Referee report for 38282, by Frenk et al. The Santa Barbara cluster comparison project: a test of cosmological hydrodynamics codes

This paper is an attempt to compare various cosmological hydrodynamics codes to each other on the same problem. There are far too many uncontrolled variables in this experiment to compare the codes to one another meaningfully, and the authors have chosen a test case whose answer is not known. There is, however, some intrinsic worth in having a weak comparison when creating new codes. The project is much more useful as a test of *research groups* (rather than of codes), as it enables critical evaluation of other studies the authors are involved in. Since the author list contains most of the cosmological gasdynamics community, this is of substantial importance, and an improved version of this paper should definitely be published.

This paper could use some improvements to the flow, especially when the codes are described. Those sections read as though each was written by a different author (as I suspect they were). Some of the code descriptions are not specific to this cluster comparison project, and this needs correction.

Publication of the improved version of this paper is probably more appropriate in the Astrophysical Journal Supplement Series than to the Astrophysical Journal, due to its length and the color figures.

1 Major points

1. The importance of having an exactly solveable problem against which to test simulation results cannot be overemphasized. Had one considered only single-run simulation results when evaluating the very first generation of gasdynamics codes, one would have concluded that all shocks were accompanied by oscillations. One doesn't realize that such a result is completely unphysical until one compares the results to the analytically-solvable Riemann (or Sod) shock tube problem, or at the very least performs a convergence test. The authors' main conclusion that all the codes are consistent (except for a few outliers) could represent the same error. The codes could *all* be wrong, especially in the absence of convergence tests. Please comment about this.

2. The authors stated incorrectly in the introduction that an *astrophysically interesting* exactly known case does not exist. In fact, the Bertschinger (1985, ApJS 58, 39) spherical collapse solution is quite relevant, and there exist separate solutions for dark matter and for adiabatic gas. Please remove or modify this comment.

3. The formation of an X-ray cluster is, by itself, too complex a problem to be used for judging the accuracy of numerical codes. It would be interesting to see results for a shock tube problem, given that shock annuli are reported in some codes in this paper, but not in others. Surely all of the codes used in this study performed either this test or the Bertschinger (1985) test, as part of their development. I would very much like to see some discussion of the reliability of each code on these test cases.

4. An excuse offered many times in this paper for disagreement is that "internal times" were mismatched (e.g. pp. 43–44). This is a symptom of the failure to agree on a universal timestep, and it invalidates most of the comparisons because it can't be separated from real differences in the modeled dynamics. Though it is never mentioned as such, the various different spatial resolutions are almost certainly another source of non-algorithmic differences between codes. I find it particularly damning that two different runs *of the same code* (Hydra) yield results as different as an SPH compared to an Eulerian code. No reasonably designed test *of codes* could possibly admit that result, and for this reason, the authors should not represent their results as a test of codes.

5. I object to the characterization of 5–20% discrepancies as reproducing properties "quite well" (p. 42), because the comparison is meaningless without an estimate of the intrinsic discreteness errors within the individual codes. Given that such numerical errors can be held to within much less than a percent for aerodynamics codes, I find it hard to believe that cosmologists can't do as well. As the authors point out, "the scatter [in global properties] seems to be due to a slight asynchrony," or in other words, the results are dominated by uncontrolled variables. The real conclusion one is drawn to by the results given in this paper is that the codes *do not* agree with each other, to the precision they are capable of. Please modify the subjective comments about agreement, or delete them entirely.

6. What the authors *have* proven is that no code tested is violently inconsistent with the other codes. This is the strongest justifiable conclusion about reliability that can be made given the structure of the test. It *cannot* be concluded that the simulations are correct, even approximately. This caveat should appear in the conclusions about "typical" massive clusters in a CDM universe (pp. 44–46).

7. I also would like the authors to deemphasize the results with unequal smoothing between the various runs, in favor of the results with uniform 250 kpc smoothing. It is not a fair comparison *between codes* otherwise. This smoothing really should have been applied to the initial conditions for a more fair comparison, but I recognize that it is unfeasible to rerun the entire project.

2 More detailed comments

p. 7 – The formation of an X-ray cluster is not the simplest astrophysically relevant problem. That would be spherical collapse (Bertschinger 1985, ApJS 58, 39).

p. 8 – Please specify the tilt used. I presume it was $n = 1$.

p. 11 – The inertia tensor is a poor statistic for the cluster axis ratios because it is dominated by the outermost noisy parts of the cluster. Better is to project all the particles onto the unit sphere, and form the inertia tensor from that.

p. 14 – Please don't make excuses for poor parallelization. It is certainly true that keeping the load balanced in a complex code is a daunting task, but the task at hand is to evaluate the reliability — not the efficiency — of the various hydro codes.

p. 14 – The use of "glass" initial conditions is inappropriate because it makes parallel Hydra's initial conditions inequivalent to the other codes (i.e., yet another uncontrolled variable).

p. 16 – "State-of-the-art" does not always imply an expensive code. It is, however, a very good point that the test at hand could be performed on a workstation.

p. 19 – The external shear applied to the box is not used in other codes, which makes it an unfair modification, though I'd suspect it makes the physics better modeled.

p. 20 – "Hours" should not be capitalized.

p. 21 – What is "apparent" resolution? Do you mean *effective* resolution?

p. 23 – Why was a sharp k-space fiter used to smooth the Bryan & Norman initial conditions? Wouldn't that create significant oscillations in configuration space, since that filter maps points to products of sinc ($\sin(x)/x$) functions? These could be unphysical seeds for substructure formation.

p. 24 – The Bode code is *not* well parallelized if it has only 50% efficiency on only 64 nodes. That's barely adequate, and it will not scale further.

p. 24 – For the Pen MMH code, is the function $\phi(\xi, t)$ there to keep the grid from twisting, as other multidimensional Lagrangian codes typically do? Why is the grid constrained to compress only by a factor of 10?

p. 25 – The Pen MMH code mentions two different initial conditions approaches – the traditional Zel'dovich approximation, and a uniform grid with variable particle masses. Which was used?

p. 26 – The mesh in the Gneden code becomes stationary if the smallest eigenvalue of the deformation tensor becomes too small. But, the top paragraph says this persists until the eigenvalue increases again. How can the eigenvalue increase if the mesh is stationary?

p. 28 – Why was a logarithmic timestep used? That seems odd. The change in timestep at $z = 9$ can spoil the order of the time integration – it could be first order, rather than second, if leapfrog is used.

pp. 29–30 – The pleasing similarity referred to is only at the coarsest level. Highly significant differences are visible at 250 kpc scales (e.g., is there one cluster center or two at $z = 0.5$?).

p. 31, fig. 8 – With different smoothings, the various codes are not comparable. The smoothings used seem also not to have much relation to the numerical errors. The Navarro thumbnail is *very* noisy, which suggests the smoothing is inappropriately small. This figure should be cut.

p. 32 – Please discuss (briefly) the apparent spherical areas of "missing" temperature in the SPH codes, as can be seen in figs. 5 and 6.

p. 32 – Which code gets the *best* estimate of the X-ray surface brightness, and why?

p. 33 – The cluster center should have been found by the same algorithm for all the codes. This could be a major source of scatter in the various profiles. At least, please list what each code did to find the center.

p. 33 – What are the "variety of resolution and timing effects," specifically? These are your sources of error, and it is very important to be explicit.

p. 34 – Lower resolution simulations find progressively smaller gas fractions. This is very interesting. Why is it so?

p. 36 – One can't tell if a cluster is triaxial from a 2-D thumbnail. Do you mean to say that the cluster is not spherical?

p. 38 – Assigning the scatter in gas densities to differences in the timing of the simulations rather than differences in the hydrodynamics is not justified by the information given. As the test is structured, it is not possible to separate these effects. The noted differences between the Couchman and Jenkins runs isn't really enough, since there are other uncontrolled variables besides the timestep (e.g., the particle masses and the change to initial conditions performed in the Jenkins version of Hydra).

p. 40 – Do any of the codes actually enforce hydrostatic equilibrium? I wouldn't expect this, since the formation of shocks is a distinctly nonequilibrium process, and the cluster is not static. Do you mean that hydrostatic equilibrium results, as stated in conclusion 3, p. 45? Please discuss this more clearly.

p. 44 – **You cannot make general conclusions about *typical* clusters from a simulation of only one.** The paper said earlier that the cluster was specifically chosen to be near virial equilibrium. That may not be typical. Among other things, that choice might affect the value of β.

p. 45 – The cluster was chosen to be near virial equilibrium, so is it a surprise that it is near hydrostatic equilibrium over most of the cluster?

p. 55 – How did you calculate the softening in the Yepes et al. PM-FCT code? PM doesn't (usually) use an explicit softening.