Authors' response to the third anonymous referee

We would like to thank the third referee for his review and for sharing all his relevant concerns regarding the neural network design and training. We will address these concerns in due course. This will help us improve the Methods section, which requires more detail.

(1) Feature selection: They used a correlation analysis and SHAP to select the features for the neural network. However, it is unclear to me whether the considered variables are all observed or simulated by the land surface model or mixed. This clarification is important. In my view, the operator should ideally be trained on observational data, as it is intended to capture the observation-derived relationships between the input features and SIF. If, instead, these relationships are learned from model simulations, any inherent biases in the model may be embedded in the operator. As a result, during assimilation, biased model outputs fed into the operator may already be implicitly corrected by the learned relationships. Consequently, the mismatch between the operator output and the observations would be artificially reduced, undermining the effectiveness of the assimilation process. Furthermore, the SHAP results show that surface soil moisture, GPP, and ET have very limited importance, while only LAI and several spatial and temporal positional features are important. It is hard to believe that SIF only depends on the structural variable, i.e., LAI, in addition to the positional features. SIF should at least correlate with some functional variables, e.g., GPP and/or soil moisture, which is partly shown by the correlation analysis. Would this be because SHAP can not consider the inter-dependence between variables? Therefore, the roles of GPP, etc., are hidden by LAI. A conditional SHAP could be used in this case. In addition, the model on which SHAP is based is not reported.

RESPONSE 3.01:

We agree with the referee that section 2.2.3 requires more information and could be clearer. In this study, we opted for full observation-driven training. We realised that L. 190-200 could be confusing. In this sentence, only LAI V1 and TROPOMI SIF correspond to observations. This section will be clarified by moving the description of preliminary tests (including Appendix A2) in a Supplement. We will focus on the actual configuration used in the paper.

The resulting SHAP values for gross primary production (GPP) and soil moisture are low. We used a simple Gaussian kernel, which assumes independence between the inputs. Using a conditional kernel would be more accurate, likely redistributing the importance of LAI towards GPP and soil moisture. However, training was performed both with and without GPP, and there was no significant change.

(2) Temporal scale of SIF: The daily SIF is used in this study, but the correlation between SIF and GPP is usually higher in a coarser temporal resolution, e.g., 8-daily. If there is no strong reason that the assimilation must be conducted at the daily scale, I would suggest using a coarser temporal resolution, which also avoids the time interpolation of LAI.

RESPONSE 3.02: The Copernicus Data Space Ecosystem's operational product is the daily TROPOMI SIF product. Using the 8-day SIF product would likely improve the correlation of SIF with GPP, and potentially with LAI too. However, this would mean losing one of the daily product's biggest strengths: its frequency. Furthermore, the difference in availability frequency between LAI-

V1 (or LAI300 from CLMS) and the 8-day SIF product would require temporal interpolation to align the dates for NN training. Therefore, using a coarser temporal resolution would not eliminate the need for linear interpolation.

(3) The training strategy: During the training, only 40% of the one-year data is used as the training set, while 60% is used for validation. This is a very unusual training-validation split, and I have never seen a split where training is less than 60%. Further, the data from 1 June 2019 to 31 May 2020 is used as the test set if this is not a typo in the text, and then it seems that the data from 1 June 2020 to 31 May 2021 is not used at all, which is strange as well. Overall, the training data is too little compared to the data for validation, test, and unused. A clarification is needed for this choice. Otherwise, significantly more data should be used for training to be expected to improve the model performance.

RESPONSE 3.03: The referee is right in saying that our data splitting is unusual. The amount of data used to train the neural network is already significant, and even 40% of it involves several millions of input/output pairs. For given batch size and loss, we have evaluated our data splitting choice by trying several partitions, from 40 % to 80 %. The resulting learning curves are displayed in FIGURE R3.1. Both the training and validation losses decrease with the amount of data used for training, but ultimately, the RMSE on the validation loss on SIF remains quite similar. The choice of data splitting does not generate many differences, and the 40 % amount of data used for training was sufficient to produce a model that generalise quite well.

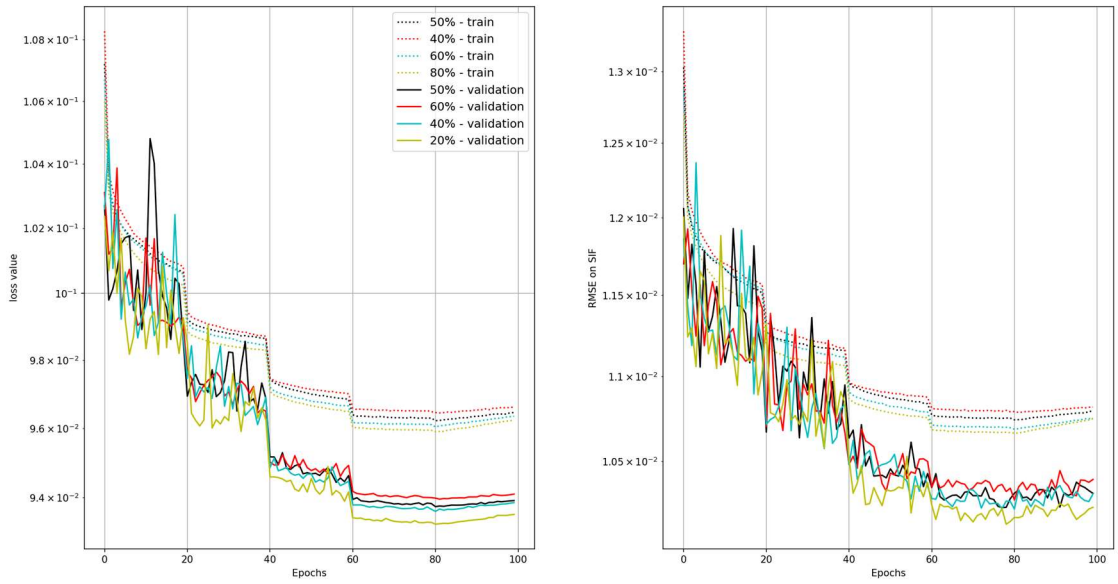

FIGURE R3.1: Learning curves for different partitions between training and validation

(4) I found parts of the manuscript somewhat difficult to follow, partly due to issues with flow and the brevity of the figure and table captions. Improving these aspects would likely enhance the overall readability.

RESPONSE 3.04: We will provide more details in the caption, thank you for your general suggestion.

(5) Ki in equations 1b and 3a are different with or without bold, which I guess represent the original Kalman filter and the extended Kalman filter, respectively. But this is not clear in the

text.

RESPONSE 3.05: These equations require a rewrite as the first referee pointed out, we will clarify this aspect.

(6) line 149: I did not understand what is the reason of 'that is why'. Is the linear assumption not correct?

RESPONSE 3.06: We propose replacing:
"That is why, in our case we use an extend Kalman filter by using the tangeant linear of the operator"
by
"We use an extend Kalman filter by using the tangent linear of the operator"

(7) -lines 160-161: I did not understand this sentence. What do you mean by 'assimilation is assessed'? What is observation space?

RESPONSE 3.07: We propose to replace:
"Assimilation is assessed when the difference in observation space between the observation and the background simulation exceeds the final departure in the observation space between the observation and the analysis simulation."
by
"Since the purpose of assimilation is to bring the model closer to the observations, the difference between the analysis and the observations in the observation space should be smaller than the difference between the prior simulation and the observations."

By 'state space' and 'observation space', we mean the mathematical spaces defined by the control and observed variables, respectively. If the observed variable is among the control variables (e.g. LAI), we are working in the same space. If not, the observed space is linked to the state space by H.

(8) line 165: unclear to me, probably because of the grammar.

RESPONSE 3.08: We propose replacing:
"For training and evaluation purposes, all areas covered by any fraction of ocean, water, snow or lakes are removed, as are all areas above 1,500 metres above sea level and areas where bare soil, rock, snow or ice are most prevalent according to ECOCLIMAP-II."
by
"For training and evaluation purposes, all areas covered by ocean, water, snow or lakes, or lying above 1,500 metres above sea level, or where bare soil, rock, snow or ice are most prevalent according to ECOCLIMAP-II, are removed."
(9) Line 238: What is the open-loop simulation? Is it the baseline?

RESPONSE 3.09: The confusion arises from the fact that these sentences are not in the correct order. We will rephrase this section. In an open-loop simulation, there is no data assimilation. The baseline simulation involves assimilating LAI-V1 into the model. The purpose of the open-loop simulation is to evaluate whether the SIF assimilation process leads to an analysis that differs from the basic model output. The baseline is used to evaluate whether assimilating SIF leads to an analysis that is close to the state of the art.

(10) Table 1's caption should include a detailed description of the different experiments, which makes it clear when looking at it alone without checking the description in the main text.

RESPONSE 3.10: To clarify, we will detail the assimilation periods by adding a dedicated column in Table 1. We will better distinguish the two LAI columns and indicate the observation errors used for LAI. We will complete the caption by mentioned the Ebro basin domain.

(11) Table 2: Why are the metrics of the training dataset shown? Not validation. Also, the limited overfitting evidence is claimed based on the train and test datasets. Although limited overfitting is true in this case, this is usually demonstrated by training and validation loss.

RESPONSE 3.11: We consider the performance of the neural network over the course of a whole year and compare it with the previous year. The learning overfitting is checked in Annex A3 (Figure A3). We will move this Annex to a supplement and add more details.

(12) Figure 6: It is unclear to me what is shown here and why it is shown. What are increments?

RESPONSE 3.12: The analysis increment is the difference between the analysis and the prior terms. This correction is added to the model simulation, which starts from the latest analysed state, to create the new analysed state. This will be explained in more detail in the data assimilation process description.

(13) lines 295-296: unclear what the relation is between the data availability and p-value.

RESPONSE 3.13: We propose replacing:
"The resulting map is displayed in Fig. 7. As only one observation is available every 10 days for calculating the correlation, all cells with a p-value higher than 0.01 were removed."
by
"The resulting map is displayed in Fig. 7. As only one observation is available every 10 days for calculating the correlation, all cells with a F-test p-value higher than 0.01 were removed. This mask is used for both the correlation and RMSE panels. This ensures that significant correlation and RMSE values are displayed."

(14) Figures 7-10: should mention that the domain is the Ebro basin in the captions.

RESPONSE 3.14: We will mention it in the caption, thank you for your suggestion.

(15) Figure 11: For each day, does the boxplot represent the values of many pixels from the in-situ measurements?

RESPONSE 3.15: We propose replacing the caption of Fig. 11 by:
"Comparison of the OL and SIF20 experiments with the SIF airborne measurements of the HyPlant instrument within the LIAISE campaign. Boxes represents the 25th and the 75th percentiles, the whiskers the 5th and 95th percentiles of the airborne SIF observations. The central horizontal red lines represent the 50th percentile and the green triangles the mean of the airborne SIF observations."

(16) line 370: What is the update rate?

RESPONSE 3.16: We propose to replace:
"However, it should be noted that the update rate yielded by TROPOMI SIF assimilation is five times higher than that of 10-day synthesis assimilation."
by
"However, it should be noted that daily TROPOMI SIF assimilation yields a higher update rate than assimilation of 10-day LAI observations."

(17) line 9: could mention neural network here already

RESPONSE 3.17: Thank you for your suggestion, we will add "neural network" in the Abstract.

(18) lines 96-97: could also mention 0.1 resolution here to guide readers.

RESPONSE 3.18: Thank you for your suggestion, we will replace "regular grid" by "0.1° resolution regular grid".

(19) line 109: 'interpolated'–>'aggregated'?

RESPONSE 3.19: We will change this to "aggregated".

(20) line 116: If the FLUXCOM remains at its original resolution, how was the training test run conducted?

RESPONSE 3.20: FLUXCOM was not used in the training process. The GPP used was taken from a pre-existing analysis over the European domain.

(21) line 183: 'maximum'. You mean peak?

RESPONSE 3.21: Yes this is the peak of the distribution.

(22) line 187: I did not understand how this function keeps the prediction positive. Would ln(SIF) lead to SIF<1 negative?

RESPONSE 3.22: SIF_l can take negative values when the SIF is less than one. What matter here is that if we apply the inverse function of SIFl we get only positive SIF values.
The inverse function writes :
SIF = Exp(SIFl) if SIFl ≤ 0, and SIF = ln(SIFl+1)+1 is SIFl > 0.

(23) lines 204-207: DOY provides the information on the seasonality but LAT and LON should help capture the spatial variability of SIF across the domain.

RESPONSE 3.23: We will complete the sentence, thank you for your suggestion:
"The metadata predictors provide information on the seasonal cycle and on the spatial variability."

(24) Figure 5: could display as a 2*2 figure to save the space. 'were'–>'where'

RESPONSE 3.24: Thank you for your suggestion, we will consider it to see if it fits better in 2x2 in two columns.

(25) Figure 6: could make use of the horizontal space more.

RESPONSE 3.25: This figure is supposed to take a single column, we will see if less margin can be taken.

(26) lines 321-322: For me, it looks like the co-assimilating SIF alongside LAI and assimilating only LAI-V1 have the similar RMSEs on LAI.

RESPONSE 3.26: In addition to Fig. 8, there is a missing reference to Fig. 10 here. In terms of RMSE, the two 'SL' configurations have lower mean values than the LAI-V1 assimilation experiment when compared with the same LAI-V1 observation product. The 25th and 75th percentile spreads are also tighter than when assimilating the LAI-V1. In the Ebro basin, co-assimilating LAI300 and TROPOMI SIF results in a lower RMSE than assimilating only LAI-V1 in the ISBA model.

(27) Figure 7: could change to a divergent color scheme that could differentiate positive and negative values.

RESPONSE 3.27: We will change the colormap to better distinguish the negatives values from the positives.

(28) Figure 8: Fluxcom should not be able to capture irrigation effects as it infers GPP largely based on meteorological conditions. Therefore, the additional water input by irrigation is likely not be incorporated. So FLUXCOM GPP cannot be treated as a ground truth here. Should reconsider the discussion of the relevant results.

RESPONSE 3.28: We will conduct a more thorough analysis of the data captured by FLUXCOM over this irrigated area. Nevertheless, its correlation with SIF and LAI observations suggests that FLUXCOM expects the vegetation to receive sufficient water during the dry period.

(29) line 339: The operator is not trained in this way. Why not obtain the SIF values for each pixel and then calculate the average value?

RESPONSE 3.29: The neural network is trained to take an ”average LAI” value over a model grid cell. During the assimilation, this average is given by the weighted summation per patch of the LAI simulated per vegetation type. In Fig. 11 we used the same approach. This section will be moved to a supplement.

(30) lines 344-345: I did not understand this. Isn’t it an emulator of TROPOMI SIF?

RESPONSE 3.30: This section is confusing because we are comparing the emulated SIF with the airborne SIF, which does not follow the same retrieval process. The key point is that the observation operator is not a surrogate model for airborne SIF measurements. Since these results are inconclusive, we will move this section to the supplementary material.