

Dear Dr. Mukund Palat Rao,

We would like to thank you for taking your time to evaluate our work and foremostly for your interesting and useful comments, questions, remarks and suggestions.

We will try to answer to your questions and we integrated your structural and grammatical suggestions into the manuscript (all answers and changes are in blue color).

Do the RFR style models (RF-F, RF-SIF-R, RF-SIF-R-PFT) outperform the RF-SIF-VI model just because they have more predictor variables (14 vs 4)? Perhaps you can present some analysis on how these models perform on 'validation' data that the models have not been calibrated on?

The RF-R model has 11 inputs, the RF-SIF-R 12 inputs, the RF-SIF-R-FTP 13 inputs and RF-SIF-VI 4 inputs. In the manuscript, we used the adjusted R^2 , which considers the number of samples and predictor variables in its computation to evaluate and compare the performance of our different models (see line 221, page 6). We also compared the models based on the RMSE, which isn't sensitive to the number of explanatory variables.

We also used out-of-bag predictions of RF to calculate the adjusted R^2 and RMSE. We separated our dataset in two datasets: 80% of the data for training and 20% for testing or evaluating the model. Second, on the training dataset we applied a 10-fold cross-validation and 20 iterations to determine the best parameters for each model. Lastly, we evaluated or tested each model on the testing dataset, which were not seen by the model before (see lines 207-213, page 6).

As these points were not clearly stated in the manuscript, the lines 219 to 222 were modified as follow :

“Ultimately, the linear relationships between SIF and GPP were compared based on the coefficient of determination (R^2), Root Mean Squared Error (RMSE), and the p-value metrics. The random forest models were evaluated and compared based on out-of-bag adjusted R^2 and RMSE”.

Page 1, Line 15: “Earliest” studies, based could be “earlier” studies or prior studies?

Line 15: changed to “prior studies”

Page 1, Line 17: “plant functional type” should be plant functional types.

Changed to “plant functional types”.

Page 2, Line 14: “which is the amount of flux carbon taken up by vegetation.” The word ‘amount’ is not needed since the work flux assumes an amount.

L 43 :Changed to “which is the carbon flux taken up by vegetation through photosynthesis”

Page 2, Line 53: “Remote sensing is widely used to upscale canopy GPP to landscape, regional, and global scales and at daily scale using reflected sunlight measured by satellite sensors”. The “and at daily scale” addition seems a bit awkward. Maybe remote sensing is widely used to upscale daily GPP to landscape....

Changed to: “Remote sensing is widely used to upscale daily GPP to landscape, regional, and global scales using reflected sunlight measured by satellite sensors.”

Page 2, Line 60: “and biochemical canopy characteristics (Dechant et al., 2020; PabonMoreno et al., 2022). Although, they suffer.” The way the sentence is framed, I think it would flow better as a continuous sentence.

Changed to: “and biochemical canopy characteristics (Dechant et al., 2020; Pabon-Moreno et al., 2022), but they suffer from contamination by atmosphere and saturation in canopy dense ecosystems and are less sensitive to diurnal and daily variations in photosynthetic status resulting from physiological responses induced by rapid changes of abiotic stresses.”

Page 3, Line 83: “Early studies relied on ground-based”, should be Earlier studies relying on...

Changed to “Earlier studies relying on”

Page 3, Line 96: “which is on board Sentinel 5-Precursor, represents a novel (???) for understanding?”. A missing word after novel? Tool maybe?

Changed to: “which is on board Sentinel 5-Precursor, represents a novel tool for understanding”

Page 3, Line 97: “it provides a quiet high temporal resolution at daily”. Quite instead of quiet. However, the word quite is not needed either.

Changed to: “it provides a high temporal resolution at daily scale.”

Page 3, Line 104: “comprehensively addressed. Owing to most”. This should be once sentence, or the second sentence should start of as, this is due to the fact or This is because....

Changed to: “However, to the best of our knowledge, an attempt to study the synergy between those variables have not been comprehensively addressed due to the fact that the relationships between structural and functional components are not linear, and have complex interactions over time and space”

Page 4, Line 8: “data products is given in Supplementary Materials in Tab S2.”. the Authors don’t need to do this, but might consider including a column for the spectral band (i.e. visible (R/G/B), NIR, etc. in the table)..

Here the spectral band column was added in tab S2 as you suggested (see supplementary Material in Tab S2).

Page 6, Line 230-232: “Overall, SIF_d was significantly related with tower-based GPP at the site level and at the daily timescale (as $p < 0.0001$ was statistically highly significant), except for IT-Cp2 site of which GPP and SIF_d relationship was insignificant and weak ($R^2 = 0.001$, $p \leq 0.60$)”. This is of course quite subjective, but despite some of the sites being statistically significant I would not call these relationships as being strong. The reason for the statistical significance and p-value is being driven by the high sample size. In particular, I would add GF-Guy to the list of sites where there is no relationship between GPP-SIF_d. The correlation needed to get an R^2 of 0.2 is around 0.15 which is still quite weak. I would also then add FR-Mej, FR-EM2, and FI-Var to the list of sites with a weak relationship. I know this is mentioned a bit later, but maybe an easier way to frame it would be to not mention the weak relationship at IT-CP2 in the beginning, but then mention all these sites together at the end of the paragraph?

This part was reframed as follows (lines 230-236): “Overall, SIF_d was significantly related with tower-based GPP at the site level and at the daily timescale ($p < 0.0001$). However, Figure 2 indicates that the slopes and intercepts of the linear regression between GPP and SIF_d are site-dependent, suggesting that the difference in plant functional types and spatial heterogeneity across sites may significantly affect the relationships between GPP and SIF_d. The strongest relationships were found at DK-Sor and FR-Fon, which are DBF vegetation type sites, with R^2 values of 0.81 ($p < 0.0001$) and 0.66 ($p < 0.0001$). The weakest linear relationships were recorded at FI-Var, FR-EM2 and FR-Mej sites, and no significant relationship was found at GF-Guy and IT-Cp2.”.

Page 7, section 4.1.2 I like the progress from the site level (Section 4.1.1) to PFT level SIF_d-GPP relationships. However, the way the PFT level relationships are presented, don’t seem to actually allow us to closely examine within PFT spread in the R^2 ’s, slopes, and intercept. For example, in Table 2, all sites of a PFT are lumped together. If the authors wish to highlight the within PFT spread, one option could be to include boxplots by PFT for the R^2 , slope, and intercept for the SIF_d-GPP relationships.

We did not include the boxplot of R^2 , slopes and intercepts of the GPP vs SIF_d for sites from the same PFT, because, we have low number of sites for some PFT, including OSH (1 site), MF (2 sites), EBF (2 sites), and GRA (3 sites). In addition, the R^2 , slopes and intercepts of the relationships between GPP and SIF_d for each PFT and site are detailed in Tab S3.

Page 9, Section 4.5, Line 284, Supplementary Fig. S1: I would recommend changing the figure to have a diverging colour-bar. The gradient colour-bar from ~ -0.65 to 1 is not intuitive to me and hard to visualize.

In the figure below, the gradient colour-bar for the correlation matrix was changed, as you recommended. The figure was added to the Supplementary Material draft.

Page 12, Table 3: would benefit from a vertical line separating RF-R, RF-SIF-R, and N, and other similar vertical line in the lower panel

The vertical lines separating RF-R, RF-SIF-R and N were added in table 3.

Page 15, Line 431: “it can be avoided”, maybe better phrased as “we don’t need to rely on land cover type....and meteorological data”?

Changed to: “we do not need to rely on land cover type and land cover change, and meteorological data.”

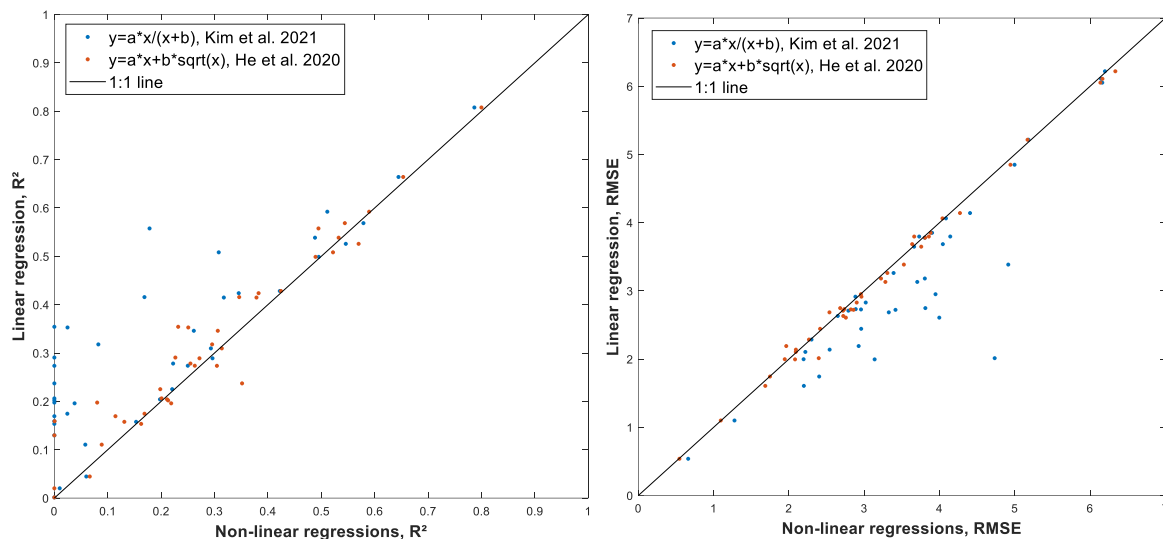
Dear reviewer,

We would like to thank you for taking your time to evaluate our work and foremostly for your interesting and useful comments, questions, remarks and suggestions.

We will try to answer your questions and we will integrate your suggestions into the manuscript (all answers are in blue color).

The authors claim there is a linear correlation between SIFd and GPP both in the site and the PFT levels. However, a quick look at the figures (2-4) shows that in most cases at some point the SIFd-GPP relation reaches saturation. The authors did not mention this even once in their manuscript. Several works are demonstrating this relation and discuss its meaning (see He et al., 2020 for example), however, the authors here ignore it and refer to it as a linear relation. Moreover, in many cities and PFT, the linear correlation is also low for the same reason.

While non-linear relationships have been shown at canopy scale (He et al. 2020, Kim et al. 2021), when using satellite data, the added noise and inherent linearization at larger scales (3.5*7.5 km in our case) makes it hard to fit non-linear model across a diverse set of sites. In our case, non-linear models do not show a clear improvement in performance, as shown in the following figure and table :



Site name	linear regression		Hyperbolic model (Kim et al. 2021)		linear+square root model (He et al. 2020)	
	R²	RMSE	R²	RMSE	R²	RMSE
BE-Bra	0.592	2.00	0.511	2.19	0.590	1.94
BE-Lcr	0.204	4.06	0.198	4.09	0.211	4.04
BE-Lon	0.274	6.05	0.250	6.16	0.263	6.14
BE-Vie	0.202	2.44	-0.160	2.96	0.213	2.41
CH-Dav	0.153	1.74	-0.597	2.40	0.163	1.75
CZ-BK1	0.354	3.38	-0.345	4.91	0.232	3.53
CZ-Lnz	0.416	3.18	0.169	3.80	0.346	3.22
CZ-Wet	0.196	3.68	0.038	4.04	0.219	3.64
DE-Geb	0.526	3.80	0.546	3.73	0.570	3.66
DE-Gri	0.278	3.26	0.222	3.39	0.255	3.31

DE-Hai	0.569	2.91	0.579	2.88	0.544	2.96
DE-HoH	0.346	4.14	0.261	4.40	0.306	4.27
DE-Hzd	0.415	2.73	0.317	2.96	0.379	2.82
DE-Kli	0.310	3.85	0.293	3.90	0.314	3.88
DE-RuR	0.174	3.80	0.024	4.14	0.169	3.86
DE-RuS	0.289	6.22	0.296	6.20	0.272	6.34
DE-RuW	0.159	2.61	-0.959	4.00	0.000	2.76
DE-Tha	0.558	1.61	0.178	2.20	0.495	1.68
DK-Sor	0.808	2.73	0.787	2.89	0.800	2.74
FI-Hyy	0.274	2.68	-0.111	3.33	0.304	2.54
FI-Sii	0.318	1.10	0.082	1.28	0.296	1.09
FI-Var	0.130	2.00	-1.146	3.14	-0.008	2.08
FR-Aur	0.225	3.65	0.221	3.66	0.198	3.76
FR-Bil	0.169	3.13	-0.160	3.70	0.114	3.28
FR-EM2	0.111	4.85	0.058	5.00	0.089	4.94
FR-Fon	0.664	2.71	0.645	2.79	0.653	2.72
FR-Hes	0.499	3.78	0.495	3.80	0.489	3.81
FR-LGt	0.428	2.29	0.422	2.30	0.424	2.27
FR-Mej	0.045	5.22	0.060	5.18	0.066	5.17
GF-Guy	0.020	2.63	0.010	2.65	-0.045	2.72
IT-BCi	0.158	6.11	0.001	6.67	0.131	6.16
IT-Cp2	0.001	2.01	-4.463	4.73	-0.397	2.40
IT-Lsn	0.538	2.10	0.488	2.22	0.532	2.09
IT-SR2	0.206	2.95	-0.991	4.68	0.201	2.96
IT-Tor	0.424	2.83	0.345	3.02	0.383	2.90
NL-Loo	0.237	2.19	-0.349	2.93	0.352	1.96
SE-Deg	0.353	0.54	0.025	0.66	0.251	0.54
SE-Htm	0.291	2.72	-0.114	3.41	0.227	2.86
SE-Nor	0.508	2.14	0.308	2.54	0.522	2.09
SE-Svb	0.198	2.75	-0.542	3.81	0.080	2.68

While the two tested non-linear models do not show significantly lower RMSE overall, they exhibit a strong instability, and cannot be accurately fitted on all sites. As our goal is to compare relationships between sites and considering the level of noise in the TROPOMI SIF data, we see no clear benefit in using a non-linear model which only brings in marginal improvement over a few sites at the expense of a loss in genericity.

I'm not sure what is the added value of the pooled graph of all PFTs vs. SIFd (Fig. 4)

Within this figure, we would like to evaluate the genericity of the relationship between GPP and SIFd across the study sites, which is demonstrated by the low R^2 value found on data pooled across all sites.

It is very hard to estimate the performance of the different models vs. EC GPP in Figure 8. Please consider reducing the size of the dots and making them transparent.

The size of the dots has been changed and we made them transparent.

Line 399: Please try to explain why the high correlation in the DBF and OSH PFT's

We briefly tried to explain this high correlation in line 399. The main explanation is that in DBF and OSH (one sample of vineyard plantation) biomes, there are explicitly marked seasonal and phenological changes compared

to EBF or ENF forest where there is greenness all time. Thus, in DBF and OSH biomes SIF signal may easily capture the variations in LAI and APAR and consequently display a high correlation between GPP and SIF_d (added to: line 405-408).

Line 410: This is not clear to me, the authors mention in the methods section that they took out the cloudy day data. Line 414: So, why not do that in your data?

The representativity of satellite SIF data which needs to be filtered for cloud coverage is indeed a limitation of the current study. This limitation is inherent to the use of satellite data, and can only be lifted through the collection of ground-based SIF data across diverse ecosystems.

Line 436: it is problematic to say there is a difference in the models while earlier you mentioned there was no statistical difference between them (line 296). Line 447: same comment as above.

The line 296 was indeed unclear. We edited it as follows:

“A paired t-test realized between the four models based on the adjusted R^2 performance revealed that the difference in mean **adjusted R^2** between RF-R and RF-SIF-R, RF-R and RF-SIF-R-PFT, and RF-SIF-R and RF-SIF-R-PFT models was not statistically significant. In other words, these three FR models have the same performance.”

We have reframed the line 436 as follows to clarify it:

“The current results show that the RF-R (surface spectral reflectance alone), RF-SIF-R (SIF_d plus surface spectral reflectance) and RF-SIF-R-PFT (SIF_d plus surface spectral reflectance plus PFT) models explain a non-significant different percentage of the variance in tower-based GPP at the daily time scale (82~86%), whereas the RF-SIF-VI (SIF_d plus reflectance based-indices) explains 75% of the interannual variabilities in GPP across all sites.”

As for line 447, it is based on a difference in relative importance rather than a difference in R^2 (see Figure 9). Unlike the differences in R^2 which aren't statistically different and cannot be interpreted, the relative importance depends on input variables and can be interpreted. This sentence was edited as follows to clarify it:

“SIF remains a better predictor of GPP than each reflectance band individually (Fig. 9).”