We would like to thank the reviewer for their extremely useful comments which have improved the paper.

Almost all the points raised have been accepted and adjustments made. A point by point response is included below after each point by the referee denoted by '*Response'.

---

1) General Comment: Admittedly, the Authors spend considerable efforts to present a comparison between the different methods which is as homogenous as possible. However, an intrinsic difficulty of this comparison, that involves as many as 25 methods, is that it is difficult to keep track of the different assumptions and level of information required by the different methods.

For instance, one of the main results claimed in this paper is that methods based on richness and abundance matching are superior for mass estimates to methods based on velocity dispersion or phase-space reconstruction.

I'm not sure I agree with this statement for at least two reasons.

First, the meaning of richness and the way of measuring it are quite different in different methods. For instance, in the NUM method measuring richness requires assigning membership from phase-space information. As such, the input information required is spectroscopy with well defined completeness criteria, i.e. much more than simple multi-band photometry as required by "standard" definitions of richness.

Second, it is not completely clear to me how the reliability of richness as a mass proxy depends on the assumptions made to populate the N-body simulation with galaxies and, therefore, on the assumed underlying relation between halo mass and intrinsic richness.

For these reasons, I invite the Authors to be more cautious about stating "tout cour" that richness and abundance matching criteria are closer mass proxies than, e.g., velocity dispersion.

**** Response:*** *methods PCN, PFN and NUM use velocities to refine membership, however, the main characteristic of that membership that they use as a mass proxy is richness. We ensured that this is noted in Table A2 and we have changed the header line in the table to 'Galaxy properties used to obtain group/cluster membership and estimate mass', rather than 'Properties used to estimate halo mass', for clarity. It is our understanding that common methods classed as 'richness' in the literature also use velocities to refine galaxy membership if spectra is available.*

We had considered the dependence of our results on the implementation of the models during Phase I, however, the general success of richness-based methods in both the new HOD2 model and the SAM2 model (which, unlike a simple HOD, does not rely on a strong richness–mass relation) indicate that the good behaviour of the richness-based methods is robust.

2) Sect. 2.3. One of the improvements of HOD2 wrt HOD (used in Paper I) lies in the more realistic assignment of velocities to galaxies. I would suggest to show the scaling relation between sigma v computed within some aperture (e.g. R200c) and mass within the same aperture. In fact, it would be interesting to have this plot (in Appendix B) for both HOD2 and SAM2. It should be analogous to Fig. B1, but for sigma v, instead of Ngal. Comparing the two plots would give an idea of the intrinsic (i.e. 3D) performances of richness and velocity dispersion as mass proxies.

**** Response:*** *we have added this to the HOD2 and SAM2 richness–mass relation in Figure B1. Here the velocity dispersion is calculated via the standard deviation of the line-of-sight velocities of all galaxies. Note that this is not within an aperture, as the parameter R_{200c} is not available in the ROCKSTAR catalogues, only $R_{360\rho}$ is (unlike M_{200c}). However, as the richness–mass relation presented is also not within an aperture, it may be more consistent to also show the velocity dispersion–mass relation without imposing an aperture.*

3) Sect. 3. I find the description of the different new methods difficult to follow in some places.
-Sect. 3.1. The description of NUM should provide few more details, although the complete description is provided in paper I: meaning of M CLE and N CLE?
- Sect. 3.2, last sentence: not clear whether the lambda-M200 relation is obtained by applying abundance matching to the mocks, or is assumed to be that calibrated by Rykoff et al. (2012) for SDSS.
- Sect. 3.3. I wonder whether RM2 needs to be included. After all, the aim of this paper is not to compare performances of cluster finders.
- Sect. 3.4. In a similar way, I wonder whether it's necessary to include SG1 and SG2. All the results based on these two methods are similar to, if not worse than, SG3. Removing un-necessary methods would simplify the quite complex presentation of the results.

**** Response:*** *we have added details to the description of NUM for clarity. For RM1, the lambda–M200 relation is calibrated using Rykoff et al. 2012, we have reworded this sentence, also for clarity. It is common to see methods in the literature under the same name (e.g., shifting gapper) but implemented by different people. For this reason, we wish to keep the methods in the paper, as it is useful to see if they are, indeed, analogous. In addition, the differences between SG1, SG2, and SG3 are greater than simply using the same set up parameters.*

4) Table 2. Why not merging with Table 1, by simply adding one more column to the latter?

**** Response:*** *agreed and implemented.*

5) Caption of Fig. 2. I suggest to include the description of the colour coding for the different methods also in the caption.

**** Response:*** *agreed and implemented.*

6) Sect. 5.2, 1st sentence: I find it quite optimistic. The fact that "the methods do not collectively under- or over- -estimate the mean true mass" is not necessarily an encouraging result. Here the point is not about making an "ensemble average" among the results obtained from different methods. First of all, even if a method recovers well the mean mass, it does not imply that masses are recovered with small bias or scatter. Furthermore, there are several methods that fail at recovering even the mean mass by more than a factor 2. I would suggest to rephrase this statement, also when referring to Fig. 6.

**** Response:*** *we agree that the wording here is too optimistic. We have taken out 'encouragingly' so that this sentence is more of a statement pointing out the fact that the systematic under-estimation of cluster mass that we saw in Phase 1 does not occur in Phase 2 with the new HOD2 (with improved velocities). We have also taken this point out of the main conclusions in the Conclusions section.*

7) p. 11, 3rd par. Quoting from the paper: "In particular, for the HOD2 catalogue, velocity dispersion based methods produce slopes close to unity. All richness based methods have slopes slightly lower than unity, whereas all but one radial-based methods produce slopes greater than unity." I suggest to provide an explanation for this different behaviour of methods based on velocity dispersion and on richness.

**** Response:*** *on further examination, this statement was not correct as PCN has a slope of > 1 (1.32) and other richness based methods are very close to unity. This paragraph has been omitted.*

8) Figure B.1, right panel. Is there a simple reason why clusters with log M < 14 and log Ngal < 1.3 are not included?

**** Response:*** *this is a feature of selecting the 800 most massive, than the 200 richest clusters from the SAM2 mock*

galaxy catalogue (this is also the reason why this feature is not prominent in the HOD2 plot).