This is a very interesting paper that deals with limitations and perspectives for the calculation of surface solar irradiance (SSI) using machine learning techniques.

The paper deals with an aspect including a number of more or less .. easy to explain, sources of errors and uncertainties. The work is high level and ends up in a publication with unique in my opinion results worth being published in AMT.

Some comments towards manuscript improvement

Abstract

At the moment the abstract is a bit like a general discussion and some metrics there, especially summarized comparisons of ML and CAMS, could be useful for a reader that will be intrigued to read more about it.

We added several metrics from the result section in the abstract:

> "We found that the data-driven model's performance is very dependent on the training set.
>
> Provided the training set is sufficiently large and similar enough to the test set, even a simple MLP has a root mean square error (RMSE) that is 19% lower than CAMS and outperforms the physical retrieval model in 96% of the test stations.
>
> On the other hand, in certain configurations, the data-driven model can dramatically underperform even in stations located close to the training set: when geographical separation was enforced between the training and test set, the MLP-based model exhibited an RMSE that was 50% to 100% higher than that of CAMS in several locations."

Introduction

What I am missing is some basic state of the art of current datasets (including CAMS) and their performance evaluation.

We updated the first paragraph of the introduction to discuss this aspect:

> "Some of these retrieval algorithms are operational and provide SSI estimations worldwide. For example, HelioClim3 (Blanc et al., 2011a) offers real-time estimations of the Global Horizontal Irradiance (GHI) over Africa and Europe. CAMS, the Copernicus Atmosphere Monitoring Service, is another near real-time service that derives SSI estimations from data collected by both Meteosat and Himawari satellites; it covers areas including Africa, Europe, and

a significant portion of Asia (Schroedter-Homscheidt et al., 2016). In the United States, the National Solar Radiation Database (NSRDB, Sengupta et al. (2018)) serves as a valuable resource, providing SSI estimates primarily from the GOES satellites. The performances of these solar radiation databases vary with the location and sky conditions; they are discussed in detail in Forstinger et al. (2023)."

Data

AERONET does not provide AOD every minute and also in cloudy days, so some clarification could be included as a short paragraph in 2.2 e.g. how AERONET data used , which wavelength for aerosol optical depth used etc.

We have updated the paragraph describing AERONET measurements:

> "As an AERONET station, it provides measurements of spectral aerosol optical depth (AOD). The AOD at different wavelengths are measured with a Sun photometer but are only valid under clear-sky conditions. Cloud screening is thus applied to the raw data and measurements are therefore only available intermittently (Giles et al., 2019). In this work, we use the AOD at 500 nm."

Section 3

It is impressive the choice of using 3 hours (12 instants) as a basic hierarchy of the method. Could you explain how this choice has been decided? isn't it 3 hours relatively .. long ?

3 hours was chosen intentionally a bit long because the MLP algorithm should be able to handle redundant or irrelevant inputs. It is indeed questionable whether 12 past instants are really needed for this algorithm; similarly, it is possible that the algorithm could benefit from a larger spatial neighborhood. The two aspects are actually very likely linked.

A rigorous ablation study could help clarify this point and, as a matter of fact, several other aspects of the algorithm could be further optimized. But we think that this would be the focus of a different paper.

The 3 set ups could be of course more complex but I personally find the choice really appropriate here.

I am a bit puzzled by the fact that the kc=1 limitation of CAMS does not have a more visual impact on the statistics. Or is it a major factor of the ML better performance ?

It is very likely not a major factor of the ML better performance as, on the contrary, we see in Figure 5 that the MLs model does not perform that well for kc>0.9.

The reason why it does not have a more visual impact on statistics (and may seem somewhat contradictory with the clear-sky results discussed later) is that there are only a few instants for which the measured clear-sky index is above 1.

We added a sentence in section 5.1.1 to remind the reader of this:

> "Admittedly, this only concerns a small portion of all instants, and, in addition, ML model tends to produce too many estimations with high clear-sky index."

Figure 5: based on the definition given in lines 80 – 85 and the aerosol issues discussed after fig. 5 there should be clear sky index higher than 1 not visible in the figure.

This was indeed a mistake in the labeling of the bins. The last kc bin is 'open'. We have updated the plot to reflect that.

Aerosols: It is clear that the ML inputs does not include any aerosol information so figure 7 is more or less expected. A very rough predictor including an aerosol climatology (more in summer less in winter) would for sure improve this negative correlation shown in fig. 7 . Especially because this has an impact on "high solar irradiance" cases.

Indeed this result could be expected, and it is likely that adding aerosol-related data to the ML model predictor would improve its performance. We added a discussion to Section 5.2.2:

> "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."

Fig. 7 needs a bit more explanation as it is not clear if the points are based on instants, hourly or daily values.

We updated the text and the Figure caption:

> "To further investigate the role of information about AOD at 500 nm in ML model under-performing for clear-sky days, we analyze the relationship between the hourly estimation error and the corresponding hourly AOD average under clear-sky conditions. This relationship is illustrated in Figure 7, which shows the distribution of the error of each retrieval model as a function of AOD 500;"

> "Figure 7. Joint distribution (2D histogram) of hourly average AOD and hourly estimation error for CAMS (a) and ML model (b). Spearman's Rank-Order Correlation between AOD and error is also given."

I find difficult to understand how the ML can outperform CAMS for clear skies in the related bins of fig. 5 and still have these aerosol related aspects shown in fig. 7.

That is because the right-most bin of kc in figure 5 does not contain only clear-sky situation. These are hourly values of kc, so many points likely see a mix of partially cloudy and clear sky. This shortcoming of the kc binning was our main motivation to select 'true clear sky instants' in Carpentras.

We have updated the text at the beginning of section 5.2 to insist on this point:

> "In this section, we focus on the performance of the two retrieval models under clear sky conditions. To clearly identify such conditions, however, the analysis done in Figure 5 is not sufficient: all clear sky situations should be contained in the right-most kc bin ([0.9 − [), but other situations (typically a mix of overshooting, clear sky and partially cloudy) are likely also contained in this bin. To rigorously select clear-sky conditions, we need 1-minute irradiance data (Section 4.3); we thus focus on the Carpentras station (Section 2.2)."

Maybe the authors could discuss:

In general it is understandable that the paper does not introduce a method to be used in different areas but it is a kind of sensitivity study on the ML performance. For this case a really unique dataset is used with a huge number of stations. However, it would be nice to comment on perspectives of an actual application of such system. Indirectly this study can assess some kind of realistic cases of limited or not, ground-based data available that can be used for applying such methods in different areas.

This was already a bit discussed in section 6.2:

> "In many regions, good quality ground measurements are too scarce for this model to be useful."

We agree that it is an important point and 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."

The whole France and so many stations is a huge area, but still could be very different than another area with different cloud/aerosol conditions which the same results with the same number of stations and analysis can vary. E.g. aerosol (not captured) effects in N. Africa will have a crucial effect on the statistics as well as areas with different and more clouds.

We have updated section 6.1 to mention the fact that AOD are more impactful in other regions:

> “This only slightly impacts the performance of the ML model in France, where the effect of AOD on SSI is relatively small, but in other regions – for example deserts (Eissa et al., 2015) – the ML model may underperform.”

Finally I can say that problems such as the spatial (difference of point/station to grid/satellite) and the temporal (15 min satellite frequency vs 1 minute measurements integrated to hourly), seem to somehow dealt in a nice way with the ML training.

Minor

“Note that, since night-time is flagged as failing QC, 30% is a high requirement”, I don’t understand this maybe you could clarify.

We have rephrased this part; we hope it is clearer:

> “In the first setup, 100 test stations are chosen randomly from those passing QC for more than 30% of the hours over the test period (2018-07-01 to 2019-06-30). In other words, QC must be passed for at least 8 hours per day on average. As night-time is always flagged as failing QC, this is a stringent requirement.”