comments:

The machine learning based techniques are very popular in recent years, and these methods do provide very good performance in various fields. And people start to question is it possilble for ML to replace the traditional physical method.

I think the authors tried to try to give some explanation from some extend. In this paper, the authors used physical method and ML based method to infer SSI from satellite images. And the authors gave an overall exploration of the MLP and analysed pitfalls and drawbacks of the method. It is very interesting that the result from MLP is better than CAMS from some aspects.

This paper is of high standard, I have some questions about some points, and hope the authors can consider.

1. There are many machine learning methods, and MLP is just a very basic ML method, why you chose this method? From the introduction part, I can't see some information about how MLP is used in the past researches, how it works in this field? Of course, this paper explained the performance of AI and physical method, but MLP can't represent AI techinque, could you explain why don't you use some other state-of-art AI techniques? I think some more information should be given in introduction part as well.

We did test some deeper architectures (2 and 3 hidden layers, some with more hidden neurons), but they didn't perform better. Following the principle of parsimony, we thus kept the simpler model (this is mentioned in section 3.3). Regarding other, more complex architectures, we were also guided by the principle of parsimony: if an MLP can do the job (and we consider it does since it already outperforms CAMS in training setup 1), there is no need to deploy a more complex model.

In addition, the dimension of the chosen input set did restrict a bit the list of adapted architectures: because the spatial extent of the input is only 3x3, using a CNN architecture would not make sense: the smallest size of a CNN kernel is usually 3x3. (Admittedly, we could have tested a RNN model to handle the 12 time steps in input.)

Most importantly, the goal of this work is to underline ML-specific validation issues, notably linked to generalization. More complex models (CNN, RNN, LSTM, Transformers, GAN, etc.) are trained the same way as MLP and therefore suffer from similar issues when it comes to generalization or out-of-domain performance. It is even likely that more complex models are more subject to overfitting and thus bad generalization (see for example *Generalizability of Neural Network-based Identification of PV in Aerial Images*, Ranalli 2023 for a related discussion).

We have explained this choice in the introduction:

> “Our objective is not to introduce a new retrieval method; hence, we have deliberately opted for a simple, fully connected architecture. This choice allows our conclusions regarding generalization to extend more effectively to the realm of complex networks (convolutional, recurrent, attention-based, generative, etc.), which are generally prone to encountering greater generalization challenges (Wang et al., 2017; Ranalli and Zech, 2023)”

We also agree that investigating more complex architecture is of interest. We have thus updated the perspective section (6.3):

> “The investigation of more sophisticated neural network architectures is also of interest and would become particularly relevant when dealing with larger input datasets. Architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), or Spatio-Temoral Transformers hold promise, especially when a broader context in both time and space is required. However, it is important to recognize that such complexity may raise the risk of generalization issues, as more complex models are generally more likely to overfit.”

2. I think another index should also be considered, that is efficiency. What is the running time of MLP and CAMS respectively? This is also a very important index to see the performance of the two methods.

That is a good point. We’ve added a subsection about it: “3.4 running time” (see the updated manuscript

3. According to 5.3.3, as to ML methods, input training dataset plays very important part in the result, it should include all the necessary information it needs. So that is why a correlation analysis between input variable and target variable is necessary.

If we were doing in-depth feature engineering, we agree that a correlation analysis between potential predictors (or features) would be the first necessary step. Here, however, we opt to keep the ML model as simple as possible and to only use data from MSG (the other predictors are calendar-derived).

Just as the authors have analysis, in some locations, ML method shows poor performance because it doesn’t have direct access to that information, is it because satellite can’t cover that area which lead to this problem?

All the areas tested are well covered by the satellite, so it shouldn’t explain bad performance. We rather think that this is due to these stations being out of the training domain.