Herrmann et al. analyse storylines of low summer NDVI events in Europe for the period 2000-2020. The storylines are estimated for up to two years prior to the low NDVI events and are analysed separately for the temperate and Mediterranean biomes. This is an extremely relevant topic, since the ongoing disturbances in Europe are in many regions unprecedented (Senf & Seidl, 2021) and the drivers of tree mortality are still elusive (Hartmann et al., 2022). The use of storylines to evaluate drivers of low NDVI events is an innovative and well-fitting approach that could have a high potential to shed light into the general processes driving these events. After reading the manuscript in detail, I believe the study unfortunately falls short of this ambition for generality, due to a number of fundamental issues with the specific methodological approach, underlying data, and the general lack of mechanistic insights, as discussed in more detail below. Moreover, the reading is at times engaging but at times it can be quite cumbersome due to the (heavy) use of confusing notation, unclear grammar and some distracting sentences that read as if the authors were addressing specific reviewers' comments, perhaps from a previous submission. That said, I believe that the analysis performed here can be of great value to the community, should these issues be addressed.

Major comments

1) Generality of the storylines & mechanistic interpretation

At the moment, the results and discussion are presented as if the authors can derive general lessons about meteorological storylines leading to reduced forest greenness in summer. Given the short length of the time record (3 years), and that very few (1-3) large-scale events dominate the overall number of events detected – 42% of individual events correspond to the 2018 drought/heat event – it is hard to accept that the conclusions can be generalized to different event types. This is even more problematic for the "consecutive events", which refer basically to the 2018/19 event (making up 82% of the "individual events"). Therefore, the event samples cannot be treated as individual events. If I understood well (see my comment below) the authors calculate the anomalies for the climate variables by selecting 10k randomly selected samples and calculate their means over the different events for each of the variables. By using a random sampling of *n* "pixel" events that are dominated by a single large-scale event (or two for the repeated events), a large fraction of the 10k samples will stem from this single event, so that their sampling will necessarily be biased. It follows that the meteorological anomalies estimated for these storylines will strongly reflect the anomalies of those 1-2 synoptic-scale events. This results in the authors possibly overinterpreting short-term departures of the storylines from the 95% confidence interval, particularly the rainy previous winter season – that occurred in much of the region affected by the 2018 event (https://climate.copernicus.eu/european-wet-and-dry-conditions) – as a general percussor of low summer NDVI events. It is unclear what would be the mechanistic link between rainy winters and dry/hot summers leading to low NDVI events and the authors do not offer satisfactory explanations (rather list a number of hypothetical processes in the discussion).

I would be curious to whether the authors could find significant signals, especially the rainy winter peak before the low NDVI events in temperate regions, if events occurring in 2018 are removed from the analysis. The authors could also try an alternative sampling approach that would explicitly correct for the sampling bias due to the predominance of few large-scale synoptic events, but this would reduce drastically the number of events (by about half for EV10 and by 80% for EV11).

An alternative could be to tone down the ambition to derive general conclusions for 2000-2020 and rather focus on highlighting the value of the storyline approach to learn about anomalies in

vegetation activity and, specifically, extremes. While the results for 2018/19 in central Europe are not necessarily new, the fact that the approach proposed here shows consistent results past studies while providing a means to assess more generally past history is a strength that could/should be stressed. For the Mediterranean case, the results might also reflect the year 2017 the most (given figure 3d), so that a similar approach/reasoning can be applied.

2) Spatial resolution

The authors analyse storylines leading to low NDVI extremes, focusing on forest ecosystems. To do this, they rely on NDVI data from MODIS at 0.05km. This is already a relatively coarse resolution for the analysis proposed, as pointed out by Reviewer 1, but the authors then further coarsen the data in the analysis to 0.5 degree spatial resolution. The authors tried to minimize this issue by estimating low NDVI values first at the 5km resolution and then imposing the condition that at least 50% of the coarser grid cell needs to register low NDVI events. However, their definition of "forest pixels" includes pixels with less than 50% of forest, and even as low as only 20% (Section 2.1). This results in many areas that are dominated by croplands or mixed tree/herbaceous cover (parts of central and eastern Europe and of Italy, southern Sweden), being included as "forests".

While I understand that the authors could not perform this analysis at the higher MODIS spatial resolution given the lack of such high-resolution climate data over Europe, it is unclear why the authors chose to aggregate results to 0.5 degrees, since ERA5's native resolution is 0.25 degree, and ERA5 land provides downscaled fields at 0.1 degree spatial resolution, much closer to the spatial scale used to define events and to distinguish (at least better) forest from non-forest pixels. The same for E-OBS, and the group of COSMO-REA reanalyzes provide temperature and precipitation fields at 6km or even 2km.

Surely, repeating the analysis at such fine resolutions (2km) would be more computationally expensive, so that keeping the 5km scale of the initial steps is probably the most feasible option.

3) Synthetic event sets and meteorological means

In appendix A and the methods, it is generally unclear if the 10k event sets are derived per pixel individually, as suggested by keeping *n* in subscript (appendix A), or across pixels or across individual events (i.e. pairs of (pixel, year)). It is also unclear how the authors calculate the means and climatologies of the meteorological variables (Line 562-63). Generally, this makes it difficult to evaluate and reproduce the analysis.

The authors state that they take 10k synthetic event sets and then (1) take the mean among all events, then (2) take a distribution of these mean values that is used to calculate climatologies. I have several questions about this step:

- It is not clear what is the size of the event set matrix, is it 3d (10k, 21, n)?
- And over which dimension is the first mean calculated, the second?
- What is the size of the resulting vector/matrix?
- Do the climatologies in T and P have seasonality?
- Over which dimension is the standardization performed?

For the consecutive event sets, how exactly is this done, do you select from EV10 subsets of synthetic events that happen consecutively? Where are the values of the different n values derived from?

Finally, the authors mention in the methods section that the sampling strategy preserves the spatial correlation of T and P fields, but this is not discussed in Appendix A.

As a consequence of these issues, it is difficult to evaluate the robustness of the method described. I suggest either describing the steps using explicitly mathematical notation, reporting the sizes of the vectors/matrices, and/or a flowchart to facilitate understanding of the exact steps performed here. It would also be good to discuss the rationale behind using bootstrapping in the methods and start of Appendix A.

4) Use of forest disturbance dataset
It is unclear why the authors introduce a whole new dataset that is only used for cross-comparison purposes. Low NDVI events do not necessarily have to correspond to crown-mortality events (D), and conversely crown mortality events might not necessarily result in low NDVI if understory vegetation benefits from the canopy opening. This discrepancy is briefly mentioned by the authors, and very clear in Figure 4, but not fully addressed. Another point not really mentioned is that D can be affected strongly by anthropogenic signals, such as management and selecting logging. Since the authors presented the goals of the analysis as to evaluate extremes in vegetation greenness, a clearly well-defined and constrained problem, I find that the comparison with D adds more confusion, rather than clarity to the study.

Specific comments:

"forest performance" is not a standard expression and is rather unclear whether the authors mean vitality, health, or any other aspect reflected by NDVI (vegetation cover, LAI, …). To be accurate, the best would be to stick to "forest greenness".

Line 23: Over what time-scales is this sentence referring to? This is not true for the past several decades – acid rain, changes in management, reforestation, elevated CO2, nutrient deposition, … there is a long list of processes that have been destabilizing forests in Europe.

Line 48-50: but in Mediterranean regions, drought can also reduce fuel load (Pausas and Ribeiro, 2013).

Line 56: intensively discussed… in the literature?

Line 59: "stressed", or "identified"?

Line 60: "drought prone region", such as the Mediterranean?

Line 63: "margins" of the growing season is an unusual expression

Line 65: increasing understanding "of …", specify what is meant here. Furthermore, the concept of storylines should be described in more detail and appropriately referenced (e.g. Shepherd et al. 2018)

Line 73-74: Other studies have attributed this to a Wave-7 pattern and a positive NAO phase (Drouard et al., 2019 and Kornhuber et al., 2019)

Line 84: what do the 90 in subscript stand for, 90-day moving average not yet mentioned, making it confusing.

Line 85: why 3 year only?

Lines 87-88: I propose swapping (2) and (3)

Lines 101-102: why the choice of this specific domain and why ignoring boreal forests?

Figure 1 caption: please add a brief description of panel b) and it is impossible for the reader to understand it without reading the methods.

Line 106: no justification about why 0.5 degree is used.

Line 116: “at forest pixels” does not seem grammatically correct

Line 119: this is the first time missing values are referred to. Where do they stem from? The use of quality control flags? And what if there are two consecutive months missing?

Line 123: in mathematical notation, the apostrophe is usually used to express the first derivative, so this notation is confusing. Why not a for anomaly?

Line 132: replace “at” by “for” or “in”

Line 134: there is no scheme presented here. Do the authors mean the “approach presented here”?

Line 135: is it a “forest grid cell” or an “atmospheric grid cell over forested pixels”? Overall, I find these definitions confusing.

Line 136 and Equation 2: how does the flag work when there at 3 months in the season? Since the authors take the minimum value over the season, does this mean that it is enough that 1 month in JJA is flagged as low NDVI? How can the authors be confident that this is a “low NDVI season”?

Line 148: correct to “reanalysis”, singular

Line 149: why interpolating ERA5 to 0.5 degree?

Line 151: if this refers to seasonal averages, should the subscript be 90d or “season”?

Line 151-159: it is not fully clear if the standardization is done also for the 90d moving windows, please clarify.

Line 167: so only 40% of the values are “not extreme”?

Line 167-169: give correlation values and respective significance

Figure 2: what do the colorbars indicate? Not mentioned in the caption.

Line 181: this is not described in the Appendix A.

Line 191: for a Biogeosciences audience, is would be good to explain what the "outermost closed SLP contour …" means.

Line 191-195: more generally, for a Biogeosciences audience it would be good to explain what additional information does this analysis bring.

Line 197: grammar "To evaluate", "for" + "ing" does not express purpose/intention.

Line 201: which is aggregated and normalized. Why is the normalization now done at the coarser resolution?

Line 216: correct "succeeding" to "subsequent"

Section 3.1.1 – if the purpose of using the D dataset is for evaluation of low NDVI events, why not compare the annual variability in D as well here and in Figure 3?

Figure 3b: the colors in the 4 quads are hardly distinguishable

Figure 3d: if the authors decide to keep D, then add the extent affected by D events in this panel as well.

Line 249: these "conceptual, technical and physical reasons" are not really thoroughly discussed in Sec 41.

Figure 4: What is D'[]?

Figure 5: mark in shaded areas the periods when the event mean is outside of the 95% CI. What do the vertical lines in a-d indicate?

Line 269: negative, but still within the 95% CI. Here, and elsewhere, the authors over-emphasize non-significant results.

Line 271: for short periods. Add "is significant for Xdays … "

Line 274: continuously negative, but still within the 95% CI. Please give duration of the periods when event mean is outside of the 95% CI.

Line 276: negative in the previous winter, but not extreme.

Line 280: DJF of which year?

Line 284: can you give a mechanistic explanation for this?

Line 304: add ", respectively," between "P'90d" and "from"
Line 315: the accumulation of dry periods is not significant

Section 3.2.3: please state clearly that this applies basically to 2018/19 in temperate regions, and make a similar assessment for the Mediterranean biome.

Line 331: the information about the fraction of these years to the event samples needs to be given much earlier in the manuscript, and please add information for the Mediterranean biome too.

Line 334: why is 2020 excluded?

Line 340: "hot" anomalies?

Line 350: again, I find in-depth mechanistic interpretation of these patterns lacking in the discussion.

Line 355: "exerts" is not applicable here, since the NDVI events have no influence on T90.

Line 356: I quite like that the authors here give specific values of the anomalies discussed. This should be done throughout the whole results section.

Line 358: what does "small" mean? That the absolute value is close to zero?

Line 364: only T, see comment above for line 315.

Figure 8: please explain what the different color shades mean (95%CI, I believe)

Discussion Section: it is surprising that the limitations related to the short temporal records and the dominance of single years in the events analysed are not discussed.

Lines 414-325: How does the analysis done here "help to characterize the nature of these events"? If this would be true, I would expect separate analyses for pixels with low NDVI and no crown mortality and pixels with low NDVI and crown mortality. Overall, this paragraph is quite distracting given that the main goal of the paper was to analyse low NDVI extremes.

Line 432-433: the grammar can be improved

Line 437: regadless of what?

Line 437-438: do the authors mean that drought early in the growing season directly "damages" forests, or simply that low P in spring promotes drier summers?

Line 440: unclear why warming in the previous 3 years would affect an instantaneous process like fire.

Line 485: can you explain in more detail how acclimation results in reduced leaf area and productivity?

Line 488-490: isn't this simply a consequence of the fact that only low NDVI events were selected? The authors did not evaluate post-event recovery trajectories separately, so that they cannot know whether increased vulnerability out competes acclimation.

Line 491: sensitivity to drought is not shown in the results, what do the authors mean here?

Line 501: what does "superior statistical modelling" mean?

Line 508: "not shown", please add these results to the supplement and this is an important point.

Conclusions

I find that the authors overemphasize the winter wet signal in the temperate biome, which first is rather short (a couple of days outside of the CI) and second cannot likely be generalized for all events. This should be toned down.

Appendix A

Line 557: the superscript *r* should be placed above *n*, and be defined again in the text here, for those readers who might start here.

Line 571: Add "for" before the "null hypothesis". Also, EV10 and EV11 have not been defined previously in the appendix, making reading confusing.

References

Shepherd, T.G., Boyd, E., Calel, R.A. *et al.* Storylines: an alternative approach to representing uncertainty in physical aspects of climate change. *Climatic Change* **151,** 555–571 (2018). https://doi.org/10.1007/s10584-018-2317-9

Pausas, J.G. and Ribeiro, E. (2013), Fire and productivity. Global Ecology and Biogeography, 22: 728-736. https://doi.org/10.1111/geb.12043

Drouard, M., Kornhuber, K., & Woollings, T. (2019). Disentangling dynamic contributions to summer 2018 anomalous weather over Europe. Geophysical Research Letters, 46, 12537–12546. https://doi.org/10.1029/2019GL084601

Kornhuber, K., Osprey, S., Coumou, D., Petri, S., Petoukhov, V., Rahmstorf, S., & Gray, L. (2019). Extreme weather events in early summer 2018 connected by a recurrent hemispheric wave-7 pattern. *Environmental Research Letters*, *14*(5), 054002.