Evaluating the consistency of forest disturbance datasets in continental USA
Laura Eifler, Franziska Müller, and Ana Bastos, *EGUsphere*

Response to Reviewer #1

R1C1: The article “Evaluating the consistency of forest disturbance datasets in continental USA” aims for a comparison of different forest disturbance recording products to identify consistency and mismatched between them. It provides an interesting first exploration of differences in ground survey- and remote sensing-based datasets and how they align in some key disturbance characteristics like the timing of the disturbance and disturbance agent. The authors put effort into data cleaning and sub setting to generate a comparable dataset. I found the language to be clear, but the manuscript is very lengthy and would benefit from a shortening and excluding double mentioning of information and putting many details into the appendix to focus the story of the study. There are, major changes/issues that I would recommend the authors to consider. I will highlight the key aspects first, and then provide line-by-line comments in a classic review style.
We thank the reviewer for the evaluation and constructive comments. Below, we provide a point-by-point reply to the reviewer comments. The manuscript is shortened now, to pinpoint the important information and focus on the story of the study.

R1C2: 1/ **Aim of the study**: My impression is, that this study is not properly focus on its aim. You do not demonstrate, why one product is more suitable than another one for a certain application and this depends highly on the (research) questions asked. It would be greatly beneficial if you demonstrate one or two use cases and evaluate the performance of the different products (singular and in combination). In some cases, a remote sensing product might be more useful, in other cases (e.g. when I want to analyse non-stand-replacing disturbances) a field-survey based dataset might be better. I am missing the overall guiding question here.
You could for example do a case study: How have disturbance dynamics developed in the U.S. over the past 20 years, compared between Western U.S. (Rocky Mountains) and Eastern U.S., or something similar. In such a setting you could explore if e.g. the Hansen dataset says disturbances have accelerated in one region, but FIA would say they have remained stable. This would also directly link to the climate change context you mentioned in the introduction.
Another avenue could be to test different ground-based surveys to generate disturbance maps and to explore the impact, different ground-based survey designs and information have on the mapping (and the combination of different surveys). This is again a different approach, but could be a point in the discussion.
I ask myself a bit what you expected, as different foci of survey methods will result in different disturbance characteristic information and information granularity. The different methods do not measure necessarily the same process.
We agree with the reviewer that the aim of the study and the purpose of using certain products could be explained and demonstrated more precisely. We will clarify the aim and adapt the analysis according to that. Furthermore, we agree that providing a demonstration of the performance of the different datasets would be valuable and will provide such a case study in a revision of the manuscript.

R1C3: 2/ **Data product choice to compare:** It would be interesting to at least compare one other remote-sensing based product with the Hansen map, I listed some products in the line-by-line comments I came across, which are more specific to the US. Also, to separate analysis approaches to compare ground-survey products and remote sensing products seem to be more useful to me. The paper should not be published with at least including one other U.S. specific remote sensing-based forest disturbance product.

Thank you for your helpful suggestion. We appreciate your recommendation to include another U.S.-specific remote sensing-based forest disturbance product in the analysis. In the revision of the manuscript, we will incorporate the North American Forest Dynamics (NAFD) Forest Loss Attribution dataset (Schleeweis et al., 2020) in our comparison.

R1C4: 3/ **Analysis approach:** Even though there are some things that are not comparable, some products must be more useful/accurate than others or useful in a wider variety of settings. As a reader I would be interested in precise recommendations when to use which product, and when to avoid which product. And if there is a clearly superior product (but only if), then this should also be made clear. But it needs backing up with numbers.
We agree with the reviewer that recommendations about specific use of given datasets are an important result of this study and we will add these to the discussion section. We note however, that choosing a superior dataset is often difficult due given the many degrees of freedom in the differences between datasets. Nevertheless, we will provide more clear statements about suitability of the datasets analysed for specific research questions. See also R2C2 and R2C3.

R1C5: You did a first exploration in how the datasets align/ misalign in some characteristics, but I am missing an analysis which quantifies how data divergence might be affected by ownership, state, surveyor, topography or other environmental drivers (remote sensing products tend to be more uncertain in mountainous regions and areas with higher cloud cover and also field plot visits are more difficult and often rarer in certain locations). You only looked at some aspects separated, but there can be interactions which are important to consider (e.g. Is the year offset for remote sensing products higher in mountainous region? Are offsets very pronounced in some states, which would point to surveyors evaluating disturbance years differently?).
We agree with the reviewer that understanding reasons for divergence between datasets is very valuable. We note that we have analyzed the divergence between datasets by ownership, shown in Figure 6 and discussed in lines L.322-L.333 and L.416-L.418. We also analyzed the influence of the surveyor measurement timing on the differences in disturbance timing, shown in Figure 7, lines L.335-L.351 and L.424-L.427. There are indeed many other aspects that can contribute to disagreements between datasets, here we focused on information that is available from the datasets themselves, i.e. that users can make use of without need for further analysis.
Following the reviewer's suggestion, we will analyze the effect of topography on the disagreements with GFC and between FIA and IDS.

R1C6: Furthermore, I am still not sure how you account for the swapping in the FIA dataset. Disturbances are very location specific and random; even when USDA claims this does not impact the ability to analyse the data, this is probably only the case for larger area representation or strata representation, but difficult for location-based comparisons. I know that we cannot change the product, but if it consistently misses disturbances, **then the authors should reflect precisely on the swapping and make this a core point in the discussion**.
Thank you, we will discuss this issue as an important factor explaining misalignment more clearly.

Line-by-line comment:

Abstract:
L.7: How do you define consistency and reliability? What is your benchmark?
Here we refer to consistency as the degree to which different forest disturbance datasets report similar disturbance timings and agents of overlapping events.

Reliability refers to the trustworthiness of the temporal data/ identification of agents reported by each dataset.
We added the definitions in the text for clarification:

> *"In this study, we focus on the continental United States and compare four datasets on forest disturbances to evaluate their consistency and reliability regarding their spatial and temporal characteristics and driven agents, when available.* ***Here, consistency refers to the extent to which different forest disturbance datasets report similar timings and causal agents for overlapping disturbance events, while reliability reflects the accuracy and credibility of each dataset's reported timing and agent information.****"*

L9: lower alignment as in spatial overlap?
Yes, thank you. We rephrased and it reads now:

> *"In contrast, comparisons involving remote sensing data show lower* ***spatial*** *alignment and a delayed detection of disturbances by satellite observations compared to ground-based inventories."*

L10: more stylistic, but I prefer when authors stay in the active form (also the text jumps between active and passive)
Thank you for pointing that out, we will avoid a form shift and stay in active form throughout the paper.

Main text:
L.16: "These services are sustainable to society …", I do not really understand what you mean with this
We changed the sentence to:

> "*These services—including e.g. provision of food and water, climate regulation, and cultural services—are essential for society and help preserve the biological diversity of forests (Thom and Seidl, 2016)."*

L.20: true, but the paper also found that disturbances have a positive impact on biodiversity, which I would also mention in this context.
Yes, thank you. We added the sentence mentioning the positive impact in that paragraph:

> *"Naturally occurring disturbances can have a positive impact on biodiversity, as highlighted by Thom and Seidl, (2016). They can also enhance certain ecosystem services, such as albedo, and in some cases, contribute to carbon storage. Additionally, Thom and Seidl (2016) found an overall positive relationship between natural disturbances and biodiversity. In the reviewed studies, disturbances had neutral to positive effects on species diversity, species richness, and habitat quality."*

L.21: This impact definition of disturbances let's me ask how you define disturbances in your study? I would have defined it as a mortality event in the first place, but depending on the context, this might be different. A clear disturbance definition would be in general helpful for the comparison of disturbance events.
We agree that a definition of disturbance is helpful for comprehensiveness. We included a definition of disturbances in the introduction:

> *"Ecosystem disturbances have been defined in various ways in the literature, with one of the most widely cited definitions provided by White and Pickett (1985), who describe disturbance as "any relatively discrete event in time that disrupts ecosystem, community, or population structure and changes resources, substrate availability, or the physical environment". Here we follow a stricter definition that focuses specifically on tree mortality*

and health, in line with the aspects of disturbance considered by each dataset. In this study, a disturbance is defined as any event that causes tree mortality or affects forest health, including both direct mortality and preceding stages of decline. In the inventory datasets, disturbances encompass not only tree death, but also early indicators such as discoloration and crown dieback."

L.29: I agree that quantifying the climate change impact on disturbances is an important field of research, but I do not see the clear connection between your study and approaching this question. This points in general to one of my main comments, that I would appreciate a comparison between different (combinations) of disturbance products to address different ecological research questions to evaluate their "performance".
We agree on the lack of connection as it stands. This will be addressed by including a case study, see R1C2.

L.34: You speak about evidence being based on sparse data, but point to continental scale datasets in the following sentence. Depending on what kind of research question you ask, the current datasets at least for Europe are already quite impressive. So, when you make this statement, I would refine it a bit more for what kind of research questions we need more and refined data. (Also, for which spatial extend, as you compare a US based product, this might not help for questions in East Asia – so how do you address certain aspects of sparse datasets here?)
We did not mean to downplay the value of the existing datasets for Europe. Nevertheless it is a fact that they have limited coverage in space and time. Given the wide interest in using remote sensing for upscaling of disturbance agent classification models (see refs, Senf et al., 2017; Seidl et al., 2011; McDowell et al., 2015; Meigs et al., 2015, Kennedy et al., 2015; Shimizu et al., 2019) we believe that there is value in quantifying uncertainties in the datasets that can be used for benchmarking and training of such models. We do not make statements about the suitability of the datasets outside of the studied area, rather we hope that our study can motivate other regional assessments wherever data is available.

L. 38: IDA is not the abbreviation for the Forest Service and I assume it is the dataset you use from that agency. Please introduce the full dataset name before you use the abbreviation (same with FIA in the next lines, thou this one is very known, but helpful to write out for readers outside the field). As a general (stylistic) hint, I always prefer to avoid the use of abbreviations where possible, as it disturbs the reading flow when the reader is not familiar with the abbreviations from before. But this is a style choice I guess.
Apologies, this is a problem with the citation manager, which we will fix. Thank you for pointing out that these abbreviations were not introduced beforehand. We will change the reference for clarity and introduce the abbreviations beforehand.

L.45: and also depends on disturbance size and underlying forest structure
Disturbance size and underlying forest structure were added to the text.

L.57: Very vague statement. What kind of analysis? Do we need better quality cameras, so the image quality to classify disturbances is better? Or do we need better training for surveyors? Or do we need more refined survey designs, to gain more consistent and high-quality datasets?
The sentence was redundant since we then cite Coleman et al. (2018) who provides such a discussion, therefore it has been removed.

L.66: I am unsure how well the choice of datasets to compare is. As not all datasets exhibit the same features you want to compare (disturbance extent or disturbance agent), the pure

comparison of the products is always leaning to be unsatisfactory. I do agree thou that it is interesting to look where they agree and disagree – is this what you mean by robustness? - but I am missing in the introduction the read thread to why you choose those different datasets (which to my knowledge partially focus on different disturbance types and agents).
Exactly, by robustness we mean the general agreement temporally and spatially. We added a section explaining why we choose these datasets and that this approach serves as a case study to assess the reliability of large-scale forest disturbance monitoring and identify potential biases or inconsistencies.

L.70: I really like the effort to look into different disturbance products to explore their potential, but feel that identifying advantages and shortcomings needs an application example to actually test the performance of the different products to answer an (specific) question. Different products can inform differently well depending on the question asked.
We agree with the reviewer that the motivation and overall goal is not clear and will refine the aim and research question of the paper, see also R1C2.

L.73 ff.: Ok, but do you test the newly compiled dataset and if it improves analysis applications? I assume you need to reduce every product to some extent to create a harmonized combined dataset, which might be not advantageous in some applications?
The new dataset provides a subset of disturbed patches based on GFC that show good spatiotemporal agreement with inventory surveys and further includes disturbance agent information. We believe it can be useful for applications such as training machine learning models that aim to upscale disturbance classification based on remote-sensing as for example done by Forzieri et al. (2021); Senf and Seidl. (2018).
We will rephrase and clarify this point in the revised version of the manuscript. Testing the newly compiled dataset for such applications would be beyond the scope of this study, but we will include this dataset in the case study proposed in R1C2 as a first assessment of its applicability to such problems.

L.77 Figure 1: I would appreciate some more map insets to get a better idea of the data sets overlay, also with the remote sensing data.
Thank you for the suggestion, we will update the figure to better distinguish between datasets.

L.80: Table1: I appreciate the structured display of the different datasets compared in the study! A row indicating the disturbance agents surveyed would be helpful, as well as information on the resolution of the remote sensing product (for readers not familiar with the dataset) and a clearer description of the data format in the data row: (e.g. that the Hansen dataset is a continuous raster and that ITMN and FIA are point based estimates). You do this a bit in the information row, but also not extensively (e.g. you describe the IDS product to describe disturbances from insect and diseases, but in the following text in L.83 you state that it includes fire and wind disturbances).
The table is now updated including the information mentioned by the reviewer.

L.90: You name mortality as an example for a damage type, which again leads me to the question on how you define a disturbance in your study? (I would have set tree mortality as a characteristic of a disturbance, but this depends on your overall definition).
A definition of disturbances is provided in reply to Line 21 above.

L.91: Is there any information on how surveyors deal with anthropogenic disturbances such as harvest and salvage logging?

No, unfortunately we have not found any information on that. We will make the point that this information would be valuable for users in the discussion of the revised manuscript.

L.94: I see the advantage of utilizing the polygon data when comparing the dataset with the remote sensing-based map, but why excluding the point data, when you also compare this dataset with plot-level data?

In our opinion, polygons provide a better approximation of the actual disturbance footprint, capturing the spatial distribution and extent more accurately than point-based records. This is particularly important when comparing with remote sensing-based datasets, which also represent disturbances as spatially explicit areas rather than discrete locations. Disturbances, such as insect outbreaks or wildfires, do not occur as isolated points but rather as spatially continuous events affecting larger areas.

L.111: The swapping is a major problem when comparing the datasets with each other. If you would do a regional or some kind of sufficient area-based comparison, this might hold. But how do you plan to disentangle the different uncertainties from the disturbance detection itself and the additional swapping?

Thank you for your thoughtful comment. We understand the concern about the swapping and its potential impact on the comparison of the datasets.

According to the FIA Database Guide (Burrill et al., 2021), the swapping process is implemented to protect landowner privacy, as required by the amendments to the Food Security Act of 1985. Specifically, “up to 20% of private plot coordinates are swapped with another similar private plot within the same county. This method creates sufficient uncertainty at the scale of the individual landowner such that privacy requirements are met. It also ensures that county summaries and any breakdowns by categories, such as ownership class, will be the same as when using the true plot locations. This is because only the coordinates of the plot are swapped—all the other plot characteristics remain the same.”

While the spatial uncertainty introduced by swapping limits the feasibility of precise, patch-level comparisons, the broader statistical characteristics at the county or regional level are preserved. Consequently, the conclusions we draw about consistency across datasets—especially when using aggregated metrics—should remain robust despite this limitation.

In our case study, we plan to include a comparison of variability in disturbance agents across different spatial aggregation levels. We hope this will help illustrate how much the swapping procedure might influence our results at varying scales.

While it is difficult to fully disentangle the effects of swapping from the disturbance detection process itself, we believe that by carefully comparing the datasets and accounting for this potential bias, we can still derive meaningful insights into their relative performance. We will clarify this point in the manuscript, ensuring that the potential impacts of swapping on our analysis are transparent and clearly explained.

L.136: I was not aware that the ITMN focuses on drought- and heat induced tree mortality or is this the dominant mortality driver recorded specifically in the US? What is the idea behind comparing the consistency in extent and agent of disturbances, when the extent is not recorded in some instances and the data products specifically focus on different agents?

Indeed this was not fully clear. As stated in their website, the ITMN aims to quantify trends in tree mortality rates globally, attribute the causes of tree mortality, and accurately predict tree mortality trends. Given that ITMN is, to the best of our knowledge, the first inventory-based global dataset of tree mortality, we aimed to evaluate its consistency with regional inventories, particularly in terms of the consistency in disturbance agent. We find, however, that ITMN as presented in Hammond

et al. (2022) did indeed attribute most of the events in USA to heat-drought, while IDS attributed coinciding events predominantly to biotic disturbances. We consider that this is an important result of our study, as it is part of the inherent uncertainty in distinguishing between disturbance agents that tend to occur together (Allen et al. (2015)).
We have now corrected the sentence to:

> *The study notes a potential bias in the dataset, as it is based solely on available peer-reviewed studies, which in the case of the USA predominantly report drought- and heat-induced tree mortality events.*

L.140: I think including the Hansen disturbance map is fine, but why did you not include other remote-sensing based forest disturbance products, which are more specific for the USA? Remote Sensing products on a large scale often come with the cost of reduced accuracy when zooming into specific areas. Also, US specific disturbance maps can account better for specifics of the region and use f.e. country level auxiliary data for mapping improvement. Some other RS products explicitly map disturbances in the US only, e.g.: Masek et a. (2013). Ecosystems. https://link.springer.com/article/10.1007/s10021-013-9669-9 and Zhao et al.(2018). Remote Sensing of Environment.
https://www.sciencedirect.com/science/article/abs/pii/S0034425718300476 (or Goward / Dungan 2020)
I am not sure if the products are publicly available, but for such a comparison study it could be worth it to contact the authors. Else there are other available datasets which could be included in a further comparison: Schleeweis, K., G.G. Moisen, C. Toney, T.A. Schroeder, C. Huang, E.A. Freeman, S.N. Goward, and J.L. Dungan. 2020. NAFD-ATT Forest Canopy Cover Loss from Landsat, CONUS, 1986-2010. ORNL DAAC, Oak Ridge, Tennessee, USA. https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1799
I would argue you need to include at least one other U.S. specific disturbance map from Landsat to make this a valid comparison.
Thank you for the excellent suggestion, this is a very valuable dataset that we were not aware of. We will include it in a revised version (see also R2C2).

L.159: I like the flowchart overview, but there are a lot of details in the graph which might be overwhelming and not necessary with a good explanation in the text. I would keep a reduced version and put this one in the attachment, but this is more of a style choice.
We condensed the information of methods applied on each dataset to be more precise.

L.160: The Method section is very detailed. I welcome the transparency and decent documentation, but would recommend to shorten it for the sake of readability and flow of the text and keep the more detailed version for the appendix. For example, the exact column codes for the agent categories are not helpful for my understanding at this point, but great for reproducibility, so for those interested this information is great for the attachment.
We agree with the reviewer and will shorten the Methods section.

L.176: Could you find out why there are these really big polygons? And did you exclude polygons post or pre-dissolving polygons with the same disturbance year and agent?
No, we did not find an explanation of the exceptionally large polygons so far. We excluded areas exceeding 2000 km2 before dissolving the data. The process of dissolving polygons that share the same DCA and survey year helps simplify the analysis by grouping adjacent disturbances that are considered to be part of the same event, helping to avoid over-fragmentation in the data. Furthermore, since disturbances can extend over multiple neighboring polygons, clustering polygons that are geographically close (e.g., within a 500m buffer) and have the same DCA and

survey year could provide a more accurate representation of a single disturbance event. This approach helps to reduce noise, streamline the data for further analysis, and ensures that the disturbance event is not artificially fragmented into smaller, disconnected units.

L.186: I find it difficult to exclude all disturbance events without an assigned disturbance agent, especially when you compare the extent or just recording of a disturbance event between data products. Also, when checking for the disturbance period/time comparison, you loose quiet some data here. What is the rational for excluding all this information?
We agree on the data loss but defend the decision of excluding the points without agents, given the goal of this study, i.e. to evaluate the consistency between datasets in terms of spatiotemporal agreement **and** disturbance agent, to guide - among other applications - the development of forest classification algorithms (see R2C2). However, we agree that information about disturbances with no specific agent is relevant, so we will add a new category “no agent” to our analysis.

L.195: I am not really convinced that due to the swapping a proper comparison of disturbance event recording and agent is possible here. The buffer only addresses the fuzzing and also here I see a lot of potential bias when comparing it with the remote sensing-based map when e.g. the FIA plot is at the edge of a disturbance and moved a mile away from the disturbance edge. Also, as mentioned earlier, you can try to assess the effect of the swapping on the disturbance detection consistency between products, but how do you account for uncertainties in the other product and the different sources of uncertainty in the disturbance recording in this dataset itself vs. the effect of the swapping? A regional comparison as I proposed in my main comments could help here.
We agree that swapping, along with the fuzzing of plot coordinates, introduces some uncertainty in terms of precise plot locations, especially at the edges of disturbances. However, the buffer approach we used helps account for the fuzziness and ensures a more consistent comparison across datasets. The choice of buffer sizes (500 and 1000 m for ITMN and 800 and 1600 m for FIA) was made to increase comparability and reduce any edge effects from fuzzing or swapping, without assuming that disturbances would be confined to an exact plot center. While there could still be some bias when a plot is at the very edge of a disturbance, the buffer serves to increase the likelihood of capturing disturbance events that might be recorded as nearby but not directly on the plot itself. This approach is common in the literature for handling spatial uncertainty in datasets with imprecise plot locations. The buffer sizes also align with the necessary privacy protections within FIA’s data, ensuring that the swapping and fuzzing procedures do not compromise the integrity of disturbance detection. The use of larger buffer zones accounts for the uncertainties caused by fuzzing and ensures that any disturbance events near a plot edge are still included in the analysis, which aligns with the objective of maintaining county summaries and other regional breakdowns that are consistent with FIA’s privacy rules. In the revision of the manuscript, we will include a regional comparison, as already addressed in the replies to the main comments.

L.215: I find this whole section a bit unstructured and got lost which product you are comparing now with which one. And do you want to compare simply the disturbance occurrence or also the disturbance extent/size when possible?
We restructured the text for a better understanding. In this case, the disturbances are only compared in terms of their occurrence but not the extent or size. However, we will include an additional analysis focusing on disturbance extent, using the IDS, GFC, and NAFD datasets.

L.239: So, you did not incorporate the information from the FIA and ITMN databases into a fused data product? I can think of various reasons why this is tricky, but could you elaborate on that choice of excluding the other datasets?

We decided to combine GFC only with IDS because IDS has the most detailed and spatially explicit information on disturbances. This will be clarified in the revised version of the manuscript.

L.248: Why reporting only the values for the smaller buffers, when the fuzzing can go up to 1 mile?
Thank you for pointing that out. We will include the corresponding numbers for the larger buffer sizes, as is also done in Section 4.2.

L.156 ff: Next to the spatial overlap of patches, it would be interesting to compare the overall disturbed area mapped by GFC versus IDS (so in the area where we know that IDS mapping took place, how much area in total was mapped by the surveyors versus identified by the Hansen Landsat based map?).
This will be addressed by the inclusion of the case study, in line with R1C2.

L.284: This whole section can be shortened significantly by only highlighting those comparisons which stands out. The rest of the information can be retrieved from the reader in the graphic and table.
We agree and shortened the section as proposed.

L.288: Why do you not compare all datasets in one graphic?
Thank you for your comment. The decision to separate the figures was made to ensure clarity and to account for the different databases used for the comparisons. The datasets have varying coverage and overlap, leading to different sample sizes in each comparison. For example, IDS and ITMN include a maximum of 1,500 records per DCA, while IDS and FIA involve up to 600,000 records per agent. By separating the figures, we aim to clearly highlight the differences in a way that is visually comprehensible for each specific comparison. We can revise the figure if the Editor deems it important, but consider that the figure is clearer as it is.

L.307: In figure 4, please write out the disturbance agents or give the full agent name in the figure description.
Thank you, we added the full names in the figure description.

L.294: In general, I would propose using a confusion matrix instead of a bar chart comparison, so we can see which agent have been labelled by another agent (is it always the same agent misalignment? – than it is likely the labelling or is the confusion of agents more diffuse?). Also I would urge to report much clearer summary statistics.
We will incorporate this information into the results for a clearer demonstration of the agreement.

L.300: Would be also interesting, to see the agent agreement/disagreement in space (with a map, points indicating by colour the agreement). Maybe there is a spatial pattern? Also, it would be interesting to know if the agent agreement depends on disturbance size or if there is a clustering in time or if this simply depends on the surveyor.
We agree with the reviewer and will provide a comparison of disturbance extend between the datasets, note that spatial explicit information is only available in IDS and, to some extent, in GFC.

L.307: This section is actually a part of the discussion and not of the result section. When you want to lead over looking into effects of ownership and the recorded disturbance year versus measurement year, then I would give this structure in the methodology section (and prime it in the introduction) and simply report your results here.
We agree with the reviewer and added this explanation to the methods section 3.1.2 FIA, it now reads:

"The FIA data provides information on plot ownership (national vs. private forests) as well as the year of tree mortality, measurement, and inventory, allowing us to assess how these variables contribute to the discrepancies observed between datasets."

L.308: What about the spatial extent/ overlap?
We will add this information, in line with replies to previous comments.

L.315: The agent attribution in the FIA and IDA doesn't happen every year (but only the disturbance mortality)? This was not clear until this point and should be stated explicitly when describing the dataset. In this case simply comparing the disturbance years might not be a valid approach anymore. How do the surveyors decide on a disturbance year at this point? And which year do you compare in Figure 3?
We acknowledge the reviewers' concerns that this inventory approach introduces uncertainty in determining the correct disturbance and/or mortality year. Unfortunately, the methods that surveyors use to assign mortality years are not documented in the data description documents. This is precisely one of the aspects we aimed to address in our analysis.
Our results in figure 7 show that this method of reporting mortality timing in between field surveys introduces a lag of ~1 to 1.5 years in the differences between datasets. Therefore, our results show that applications that rely on FIA data for mortality timing assessment (e.g. climatic drivers) need to control for these uncertainties.
In Figure 3, we present the disturbance year from IDS (the only one provided), the mortality year from FIA (or the dataset used), and the respective temporal information available from ITMN and GFC, as mortality and stand loss respectively. These were the only time-related data available from the datasets.
To address this issue, we will clarify the temporal uncertainties associated with disturbance year reporting in the Methods and Discussion sections. In Figure 3, we specifically compare the IDS "survey year," the FIA "mortality year," and the "mortality year" from ITMN and "loss year" from GFC, as these are the only available indicators. The accuracy of these temporal records should be carefully considered when interpreting the results.

L.316: Yes, but that comes again down to what you define as a disturbance and without knowing the product too much in detail, there is probably a higher uncertainty due to image quality and training data quality.
The disturbance definition was added in the introduction.

L.321: Here I would be right away interested in the number of points which you actually compare (so information from line 329 ff.). Also, it is not clear to me which numbers you compare here until you go more in detail from L 329 ff.
We updated the section, including the total numbers right at the beginning of the paragraphs which are data. Set. Comparison by dataset. Comparison, so first. FIA and IDS and then FIA and GFC:

*"**In the FIA events overlapping with IDS, a total of 989166 events were recorded, with 153987 (16%) occurring in privately owned plots and 835179 (84%) in public plots.** We find that national forests tend to show smaller differences in the reported timing of disturbance, than privately owned forests, with mean differences of -0.9 years and -1.7 years respectively. However, we find a better **temporal** agreement overall **without separating private and national plots (as analyzed in Section 4.2).** Furthermore, we find a larger spread in the differences of event timing in **national** forest plots than over **private** ones, **with mean disturbance years of -0.9 and -1.7 and** standard deviation of the differences of 6.8 years and 7.3 years respectively. **However, the mean disturbance**

values for the separate ownership categories are higher compared to the mean for all events combined, while the standard deviations remain relatively similar, differing by only about half a year. *The data is unevenly distributed in national and private plots, leading to sub-group specific characteristics and larger differences in timing compared to the results of all data.*

For FIA events overlapping with tree loss by GFC, 85105 events were recorded, with 50358 (59%) in privately owned plots and 34747 (41%) in public plots. *In the case of GFC, the mean differences in the reported timing are very similar for both350 privately owned (1.3) and public forest plots (1.3 years), being between the values found for the FIA-IDS comparison. The spread in the temporal differences is slightly smaller for public plots (7.1 years) than for private ones (7.6 years), and always larger for the FIA-GFC comparison than the FIA-IDS comparison for each ownership group."*

L.325: What is meant with all data?

"All data" refers to the analysis without separating into private and national, meaning the analysis in Section 4.2 (Temporal agreement).

Thank you for pointing that out, we updated the sentence, now it reads:

"*However, we find a better temporal agreement overall without separating private and national plots (as analyzed in Section 4.2).*"

L.326: This doesn't seem like a big difference to me looking at the distribution spread (it is just a few months), so I would claim based on these results that it is the same between private and public. But did you take a test on the differences, which also accounts for the different sample size?

We haven't done a test on the differences but we will analyse the results when accounting for sample sizes. The statement also refers to the mean values, therefore we rephrased it for more clarity:

"*Furthermore, we find a larger spread in the differences of event timing in* ***national*** *forest plots than over* ***private*** *ones,* ***with mean disturbance years of -0.9 and -1.7 and*** *standard deviation of the differences of 6.8 years and 7.3 years respectively.*"

The fact that the mean difference is larger than six months and that data have annual resolution implies that on average the data can disagree by one year, which might be important for studies aiming to analyse climatic drivers of given disturbance events (e.g. the contribution of heat-drought to biotic disturbances).

L.327: Can you elaborate on that? The SDs for all comparisons range between 6.7-7.4 years (if I did not misread it), that's pretty much the same isn't it?

The standard deviations differ by only about half a year, but the mean disturbance years show a more noticeable difference, with -0.9 for national plots, -1.7 for private plots, and 0.7 for all data combined. We have revised the sentences for better clarity, following the comment above.

"However, the mean disturbance values for the separate ownership categories are higher compared to the mean for all events combined, while the standard deviations remain relatively similar, differing by only about half a year."

L.334: Again, did you test that in any way? That does not seem like a big difference (between public and private) to me, especially when the agents are not recorded annually. In general, it would have been helpful here to test for variable impact with a modelling approach. How much of the lag in disturbance detection is explained by the dataset combination compared, the ownership, maybe even topography/landscape/ tree species information and the state and year.

Thanks for the suggestion. We agree that a statistical modelling approach to quantify the relative importance of each factor could be valuable. We will include this in a revised version of the manuscript.

L.335: This information belongs into the dataset description, as I assumed up to here (or line 315) that the agent was recorded annually as well. This makes a big difference and introduce immense uncertainty, so far that I am not sure if you rather measure the surveyor differences than anything else.
We added these introductory sentences about the different years to the data section, to focus on the results.

L.338: How is a plot measured across different years?
According to the description on the FIA website, the program maintains a national network of permanent plots that are remeasured on a rotating basis every 5 to 10 years, depending on the region. Sampling intensity and remeasurement rates can vary across and within FIA regions, states, and measurement years due to operational and budgetary constraints (FIA, 2025; Hou et al., 2021). The specific field procedures used during remeasurement are not fully detailed here.

L.343: Again, how do the surveyors decide if a tree or tree cohort dies 15 or 17 years ago? There should be some information in a recording/survey protocol about that, otherwise this seems rather arbitrary.
The definition in the FIA Database guide defines the *Inventory year* as:
"The year that best represents when the inventory data were collected. Under the annual inventory system, a group of plots is selected each year for sampling. The selection is based on a panel system. INVYR is the year in which the majority of plots in that group were collected (plots in the group have the same panel and, if applicable, subpanel). Under periodic inventory, a reporting inventory year was selected, usually based on the year in which the majority of the plots were collected or the mid-point of the years over which the inventory spanned. For either annual or periodic inventory, INVYR is not necessarily the same as MEASYEAR."
We will add a summary of this information in the methods section.

L.365: In this case you need to correct table 1 as you state annual recordings there (or it is easy to misunderstand).
Thanks for pointing that out, we corrected it in the table.

L.385: But seems to be the similar case for the IDS dataset, that mapping only happens when disturbances reach a certain extent and severity.
That may be true for aerial detection surveys, but they also include in-field surveys, which can capture even small changes. Since these surveys are conducted at a much lower altitude than satellite imagery, smaller disturbances may be easier to detect compared to satellite images with a 30 m resolution. This argument primarily addresses the detection capabilities of satellite imagery, which, in our opinion, are more limited by the size of a disturbance than by IDS. To address this point more directly, and in line with previous comments, we will include an analysis of disturbance extent and size.

L.388: Exactly, and it depends on the process(-scale) you want to research. Hence a demonstration of the dataset on a simply question would be helpful to evaluate the suitability of the datasets (and dataset combinations) to evaluate that.
This is in line with R1C4 and we will provide recommendations more clearly, including on this point.

L.394: The buffer helps to identify presence or absence of the disturbance event registration, but does not enable you to compare the disturbance area (overlap), which is an extremely important information when analysing disturbances and their impact on the environment. You cannot do that with all datasets, as the information is missing in two of them, but as you also compared IDS and GFC, I would discuss that here a bit more.

Since our analysis does not focus on comparing the exact extent of disturbance areas, and IDS and GFC are inherently more spatially explicit than point-based datasets, we did not apply a buffer to them. Point-based data do not capture the full extent of a disturbance event, which is why we introduced buffer zones to increase the likelihood of detecting overlaps. We will analyse the % overlap between disturbances in IDS and GFC and expand the discussion to clarify this point.

L.396: I do am surprised by that! Do they include the condition of a plot being disturbed as a criterion for the swapping? (so only disturbed plots get swapped?)

Swapping of the private plot coordinates is done with a similar private plot within the same country, FIA User Guide states: swapped plots are chosen to be similar based on attributes such as forest type, stand-size class, latitude, and longitude.

They swap according to the ownership while accounting for similar attributes, therefore the condition and disturbance year shall be similar in our understanding.

L.410: You need to elaborate/discuss more what drives the difference in survey and disturbance recording year, as this is such a big uncertainty and driver of this pattern. If there are no survey manuals or guidelines by the USDA for the recording of the actual disturbance year, I would not use that information at all as it seems rather being a guess.

We will clarify this point following the results of the statistical analysis in response to R1 L.334.

L.446: You should have reduced the agent categories to the one with the lowest granularity (so the broadest agent definition).

We have considered this point carefully before the analysis. , While FIA claims to provide detailed disturbance agent information, it uses broader categories than, and not fully aligned with, IDS. For example, bark beetles and defoliators in IDS are grouped under "other biotic" in FIA. To show this discrepancy, we kept it in these categories.

We added a short statement highlighting the issue:

> *“For instance, the "other biotic" category in FIA could refer to disturbances caused by bark beetles or defoliators in IDS, which may increase the overlap between the datasets. However, it could also represent different biotic agents, leading to discrepancies in the comparison. This ambiguity highlights the challenges in comparing datasets with differing levels of agent-specific detail.”*

Furthermore, we will include an additional column to the IDS of “Broad_DCA” (insect) while keeping IDS_DCA with finer granularity, maintaining the distinction between bark beetles and defoliators. We will compare the broader agent categories and finer classes. In our opinion these details are important, as they could be crucial for other analyses, given the significant differences in the temporal and spatial patterns of different insect species.

L.446: Yes, but this is also as surveys have a hard time to differentiate drought stress and mortality induced by drought in the field or on the screen. You only have a chance to tear apart stressors or initial disturbance and actual or secondary disturbance leading to mortality by a more detailed research setup.

We agree with the reviewer that this is an inherent physical challenge and a general limitation of aerial surveys is their difficulty in detecting long-term, slow disturbances and disentangling the actual cause of mortality, particularly in cases of compound disturbance effects (McDowell et al.

(2015)). We will add a section to the text highlighting this as an important area for further research and potential improvements in survey methodologies.

L.472: That is not true, salvage logging and human interventions in general increase the chances for a disturbance detection, but disturbance detection does not rely on human intervention.
That is true, we corrected that sentence:

> *"Furthermore, the requirement of a significant threshold for tree loss detection, as applied in the GFC and other studies, may be influenced by human activities, such as salvage logging, which can increase the likelihood of detecting disturbances but are not essential for disturbance detection itself."*

L.481: From which finding do you derive that inventory-based datasets show the highest reliability? Which dataset do you define as the "trusted" benchmark datasets, which is the most reliable?
We agree that this was not clear. Our goal here is not to provide a single all-purpose benchmark, but rather identify the different components of uncertainty and their influence on the agreement between different datasets. The choice of relevant reference dataset will be case dependent, as also highlighted by the reviewer above. In the revised text, we will make this point and provide recommendations for specific applications (see reply to R1C4).

L.482: Most medium-resolution RS disturbance products have a hard time or no chance at all to identify lower severity disturbances. So, if you want to research/include those in an analysis, you have to turn to other products anyways (this is no new information). I would have been interested to see in which way you can combine remote sensing and ground-based surveys, to gain more insights about disturbance dynamics. This leads back to the request of demonstrating the use of the datasets on different research questions and to evaluate their performance.
Indeed that is a big challenge recognised by the community. There is a wealth of literature on using RS to detect and classify biotic disturbances, which are typically low severity or slow progression (Senf et al., 2015; Senf et al., 2017; McDowell et al., 2017; Trumbore et al., 2015; Meddens et al., 2014; Simard et al., 2012). However, as the reviewer states, this is very challenging. Part of the reason is the lack of reference datasets (Kautz et al., 2017). Our goal here is to provide guidance to the community on the uncertainties associated with the different datasets that are typically used to train forest classification models (Forzieri et al., 2021; Senf and Seidl., 2018; Meddens et al., 2012), or to assess climatic drivers of disturbance occurrence (e.g., Hammond et al., 2020). See also reply to R1 L34.

L.485: I am not sure how much your findings support this statement. As you compared different details in agent recordings, you say more about the differences in the record method, but not about the reliability in the agent detection. You would have to compare the least detailed agent classes and see how those align.
For IDS and ITMN the degree of detail is the same, only for FIA, this is not fully possible, because the disturbance agents are not fully aligned. We will provide results for a less detailed aggregation level: biotic, wind, fire disturbances.

L.491: A demonstration on using the different ground-based surveys (and their combinations) to generate a remote sensing-based disturbance map in order to understand the implications of the survey uncertainties would have been helpful. I am not sure what to draw from the current analysis.
It is not clear if the reviewer proposes to generate new RS-based disturbance maps outside of the patches covered by the inventory datasets. That would require training a classification model which is beyond the scope of this study, as here our goal was to provide guidance on uncertainties (see reply to comment in L482). Nevertheless, we consider that our GFC-IDS combined dataset

provides a subset of RS-based and inventory-informed disturbance map that can support the development of such maps (see reply to R1C2).

L.495: This is not really helpful to design methods for disturbance mapping per se. The question on how to set up a disturbance mapping design depends so much on the use case of the data and the research questions which we aim to answer. I am not sure what should follow from this. Should the agencies align their surveys and bundle their efforts, to create one more detailed disturbance data product? In what information dimension (year of disturbance, agent, disturbed area) is which dataset more preferable? Or which dataset combination?

We politely disagree. Information about uncertainties is a crucial part of any model development, including forest disturbance classification models, such as the existing RS based datasets cited in the manuscript. Given the increasing interest in using machine and deep-learning methods in such applications, proper assessment of the reliability of the training datasets is crucial. Note that in such methods, uncertainties can even be part of the modelling approach, if they are known (e.g., through the use of fuzzy labels or other uncertainty representations). Based on our results, there is a number of recommendations that can be made for agencies and users, for example:

1. Better alignment of disturbance agent types, whenever possible (agencies), to ensure consistency across datasets and improve comparability
2. Improved description of subjective decisions, e.g. when reporting tree mortality between inventories (agencies)
3. Explicit consideration of the effect of data collection methods in statistical applications, e.g. property status for FIA data (users)
4. Consideration of temporal uncertainties in disturbance timing when using data to quantify disturbance impacts (e.g. carbon losses) or drivers (e.g. climatic)
5. Better consideration of uncertainties due to disturbance agent classification in reference data, e.g. by performing analyses at coarser agent aggregation level

References:

Schleeweis, K.G., G.G. Moisen, T.A. Schroeder, C. Toney, E.A. Freeman, S.N. Goward, C. Huang, and J.L. Dungan. 2020. US National Maps Attributing Forest Change: 1986–2010. Forests, 11(6), p.653. https://doi.org/10.3390/f11060653

Senf, C., Seidl, R., & Hostert, P. (2017). Remote sensing of forest insect disturbances: Current state and future directions. *International journal of applied earth observation and geoinformation*, *60*, 49-60.

Seidl, R., Fernandes, P. M., Fonseca, T. F., Gillet, F., Jönsson, A. M., Merganičová, K., ... & Mohren, F. (2011). Modelling natural disturbances in forest ecosystems: a review. *Ecological modelling*, *222*(4), 903-924.

McDowell, N. G., Coops, N. C., Beck, P. S., Chambers, J. Q., Gangodagamage, C., Hicke, J. A., ... & Allen, C. D. (2015). Global satellite monitoring of climate-induced vegetation disturbances. *Trends in plant science*, *20*(2), 114-123.

Meigs, G. W., Kennedy, R. E., Gray, A. N., & Gregory, M. J. (2015). Spatiotemporal dynamics of recent mountain pine beetle and western spruce budworm outbreaks across the Pacific Northwest Region, USA. *Forest Ecology and Management*, *339*, 71-86.

Shimizu, K., Ota, T., Mizoue, N., & Yoshida, S. (2019). A comprehensive evaluation of disturbance agent classification approaches: Strengths of ensemble classification, multiple indices, spatio-temporal variables, and direct prediction. *ISPRS Journal of Photogrammetry and Remote Sensing*, *158*, 99-112.

Kennedy, R. E., Yang, Z., Braaten, J., Copass, C., Antonova, N., Jordan, C., & Nelson, P. (2015). Attribution of disturbance change agent from Landsat time-series in support of habitat monitoring in the Puget Sound region, USA. *Remote Sensing of Environment*, *166*, 271-285.

Forzieri, G., Girardello, M., Ceccherini, G., Spinoni, J., Feyen, L., Hartmann, H., ... & Cescatti, A. (2021). Emergent vulnerability to climate-driven disturbances in European forests. *Nature communications*, *12*(1), 1081.

Senf, C., & Seidl, R. (2018). Natural disturbances are spatially diverse but temporally synchronized across temperate forest landscapes in Europe. *Global change biology*, *24*(3), 1201-1211.

Allen, C. D., Breshears, D. D., & McDowell, N. G. (2015). On underestimation of global vulnerability to tree mortality and forest die-off from hotter drought in the Anthropocene. *Ecosphere*, *6*(8), 1-55.

Pickett, S. T. A., and P. S. White, editors. 1985. The ecology of natural disturbance and patch dynamics. Academic Press, New York, New York, USA.

Coleman, T. W., Graves, A. D., Heath, Z., Flowers, R. W., Hanavan, R. P., Cluck, D. R., & Ryerson, D. (2018). Accuracy of aerial detection surveys for mapping insect and disease disturbances in the United States. *Forest Ecology and Management*, *430*, 321-336.

Burrill, E. A., DiTommaso, A. M., Turner, J. A., Pugh, S. A., Menlove James, C. G., Perry, C. J., and Conkling, B. L.: The Forest Inventory and Analysis Database: Database Description and User Guide for Phase 2 (Version 9.0.1), U.S. Department of Agriculture, Forest Service, 29, 1026, http://www.fia.fs.fed.us/library/database-documentation/, 2021.

FIA, 2025. Forest Inventory and Analysis. Available online at https://research.fs.usda.gov/programs/fia.

Hou, Z., Domke, G. M., Russell, M. B., Coulston, J. W., Nelson, M. D., Xu, Q., & McRoberts, R. E. (2021). Updating annual state-and county-level forest inventory estimates with data assimilation and FIA data. *Forest Ecology and Management*, *483*, 118777.

Kautz, M., Meddens, A. J., Hall, R. J., & Arneth, A. (2017). Biotic disturbances in Northern Hemisphere forests–a synthesis of recent data, uncertainties and implications for forest monitoring and modelling. *Global Ecology and Biogeography*, *26*(5), 533-552.

Hammond, W. M., Williams, A. P., Abatzoglou, J. T., Adams, H. D., Klein, T., López, R., ... & Allen, C. D. (2022). Global field observations of tree die-off reveal hotter-drought fingerprint for Earth's forests. *Nature communications*, *13*(1), 1761.

Senf, C., Campbell, E.M., Pflugmacher, D., Wulder, M.A., Hostert, P., 2017. A multi-scale analysis of western spruce budworm outbreak dynamics. *Landscape Ecol.* 32, 501–514.

Trumbore, S., Brando, P., Hatmann, H., 2015. Forest health and global change. *Science* 349, 814–818.

Meddens, A. J., & Hicke, J. A. (2014). Spatial and temporal patterns of Landsat-based detection of tree mortality caused by a mountain pine beetle outbreak in Colorado, USA. *Forest Ecology and Management*, *322*, 78-88.

Simard, M., Powell, E. N., Raffa, K. F., & Turner, M. G. (2012). What explains landscape patterns of tree mortality caused by bark beetle outbreaks in Greater Yellowstone?. *Global Ecology and Biogeography*, *21*(5), 556-567.

Meddens, A. J., Hicke, J. A., & Ferguson, C. A. (2012). Spatiotemporal patterns of observed bark beetle-caused tree mortality in British Columbia and the western United States. *Ecological Applications*, *22*(7), 1876-1891.