In SCHISM and test casesTest_GEN_MassConsv*

Mass conservation in SCHISM and test cases

Recently both Richard and Hai raised some questions about mass conservation in SCHISM, especially when settling/swimming vel is present. In light of this, I've created a branch settling velocity and added 2 new tests (Test_GEN*) to the suite to help us collectively diagnose the issue. My expectation is to merge this dev branch back to trunk after we test it for sediment transport.

Mass conservation statement in SCHISM

The mass conservation in SCHISM hinges on volume conservation in each prism:

$$\sum_{j\in S^-}|Q_j| = \sum_{j\in S^+}|Q_j| \quad (1)$$

where Q_j is the *outward* flux from a particular prism. This ensures constancy (i.e. *C*=const is a solution for advection eq.) and often monotonicity, and underpins volume conservation (note that volume is mass with concentration of 1).

The FV eq for the transport eq. is (focus on advection only):

$$C_i^{n+1} = C_i^n - \frac{\Delta t}{V_i}\sum_{j\in S} Q_j C_j + \cdots \quad \text{(i=1,..., ne)} \quad (2)$$

where C_j is a properly defined face value at face *j* (e.g. upwind etc). This is the starting point for all schemes (WENO, TVD) even though the high-order schemes become more complex in form. Note that the only requirement for C_j is that it's unique at a face, and it's the surface prism value when *j* is the surface so it can be either upwind or downwind depending on the surface movement. Similar ambiguity exists at all horizontal and vertical boundaries. However, at the horizontal bnd, the flux is either 0 (land) or the B.C. is prescribed for inflow, and so the upwind value is given. The flux at bottom is not always 0.

Summing up prism by prism (over *i*) and denoting horizontal open bnd's as HOB, we have:

$$\sum_i V_i C_i^{n+1} = \sum_i V_i C_i^n - \Delta t \sum_{j\in FSUBOT} Q_j C_j - \Delta t \sum_{j\in HOB} Q_j C_j + sources \quad (3)$$

Note that V_i is from previous step n, not n+1. The 2nd RHS term is because $Q_j C_j$ cancel out at all vertical faces except at surface/bottom. This term represents the contribution from surface/bottom flux and is

supposed to account for the movement from *n* to *n*+1. However, this balance is not precise (time truncation error) (in the case of itr_met>2, there is also additional split error; all schemes also have round-off errors). Therefore mass conservation is only good up to time truncation error, and so conservation error is a function of accuracy. On the other hand, the exact mass balance should be:

$$\sum_i V_i^{n+1} C_i^{n+1} = \sum_i V_i C_i^n - \Delta t BOT - \Delta t \sum_{j\in HOB} Q_j C_j + sources \qquad (3b)$$

Because F.S. movement does not change mass, but some bottom exchanges (BOT) like in sediment will change mass. Note that adding a constant to all vertical fluxes should not change mass.

If we treat the settling vel. implicitly as Richard H. proposed with upwinding:

$$\frac{\partial(w_s C)}{\partial z}|_k = \frac{(w_s)_k C_{kup} - (w_s)_{k-1} C_{(k-1)up}}{\Delta z_{k-1/2}} \qquad (4)$$

and take care of the surface/bottom b.c. (e.g., no flux and zero out the settling vel there w_s=0 at *k=kbe* or *Nz*, or mixed b.c. at both boundaries), the conservartion statement is the same as (3). However, large settling vel would create stability issue even with implicit scheme as the matrix is ill conditioned at bottom/surface (no longer diagonal dominant because w_s=0 there). This seems especially severe for upward swimming. Decreasing dt helps.

Eq. (3) is only the general case; we discuss in detail different transport solvers in SCHISM below.

Conservation statement for itr_met=2

This option actually offers 2 sub-options: upwind and TVD, and there are differences between the two that have some implications for conservation checks.

The starting point for both options is:

$$C_i^{m+1} = C_i^m - \frac{\Delta t'_m}{V_i}\sum_{j\in S} Q_j C_{jup} + \frac{\Delta t'_m}{V_i}\sum_{j\in S} |Q_j| \frac{\varphi_j}{2}\left(C_i^m - C_j^m\right), \quad (i=1,\ldots, \text{ne};\ m=1,\ldots,Nt) \qquad (5)$$

where is $\Delta t'_m$ the sub time step, C_{jup} is the upwind value except at surface (where it can be downwind value when the surface is falling), and the limiter $\varphi_j = 0$ at all 3D boundaries. The time level for is intentionally unspecified, as it depends on upwind/TVD choice. Note that Eq. (5) is different from the formulation used in the trunk code although both are mathematically equivalent due to volume conservation. However, Eq. (5) avoids truncation errors in the volume conservation and therefore gives better conservation results (changed implemeted in the branch).

Pure TVD case

In this case we have $C_{jup} = C_{jup}^m$. Summing over all prisms *i* gives (since $\varphi_j = 0$ at all 3D boundaries):

$$\sum_i V_i C_i^{m+1} = \sum_i V_i C_i^m - \Delta t'_m \sum_{j\in FS} Q_j C_{FS}^m \qquad (6)$$

where '*FS*' stands for free surface. Summing over *m* gives:

$$\sum_i V_i C_i^{m+1} = \sum_i V_i C_i^m - \sum_{m=1}^{N_t} \Delta t'_m \sum_{j\in FS} Q_j C_{FS}^m \qquad (7)$$

Eq. (7) is the conservation statement.

Pure upwind case

Eq. (5) becomes:

$$C_i^{m+1} = C_i^m - \frac{\Delta t'_m}{V_i} \sum_{j \in S} Q_j C_{jup} \quad (8)$$

And we further split the faces into horizontal and vertical faces, since we treat the latter implicitly:

$$C_i^{m+1} = C_i^m - \frac{\Delta t'_m}{V_i} \sum_{j \in S_H} Q_j C_{jup}^m - \frac{\Delta t'_m}{V_i} \sum_{j \in S_V} Q_j C_{jup}^{m+1} \quad (9)$$

Summing over *i* and *m* gives a similar conservation statement:

$$\sum_i V_i \, C_i^{m+1} = \sum_i V_i \, C_i^m - \sum_{m=1}^{N_t} \Delta t'_m \sum_{j \in FS} Q_j C_{FS}^{m+1} \quad (10)$$

with the only difference in the implicit term in the last term.

Conservation statement for itr_met=3,4

TODO. This case involves split and it's not easy to convert into the upwind form in (5). However, tests below demonstrate the conservation is still good.

Sediment

After consulting with Delft3D manual, the vertical b.c.'s are:

$$\kappa \frac{\partial C}{\partial z} + w_s C = 0, z = \eta \quad (11)$$

$$\kappa \frac{\partial C}{\partial z} = -E, z = -h \quad (12)$$

i.e., there is a balance between diffusion and settling at boundaries, *E* is erosion flux (known). The presence of settling in (11) is crucial for the matrix conditioning (which is missing in upward swimming case). The derivation of Rouse profile also requires this extra term.

However, the presence of settling in (11) results in extra terms in the conservation, which now must include a bottom term as mass is leaked into bed. Conservation statement needs to be revised (Hai).

I've tested this new approach with trench migration and results are similar. More tests are needed (Hai please do those).

Upward swimming

In the case of upward swimming, the truncation error is much larger, because large concentration/gradient is only found near surface, which exacerbates the error due to surface movement. So we'll have to take care of this case with some special treatment.

With no-flux b.c. at bottom and surface and 0 swimming vel (w_s) there, the discretized eq @surface prism is (including diffusion term):

$$C_N^{n+1} = C_N^n + \frac{A_i \Delta t}{V_i}\left[-\left(\kappa \frac{\partial C}{\partial z}\right)_{N-1} - (w_s C)_{N-1}\right] \quad (13)$$

where we have used upwind value for *C* in settling term. Note that $w_s < 0$ in this case. Therefore the off-diagonal is enhanced by the settling term without a compensating term in the diagonal, which may lead to oscillation or instability. This is what I observed for Case 1 over a long period of integration.

I think this case is mathematically ill posed, due to the sudden shutdown of w_s near the boundaries. A proposed approach is to impose b.c. like (11) at the level *N*-1, since the diffusivity κ is somewhat arbitrary. Any other ideas?

A global mass correction scheme

From Eq. (3) and (3b), we can develop a global mass correction scheme as follows. The 3rd terms in Eq. (3b) can be calculated to give the exact mass increase (from net inflow from horizontal open boundaries) (ΔM_0) during the advection process. If the splitting method is used(itr_met=3,4), we compute the mass change, ΔM_1 after 1st and 2nd substeps (advection). The difference $\Delta M_1 - \Delta M_0$ represents the 'error' from advection. Then we can compute the mass before and after the surface level adjustment from n to n+1 (call levels*()) and thus the difference, ΔM_2, which represents another source of 'error'. Therefore the total deficit is

$$D = \Delta M_2 + \Delta M_1 - \Delta M_0 \quad (14)$$

Normally the 2 errors are of opposite sign and roughly cancel out each other due to surface movement but as stated above the balance is not precise. Once we know the deficit, we can correct the tracer concentrations at each prism with a constant inflation factor, γ:

$$\widetilde{C_i} = \gamma C_i^{n+1} \quad (15)$$

that satisfies:

$$\sum_i V_i^{n+1}\left(\widetilde{C_i} - C_i^{n+1}\right) = -D \quad (16)$$

Therefore:

$$\gamma = 1 - \frac{D}{\sum_i V_i^{n+1} C_i^{n+1}} \quad (16)$$

This scheme obviously will violate constancy but since the total mass is a large number the correction factor is very close to 1. The benchmark test results demonstrate that the scheme is able to enforce conservation to machine precision.

Test cases

We have many test cases that show the conservation for T,S is good with reasonable model setup (i.e. as long as the results are accurate enough), at least for short term. Below are new results for long term.

Case 1: lock exchange with wind (Test_GEN_MassConsv in test suite)

The set up is in a closed box, with time-varying surface wind (nws=1). Transport scheme itr_met=3 with eps1_tvd_imp=1.e-9. A generic tracer (GEN) is used with i.c. of 1 everywhere with various settling vel.

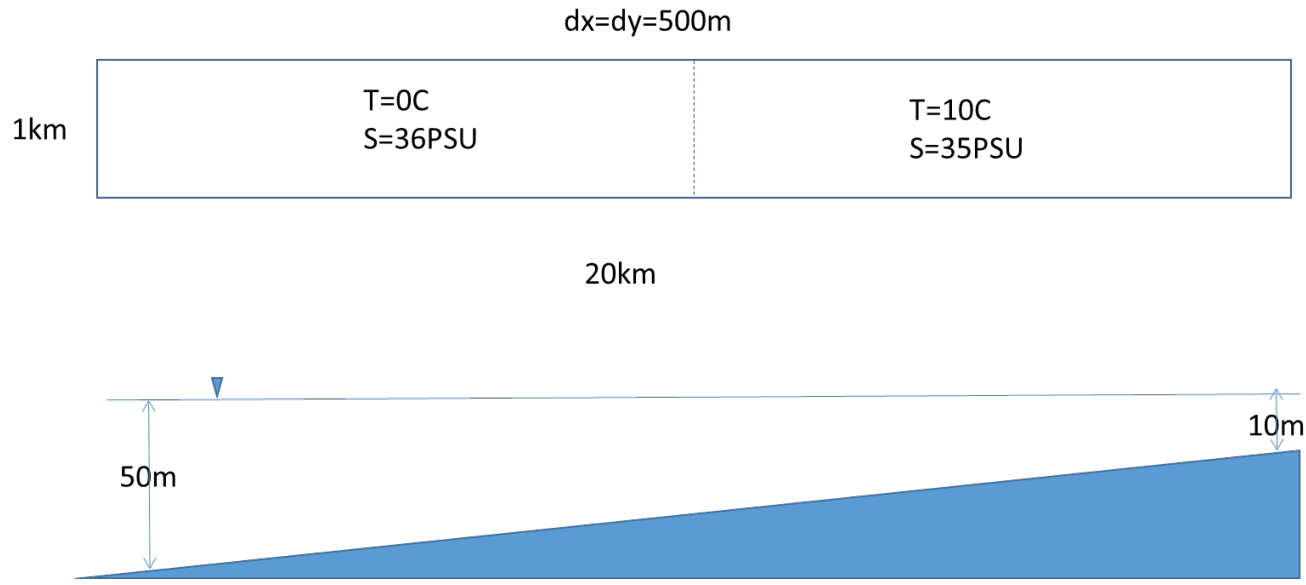

Fig. 1: model setup. Initially both T,S have gradients.

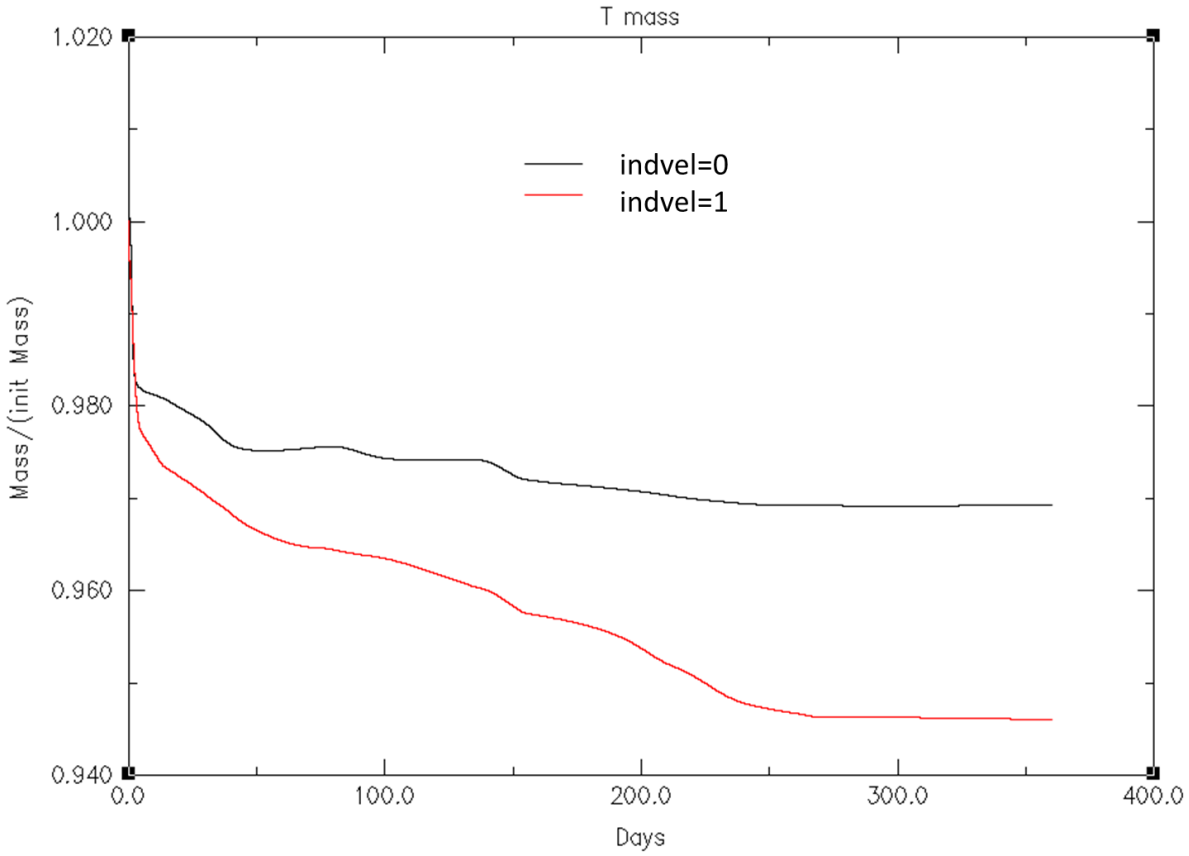

Fig. 2: time series of total T mass ratio (S mass shows better conservation). Max error<6% over a year.

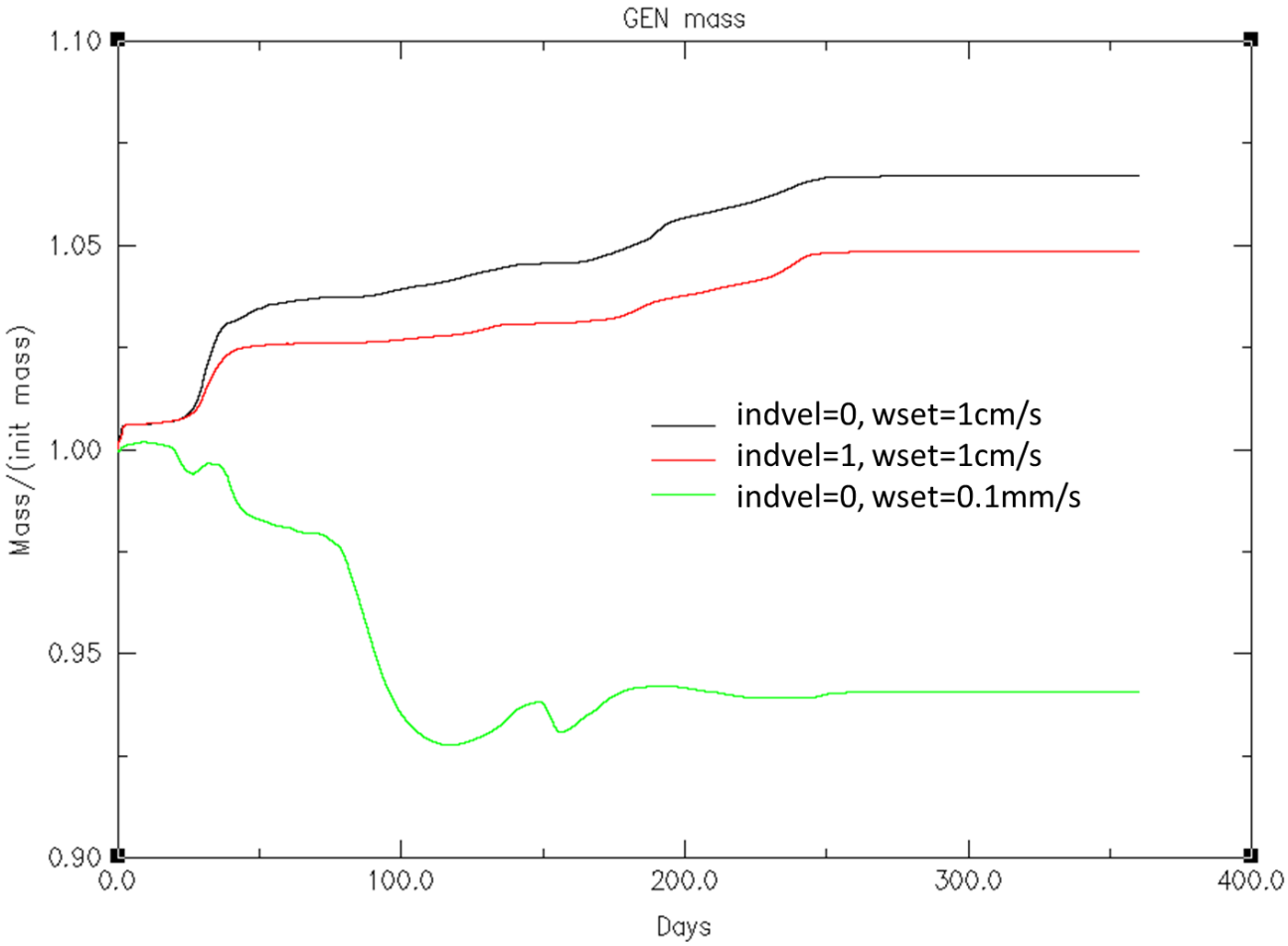

Fig. 3: time series of total GEN mass ratio. Max error=8%.

Discussion: also tried with an ambient stratification of *T* initially, and itr_met=2, with better results. Note that large gradients still exist for GEN at the end of run (all settled near bottom), and also that GEN mass may increase/decrease depending on parameter choices. With dt=200s, errors are larger (and it even triggered instability for GEN with larger w_s) so it should not be used in practice. The convergence tolerance for TVD[2] needs to be smaller for larger w_s for stability.

Results for mass balance at each time step

Here we check Eq. (7) and (10) at each time step. If we correct the total mass with the surface terms there (last term on RHS), the masses at *n* and *n*+1 should be the same before level update. The balance is at machine accuracy level (with max=1.e-14 [mass unit] for both), and so the code is doing what it's designed to do. Of course the time truncation error is still there.

Case 2: Tidal channel (Test_GEN_MassConsv2 in test suite)

The setup is similar to Case 1 except with an open bnd at the left end. The channel is of dimension 50km(L) x 2km (W), with a linear slope from 100m@x=0 to 40m @ x=40km. There is a M2 tide of amplitude 0.5m at the open bnd. One class of GEN tracer is initially placed near the right bnd with a settling vel. T, S are constant. The total mass should be conserved until tracer reaches the open bnd, which did not happen in 300 days.

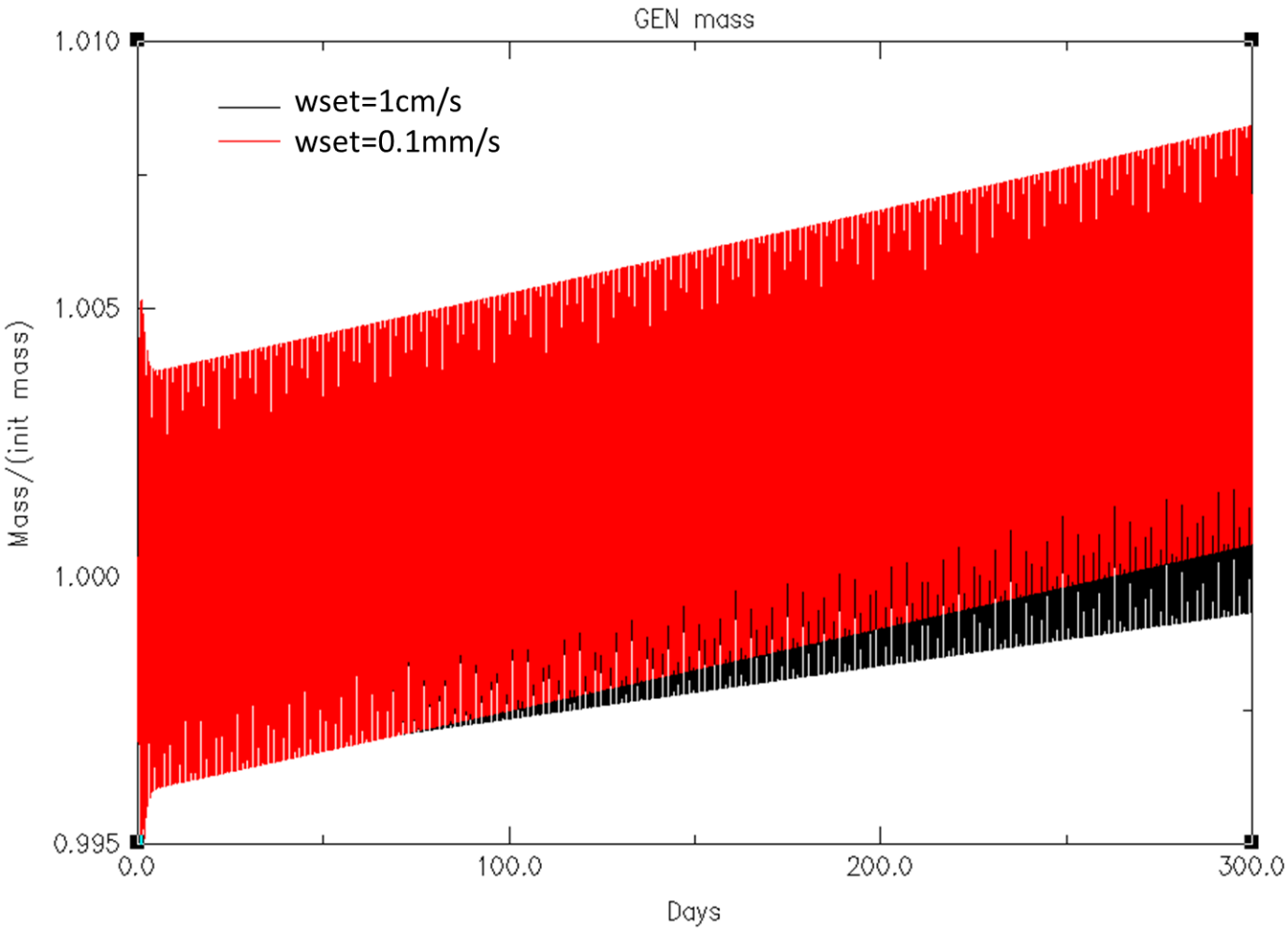

Fig. 4: time series of total GEN mass ratio. Max error<1%.