The authors present a study of the globular cluster (GC) candidate populations of the Auriga simulations and compare some of their properties with the GC systems observed in the Milky Way and M31. They find that the GC populations in the simulations, defined as stars older than 10 Gyr, do not reproduce the metallicity and radial distributions, and overproduce the total mass of the observed GC systems, which is argued to reflect trends on the environmental dependence of the formation and disruption of bound clusters.

The results of the paper clearly demonstrate that the technique of simulating stellar clusters in cosmological hydrodynamical simulations by tagging stellar particles solely based on their ages does not reproduce key properties of the observed GC systems in the Milky Way and M31. This result is extremely interesting and should be published. However, the authors do not state this conclusion anywhere. In addition, the authors investigate the similarity between the simulated GC systems and the observed ones by comparing several properties, without ever stating why are these properties relevant or what implications arise from finding similarities or differences in them.

Overall, the paper presents an important result for the field, but it feels unstructured and would benefit from a more coherent flow and discussion. The authors state different results without a proper discussion of their implications for the wider context. For that reason, a moderate revision of the paper is needed. I have some general comments regarding style, several major comments and suggestions and some minor comments, that I describe below and hope that are helpful to the authors.

General comments about style:

- Numerical quantities should be described in arabic numbers, not words. So it would be “an additional 30 GCs” and not “an additional thirty GCs” (Sect. 3.1).
- In order to facilitate reading, when describing physical quantities it is advisable to use first their name and then their symbolic representation if needed be. E.g. it would be “… in the bin with metallicities in the range $ \rm [Fe/H] \in [-2.5, -1.5]$ and …“ and not “ … in the bin where -2.5 < [Fe/H] < -1.5 and …” (Sect 4.3).
- Also to facilitate reading, the use of parentheses in sentences should be kept to a minimum. Sentences like “The purple (magenta) cross again shows the mean of the MW (M31) GCs, …, but solid (open) dot now shows our calculation of the mean values of the ‘red’ metal-rich (‘blue’ metal-poor) population.” in Sect. 4.2. would greatly benefit from this.
- Avoid colloquial expressions such as ‘Having said that, we do find …’ or ‘The main take-away from this plot is …’ and instead consider more formal constructions such as ‘Despite the lack of a metallicity cut for … it is interesting that …’ or ‘The main result of this figure is …’.
- It is MNRAS style to refer to the figures, tables and sections of their present work with a capitalised word, i.e. Figure 1, Table 1 or Section 1, and without it for references in other papers.
- Acronyms such as MW or GC need to be introduced the first time they appear in the text.
- Regarding the style of the figures, they should be as colourblind-friendly as possible, i.e. avoid using red and green or blue and yellow symbols with the same shape or style.
- Also, purple and magenta are quite non-standard colours that might confuse the readers for being too similar.
- In addition, it is custom in the stellar cluster field to leave the colours blue and red to represent the metal-poor and metal-rich cluster populations, specially in metallicity distributions, so it is a bit confusing to see them inverted in Fig. 3.

Major comments and concerns:

The goal of this work is to investigate whether the star formation model implemented in the Auriga simulations produces realistic GC populations compared to those observed in the MW and M31. Here is where two of my main concerns are.

1. The introduction lacks of a general description of what are the relevant properties when discussing GC systems, and why are they interesting or what implications to models of cluster formation and disruption there could be learned from them. So it is never clear why the authors discuss these particular properties.
2. In their work, GC candidates are defined as all stellar particles within the virial radius of the most massive sub halo that are older than 10 Gyr (z~2). This is an ad-hoc limit that is not well motivated, as this definition neglects 1 GC from the Milky Way system and 24 GCs from M31,

and places constraints on some of their results by neglecting the most recent star formation (z<2). For that reason, this definition needs further justification.

In addition to these points, I have several other major comments:

- The authors do not justify their modelling philosophy. There is a large variety of literature on the environmental dependence of cluster formation and evolution (see e.g. Brodie & Strader 2006 and Forbes et al. 2018) that is not discussed. On face of these studies, the authors need to justify why tagging stellar particles is the best approach to study stellar cluster formation and evolution in a cosmological simulation. In addition, the present work lacks a discussion on how would the results change if the authors considered a more elaborate shape for their selection function (either based on ages, metallicities, positions or the mass of their natal galaxy). As discussed in sect. 7-c in Forbes et al. (2018), the results of particle tagging will be sensitive to the exact subpopulation of particles that is tagged, so this point needs to be explored and discussed. I would suggest that the authors could do that in a new section 5.3., and then state clearly whether their particle tagging approach reproduces the observed GC populations or not.
- In Sect. 3 the authors describe with great detail the observational data used and the two samples considered: the general catalogues and the one with age-measurements for both the Milky Way and M31. However, it is unclear to me why the numbers of observed objects for each sample differ in each of the figures where these are shown.
- In order to compare the different GC systems, the authors chose to re-scale the spatial distributions by the virial radius of the Milky Way and discuss in terms of absolute distances. Nonetheless, the virial radius describes the size of the dark matter halo, but it is not a good description of the gas disk from which stars form. As an example, the median radius of the GC population in the Milky Way is 5.9 kpc (Harris 1996), which is clearly not related to the virial radius of the halo, but rather to the stellar length scale of the Galaxy. For that reason, it would be more informative to show all the radial distributions in terms of radius over a stellar length-scale (r/r_star), such as the half-mass stellar radius, and compare the different GC systems in terms of that quantity instead of using absolute distances.
- The discussion on the implications for the formation and evolution of clusters often feels a bit out of place and forced. Because of the existing degeneracy between both mechanisms, the authors should try to be clear on their assumptions before suggesting trends.

And here I list the rest of my comments for each section:

Introduction:

The current introduction of this work contains a long discussion on many different physical formation models for GC formation, but there is no further discussion on how the results obtained help to discriminate between these models. In addition to such a descriptive review of formation models, the introduction of this paper should try to consider the following questions: What are GCs? What properties do GC systems have and what are they characteristics? Why is discussing them relevant, i.e. do they tell us anything that might help disentangle cluster formation and evolution models? Has there been previous work on the topic? What did they learned? What were the benefits and limitations of their approaches? What is it going to be studied in this work? Why is the authors approach relevant or better than others? And in the conclusions the authors should go back and discuss what implications their results have on cluster formation and evolution models.

What does it mean to see “whether the model produces enough stellar mass with the right properties to allow for formation efficiencies lower than unity and the expected (dynamical) mass loss over the cluster lifetime”? None of these concepts have been described in the introduction.

Sect. 2 -

It would be interesting to also quote the halo mass range of the Auriga simulations to indicate that they can be a good description for Milky Way-mass galaxies.

The last paragraph does not make sense without having described first which are the properties to describe GC systems and why are they relevant.

Sect. 3 -

There should be also a brief description of Sects 3.3, 3.4 and 3.5, and the authors could also state here which are the properties that will be discussed.

Sect. 3.2 - This section would benefit from a more coherent structure, as it currently feels just like a listing of surveys. The authors should try to answer: which properties are relevant to have for each GC? From which catalogue are they obtained and why? How are they measured? And for quantities that need to be derived like the mass, how did they calculate it?

Sect. 3.3 - The statistical analysis is not clear; the uncertainty on the mean should include the uncertainties of the individual measurements. Also, the authors should avoid using the term 'dispersion' and instead use 'standard deviation', as the dispersion of a distribution can be described by several statistical descriptors.

Regarding the sentence: "However, the two radial distributions are not statistically consistent with being drawn from the same underlying distribution due to substantial differences at intermediate radii" - Why would these two quite different GC systems arise from the same distribution?

The bottom line of this section is that the Milky Way and M31 host GC systems that differ in their number, ages, metallicities, radii and masses, and I am missing a brief discussion as to why that could be.

Sect. 4 -

The authors should specify the number of stars that are chosen as GC candidates in the Auriga galaxies, and compare those numbers with the overall stellar population.

In general when discussing quantities, the authors should indicate whether it is the mean or median among the simulations, or if they refer to the absolute value of one of the galaxies.

Sect 4.1. - 3rd paragraph - The authors could briefly discuss the offset towards lower metallicities in the age-selected stellar sample as a reflection of the enrichment history of the galaxy. By cutting those stars younger than 10 Gyr, they are neglecting the more metal-rich stellar population, which shifts the mean metallicity of the sample towards lower values.

Sect 4.1. - End of the 3rd paragraph - This is the first time that the bimodality in metallicities of the GC system in the Milky Way is mentioned, but it should be discussed in the introduction when describing the properties of GC populations. In addition to that, Usher et al. (2012) shows the great variety in GC metallicity distributions present in the sample of early-type galaxies of the SLUGGS survey, which contradicts the classical idea that all the metallicity distributions of GCs should be bimodal.

Figure 4 - The authors should indicate what is the metallicity cut between the metal-poor and metal-rich GC subpopulations in the Milky Way.

Sect. 4.1. - Last paragraph - The trend of decreasing formation efficiency with increasing cluster metallicity might be explained by the cluster formation efficiency model by Kruijssen (2012), in which the pressure dependence leads to lower metallicity clusters forming from higher efficiencies because they form earlier from higher pressure environments. Have the authors looked into it?

Sect. 4.2. - The mean might not be the best statistical estimator to describe the spatial distributions, as it will be biased by the most distant objects. How different are the mean and the median in the Auriga galaxies? A more observer-friendly quantity would be the median, which represents the half-number radius, and if there is no trend of mass with radius, it can be converted into a half-mass radius, which in observations can be measured from the half-light radius.

Figure 6 - The authors should indicate the units of the galactocentric radius, or indicate it in terms of the length scale, e.g. log10(r/r_star).

Figs. 5 and 7 - The standard deviation is generally not recommended as a statistical dispersion estimator when representing medians. Instead, some distance between the percentiles is used (e.g. 10-90th or 25-75th).

Sect. 4.3. - The results of Figure 8 for the Auriga simulations are not discussed at all. Please remove the figure if it is not necessary for the discussion or add a brief discussion of it.

Sect. 4.3. - Third paragraph - I do not see in Figure 9 the claims about the in situ or accreted GC candidate populations. Are the authors referring to previous figures?

Figure 9 - If the upper right corner is masked because there is no observational data, then those boxes should be empty for clarity.

Figure 2, 8, 9 - The upper and right axes also need axis labels and units.

Sect. 4.3. - End of the fourth paragraph - I am missing references that support the discussion regarding the disruption of stellar clusters.

Sect. 5

Overall comment: The discussion in this section feels unconstrained as it misses some final conclusions.

Sect. 5.1. - 1st paragraph - Regarding the discussion on whether GCs contribute to the stellar halo, the authors could discuss some recent observational studies like Deason et al. (2015) or Conroy et al. (2019).

Sect. 5.1. - 2nd paragraph - As mentioned already, Usher et al. (2012) very nicely shows the great variety in metallicity distributions of the GC populations in the sample of galaxies in the SLUGGS survey. They also show that the metallicity distributions obtained from the (generally bimodal) photometric colour distributions do not always correspond to the real metallicity distributions obtained from the spectra. So obtaining a perfectly bimodal metallicity distribution might not help understand how GC populations in general form. The authors should consider this point in their discussion.

Sect. 5.1. - 3rd paragraph - Regarding the results on accreted GC candidates being more metal-poor: this result comes from the age cut of 10 Gyr, as the accreted metal-rich stars are too young for being selected as GC candidates. The reason is that massive galaxies (like the central in which the in situ population is forming) form stars earlier and enrich faster than the satellites that get accreted before z=2, that will be more metal-poor. This needs to be stated more clearly in the discussion.

Sect. 5.1. - 3rd paragraph - The sentence “The Auriga simulations do not support …” require references to studies suggesting that hypothesis.

Sect. 5.1. - End of 3rd paragraph - The classification suggested by Kruijssen et al. (2019) would not work for any of the Auriga simulations because z=2 roughly corresponds to 10 Gyr ago, so none of their GC candidates (older than 10 Gyr) could also be considered accreted. This needs to be clarified in the final sentences of the paragraph.

Sect. 5.1. - 4th paragraph - For which galaxy do Brodie & Strader provide those numbers? The interpretation of the specific frequency is often non-trivial, as it is degenerate between the formation and disruption of stellar clusters.

Sect. 5.1. - Footnote 8: Please be specific when describing quantities. It is the specific frequency, not quantity T.
Sect. 5.2. - What are the implications of the results discussed by the authors?

Sect. 6.

Many readers read the summary and conclusions section before they look at the rest of the paper, so the answer to the following questions should be clear: What is the goal of this study? Why is it relevant? How do the authors investigate it (use of simulations, observational data, comparing data, …)? What are the main results found? What implications to the bigger picture do they have? Are there future prospects or avenues for this type of study?

1st point - The authors should provide a more quantitative description of the quantities discussed so the reader can evaluate by how much something is larger or smaller.

4th point - This discussion is confusing because of the degeneracy between both effects.

Minor comments:

Introduction:
- 2nd paragraph: It might be more descriptive to use ‘metal-poor’ instead of ‘blue’, and quote the the metallicity cut considered

Sect. 3.1.
- The mass-to-light ratio should have units of MSun/LSun.

Sect. 3.4.
- The galactocentric radius should be represented by a lower case ‘r’
- There are missing error bars and references for the virial radii of the Milky Way and M31

Figure 3:
- The reference to Harris 2001 should not have parenthesis.
- typo: "GG candicates” should be “GC candidates”

Sect. 4.1.
- “We investigate the model implemented in Auriga … ”: which model do the authors refer to?
- “The top half of the left figure … ”: it is not clear what this refers to

Sect. 4.2. - 3rd paragraph
- typo: The sentence “The lowest mass surplus … ” has an extra ‘the’

Sect. 5.1. - Last paragraph
- typo: ‘mass los rates’ should be ‘mass loss rates’