Response to Reviewer 2

We would like to thank the reviewer for their valuable time and feedback. The reviewer's comments are in ***Times (italic and bold)***, our response is in Red Arial.

This manuscript addresses a well-recognized issue in ground-based solar radiation QC: The globally uniform thresholds of traditional ERL tests are overly conservative and frequently insensitive to localized environmental variations. Through a newly proposed method, this work effectively tightens conservative limits. This paper is well-written, logically sound, and worthy of publication in AMT after minor revisions. Below are my detailed comments.

We sincerely thank the reviewer for their careful reading of our manuscript and for their accurate summary of our work.

1. Eqs (1)-(3) and Eqs. (5)-(7) both contain the cosine of zenith angle. Would it be clearer if the notations are standardized? The same applies to the notation in Line 352.

We thank the reviewer for their careful review. We have standardized the notation during revision, where \mu_0 is now used consistently throughout the manuscript.

2. In Section 2.2.1 and Fig. 2 of the manuscript, there is an inconsistency regarding the feature variable abbreviations. While the figure caption denotes the feature variable as B_n, the corresponding label inside the actual illustration of Fig. 2 is marked as B_h.

Indeed, the figure caption contains a typo, which has been corrected during the revision; it should be B_h.

3. It is unclear to me why the number of harmonics is set to 25. If it is simply the yearly cycles that need to be identified, a few harmonics should work, given the regularity of the seasonal component.

Thank you for this comment. Reviewer 1 has raised the same issue. Therefore, we are pasting our earlier response here:
"The choice of m=25 was determined through an empirical sensitivity analysis aimed at balancing the goodness of the seasonal fit against the risk of overfitting. When testing lower harmonic orders (e.g., < 10), we observed that the seasonal fit was insufficient—effectively underfitting—and failed to adequately capture the true amplitude of the annual radiation cycle. Conversely, significantly higher orders introduced high-frequency artifacts that did not correspond to physical climatological signals."

4. In Section 3.2, the authors utilize the L-BFGS-B algorithm to optimize the coefficients of the structural limit curves. However, the choice of the initial guess for this iterative solver is not specified.

Thank you for this comment. For each component and cluster, L-BFGS-B starts from a=1, b=1. We have added this information in the revised manuscript. Furthermore, our work is reproducible, i.e., the code is on GitHub, similar algorithmic details can be found there.

5. The objective function for parameter optimization is designed to maximize the F1-score. While this is mathematically reasonable, the manuscript would benefit from a brief discussion on its meteorological and practical implications. Specifically, the authors should explicitly justify why a balanced F1-score is preferred in this radiation QC framework.

F1 is used because QC must balance rejecting true outliers (recall) and keeping physically valid low irradiance (precision). Recall-only limits are too aggressive and discard legitimate cloudy/polluted observations, as with the ERL; precision-only limits are too lenient and miss inconsistent G_h, B_n, and D_h points. *F1* targets a compromise suited to BSRN, where both data quality and retention matter. The isolation forest on (k_t, k_d, k_b) supplies the multivariate reference for that trade-off. We have added a brief note on this in the method section.

6. From Table 3, it shows that the values of parameter c_G are often very small, i.e., near zero. This is inconsistent with the original test, where the parameter c should reflect the measurement uncertainty, as also mentioned by the authors in the introduction. At the same time, the values of c_B are capped at 10. This suggests an issue with the optimization setup.

Thank you for raising this very important criticism. Indeed, our initial plan is to automate the selection for all three parameters in the inequality. However, we do agree that this violates the physical meaning of these tests. Therefore, in the revised version, we have modified our code and method, such that they only perform optimization on *a* and *b*, but keeping *c* the same as the original ERL test. Because of this modification, several tables and figures need to be, and have been, updated. That said, the findings and conclusions remain valid and unchanged.

7. The authors should explain how the proposed radiation classification can be applied to other locations.

We thank the reviewer for raising this interesting comment. The seven radiation regimes are identified, in an *unsupervised* fashion, from the BSRN data: PCA reduces the 13 extracted irradiance features, and Ward hierarchical clustering groups stations by multivariate similarity, without predefined geographic or climatic labels. The regimes are therefore defined by observed radiation statistics, not by location. Once this classification is fixed, it can be applied to new stations or locations as a *supervised* assignment step: extract the same 13 features from 1-minute G_h, D_h, and B_n, standardize and project them into the PCA space, and assign the site to the nearest regime. We have added the same explanation in the revised manuscript.

8. The authors are expected to append some qualitative scientific remarks in the discussion or future outlook section to enhance the comprehensiveness of this methodological framework. For instance, can this approach be used to handle other QC tests, such as the closure test?

Thank you for this interesting remark. We agree and have added some qualitative remarks to the conclusion. In short, the framework is not limited to single-component limits, the same regime-based logic could extend to other QC tests—including regime-specific closure tolerances (e.g., stricter under stable clear-sky regimes, relaxed under variable marine/polluted conditions) and tracker consistency. That would move from "one global ERL" toward a multi-test, climatology-aware QC suite. We note this as a natural next step for future work.

9. Comparing Fig. 7 (Class 6, Bn) with Fig. A6 (top two panels), the scatter plots look different. Why? It seems to me that the latter contain more points.

Thank you for your careful review. The two figures use different data by design. Fig. A6 plots all 1-minute data for GAN (2015) and GUR (2018) separately to show aerosol attenuation (~165k–200k points per panel). Fig. 7 plots the subsampled fitting dataset (top 5% per zenith bin + 10% random, then 20% for clarity), with both stations pooled across all years (~28k points total). Fig. A6 illustrates the physical regime; Fig. 7 validates the fitted QC limits. The lower density in Fig. 7 is from deliberate subsampling, not inconsistent data.