GENERAL COMMETS:

1. I think this manuscript addresses a very important point, the pitfalls and drawbacks of AI for the retrieval of SSI. For example, the analysis of the results at the Mediterranean stations presented in this manuscript illustrates the challenges associated with machine learning. The net is only able to learn from local relations between reflection, surface albedo, atmosphere and SSI. It knows nothing about physics. Hence, it has to be expected that the performance decreases significantly if it is applied in regions with quite different conditions concerning aerosol load, cloud types, H20 and surface albedo. Within this scope it has to be taken into account that in many regions almost no in-situ data are available for the training or retraining. Hence, no learning of regional relations is possible then, but physical retrieval methods do not have these problems and it would be interesting to see the results between the current network with CAMS in Africa. Hence, the question why AI is needed for the retrieval of SSI should be addressed in more detail in the manuscript..

   The results presented in this work strongly suggest that such a model is NOT applicable to regions with less dense measurement networks. We had already suggested this in the conclusion ("In many regions, good quality ground measurements are too scarce for this model to be useful."). To make it clearer, we added the following:

   > "Therefore, while the ML model tested in this work could easily be adapted to be used operationally in France, it is unlikely that it can be extended to most other regions of the globe."

   Further, it is difficult to know what the net has really learned (black box approach), If we do not know what the net learns, we can't learn either (and our intelligence might expire on the long run). This should be discussed in more detail as well, based on the results presented in the manuscript. These points are partly addressed in the conclusion (e.g. L370ff) but should be discussed in more detail.

   We fully agree with the remark about the "black box" approach. We have added a new paragraph to the conclusion:

   > "Finally, we must remember that machine learning models are often opaque, making it difficult to understand how they make their predictions. This means that it is unlikely, at least in the short term, that we will be able to derive new physics from these models. If we focus only on machine learning, we may limit our understanding of the world around us. We therefore firmly believe that the research community should continue to invest in the development and improvement of physical retrieval models."

2. Poor performance of AI could also result from wrong training or training architecture. However, the comparison with the established CAMS shows that the training has been done well (5.1.1). This is very good, because it shows that the discussed pitfalls are not due to failures in the training or the chosen training method.

3. Figure 5: It seems to me that the main benefit of the machine learning is that it corrects differences in SSI induced by the different viewing geometries of ground based and satellite observations. We are aware of this effect, as significant differences are apparent when SSI retrieved from Meteosat East is compared with those from Meteosat prime for the same regions. So far these effects are not considered in many physical methods e.g. in CAMS, but it might be possible to implement appropriate “slant column” geometry corrections, which would increase the comparability of ground-based and satellite-based SSI. Please discuss this issue.

   It is possible that the better performance of the ML model (in training setup 1) stems from its ability to handle different viewing geometry. However, our results cannot confirm or infirm this hypothesis. Because we use a simple neural network, however, it is in our opinion unlikely that the model is able to correct for e.g. the parallax effect.

   A more thorough investigation would be necessary to determine how sun and satellite geometry is handled by the network. We believe that this would be a very interesting topic for future work. Such a study would also need to look into recent improvements in physical models that also account for viewing and solar angle.

4. Please consider that other physical retrieval methods might perform better or worse than CAMS, hence that the network might have a lower/higher performance when compared with other models. Please mention this briefly.

   We mentioned this in section 2.3:

   > “It should be noted that other physical retrieval methods might outperform CAMS (Forstinger et al., 2023). It remains, nonetheless, a state-of-the-art retrieval model.”

5. 5.2.2 Impact of aerosols: This is not a really a fair analysis, AOD (and H20) have not been given at predictors for the learning, hence the network could not learn anything about the relation of AOD variations and SSI, SAT reflection. It can just learn locally some kind of mean clear sky state. Contrarily AOD is used in CAMS as “predictor”. Please mention that the performance of ML might be better if AOD data were used as predictor in addition. Of course, it is not easy to find an accurate AOD raster data set, but this problem concerns AI as well as phyical

methods. Further, here, AOD from Aeronet is used, which is not available for CAMS elsewhere. Hence the capability of CAMS (or any other sat retrieval) concerning AOD variations is probably much lower as in the example. This should be also mentioned.

We added a discussion to the section 5.2.2 impact of aerosols:

> “This result is somewhat expected, as CAMS model integrates some information about the AOD (through McClear), whereas ML model does not. Adding AOD-related predictors to the neural network may help decrease the performance gap between the two methods for clear-skies.”

Regarding the fact that CAMS uses AOD raster data and not actual measurements, we added a footnote: “The remaining correlation may come from the fact that CAMS uses modeled AOD, that sometimes deviate from the truth.”

6. Please add more information about the in-situ data. Do they all have the same maintenance, calibrations cycles and so on. Hence can the same accuracy be expected for all pyranometers ?

   Unfortunately, we only have access to limited information about this. The measurement stations used in this work are all operated by Météo-France which performs regular checks and calibration. However, to the best of our knowledge, there is no synthetic information describing the schedule of these procedures.

   This was a motivation for the thorough and conservative QC applied to these stations, summarized in Appendix and thoroughly described in https://doi.org/10.1016/j.solener.2023.04.037

7. Throughout the manuscript. Please avoid the separation between physical methods and clear sky index methods. They are physical methods as well !

   Done!

DETAILED COMMENTS:

1. Abstract: “the first of which is likely solar energy”. This depends on the viewpoint. Please delete “the first of” and rephrase accordingly, it is also quite important for climate, tourism,...

   This was removed.

2. Abstract: "For long, the emphasis has been on empirical models (simple parameterization linking the reflectance to the clear-sky index) and on physical models” The use of the clear sky index follows also physical laws, hence please rephrase. Please see also the general comments.

We changed to:

> "For long, the emphasis has been on models grounded in physical laws with, in some cases, simple statistical parametrizations."

3. L25 "...index methods without explicit physical cloud models", L29 "empirical" please see 2.) and general comments . The use of the clear sky index follows also physical laws and the cloud index is a measure for the cloud transmission, thus, not without physical cloud model, please rephrase

   We changed to:

   > "from the earlier cloud index methods (Cano et al., 1986; Rigollier and Wald, 1998) to more recent approaches relying on advanced radiative transfer models (Xie et al., 2016; Qu et al., 2017). »

4. Line 55 "z. 4 by 5 km". it might be closer to 3.2x5.5 please check.

   There was a mistake. According to "MSG Level 1.5 Image Data Format Description (figures 10 and 11)", it is actually ca N-S 6 km and E-W 4 km. We updated the text:

   > "MSG channels have a temporal resolution of 15 minutes and a spatial resolution of 3 km at Nadir (0,0)[1], which above France corresponds to pixels of ca. 4 by 6 km (in the E-W and N-S directions, respectively) (EUMETSAT, 2017)"

5. L104: "ML model must be fully online" Please explain why ?

   This is for the comparison with CAMS to be fair – although it is currently only available after a certain delay, CAMS only uses past and present data to deliver estimations. We change 'online' to 'real-time', to make it clearer clearer.

6. L 195 "Three tricks are applied:" Please use a more appropriate term instead of tricks.

   We changed to "Three techniques are further applied:"

7. L 370 please consider to add surface albedo here

   We already mention albedo at the end of 6.2 – we intentionally left it out of the first sentence (formerly L370), because as far as we know, it is only one of the factors that impact generalization. We, however, added a sentence to insist on the need to understand which factors impact generalization:

   > "Understanding the factors that describe the similarity between two locations should be an important aspect of future research."

8. L 385: Another option is to improve the physical methods, without AI e.g. as demonstrated by HelioMon. The accuracy of HelioMont is already close to that of BSRN stations, why fuss with AI ? Please consider to add this option to the manuscript. In the Alps it is questionable if any network would be able to learn the complex relations for all regions, because taking the spatial heterogeneity into account there are not enough ground stations.

   We agree that the community should keep increasing physical models. In response to your first comment, we added a paragraph in that sense in section 6.3:

   > "Finally, we must remember that machine learning models are often opaque, making it difficult to understand how they make their predictions. This means that it is unlikely, at least in the short term, that we will be able to derive new physics from these models. If we focus only on machine learning, we may limit our understanding of the world around us. We, therefore, believe that the research community should continue to invest in the development and improvement of physical retrieval models."