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

Reviewer 2

We are deeply grateful to reviewer 2 for taking the time to read our manuscript, their kind words on the importance of the study and the very thoughtful comments they provided that have helped tighten up the text and analysis. We have provided a point-by-point response to their comments below.

General Comments

Previous studies have shown that satellite-driven models of gross primary productivity (GPP) often underestimate the observed response of GPP to drought conditions. This likely is because the vegetation indices (VIs) and products that the models are using may not be responding to drought as expected. This study evaluates broadleaf and coniferous forest responses to recent droughts in Europe, comparing the response of GPP from 18 flux tower sites to the responses of a suite of MODIS products (mostly VIs). The authors find that many of the VIs (broadly defined here to also refer to other satellite remote sensing products) commonly used in GPP models (e.g., FPAR) do not respond sufficiently to droughts (defined by the Standardized Precipitation Evapotranspiration Index at 3-month span, SPEI-3) or inconsistently at best. They find that land surface temperature consistently responds to drought periods, which may be as expected, but that it would be worthwhile to include in GPP models. Furthermore, other vegetation indices such as the older and simpler NDVI have more consistent responses to droughts.

Initially, I thought this study was retreading old ground, but instead it appears to do a rather nice job at succinctly trying to diagnose why GPP models may not be getting the response to drought in temperate forests quite right. It could use some revision to make its main points clearer and tighten up the analysis, but I believe it will be an important contribution to the GPP literature once edited. I suspect it may be contentious by some in the community, especially given its relatively limited analysis compared to the application scope of global GPP models, but I think that it is a needed case study to kickstart discussions on how to model GPP better using satellite data.

Specific Comments

1. This study appears to treat all SPEI-3 droughts the same, and is attempting to find if there is an effect of drought on vegetation indices regardless of drought severity. Was there any ranking or hierarchy in terms of the severity of drought (how negative SPEI-3 is) and whether the effect of drought manifested in a particular VI? There may be a threshold sorting in which drought may always develop in some variable (LST) but less frequently in others (NDVI, NIRv) depending on the intensity of the event. There is some brief discussion related to drought severity at line 320, comparing broadleaf and needleleaf sites, but this could be expanded.

Thank you for this suggestion. We did not add a drought severity term to the original mixed models used to assess whether each VI responded to drought. This was due to the small sample size which precluded the use of any additional variables in the model (otherwise the models would not converge). Similarly, we tried fitting another set of mixed models with average VI z-scores during drought as the independent variable and minimum SPEI during drought and forest type as dependent variables but the random effects could not be calculated, likely due to the lack of close relationships and the small sample size. Therefore, we resorted to performing a simpler analysis and calculated the spearman's rank correlation coefficients between minimum SPEI during drought and average VI z-score during drought to assess the relationship between drought severity and the VIs. However, we did not observe any threshold sorting depending on the severity of the drought and the correlations between mean VI z-score and minimum SPEI were generally weak. We have updated the method to describe the new analysis:

l.286 *"We also assessed whether there was an effect of drought severity on mean VI z-score during drought by calculating spearman's rank correlation coefficients for each VI versus minimum SPEI during drought (using drought dates based on GPP $Date_{start}$ and $Date_{end}$). Correlations were calculated separately for broadleaf and coniferous sites, using all same drought periods as included in the mixed effects modelling analysis."*

Since the results mostly showed weak correlation between the VIs and drought severity, we added the new Figure to the Appendix as Figure A2:

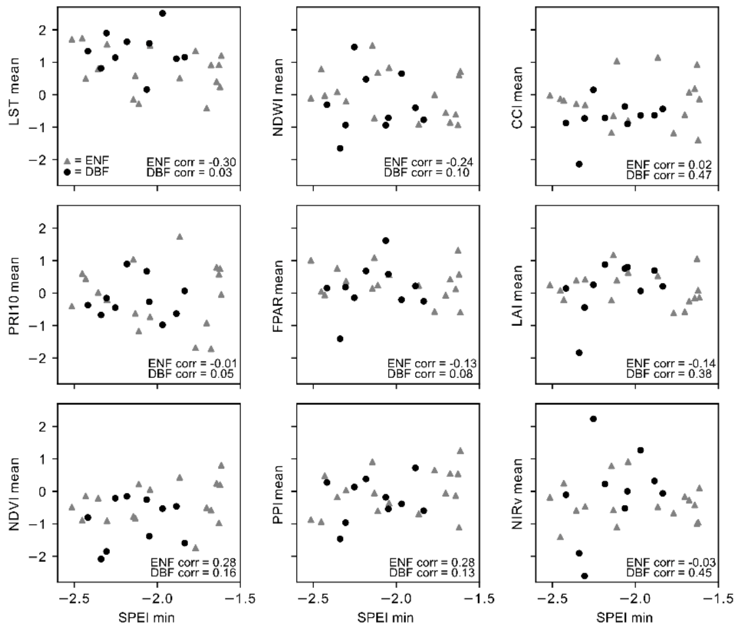


Figure A2. Mean VI z-score during drought versus minimum SPEI during drought for coniferous and broadleaf sites, with spearman's rank correlation coefficients for each forest type.

We also added text to the results section describing the new figure:

l. 371 *"Correlations between mean VI z-score during drought and drought severity (minimum SPEI) were weak for most VIs except LST at coniferous sites (spearman's correlation coefficient = -0.3) and CCI, LAI and NIRv at broadleaf sites (0.47, 0.38 and 0.45, respectively; Fig. A2). We did not observe any thresholds in VI behaviour in response to drought severity (Fig. A2)."*

Finally, we added text to the discussion section describing the new results:

l.446 *"The large variability (i.e. large error bars in Fig. 3) in drought response timing for each VI as well as the mostly weak relationships between VIs and drought severity (Fig. A2) we observed were likely driven by the highly variable sites and drought conditions we compared (e.g. differences in topography, aridity, mean annual temperature and precipitation, soil type, soil water storage, tree species, tree age as well as drought length, severity and timing)."*

2. Last comment in Abstract states: "Remote sensing-based GPP models based on these structural indices alone may thus overestimate drought impacts on GPP and underestimate forest resilience to drought." Most GPP models include at least some form of meteorological data in additional to veg structure metrics. Anything specific? This could be directly tested (see comments below.)

Thank you for the comment, we agree that most GPP models include some form of meteorological data so we have removed this sentence from the abstract.

3. Any particular reason MODIS GPP (MOD17) or other MODIS-derived GPP products like FLUXSAT were not evaluated? This analysis does not need to be done here since it has been evaluated in many previous studies (like Stocker et al 2019, as mentioned) on model development, but it does need to be carefully stated (likely something related to "how are drought impacts being expressed in forests in Europe that are measurable from satellite remote sensing?")

Thank you for the suggestion, we have added text to justify our decision to the end of the introduction as follows:

l. 108 *"We analysed the behaviour of the VIs that are used as the basis for many commonly used GPP models (e.g. MOD17 and FLUXSAT) rather than the models themselves. This allowed us to analyse to what extent drought impacts on European forests are measurable purely from satellite remote sensing."*

4. In addition, why MODIS and not Sentinel-2? Is this to keep a focus on global products? Just need more of an explanation in the Discussion.

Thank you for the suggestion, we have added text to justify this decision in Section 2.2. of the methods:

l. 127 *"We opted to use MODIS instead of Sentinel-2 due to its much longer archive that allowed us to analyse more different drought periods and ensure that enough sites had drought, reference and post-drought periods that met our criteria. In addition, Cai et al., 2021 showed that forest GPP models were similarly accurate when based on MODIS or Sentinel-2 data."*

5. Line 411: There is some circularity in LST being predictive of SPEI-defined drought, since the SPEI's PET is derived from modeled temperature and solar radiation variables in ERA5-Land that would directly covary with LST in most situations. This should be noted, but I also agree and think it is a good suggestion that LST should be included in GPP models for drought relationships.

Thank you for pointing this out, we have added text regarding this point to the discussion:

l. 472 *"We also note that the high response rate of LST to SPEI-defined drought is not unexpected as SPEI is calculated using temperature and solar radiation variables that likely co-vary with LST"*

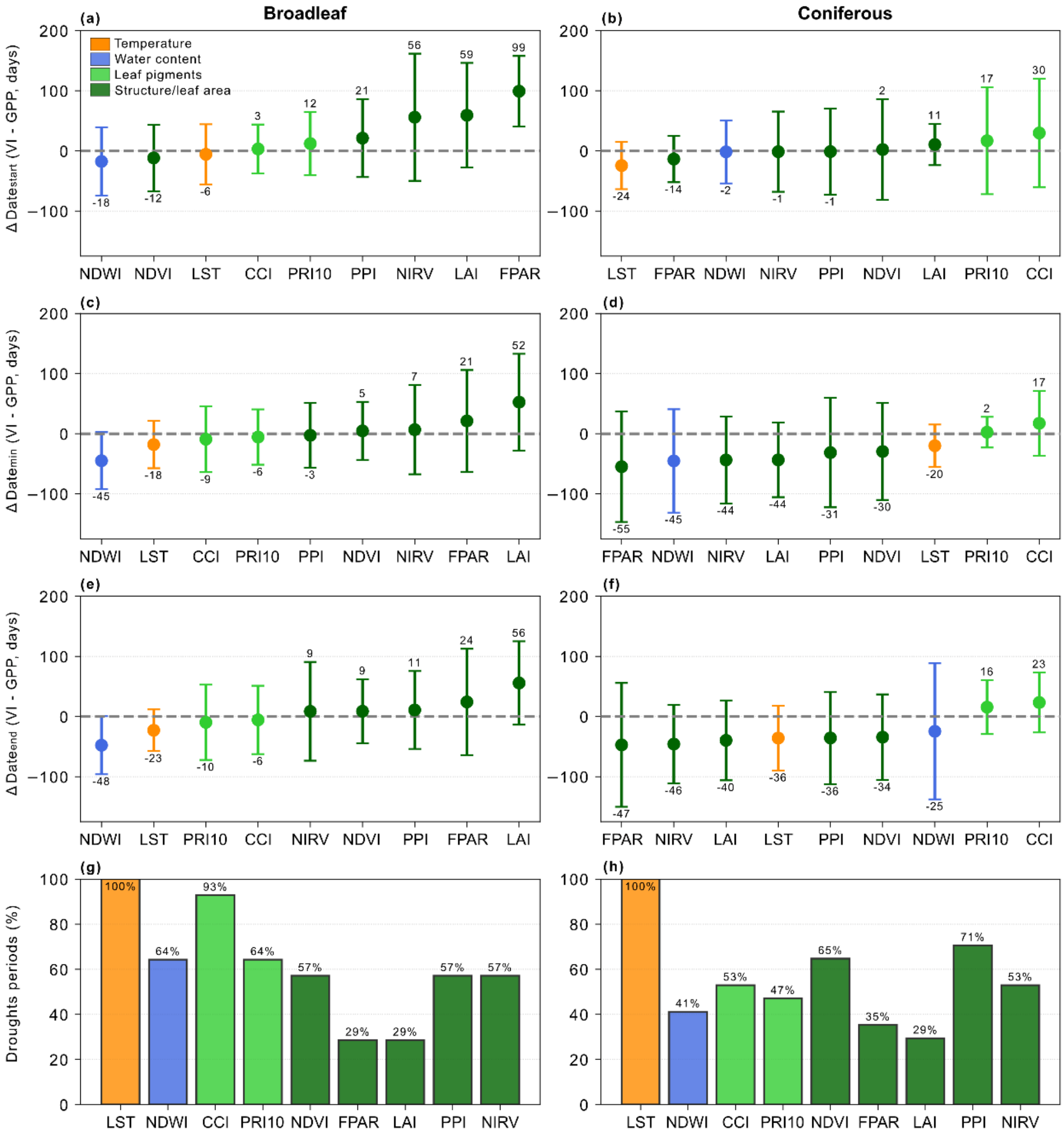
6. Line 474: “We suggest that these indices should not be used to capture drought effects on GPP.” This is a big suggestion since FPAR is used for the MODIS GPP product. Would you be willing to suggest that NDVI would be a better variable to use instead of FPAR in this product?

Thank you for pointing this out. We agree that the original wording was too broad given that FPAR is a core input to the MODIS GPP model. Our intention was not to suggest that FPAR should be removed from MOD17, not that NDVI should replace FPAR. MOD17 uses FPAR within a light use efficiency model that includes meteorological constraints. Therefore FPAR is not expected to represent drought stress alone. We have revised the text to clarify that our result concerns the use of MODIS LAI/FPAR as stand-alone indicators of drought-related GPP reductions.

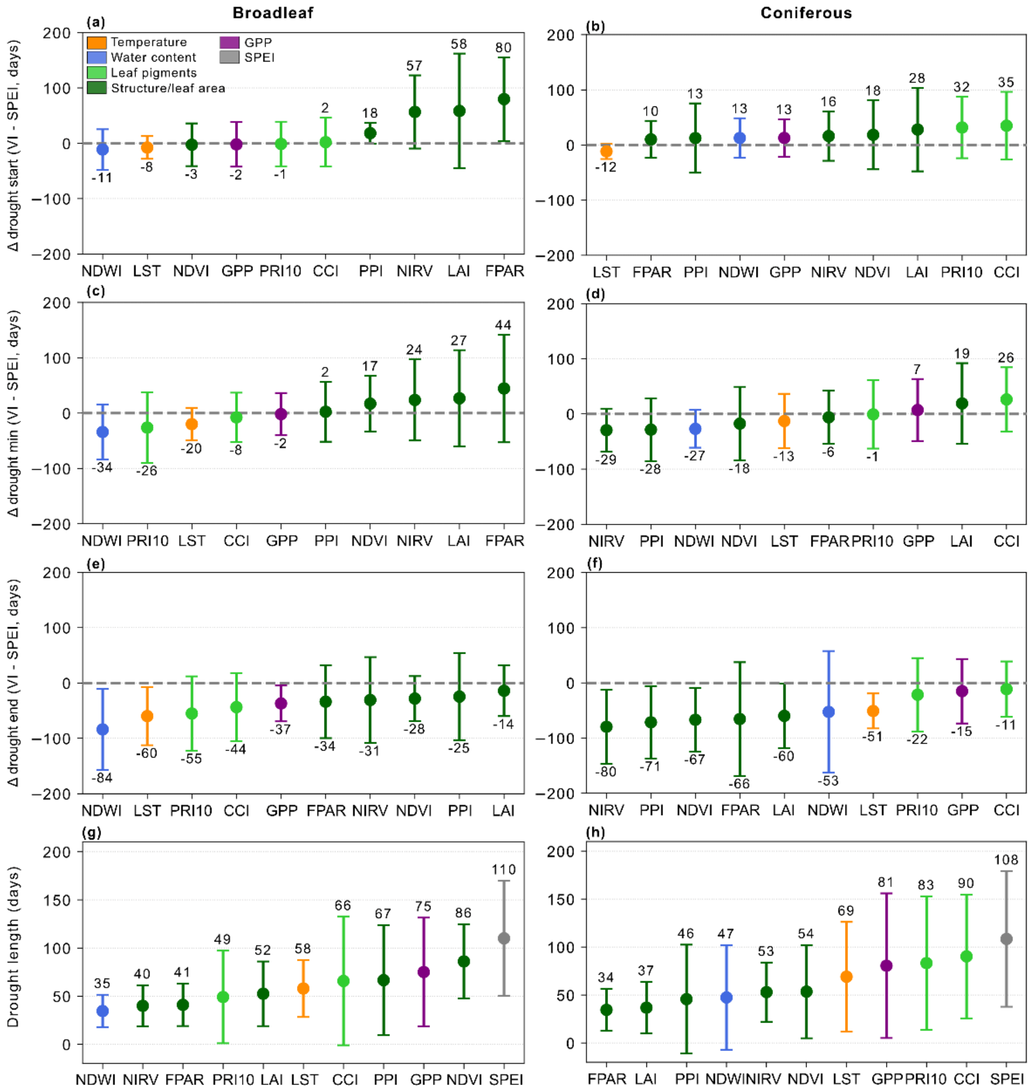
l. 534 *“Our results suggest that MODIS LAI and FPAR should not be used alone as indicators of drought-related GPP reductions in forests and must be combined with other VIs or meteorological constraints.”*

7. I would change Figure 3a-f and Figure 4a-f into box plots or into points for the mean with SD ranges. The very small bars, which the dates that are actually close to the measured GPP and SPEI dates, are what should be the focus of the plots (in my view), and the small size of the bars detracts from this being the emphasis.

Thank you for the suggestion. We have changed Figures 3a-f and 4a-h as suggested:



Updated Figure 3



Updated Figure 4

Technical Comments

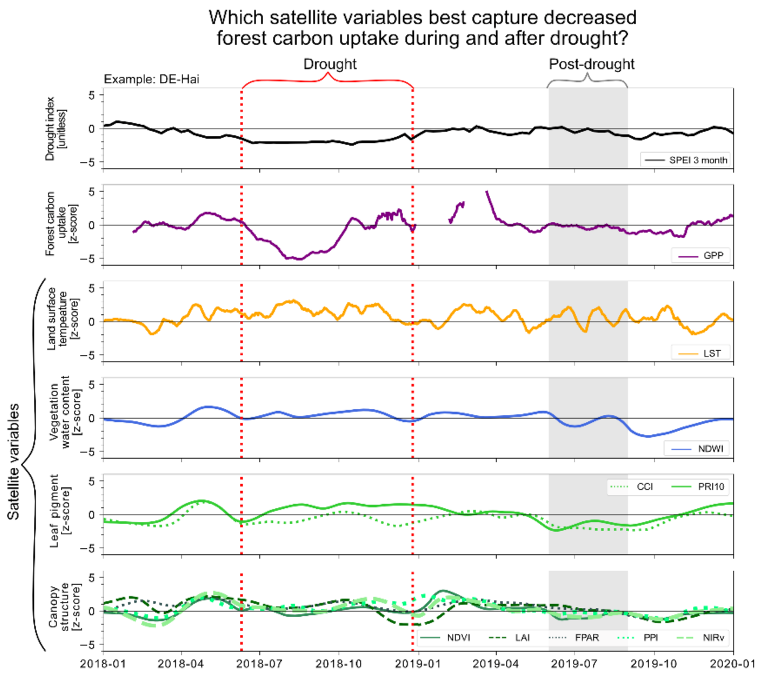
1. Abstract needs to be more explicit about where the GPP numbers are coming from, please state that they are coming from eddy covariance sites using data from FLUXNET (or ICOS, whichever the authors prefer).

We have added the following clarification to the abstract:

l. 15 *"...to capture drought-induced reductions in GPP from ICOS and FLUXNET eddy covariance data at 18 forest sites across Europe."*

2. Graphical abstract: Need to specify the units on the y-axis (these are SDs?)

We have added the units to the y-axis



3. Line 81: Introduction "Yet no previous studies have quantified the lag between drought start and the timing of GPP and VI response for a wide range of different indices in forest ecosystems." This is a key justification for the study, please include some variation of this statement in the Abstract if possible.

We have added the following sentence to the abstract:

l. 13 *"However, the length of time lags between drought start, GPP and VI response in forests have not been quantified or compared among VIs."*

4. Line 43-45: I know the references for these VIs are listed in Table 1, but the references should be here too because this is where they are first introduced and Table 1 doesn't appear until much later in the document.

We have added references for all the VIs the first time they appear in the introduction.

5. Line 115: Why was a 15-day moving average needed/chosen for LST?

Since LST varies significantly at the daily scale, we used a 15-day moving average to remove day-to-day variations in LST and emphasize the longer-term fluctuations due to drought. We tested moving averages using shorter window lengths but found that a 15-day moving average gave a good compromise between removing high frequency daily variation but not overly smoothing the data. We have added text to clarify this decision:

l. 133 *"We smoothed the LST data using a 15-day moving average to remove day-to-day variations and emphasize longer-term changes due to drought".*

6. Line 119: Why were NDVI and EVI calculated instead of using the existing MODIS products (MOD/MYD13)? This is likely what many MODIS users would likely use instead.

We used the MCD43A4 product, which is the Nadir Bidirectional Reflectance Distribution Function (BRDF)-Adjusted Reflectance (NBAR) dataset, to calculate many of the VIs 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 vegetation index time series. It is the highest quality MODIS product available for the bands of interest. We have added text to the methods section to clarify our data choice as follows:

l.138 *"We opted to use the MCD43A4 product because it provides reflectances at nadir corrected for changes in viewing angle."*

7. Line 119: Were these daily or 8-day MODIS reflectance products, or something different? I assume this is the 8-day product since it's smoothed using TIMESAT to make daily products. Any more details that can be provided about settings for the TIMESAT smoothing? Spline spans, knots, order, etc.

As stated in Table 1, these were daily MODIS reflectance products (except for FPAR and LAI which were 4-day). Even though the product is daily it still contains gaps and noise due to e.g. cloud cover that have to be filled or smoothed using TIMESAT. We have added text to clarify which settings were used in TIMESAT

l. 143 *"We used the following settings in TIMESAT for all bands: spline smoothing factor = 5000, number of envelope iterations = 1, adaptation factor = 0."*

8. Line 126: Was this a regular linear regression slope, or Sen's slope with a Mann-Kendall test (which would be more appropriate for time series and less sensitive to outliers)?

It was a regular linear regression slope but as stated on l. 147, we fit it to the 90^{th} percentile of the data which should be less sensitive to outliers than for example, using the mean. We had not heard about Sen's slope before and will use it in future work.

9. Line 134: Please define MAD

It is already defined when it is first used (now on l. 141, previously on l. 121).

10. Line 135: Similar to earlier comments, form of slope and test calculation? Why was 15-days chosen for smoothing?

We fit a standard linear regression to all the data with a Wald test to determine the significance of the slope. Since the GPP exhibited large daily fluctuations, we used a moving average to remove this high frequency noise and emphasize longer-term changes due to drought. We tested moving averages using shorter window lengths but found that a 15-day moving average gave a good compromise between removing high frequency daily variation but not overly smoothing the data. We have edited the text to clarify these points:

l. 156 *"To remove long-term trends in GPP we fit a standard linear regressions to all data (daily resolution) in each time series. If the slope of the regression was significantly different from 0 (based on a Wald test), the GPP was detrended. The daily GPP timeseries was then smoothed* *using a 15-day moving average to remove high frequency noise and emphasize longer-term changes due to drought."*

11. Line 151: Any reference for discarding values outside +/- 6 z-score SPEI?

We did not discard any SPEI values, only the GPP or VI z-scores that were +/- 6 were discarded. We chose to use +/- 6 as a threshold because it represents values that are 6 standard deviations above or below the mean which represents extreme outliers. We have clarified this point:

l. 173 *"Any GPP or VI z-score values < -6 or > 6 were discarded as outliers".*

12. Line 160: These z-score values selected are reasonable, but why choose different thresholds for through based on SPEI (-1.5) and GPP (-1)? What happens if these values are adjusted? Don't necessarily need to do some kind of sensitivity analysis here, but would be good to know or reference.

We had to use a less strict threshold for GPP than SPEI to ensure that we still had a sufficiently large sample size to be able to analyze while making sure that GPP was still noticeably negatively affected by drought. Using -1.5 for GPP eliminated too many droughts/sites. We have clarified this point:

l. 188 *"We used a different threshold for GPP compared to SPEI to define drought to ensure a large enough sample size remained for the analysis."*

13. Line 205: I would think that the droughts where all the VIs were less than -1 z-scores would be an indicator of the severity of these particular droughts. A broad impact like this may not be normally expected. Might be worth including some comments on this in the Discussion.

We carefully considered this suggestion but could not find a suitable place in the discussion to include such a comment without adding significantly more text. It does not fit within the current structure of the discussion.

14. Line 215, Figure 2, panel b: How can we be sure this decline in NDVI preceding the SPEI identified drought is also a drought and is related? NDVI z-score could have dropped before the drought event for an unrelated reason and might not have anything to do with the drought itself. If it is likely a drought, is this a suggestion that SPEI-3 is too long a time span and SPEI-1 month would be more appropriate due to the quickness of the response?

Unfortunately, we don't have enough ancillary data from the sites to know if VI declines preceding drought at some sites are due to the drought or other factors. We assume that VI decline may precede the start of SPEI drought since SPEI-3 is calculated using the previous 3 months, i.e. it is already getting hot and dry before we define the start of drought and therefore the trees may already show signs of responding. It is difficult to say whether SPEI-1 month would be more appropriate without doing a full analysis. Nevertheless, the fact that GPP (at least for deciduous sites) and most of the VI's $Date_{end}$ occurs before SPEI-3 $Date_{end}$ suggests that a shorter SPEI aggregation time may work better. On the other hand, the SPEI also becomes noisier, i.e. includes more high frequency variation, at shorter timescales which may make it harder to define drought periods. We have added text to the discussion to clarify this point:

l. 434 *“In fact, at broadleaf sites, GPP stopped responding to drought on average over a month before the drought finished as defined by SPEI, which suggests an even shorter SPEI aggregation period may work better for these sites. However, using a shorter aggregation period introduces more high-frequency variation into SPEI which may lead to difficulties identifying sustained and severe drought periods.”*

15. Line 318 and Figure 6 caption near Line 355: These groupings are tree genus, not tree species.

We have corrected the text in both places.

16. Line 370: Figure 7 may be better to include in a supplemental section since no significant differences were found for any pre- and post-drought comparisons

Although no statistically significant differences were found we still feel that there are some important patterns that emerge in Figure 7 and that it hence should remain in the main text. However, we have toned down the language used to describe these results from Figure 7 in the 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, even though GPP 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 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”*

17. Line 433: Pinus and Fagus are both genera, not species.

We have corrected all mention of species to genera where appropriate.

18. Line 486: I think this should be referencing Figure A2 instead. Figure 2a only shows NDVI, while A2 shows all the VIs for DE-Hai.

We have corrected the Figure reference.