. It seems that the actual topic of benchmarking was completely sidetracked by the need to explain everything from scratch. Additionally, the discussion is still missing salient points in my eyes and does not even begin to treat LEM algorithms that have a big influence on the results presented here.

The concept of benchmarking that is executed here does not seem to be adressing the needs that are put forward in the introduction. The authors define the scope of their concept of benchmarking to a "known answer (that) may derive from the analytical solution to the model governing equations" (L. 21f) and the idea that "model performance on benchmark problems indicates (that) the model can reliably solve that type of problem" (L.23f).

Concerning the first statement, I expected that for benchmarking purposes as they are referenced in the introduction, model results should be compared against a "solution" that is independent of the model, for example a natural dataset or an independent analytical solution, or even against a specific model that is considered state of the art. Deriving a prediction from the model equations itself and testing it just tells us if the model can solve its own model equations correctly, but gives us no measure how well we approximate the natural process. To be frank I expect that model developers test routinely if their model works technically before release. Unless a formerly unknown problem turns up, I do not see the benefit of revisiting these test.

Concerning the second statement, the first issue that should be adressed is which transport law comes closest to reproduce natural processes. There are different models around that include or exclude sediment transport and have different concepts to deal with how sediments are transported and deposited. I at least expected a short discussion that highlights why the detachment-limited model is considered a good candidate. If purely detachment-limited erosion turns out to be the worst approximation, comparatively small deviations that arise from different numerical implementations probably do not matter so much anymore. Concentrating on a specific transport law before being sure that it is a useful law is a bit like saddling the horse from behind.

Another issue with the tests is the focus on purely numerical issues concerning accuracy and stability that arise from the discretization scheme. These problems are not particular to LEM software or numerical models, but to every mathematical and numerical scheme that uses discretized difference quotients and iterative steps. I think it really should not be surprising that the quality of model results is compromised when i.e. the temporal resolution of the model is coarsened too much. Also I do not see the sense in demonstrating that ignoring an established stability criterion for dt leads to wrong and unusable results. To be frank, I learned about these issues in my first lecture on numerical methods and every book on the topic covers these issues and how to handle them. Most of the shown issues are vastly improved or even disappear if a reasonable dt is chosen. (I will point out a few examples in greater detail in the line comments further down to illustrate this point.) Actually, I expect that a benchmark experiment for inter-model comparison is designed in such a way that numerical inaccuracies arising from the scheme are reduced as much as it is possible and sensible, so that the actual model differences with regard to model equations and algorithms are brought out. As a model user interested in the capabilities of other models, the comparison of essentially wrong results or results obtained outside of a sensible parameter range does not help me much. Also there is a massive backlog of mathematical literature that dates back decades (The associated paper from the namesakes of the Courant-Friedrichs-Lewy criterion is from 1928!) that is basically ignored in the discussion. The convergence and accuracy of discretized schemes with regard to their analytical solutions (especially concerning the advection or wave equations that the detachment-limited model belongs to) are better understood than shown here. What I also miss in the discussion is a full comparison of the model, including the algorithms (flow routing, handling of sinks, etc.) that actually make these models "landform evolution models" and not just "advection models". That these other algorithms play quite a big role for the result of identical experiments becomes very clear in Figure 2 and 3, where the difference between results using different software is quite obvious, but the the same model software just using different numerical schemes produces very similar topographies.

The practical benefit of the tests is additionally not clear to me. What is the significance of testing the time to steady state or comparing different steady state solutions that evolved from random initial topographies if we do not have a measure which of these solutions is the "truest"? The model results differ only by divide migration, but all results are valid steady state solutions. What the study lacks here is an independent solution that gives a measure which of the LEM softwares gives a closer approximation of nature.

In conclusion, the authors essentially demonstrate that the selected models work as they are intendend, which is reassuring, but not surprising considering that the model softwares are used for quite some time. The study mainly focuses on the investigation of different numerical schemes that solve an advection equation, but cannot add anything new here. Known issues are not adequately discussed and relevant theoretical background is often omitted. The shown results essentially confirm established mathematical theory and demonstrate known issues. Many of the differences shown here are the direct result of using a dt that is intentially unstable or too large to adequately resolve changes in the process rate. For an inter-model comparison, I expect that models are handled correctly in such a way that accuracy is optimized and on equal grounds. If this issue is remedied, I do not expect that there is much left to discuss, considering that the models are very similar and that the authors are unwilling to discuss the influence of other model algorithms. The latter is also part of the reason that I do not see a real inter-model comparison of LEM models here because most of the algorithms that affect the outcome of different models are not discussed at all. Additionally I miss an independent criterion for the quality of results. Differences are pointed out, but there is no measure that highlights one model over the others.

Based on the overall impression I recommend rejection of the manuscript.

The following section provides a few line comments that underline the reasons that led to my recommendation to reject the manuscript. It is not exhaustive, I focused on examples that underline my criticism. Additionally, since the authors expressively state that they want to initiate a discussion among the greater community, I felt motivated to add a few thoughts that arise from my own modelling experiences.

1 Major Remarks

1.1 Title

Is this a short communication? The structure of the manuscript, especially the terminology section and detailed explanation of numerical algorithms and model equations of the different models gives a strong impression of a review paper.

1.2 Introduction

L.21: The comparison against predictions stemming from the model equation just tells if the model can correctly reproduce the assumptions that are put in at the beginning, but gives no indication if the model equation can approximate the natural process. I specifically miss a discussion on expectations of how well the detachment-limited erosion law represents nature.

L.33: These references are maybe misleading as role models for this study, since they are concerned with benchmarking models based on different governing equations and comparison against independent data like natural datasets (historical climate data) and analogue experiments (critical taper sandbox models).

L.51: I am puzzled about the "statistical" nature of deviations in outcomes of the same model simulation used in different model softwares. Do the authors do not want to imply that the models have a random component? I assume that LEM simulations give reproducable results as long as no parameters are changed.

Terminology section This section seems more appropriate for a review paper or tutorial and might not be useful here. I understand that the authors want to bring everybody on the same page and want to open the discussion on appropriate benchmarking experiments for members of the community that are not intimate with the internals of numerical models, but I expect that people interested in computer models have a good idea about at least some of the terms.

Figure 1: I have trouble with this diagram. Assuming that the colors indicate groups of models that are compared against each other, it seems that the CHILD model is not compared against anything?

L.189: The variable A(x) is spatially uniform?

L.195f: The following introduction to numerical schemes seems not to be really used at a later point to explain the signatures and differences between the modeling softwares. Is it really necessary?

L.196: The explicit scheme is not limited to regular grids and the discretization of the raster with a constant dx is not valid for the voronoi, which has a variable dx. I do not think this section should be expanded, but it does certainly not help if the information is incomplete with regards to all the model variations that are tested here.

L.203: I think the implementation of this particular implicit scheme was formerly proposed by Hergarten & Neugebauer (2001, doi 10.1103/PhysRevLett.86.2689).

L.236f: In Fig. 2 and 3 and thought it seems quite obvious that these differences in other model algorithms seem to be responsible for the larger part of the differences between the results of different model softwares. The authors note themselves in L. 321f that the strong similarity in river networks for the two TTLEM simulations is likely due to the same algorithms for flow routing, etc.. Considering that the TTLEM implemented the same implicit scheme from Landlab, the best explanation for the large differences there is the great importance of these other algorithms. How can the authors uphold their confidence that these other factors can be ignored? From my own experience I expect that differences in flow routing, calculation of discharge and handling of sinks will strongly affect the migration of drainage divides.

L.290: The merit of purposely working with a dt that results in an unstable simulation eludes me. Unstable means that the results systematically drift away from the analytical solution and errors accumulate, so the results should be considered wrong and put into the bin. I am absolutely not a fan of treating instability as a kind of inaccuracy, because this implies that results are still considered to be somehow usable. Maybe I get the intent of the authors wrong here, but establishing that model results are unusable and then continuing to evaluate them sends a wrong signal here.
Another thought: 25 or 100 kyr are impratically (not to say ridicilously) large timesteps. Under changing boundary conditions, erosion rates undergo a gradual change and the time resolution must be fine enough to adequatly resolve that. Otherwise the result can be such stair stepping in the river profile as shown in Fig. 7 later in the manuscript.

L.325: The predicted slope-area relationship in steady state is essentially derived directly from the model equation for the SPPE model. The TRT model, however, essentially adjusts how the erosion rate is calculated to counteract numerical diffusion that smears knickpoints. Doesn't that mean that the analytical model prediction must also be adjusted here to make a comparison on equal grounds? Alternatively, are slopes and catchment sizes recalculated/averaged according to their usage in the model equation? If not, that might at least explain the scatter. I wonder about the flattening, but my first guess would be that this is the result of adjusting the calculation and not numerical inaccuracy. If the authors actually want to imply that the predicted slope area relationship is somehow reprasentative for natural systems, I definitely missed the discussion (and

I would probably disagree, remembering natural datasets where the relationship breaks down at small drainage areas).

L. 329f: It is not surprising that numerical models with slightly different algorithms produce slightly different results even if the model equations are the same or very similar. The authors do not present a discussion that pins down the differences adequatly. Another thought occurs: how do we know which of the solutions is the "truest" without an independent measure, e.g. from a natural dataset or a pointwise analytical solution or something comparable? The slope area plots imply that the topographies are equivalent in their properties. So what do the observed differences even mean?

L.334: Just an innocent comment, but I do hope that any decent model developer runs these test routinely before releasing a model and that it is not necessary to officially confirm this. Does not hurt to run these test especially when new to the software, but if everything is in order the main benefit was getting used to correctly handling the software.

L.335: Model accuracy strongly depends on grid and time spacing (e.g. the smaller, the more accurate), so most models should technically accomplish an arbitrary level of accuracy (in the bounds of round off and truncation errors). The question is not so much if a model can achieve a certain accuracy, but rather how long it takes to compute. It would have been nice if this aspect of the models was discussed.

Section: variable time step simulations I do not see the point in this demonstration with different timesteps. As mentioned before it is a fundamental property of discretization schemes that they are more accurate the finer the resolution, so differences in model results are expected. If the authors could present a method that allows to quantify the absolute accuracy of each model, the models could at least be ranked by the dt that is necessary to achieve a certain level of accuracy (computing time!).

L.344: I think this is the one part of the manuscript that troubles me the most. Although it is clearly stated that the solution is unstable, the results are repeatedly treated as if it is merely a matter of accuracy in the following part of the manuscript. The main difference is that in a stable simulation, the numerical approximation converges to the analytical solution and accuracy can be estimated and is well-behaved. Errors in unstable simulations accumulate over time and the results are essentially broken. The hole in the DEM shown in Fig. 5d is a very clear indicator (looks like NaN values?) and it should be clear that this data cannot be used, regardless of whether a part of the topography generated at smaller catchments sizes looks halfway plausible. Again, I also do not see what this demonstration adds to the topic of the paper. The shown stability issue is well-known for the type of model equation used here and the problem can be avoided by simply obeying the stability criterion.

L. 358 I do not see what the authors mean by instability is "not an artefact of a voronoi grid". Is this meant to mean that the voronoi grid does not show the problem in an obvious way or that the instability is not a result from using a voronoi grid? The latter is not to be expected. The stability actually surprises me. Voronoi grids have a variable dx, so a slight shift in the stability criterion is to be expected depending on the combination of catchment area and dx, but this does not explain why the result looks so good. I assume that dt is lowered internally here as well to ensure stability, in which case it is not surprising at all that the Voronoi model works relatively well even if a much too large dt is used as input.

L. 358 It is a hallmark of the implicit scheme that it has much better stability than the explicit scheme. The stability criterion is derived for explicit schemes, so it is not surprising that it is not binding for an implicit scheme, either.

L. 363 I assume that this issue arises mainly when starting from random values. Erosion rates are essentially random in the beginning. So the relative lowering of neighboring sites in one erosion step strongly depends on dt. Accordingly, the site that ends up as a lower neighbor at each timestep also depends on dt and how flow directions change then also depends on dt. I expect that this will have a discernible effect on divide migration and river network organisation. But what is the harm of that? If we generate topography from random values, the main goal is probably to obtain a valid steady state topography, not a specific network configuration.

L. 377 Instability should not be mixed up with inaccuracy.

L. 380 The smoothing of knickpoints is caused by numerical diffusion. I do not see a "sharp transition" even for lower timesteps and I do not expect to, since the problem cannot be solved by decrease in dt. That is the reason for the more complx knickpoint preserving algorithm that was also tested by the authors. Numerical diffusion is mentioned in the model decription there, but it should be brought into the right context again here.

L. 390 Maybe the statement falsely connects to the next sentences, but: Since you establish in the following that the stair stepping is the result of 12.5 / 50 m instantaneous uplift at the beginning of each timestep, why is it a signal for numerical instability? Looks to me that the model does exactly what it is programmed to do. I assume that uplift for the whole timestep is added first, creating the large step. The subsequent part of the computation then calculates fluvial erosion in multiple steps for the same timestep (decreased dt for a stable solution). It looks to me that the knickpoint generated by the uplift is advected in upstream direction at roughly the correct velocity and it also falls closely to the predicted slope area relationship. What else can the model do but produce steps under these circumstances? I recall that stairstepping is actually a sign of reaching the threshold of stability, as it is also mentioned later in the manuscript (L. 469), but I do not think that this is the case here. Otherwise the authors should be clearer as to the nature of the signs of instability that they see in the diagram.

L. 399 After the whole effort of using the stability criterion, it seems problematic if the models have the capability of reducing dt internally in order to intercept the error of the user and improve stability. It explains why so many of the results for too large dt look indecently good, but it also means that the dt put into the model is not the dt that was actually used and the observations regarding relative stability are tied to essentially unknown dt. Again it seems that the crucial algorithm that affects result is not properly discussed and interpreted.

L. 443 The graphs for the lower dt that result in stable simulations are either overlapping or very close together, suggesting that the time to steady state does not vary significantly if dt is chosen within reasonable bounds. That the larger dt take longer to reach steady state is probably owed to the stairsteps that are carved into the landscape. That should take some additional time to smooth out. Since we already found several good reasons why these larger dt are very unpractical anyways, they should probably be left out of the comparison. The comparison is not on equal grounds and unjustifiably and artificially distorts the differences between the models.

L. 461 I see similar oscillations in my own steady state topographies that are probably due to the same cause. I use a different software, but the same model equations and similar algorithms. The effect becomes more pronounced on larger grids. I do not think that the issue is a mark of inaccuracy, I rather suspect that the problem lies with the discretization of the grid. In a steepness index map, there are very localized ks anomalies of a few pixels size (both too high and too low in direct neighborhood) at some of the peaks. It looks to me that the tributaries meeting at the peaks cannot achieve a steady state elevation that satisfies the steady state requirement for all streams simultaneously. I see a relatively fixed and small number of changes in flow directions and it seems to me that the grid just switches between two or maybe more states that oscillate around the unattainable ideal steady state, but the changes seem to be restricted to these tiny anomalous areas. I assume that it is mainly a consequence of a fixed gridspacing. Stream lengths cannot add up properly at the one end and the "residuals" concentrate at the drainage divide. A variable dx (or a finer subgrid that allows a better control for the location of the drainage divide) would probably solve the problem.

L. 463 Time to steady state metrics are better comparable if dt is chosen in such a way that the accuracy of the numerical scheme is good enough that it does not significantly influence the results. I am not sure how sensible it is to do so as a concept as long as we do not have an independent measure what the "correct" timespan should be and while the reasons for delays are left largely undiscussed.

L. 469 At this point it should be mentioned that accuracy comes at different levels and there is no black and white concept of wrong or right here. The smaller dt, the more accurate the solution (within the limits of round off and truncation errors, etc.). Also, this is one of several examples where I am unsure if the authors use a term in its mathematical/computational or colloquial sense.

L. 474 Just no. Accuracy can be tweaked according to what we need as long as we know that the solution is well-behaved and follows the analytical solution, but instability is an avoidable and unpredictable error source and should always be reduced as far as possible.

L. 495 As far as I see it the diagrams show quite clearly that numerical results obtained using a dt below the Courant criterion are very similar compared to results obtained with a larger dt... And just to clarify, the criterion only ensures stability of the numerical model and does not ensure a specific accuracy. The differences observed in the results for different dt can be eliminated How much farther dt must be reduced for a certain level of accuracy depends on the numerical scheme. I expect that the explicit scheme needs a higher reduction than shown here, but the results for the TRI, TRT, LVI and LHI overlap quite nicely and indicate that the model result probably would not change significantly anymore if dt was decrease.