Wilma H. Trick
Max-Planck-Institut für Astronomie
Königstuhl 17
69117 Heidelberg, Germany

Prof. W. Butler Burton
Associate Editor-in-Chief, The Astrophysical Journal
Professor Emeritus, Leiden University
National Radio Astronomy Observatory

Heidelberg, ??th April 2016

Subject: Answer to the Referee Report on "Action-based Dynamical Modelling for the Milky Way Disk"

Dear Prof. Burton,
Dear anonymous referee,

thank you very much for having such a very throughout look at our paper draft and preparing this very detailed, informative and helpful report. We made extensive efforts to account for all the referee's comments and have rerun several tests. The latter is also the reason why it took us so long to re-submit the paper. We attach the original letter and report below to which we added our answers to the referee's comments in blue and italics. We sincerely think that the content of the paper improved a lot with the referee's help—and we hope you agree.

Kind regards,
Wilma Trick

Attachments:

- `Trick_MW_Modelling_revised.tex/.pdf` - revised paper file ready for publishing (without any marked text or annotations), if the referee agrees.
- `Trick_MW_Modelling_revised_all_Changes.tex/.pdf` - file in which all changes with respect to the previous version of the paper are marked (new text = green, deleted text = red, annotations = blue).
- `Trick_MW_Modelling_revised_for_referee.tex/.pdf` - file for the referee with the deleted text excluded to be more readable (new text = green, annotations = blue).

Ms. Wilma H. Trick
Max-Planck-Institut fuer Astronomie
Koenigstuhl 17
Heidelberg, Baden-Wuerttemberg 69117
Germany

December 18, 2015

Title: Action-based Dynamical Modelling for the Milky Way Disk

Dear Ms. Trick,

I have received the referee's report on your submission to The Astrophysical Journal, and append it below.

I hope that you will agree with my assessment that the report is constructive in tone, but you will note that the referee asks that a number of issues be addressed before publication in the ApJ.

Following the referee's report, I have also appended some comments regarding the statistical aspects of your manuscript. ApJ submissions with a statistical component are routinely previewed by Dr. Eric Feigelson, the member of the ApJ editorial board with a broad expertise in statistical matters in an astrophysical context. Perhaps these comments will be of use to you as you prepare your revision. If you have any questions about the comments, however, please direct them to me, not to Dr. Feigelson. Please consider the report carefully. When you resubmit, please include a detailed cover letter indicating point-by-point your responses to the referee's report, and also indicating any other changes you have made to the text. Reviewers find it helpful if the changes in the text of the manuscript are easily distinguishable from the rest of the text. Therefore we ask you to print changes in bold face; this highlighting can be removed easily after the review.

$\Longrightarrow$ All new text in the attached PDFs is printed in green. Annotations, that should be removed for the final version, are written in blue. Text that was part of the original version of the paper but was removed for the revised draft is printed in red. (See list of attachments above.)

And two minor matters:

- Please adhere to the ApJ instructions to authors (see `http://aas.org/journals/authors/common_instruct#_Toc3.2`) regarding the abstract –"The abstract should be a single paragraph of not more than 250 words ..."
  $\Longrightarrow$ We have tried to significantly reduce the length of the abstract (according to our count it is now at 254 words). We don't believe however that we can reduce it even further while keeping it also informative and readable.

- Please also use the appropriate journal abbreviations in the reference list: these abbreviations for refereed journals are given on the ADS site `http://adsabs.harvard.edu/abs_doc/refereed.html`, and on the site `http://adsabs.harvard.edu/abs_doc/non_refereed.html` for non-refereed

publications.
$\Longrightarrow$ Our reference list is now using bibtex and the bibtex style file "aasjournal.sty". We hope that makes everything automatically correct.

Click this link (it will work one time) to upload your revised manuscript: [...]

Alternatively, you can also log into your account at the EJ Press web site, http://apj.msubmit.net. Please use your user's login name: wilmatrick. You can then ask for a new password via the Unknown/Forgotten Password link if you have forgotten your password.

The policy of The Astrophysical Journal is to view manuscripts as withdrawn if no revised version is received within six months after the most recent referee's report goes to the authors.

If you have any questions, please contact me.

With best wishes,

Butler Burton

Prof. W. Butler Burton
Associate Editor-in-Chief, The Astrophysical Journal
Professor Emeritus, Leiden University
National Radio Astronomy Observatory

++
The text of the review is appended below.
==

1 Reviewer's Comments:

This paper details and tests a framework for constraining the Milky Way potential using action-based dynamical distribution functions. The work very much builds on previous work from two of the authors as well as the dynamics group in Oxford. However, for the first time the machinery and its assumptions are quite rigorously tested and the paper thoroughly discusses the limitations and downfalls of the methods.

I believe that this work should be published but that there are parts of the paper where the presentation could be clearer and where the discussion could be expanded or reduced. Additionally, I think that this paper will be a good reference paper for those who will fit dynamical models to the Gaia data, but there are a number of places where the paper could be made more 'usable' without much effort.

My main complaint with the paper, however, is that it doesn't tackle the 'real' problem for three reasons:

- 1. The selection volumes considered seem unrealistic. It is unclear how

they relate to what one might use with the Gaia data.
$\Longrightarrow$ *The treatment of selection functions that are 'realistic' in their complex angular patterns has been already demonstrated in Bovy & Rix (2013) and Bovy et al. (2015). While the pattern of the survey area on the sky may be complex, it is usually precisely known. It is the uncertainties (e.g. incompleteness) in the line-of-sight, or magnitude directions that is most prone to systematic mis-assessment. This is why we have focussed on this here (with a simplified angular pattern). We now state this reasoning in Section [ADD A SENTENCE] We understand the referee's concern though, and we added a small paragraph in Appendix B that tries to motivate the spherical selection function used in almost all tests in this work. For contiguous surveys like Gaia or Hipparcos we would of course not have a survey volume with a sharp edge and the survey selection function is more likely to be a magnitude cut. But for intrinsic standard candles like red clump stars a magnitude cut translates to a distance cut. (The very idealized wedge-like selection functions are only used in one test, Test 4, and that one was not supposed to show realistic scenarios, but rather investigate an intrinsic aspect of action-based modelling.)*

- 2. It is unclear how the error distributions relate to the anticipated errors from the Gaia data.
  $\Longrightarrow$ *This is a good point and we included two paragraphs and a footnote in Section 3.4, that try to address this. The proper motion errors we considered were inspired by the range of proper motion errors achieved by ground-based surveys. We estimated also the Gaia errors for suitable giant tracer stars and Rene Andrae used the GUMS catalogue to estimate for us the magnitude limit required to ensure that the distance errors for all Gaia stars in a sample would be small enough for our method to work.*

- 3. The majority of the tests seem to be done with the isochrone potential. This is obviously done for speed reasons. A few cases are tackled with the KKS-Pot and one case with the MW13-Pot.
  $\Longrightarrow$ *This was indeed a significant limitation of our work (though clearly stated as a caveat). Four out of six aspects investigated in this paper were done with the* `Iso-Pot`*. To alleviate this, we reran two of the more important aspects (on DF misjudgments and on SF misjudgements) with a more realistic Galaxy potential with disk, halo and bulge, the* `DHB-Pot`*. (All tests with the* `MW13-Pot` *were also replaced with the* `DHB-Pot` *to have no more than four different potentials in the paper.) Because of the immense computational effort we did however not rerun the tests on the measurement uncertainties with the* `DHB-Pot`*. There we still use the* `Iso-Pot`*. We also think that for testing the effect of measurement uncertainties the form of the potential should not be that important. Results of the new tests using the* `DHB-Pot` *are presented in Figures 1, 2, 6, 7, 11, 12, 13, 14 and Tests 1, 4, 5, 7.*
  It is then not clear how the precision of the potential parameter recovery for the isochrone relates to more realistic potentials (e.g. will the circular velocity always be so well constrained irrespective of the potential form?).
  $\Longrightarrow$ *Now, that we reran some of the tests with the more realistic* `DHB-Pot`*,*

we found that the $v_{circ}(R_\odot)$ parameter is indeed recovered to a very similar precision as for the simplified `Iso-Pot` *test case. Overall, the new results with the* `DHB-Pot` *are very similar to the* `Iso-Pot`*—for example for which differences in DF and misjudgment of the SF the recovery of the potential broke down—which is very encouraging for the transferability of the results, also considering the tests that we could not repeat with the* `DHB-Pot`*.*

The issue of computational speed is mentioned in the discussion but I think this issue is slightly glossed over.

$\Longrightarrow$ *We expanded the section on "Computational speed" in the discussion, which lays out how computation speed limits the potential models that we are currently able to use in RoadMapping. We also included a measurement how long some of the analyses in the paper need on how many cores.*

Using more realistic potentials in all cases would obviously be much better and would also test the 'Staeckel fudge' apparatus more fully. The improvements to this apparatus are highlighted in the paper but I don't think that it is shown that it has been fully tested.

$\Longrightarrow$ *The* Staeckel fudge *is used in all similar studies, e.g., Bovy & Rix (2013), Piffl et al. (2014), Sanders & Binney (2015), Das & Binney (2016). The* Staeckel fudge *was also recently reviewed and discussed in Sanders & Binney (2016). Because the* Staeckel fudge *is "state of the art" and has been tested, used and discussed by so many people before us, we did not test it in detail.*

I think that the heavy use of the isochrone potential should be flagged in the abstract.

$\Longrightarrow$ *As it is now only two of the tests that still use the* `Iso-Pot` *we prefer not mention the* `Iso-Pot` *specifically in the abstract.*

It would be good if some of these concerns were brought to the fore and highlighted in the paper. I have detailed places in the paper where things could be changed. There is also a short list of typos at the end.

2 "Abstract"

- 1. 4th sentence – explain what 'slightly wrong' means more quantitatively.
  $\Longrightarrow$ *We had to delete this sentence because the abstract was too long. We tried however to give some qualitative statements for all the results we are summarizing in the abstract.*

- 2. 5th sentence – Are the constraints of high precision on the potential or DF? – clarify
  $\Longrightarrow$ *This sentence was also removed. In the original abstract however we meant that the constraints on both potential and DF parameters are precise.*

- 3. The tests referred to in the abstract have been performed independently but the lists suggest that you have shown that when all the listed

conditions are satisfied the constraints are of high precision. This should be clarified.
$\Longrightarrow$ *This is a fair point. We also had to remove the sentence about precision, but we mentioned that we investigated "isolated test cases" for "idealized mock data" and added an "either ... or" between the results on "unbiased estimates". We hope that makes it clearer.*

- $\Longrightarrow$ ***Additional changes:*** *The abstract was partly rewritten, to (a) meet the limit of 250 words, (b) focus more on the key results of the work, (c) make it clearer that we used idealized test cases.*

3 "1. Introduction"

- 1. Magorrian (2014) has provided a framework for constraining the potential without assuming a particular parametrized form for the DF. Whilst Magorrian's method is computationally intensive, it should be referenced in the introduction as it relates to the later discussion of choosing a particular DF parametrization.
$\Longrightarrow$ *Thanks for the suggestion. We included the reference in the paragraph on different modelling methods in the introduction. We also added a little paragraph in the discussion of different modelling methods.*

- 2. At the end of the introduction I think that you should refer people more strongly to the results section as much of section 2 is presenting a framework that appears elsewhere.
$\Longrightarrow$ *Most of the introduction on page 2 is a list of the different questions we try to investigate in the result section of this paper paper. But to make sure that the reader knows immediately what the key results are, we added a sentence pointing again to the corresponding sections at the very end of the introduction.*

4 "2. Dynamical Modelling"

4.1 Section 2.2 - Now: "2.2 Actions" & "2.3 Potential models"

- 1. A slightly fuller introduction of the actions is merited. Mention why the actions are introduced. What advantages do they present? Also state that the action-angles are canonical – this is important later for transformation of the pdf.
$\Longrightarrow$ *We completely agree. We added a corresponding section.*

- 2. The third sentence isn't quite right – the most general are the triaxial Staeckel potentials of which the axisymmetric Staeckel potentials are special cases and all spherical potentials are special cases of these. THe isochrone potential is the most general potential in which the actions are not computed as a quadrature.
$\Longrightarrow$ *Thanks for making it clearer for us. We changed the text accordingly.*

- 3. The potential discussion could be put into a separate section. Also mention that the circular speed at the Sun is the same for all three potentials.
  $\Longrightarrow$ *Done.*
  Re-iterate that the reason for using the isochrone and Staeckel potential is the ease with which the actions can be computed.
  $\Longrightarrow$ *Done. We added a paragraph in Section 2.3 summarizing the advantages of the different potentials we choose to investigate in this paper.*

- $\Longrightarrow$ ***Additional changes:*** *We removed the* `MW13-Pot` *completely from the paper and replaced it with the* `DHB-Pot` *everywhere, also in Table 1 and Figure 1. We wanted to have only four different potentials in the paper, and the* `DHB-Pot` *was much better suited to rerun some of the tests in short time than the* `MW13-Pot`*, for which the analysis is very slow due to the use of an exponential disk.*

4.2 Section 2.3 - Now: "2.4 Stellar distribution functions"

- 1. Guiding-center should not be in brackets as this is important.
  $\Longrightarrow$ *Done.*

- 2. The final sentence of the left-hand column of Fig 3 should read $L << L_0$.
  $\Longrightarrow$ *Thanks for pointing out that typo. We wrote* $L < L_0$ *instead.*

- 3. Top of page 4 – explain what X is in the text.
  $\Longrightarrow$ *Done.*

- 4. Do you interpolate in log density?
  $\Longrightarrow$ *Yes, we did. We now mention it in the paper as well.*

- 5. The footnote says 'should be chosen as' – add a forward reference to Fig. 4
  $\Longrightarrow$ *Done.*

- $\Longrightarrow$ ***Additional changes:*** *We added a paragraph at the beginning of this section which discusses why the Jacobian = 1 (which we mentioned in Section 2.2 as requested by the referee) allows us to consider DF(x,v) and DF(J) equivalently as probability of a star to be at (x,v).*

4.3 Section 2.4 - Now: "Appendix B: Selection Functions"

- 1. An entire section dedicated to this topic seems unnecessary. See below.
  $\Longrightarrow$ *We moved this section to the Appendix (because we also wanted to further motivate the "unrealistic" selection functions we use in this work).*

4.4 Section 2.5 - Now: "Appendix A: Mock data"

- 1. As this section only explains technical details rather than testing the apparatus I think that this section can be put in an appendix along with Fig. 2 and 3. Fig 2 and 3 are illustrative but I think that they are similar to BR13 Fig2 and 3 so do not need to appear in the main body. I don't

think it is a 'test' so should be removed from Table 3.
$\Longrightarrow$ *We agree. We moved the mock data section and figures to the appendix and removed the parameters belonging to Figure 2 + 3 from the 'tests' Table 3.*

- 2. The discussion of selection on very errorneous x coordinates is interesting but surely this isn't the way the data will actually be handled?
$\Longrightarrow$ *We are not 100% sure that we understand the referee's main point. If one considers several percent of distance errors at the edge of the survey volume, the volume from which stars can get scattered outwards or inwards due to random noise is larger than the survey volume. This is why we create perfect data in a large volume, before we first perturb it according to the measurement errors and then select the stars that would enter the actual survey volume and therefore data catalogue. We realise it's idealized and there won't be a sharp distance cut. But scatter in and out of survey volume will play a role and we explore an ideal case.*

4.5 Section 2.6 - Now: "2.5 Data likelihood", "2.6 Likelihood normalisation" & "2.7 Measurement uncertainties"

- 1. The selection function can be briefly mentioned at the beginning of this section and stated that you assume here for simplicity it is a function of $\vec{x}$.
$\Longrightarrow$ *Done.*

- 2. The Jacobian from J,theta to x,v should be mentioned here.
$\Longrightarrow$ *We already put a little bit of the implications of the Jacobian for the DF at the beginning of Section 2.4 "Stellar distribution functions", but we added also a footnote here that explains why we can integrate DF(J) easily over (x,v) in the likelihood normalisation.*

- 3. pdf should be defined in a separate equation.
$\Longrightarrow$ *Done.*

- 4. Figure 4. – it wasn't clear to me that the 'truth' normalization used a high enough set of parameters. $N_x = 20$, $N_v = 56$ and $n_\sigma = 7$ only seem slightly larger than the values actually compared to.
$\Longrightarrow$ *Yes. We reran this test with a higher accuracy $N_x = 32$, $N_v = 68$, $n_\sigma = 7$ for the "true" normalisation. We also used the* `DHB-Pot` *in this plot instead of the* `KKS-Pot`*, which was used in the original plot. The reason for the change is, that the* `DHB-Pot` *is used in more tests. We were also interested how this test depends on the kinematic temperature of the population. And as there seems to be a dependence, We added a new row of panels for the cool population in this plot. It turned out that for the cool population the accuracy condition was not satisfied for all volumes. So overplotted little black stars that show which volumes are actually used in tests of this paper, to assure the reader that for the actual tests the condition was satisfied.*

- 5. Make the normalization discussion a separate section.
$\Longrightarrow$ *Done.*

- 6. The discussion of the likelihood normalization should reference and compare with McMillan and Binney (2013) as the discussion is very similar.
  $\Longrightarrow$ *Thanks for the suggestion. We included a paragraph where we compare their work on the likelihood normalisation with ours.*

- 7. Is there any general advice on how to choose N_x, N_v and N_σ? The authors have shown it is OK for the mock datasets but do I have to redo the authors' exercise when I have a real dataset?
  $\Longrightarrow$ *The short answer is unfortunately "yes", you would need to go through the full exercise. We think if you use a similar DF and potential than we did in Figure 2, you could use Figure 2 to look what accuracy you need for your number of stars. Or use our values as starting values and optimize later. But otherwise you need to check on a case-by-case basis if your accuracy was high enough. We added a corresponding sentence in the paper and explain (in a footnote) why it depends on both kinematic temperature and also the particular potential model what accuracy you need.*

- 8. Put error discussion in separate section.
  $\Longrightarrow$ *Done.*

- 9. Reduce size of caption for Fig 5. More of the details could go in the text.
  $\Longrightarrow$ *Done.*

- 10. Equation (15) is a novelty. It is troubling that the tests that use this approximation all seem to use the isochrone but the approximation is still necessary. Is that because it is computationally awkward to calculate this integral or just very slow?
  $\Longrightarrow$ *In our simple test scenarios it should be indeed computationally possible to convolve the selection function with the homoscedastic uncertainties and not use the approximation in Equation (15) - independent of potential. However, in each realistic case (heteroscedastic errors, more complex selection functions, etc.) the proper treatment would be actually computationally way too expensive as the normalisation has to be calculated for each star separately. We will therefore never actually want to use the proper likelihood formula, and always use the approximation. Ignoring measurement uncertainties in the normalisation is by the way also done in similar studies, like McMillan & Binney (2013) and Das & Binney (2016). We are the first to test this approximation. That we used the* `Iso-Pot` *instead of a more realistic potential was simply for its computational speed, because we wanted to investigate a large number of mock data sets at different accuracies to be able to do some statistics. If it would have been feasible, we would have done the test with a more realistic potential. We still think that it should not make a huge difference. In the tests that we run with both* `Iso-Pot` *and a more realistic* `DHB-Pot` *we always got pretty similar results - also quantitatively. We added a sentences in the section, to stress more the computational expense when NOT using the approximation and added a caveat in the discussion that we tested the approximation with the* `Iso-Pot` *only.*

- 11. The penultimate sentence of this section contradicts the previous sentence without validation. Why is this?
  $\Longrightarrow$ *Thanks for pointing it out. That was indeed confusing. After running more tests for both the* `Iso-Pot` *and* `DHB-Pot` *we think the reason is, that the recovery of* $v_{circ}(R_\odot)$ *behaves in the* `Iso-Pot` *always as expected and similar to* `DHB-Pot`*. To not get biases in* $v_{circ}(R_\odot)$ *one needs more* $N_{samples}$ *for higher* N_**, as expected. For the isochrone scale length* b *this trend was not so clear. We think that this is a specific property of the* `Iso-Pot`*. We rewrote those two sentences in the paper and specified a bit more closely what we observed for smaller* N_* *and different parameters.*

- $\Longrightarrow$ ***Additional changes:*** *In the section "Data likelihood" we added a paragraph that argues why we chose to use a Bayesian framework, to address the comments by Dr. Feigelson (see below).*

4.6 Section 2.7 - Now: "2.8 Fitting procedure"

- 1. I liked this section – it was well thought out and informative.
  $\Longrightarrow$ *Thanks. :-)*

- 2. Here a fixed sampling is used for the error samples. I think again you should reference McMillan & Binney (2013) as they discussed the numerical stability of this method.
  $\Longrightarrow$ *We referenced this paper, but not in this section, which is more about fitting and less about measurment errors. We mentioned it in "2.7 Measurement uncertainties" as a similar approach using MC sampling for convolution with the measurement errors.*

5 "3. Results"

- 1. It is stated that the breakdown of axisymmetry and steady state assumptions is not explored. I wonder as well about the impact of resonances, particularly when the data are very high quality. This cannot be explored in the current setup as the data are generated from an action-based DF but perhaps should be mentioned as a potential limitation of the approach.
  $\Longrightarrow$ *Thanks for this suggestion. We added a corresponding comment in this section and also mentioned that we are planning to further investigate non-axisymmetries by applying RoadMapping to a galaxy simulation in a future paper.*

- $\Longrightarrow$ ***Additional changes:*** *As we moved the Mock data section to the Appendix, we stress here explicitly that we use mock data for our tests.*

5.1 Section 3.1 - "Model parameter estimates in the limit of large data sets"

- 1. This seems a good sanity check but should it be published? Fig 6. seems sufficient to me to demonstrate that your code works. I don't think the paper would miss this section.

$\Longrightarrow$ *We wanted to have one section in this paper that summarizes the (statistical) implications of using large data sets for RoadMapping separately and not just in the context of "Likelihood normalisation" as in Section 2.6. But as you are in principle right and it is mostly a sanity check, we edited the whole section to make it much shorter. We also removed the figure, that shows how the width of the pdf decreases with* $1/\sqrt{N_*}$*, completely from the paper and only mention the result in one sentence.*

5.2 Section 3.2 - "The role of the survey volume geometry"

- 1. I understand that the selections used in Fig 9 are illustrative but the pink selection just doesn't seem realistic. I think Fig 8. is a sufficient demonstration of the difference between different selections. Fig. 9 doesn't add anything and is barely discussed in the text. Also, without observational uncertainties (which will be greater for the more distant boxes) the discussion seems superficial. I would consider removing this.
  $\Longrightarrow$ *We understand the reviewer's opinion. However, we wanted to make a slightly different point by using these survey volumes. We were interested in how different regions of the Galaxy intrinsically constrain different potential parameters, and if we would loose a lot of information on the potential if we excluded certain regions from the analysis (for example because of large dust extinction in the disk or due to other reasons). Why this plot and its discussion seem uninteresting is because of the result: The role of the survey volume geometry and position is not that important! It is ONLY the other obvious factors (the number of stars, the measurement uncertainties and the size of the survey volume in general) that one should consider when choosing an observed volume—independent of the volume shape. This was actually not obvious to us before we made this test. And maybe it is also not for the reader. Therefore we would like to keep the plot. We hope you agree. We tried to stress our point a little more in the paper, and we also changed the order in which the plots are discussed in this section, to reduce confusion.*

- $\Longrightarrow$ ***Additional changes:*** *In Figure 9 (which is now Figure 6) We replaced the* `MW13-Pot` *analyses with analyses for the* `DHB-Pot`*. The original* `MW13-Pot` *analyses were totally fine, but we wanted to remove this potential from the paper to have only four different potentials. Also, the* `DHB-Pot` *allowed us to fit* $v_{\rm circ}(R_\odot)$ *too. This parameter is not shown in Figure 6, because its recovery is very similar (qualitatively and quantitatively) to the* `Iso-Pot`*'s* $v_{\rm circ}(R_\odot)$*. We mention this also in the caption of Figure 6.*

5.3 Section 3.3 - "Impact of misjudging the selection function of the data set"

- 1. Isn't the reason for the cold population being more robust that it doesn't have as many stars at large distance as the hot population so it is less affected by the cuts? I suppose this not necessarily true for lines-of-sight in the plane.

$\Longrightarrow$ *That is a really good (and surprisingly simple) explanation! Thanks! We changed the text accordingly. We believe the recovery of the radial profile is less affected by the cuts, because we removed stars symmetrically at both small and large R. Therefore the difference in tracer profiles in the vertical direction is more important.*

- $\Longrightarrow$ ***Additional changes:*** *As part of our efforts to make the paper considering more realistic cases, we replaced all analyses in this section with the new analyses using the Galaxy-like* `DHB-Pot` *instead of the spherical* `Iso-Pot`*. As the* `DHB-Pot` *is slower to use than the* `Iso-Pot` *we could not make the same number of analyses—but the observed trend is already so similar to the* `Iso-Pot`*, that we think that is fine. We added a paragraph which discusses, in addition to the new* `DHB-Pot` *figure, also the results that we got previously for the* `Iso-Pot`*. We also removed the figure displaying the analyses with marginalization over v_T, because showing this plot was not necessary for the* `DHB-Pot` *anymore. Its result is however still discussed in this section.*

5.4 Section 3.4 - "Measurement uncertainties and their effect on the parameter recovery"

- 1. It would be nice to state how the considered errors are related to the anticipated Gaia errors or other surveys.
  $\Longrightarrow$ *Good point. We should have included such a discussion already in the first version of the paper... We added a paragraph that discusses the proper motion errors of ground-based and space-based surveys (and a footnote that mentions a few more details). The proper motion range that we considered is the range that can be currently achieved with ground-based surveys. One of our results is, that for large distance errors our likelihood approximation is not unbiased anymore. We estimated what kind of distance errors we would expect for red clump stars in Gaia, and Rene Andrae helped us to work out a magnitude limit which would keep the distance errors of all Gaia stars below the required limit.*

5.5 Section 3.5

- 1. I think this and section 3.6 are the most valuable in the paper as they really explore potential systematics. In my opinion, these are the key results.
  $\Longrightarrow$ *We think so too. We stressed this in both the abstract and at the end of the introduction.*

- 2. Fig 15. – it would be interesting to see the difference between the fits and the truth. Do the fits break down in particular places?
  $\Longrightarrow$ *Good idea. We think by "places" you meant indeed the (R, z) position, right? As we already had a plot that compares the velocity distributions of the mock data with the best fit, but none showing the spatial distribution, we added a new plot that shows for one example data set the spatial residuals between mock data and best fit. We think that illustrates even better how the mock data did indeed not follow a single qDF anymore and had a different radial and vertical density profile.*

- $\Longrightarrow$ ***Additional changes:*** *We reran the tests in this section with the more realistic* `DHB-Pot` *(the original tests had been with the* `Iso-Pot`*) and replaced all plots accordingly. The results are quantitatively and qualitatively very similar to the original results. We added a short paragraph pointing out, that the results are therefore independent of the choice of potential. (The only difference was, that h_R was slightly less close to the true parameters of the* `hot` *qDF, so we removed a corresponding statement that claimed that the influence of the* `hot` *population is more important. While this could be still true, it wasn't that obvious from the plots to us anymore.)*

5.6 Section 3.6

- 1. Fig 19 is difficult to interpret. Is it possible to display the difference?
  $\Longrightarrow$ *We thought about it, but to be able to show the results for two analyses (*`hot` *and* `cool` *population) and to show not only the best fit but also the range of probable models as given by the pdf in only a few figure panels the way how we plotted it seems to be the optimal way. We added however two more panels in this figure that show not just iso-density contours as the old panels, but quantitative cuts through the density distribution. We think that should visualize the result even better.*

- 2. The fact the density is not well recovered seems interesting as it points to possible biases in the surface density of the disc/dark matter measurements if one uses the wrong potential. It would be good to have the discrepancy quantized in the text.
  $\Longrightarrow$ *That was a great suggestion! Thanks! It turns out that we really get some biases of $\sim 10\%$ in the disk/halo fraction at the Sun. When looking further into it we found however, that the disk/halo fraction is almost perfectly recovered at the radius where most of the stars are positioned. That the fit is best where most stars are located was to be expected - but it is still nice to see that the modelling behaves indeed as expected, even for a wrong potential model.*

- 3. I think Fig. 20 could be removed. As mentioned it doesn't make sense to compare the DF parameters between different potentials so I am not sure what Fig 20 is telling us.
  $\Longrightarrow$ *You are right that this figure does not show any new Science results, but we think it is illustrative for the reader, so we would like to keep it. To make it a bit more interesting, we overplotted the qDF parameters with the physical scale lengths and velocity dispersion we measured from the mock data. Now it demonstrates two things: (i) The actual physical velocity distribution is pretty well recovered by the fit. (ii) The action-based DF parameters in different potentials do not have to agree with each other or with the physical velocity distribution. For readers that are new to action-based modelling this might be not absolutely obvious and it would be worth to remind them. We hope you agree that we keep this plot. We also enhanced the discussion of this plot a bit.*

- $\Longrightarrow$ ***Additional changes:*** *The biggest change in this section is, that we reran the analyses. The reason for this was, that we made some minor but*

necessary modifications/corrections in the code, concerning the numerical accuracy of the likelihood normalisation and the action interpolation grid. We tested a bit if those small changes made a difference. Some of the tests we reran anyway and we did not see a qualitative change in the results - only in this test (for which the changes we made mattered most). Especially at small radii and in the mid-plane the recovery of the potential is now much better than before and the constraints are tighter. As the recovery of the flatness of the disk is now a better, we also removed a corresponding sentence claiming otherwise from the text.

5.7 Section 3.7

- 1. Should this section be moved to the discussion section?
  $\Longrightarrow$ *We thought about it, but we think it contains an—admittedly less important—result ($\longrightarrow$ that by comparing the differences of cool and hot populations in different figures in this paper, they seem overall equally well suited for dynamical modelling), but a result nevertheless, and should therefore stay in the result section in our opinion.*

- $\Longrightarrow$ ***Additional changes:***

  - $\Longrightarrow$ *We removed a sentence that claimed more or less that a reliable rotation curve measurement also constrains the potential better. But as this is only true in the midplane, we thought we rather remove it.*
  - $\Longrightarrow$ *We removed a whole paragraph from this section. This paragraph was discussing in which regions hot and cool populations constrain in Figure 16 the potential best. After we reran the tests in Figure 16 with some improvements in numerical accuracy (see above), there was no clear indication anymore that hot populations would constrain the halo better etc. While this could still be true, we did not see it directly in our plots. That we saw it in the old plot might have been a coincidence. We trust the new—and actual tighter and better—results more.*

6 Summary & Discussion

- 1. Perhaps add statements comparing the errors explored to those anticipated from Gaia.
  $\Longrightarrow$ *We are discussing it in in Section "3.4 Measurement uncertainties and their effect on the parameter recovery" in particular, but we also added a statement and reference here. We also included a paragraph on the caveat that we did the tests in the error section only with the* `Iso-Pot`.

- 2. The two approaches mentioned at the end of 'Gravitational potential beyond the...' are stated as formally similar but I think it is clear that one is better than the other. The true Staeckel approach limits you to potentials with the same foci. This is an obvious limitation and has been discussed before.
  $\Longrightarrow$ *You are right. After reading up a little bit more about it, we agree. Thanks for pointing it out. We changed the text accordingly.*

- 3. The definition of X in f(J,[X/H]) doesn't seem to make sense.
  $\Longrightarrow$ *There was a typo in the text... Apart from that we mean the whole abundance space of different elements by [X/H]. We tried to make that more clear in the text.*

- 4. The first section in future work is very interesting. Use of different DFs and potentials as explored in this paper is interesting but a true test of the apparatus on a more realistic galaxy would make the 'RoadMapping' tool much more attractive.
  $\Longrightarrow$ *We are working on it. :-) Stay tuned.*

- 5. I think that the final two questions of the future work section are weak. Clearly the rotation curve is only describing the in-plane force not the force everywhere. Parametrizations will naturally convince you that the rotation curve is well measured but I think there is a lot more flexibility. Also, the advantage of using the approximate actions is that more realistic potentials can be considered.
  $\Longrightarrow$ *You are right. Thank you for making things clearer for us. We changed the discussion accordingly and we also thought of some more interesting science questions to conclude the paper with.*

- $\Longrightarrow$ ***Additional changes:***
  - $\Longrightarrow$ *We added a paragraph in the sub-section on "Computational Speed", where we mention how long some of our analyses run—to address the referees concern that we would gloss over the problem a bit. We also discuss a bit more how computational speed restricts us to potentials with close-form expressions for Φ at the moment.*
  - $\Longrightarrow$ *Some of the results that we summarized in the discussion seem to hold not any longer after we reran some of the tests. We removed them from the discussion. In particular: (i) The statement of the VERY strong robustness against SF misjudgements and its probable reasons was rewritten and mitigated. (ii) The statement of different populations probing different regions of the galaxy was rewritten and mitigated. (iii) That the qDF h_R is especially important for recovering the potential scale length. (While this could be still true, we do not really see it in the results of this paper anymore.)*
  - $\Longrightarrow$ *We included a little paragraph mentioning Magorrian (2014) and his approach to marginalize over non-parametric DFs, as suggested by the referee.*
  - $\Longrightarrow$ *We mentioned also the very recent paper by Payel Das and James Binney in one sentence.*

- $\Longrightarrow$ ***Acknowledgements:*** *We added two more people to the Acknowledgements we would like to thank.*

6.1 Table 3

- 1. Can you add a summary column that summarizes the result? This would make the paper much more 'usable'.
  $\Longrightarrow$ *Done. Great idea.*

- $\Longrightarrow$ ***Additional Changes:*** *Parameters of tests that we reran were changed in this table, in particular in Tests 1, 4, 5 and 7. We also removed two rows completely from the table - they were about tests that were considered not so important by the referee. The corresponding plots were either removed or moved to the Appendix.*

7 Typos

$\Longrightarrow$ *Thanks for pointing them out.*

Abstract:
1. 3rd sentence – 'rules of thumb' for how data, model and machinery most affect ... and DF.
$\Longrightarrow$ *Sentence was removed anyway.*

Introduction:
1. Start of 3rd para: 'to constrain' → 'constraining'
$\Longrightarrow$ *Done.*
2. Start of penultimate para: 'to restrict' → 'restricting'
$\Longrightarrow$ *Done.*

Table 1:
1. 'troughout' in caption.
$\Longrightarrow$ *Done.*

Section 2.3:
1. Second sentence: 'about'→'on'
$\Longrightarrow$ *Done.*
2. 'the circular orbit' to 'near-circular orbit'?
$\Longrightarrow$ *Done.*
3. Top of page 4 two 'in's
$\Longrightarrow$ *Done.*

Section 2.6:
1. Top of right column page 7 – replace 'besides' with 'not only... but also'
$\Longrightarrow$ *Done.*

Section 2.7:
1. Need '(MCMC)' after MCMC
$\Longrightarrow$ *Done.*

Fig 14:
1. 'pest'→'best'
$\Longrightarrow$ *Done.*

Fig 15:
1. 'refereed'→'referred'
$\Longrightarrow$ *Done.*

Section 3.5:
1. Second paragraph right column page 13 – 'sun'→'Sun'
$\Longrightarrow$ *Done.*

Table 3:
'analyis'→'analysis'
$\Longrightarrow$ *Done.*

8 Comment from Dr. Eric Feigelson

==========
Below are comments on statistical aspects of the manuscript: ApJ submissions with a statistical component are previewed by Dr. Eric Feigelson, the member of the ApJ editorial board with a broad expertise in statistical matters in an astrophysical context.
++++++++++

An elaborate Bayesian inferential procedure is described in sec 2.6-2.7 for parameter estimation with results shown in Fig 6. But with uninformative uniform priors and a simply unimodal likelihood with a nearly multivariate normal distribution, this effort is unnecessary. The same result would be obtained with maximum likelihood estimation via the EM Algorithm (probably $\gg$100 iterations with convergence guaranteed by theorem) and parameter uncertainties estimated from the Fisher Information Matrix. The confluence of Bayesian and MLE procedures in this case should be presented.
$\Longrightarrow$ *We have added a small paragraph in Section "2.5 Data likelihood" that mentions that in the limit of non-informative priors there is no big difference between maximum likelihood estimation and Bayesian inference. But we also gave an argument that we expect to use in the future more informative priors and then the Bayesian inference procedure will be the proper way to approach the problem.*