Review of –"Imputation of missing IPCC AR6 data on land carbon sequestration"

Summary- The AR6 database and its results are widely used to understand and analyze future climate mitigation and adaptation pathways. Since some IAMs used in the AR6 database model/report net land use emissions rather than gross, data on carbon sequestered on land is often missing. In this paper, the authors conduct an imputation/statistical interpolation to calculate data on carbon sequestered on land for the AR6 database where not reported. Rather than developing a method to convert the net land use to gross, the authors adopt a statistical approach which directly calculates the carbon sequestration from land.

The land carbon component is indeed a key component of the database. Therefore, I find the work of the authors important. However, the paper as presented seems to be related to a specific method of interpolation applied an existing set of data rather than the creation of a generally usable dataset (which would put this paper outside the scope for a data journal like ESSD) (**See Major comments 2, 4**). Moreover, I had several questions and concerns regarding the general applicability of these methods outside of interpolating the current state of the AR6 database (**See Major comments 1,2,3**). Moreover, as a minor but important point- while the length of the paper does not automatically equate with quality (and the paper is clearly written), I found that the paper is too brief in its explanations regarding its variables, methods (As a simple example, it contains no methods section neither in the main text or the supplementary file which is usually critical for a journal like ESSD) (**See Major comment 5**). Please see all detailed comments below. **Given the comments, I recommend rejection and resubmission at a later date**. However, this is subject to the editor's discretion. All my comments are meant as constructive and in good faith to my colleagues in the field.

Major comments-

1. **Scale of analysis**- One of the fundamental questions I had was whether this imputation exercise is conducted at a global scale or at a regional scale? The actual dataset released contains only the global results (https://zenodo.org/records/10696654). Also, from the text, I interpreted all results as global? i.e., the independent variable is the carbon sequestered on land globally? If this is the case, this seems to be a shortcoming of the approach since it ignores regional heterogeneity. All IAMs included in the AR6 database produce regional emissions results. An imputation such as this seems to consolidate all the underlying regional dynamics to a regression which I find unconvincing. This is especially true in an age when studies are focused on deriving fine resolution (pixel level) results from regional ones. Can the authors comment on this? How valid would this method be if applied to the regional scale? Also see comment no 3 below. If regional results are calculated, kindly discuss them in the manuscript (See sub-comments of comment no 5).
2. **General applicability of this method**- Since both the dependent and independent variables are coming from the same version of the AR6 database, it seems to me that this

method really just begins and ends with the current state of the AR6 database. The IAMs underlying this database are constantly evolving and several of these IAMs are now focused on developing gross emissions pathways from the land sector. If a new version of the AR6 database were to be released, would this method still be valid? If the answer is no, then the data created and presented here is not a dataset with general usability. Rather , it is just a method of extension of the current AR6 database.

3. **Applicability beyond AR6-** Related to the above point, the Global Carbon Project (GCP) which releases its carbon budget analysis that is annually published in ESSD (https://essd.copernicus.org/articles/15/5301/2023/), includes land use emissions and carbon sequestration historically, globally. In fact, in more recent versions, there are also national emissions and sequestrations included. Would the current method produce reasonable results for a database such as the GCP's? Since the GCP is an equally well known dataset/exercise with a richer historical dataset, I would recommend that the current method be tested on that dataset to justify general usability of this method and data. Note that the national and global land use emissions and sequestration numbers are also available here- https://www.globalcarbonproject.org/carbonbudget/archive.htm
4. **Interpolation of existing data or new data**- Given the above points, I believe rather than a generalizable imputation, the authors have rather presented an interpolation method for existing data on the AR6 database. While this is not an issue, this would classify this paper as a methods paper rather than the development of original and novel data. In which case, this paper is not a good fit for the given journal which is generally meant for data descriptors. Finally, I would describe the methods of the authors as an interpolation rather than an imputation. That seems like a more precise description of the methods in the paper.
5. **Lack of description of methods (and results)**- While the paper provides a summary of the methods, the, they are never really described in any detail in the manuscript. As a simple example- ESSD papers generally contain a detailed methods sections which describe and justify the methods underlying the data which increase usability and reproducibility. This paper does not describe the method used in much detail which I find concerning. I have added specific comments related to the same below-
   - **i)** Can the authors show a scatterplot of the x and y variables underlying the R squared values shown in Figure 2? Can these be separated for the training and the testing dataset?
   - **ii)** What are the actual equations used for each of the methods/models? What were the final coefficients chosen for the gradient boosting method?
   - **iii)** I checked the AR6 database, and it seems regional results are available. Therefore, related to comment 1 above, is the regression trained on global or regional results? If it's trained on regional results, can the regional heterogeneity be discussed in the manuscript? Does the regression fit all regions in the same way?
   - **iv)** How and why is the current independent variable (IV) chosen? Were other IVs tested? Does performance change when different IVs are chosen?

v) The selection of the four final models for hyper parameter testing are described in a few short sentences. Can the details regarding the selection of the four models be added to the paper or supplementary information?

vi) I had trouble understanding Figure 3. What are the categories described in the figure? Can the authors describe what each of the categories represent? Why do C8 category emissions differ so much from observed compared to other categories? Was training and testing differentiated by category?