Reply to anonymous Referee#2

We thank the reviewer for the comments/suggestions that much improved the manuscript. We reply to referee#2 comments point by point.

Comment1:

"The study design introduces nutrient treatments which influence species composition GPP to some degree but the successive analyses discard this structure when testing the skill of reflectance metrics as predictors of GPP. GPP varies over time, across replicates and across treatments but all of this variability is pooled when testing the reflectance metrics. The paper might provide more insights by structuring the analysis to test treatment and temporal variability separately. Perhaps a repeated measures ANOVA could help, for example, by testing for significant effects of date, treatment, and one reflectance metric on variability in GPP. This would be akin to the linear mixed effect analysis of GPP and PAIgr".

Answer**:** The ability of reflectance metrics in predicting GPP were analyzed altogether without considering different treatments and dates because the objective of the study was to test the skill of spectral information to represent GPP across treatments and time. However no statistical differences were found in GPP among treatments (see section 3.3).

Comment2:

"GPP and PAIgr both vary over time and PAIgr varies across treatments (GPP appears to as well but apparently the statistical testing does not support this). These patterns are displayed well with Figs 2 and 3 but what is missing is display of a scatter of GPP versus PAIgr, and also of GPP versus (selected) reflectance metrics. These relationships should be shown, with symbols that differentiate the treatments and display individual replicates. The relationship that emerges (slope) would offer insights about the effective light use efficiency per unit green leaf area. The term 'effective' here refers to the combination of a maximal LUE with any limitations by water, light, or nutrients. This is the sort of parameterization that would be needed in a functional model. In fact, it would be interesting to test if any of the reflectance metrics have skill in predicting variability in GPP / PAIgr, thus capturing patterns in LUE rather than just green plant area."

Answer: We thank the referee for this valuable comment that improved our insight on the relationship between green plant area and GPP. The scatter plot produced is shown below.

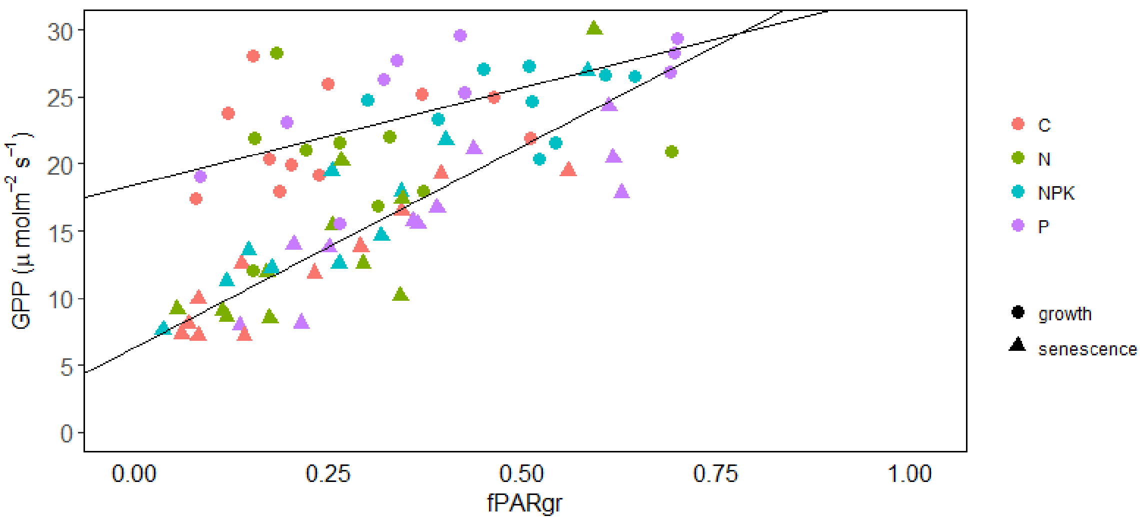

Figure caption: Relationship between GPP and fPARgr observed in grassland subjected to different fertilization treatment (C, N, NPK and P). Measurements performed during the vegetation growth are indicated as circles (corresponding to measurement days 1 and 2) while those realized during the senescence period are indicated as triangles (measurement days 3 and 4). Linear regression lines were fitted separately to the two periods (GPP= 14.48 *f*PARgr + 18.44, R^2=0.39, P<0.001 and GPP= 30.08 *f*PARgr + 6.39, R^2=0.65, P<0.001, respectively for the growth and the senescence period).

The GPP-fPARgr relationship showed to be different during the vegetation growth (circles in the figure) and the senescence phase (triangles in the figure).
The differences observed in the slope of the regression lines reveal the occurrence of marked changes in the "effective LUE" along the growth life cycle and confirms previous results (Nestola et al., 2016; Perez-Priego et al., 2015).
The figure will be integrated in the manuscript and the text modified accordingly.

In addition, we also verified the correlation between LUE and selected vegetation indices calculated (GNDVI, MTCI,NDWI, PSRI) (See table below). For all vegetation indices correlations are weak.

Table: Correlation coefficient (*R*) between LUE and selected vegetation indices observed in herbaceous plots undergoing different fertilization treatments.

VI	*R*
GNDVI	-0.145
MTCI	-0.076
NDWI	-0.214
PSRI	-0.022

Altogether these results further confirm the need for an empirical approach, instead of a classic LUE model, to assess the ability of spectral retrievals for GPP estimation.
We agree that the approach has limited applicability as compared to a full functional model. On the other hand it allows testing different spectral metrics, independently of their ability to represent green plant area or LUE.

Comment3:

"Table 5 shows the skill of various multivariate linear models that include a suite of reflectance metrics selected to represent those available from different observing system types. This is a highly empirical approach to analysis and does not seem particularly useful in my opinion. The results are likely to be very heavily tuned to the specific dataset on hand and is not likely to be generalizable beyond the current study. For example, the Hyp-B step one selection includes a simple, linear model involving 13 unique bands. Biophysical or ecological functional models tend to use one or two metrics to represent structural (PAIgr) and functional (LUE) attributes of an ecosystem's capacity for primary production. This paper's approach throws every possible indicator and combination at the variability in the data and thus lends little practical insight into the theory with very limited capacity for transferability. A more thoughtful approach grounded in theory and practice would be more useful."

Answer: We agree with the referee that our results are tuned to our data set and results cannot be generalized. However, our dataset concerns herbaceous communities with a wide range of species composition and growth rates, resulting from the fertilization treatments. Hence, our empirical results can be considered valid for Mediterranean grasslands in a wide variety of situations. Also, the theoretical approach based on the LUE model already showed, in several studies (examples are reported also in the manuscript introduction), a limited capacity in expressing dynamic GPP changes in grasslands. This is also confirmed by the GPP-*f*PARgr relationships and the low correlation between LUE and VIs observed (see answer to comment 2).

We agree with the reviewer that considering reflectance bands, obtained by grouping highly correlated reflectance can be impractical but, in our opinion, it presents the advantage of indicating areas of the spectra of potential interest for GPP estimates in grasslands. For example, our results show a large influence of bands in the SWIR region in both step one and two (see Table 5). These results provide basis for further studies in the area.

Comment4:

"The study's test of linear models includes VIs and bands, but not band ratios. Given that the approach is highly empirical in nature, there does not seem to be a good reason to omit band ratios or other simple mathematical combinations of bands (e.g. unique normalized difference ratios). Testing a wider range of combinations could be warranted to see if any other indicators happen to rise to the top in terms of predictive skill."

Answer**:** As explained in the manuscript, we considered VIs that are widely used. We didn't want to expand even more the set of explanatory variables, which was already large, to be able to interpret our results in terms of simple bands and well-known indices, while considering a large pool of predictors. Moreover, there is in general a 1 to 1 relation between VIs and band ratios. For instance, NIR/RED=k implies that NDVI=(k-1)/(k+1). Therefore, band ratios are implicitly incorporated into the model (although this is not equivalent to incorporating them as additional explanatory variables).

Comment5:

"The paper's interpretations and conclusions suggest that bands are better than VIs as predictors of GPP but this is not reasonably supported by the quantitative results. Table 5 shows a small, marginal, and questionable increase in adjusted R2 for Hyp-B compared to Hyp-VIs, and a decrease in adjusted R2 for S2-B compared to S2-VIs. In any case, the differences in explanatory power over all of these cases is less than 0.0247 R2, or 2.5% of the variability in GPP, indicating that all are equally good at predicting GPP. For L8, a case might be made, however the band metric has many more variables thrown at the problem (6 bands compared to just NDVI) and when these other bands are included in a step two selection, the NDVI model with bands had high skill. Surely bands and VIs are equally skillful for the other observing system types. Corresponding edits need to be made to section 4.2."

Answer**:** We thank the referee for this thoughtful comment on the discussion of results shown in table 5. The text was changed accordingly. The modified text is:

4.2 Are spectral bands better GPP estimators than VIs?

Our results suggest a marginal improvement in GPP estimates obtained adopting bands (Hyp-B, S2-B, Table 5) instead of vegetation indices (Hyp-VIs, S2-VIs, Table 5). A larger impact was observed in the case of L8-VIs+B models when compared with the L8-VIs model which included only NDVI (L8-VIs, Table 5). Although normalized VIs are important in establishing strong relationships between biophysical and optical properties of vegetation, our results showed that the selection of the proper band is equally important to the mathematical formulation of the indices for the explanatory power of spectral retrievals as predictor variables. Previous studies comparing the explanatory power of VIs and bands in grasslands evidenced also the importance of the selection of the proper spectral range (Balzarolo et al., 2015; Matthes et al., 2015). The approach adopted in our analysis assured a high correlation among responses within a band, which determined that spectral bands were representative.

Comment6:

"One of the advantages of VIs is that they normalize for a wide range of background reflectance, sun-sensor geometry, and atmospheric effects in ways that direct bands do not. This point seems to be lost on the authors and is important for developing indicators that can be transferred to remote sensing (space or airplane) over large areas and across large gradients in surface and atmospheric conditions. Discussion about this should be included in the paper."

Answer: We thank the referee for this valuable comment. The following sentence was included in the text.

However, VIs offer, in comparison with spectral bands, the advantage of providing spatial and temporal comparable representation of vegetation features, since differences resulting from background reflectance, sun-sensor geometry and atmospheric effects are minimized by normalization of spectral values.

Comment7:

"Akaike or Bayesian Information Criteria need to be adopted to evaluate the relative skill of the selected linear models, penalizing models that select more variables."

Answer**:** We agree that models should be compared with indicators that penalize complex models and prevent overfitting. This is why we used the adjusted R^2 instead of the standard R^2. The optimization techniques we considered (LEAPS in particular) can be applied using adjusted R^2 (as we did), AIC, BIC or other similar criteria. It is easy to show that stepwise selection always chooses, at each step, the same variable to be excluded or included in the model regardless of the criteria (AIC, BIC or adjusted R^2). In that restricted sense, they are equivalent among them but distinct from R^2. Our preference for the adjusted R^2 over BIC or AIC was determined by the fact that we applied bootstrap techniques (available for the adjusted R^2) to derive confidence intervals for the goodness-of-fit of the models we fitted to the data.

Comment8:

"It is worth noting that soil moisture is essentially equal on all four reflectance observation dates, while temperature increased steadily from the first to the last observation date. Correspondingly, the statement on P12, L19 that suggests that the Hyp model represents changes in canopy water content might need to be revised. Canopy water content was not observed and soil water content did not differ over the four sampling dates. It is possible that canopy water content differed substantially from soil water content over this time series but that has not been established with quantitative, direct observations."

Answer: The aim of the mentioned test is not to establish a quantitative relationship between canopy water content and the vegetation indices NDWI and WBI but to underline the known ability of these vegetation indices as indicators of canopy water content. On the other side, differences observed between fPAR and fPARgr reveal a progressive increase of dry biomass, which is related with the senescence process that depends more on increased temperatures than soil water content. The sentence was changed into:

The Hyp model also put in evidence the importance of changes in canopy water content, as both the NDWI (Gao, 1996) and the WBI (Penuelas et al., 1997) were included in the model. Considering changes observed in NDWI along the experiment and the good correlation observed between NDWI or WBI and GPP it is reasonable to assume that GPP is largely affected by the progressive senescence of vegetation, (Balzarolo et al., 2015; Vescovo et al., 2012). In a previous study, Vicca (et al., 2016) found that NDWI was able to estimate GPP in semiarid grasslands better than other indices, allowing to distinguish the effect of drought.

Comment9:

"It is unfortunate that the study did not include an additional observation period in the mid to late June as PAIgr continued to decline."

Answer**:** After our last measurement PAIgr declined very fast, in mid June all plots were already dried. There was no time for a further round of measurements.

Comment10:

"The introduction is very well written and cited. One paper that might be useful to add to the framing and discussion is that of Asner et al. 2004 in PNAS ("Drought stress and carbon uptake in an Amazon forest :")".

Answer**:** The following sentence was inserted in the introduction.

A promising new technology is the use of spaceborne imaging spectroscopy. The hyperspectral resolution of these images allows to identify canopy properties with higher sensitivity than traditional vegetation indices. For example, the use of spaceborne imaging spectroscopy was able to detect changes in canopy leaf area and water stress in the humid tropical forest, while NDVI and other vegetation indices failed (Asner et al., 2004).

Comment11:

"Page 5, L12: "All nutrients were added at" seems to suggest N for nitrogen, or is N for nutrients here?"

Answer: The referee is right; it should be "All nutrients were added at a rate of 10 g. m^{-2} yr^{-1}…". In the text N refers to nitrogen. The sentence was corrected.

Comment12:

"Measurement of soil moisture at only 10 cm depth may not be adequate to represent the soil water content being experienced by the grassland plants. It would be best to also measure a deeper profile of moisture."

Answer: We agree with the referee that the soil water measurements along a soil profile would have provided more detailed information. However our grassland is dominated by annual species with winter seasonality. In these systems mean rooting depth is approximately 20 cm (Schenk and Jackson, 2002) with most of the rooting zone at 10 cm depth as observed in a similar grassland in Portugal (Jongen et al., 2013).

Comment13:

"It is interesting that the nutrient treatments allegedly altered the functional composition of the grassland plots, however pre-treatment data are not presented and this would be essential to demonstrate that the compositional shifts were due to the treatments themselves. Unless it can be established with data, the corresponding statement (P9, L23) should be corrected to omit suggestion that the treatments caused the compositional differences."

Answer**:** As mentioned in the manuscript, the grassland is part of the Nutrient Network experiment (www.nutnet.umn.edu) which is still on going. Since 2012 (pre-treatment), species composition is evaluated every year. In 2012 species composition showed a large dominance of forbs (73%), while graminoids (24%) and legumes (3%) were less represented. These results are only partially similar to the observed in the present study (Table 3) in which forbs were the dominant functional type (56.85%), and grams and legumes have similar proportion (21.22% and 21.93% respectively). However, it must be also considered that species composition can change from year to year as a result of precipitation amount and distribution as observed also in previous studies (DeMalach et al., 2017). The sentence was rephrased into:

Plant species composition has been measured every year since 2012 (pre-treatment) under an ongoing long-term nutrient addition experiment on this grassland site. In line with results from that study (Nogueira et al., personal communication), the fertilization treatments influenced the functional composition of grasslands (Table 3). In the NPK treatment the percentage of graminoids was higher than in any of the other treatments.

P treatment showed a higher percentage of legumes and in the C and N treatments forbs were the dominant functional group.

Comment14:

“It is surprising that treatment effect was significant for respiration but not for GPP considering that both have similar spreads and error bars. Double check results here.”

Answer**:** Results were double checked and they are correct.