Reply to Reviewer #2

We thank the reviewer for their careful reading of the manuscript and for their constructive comments and suggestions. Their comments have helped us improve both the clarity and the scientific quality of the manuscript. Below we provide a point-by-point response to every comment. Reviewer comments are shown in **bold** font while our responses are given in normal font.

"This manuscript presents a machine-learning-based framework for diagnosing and correcting irradiance residuals in the CAMS Solar Radiation Service (CRS) using XGBoost and SHAP-based feature attribution. The study addresses a relevant topic and provides an interesting combination of operational irradiance products, explainable machine learning techniques, and physical interpretation of model residuals. The manuscript is generally very well written and the proposed framework shows promising potential for improving solar radiation products and investigating the factors associated with their errors.

Some aspects of the methodology and interpretation require possibly further clarification. In particular, the distinction between predictor importance and physical error attribution should be discussed more carefully, as SHAP values quantify associations with the model residuals but do not uniquely identify or describe their underlying physical causes. Additionally, some conclusions regarding aerosol-related error sources appear a bit stronger than directly supported by the presented analysis. I also believe that the manuscript would benefit from additional discussion of representativeness effects, uncertainties associated with the residual definition, and the robustness of some conclusions.

Furthermore, the presented SHAP framework offers substantial potential beyond the analysis shown here. A station-level and/or seasonal investigation of the SHAP contributions could provide valuable insight into the geographical and temporal variability of the dominant error drivers and would further strengthen the physical interpretation of the presented results. Such analyses may also represent an interesting direction for future work.

I think this is a very interesting work worth being published. Some comments below might be useful for improving the manuscript."

We thank the referee for the positive assessment of our work and for the constructive comments. We are pleased that the referee finds the proposed framework relevant and potentially useful for improving solar radiation products and understanding their error characteristics.

We agree with the referee that additional care is needed when interpreting SHAP-based feature importance in terms of physical error attribution. In the revised manuscript, we have clarified that SHAP values quantify associations between the predictors and the modelled residuals rather than uniquely identifying the underlying physical causes of the errors. We have also revised several statements related to aerosol effects to avoid overly strong causal interpretations and added discussion of representativeness effects, measurement and retrieval errors, and other contributors to the analysed residuals. Furthermore, we have expanded the discussion of the limitations and interpretation of the results and added station-level and seasonal SHAP analyses as promising directions for future work.

“The detailed comments are provided below:”

“Line 65: “The CRS provides...” Please introduce the full name of CRS at its first occurrence.”

Revised so that the full name of the acronym CRS is now introduced both in the title of the section and at its first occurence in this section.

“Line 67: Please specify the temporal coverage of the CRS dataset used in this study, as well as its temporal and spatial (latitude/longitude) resolution.”

We have defined the temporal coverage (February 2004 to present for Meteosat and January 2016 to present for Himawari field of view) and temporal (1 minute) and spatial resolution (as indicated by the CAMS CRS uses various input data resolutions and the final radiation data is interpolated to a single point of interest) of the CRS data in the revised manuscript.

“Line 73: Please introduce the acronym “AOD” after the first occurrence of “aerosol optical depth”.”

Revised to introduce the acronym AOD.

“Line 74: Please introduce the acronyms “TCWV” (total column water vapour) and “TCO or TCO3” (total column ozone).”

Revised according to reviewer suggestion.

“Line 80: Please provide more details regarding the CRS v4.6 service, including a brief description and appropriate reference(s).”

More details, description and references added.

“Line 86: Please introduce the acronym “SZA”.”

SZA introduced in the revised version.

“Figure 1: It would be useful to provide additional information on the temporal availability or number of observations per station of each network. This could be visualized, for example, using marker size or color shading while retaining the current network classification.”

Instead of adding the information to Figure 1, we have included the number of data points for each station to Table 1. We also added the temporal coverage (year 2021) to the Section 2.2.

“Lines 112-113: The ground-based measurements were averaged within centered ±15 min intervals the actual pixel observation time. Given the strong temporal

variability of solar irradiance, particularly during morning and afternoon periods when SZA changes rapidly, have the authors investigated the sensitivity of the results to shorter averaging windows (e.g. ±5 or ±10 min)? I believe that a brief sensitivity analysis or discussion would strengthen the methodology."

We have added discussion about temporal averaging to Section 2.2: "The 15-min averaging window was selected to reduce the influence of short-term fluctuations in surface irradiance caused by cloud variability and to improve representativeness between point measurements and CRS estimates. While shorter averaging windows would better preserve rapid irradiance changes, they would also be more sensitive to timing mismatches and sub-pixel cloud variability. The selected averaging interval therefore represents a compromise between temporal resolution and collocation robustness.“.

"Section 2.3: The theoretical formulation presented in Eqs. (1)–(3) links the residual error to deviations between the ideal and approximate CRS radiative transfer framework. However, the target variable is defined as the difference between ground-based observations and CRS estimates and therefore also contains contributions from measurement uncertainties, retrieval errors, and representativeness errors associated with the comparison between point (or station) observations and grid-scale estimates. For example, under broken-cloud conditions, a point measurement may observe a cloud-free period while the corresponding CRS pixel is partially cloudy (or vice versa), resulting in substantial residuals that are not directly related to deficiencies in the radiative transfer model itself. Please discuss how these additional sources of uncertainty may affect the interpretation of the ML-predicted residuals and the attribution of errors to specific CRS input parameters."

We thank the referee for raising this highly important point. We agree that the residual analysed in this study is not solely determined by deficiencies in the CRS radiative transfer modelling but it also contains contributions from additional uncertainty sources associated with the comparison between modelled and observed irradiances.

We have revised the manuscript accordingly. We added to Section 2.3 examples of potential error sources ("...including, for example, the measurement uncertainties, retrieval errors, and representativeness errors.") and added the following discussion to Section X.X:
"It should be noted that the residuals analysed here contain not only the deficiencies in the CRS radiative transfer modelling, but also measurement uncertainty, representativeness differences between point observations and satellite-scale pixels, and collocation and retrieval errors. Consequently, the SHAP analysis identifies the factors associated with the total residual rather than isolating individual physical error sources.".

We further clarify that, given the high quality of the ground-based radiation observations and the fact that the residuals are analysed as a function of CRS input variables, we expect a substantial fraction of the residual variance to be related to deficiencies in the CRS modelling framework. However, the additional uncertainty sources discussed above cannot be completely separated from the residual and should therefore be considered when interpreting the SHAP-based error attribution results. Therefore, we also added the following to the results section: "Nevertheless, given the high quality of the ground-based observations used in this study and the fact that the residuals are analysed as a function of CRS model inputs, a substantial fraction of the explained residual variability is expected to originate from deficiencies in the CRS modelling framework. However, the different contributions to the residual cannot be fully separated in the analysis and should therefore be considered when interpreting the SHAP-based feature attributions."

"Lines 165-166: The authors state that the XGBoost hyperparameters were "tuned through experimentation". Please provide additional details regarding the hyperparameter optimization procedure, e.g. including e.g. the evaluation metric used or the range of tested values."

We agree that the original description was too brief. The hyperparameters were selected through a manual tuning process using the validation dataset, with the validation RMSE as the selection criterion. During this process, several combinations of the main XGBoost hyperparameters were evaluated. While a complete record of all intermediate trials is no longer available, we have revised the manuscript to clarify the tuning procedure and to report the final hyperparameter values used in the study.

"Section 2.5: Several predictors used in the analysis are expected to be strongly correlated (e.g. Cloud Optical Depth, Cloud Coverage, and Cloud Type, as well as some of the surface reflectance parameters). Could this affect the SHAP-based feature attribution and the interpretation of the feature importance rankings?"

We agree that correlations between predictors can affect the interpretation of SHAP-based feature attributions. However, SHAP values are derived from Shapley values in cooperative game theory and provide a principled framework for distributing feature contributions while accounting for interactions among predictors. Consequently, SHAP is generally considered a robust and theoretically well-founded approach for feature attribution in complex machine-learning models. Nevertheless, in the presence of strongly correlated predictors, the attribution assigned to individual variables may not be unique and some uncertainty may still remain regarding the exact distribution of importance among correlated features. Therefore, in this work, we have also considered broader feature groups (aerosol, cloud, surface, gas and others) and overall patterns when interpreting the results. We have revised the manuscript by adding the following discussion of this limitation to Section 2.5:
"Although SHAP provides a theoretically well-founded framework for feature attribution based on Shapley values from cooperative game theory, strongly correlated predictors may share explanatory power, resulting in some ambiguity in the attribution assigned to individual variables. Consequently, the exact ranking of individual predictors should be interpreted with caution. In this study, we therefore also consider broader feature groups when interpreting the results, in addition to individual features."

"Lines 200–201: Since the evaluation of the original CRS products is presented later in the manuscript, consider briefly summarizing the baseline CRS performance earlier in the Results section to provide context for the subsequent machine-learning-based corrections."

We agree that a brief description of the baseline CRS performance provides useful context for the subsequent SHAP analysis. We have therefore added a short summary of the baseline performance at the beginning of the Results section, with a reference to the detailed evaluation presented later in the manuscript.

"Line 209: I think it would be better to introduce the acronym "SZA" in Section 2.1."

In the revised manuscript, acronym SZA is now introduced in Section 2.1.

"Lines 222-224: The authors suggest that the importance of AODAM and AODNI may indicate inaccurate optical properties of ammonium and nitrate aerosols in the CRS model. However, SHAP values primarily quantify the contribution of predictors to the ML-estimated residuals and do not directly identify the underlying physical cause of the error. The observed importance of AODAM and AODNI may also reflect aerosol retrieval uncertainties, interactions with cloud properties, or representativeness errors associated with relatively "polluted" environments."

We agree that the original wording was too strong in directly linking the importance of AODAM and AODNI to inaccuracies in aerosol optical properties. SHAP values quantify associations with the machine-learning-estimated residuals and do not uniquely identify the underlying physical causes. We have therefore revised the manuscript to clarify that the observed importance of AODAM and AODNI may reflect multiple factors, including aerosol retrieval uncertainties, aerosol-cloud interactions, representativeness effects, correlations with other predictors, or deficiencies in the aerosol representation within the CRS modelling framework. The analysis now reads:
"Aerosol, surface, and gas variables form the next group of influential predictors. Aerosol species AODAM and AODNI show substantial contributions but no consistent SHAP–feature pattern. This finding is interesting because these two inorganic aerosol species, ammonium and nitrate, were recently added to CRS. Their elevated importance suggests that ammonium- and nitrate-related aerosol information is associated with the analysed residuals. However, SHAP values alone do not identify the underlying physical causes of the residuals. Possible explanations include uncertainties in the aerosol retrievals, representativeness errors, correlations with other predictors, or deficiencies in the aerosol representation within the CRS modelling framework for example in cases of polluted environments."

"Lines 259-260: The conclusion that aerosol-related improvements are particularly important under low-SZA conditions should be interpreted with caution. As noted above, the lowest SZA bins contain substantially fewer observations and originate from a limited subset of stations."

We agree that the interpretation of the lowest SZA bins requires additional caution because these bins contain substantially fewer observations than the higher SZA bins and originate from a more limited subset of stations and geographical regions. The increase in the relative importance of aerosol-related predictors under low SZA conditions is associated with greater statistical uncertainty and may not be fully representative of the entire dataset. We have revised the manuscript to explicitly acknowledge this limitation and to avoid too strong conclusions regarding the role of aerosol-related improvements under low SZA conditions.

"Lines 265–266 "The lack of clear SHAP–feature dependencies for aerosols and surface parameters high lights the importance of accurately representing their interactions with cloud fields and scattering geometry.": The conclusion that improvements in aerosol optical properties would provide substantial benefits for reducing DHI errors is not entirely clear from the presented SHAP analysis. Earlier in the section, the manuscript notes that AODNI and AODAM do not exhibit a consistent SHAP–feature relationship. If the observed importance of these variables primarily reflects interactions with cloud fields and scattering geometry, as suggested in the following sentence, please clarify why deficiencies in the aerosol optical properties themselves are considered the most likely explanation."

Thank you for this important observation. We agree that the original wording was too strong in linking the observed aerosol-related predictors' importance directly to deficiencies in aerosol optical properties. As noted by the referee, the SHAP analysis identifies associations between the predictors and the machine-learning-estimated residuals, but does not uniquely identify the underlying physical causes. Also the lack of a consistent SHAP–feature relationship for AODNI and AODAM does not provide sufficient evidence to conclude that inaccuracies in aerosol optical properties are the primary source of the observed residuals.

We have revised the manuscript to adopt a more cautious interpretation. The discussion now emphasizes that the importance of aerosol-related variables may reflect a combination of factors, including aerosol retrieval uncertainties, interactions between aerosols and cloud fields, scattering geometry effects, representativeness errors, correlations with other predictors, and potential deficiencies in the aerosol representation within the CRS modelling framework. Direct attribution to aerosol optical properties has been removed.

"Lines 317–319: The importance of the BRDF-related parameters for the BNI residual is somewhat unexpected given the weak direct influence of surface reflectance on the direct beam component. Please further discuss the physical interpretation of this result."

We agree that the importance of BRDF-related parameters in the BNI analysis is not immediately intuitive given the relatively weak direct influence of surface reflectance on the direct beam component. We therefore expanded the discussion in the manuscript.

We note that SHAP values identify predictors associated with the machine-learning-estimated residuals rather than direct physical causes of the errors. Therefore, the importance of the BRDF parameters does not necessarily imply a strong direct radiative influence on BNI. Instead, these variables may, for example, act as indicators of surface and environmental conditions associated with larger BNI residuals. For example, surface reflectance characteristics are linked to surface type, cloud retrieval performance, aerosol conditions, and viewing geometry, all of which may influence the accuracy of the CRS estimates. We have revised the manuscript accordingly and now interpret the BRDF-related importance more cautiously as an indirect association rather than a direct effect on the direct beam component.

"Lines 323–325: Given the extremely low AODs reported for this case (e.g. AODAM = 0.006 and AODSS = 0.006), please discuss the physical significance of

the aerosol-related SHAP contributions. In particular, it is unclear whether these aerosol loadings are sufficiently large to produce meaningful radiative effects, or whether the associated uncertainties may be of the same order as the reported values."

We agree that the original interpretation was possibly too strong given the very low aerosol optical depths in the selected example. At such low aerosol loadings, the direct radiative impact of the aerosol components is expected to be small. Consequently, the SHAP contributions may not be interpreted as direct evidence that the CRS underestimates aerosol extinction.

We have therefore revised the manuscript with a more cautious interpretation. The discussion now emphasizes that, despite the low aerosol optical depths, aerosol-related variables remain important predictors within the machine-learning model. Their SHAP contributions may reflect interactions with other predictors, retrieval errors, representativeness effects, or broader environmental conditions associated with the analysed residual rather than a direct aerosol radiative effect.

"Section 3.4: The selected case study corresponds to a heavily overcast scene (Cloud Optical Depth = 77.6, Cloud Coverage = 100%) with very low AODs. While this example effectively demonstrates the dominant role of cloud-related predictors, it provides limited insight into the interpretation of aerosol-related SHAP contributions. Consider including an additional case characterized by relatively high aerosol loading and relatively low cloud influence, which would better illustrate the role of aerosol-related predictors and complement the current cloud-dominated example."

We agree that the original case study, while effectively illustrating the dominant role of cloud-related predictors, provided limited insight into the interpretation of aerosol-related SHAP contributions because it corresponded to a heavily overcast scene with very low aerosol loading.

To address this comment, we have substantially revised Section 3.4 by extending it to include a second case study representing contrasting atmospheric conditions. The additional example corresponds to a cloud-free scene with a high aerosol optical depth (about 0.8), allowing the role of aerosol-related predictors to be examined under conditions where aerosol effects are expected to be much more pronounced. The figures have been updated accordingly to present both case studies side by side.

Together, the two examples illustrate how the SHAP framework provides physically interpretable explanations under two distinctly different atmospheric regimes: one dominated by cloud effects and one characterized by elevated aerosol loading.

"Lines 346–360: While substantial improvements are reported for DHI and BNI, the improvements for GHI appear relatively modest (RMSE reduction from 105.3 to 103.0 W m^{-2} and R^2 increase from 0.873 to 0.879). Is there a physical explanation ?"

The more modest improvement observed for GHI compared with DHI and BNI likely reflects the composite nature of GHI, which consists of both direct and diffuse irradiance contributions. In practice, errors in the direct and diffuse components may partially compensate each other in the original CRS product, resulting in a relatively smaller net error in GHI. Consequently, even when the machine-learning correction substantially improves the individual DHI and BNI estimates, the resulting improvement in GHI may be more limited.

Furthermore, GHI is the primary quantity provided and routinely validated by the CRS and therefore already exhibits relatively high baseline performance. As a result, there is less room for improvement compared with DHI and BNI. We have added discussion of these factors to the manuscript.

"Lines 363–364: The statement that the XGBoost-based error model can "effectively learn and correct systematic deficiencies in the CRS radiation products" appears somewhat strong. Since the residuals may also contain contributions from measurement uncertainties and representativeness errors."

As discussed elsewhere in the revised manuscript, the analysed residuals contain not only deficiencies in the CRS modelling framework, but also contributions, for example, from measurement and retrieval errors, collocation errors, and representativeness effects. We have therefore revised the text to avoid implying that the machine-learning model exclusively learns and corrects deficiencies in the CRS radiation products. Instead, the revised wording emphasizes that the model learns and corrects systematic patterns present in the overall residual between the CRS estimates and the ground-based observations.

"Lines 373–374: The statement that the feature importance rankings are robust across solar elevations should be interpreted with caution, particularly for the lowest SZA bins, which contain substantially fewer observations and represent a geographically limited subset of the dataset."

We agree that as the lowest SZA bins contain substantially fewer observations than the higher SZA bins and originate from a more limited geographical subset of the dataset, the conclusions regarding the robustness of the feature importance rankings across all solar elevations should be interpreted with some caution. We have revised the manuscript accordingly and now explicitly acknowledge this limitation when discussing the SZA binned results.