Authors' response to Editor and Referees' comments on the revised manuscript "Using deep learning to assimilate sun-induced fluorescence satellite observations in the ISBA land surface model"

Preprint egusphere-2026-941

Pierre Vanderbecken, Jasmin Vural, Oscar Rojas-Munoz, Sébastien Garrigues,
Bertrand Bonan, Cédric Bacour, Uwe Rascher, Bastian Siegmann,
Patricia de Rosnay and Jean-Christophe Calvet

(bold +italic = referee comment)

Authors' response to the Editor

We would like to thank the editor for reading the manuscript and for his valuable comments and suggestions. The following provides answers to the questions and details the improvements that were made to the manuscript.

(1) The review and discussion of this article may proceed. However, please note that the similarity report came back with a high amount of duplication in Section 2.2 with this article here: https://hess.copernicus.org/articles/22/3515/2018/. While we understand that there are going to be similar processes used in the data assimilation procedures, we ask the authors to use a more paraphrasing in this section.

RESPONSE: Thank you for this comment. Section 2.2 provides a description of the ISBA land surface model and the LDAS-Monde assimilation system. These are the same tools used in our article. We have rewritten the section to make the description of this model less similar to that in the 2018 article. Furthermore, this section has received numerous comments and suggestions from the referees, and the resulting modifications now distance the manuscript further from the 2018 article.

(2) Please ensure that the colour schemes used in your maps and charts allow readers with colour vision deficiencies to correctly interpret your findings. - Fig 12

RESPONSE: All figures were created with this specific goal in mind, using adapted colourblind colour maps from the "cmcrameri" Python package. Where necessary, we added markers or hashing to ensure everything could be interpreted, even when printed in black and white. The timeseries displayed in Figure 12 differ in terms of colour and line style (dots, dashes and solid lines, respectively). We acknowledge that the figure is stacked, which makes it difficult to read. We have modified the figure's shape to improve its readability.

Reviewers 1 and 3 notably raise concerns about the assimilation and pre-analysis feature selection clarity. Reviewer 3 also detailed concerns about how the training and validation was carried out, which I agree appears different from some conventions.

RESPONSE: In response to the reviewers' comments, we have simplified the introduction, completely revised the set of SEKF equations and clarified the NN observation operator section by moving the description of the preliminary tests (including Appendix A2) to a supplement. This

allows us to focus on the configuration used in the paper. Figure R3.1 (new Figure S5) shows that both training and validation losses decrease with the amount of data used for training. Ultimately, however, the RMSE of the validation loss on SIF remains quite similar. The choice of data splitting does not generate many differences, and using 40% of the data for training was sufficient to produce a model that generalises quite well.

-The figure 6 caption should provide more detail about what "Mean monthly analysis increments on the LAI" are.

RESPONSE: In the new version of Fig. 6 (now Fig. 5), we have added a caption indicating the name of the experiment and the domain under consideration. Increments are now defined in Eq. (1).

-The results around figure 8 discuss differences between assimilation scenarios. Please ensure statistical significance tests are provided as a part of the differences.

RESPONSE: We realised that the words 'significant' and 'significantly' were used too often in the manuscript. Alternatives were found. Where it meant 'statistically significant', Fisher's test p-values were indicated.

Authors' response to the first referee: M. Alban Farchi

We would like to acknowledge the meticulous review of our article done by M. Farchi, and we thank him for his relevant remarks and comments that have helped us to improve the article content further.

(1) On several occasions, the text alternated between two topics, which can be sometimes confusing:
• In the introduction, between physical-and machine learning aspects of your problem.
• L91-92 between the description of the retrieval process and a post-processing that you apply (discarding negatives values + averaging)
• L212-213 between description of results and conclusion
• L362-367+ between NN results on the test (RMSE and correlation) and the application of the NN.
I would advise to reorganise the text to avoid giving the reader the impression to go back-and-forth between topics, and also to take this opportunity to make the logical connections more visible. For example, it is only after reading the whole introduction that I understood the logical links between the various aspects of the problem.

RESPONSE 1.1: Thank you for these comments. We simplified the Introduction by removing L. 40-46. This ensures a better transition from SIF to assimilation.
About the specific lines mentioned:
• L92: we gathered all sentences related to the instrument together, as well as those related to the retrieval. Our own post-processing work is now summarised as:
"In this study, since our model domain is defined on a regular grid, the SIF observations were interpolated onto the model grid. Only all positive instantaneous measurements for a given day covering a grid cell in the domain are averaged to provide a value for that grid cell." It is now written on L. 90-93 of the revised manuscript

• L 212-213: We agree. Presenting results in the Methods section is confusing. We moved the intermediate tests in a Supplement and now focus on the final configuration that was used to produce the results. Namely, use of LAI observations as a predictor for training the NN, together with DOY, LAT, LON as metadata predictors. This is now discussed in the first section of the supplementary materials and a reference to it is placed on L. 189 of the revised manuscript.

• L362 -367+: We agree. We moved the following sentence at the beginning of the next paragraph: "The neural network was then successfully used …". It is now on L. 373 in the revised manuscript.

(2) Eq 1-3: at first, I thought that there was a typo in 1a (that the argument if the model should be the analysis and not the background). It is only after reading Eq. 2 that I understood that it is not. Overall, the presentation method is highly confusing, first because there is a temporal aspect, which means that the method should be called "smoother" and not "filter" (I can understand that, for historical reason, you would prefer call it a filter, but you should at least mention this point), and second because the temporal aspect is never really explicitly explained: is the control vector the state at t=i or t=i-1? From what I understand, the fact that M appears in the definition of the Kalman gain means that the control vector is the state at time=i-1, which means that there is a typo in 1b, it should be:

$x^a_{i-1} = x^b_{i-1} + K. . .$
or equivalently:
$x^a_i = M\, x^a_{i-1} = x^b_i + M\, K. . .$
in a case of a linear model.
Then, the sentence L 557-558 "The initial state is determined by the analysis performed over the previous 24-hour assimilation window." is confusing. The initial state for what? Or is this sentence only here to explain that the time step between t=i-1 and t=i is 24h?
Finally, there is no mention about cycling. Are you using cycled data assimilation or is the background reset at each assimilation cycle?

RESPONSE 1.2: The reviewer is right, there is a mistake in the SEKF equations. We replaced this set of equations by those published in the Supplement of Bonan et al. 2020. The section 3.2 starting on L. 130 in the revised manuscript was updated accordingly and presents equations equivalent to your second suggestion.

(3) The LAI acronym introduced multiple times (L30 & L37)

RESPONSE 1.3: Thank you for pointing this out, the last part of the introduction was removed. The leaf area index is now only introduced on L. 30 in the revised manuscript.

(4) L37 "SIF observations" later in the introduction, these are called "TROPOMI SIF retrieval" (L46-47, 50). Are you referring to the same object? Please try to be consistent with your nomenclature.

RESPONSE 1.4: We agree that it is confusing. We replaced:
"To enable more real-time and operational monitoring of vegetation, the 8-day TROPOMI SIF retrieval has been assimilated into the IFS land surface model developed at ECMWF at a global scale (Garrigues et al., 2026). In this study, we propose to constrain simulations from the ISBA (Interactions between Soil, Biomass and Atmosphere) land surface model, which is developed at Météo-France and available through the SURFEX (SURFace EXTernalisée) platform (Masson et al., 2013; Decharme et al., 2013), by assimilating the daily retrieval of TROPOMI SIF."
by
"To enable more real-time and operational monitoring of vegetation, Garrigues et al. (2026) have assimilated the 8-day SIF retrieval reprocessed from raw TROPOMI data has into the IFS land surface model developed at ECMWF at a global scale. In this study, we propose to constrain simulations from the ISBA (Interactions between Soil, Biomass and Atmosphere) land surface model, which is developed at Météo-France and available through the SURFEX (SURFace EXTernalisée) platform (Masson et al., 2013; Decharme et al., 2013), by assimilating the daily retrieval of TROPOMI SIF instead of the 8-day TROPOMI SIF product used by Garrigues et al. (2026). Unlike Garrigues et al. (2026), our land surface model simulates LAI." This is on L. 40 to 45 in the revised manuscript.

(5) L40 a transition is missing before the second paragraph.

RESPONSE 1.5: Thank you for pointing this out. We deleted the first 6 sentences of this paragraph ("Advances in … a large database ()"). This ensures a better transition to the work of Garrigues et al. (2026).

(6) L42 "greater accuracy" → "higher accuracy"?

RESPONSE 1.6: Yes we agree, it should be “higher accuracy”. Nevertheless these lines have been removed in the revised manuscript.

(7) ”better suited to the requirements of an operational system” this statement is questionable. Taking the example of meteorology, machine learning weather prediction models sometimes show a lack of physical consistency in their forecasts, which is typically not what you would like in a operational setting.

RESPONSE 1.7: We would like to thank the referee for raising this issue. We deleted the first 6 sentences of this paragraph (“Advances in artificial intelligence … a large database (…)”).

(8) L44 ”are beginning to emerge” the article cited has been published 6 years ago, I wouldn’t say that it is the beginning.

RESPONSE 1.8: We would like to thank the referee for raising this issue. We deleted the first 6 sentences of this paragraph (“Advances in artificial intelligence … a large database ()”).

(9) L45 ”Deep learning is based on the supervised training” there are many examples of applications of semi-supervised or even unsupervised learning in the geosciences.

RESPONSE 1.9: We would like to thank the referee for raising this issue. We deleted the first 6 sentences of this paragraph (“Advances in artificial intelligence … a large database ()”).

(10) L46 ”To enable more real-time and operational monitoring. . . ” Should’nt you start a new paragraph here?

RESPONSE 1.10: We would like to thank the referee for raising this issue. We deleted the first 6 sentences of this paragraph (“Advances in artificial intelligence … a large database ()”).

(11) L54 ”seems more suitable” at this point, this is hypothetical, as it remains to be proven that such ML-based operator can be constructed and possesses sufficient accuracy for the task. Perhaps replace ”seems” by ”could” or ”would”?

RESPONSE 1.11: We thank you for this relevant remark. We replaced
“An alternative machine learning-based operator for SIF seems more suitable.”
by
“An alternative machine learning-based operator for SIF could be more suitable.” Now on L. 49 in the revised manuscript.

(12) L63 ”the methods including the dataset” I wouldn’t say that a dataset can be considered as a method.

RESPONSE 1.12: Thank you for pointing this out. To clarify, we reorganised the paper by distinguishing a Data section from the Methods section. The section 2 of the revised manuscript is now dedicated to datasets and the section 3 presents the methods. The introduction has changed accordingly on L. 58 to 64.

(13) Figure 1: For panel(a), I would suggest to use a rectangular projection (like PlateCarree

for example) to be consistent with the domain used. I assume that the colourmap is the same for all three panels (it is not explicitly mentioned). With this choice of colours, the map is very difficult to read. I would advise to choose a so-called categorical colourmap. Also, I am not sure whether the figure is intended to be single or double column, but I have the impression that a lot of space is lost with padding and marging.

RESPONSE 1.13: We agree. We have simplified Figure 1 as follows.

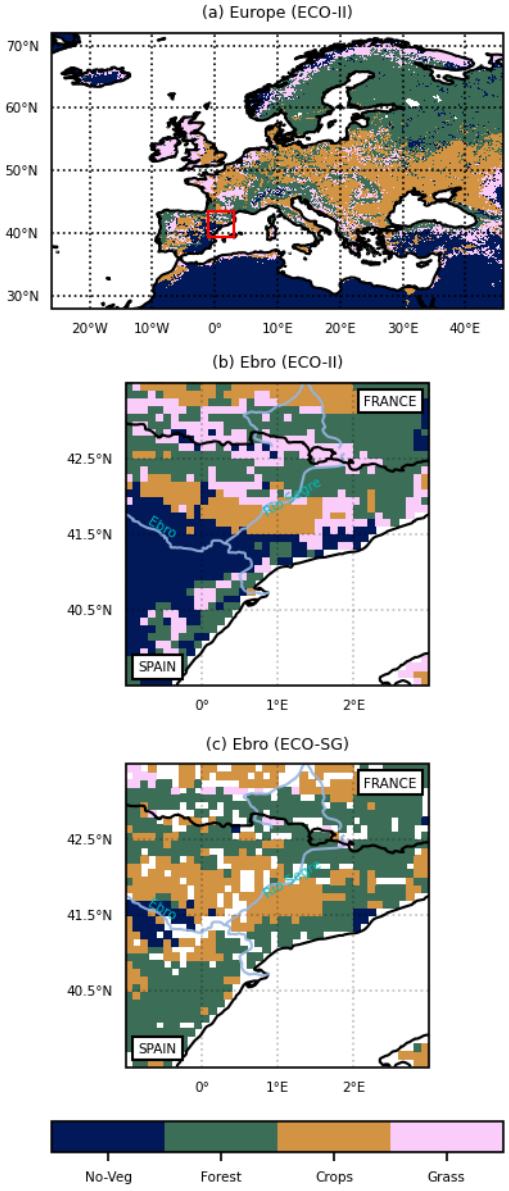

FIGURE R1.1 - Dominant Plant Functional Type according to land use maps. (a) Domain Europe with ECOCLIMAP-II [26◦W to 46◦E, 28 to 72◦N], the delimited area in red represents the Ebro basin subdomain. (b) The Ebro basin subdomain [1◦W to 3◦E, 39.5 to 43.5◦N] with ECOCLIMAP II. (c) The Ebro basin subdomain with ECOCLIMAP-SG. Four surface types are shown: grass (C3 grassland, C4 grassland), crops (C3 croplands, C4 croplands, flooded crops), forest (deciduous broadleaf forest, coniferous needleleaf forest, evergreen broadleaf forest), No-veg (bare soil, rocks, snow and ice, peat and wetlands).

Figure R1.1 corresponds to the new Figure 1 in the revised manuscript. It was adapted to be on

the horizontal rather on the vertical.

(14) L86-91 I would reformulate these sentences. As a non-expert in land surface modelling, I got the impression that the same process is described three times and was totally unable to understand exactly what is done here.

RESPONSE 1.14: Yes. We replaced:
"The retrieval process involves the statistical modelling of TOA radiances over vegetation, while excluding the signal induced by the underlying bare surface. Surface vegetation is described using the MODIS2018 database (Friedl and Sulla-Menashe, 2015). Using prior knowledge of the solar-induced fluorescence spectrum, the algorithm extracts the SIF-related part of the TOA radiance signal (Guanter et al., 2015). The estimated SIF is then computed by subtracting the expected radiance of the bare soil from the TOA signal. This process ensures that the distribution of TROPOMI SIF observations is centred on zero over bare soil."
by
"The retrieval process involves statistically modelling TOA radiances over vegetation while excluding the signal induced by the underlying bare surface. The MODIS2018 database (Friedl and Sulla-Menashe, 2015) is used to identify areas of vegetation that are relevant for SIF retrieval. Using prior knowledge of the SIF spectrum, the algorithm extracts the SIF-related part of the TOA radiance signal (Guanter et al., 2015). This process ensures that the distribution of TROPOMI SIF observations is centred on zero over bare soil." This was changed in the revised manuscript (L. 81-85).

(15) L91-92 "This means that negative values for TROPOMI SIF can occur, which we have discarded". I am a bit lost here (most probably because I don't know what exactly SIF is): is SIF supposed to be positive?

RESPONSE 1.15: Thank you for pointing this. We replaced:
"This means that negative values for TROPOMI SIF can occur (…)"
by
"This means that non-physical negative values for TROPOMI SIF can occur." (L. 85 in the revised manuscript).

(16) L98-100 At this point, it is unclear for what these additional measurements will be used.

RESPONSE 1.16: We agree. We moved the description of the LIAISE HyPlant dataset to the supplement. This is now in the second section of the supplementary materials and mentioned on L. 307-309 in the revised manuscript.

(***17) L107-110 "if an observation is missing, the gap is not filled" and "using an arithmetic average if at least 50% of the grid points are observed (i.e. half the minimum amount)" this whole part is confusing. First, I think that it would be good to remind the reader why observations can be missing and if we are speaking of only parts of the observation missing or entire maps missing. Then, depending on the answer to that question, does the order of interpolation (space or time first) matter? Finally, the 50% threshold does not really make sense to me.***

RESPONSE 1.17: Thank you for your remark and for highlighting that this paragraph requires more detailed explanations.

LAI observations are available at a resolution of one kilometre, at best, every 10 days. In order to align with our daily TROPOMI SIF product during the NN training process, the LAI maps undergo linear time interpolation. This is only performed between two consecutive LAI observations separated by 10 days. If a LAI observation is missing, the gap is not filled. As in previous studies (Albergel et al., 2010; Barbu et al., 2011; Corchia et al., 2023, among others), LAI observations are aggregated to the spatial resolution of the model grid cells using an arithmetic average, provided that LAI observations cover at least 50% of the grid cell.
We checked that the order of interpolation (space or time first) does not impact the final maps.
These explanations were included in the revised version of the paper. This is mentioned on L. 97 to L. 101 in the revised manuscript.

(18) L111 "instead" is repeated

RESPONSE 1.18: Thank you, it was removed.

(19) L112 "is spatially interpolated the same way" no time interpolation for these observations?

RESPONSE 1.19: Yes. We replaced:
"This provides data every 10 days and is spatially interpolated in the same way."
by
"This provides data every 10 days and is spatially and temporally interpolated in the same way."
(L. 105-106 in the revised manuscript).

(20) Figure 2: I would advise to use a grid, otherwise the plot is difficult to read. You can also set the green axis range to (0,5) so that the tick positions are exactly the same for all three axes. Furthermore, is there a real interest to present the data as points (with fancy markers that can be distracting) instead of lines? In the caption, you mention the domain Fig1, I would suggest to add the (a) label to remind the reader that it is the median over the European domain. Finally, are you showing the statistics for the raw or interpolated product?

RESPONSE 1.20: We thank the referee for these suggestions. We now use lines and we modified the axes to make the figure more readable. We replaced "Fig. 1" by "Fig. 1a". Regarding the displayed statistics, the LAI and SIF are interpolated products, whereas the GPP remains on its original grid (pre-interpolation). These changes are adopted in Figure 2 of the revised manuscript.

(21) L 118 "which is present only in the TROPOMI SIF and LAI V1 datasets". I only saw that spike for the LAI V1 in the figure after zooming (the green cross is almost hidden by a blue triangle). I think that this is a good example why lines would be perhaps more appropriate for this figure.

RESPONSE 1.21: Yes, we now use lines instead of points (see Figure 2 in the revised manuscript).

(22) L 136-137 "This process involves identifying the optimal control vector according to a model that minimises discrepancies between observations and a prior background estimates" This needs to be reformulated. As is, one understands that the choice of the control vector is part of the DA system, that the model does the minimisation, that discrepancies are computed between the background and observations, and that the background is to be optimised! Furtermore "prior background" seems redundant.

RESPONSE 1.22: Yes. We replaced:
"This process involves identifying the optimal control vector according to a model that minimises discrepancies between observations and a prior background estimates."
by
"This process involves updating the control variables in the ISBA model by minimising discrepancies between the variables and observations, and between the variables and a prior estimate." (L. 132-134 in the revised manuscript).

(***23) L 139 "(SEKF)(Mahfouf and Bilodeau, 2009)" → "(SEKF, Mahfouf and Bilodeau, 2009)"***

RESPONSE 1.23: We corrected this, thanks for pointing this out. The correction appears on L. 135 in the revised manuscript.

(24) L140 "x, where the subscript denotes the temporal step" Please reformulate. As is, you are referring to a subscript in "x", which has no subscript.

RESPONSE 1.24: Yes. We replaced:
"Thereafter, the control vector will be referred to as x, where the subscript denotes the temporal step."
by
"Thereafter, the control vector will be referred to as x_i, where the subscript denotes the temporal step." (L. 135-136 in the revised manuscript).

(25) L144 "where he superscripts "a", "b", and "o" stand for analysis, background and observations, respectively." I think that you can remove the "o" superscript. Also, I would use an italic font for "a" and "b" here, to match what is used in the equations.

RESPONSE 1.25: We changed the notations and fontstyle used in the equation to be consistent (L. 140-141 in the revised manuscript).

(26) Eqs 3a & 3b, K should be in mathbf

RESPONSE 1.26: Thank you for pointing this out, it was corrected.

(27) L160 is "assessed" the appropriate word here?

RESPONSE 1.27: Yes. We replaced:
"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." (L. 156 -159 in the revised manuscript, including the introduction of the residuals and innovation).

(28) Before section 2.3 (but also before section 2.2) I think that it would be good to provide some transitions.

RESPONSE 1.28: Thank you for this suggestion, we revised the structure of the paper and we added transition sentences (L. 127-129 and L. 160-151).

(29) L172-173 "we will train a single neural network to cover all pixels." This gives the impression that you design a NN that process the map at once, which, as far as I understand later on, is not the case: each grid point is processed independently.

RESPONSE 1.29: This sentence was rephrased. Although an NN is applied to each cell independently, only one NN is trained on all the model grid cells. We are not training a dedicated NN for each grid cell in the model. New sentences are now in the revised manuscript on L. 168-170.

(30) L174 "probability of belonging to a given category" Now I am confused, I thought it was a regression problem, why do you mention classification?

RESPONSE 1.30: Yes. We replaced:
"A feed-forward neural network's basic architecture comprises an input layer, one or more hidden layers of neurons, and an output layer, which provides an estimate or probability of belonging to a given category."
by
"A feed-forward neural network's basic architecture comprises an input layer, one or more hidden layers of neurons, and an output layer, which provides an estimate." (L. 170-171 in the revised manuscript).

(***31) L183 "with a maximum close to zero" do you mean here the maximum of the SIF values or the mode of the distribution?***

RESPONSE 1.31: Yes. We replaced:
"The distribution of TROPOMI SIF values is similar to a log-normal distribution with a maximum close to zero (see in the appendix)."
by
"The distribution of TROPOMI SIF values resembles a log-normal distribution where the peak of the mode of the distribution is close to zero (see Fig. S1 in the supplement)." (L. 178 -179 in the revised manuscript).

(32) L184 "below zero predictions" → "negatives predictions"

RESPONSE 1.32: Yes. "below-zero" was replaced by "negative". (L. 180 in the revised manuscript).

(33) L 184-185: "we applied the following strictly increasing function to the TROPOMI SIF values to predict. "The sentence should end with ":". Also, to prove your point, it would be perhaps easier to give the expression of the inverse function and show that it has positive values, no?

RESPONSE 1.33: Yes. The inverse function writes :
$SIF = Exp(SIFl)$ if $SIFl \leq 0$, and $SIF = \ln(SIFl+1)+1$ is $SIFl > 0$. Both the function and its inverse are presented in the revised manuscript in Eq. (3a-b).

(34) L190-200: I am confused here. First, you should mention explicitly what is coming from observations and what is coming from simulations. I guess that GPP, LAI V1, TROPOMI SIF are observations, but for the rest it is not explicitly said. Then, you mention a simulation database: where does it come from? Did it already include assimilation? Does it use the same model as the one you are using for the assimilation experiments? This last point is especially important: if the NN is trained to predict observations based on simulated variables (i.e. more or less something close to a background), then in a data assimilation experiment I would expect the innovation (y-Hxb) to be essentially noise, right? In the end, I don't think it is an issue, because you only selected LAI-V1 (which, as far as I understood, is an observation product) plus some extra predictors, but this is definitely confusing.

RESPONSE 1.34: Yes. In this sentence, only LAI V1 and TROPOMI SIF correspond to observations. This section was clarified by moving the description of preliminary tests (including Appendix A2) in a Supplement. We now focus on the actual configuration used in the paper. This section was moved to section 1.2 of supplementary materials. It is mentioned on L. 189 of the revised manuscript.

(35) L204-207: I did not understand this paragraph, in particular connections.

RESPONSE 1.35: Thank you for pointing out this missing information. This section lacks a reference to Annex Figure A2, which displays the average Pearson correlation between each input and output. This step is intended to identify possible linear relationships between certain variables. This section was clarified by moving the description of preliminary tests (including Appendix A2) in a Supplement. This is Fig. S2 in the Supplement.

(36) Fig 4: perhaps it would be better to present these results as a table instead of a figure?

RESPONSE 1.36: We thank you for your suggestion, we now show the Table associated to this Figure. This section was clarified by moving the description of preliminary tests (including Appendix A2) in a Supplement. The figure was converted into a Table in the supplementary materials (Table S1).

(37) L214: ZSE and LON seem to have similar impact. Why not dropping LON as well?

RESPONSE 1.37: ZSE seems to have a little less impact, and keeping both coordinates as metadata predictors felt more logical.

(38) L215 "even if the observation operator has learnt the model biases": how could the observation learn the model biases, since it only relies on observation products in the training?

RESPONSE 1.38: There is a misunderstanding due to the lack of clarity in this section. We meant that, since the NN is trained using LAI observations and LAI is the only biophysical predictor, the NN can be used with any land surface model. The following sentence was deleted:
"This configuration ensures that, even if the observation operator has learnt the model biases, the assimilation of SIF will not simply adapt to the model and that it will remain robust to any improvements to the model."

(39) L217 " to estimates" → "to estimate"

RESPONSE 1.39: Thank you for pointing out this typo, it was corrected.

(40) L218-225: I would reformulate this paragraph a bit to make it a bit smoother. Also, you mention "from 1 june 2018 to 31 May 2021" for the full data, but it is mentionned in 2.1.2 that Troposif provides data from May 2018 to December 2021. Did you discard part of this data? Also the final year is not used at all?

RESPONSE 1.40: Yes, this was a mistake. The full data range used is from June 2018 to December 2021, including these dates. This was corrected. LAI-V1 was used for training until June 2020. Therefore, the training/testing dataset could not extend beyond June 2020. However, the availability of LAI-V1 did not constrain the assimilation experiment, which is why it was conducted up to December 2021. To make it clearer, the period of the different experiments runs are now provided in Table 1 of the revised manuscript.

(41) L228-229: "As mention earlier" → "As mentionned earlier"

RESPONSE 1.41: Thank you again, we corrected this typo.

(42) L227: our focus is on a small area" it is unclear whether the small domain is used only for the assimilation or for the selection of the training dataset as well? Also, it is unclear whether the forecast model used in the assimilation experiments is able to run on the small domain only or needs to be run on the whole European domain?

RESPONSE 1.42: Training was carried out across the entire European domain. However, the assimilation experiments were only conducted over the Ebro basin. This was made clearer in the revised paper. It is now mentioned in the captions of figures and Table 1. This is now explicit on L. 209-210 and on L. 272 in the results section of the revised manuscript.

(43) L231 "The assimilation period runs from 1 January 2018 to 31 December 2021" I thought that the Tropomi data only started in May 2018?

RESPONSE 1.43: You are right, the TROPOMI data were not available in early 2018, but we also run an assimilation experiment involving the LAI product, which was available. We have chosen to run the experiment for a full year.

(44) L233: why did you change the land cover map? If the second is more accurate and more recent, then why not using it for the training database?

RESPONSE 1.44: At the time of the study, ECOCLIMAP-II-driven ISBA simulations were the only ones available for the entire European domain. As ECOCLIMAP-II tended to prescribe excessive bare soil in the Ebro basin region, the ECOCLIMAP-SG updated framework was included for the data assimilation experiments.

(45) L241-243: It is not entirely clear in the text that you are talking about the observation error for the SIF observations (and not for the LAI observations).

RESPONSE 1.45: Yes. We replaced:
"To determine the amount of observation error to consider,…"
by

"To determine the amount of SIF observation error to consider,…" This change can be seen on L. 225 in the revised manuscript.

(46) L236-246: you don't mention the dates for the experiments assimilating SIF. EDIT: From the caption of figure 7, I get the impression that the SIF assimilation experiments cover June 2018 - May 2020, which includes the training set of the NN (June 2018 - June 2019). I would heavily advise to remove this period from the statistics.

RESPONSE 1.46: To clarify, we detailed the assimilation time periods by adding a dedicated column in Table 1. We produced another version of Figure 7 (new Fig. 6), where the different sub-periods are distinguished (June 2018-May 2019, and June 2019-May 2020). The new Fig. 6 in the revised manuscript presents the two different periods in addition to the combination of both.

(47) Table 2: is there an interest in including the line without data points? Also, you haven't defined ubRMSE (I haven't checked, but there might be other undefined acronyms)

RESPONSE 1.47: We checked throughout the paper that all the acronyms are defined. "ubRMSE" stands for "unbiased root-mean square error" and is equivalent to standard deviation of differences. The "zero" lines have been removed in Table 2 of the revised manuscript. The ubRMSE is now defined on L. 207 of the revised manuscript.

(48) Fig. 5: I would advise reducing the padding of the figure overall (I don't think that the ticklabels are needed) and using continuous (instead of discrete) colourmaps. Perhaps you could also change the projection to PlateCarree. Furthermore, I don't understand the choice of a diverging colourmap for correlation, unless there is something specific happening when the correlation is 0.5? Finally, even if this figure is intended for one column, I think you can use a bit more horizontal space.

RESPONSE 1.48: Thank you for your suggestions on how to improve this figure. We adjusted the settings for the plots. Regarding the colour maps, we ran several tests and found that the current settings, which are based on discrete colour maps, produce the most readable figures (new Fig. 4 in the revised manuscript, with adjusted size and projection).

(49) L279: "The saved weights of the neural network are implemented in the assimilation scheme" please reformulate.

RESPONSE 1.49: Yes. We replaced:
"The saved weights of the neural network are implemented in the assimilation scheme as an observation operator for the SIF."
by
"In the first phase, the NN was trained across the entire European domain. In this section, the saved weights of the neural network are implemented in the assimilation scheme as an observation operator for the SIF." This correction appears on L. 268-269 in the revised manuscript.

(50) L281: "LAI representation will be our main focus on the control vector variables" I don't understand what you mean here?

RESPONSE 1.50: Yes. We replaced:
"As the observation operator only takes LAI as inputs, LAI representation will be our main focus on

the control vector variables."
by
"Since the observation operator only considers the LAI among the control variables, the accuracy of the LAI representation is the main criterion for evaluating the assimilation results." This correction appears on L. 270 to 272 in the revised manuscript.

(51) L282: "the monthly-averaged maps over the C3 crops" It is unclear to me whether the figure will show the increments only over this land category or over the entire domain. Also, is there a reason why you want to focus on this category?

RESPONSE 1.51: Figure 6 only shows the increments for this specific C3 crops vegetation type. The reason for focusing on this category is twofold: firstly, the NN tends to perform better where this category is dominant; and secondly, the Ebro basin is an agricultural area which is typically characterised by this specific vegetation type. This enables us to observe that LAI increases in known irrigated areas in summer and that this increase is applied to the correct type of cover. To clarify Figure 6, we removed the grid cells displaying less than 10% C3 crops (see FIGURE R1.2 below). Figure R1.2 corresponds to the new Fig. 5 in the revised manuscript.

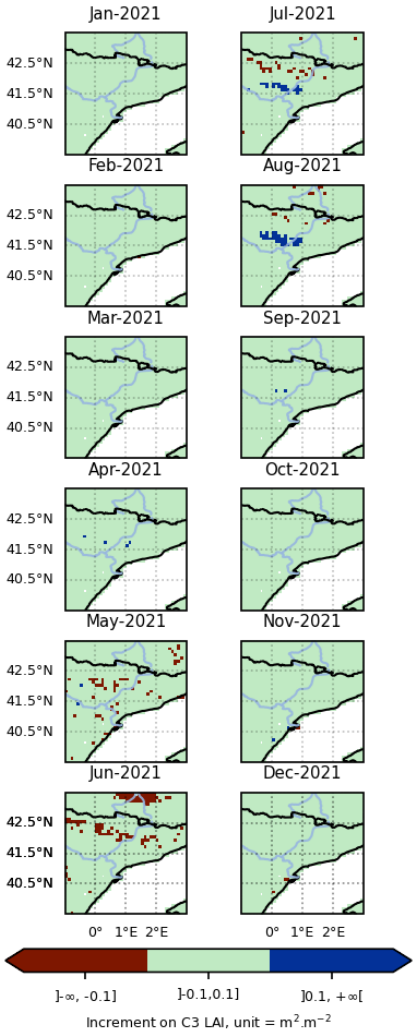

FIGURE R1.2. Mean monthly analysis increments of the SIF20 experiment on the LAI of C3 crops over the Ebro domain due to SIF assimilation during the year 2021. Grid cells displaying less than 10% C3 crops are masked.

(52) Figs. 6 & 7: Same as for the previous figure, I think that you can remove the ticklabels and reduce padding overall. For this data, you absolutely need a diverging colourmap centered on zero.

RESPONSE 1.52: A linear colour map was used to meet the colour blindness requirement. If we use a diverging colour map, the difference between the two extreme values will not be significant enough. We now use another colour map. Figs. 5 and 6 have been updated in the revised manuscript, we hope they are now clearer.

(53) L290: You mention an RMSE here. Can you be explicit and say between what and what this RMSE is computed? From the caption in the figure, I would assume that this is the RMSE between the LAI observations and the LAI analysis. I think it would be good to remind the reader that LAI observations are not assimilated in this setup and hence that they can be used to evaluate the analysis, provided that their errors are independent of the errors of the SIF observations.

RESPONSE 1.53: The referee is right, the RMSE is computed between the LAI from the analysis SIF20 and LAI-V1 observed, in units of $m^2\ m^{-2}$. In this experiment, the LAI is not assimilated and thus we can use the LAI-V1 as reference. We have improved this section when describing Fig. 6 (former Fig. 7). The new description of the results presented in Fig. 6 of the revised manuscript starts on L. 284 and ends on L. 298.

(54) L295: "all cells with a p-value higher than 0.01 were removed" can you please specify which statistical test is used to compute this significance test? Is this mask also applied in the delta(RMSE) panel?

RESPONSE 1.54: In our case, we use the F-test. The p-value is the result of a test of the null hypothesis that two normal populations have the same variance. This enables us to eliminate grid cells with insufficient data. This mask is used for both the RMSE and correlation panels. This is explicitly mentioned on L. 282-284 in the revised manuscript.

(55) L300: "are closer to the observed values" according to which metric?

RESPONSE 1.55: In this case both metrics are used (i.e. RMSE and correlation). It is now clarified on L. 299-300 in the revised manuscript.

(56) L306: "the reference LAI-V1 experiment" I think that you should be careful with the results of that experiment since here you are evaluating against LAI observations, i.e. the observations that are assimilated.

RESPONSE 1.56: In Fig. 10 (former Fig. 9), the LAI-V1 experiments is used as a benchmark for other experiments: SIF01, SIF20, and SIFFR.

(57) Section 3.2.2: same remark for the experiments here.

RESPONSE 1.57: In Fig. 10 (former Fig. 9), the LAI-V1 experiment is used as a benchmark for other experiments: SIF20, SIFFR, SL20, and SLFR.
In the revised manuscript (L. 221-224): "Another reference experiment we use is the baseline experiment, which is referred to as the LAI20 experiment in Table 1. The LAI20 experiment will use

the LDAS to assimilate LAI-V1 from 1 January 2018 to June 2020, and provide a state-of-the-art analysis to evaluate the specific benefits of assimilating SIF data. The initial conditions were spun up for 20 years and shared by these to reference experiment and all SIF assimilation experiment described thereafter."

(58) Fig. 11: The caption is not very clear. It doesn't give a description for (a). Which analysis is plotted? I assume the whisker show the in-situ measurements but it is not mentioned.

RESPONSE 1.58: Sorry for that. We now describe the HyPlant campaign in the supplement and we replaced the current caption of Fig. 11:
"Comparison to the in-situ measurements of SIF during the LIAISE campaign (OL: Open-loop, AN: analysis). Box represents the 25th and the 75th percentiles, the whiskers the 5th and 95th percentiles. The red line represents the 50th percentile and the green triangle the mean."
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." The new Fig. R1.3 and captions correspond to Fig. S6 of the supplement.

(59) L337-338: "we must compare all in-situ SIF measurements on a given day with the simulated SIF value resulting from the analysis for the two situations shown in Fig. 11(b)" I don't understand this.

RESPONSE 1.59: It is a typo, it should read "for the two simulations" (i.e. SIF20 and OL). It was corrected in Section 2 of the supplementary materials.

(60) L340-341: "the values are closer overall to the average airborne measurement value": but I thought that absolute values should not be compared because of different wavebands?

RESPONSE 1.60: We have produced a clearer version of Figure 11 (FIGURE R1.3 below), in which the TROPOMI SIF data and model experiments are more visible (FIGURE R1.3c). The evolution of SIF20 appears to be more consistent with the airborne SIF time series than that of TROPOMI SIF and OL. However, this could be an artefact due to the lack of TROPOMI SIF data on 17 and 22 July 2021. As these results are not very conclusive, we moved this section to the supplement.

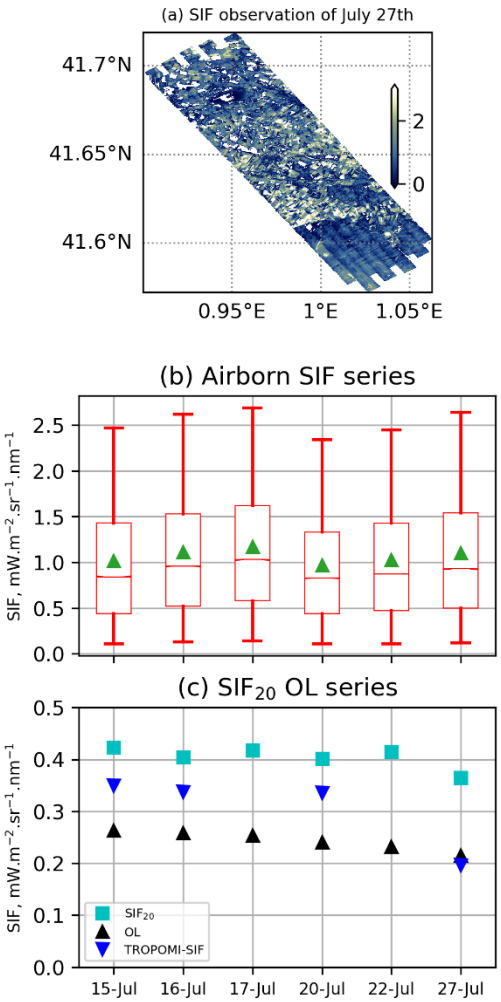

FIGURE R1.3 - Comparison of the OL and SIF20 experiments with the SIF airborne measurements of the HyPlant instrument within the LIAISE campaign. Boxes represents the 25th and the 75th percentiles, the whiskers the 5th and 95th percentiles of the airborne SIF observations. The central horizontal red lines represent the 50th percentile and the green triangles the mean of the airborne SIF observations.

(61) Section 4.1 After reading the entire section, I am confused. I don't understand what you want to demonstrate here. The figure seems to show that the output of the NN more or less matches the values of TROPOMI SIF (which is what it has been trained for) and does not match another instrument: this is absolutely to be expected, or am I missing something here?

RESPONSE 1.61: 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 not very conclusive, we moved this section to the supplementary material.

(62) L361: "Using this simple architecture" you are presenting the setup here, not really the architecture.

RESPONSE 1.62: We agree. "architecture" was replaced by "setup".

(63) L362: "no overfitting evident" -> "no evident overfitting"?

RESPONSE 1.63: Thank you for this remark. This was changed. L. 368 in the revised manuscript.

(64) L370: "update rate yielded by TROPOMI SIF assimilation is five times higher than that of 10-day synthesis assimilation" I don't understand this point.

RESPONSE 1.64: Yes. We replaced:
"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." L. 376-377 in the revised manuscript

(65) L374: "Using a neural network operator instead of a physical model makes propagating uncertainties difficult" What do you mean here? The uncertainties are all accounted for the in the B and R matrices, no?

RESPONSE 1.65: Yes. We replaced:
"Using a neural network operator instead of a physical model makes propagating uncertainties difficult, so quantifying observation errors relies more on empirical choices."
by
"This study 380 has shown that, when assimilating SIF alone, an observation error of 20% leads to the best assimilation performance. In a co-assimilation framework incorporating LAI, the representation of vegetation variables (i.e. LAI and GPP) improves further, demonstrating the value of redundancy in observations used in a data assimilation system. However, the best performance is obtained using a more complex error description (FR)." On L. 380 to 383 in the revised manuscript.

(66) L375-376: "It is also important to understand that the neural network is trained to emulate observations with instrument uncertainties rather than the true physical variable, making quantifying the observation error of the true value challenging". I don't fully agree with that statement. The instrumental errors are (in principle) unpredictable and hence the NN cannot learn them. It is true that the use of a NN will introduce modelling error (because the process from state to observations is not perfectly represented) but this is also the case for any observation operator.

RESPONSE 1.66: Yes. We replaced:
"It is also important to understand that the neural network is trained to emulate observations with instrument uncertainties rather than the true physical variable, making quantifying the observation error of the true value challenging."
by
"Further research is required to quantify observation errors accurately." This appears on L. 383-384 in the revised manuscript.

(67) L416-417: "information that cannot be changed" this is unclear: why couldn't you change the values of the support parameters?

RESPONSE 1.67: Designating 'metadata predictors' as 'support parameters' was not relevant. We replaced:

“These support parameters provide information on the expected seasonal cycles. Yet, their use is questionable since they will provide information that cannot be changed, which could prevent the model from generalising well.”
by
“These metadata predictors provide information on the expected seasonal cycles.” It was changed in the supplementary materials of the revised manuscript.

Authors' response to the second anonymous referee

We thank the anonymous referee for his/her review and all the relevant remarks and questions that helped us to further improve this manuscript.

(1) Lines 40-45: while I have experience is 'classical' machine learning (Random forest) or data assimilation, deep learning is something new for me. Please state briefly (and explicitly) what the difference is between deep learning and more classical machine learning approaches.

RESPONSE 2.01: Thank you for pointing out this lack of description. The term 'deep learning' refers to the fact that we are using a neural network with several layers. Training over a large database can be either unsupervised or semi-supervised. Here, we use supervised training, since we feed the neural network labelled data. These sentences were removed, for the sake of clarity.

(2) Line 46: if near-real time is the concern, you could have used near-daily SIF data. What is the advantage of using the 8-day TROPOMI? Please clarify.

RESPONSE 2.02: In this study, we do not use the 8-day TROPOMI data, we use daily TROPOMI SIF data instead. This now appears more clearly in the Introduction, L. 40 to 45 of the revised manuscript.

(3) A thought: the TROPOMI SIF data have never been properly calibrated/validated. They also contain quite some negative values, so it is likely that some of its values are biased. The traditional answer here is that the variability in SIF contains most of the information, as opposed to the absolute SIF value.

RESPONSE 2.03: Thank you for sharing this with us. The negative values are definitely an issue. We chose to remove them before creating the database and assimilating the observations. This also biases the product to some extent, since it was initially statistically calibrated to be unbiased over bare soil areas. We have also decided to keep this strictly positive rule for SIF in the observation operator design. As you said, we believe that enough information in the SIF is still important to be used anyway. Another approach would be to use the product with negative values and then train the neural network to match the zero bias over bare soil (allowing it to take on negative values).

(4) Line 97-100: the airborne campaign is not mentioned in the introduction. How important is it for the research question? The 'why an airborne campaign' needs more context.

RESPONSE 2.04: The HyPlant airborne campaign is here to provide context for the discussion of the limitations of observation operators. This part was moved to a supplement. This is now the second section of the supplementary materials.

(***5) Line 106: the introduction (line 46) mentions an 8-day SIF product, while this line mentions a daily SIF product. Please clarify.***

RESPONSE 2.05: In this study, we do not use the 8-day TROPOMI data, we use daily TROPOMI SIF data instead. This was made clear in the Introduction.

(6) General thought: section 2.2 within methodology would be easier to understand if there were some kind of visual. Figure 3 is helping a bit, but it focusses on the neural network itself, rather than on the assimilation.*

RESPONSE 2.06: Thank you for your comment. We included a new figure presenting the assimilation. New Fig. 3 in the revised manuscript presents a schematic view of both the NN observation operator used and the assimilation cycle.

(7) Line 193: why having two LAI datasets? LAI and LAIv1?

RESPONSE 2.07: One LAI is an observation product (LAI-V1) and the other is the LAI produced by ISBA model. This part was moved to a supplement. This is now in Section 1 of the supplementary materials.

(8) Line 195: when doing machine learning for SIF, I often note that not DOY, but sin(doy/365*2pi) and cos(doy/365*2pi) are more potent predictors compared to doy itself. What's the reasoning for not using these?

RESPONSE 2.08: Cycling the DOY metadata predictor could actually improve the NN training. More research is needed to address this question.

(9) Line 208: year missing in reference.

RESPONSE 2.09: Thank you for pointing this out, we corrected as: Lundberg and Lee (2017).

(10) Line 228: leave out 'as mentioned earlier in this study'. Be either specific on where it was mentioned or leave it out altogether (I recommend the latter)

RESPONSE 2.10: Yes. We replaced:
"As mention earlier in this study, this area comprises heavily irrigated croplands which are usually not accurately simulated by the land surface model"
by
"This area comprises heavily irrigated croplands which are usually not accurately simulated by the land surface model". This appears on L. 211-212 in the revised manuscript.

(11) Section 2.3.3 the term 'learning process' does not properly cover the content of this paragraph. Consider something like 'feature selection' or 'training variable identification'. My main point is, 2.3.3. does not describe the neural network itself, but rather the selection of the variables and the train-test split.

RESPONSE 2.11: This part was moved to the supplementary materials.

(12) Section 2.4: role of airplane data in the assimilation is not clear here. The methodology just mentions the existence of the dataset, but not its purpose within the study.

RESPONSE 2.12: The description of the airborne campaign was included here to bring together all the SIF datasets used in the same section. However, it is not referenced until the discussion section. We moved this description to a supplement.

(13) Line 236: a detail, but I am struck the use of the future tense here. Why would you 'will run'. Haven't you done it? I presume you mean it's a continuous (and therefore unfinished, but already initiated) process. The future tense made me think the process wasn't even started yet.

RESPONSE 2.13: We removed the future tense when it was irrelevant.

(14) Line 239: repeat what precisely you mean by the 'baseline experiment'.

RESPONSE 2.14: Thank you for pointing out this missing, we now repeat that the baseline experiment corresponds to "LAI-V1" (Table 1).

(15-16) Line 243-244: again a future tense. Please don't

RESPONSE 2.15: We changed the tense in both case, thank you for pointing this.

(17) Line 256: I agree on the good agreement, but the word 'good agreement' looks sloppy. Just write a metric here.

RESPONSE 2.16: Yes. We replaced:
"As shown in Table 2, the values based on the training and test datasets are in good agreement."
by
"As can be seen in Table 2, the skill score values obtained for the training and test datasets are similar."
Also, we improved the readability of Table 2. The revised Table 2 provides a more explicit split between the results obtained with the training dataset and the testing dataset. The overall skill score are in bold. The correction appears on L. 244 in the revised manuscript.

(18) Line 260: I agree with this interpretation but Huber loss is a bit of an esoteric term and I want to encourage you to make this paper readable for a larger audience than the people that are familiar with both SIF and Huber loss. Explain it a bit more in conceptual terms.

RESPONSE 2.17: Yes, we now define the Huber loss in the Methods section. Huber loss definition is provided in Eq. 5 in the revised manuscript.

(19) Line 263: how do you determine a bias? What is the reference here? The difference between the TROPOSIF measurement and the modelled SIF? Or difference with the airplane?

RESPONSE 2.18: As written in Table 2, the bias we are referring to is the mean difference between NN estimates of SIF and observed TROPOMI SIF. Thank you for pointing out its absence. We included it in the main body of the article. It is mentioned on L. 206 and in Table 2 of the revised manuscript.

(20) Figure 5: the non-existent correlation above the 60th parallel is striking. Why would that be? Presence of snow?

RESPONSE 2.19: It is mainly due to the lack of observations during the winter at these latitudes.

(21) Figure 5: I noticed you cut out mountainous regions. Why?

RESPONSE 2.20: To avoid any frozen soil where the assimilation cannot be done.

(22) Conclusion: really strong, but it ends on a false note. What's stopping you from adding additional variables (you have e.g., temperature maps to your disposal) or you can tune the hyperparameters.

RESPONSE 2.21: Yes; We replaced:
"This study assesses the benefits of incorporating TROPOMI SIF products into a land data assimilation system via a deep learning operator, which opens up new possibilities for systematically assimilating such products. As the architecture and training of the neural network operator are not overly specific to the TROPOMI SIF product, it would be interesting to adapt it for the assimilation of other SIF products. Better optimisation of the hyperparameters or the use of additional variables could enhance the neural network's accuracy."
by
"We demonstrated the advantages of integrating TROPOMI SIF products into a land data assimilation system using a deep learning operator. This opens up new possibilities for the assimilation of new SIF products. Although the neural network operator was designed for TROPOMI SIF, it can be adapted for other SIF products. We also addressed the co-assimilation of SIF with LAI, which could be extended to other satellite products, such as land surface temperature." This change appears on L. 386 to 389 in the revised manuscript.

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 have addressed these concerns. This helped us improve the Methods section, which required more details.

(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 was clarified by moving the description of preliminary tests (including Appendix A2) in a Supplement. We now 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. This is now discussed in section 1.2 of the supplementary materials.

(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. Both Fig. R3.1 and the discussion are now presented in the section 1.3 of the supplementary materials.

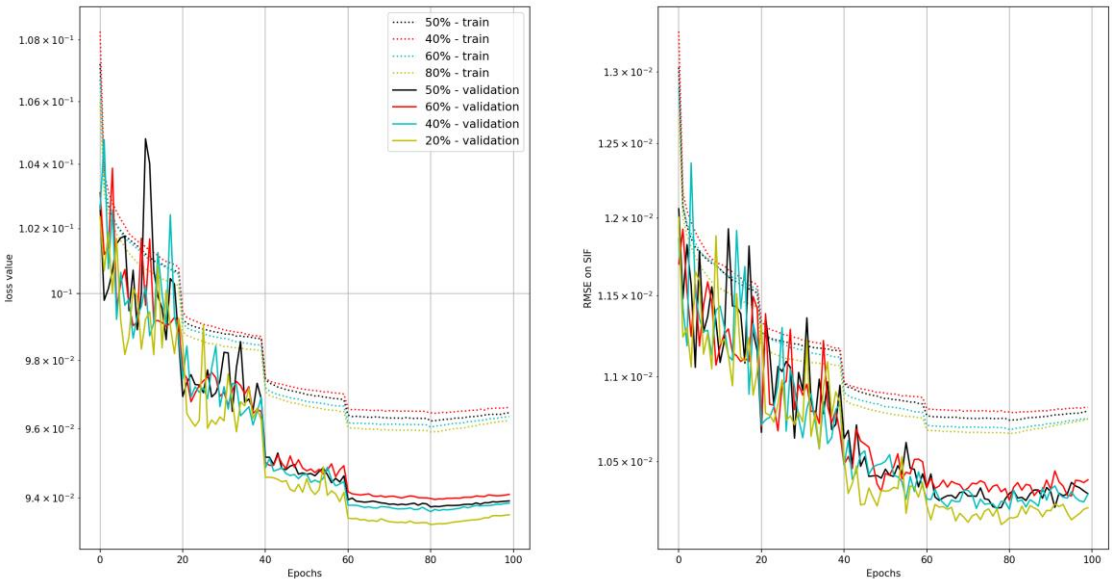

FIGURE R3.1: Learning curves for different partitions between training and validation

(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 now provide more details in the Figure captions, 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. This has been clarified in the revised version of the manuscript.

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

RESPONSE 3.06: In Section 3.2, the Jacobian matrix is now defined differently.

(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 replaced:
"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. See L. 156-159 in the revised manuscript.

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

RESPONSE 3.08: We replaced:
"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." This change appears on L. 163-164 in the revised manuscript.

(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 have 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. L. 219-224 of revised manuscript clarify why we are

using two different reference experiment (i.e. open-loop OL and LAI20 baseline) and their purposes.

(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 detailed the assimilation periods by adding a dedicated column in Table 1. We better distinguish the two LAI columns and indicate the observation errors used for LAI. We have completed the caption of Table 1 by mentioning 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 was checked in Annex A3 (Figure A3) and this Annex was moved to a supplement and add more details added. The validation and training loss are now presented displayed in the supplementary materials. The differences between the skill score obtained using data from the training dataset and the one using the testing dataset is made clear in the revised Table 2.

(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 is now explained in more detail in the data assimilation process description. The increment definition is given by Eq. 1b in the revised manuscript.

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

RESPONSE 3.13: We replaced:
"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. 6. 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 ensures that significant correlation and RMSE values are displayed. This mask is used for both the RMSE displayed in Fig. 6(a-c-e) and the correlation panels displayed in Fig. 6(b-d-f)." Note that in the revised manuscript, Figure 6 corresponds to former Figure 7. This correction appears at L. 283-286 in the revised manuscript.

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

RESPONSE 3.14: We mentioned it in the Figure 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 replaced 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." The revised Figure is now in the second section of the supplementary materials.

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

RESPONSE 3.16: We replaced:
"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." This change appears on L. 376-377 in the revised manuscript.

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

RESPONSE 3.17: Thank you for your suggestion, we added "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 replaced "regular grid" by "0.1° resolution regular grid". This appears on L. 90 in the revised manuscript.

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

RESPONSE 3.19: We changed 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. This part is now in the Supplement, in which we explain that we never used FLUXCOM in the training process. We also briefly describe the GPP, ET, and H, LST, …, simulations used in these preliminary tests.

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

RESPONSE 3.21: Yes this is the peak of the distribution. We replaced "maximum" by "peak" in this sentence, now in the Supplement.

(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 matters 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. The new Eq. 3b in the revised manuscript now clarifies this.

(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 completed the sentence, thank you for your suggestion:
"The metadata predictors provide information on the seasonal cycle and on the spatial variability."
This now appears in the Supplement.

(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 improved this Figure (now Fig. 4).

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

RESPONSE 3.25: This figure is supposed to take a single column.

(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 (GPP correlation), there is a missing reference to Fig. 10 (LAI RMSE) here. In terms of LAI 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. In the revised manuscript, Figs. 8 and 10 are now Fig. 8 and 9. Several modifications were made in Section 5.1 to clarify this.

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

RESPONSE 3.27: We changed the colour map to better distinguish the negatives values from the positives. The new Fig. 6 adopts a new colour scale in the revised manuscript.

(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 conducted a more thorough analysis of the data captured by FLUXCOM over this irrigated area. Its correlation with SIF and LAI observations suggests that FLUXCOM expects the vegetation to receive sufficient water during the dry period. Indeed, FLUXCOM-X is trained on meteorological conditions but also on MODIS land temperature and vegetation index. We provide in section 3 of the supplementary materials the monthly mean GPP over the Ebro basin for the years 2018 to 2021. In every year, the irrigated crops in the Ebro basin lead to a local increase in the GPP in the late summer. Even though, no information about the artificial irrigation is provided to FLUXCOM, like for our assimilation of SIF, MODIS remote sensing observations seem to have

captured this pattern and brought it in the FLUXCOM product. This is mentioned on L. 358-359 in the revised manuscript.

(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 was 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 moved this section to the supplementary material. This is now in the second section of supplementary materials.