We are grateful to the reviewers for their constructive and useful comments that led us to improve the paper in several points. We respond to specific comments raised by the reviewers in following.

Reviewer #1: This is an interesting article discussing detailed characterisation of two binary asteroid systems, (66391) 1999 KW4 and (88710) 2001 SL9, based on a long-term photometric campaign. The authors detect a drift in orbital elements of mutual orbits in both systems and explore possible explanations, including BYORP as possible mechanism affecting the mutual orbits. This is a well-rounded sudy with with sound methodology, and can be recommended for publication. I will focus on the observational aspect of the work, and I only have minor comments.

p. 8: I would like to know what step sizes were selected for the mean anomaly drift search.

A: The steps in ΔM_d were 0.005 deg/yr2 and 0.01 deg/yr2 for 1999 KW4 and 2001 SL9, respectively. We added this information to the paper.

p.9: The last couple sentences of the first paragraph are confusing. I assume that the range of investigated pericenter drifts reflect the values that would be possible assuming realistic parameters of the system, but the wording might need adjusted.

A: This assumption is correct; we adjusted the wording accordingly.

In the paragraph describing the RMS for the fit of SL9 models the solutions with higher RMS are rejected as they "do not appear real". Is this only because of the higher RMS? I would suggest making this clearer or including any potential additional reasons for the claim.

A: See our response to a similar comment by the second reviewer.

Figure 3: Is the period used to plot the dashed-line model optimised assuming zero mean anomaly drift, or is it the same period as for the best-fit solution with non-zero anomaly drift (i.e. period at the model epoch)?

A: The former is the case. The dashed-line shows the model with the mean anomaly drift fixed at 0, but all other parameters fitted, including the orbital period. We adjusted the caption of the figure to emphasize this.

Figures 4 and 8: I think those figures are slightly misleading. Which orbital period was used to plot the points? The orbital period at model epoch for the best-fit solutions which also include a non-zero mean anomaly drift, or the orbital period corresponding to the solution assuming zero mean anomaly drift in in figures 1 and 2? If the former applies, I think the captions to the figures and text on page 9 need adjusting to make it clearer.

A: We adjusted the text on page 9 and the figures captions to make it clearer.

p. 11: Reference to Warner et al. 2009 is included, but the bibliography data is missing in the References section.

A: We added the reference to the section.

p. 12 The radar results were used to derive the diameters ot the components, and the size ratio obtained for KW4 is compared with the results of the radar study (Ostro et al. 2006). Could the radar shape model of the primary along with the size-corrected radar model of secondary be also used to verify the mean anomaly drift?

A: A synthetic primary lightcurve computed from the radar shape model, assuming a uniform surface light scattering, does not fit the observed primary lightcurve well. (We note that we have found similar discrepancies between synthetic lightcurves computed from primary radar shape models and observed lightcurves for other binary NEAs as well. It appears that either the radar shape models are not precise enough, or the primary's surfaces don't have uniform light scattering.) So using the primary shape model would make the fit worse instead of verifying it. And the radar shape model of the secondary, besides its underestimated size, is even affected by much larger errors due to low SNR, as is noted before the end of the Section 3 of the paper.

p. 13 It is unclear what the "actual light scattering model" is.

A: We used a combination of Lommel-Seeliger and Lambert scattering, as stated in the beginning of the Section 2.2. We adjusted the text on p. 13 (now page 15) accordingly.

p. 14 I would suggest including the paremeters derived by Pravec et al. (2006) to illustrate the agreement with presented results.

A: We include the parameters from Pravec et al. (2006) to the text.

p. 17 Please clarify if "perturbing the vertices vertically" refers to perturbations along surface normals, radially, or along another direction.

A: The perturbations were radial, we adjusted the text to clarify that.

p. 19 Are the ellipsoid models developed for the two components insufficient for BYORP calculations? The opening of section 5.2 is confusing, as section 4 discusses possible near-spherical shape for the secondary.

A: The BYORP effect requires the secondary body to be of irregular shape (see Cuk and Burns, 2005), the ellipsoidal approximation is therefore insufficient in this case. We adjusted the opening of the section 5.2 to clarify that.

Appendix A: The NIR spectroscopy of KW4 is discussed here. However I feel some additional information is needed. What was the time span of spectral observations? Also, the authors obtain 37 spectra, yet only one is analysed, is this a sum of the 37? Figure 14 could benefit from including illustration of mean spectra of Q (and maybe O) type to support the discussion.

A: We added the time span (5:30-7:55 UTC) to the text and clarify that the analysed spectrum is an average of all obtained spectra. We also added mean Q and O spectra

to Fig. 14.

Reviewer #2: I recommend that this paper be revised substantially. This submission claims to have detected the quadratic drift of the mean anomaly of a binary asteroid mutual orbit in time for two systems. This is a significant achievement, however the evidence as presented is problematic. In particular, the submission does not adequately demonstrate with a statistical test that the claimed best fit solutions are significant despite the claim of significance on page 9. For Moshup 1991 KW4, the best fit solution as shown in Figure 3 is visually convincing particularly compared to the non-drifting mean anomaly case. However, for 2001 SL9, the data as shown in Figure 7 is far from convincing and, visually, all three shown cases look like they fit the data equally well. In order to be convinced, the authors need to not just calculate the value of a goodness-of-fit measure but describe the uncertainty of such a measure. Currently, the authors use the RMS magnitude residuals as a goodness-of-fit measure, but they do not report the statistics of this measure, so it is unclear as to whether a 0.0015 mag difference is significant; just comparing that number to the uncertainties of a given measurement or average of measurements would be something. A better solution would be to conduct either a Pearson chi-squared test or a G-test between the observations and the model so that the likelihood of the difference between two models being due to chance could be directly evaluated and significance assessed. The values and uncertainties of each measurement would be used to determine the goodness-of-fit for each choice of model orbit parameters and then the chi-squared or G-test metric would follow a chi square distribution and a p-value could be assessed, which will inform if the difference is statistically likely.

A: The differences between the observed and simulated data in the binary asteroid photometry are dominated by systematic effects – mainly due to the model simplifications. It results in that the fit residuals of nearby points are correlated. The statistical tests assume that the residuals are random and normally distributed, which is not justified here. To overcome the problem, so that we can use the chi-square test, we adopted a strategy described in new section 2.3.

We also enlarged the number of presented lightcurve sessions (Figs. 9 and 10 in the revised version) in order to emphasize the differences between the synthetic curve of the model with ΔM_d fixed at 0 and the observed data. We also point out that the differences are systematic in the sense that the synthetic curve for $\Delta M_d = 0$ is shifted in time with respect to the observed data and that this shift is different for each observed apparition. Moreover, these shifts evolve in time so that they are consistent with the quadratic drift of the mean anomaly, as shown in Fig. 11.

In the abstract, it may be helpful to provide the semi-major axis in km to give context to the mean rate of change of the semi-major axis or to calculate the semi-major axis doubling/halving timescale. Instead of using "internal" consider using "mutual" to describe the dynamics of the binary components about their shared center of mass as opposed to their joint motion around the Sun.

A: We added the semi-major axes in km to the abstract.

We adjusted the wording to "mutual two-body dynamics" in the Introduction.

The values used in the calculations in Section 5.1 appear to be different than the values in Table 4. Why is this? Wouldn't it be best to introduce a single set of nominal parameters and use those parameters for all calculations except for when one is explicitly varying them?

A: We updated the semimajor axis to agree with the value in Table 4 and explained in the footnote 3, why R_{mean} doesn't agree with $D_{2,\mathrm{V}}$ from Table 4.

Given the claimed best fit semi-major axis drift and estimates of all the other binary asteroid parameters for 1991 KW4, it seems appropriate for the paper to provide an estimate of the BYORP coefficient if tides were insignificant, if Q/k is consistent with the Taylor & Margot (2010), and if Q/k is consistent with the Schierich et al. (2015) estimate.

A: We added these estimates to the end of Section 5.1.

It's important to note that the Taylor & Margot (2010) estimate of Q/k is of a lower bound because they assume a maximum tidal evolution timescale – this should be made clear in the text. It's also important to note that the Q/k estimate from Schierich et al. (2015) is good for the 1996 FG3 system but theory suggests that it should scale with size. Goldreich and Sari (2009) hypothesize that k goes as R for rubble piles, Jacobson & Scheeres (2011) fit binary systems to find that Q/k goes as 1/R and hypothesized that k went as 1/R, but, lastly, Nimmo and Matsuyama (2019) explain why Q should go as R^2 so that when k goes as R as Goldreich & Sari (2009) surmised, Q/k goes as 1/R as Jacobson & Scheeres (2011) discovered. Thus, it's important to scale the Q/k value regarding 1996 FG3 in Schierich et al. (2015) to the binaries at hand.

A: We noted that Q/k from Taylor & Margot (2010) is a lower bound in the text and replaced equalities with inequalities on corresponding derived values.

The works cited indicate that Q/k goes as R, not as 1/R. We assume that that was a typo in reviever's comments. We therefore scaled the Q/k estimate from Scheirich et al. to KW4 and SL9 accordingly and updated the text and values.