Manuscript 38282 by Frenk et al.

Dear Ethan,

I am enclosing a revised version of the paper originally entitled: "The Santa Barbara cluster comparison project: a test of cosmological hydrodynamics codes." My colleagues and I have had some difficulty with the referee's report because of what seems to be a fundamental misunderstanding of what the paper is about. Many of the referee's comments follow from an impression that the purpose of this paper is to present a test of *numerical hydrodynamics codes per se*. In fact, this was never our intention. The purpose of the project was to compare the *output* of simulations of the formation of a galaxy cluster performed with different techniques and with the kind of assumptions usually made in standard simulations found in the literature. For the most part, the codes used have already been individually tested (much along the lines proposed by the referee). Thus, we do not, in this paper, ask how well a particular code models cosmological hydrodynamics in general. Instead, we ask the question: how reliable and reproducible are the results of simulations of galaxy clusters in a hierarchical clustering model?

We cannot agree more with the referee that testing a code against analytical solutions (shock tubes, similarity solutions, etc) is a very important thing to do. We take this for granted in our paper and go one step further, comparing the performance of previously validated codes in modelling the hierarchical formation of objects by collapse and mergers in a cosmological setting, an astrophysically interesting problem for which no analytical solution exists. Although we feel that this point was clearly made in the paper, the title could have been misleading. We have therefore changed it to "The Santa Barbara cluster comparison project: a comparison of cosmological hydrodynamics solutions." We have also made substantial revisions to the text in order to make the aim of the paper completely clear from the outset. This has been done primarily in Sections 1 and 5. We are pleased that the referee regards our paper as being "of substantial importance" and that "an improved version of this paper should definitely be published." We detail below the improvements that we have made in response to the referee's comments.

If the paper is accepted, we would like to insist that it be published in the main journal rather than in the Supplement. This is because we believe the paper is of general interest and contains significant scientific results (eg the convergence of the cluster dark matter density profile, the trustworthiness of calculations of X-ray properties, etc.) Furthermore, a previous paper of this type (Kang etal 1996) was published in the main journal. (We do not understand the referee's remark about appropriateness for the Supplement on account of length and colour figures.)

We now describe in detail the way in which we have dealt with the referee's comments. The latter are reproduced below in italics.

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.*

   As mentioned earlier, we agree that testing codes against analytical solutions is essential. This has been done, and reported elswehere, for all the codes used in this paper. We agree also that all the simulations presented here could be wrong and we have added a comment to this effect in Section 5 (p. 42)

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.*

   We have made the appropriate modifications and added a reference to Bertschinger's solution in Section 1.

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.*

   Please see earlier comments, including reply to point 1 above.

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.*

   We have made clear that our results do not represent a test of codes. We emphasize that the purpose of this paper is to carry out a comparison that includes not only the hydrodynamic simulation techniques themselves, but also individual choices of boundary conditions, resolution, internal variables (such as the integration timestep), and even the definition of cluster center. We do this because we wish to assess, for example, how well is the X-ray luminosity of clusters determined from real life simulations in which these choices need to be made. We emphasize that it is not our intention to test the accuracy of any individual code since this has already been done elsewhere.

   Please refer to the revised Sections 1 and 5.

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.*

   We have modified the offending remark and, as already mentioned, we have made clear what the aim of the paper is.

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).*

   This caveat has now been included.

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.*

   Results with unequal smoothing are only used in connection with the qualitative discussion of images in Section 4.1. Only 2 out of 8 images discussed there have unequal smoothing. All the others have a uniform 250 kpc smoothing.

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).

We have made the appropriate revisions to the Introduction.

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

It was $n = 1$ and this has now been specified in Section 2.

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.

We are only interested in the shapes inside the virial radius where the statistics are good.

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.

We have removed the relevant passage.

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).

Please see earlier comments about the nature of the intended comparison, particularly the response to point 4 above.

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.

Agreed.

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.

Please see earlier comments about the nature of the intended comparison.

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

This has now been corrected.

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

Yes, the correction has been made.

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.

It was chosen simply for expediency. Bearing in mind that this filter is not optimal, as the referee points out, this choice was mentioned explicitly in the relevant section. Subsequent tests using a smaller version of the cluster showed that only small differences result when a tapered filter is used instead.

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.

OK

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?

The potential grid displacement function, $\phi(\xi,t)$, ensures the regularity of the grid and its specific form guarantees that the grid will remain irrotational. The compression factor was limited to a factor of 10 because the computational cost increases linearly with compression factor. For the purposes of this comparison, larger compressions did not seem warranted.

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?

The Zel'dovich approximation – this has now been clarified.

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?

In practice the mesh never becomes stationary: as it decelerates, it approaches (but never fully reaches) a locally stationary state. This has been clarified in the text.

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.

A logarithmic timestep is appropriate for a simulation done in physical coordinates while the dynamics are dominated by the expansion of the Universe. The change in timestep at z=9 does result in one timestep which is only first order accurate, but the resulting error is negligible.

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 $z = 0.5$?).

Although there is obviously a subjective element in the meaning of the word "pleasing," we have modified the offending sentence. The next sentence in the text lists the cluster properties that we found pleasingly similar (size, shape and orientation of the main mass concentration.) The beginning of this section would have made it clear that the discussion of the images is qualitative; the quantitive results come later.

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.

We agree with the first remark and this is why six out of eight images have exactly the same smoothing. However, as explained in Section 2.2, we felt it was important to let each simulator choose a smoothing that best suited his data. As we state in that section, the images with the same smoothing were used for "a uniform comparison of all the models," while the smoothing of choice was introduced "in order to display the results of each techique in the best possible light." In view of this we do not feel that Navarro's image with his smoothing of choice 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.

An explanation has been provided in the caption to Figure 5.

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

We do not know the precise answer to this question because, as the referee points out, it is possible that all the simulations are wrong. We suspect, however, that the simulations that resolve the region where most of the X-rays are emitted (200-500 kpc) give the best estimate. This is discussed on p. 41.

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.

We have given above (and clarified in the text) the reasons why we deliberately did not specify an algorithm for defining the cluster center. This is in keeping with the intention of the comparison carried out in this paper. We have now described the algorithm used in each simulation when describing the various codes in Section 3.

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.

We have clarified this in p. 34.

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

We believe that this simply reflects the fact that, in some of the codes, the resolution is somewhat poorer for the hydrodynamics than for the N-body dynamics. We have added a comment to this effect.

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?

Yes, we have corrected this.

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).

The text says that "it seems likely that at least some of these differences result from the differences in the timing of cluster collapse..." Surely, this proposition is correct, but we have now stated explicitely that differences due to timing and due to the treatment of hydrodynamics cannot be disentangled.

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.

The codes do not literally enforce hydrostatic equilibrium. We have rephrased the sentence to make our intended meaning clear.

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.*

The statement in the paper regarding our choice of cluster as being close to virial equilibrium at the present day was, unfortunately, incorrect. In fact, we simulated the first cluster that was generated by our initial conditions algorithm because it happened to look relaxed. Had the final cluster turned out to be double, for example, we would have rejected it.

We have made this clear in the text and we have also revised the statement about how typical we expect our results to be.

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?

See answer to preceeding point.

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

The value quoted was obtained by comparing the force returned by the code for a pair of particles with the exact force for a pair of softened Plummer spheres. These two agree at a pair distance of approximately 2.4 times the cell size. We have clarified in the caption to Table 1 that ϵ denotes the effective gravitational force resolution which is only equal to the softening length in some cases.