Comment on egusphere-2022-465
Anonymous Referee #1

Referee comment on "UVBoost (v0.5): an erythemal and vitamin D-weighted ultraviolet radiation estimator based on a machine learning gradient boosting algorithm" by Marcelo de Paula Corrêa, EGUsphere, https://doi.org/10.5194/egusphere-2022-465-RC1, 2022

Review of GMD manuscript egusphere-2022-465

Title: UVBoost (v0.5): a hybrid radiative transfer and machine learning model for estimating ultraviolet radiation

Author(s): Marcelo de Paula Corrêa

MS No.: egusphere-2022-465

MS type: Model description paper

General Comment: The article describes in detail the development of a UV radiation model (UVboost v0.5) based on SML (Supervised Machine Learning) tools. Seven different methods where chosen and after a process of cross-validation, the CAT Categorical Boosting (CatBoost) technique gave the best results, showing the best accuracy at the lowest computational cost. The accuracy of the model is also based on the chosen physical radiative transfer model TUV, well recognized in the radiative transfer community that works in the UV radiation field. Certainly the model UVboost v0.5 has a clear application, with well defined limits: clear skies and to predict or to estimate a great number of data (well spatial or temporal) and not to predict very specific and accurate situations.

The paper is well structured and well written but a revision is needed for next publication.
The first thing is related to the employed methodology, the SML methods. These methods are classified as part of the Artificial Intelligence (AI) but different classifications appear in the literature, as for instance Neural Network method. I’m not expert on them and hence a short classification (table or scheme) in section 2 would be welcome.

Also, I think that the author needs to check if other models of this type have been published in the literature for solar radiation or UV prediction/estimations or related with this topic. There are no references to this in the article. For instance I have seen the paper “Review of photovoltaic power forecasting by J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F.J. Martinez-de-Pison, F. Antonanzas-Torres. Solar Energy 136 (2016) 78-111, which in my opinion is related to this discussion.

The range of values selected for the AOD in Table I is very high, from 0 to 15. Therefore my question is if these values are spectral (given for a wavelength) or are integrated or broadband values corresponding to the integration over the whole spectral UV range of the model. Values of AOD greater than 1-2 (for a given wavelength or spectral AOD) are already very high (although used and measured values: i.e. AERONET in China) and are into the conditions of multiple scattering where the use or application of Beer-Lambert law is not correct. Values greater that 1-2 are very rare (not frequent) as can be seen in the values used in this article for the comparison. The use of the proposed high range of AOD values may disturb the physical model and create a high number of simulations that are not needed, which may also disturb the applied Machine Learning methods. Therefore, explain this problem about the values of AOD.

I recommend and subsection where briefly describing the model (input-output-main core) in order to be run by the users, since the various information is scattered throughout the text. The model calculate UVI index and “Vitamin D weighted irradiance” but it is not clear for me if the model gives the UVER (W/m2) as output. As mentioned by one of the referees this type of models based of SML are difficult to replicate.

For all of this, I considers that he paper may be accepted for publication after the recommended revision.