Reply on RC2
Marcelo de Paula Correa

Author 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-AC4, 2022

Dear anonymous reviewer #2,

I thank you for the careful review of my paper. You proposed a series of reviews, which were mostly fulfilled. There were some suggestions that were partially answered. However, an appropriate argument for not completing the request was provided by me. I am open to answer and try to solve any further questions that may arise with regards to this manuscript.

- Major Comments:
  - I believe that the discussion of UV health effects and machine learning methods in the manuscript are sound and the author demonstrates expertise in both areas.
  - However, the basis of this machine learning approach is a database from a radiative transfer model.
  - In this area the author consistently uses wrong terminology and applies the wrong methods to the problem at hand, which suggests insufficient expertise in this field.
  - As the computation of the regression data is the fundamental basis of the manuscript, I cannot recommend a publication.
  - A further major and related issue is a completely inadequate description of the atmospheric input data for the radiative transfer model.

Answer: Again, I would like to the thank for the clear review of my manuscript. I feel that these recommendations, once addressed, strengthen the manuscript. I hope the revised version is now suitable for publication and look forward to hearing from you in due course.

Firstly, it is important to clarify that I used the terminology “hybrid model” for UVBoost with the best of intentions. After all, it is an input (SZA, TOC, AOD)/output (irradiances) code with a core based on an SML treatment (CatBoost). Despite its simplicity, it is fast, very efficient and, above all, it provides very accurate results.

Therefore, I strongly disagree that the methods are wrong and that the study is a mere
regression calculus. The article was carefully written and my objective was to contribute to an important scientific gap, such as the use of SML techniques in erythemally/vitamin D weighted UV Irradiance prediction.

If the reviewer considers this work a simple application of third-party codes, I agree with the executive editor's recommendation to forward this article to the EGU journal ESSD (please check the interactive discussions).

- This is compounded by a variety of further issues:
- The title of the manuscript mentions a hybrid radiative transfer and machine learning model. This would imply that the model somehow combines traditional radiative transfer techniques and machine learning, which is not the case. This misleading title should be corrected.

**Answer:** Again, I used the terminology "hybrid model" for UVBoost with the best of intentions. According the Merrian-Webster Dictionary, the term 'hybrid' is defined as 'something heterogeneous in origin or composition'. The proposed model is composed by a combination of two heterogeneous methods: a traditional radiative transfer code and a SML regression technique.

However, if this title gives a wrong impression, I suggest a new title such as "UVBoost (0.5): Erythemal and Vitamin D weighted UV radiation estimator based on a Machine Learning gradient boosting algorithm".

- While the introduction section is overall well written and comprehensive, it is missing a paragraph on existing machine learning approaches to radiative transfer, as this is a quite mature field with lots of active development.

**Answer:** The author thanks the reviewer for pointing out this deficiency that were corrected in this new version. Please check the new paragraphs included in Section 1 – Introduction:

"ML methods have also been used for solar radiation prediction; and, this use had increased significantly in recent years. A review published about 5 years ago showed that the use of some ML techniques such as boosting, regression tree or random forest for solar radiation prediction were still rarely used (Voyant et al., 2017). Currently, ML techniques have been applied in different areas, such as solar panel power prediction (Zazoun, 2022), satellite radiation estimates (Cornejo-Bueno et al., 2019), and, as a complement to surface radiation measurements in different spectral bands (Feng et al., 2020; Narvaez, et al., 2020). After all, most weather stations cannot reliably provide global solar radiation observation data and finding an accurate way to predict this is very important.

A recent and in-depth review study with more than 230 scientific publications in 20 years showed that: a) data quality control before model prediction is essential; b) methods that use training models are more accurate than methods based on statistical filters; c) novel and combined ML techniques tend to be future hot topics (Zhou et al., 2021). These elements are all addressed in this work. As there are still very few studies on UVR prediction using ML methods (e.g. Wu et al., 2022), this study offers a new contribution on the topic."
- line 85: Please provide adequate proof or a reference for the claim that UVR fluxes cannot be properly predicted with cross-validation techniques.

**Answer:** I appreciate the reviewer's concern with this detail. But, in line 85, I refer to the cross-validation techniques applied to the MLR (However, even when using cross-validation techniques one cannot properly fit an MLR model to UVR predictions). Table 3 shows these cross-validation statistics. So, I think that this is an adequate proof. For clarity, I complemented the sentence with: "as we will see later in this article".

- line 170: TUV is a radiative transfer model, not a database in itself. What you are probably referring to here is the TUV climatology that is available for download.

**Answer:** I apologize for my lack of attention and I thank the reviewer for pointing out this error. The sentence was fixed: "The database for testing and training was built from calculations performed by the RTM TUV v5.3.2 (Madronich and Flocke, 1997; NCAR, 2022)." PS: I also put the TUV link in the references according a "minor revision" recommendation.

- line 173: Which UV irradiances were calculated here, surface, TOA, or something else?

**Answer:** I recognize that this was not properly explained in the text. UVBoost estimates downward UVR at surface. I clarified this information in the text.

- line 179: How can you vary the total ozone content, but keep the atmospheric structure (i.e. the ozone profile) constant?

**Answer:** I thank the Reviewer for pointing out this lack of information. The atmospheric profiles were changed according the geographic position. For this reason, I included the following information in the paragraph: "The TOC atmospheric vertical profile was adjusted by the geographic position according the AFGL Reference Model Atmospheric Profiles (Anderson et al., 1986; Gordon et al., 2022). The following vertical distributions were used: tropical atmosphere profile for the gridpoints between the equator and 30° latitude, mid-latitude profile between 30 and 60° latitude, and subarctic profile above 60° latitude."

- line 180ff: If scattering clouds were not considered in this study, then the application of DISORT with 8 streams makes very little sense. A pure emission solver would be sufficient to accurately account for the ozone absorption of the UV radiation. The only phenomenon that would necessitate a scattering radiative transfer solver is Rayleigh scattering, which is not the dominating factor here, compared to the ozone absorption.

**Answer:** The reviewer's question is pertinent, but I would like to clarify my option for more complex calculations. I used the DISORT application with 8-streams to be as accurate as possible in view of the presence of aerosols and molecular scattering. Results using simpler methods (e.g. 2-fluxes delta-Eddington, etc) may be reasonable for most no-aerosol clear-sky calculations. However, my goal was to build a robust and very accurate
- A thorough description of the atmospheric profile input data for the TUV radiative database for training.

- A description of the crucial ozone spectroscopy used in the radiative transfer model of this study is completely missing.

**Answer:** I disagree with the need to incorporate a discussion on the ozone spectroscopy. While I agree that the paper ozone spectroscopy is an interesting aspect in and of itself, my paper focuses on the SML method for the UVR estimation based on atmospheric parameters input data, such as TOC and AOD. The ozone spectroscopy used in the TUV has already been well discussed in the references of the code itself.

Therefore, I believe that adding this dimension would not contribute to the scope of my paper. A focus on the ozone (or other gases) spectroscopy could be a good topic for a follow-up study, e.g., on the use SML for infrared radiation inference.

- line 181: How can large variability and spatio-temporal complexity for planet cloud cover mean less accurate RTM results?

**Answer:** I thank the reviewer for his concern on this matter. It is well-known that clear-sky RTM calculations are more precise than cloudiness calculations. In general, the results of the clear-sky calculations from most RTMs are nearly identical (Aumann et al., 2018). Using cloudy observations in forecast models is difficult. Firstly, clouds show large spatio-temporal variability. Besides, cloud physics uncertainties, large variability of the vertical distribution of ice and liquid water in clouds, challenge of cloud structure representation, 3D effects, and cloud overlap assumptions. I included this comment and the reference in the paper.

- line 182: Does your model use the Cloud Modification Factor you mention to correct for clouds?

**Answer:** No, UVBoost estimates only clear-sky irradiances. Cloud Modification Factor (CMF) is defined by the ratio between the measured UV radiation in a cloudy sky and the simulated radiation under cloud-free conditions (Foyo-Moreno et al., 2001). In fact, the CMF is used to estimate, in an approximate way, the radiation attenuation caused by cloud cover. In this case, the effect of clouds is given by the product of clear sky irradiance, previously calculated by an MTR (or, in this case, by the UVBoost), and the cloud modification factor (CMF).

CMF may be estimated by using clear-sky RTM simulations and ground-based observations. Or even, using look-up tables for different cloud types (e.g.: http://i115srv2.vu-wien.ac.at/UV/booklet/par_4.htm). In fact, UVBoost can even be used for this type of study.

transfer model is completely missing.

**Answer:** It was fixed. Please check the answer for line 179.

- Subsection 2.3, line 212: The dataset does not include a validation partition to check its generalization capability.

**Answer:** Dear reviewer, SML datasets used in the UVBoost development, including training and test files, are available at https://doi.org/10.5281/zenodo.7027724.

- **Minor Comments:**

  **Answer:** I thank you for this careful review. You proposed a series of minor reviews, which were all fulfilled.