Reviewer 1

We sincerely thank reviewer 1 for taking the time to thoroughly read our manuscript and provide constructive feedback that has helped improve the quality of the study. We have provided a point-by-point response to their comments below.

This manuscript evaluates the ability of multiple MODIS-derived vegetation indices (VIs) and related satellite variables to capture drought-induced declines in forest gross primary productivity (GPP) across 18 European forest flux sites. By comparing the timing and magnitude of drought responses in GPP and 12 VIs, the authors show that LST most consistently tracks drought impacts across both broadleaf and coniferous forests, while structural indices such as NDVI, PPI, and NIRv often exhibit delayed or persistent post-drought responses. The topic is interesting, and the manuscript combines flux observations and remote sensing data in a potentially valuable way. However, several issues currently limit the robustness of the conclusions, I therefore recommend a **Major Revision**. Detailed suggestions follow below:

1. **Section 2.2:**

Most of the VIs used in this study are derived from a single MODIS product family (MCD43A4), which raises the concern that the reported conclusions may be strongly influenced by the characteristics of a specific data source. The authors are encouraged to include an additional MODIS dataset to verify the robustness of the results. In addition, the LST data were not directly derived from standard MODIS LST products (e.g., MOD11A2) but from the FluxnetEO database. Therefore, the superior performance of LST reported in this study may partly reflect the quality or processing advantages of the FluxnetEO product rather than the intrinsic advantage of LST itself. This issue should be clarified and discussed more carefully.

Thank you for this suggestion. We agree that product-dependence is an important issue, however our objective was not to compare MODIS product families but to evaluate drought responses in indices using the most consistent reflectance data available. We used the MCD43A4 product, which is the Nadir Bidirectional Reflectance Distribution Function (BRDF)-Adjusted Reflectance (NBAR) dataset, because it corrects for the effects of differences in view angle and directional reflectance between images by providing data at nadir. This is an important correction because MODIS has a wide swath with viewing angles that differ significantly from nadir which can introduce significant variation into VI time series. It is the highest quality MODIS product available for the bands of interest. We have now clarified this choice and acknowledge that product-specific processing may influence the results by adding the text below to the discussion.

We decided to use the LST data from the FluxnetEO database for a similar reason, because it has undergone a correction to provide LST data at nadir, which thus makes it more stable over time and more comparable to the MCD43A4 data. We have now clarified these choices and acknowledge that product-specific processing may influence the results by adding the following text to the methods and conclusions.

l.127 *"We opted to use the MCD43A4 product because it provides reflectances at nadir corrected for changes in viewing angle."*
l.132 *"Similar to the other VIs, we opted to use this product because it provided LST at nadir."*
l. 602 *"Future work should also test whether the relative performance of the VIs we tested remain consistent when using alternative MODIS or Sentinel products."*

We do agree that the superior performance of LST may be partly due to the additional quality control and gapfilling used in the FluxnetEO product. However, this is unlikely to be the main reason that LST performed better than the other indices, as there are other studies indicating the importance of LST for modelling GPP (e.g. Wu et al 2010) and its superiority in responding to drought over traditional greenness indices (e.g. Hoek van Dijk et al 2023). Also from a mechanistic point of view, there are good arguments that LST is a direct indicator of drought response. See line 454: "These results are supported by plant physiology experiments showing that stomatal closure is the first drought response of C3 plants that limits photosynthesis in the early phase of drought (Flexas and Medrano, 2002) [and increases LST]. Reductions in leaf chlorophyll to carotenoid ratios, which indicate the downregulation of photosynthesis due to stress, then occurred within 3 to 12 days after GPP Datestart (Fig. 3a)." We have added text to the discussion to clarify this point as follows:

l.470 *"Although the superior performance of LST over the VIs may partly be due to the additional quality filtering, nadir correction and gapfilling that was included in the FluxnetEO product, it is unlikely to be the main reason [...] Our results are supported by Hoek Van Dijke et al. (2023) who found that LST increase during drought matched the timing of stomatal conductance decline due to drought at forest sites better than NDVI, EVI or NIRv."*

Hoek van Dijke et al., 2023 Comparing forest and grassland drought responses inferred from eddy covariance and Earth observation, *Agricultural and Forest Meteorology,* doi: 10.1016/j.agrformet.2023.109635

Wu et al., 2010. Comparison of multiple models for estimating gross primary production using MODIS and eddy covariance data in Harvard Forest, *Remote Sensing of Environment,* doi: 10.1016/j.rse.2010.07.012

2. **Section 2.2 and 2.3:**

The preprocessing procedures for the MODIS data should be described in greater detail. For example, the manuscript does not clearly explain how cloud contamination was removed or corrected. Similarly, the criteria used to define "good quality" flux observations (Line 131) should also be explicitly described.

Thank you for pointing this out. We have moved Table A1 into the main text (now Table 2). This table describes how the quality control layers provided with the MODIS bands were used to filter the data to remove clouds or bad quality data points. We also updated the table to clarify that the MCD43A4 product is a cloud-free product (i.e. cloud removal occurs during product creation). We have added the following sentence to clarify the satellite data filtering process:

l. 137 *"The data were filtered to remove clouds and low quality data points using the provided quality control layers as described in Table 2."*

Table 2. Quality filtering information for each satellite variable

Variable	Quality filtering
LAI, FPAR	Bit 0 = 0 (only good quality data), bits 3-4 = 0 (no clouds present), bits 5-7 <= 1 (main RT method used with or without saturation)
PRI10, CCI	Status_QA bits 0-2 = 1 (no clouds), bits 5-7 = 0 | 3 (adjacency mask normal/clear or adjacent to single cloudy pixel), bit 8 = 0 (Aerosol optical depth level is low). Cosine view zenith angle > 0.7071
NDVI, EVI, EVI2, NIRv, NDWI, NMDI	BRDF_Albedo_Band_Quality < 3 (only good quality inversions) for each band separately, Snow_BRDF_Albedo = 0 (snow-free pixels only), BRDF_Albedo_LandWaterType = 1-3 (land cover is land, coastlines or shallow inland water). MCD43A4 is a cloud-free product
PPI	BRDF_Albedo_Band_Quality < 3 (only good quality inversions) for Red and NIR bands. MCD43A4 is a cloud-free product
LST	See Walther et al. (2022) for details

For the flux observations, we have edited the text to clarify that we used the NEE_VUT_REF_QC quality control flag which indicates the percentage of daily observations that were measured or had good quality gapfill. This is a standard quality control flag that is used for flux measurement processing. The text now reads:

l. 151 *"The daily nighttime partitioned GPP (GPP_NT_VUT_REF) was used and filtered to exclude days which contained less than 50% measured or good quality gapfill data according to NEE_VUT_REF_QC (i.e. the quality control flag for the net ecosystem exchange from which GPP_NT_VUT_REF is modelled)."*

3. **Section 2.5:**

In Line 155, the removal of highly collinear variables should be conducted using more standardized approaches, such as variance inflation factor (VIF) analysis, rather than relying solely on pairwise correlation coefficients.

Thank you for the comment. VIF analyses are commonly used in multiple linear regression analyses since they can test for collinearity among multiple dependent variables at once. However, our analysis does not perform any modelling using more than one VI at once as dependent variables, they are always modelled separately. Therefore, we think that a variance inflation factor analysis would not support the analysis as suggested.

4. **Section 3.1:**

I found the interpretation of Figures 3 and 4 somewhat confusing. For example, Figure 3 suggests that NDVI Datestart occurs approximately 12 days earlier than GPP Datestart, whereas Figure 4 indicates that NDVI occurs about 3 days earlier than SPEI and GPP about 2 days earlier than SPEI, implying only a ~1 day difference between NDVI and GPP. The relationship between these two figures and the interpretation should therefore be clarified more carefully.

Thank you for pointing this out. We agree that the interpretation of the Figures is confusing, so we have added the following text to clarify the figure interpretation to section 3.1:

l. 326 *"We note that the difference (in number of days) between a VI and GPP $Date_{start}$ (or $Date_{min}$ or $Date_{end}$) will be different in Figure 4 compared to Figure 3. This is because GPP responds to all drought periods (as per the study design) whereas all the other VIs except LST respond to a smaller number of droughts (Figure 3g and h). Therefore, the average time differences shown for GPP in Figure 4 are calculated for a larger sample size than was used when comparing* *responses between GPP and each VI in Figure 3. Hence the difference in the timing between a VI and GPP should not be compared between Figures 3 and 4*."

5. **Section 3.2:**

In Fig.5, drought periods are effectively defined based on reductions in GPP, and the same periods are then used to evaluate the responses of other VIs. This framework may not provide a fully fair comparison among variables. It would be more appropriate to define drought periods independently using SPEI and then compare the responses of GPP and all VIs within the same externally defined drought window.

Thank you for pointing this out. We chose to focus on drought periods that have a negative impact on GPP as suggested by Reichstein et al., 2013 who emphasized the importance of defining extreme events not only based on meteorological parameters but also based on their impact on the ecosystem. A similar approach has been used in other studies assessing drought impacts on GPP such as Yu et al., 2022.

In our study, we focused specifically on droughts that negatively affect GPP to help narrow the scope of the paper. Drought can also cause no impact or even a positive impact on GPP but including these types of drought in the analysis would have led to confusing results since they would likely cause no change or a positive change in VIs during drought, which is the opposite to what we were expecting when drought negatively impacts GPP. Focusing only on periods when GPP was negatively affected allowed us to examine the time lags between GPP and VI response to drought and thus help identify why some VIs are not able to capture GPP declines due to drought. The ultimate goal with this type of study is to improve our ability to model GPP, not SPEI, so it is important to assess how the VIs behave relative to GPP. We interpret this analysis as testing whether VIs capture periods of observed GPP decline during meteorological drought, rather than as an independent test of VI-based drought-detection.

We have added text to the manuscript in Section 2.6 to clarify this point:

l. 190 "*We chose to focus on drought periods that have a noticeable negative impact on GPP as suggested by Reichstein et al., (2013) and use a similar definition as previous studies such as Yu et al., 2022. This was also important to help narrow the focus of this study since drought can also cause no impact or even a positive impact on GPP which would lead to different responses in the VIs. Focusing only on periods when GPP was negatively affected allowed us to examine the time lags between GPP and VI response to drought and thus help identify why some VIs weren't able to capture GPP declines due to drought.*"

Reichstein et al., 2013. Climate extremes and the carbon cycle, *Nature,* doi: 10.1038/nature12350

Yu et al., 2022. Contrasting drought legacy effects on gross primary productivity in a mixed versus pure beech forest. *Biogeosciences,* doi: /10.5194/bg-19-4315-2022

6. **Section 3.3:**

In Fig.7, the interpretation currently relies largely on visual comparison. More rigorous statistical significance testing should be included to support claims regarding post-drought differences and potential decoupling between GPP and VIs.

Thank you for the comment. As stated in l. 297, we performed paired t-tests (or Wilcoxon signed rank tests if the data were not normally distributed) to test for significant differences between reference post-drought years. However, none of the tests showed any significant differences, as stated on l. 410. Due to the small sample sizes (13 coniferous pairs and 6 deciduous pairs) we cannot perform any more complicated statistical significance testing. We have therefore toned down the language used to describe these results in the abstract, discussion section and conclusion to highlight that our statistical tests didn't give any significant results. We have edited the text to read as follows:

Abstract:
l. 24 *"VIs related to canopy structure (NDVI, Plant Phenology Index and NIRv) remained low in the first year following drought at both broadleaf and coniferous sites (although not statistically significant), while GPP often recovered to long-term mean values, implying possible post-drought decoupling between GPP and these VIs."*

Discussion, Section 4.5:
l. 562 *"We observed that several VIs commonly used to model GPP (NDVI, PPI, NIRv) showed noticeable but not statistically significant post-drought legacy effects that were not reflected in GPP (Fig. 7). These results highlight that decoupling between GPP and VIs may occur post-drought."*

Conclusion:
l. 597 *"We also found noticeable but not statistically significant drought-legacy effects in all the greenness-related VIs that are commonly used to model GPP (NDVI, PPI and NIRv), that are not reflected in flux tower GPP at all sites. The decoupling between GPP and these VIs could lead to underestimation of GPP and the resilience of the forest carbon sink during the post-drought years at some sites"*

7. **Discussion:**

The Discussion section should be reorganized. At present, much of the discussion mainly repeats the Results rather than providing deeper interpretation, synthesis, or broader implications. In addition, given that different VIs exhibit contrasting drought responses and recovery patterns, it would be more effective to compare and discuss the different VIs together within a Section, rather than discussing each index separately.

Thank you for this suggestion. We have removed text that was repeating the results on lines 431-433, 442-443, 513-514, 561 and 565-570 to ensure the discussion is focused on deeper interpretation of the results. We also want to highlight that VIs are already compared within the different sections of the discussion, e.g. Section 4.4 compares and contrasts the behaviour of NDVI, PPI and NIRv during drought. By dividing the discussion into sections based on the VI categories, we have provided a logical structure that helps break up the text. It also creates a framework to link the behaviour of each group of VIs to physiological processes and other studies to help provide an explanation for our results. We carefully considered alternative structures but concluded that grouping the Discussion by VI response category provides the clearest link between spectral behaviour, physiological interpretation and implications for GPP modelling