The manuscript "Improving interpretation of sea-level projections through a machine-learning-based local explanation approach" provides a novel approach to analyzing multi-model sea level projections, using machine learning. The method put forward in this study has two parts: (1) using ML models as surrogates for the projection ensemble, (2) explaining the role of various modeling assumptions in generating the differences between sea level projections. The method is novel for our community and provides a potentially powerful way to analyze MIP output in a way that can guide potential future development (though this part needs more elucidation). I have some questions about particular choices in the methods and in general the writing would benefit from clarification. However, overall I find the contribution to be interesting and appropriate for The Cryosphere.

Major questions:
1. Within the context of an uncalibrated multi-model ensemble, the explanation approach adopted here explains why individual models deviate from the median projection. However, it does not say how an individual model deviates from the "truth" (since the models are not compared against reality or some benchmark). Thus, the authors need to be clear about this fact in their claim that this method can help in model development. How can it help in model development? If modelers now that factor X is causing their model to deviate from the ensemble median in some way, how will this facilitate development or adjustment of factor X? The authors needs to be more specific about this claim, which is essentially the motivation for using this method.

2. As explained in the manuscript, the SHAP method is always guaranteed to produce a set of quantified contribution (mu) that add up to exactly the total projection. This has some nice mathematical properties, but it also means that the quantified contributions are themselves dependent on the choice of which factors (x) to include in the analysis. For complex ice sheet models, this includes many, many more differences than the six included in this analysis. This leads me to three questions:
(a) The criteria you used to select which factors to include was those "without empty entry in Table A1 and with a sufficient number of variation across the models." The first makes sense - you need labels (though perhaps some discussion would be in order about whether it is possible to add labels by talking to the modelers response for these simulations). The second is unclear to me - it seems that you left out some potentially very important factors like initialization year and bed topography used with complete labels. Ultimately this selection, which would seem to be very important in the overall, seemed quite ad-hoc.

(b) Ultimately, I think much of the issue with (a) could potentially be solved by adding one more factor, which is the model itself (e.g., AWI-ISSM, NCAR-CISM, etc.). There are many modeling choices that are different between those models (e.g., how certain processes are parameterized, etc) which is not described by the six existing categories. Including this as a category would determine how large these inter-model differences are. Hopefully, they would be small contributors to differences, but if they aren't small, then this indicates that more effort is required to label these differences in a meaningful way. Right now, these hidden factors are likely being lumped in together with other factors where there are inter-model differences.

(c) Ultimately, there needs to be much more discussion of the sensitivity of the results to the choice of factors to include in the analysis. This can be done by a leave-one-out validation exercise to see how the results change if one of the "known" factors is converted to a "hidden" factors. This would help other researchers hoping to do a similar analysis to decide how to make this choice of factors (or how much effort to put into labelling unlabelled factors) at the beginning of the exercise.

3. One aspect of the procedure that I found confusing is the procedure for training and then selecting the ML models. First, it is unclear why an ML surrogate is need. The manuscript says "our knowledge on the mathematical relationship f(.) is only partial and based on the n MME results." I'm not exactly clear on why this is a problem, so more explanation would be appreciated. Furthermore, you then fit a range of ML models and for each MML result, you pick a different ML model fit depending on which does a better job at reproducing this particular result. It is unclear why you do it in this way as opposed to just picking the ML model which best fits all the results. In particular, this seems to conflict with your prior statement that you need knowledge on "the mathematical relationship f(.)" in order to perform the explanation part of the procedure. Overall, this whole aspect of the procedure is quite confusing and could do with clearer explanations as to its purpose in the overall study.

4. Similar to point #3, the introduction of the ML model quality metrics (Q, RAE, MAE) is confusing. It is unclear exactly how each of these metrics is used in order to perform model selection. Is MAE even used? If not, then its not clear why it is introduced.

Other points:
Line 10-12: confusing sentence  
Line 24: can be->are  
Line 28: ice-sheet->ice sheets  
Line 33: the key to what?  
Line 33: its not clear that this method can tell us why the numerical delivered some particular results (in the absolute sense), but rather why it differs from other model results  
Line 35: "this" view? which?  
Line 39: how particular modelling  
Line 43: positive with respect to model sensitivity?  
Line 60: local importance of what?  
Line 80: Is this a test case or just the point of the study?  
Line 92: more discussion could be useful about why GCM choice is not included. I can see that this would be done to reduce complexity, but ultimately GCM choice is important to the full ensemble of results from ISMIP. At a minimum, better explanation of why this difference is omitted would be helpful.  
Line 105: explain further about "preliminary groupings"  
Line 127: the f(.) notation is generally very confusing to me, not sure what "." means in this context  
Line 151: why are these particular ML models used? Also, I think its a stretch to call a linear regression an ML model (I get that its a simple benchmark, but most people would just call it a statistical model)  
Line 157-158: not sure what this sentence means  
Line 163: explain what you mean by "strong theoretical basis"  
Line 181: the objective of this section  
Line 181: ...by h_theta through the use of...  
Section 3.3: I generally really liked this explanation of Shapley approach  
Line 237: from the global mean prediction to what?  
Line 235: The definition of the Shapley values guarantees that the sum...  
Line 242: where is equation (5)?  
Section 3.4: it would help here to give an example of the sort of dependence between inputs you are thinking of. I know you talk about this later in the results, but its hard to understand what you mean by dependence here without a concrete example.  
Line 266: the procedure describe in...  
Line 269: sensitivity of projected SLR to modeling assumptions at difference levels  
Line 295-298: The results described here are confusing, because from Figure 5a, it seems like the best performing models are all RF.  
Line 304: I could not find anything outlined in orange (maybe I'm not looking hard enough...)  
Line 319: the use of "perform" here is confusing  
Line 324: sensitivity of what?  
Figure 6: Can you add a label to each panel indicating which model experiment it is. Also, it seems like it would make sense to color bars which are "negligible" (in terms of being less than the error) as grey or something similar to indicate that their importance is below the threshold that you have defined as being interpretable. This would help reduce the amount of information the reader has to absorb when looking at this very interesting but dense figure.  
Line 359: large absolute value of kappa  
Line 365: the results are non-existent  
Fig 7 and Line 371: I'm not sure the smooth fitting is defensible. There is clearly overfitting happening here, and you even talk about (for example) mu=+3 cm in Figure 7b, when such values are clearly only a product of the smoothing in an area of parameter space with no simulation.  
Figure 7b: the dip in mu between 2 and 4 km resolution make me wonder whether the modeling of input dependence has completely removed this effect. My guess is that the three results with 3 km resolution are all a single type of model which exhibits
idiosyncratic behavior for other reasons beside resolution. This gets to point 2b above, but also why I think it is inadvisable to fit a smooth curve to these plots.

Line 383-384: Confusing sentence

Figure 9: this figure needs better explanation as I am not sure what the box plots represent compared to the red dots, etc.

Line 410: which provides narratives about the role of various modeling choices in generating inter-model differences in the GrIS MME

Line 419: affected by too coarse grids (here coarser than 5 km)

Line 424: could not have

Line 431: comparison of ANOVA with

Line 435: it would be useful to quantify the extent to which this problem is alleviated or not - several prior comments relate to the possibility that input dependence is still present in results

Line 436: for the high performance of our approach is

Line 443: In this study, we described the use

Line 444: assumptions in sea-level projections

Line 455: global effects of what?