Editor

Dear authors,

We have received two reviews on the revised submission. We thank you for your patience as two separate reviewers, late in the process, let us know they were no longer able to review the paper. This delayed the process. We were able to find a new expert reviewer 2 who stepped in later.

Both reviewers have a positive outlook on the paper, and reviewer 1 appears appreciative of the efforts made in the first round. However, both notably point to **need to complete the narrative with a change in framing**. For example, reviewer 1 points out the **need for a case study/example to demonstrate the value of combining these datasets.**

Therefore, given the consistency of this comment from both reviewers, we are returning these reviews and request revisions on all points and particularly focused on this framing point.

Sincerely,
Andrew Feldman

Dear Editor,
Thank you for handling the manuscript and for the constructive feedback.

We revised the case study by removing the NDVI analysis, as indeed it was too closely aligned with the remote-sensing datasets and did not provide an independent line of evidence. Instead, we now present a trend analysis for each dataset in the western and eastern United States (see Section 4.6), which aligns with Reviewer's #1 comments.
Additionally, we removed the section on the combined dataset, as it did not substantially contribute to the overall findings of the study and it can be reproduced using the code we will provide alongside the manuscript.

We identified a minor error in the preprocessing of the NAFD data, which primarily affected comparisons between NAFD and IDS as well as NAFD and GFC. After correcting the issue, we repeated the analysis and updated the corresponding figures and results. The temporal agreement metrics changed slightly - most notably a stronger negative mean for IDS–NAFD and a shift from negative to positive differences for NAFD–GFC.

Sincerely,

Laura Eifler on behalf of the co-authors

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

Response to Reviewer #1

The revised manuscript “Evaluating the consistency of forest disturbance datasets in continental USA” shows considerable improvements and addresses several of the previous suggestions. The overall structure has become much clearer, and the figures and presentation of results are greatly improved. I particularly appreciate the effort invested in comparing different disturbance agents, which is an important and valuable addition. These revisions substantially strengthen the manuscript.
Nevertheless, I still see several major issues that should be addressed before publication. Below I highlight the key aspects first, followed by detailed line-by-line comments.

1/ Still no application example
The manuscript would benefit from a more **concrete application example to demonstrate the added value of the comparative analysis.** At present, the study documents the inconsistencies among disturbance datasets, but does not show **how these differences matter for answering specific ecological or management questions**. For example, applying all datasets (and in varying combinations) to the question: **How did the amount of area disturbed and prevalence of disturbance agents changed 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 make the results more tangible and **help clarify to what extent disagreements among datasets lead to different scientific or management conclusions, which is ultimately of interest to the community.**
Thank you for this suggestion. We replaced the NDVI analysis with a trend analysis across the different datasets. Following the reviewer’s suggestion, we now compare the trend in event count and area affected per dataset in the study period (2001 to 2010), relative to the decadal mean in the East and West U.S (to allow for comparability across datasets). This is presented in the new Section 4.6. We analyse the overall trends per dataset, as well as the trends per disturbance agent (where possible, since not all datasets report the agents).

2/ NDVI trends are no justifiable way to compare the datasets
The use of MODIS NDVI time series as a comparative baseline is problematic and **does not provide a robust or independent validation framework**. NDVI is not equally sensitive to different disturbance types and severities, often underestimates low-severity disturbances, and is easily confounded by phenology or mixed pixels. Further, GFC itself is partly derived from NDVI spectral changes, which introduces methodological circularity: by design, GFC (and therefore IGSD) will align better with NDVI than datasets based on field inventories or aerial surveys. In addition, the spatial resolution mismatch between MODIS NDVI (500 m) and Landsat-scale disturbance products as well as polygon delineated disturbances introduces further uncertainty. For these reasons, NDVI anomalies cannot be considered a justifiable or independent way to compare disturbance datasets. Instead, you would probably need to develop your own independent baseline dataset for the time period where all datasets overlap with a convincing method justifying the superiority of the new dataset to the

others to than benchmark all datasets against this one for a certain or several subregions. But I would say this is a different study in and of itself and out of scope. A case study would instead examine whether datasets lead to different conclusions about disturbance regimes (e.g., temporal trends, drivers, or spatial distribution) in a specific context (see point 1).
We agree that the NDVI is not as sensitive to certain disturbances agents and severities and that GFC is not independent from this analysis. We therefore removed the NDVI section as a case study. Instead, we analyse the trend within the study period per dataset in Section 4.6 (see R1C1).

3/ Narrative framing and discussion of implications
The narrative of the manuscript would benefit from reframing. At present, the text repeats many of the results but **does not sufficiently discuss their implications for applications**. The analysis highlights that disturbance datasets differ considerably in timing, spatial overlap, and agent attribution, yet the discussion does not fully explore **what this means for users in different contexts. As currently written, it remains unclear whether the authors wish to motivate the fusion of multiple products into a more comprehensive dataset, or instead to emphasize the need to improve the quality and reduce the uncertainty of ground-based data to provide better training and validation inputs for remote-sensing disturbance mapping.** These are two distinct directions, and clarifying the intended message would strengthen the paper (*or discussing how both avenues would be helpful and should be explored further*). **I would encourage the authors to position their work less as offering a direct solution, and more as sensitizing the community to the inherent differences among disturbance products, and how these differences should be considered when applying them in models or analyses.**
Thank you for the detailed comment. We restructured the framing to put more emphasis on the need for a more consistent and balanced ground-based disturbance assessment. We also aim to caution the research community about the biases present in these datasets and the risk of obtaining biased results depending on which dataset is used.
We emphasize the aim in the Introduction:

> *L. 125-129: By comparing these five forest disturbance datasets, we aim to highlight the variability and underlying uncertainties associated with different observation systems. Our goal is to provide a systematic assessment of their consistency in disturbance extent, timing, and agent attribution. Through this comparison, we seek to identify dataset-specific limitations and offer guidance on their use, for instance, by incorporating uncertainty ranges for disturbance timing or by combining complementary datasets to improve spatial reliability.*

And in the discussion:

> *L 558: Below, we discuss how the uncertainties underlying the different datasets explain these mismatches, and provide guidance for users on how these differences might affect analyses and interpretations.*

Abstract
L. 3: I would argue we do not necessarily need “globally consistent” datasets, as these often fall short of consistent definitions and struggle with alignment across different forest ecosystem types, which have inherent different disturbance dynamics and are impacted in different ways and intensities by climate change. Also, you compare products for the US, not globally, so I would imply scale down a bit here.

Thank you for the comment, but we politely disagree. Globally distributed datasets with consistent methods can be of high value for certain large-scale applications, such as risk assessments, analysing disturbance regimes (similar to what exists for fire, e.g. Archibald et al. 2012) and their changes, or developing models. This point is not original and has been raised previously, e.g. by Kautz et al. (2017). We acknowledge that indeed our analysis does not have global scope, so we rephrase to

> *L. 3-5: Despite the growing concern about these trends, the lack of consistent and temporally continuous data on forest disturbances* ***at large spatial scales*** *hinders our ability to accurately characterize changes in disturbance regimes and respond to these changes.*

L. 8 different mapping/recording approaches

Thanks, we added the word "mapping" in front of approaches to be clearer in that statement.

L. 13: do all products really have a temporal granularity of 1 year? Already the ITMN cannot really adhere to that right?

The temporal granularity of ITMN is considered to be one year. Hammond et al. (2022) provide two supplementary datasets: Supplement 1 includes the variable *mortality_year*, while Supplement 2 lists *Mortality Year(s)*. For our analysis, we used *mortality_year* as the reference year of mortality reported by the source study. While *Mortality Year(s)* may cover a range of years, *mortality_year* provides a single representative year, which we adopted as the event year. Thus, although the temporal resolution is annual, the dataset itself is not updated on an annual basis.

L. 19: In general, the writing style is still a bit "wordy", you can shorten as for example: "but with more pronounced differences at smaller scales and when accounting for disturbance agents/types"

Thank you for the suggestion, we adopted the suggestion, it now reads:

> *L. 18: The datasets show similar trends in total disturbed extent over conterminous USA (CONUS) for the common period of 2001-2010,* ***but with more pronounced differences at smaller scales, and when accounting for disturbance agents.***

L.21 This sentence is a bit confusing, with values you mean years? This half sentence is a bit out of context for me: "i.e. within their temporal resolution, for most datasets, but with a spread across individual events of ±1–4 years"; this one need rephrasing

Thank you for pointing that out, we rephrased the sentence:

> *L. 19: The datasets agree well in disturbance timing: the mean difference is less than one year, while the variability in differences ranges from about 1 to 4 years.*

L. 24: What do you mean with an "advanced detection"?

Thank you for pointing that out, we mean an earlier detection.

> *L. 22: The satellite-based datasets tend to show an* ***earlier*** *detection of disturbance events, compared to the other datasets, possibly due to the inconsistent revisiting times of the inventory datasets (FIA and IDS).*

L. 26: With spatial agreements of 24 to 58%, I would not call this "agree well", but more moderate to low agreement

Thank you for the comment. It now reads:

L. 25: Our results show that although the datasets exhibit reasonably good agreement in disturbance timing, their spatial correspondence is considerably lower.

Introduction:

L. 50: You wrote before that Thom and Seidl, 2016 found a general negative effect of disturbances on all ecosystem services and here state that they can enhance certain ES. This needs to be more consistent.

Thank you for pointing that out, we agree that it was confusing to read and we rephrased the paragraph. It now reads:

L. 43-49: Thom and Seidl (2016) found that the impacts of disturbances on ecosystem services are generally negative across all categories of services. Given their impact on forest productivity, growth, mortality and composition, disturbances can further feed-back to climate through changes in forest carbon balance (Bowman et al., 2009; Hicke et al., 2012). ***At the same time, Thom and Seidl (2016) also reported that disturbances can have beneficial effects on biodiversity, showing neutral to positive impacts on species diversity, species richness, and habitat quality. These findings highlight the complex nature of disturbances, which can simultaneously compromise ecosystem services and support biodiversity.***

L. 53: As you frame it, it seems that only climate change is leading to more/changing disturbance patterns and impacts. But humans are an equally important factor in how disturbance regimes change and there have been considerable changes in forest land use patterns in the recent decades as well (land abandonment, changes in harvest practices). You should at least mention this to make clear that we see the compound effect of several global change processes.

We agree that this paragraph focused mainly on climate change, which is indeed not the only driver for changes in disturbance patterns. We added the following paragraphs:

L. 55-62: Beyond the effects of climate change, anthropogenic factors also play a key role in shaping disturbance patterns. Changes in land-use practices, forest management, and afforestation efforts modify forest extent, composition, and age structure, ultimately influencing how forests respond to disturbance (Seidl et al., 2011).

Disturbance impacts can intensify when extreme climatic events coincide with high forest susceptibility or other preconditioning factors (Seidl et al., 2011; Bastos et al., 2021). Understanding and quantifying the combined effects of compound drivers is essential for predicting and mitigating future changes in disturbance regimes and their impacts on forest ecosystems (Bastos et al., 2023).

L- 94-97: This does not seem consistent. You make the argument for data integration (so combining ground surveys and remotely sensed data) and the sentence later you state that we need to reduce the uncertainty of ground-based datasets to train better models for remote sensing-based disturbance maps. These are two things: data fusion vs. uncertainty propagation

Thank you for pointing that out. We rephrased the whole paragraph, it now reads:

L. 94-98: Uncertainties in aerial detection, such as year-to-year variability (Hicke et al., 2020) and accuracy limitations (Coleman et al., 2018), highlight two needs.

First, integrating ground-based observations with high-resolution satellite imagery might improve the consistency, accuracy, and detail of agent information of disturbance detection through data fusion.
Second, a better quantification and understanding of the uncertainties within existing ground-based datasets remains essential, particularly as these datasets are increasingly used to train machine learning models that extrapolate disturbance patterns across broader regions (Senf et al., 2015; Forzieri et al., 2021, 2023; Patacca et al., 2023; Schleeweis et al., 2020).

L.114-115: With pointing to the forests structure, you probably want to hint at the challenge which spectral disturbance detection has with low severity disturbances and a lot of advances regeneration under the disturbed forest canopy. I would make this more explicit.
Indeed, we added a short paragraph explaining the limitations concisely:

L. 114-117: This limitation arises from the reduced capacity of optical remote sensing data to capture small-scale or subtle disturbances, such as low-intensity selective logging that removes only a few trees. These minor canopy openings are spatially diffuse and short-lived, making them difficult to detect from spectral signatures in satellite imagery. Sub-canopy structural changes likewise remain largely invisible to optical sensors (Gao et al., 2020).

Data:

L.137: I really like the improvements on the map and the direct comparison of the disturbance data products!
Thank you for the feedback, we are glad that the figure is much clearer now.

L.191: ", help track surveyor positions," – I am not sure what you mean by that.
The geo-referenced base layers should help to track the position of the surveyor during aerial surveys, improve the detection and avoid duplicate mapping. We rephrased the section for more clarity:

L. 191-193: During aerial surveys, geo-referenced base layers such as aerial photographs, topographic maps, and near-infrared imagery are used to track surveyor positions, improve disturbance detection, and avoid duplicate mapping of previously recorded damage.

L. 193-194: Does that mean you take the disturbance year recorded in this dataset as the "gold standard" and compare all other disturbance datasets towards this? If yes, please make this more explicit and elaborate why you assume that this is the most reliable disturbance year estimation.
We do not take the IDS as the gold standard, since we compare all pairs of datasets, by subtracting the mortality year of one dataset from the other. Likewise, for the spatial agreement we compare pairs of datasets, or report trends for each individual dataset so that they can be compared with the other datasets.
For IDS polygons, we used the survey year as the mortality year, as no other year is provided.
We added a short part at the end:

L. 194: We focus specifically on polygon data and use the survey year - the year in which the aerial survey was conducted - as the disturbance year for all records, ***since no explicit mortality year is provided in the dataset.***

L. 205 – 207: Please give one or more references for these statements.

We added references to the statements. It now reads:

> *L. 206-212:* ***The dataset does not specify the causes of forest loss, but it includes all disturbances that result in stand-replacing changes.*** *For example the tropics are predominantly affected by the prevalence of deforestation dynamics* ***due to shifting agriculture and commodity-driven deforestation (Hansen et al., 2013; DeFries et al., 2010; Curtis et al., 2018). In extratropical regions (temperate, boreal), tree cover loss is determined by forestry, fires, logging, diseases, and storms, at more moderate rates (Hansen et al., 2013; Potapov et al., 2008; Curtis et al., 2018; Sommerfeld et al., 2018).***

L. 231 – 234: I do not understand at this point, why you look at vegetation trends on MODIS scale and how you want to include this into your analysis.

Given the changes in reply to R1C1 and R1C2, these lines have been deleted.

Methods:

L.260: Is there non-accessible forest land? Natural disturbances happen independent of accessibilities, but I simply want to point out, that you should include all available forest categories.

That's a very good point. The condition table of FIA includes the condition status code (COND_STATUS_CD) which we used for filtering the data. They include five codes:

1 - Accessible forest land (at least 10% canopy cover by live tally trees/ species)
2 - Nonforest land (land that has less than 10% canopy cover of tally tree species)
3 - Noncensus water (Lakes, reservoirs, ponds etc of 1 to 4.5 acres size)
4 - Census water (lakes, reservoirs, ponds etc of >= 4.5 acres size)
5 - Nonsampled, possibility of forest land (Any portion of a plot within accessible forest land that cannot be sampled is delineated as a separate condition. There is no minimum size requirement. Reasons might be e.g. outside U.S. boundary, ocean, wrong location, lost plot, lost data, denied access area)

The proportion of group 1 to 5 on the data in the whole time period (1957-2022) are:

Table R1.1: Mean percentage of shares in the whole data of FIA of different condition status codes.

Code	Mean across states [%]
1	42
2	51
3	0.5
4	3
5	3.5

Codes 3 to 5 make up a small proportion of all plots, and since class 5 has a high uncertainty by being possibly forest land, we want to be sure that the used data is only forest and therefore we filtered for the condition 1.

L.287: So, you really only look at tree mortality events? This is a bit confusing as in the introduction you state, that you include both “direct mortality and preceding stages of decline”.

We are sorry for the confusion, indeed we only looked at mortality events. We changed the sentence in the introduction and removed the second part, it now reads:

> *L. 42: For comparability across datasets, here, we consider disturbance as any event that causes tree mortality.*

L.293: Have you tested if your results change, when you identify the “mode” disturbance year among the pixels? Because the remote sensing-based disturbance maps can more likely detect multiple disturbance events next to each other or more recent events than those recorded in the ground-based surveys. You can not fully account for that, but I would test if your results are stable with the most present disturbance year per polygon, instead of the most recent one. At the same time, how do you deal with situations, where you have different disturbance types recorded in NDAF?

Indeed, you are right that the results are more stable using the mode disturbance year. We actually had selected the year which is most common (most frequent) and not the most recent one, so the analysis was correct, but was reported wrong. It now reads:

> *L. 296: **The most common mortality year** among these pixels is assigned as the representative disturbance year of the GFC and NAFD.*

L. 295: per region as in per state?

Thank you for pointing that out. We removed this last part of the sentence, since it was based on the first version when the computation was done from the perspective of IDS, using the CONUS regions from this dataset. Now, that statement is not valid anymore.

L. 311-312: You should bring that earlier in this paragraph, against which datasets you plan to compare FIA to.

Thank you for the suggestion, we reconstructed the paragraph. It now reads:

> *L. 311: We compare the differences in disturbance timing between FIA and the three spatially explicit datasets--IDS, GFC and NAFD. The distributions of temporal lags are assessed for different sub-sets of the data: (i) events in public and privately-owned land reported by FIA, (ii) events in FIA with coincident inventory/measurement and mortality year, and events where these do not coincide, (iii) events in different states.*

L. 326: Why are you only modelling the uncertainty for the time lag and not for the spatial correspondence and agent attribution? This seems inconsistent.

This is a good point, but contrary to spatial and temporal uncertainties, modelling uncertainty in agent attribution would require to develop a model of all possible pairs of correct/incorrect attribution, e.g. wind with all other disturbances, and so on, for all different datasets, and identifying relevant predictors. Especially the latter aspect is not straightforward, as the datasets do not necessarily provide information that could be used as a predictor of differences in agent attribution. Instead we provide the confusion matrices (Figure 5) and the spatial differences in DCA agreement (Figure A5).

L. 329-330: That’s not directly about when disturbances are detected, but rather whether or how well they’re detected at all (classification accuracy). Accuracy differences can cause

some apparent lags (e.g., if one dataset misses disturbances until later, or misclassifies them, creating a pseudo-lag), but this causal link isn't made explicit in the text.

Thank you for pointing this out, the sentence has been revised:

> *L. 331-334: Based on the previous analysis of temporal lags for FIA, the contribution of different uncertainty factors to the temporal lags with other datasets are analysed, specifically: ownership status (ownership), timing of inventory/measurement relative to reported mortality (meas_lag), administrative differences in data collection, in this case per state.* ***These factors can affect the detection timing, but also might result in accuracy errors, which are then reflected in temporal lags in co-located disturbances.***

L. 334-335: You say: "best performing model is selected through ANOVA analysis using the lmer4 package in R" → ANOVA itself doesn't select the best model, you're likely doing likelihood ratio tests or AIC-based comparison. ANOVA (Analysis of Variance) tests whether the means of groups differ significantly.

Thank you pointing that out, indeed it is what we were doing. We rewrote the paragraph. It now reads:

> *L. 336-338: For each pair of datasets, we fit a set of models in a step-wise manner with increasing number of predictors.* ***We compare the models using an AIC-based comparison implemented in the ANOVA function of the lme4 package in R, and select the best-performing model based on these comparisons.***

L. 345 – 359: I am not sure at this point why you provide such a merged dataset. Do they provide the most accurate representation of the disturbances? GFC might cover a more recent time window, but the NAFD datasets in total is probably better suited for the USA and covers a longer time series of disturbances, which is important when thinking about quantifying baseline disturbance regimes or disturbance regimes in general. It comes down to the question: What is your aim to do with this dataset? This context is still not provided and the dataset fusion seems randomly placed here. Also, you assume IDS (agent detail) + GFC (recent coverage) = best of both worlds. But you don't address differences in disturbance definitions, spatial scales, or error structures between the two and you do not account for errors in both datasets (error propagation into the new dataset).

After consideration of the review, we decided to remove the part of the new compiled dataset from this study. The initial idea was to have a dataset for others to use with higher certainty of specific events. Based on GFC, which has a good temporal certainty and the IDS which is very detailed in information. Since this is based on two publicly and freely available datasets, and it can be reconstructed using the code provided with the manuscript, we remove it for simplicity.

L. 361- 374: This is not really a case study. With a case study, you define a research question for a specific area. An example could be: Have disturbances increased in area affected and frequency over the period x to y? Than you can use all available datasets to see how they would answer this question individually and then in different combinations to see, if they tell the same or a different story. This way you can find out if we would infer different results from different disturbance datasets, which tells me more about its uses and applicability in a scenario. Also, in this context you can discuss why different datasets might show the same or different trends. Also, NDVI is not a strong disturbance proxy in the context of all your datasets, as NDVI is not always sensitive to disturbances (e.g.,

low-severity events, non-foliage impacts, mixed pixels). Further GFC itself relies on NDVI to detect loss. Comparing GFC/IGSD disturbances to NDVI signals leads to methodological circularity.
Also, when creating the IGSD dataset, you're effectively pre-filtering to the "best-behaved" cases/ most reliably detected disturbance events, which makes IGSD look more consistent with NDVI by design.
Also, you have huge spatial mismatches when you calculate the NDVI on MODIS pixels and compare them with Landsat-scale detected disturbances and polygons of varying sizes.
In line with the answer to R1C1 and R1C2, we replaced the NDVI analysis with a trend analysis across the US in Section 4.6.

Results:
L. 385: Figure 3: I like the idea of the graph, but please align the y scale so a direct comparison between datasets is more feasible. Also including the amount for disturbance events by FIA in this graph is not helpful, as it does not directly translate into area disturbed. I would propose to have two graphics next to each other, one for area disturbed and one for the number of events/patches. (or do this mirrored as bar charts with inverted coordinates, so one direction per year shows the area disturbed and the other direction the number of events/patches.)
Thank you for the comment. We decided to keep the order of the plots in one column for visibility, but added a comment to the caption to be aware of the different metrics displayed. The y axis of FIA now shows the count x 1000 to display smaller numbers. We aligned the y axes of IDS, GFC and NAFD, to ensure a faster and clearer comparability. Here the updated figure and caption:

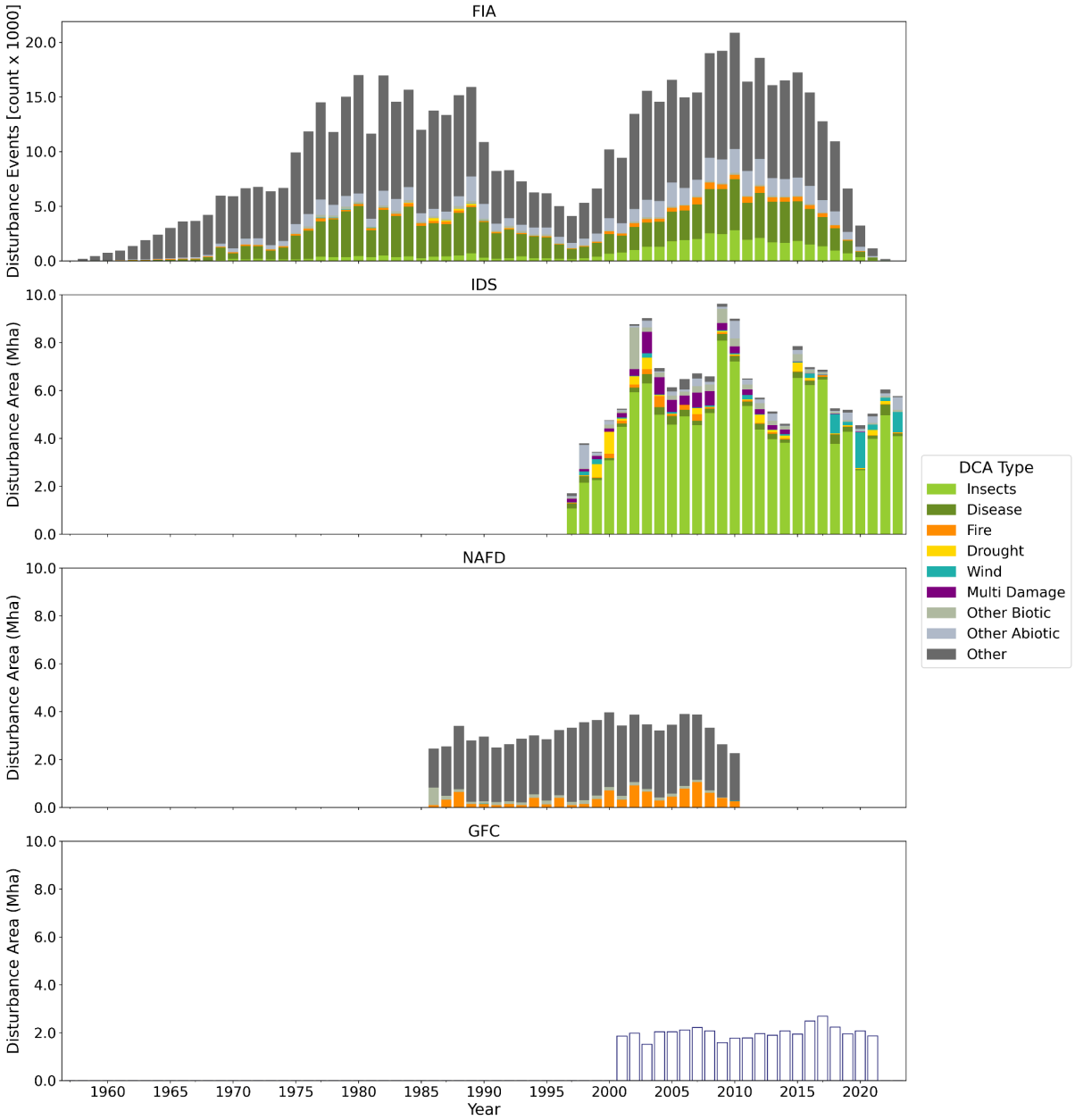

Figure R1.1: Time series of disturbance records from the original datasets, showing event counts for FIA and disturbance area (in million hectares, Mha) for IDS, NAFD, and GFC. Where available, disturbances are categorized by DCA type; GFC data do not include this classification. The full temporal coverage of each dataset is displayed to illustrate their respective time ranges, with the study period from 2001 to 2010 highlighted. ITMN is excluded due to the limited number of disturbance events within the CONUS region. ***Note the differences in metric (event count vs. affected area).***

L. 388-390: Can you also report next to the events how much area covered each survey recorded in direct comparison for the same period?

We updated the paragraph with numbers of the extent. It now reads:

> *L. 373-378: Figure 1 further illustrates these differences within a focused subset region. In this area, FIA reports 249 disturbed plots, whereas ITMN records only 10 events. T****he subset area covers 3613.3 km².*** *Among the spatially explicit datasets, IDS detects 9782 disturbance events* ***and maps the largest affected area (4,831.8***

km²)**. GFC identifies 35280 events, while NAFD reports the highest number of disturbances with 221153 individual records - reflecting its finer spatial granularity compared to GFC. **Despite the differences in event counts, GFC and NAFD both show similar spatial disturbance coverage, affecting 510.9 km² and 408.3 km², respectively. However, NAFD has a higher number of smaller disturbance patches.

L. 391-392: I would appreciate numbers here, better even distributions of disturbance patch sizes per survey type.

Following the upper comment, we also added information on the patch sizes:

> *L. 378-381: **Within the subset region, NAFD has a mean patch area of 0.18 ha (±16.35 ha). Because of the aggregation method for GFC, the mean patch area is a lot larger, with 2184.69 ha (±5951.09 ha). IDS patches are on average 47.08 ha in size with a standard deviation of ±151.45.***

L. 393: The whole paragraph on spatial agreement is very long and hard to read. You do not need to write out every information in text, especially when you provide a table at the end. Describe the big pattern (most datasets do not align well, and highlight where they align more). The rest we can read out of the table. This is a general comment for the whole result section and this way, you can considerably shorten the text.

Thank you for the comment. We shortened the results section, highlighting only the most general findings.

L. 428: Table 3: Do the spatial overlap fraction [%] refer to % disturbed area overlapping of the horizontal or vertical listed dataset? (i.e. does 2% of the GFC area align with IDS or 2% of IDS aligns with GFC?)

The fraction corresponds to the horizontal listed datasets, so e.g. 2% of total IDS area overlap with GFC.

Thank you for pointing that out, we clarified it in the caption. It now reads:

> *Table 3: Unique overlapping disturbance events across all dataset pairs. Point-based datasets are evaluated using different buffer sizes. For point-based datasets, the numbers indicate the total number of events per dataset overlapping with another. For spatially explicit datasets, the first row in each block reports the total number of overlapping disturbance events, while the second row gives the proportion of each dataset's disturbed area that overlaps with the dataset in the respective column (i.e., percentages refer to the dataset of that row). **For example, among the overlapping events, only 2% of the total IDS area overlaps with the GFC-affected area.** The right column* Total *presents the total number of events from the original datasets in the study period.*

L. 477: Figure 5: I really like the figure this is super interesting and insightful!

Thank you very much!

L. 499: What is the overall share of private and publicly owned forests in general in the dataset?

In the study period (2001 to 2010) the share is 64% (100620 plots) to 36% (57613 plots) of private and public plots respectively. Including all years (1957 to 2022), the difference increases to 74% private (938229 plots) to 26% (331749) plots.
We added the overall share in Section 4.5.1:

> *L. 461: In the study period (2001 to 2010) the share is 64% (100620 plots) to 36% (57613 plots) of private and public plots in FIA respectively.*

L. 521: Table 5 could go to the supporting information. To me the differences do not seem that pronounced in this case and simply describing that in the text and referring to the table in the supporting information should be fine and give more focus on the key results.
Thank you for the suggestion, we agree and moved Tables 5 and 6 to the Appendix. They are now Tables B2 and B3.

L. 539: Given that ownership does not really explain a lot of variability, why did you keep your extensive analysis of this factor in the section before? This seems not proportional, as this analysis for the effect of state (and maybe different survey (teams) per state) seem to have a much larger impact on uncertainty in disturbance record timing.
Thank you for the thoughtful consideration. We agree that ownership explains relatively little of the overall variability in temporal uncertainty compared to state-level effects. However, we retained the ownership analysis for two reasons. First, ownership is an intrinsic attribute of the FIA data and was raised as a potential source of uncertainty – particularly in relation to plot swapping and fuzzing – in the first review round. Including this analysis allows us to explicitly demonstrate that ownership contributes only marginally to temporal lags, thereby addressing this earlier concern. We also believe this is important information for users.

Second, while the overall effect size is small, ownership-related differences are not uniform across states. When inspecting the per-state patterns, some states exhibit meaningful differences between public and private forests (e.g., larger lags in private forests in Wisconsin and Indiana, and smaller lags in Nebraska). These cases motivate retaining the analysis to illustrate that ownership effects exist but are highly context-dependent. We can make these patterns more transparent by adding the per-state figures to the manuscript if desired. We added the per state lag in private and public forests to the Appendix (Figures A7, A8, A9 in the new manuscript). Below is an example of the FIA-IDS comparison:

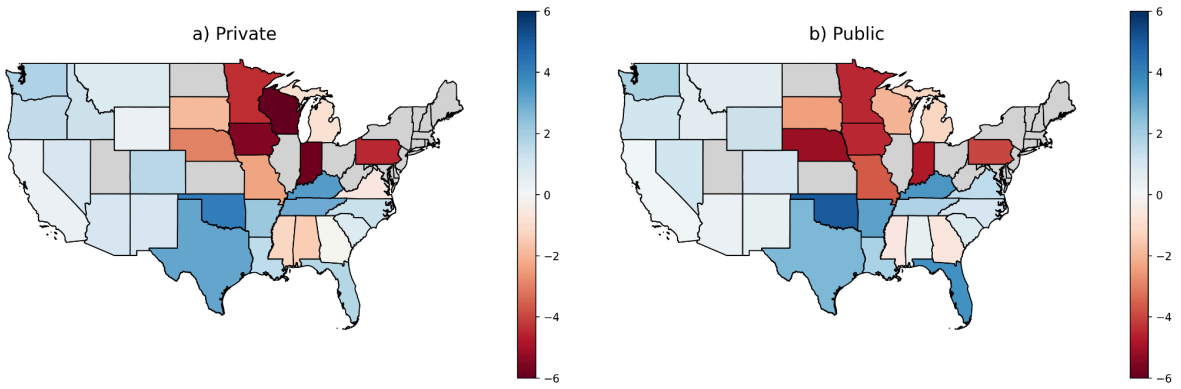

Figure R1.2: Comparison of mean temporal lag between FIA and IDS across U.S. states, grouped by the difference between public and private forests: (a) private forests and (b) public forests. Negative

values (in red) indicate earlier mortality reporting by FIA relative to IDS, while positive values (in blue) indicate later reporting by FIA.

We added a sentence in the results section:

> *L. 472-474: A per-state comparison of lags for private and public forests (Figures A7, A8, A9) shows generally consistent patterns across the U.S., but also highlights state-specific differences and a clear contrast in the magnitude of timing differences between the western and central–eastern states.*

L. 549 - 551: I agree with your interpretation, but this rather belongs to the discussion. Another factor could be, that complex topography makes disturbance extent detection and extent identification more difficult due to shadows and image quality issues at slopes for the IDS.

Thank you for pointing that out, we moved that part to the discussion. It is now in:

> *L. 666: This pattern is also reflected in our model results for FIA, which show a strong association with elevation. In situations where accessibility constrains field surveys, remote sensing data can help bridge this gap by providing consistent observations even in remote areas.*

L. 560: This is a very vague aim. What are drivers of disturbances? (harvesting pressure, climate change, land use reforms?) And what kind of predictive models? Where disturbances will likely occur in the future or not? You need to become more specific here and then make an argument, why your combined dataset is better equipped to the/one job than the other datasets (especially as you so far as I understood did not address the specific uncertainties associated with each dataset – combining propagates errors and uncertainties into the new one).

Since we removed the NDVI comparison, this statement is no longer the manuscript.

L. 565: How big is the area covered by this analysis and how much area disturbed per dataset?

The area in this analysis section is 3613.3 km^2. In the new manuscript, we removed the NDVI analysis, see also R1C1 and R1C2.

L. 566: Figure 7: I like the figure, but as I argued in my comments to the method, I do not trust NDVI here as a single disturbance proxy to compare all these datasets, but like the idea to map all datasets on comparable metrics to see how these trends behave. But you will always run into the problem of circularity, as the only datasets you will get across all areas are probably those remote sensing data sources, which have been used for the remote-sensing based disturbance products. Hence these will most of the time align better with this metric, as they base on them, while field-based products will include subtler and non-stand-replacing disturbances, which are hard to detect via remote sensing. Hence this is not really a valid comparison.

Thank you for the thought. In line with R1C1 and R1C2 we removed the NDVI analysis and replaced it with a trend analysis comparing East and West U.S. in terms of event count and area affected per dataset, and if possible per disturbance agent. It's now Section 4.6.

Discussion:

L. 584-585: I do not really agree with this statement. Bark beetles affect the whole disturbance legacy structure, which in turns is an important factor in forest recovery. Additionally, can fires equally affect the canopy, but as ground fires, fires do not necessarily affect the regeneration capacity of forests (in some systems it is even necessary to initialize forest recovery, as f.e. in the southern Rockies). Please remove or revise this statement and add references for your statements.

Thank you for pointing that out. We rephrased the whole introductory paragraph of the discussion. It now reads:

> *L. 547-551: Forest disturbances have multiple causes and affect forests in varied ways**, shaping how they can be observed and interpreted. Some agents, such as windstorms, cause abrupt structural damage (Forzieri et al., 2020), while others, such as insects or drought, may act more gradually or interactively, altering forest composition, vitality, and recovery potential (Meddens et al., 2012; Kurz et al., 2008; Clark et al., 2016). These differing disturbance mechanisms influence not only their ecological consequences but also how readily they can be detected, attributed, and quantified in observational datasets.***

L. 590: Please also make a statement on spatial agreement in this context.

We agree with this comment and have added a summary of the spatial agreement.

> *L. 552-556: Overall, we observe a relatively good average agreement (i.e. good accuracy) among the datasets in terms of disturbance timing and agents, but with considerable variability across individual events (i.e., low precision).* ***The spatial agreement is generally lower: while IDS captures most of the area detected by the two remote sensing datasets, only a small fraction of IDS area – at most 2% – overlaps with the remote sensing products. Between the remote sensing datasets themselves, spatial agreement is only moderate.***

L. 593- 594: You did not revise the methods, which would mean to improve methodological aspects in individual disturbance detection methods and to create/map a new disturbance dataset, with demonstrating its improved performance. You compare different datasets and look into variables, which might explain disagreement in datasets. Also, with setting IDS as your reference dataset, you assume (in this context) that it is closest to the “true disturbance state”, so you have something to compare it to. But this is in and of itself flawed, as also this dataset comes with uncertainties and errors. This is hard to come by, but should definitely be discussed at some point. At the same time, having this in mind, you do not quantify uncertainties of the methods themselves, but you look into the (dis)agreement between datasets. That is a difference and should be made clear. (but of course, you describe the potential sources for uncertainties in the datasets!)

We agree that this was not worded correctly. We shortened that paragraph and rephrased the goal of the study.

> *L. 556-559: The results show that differences between the datasets can be attributed to inherent uncertainties in detection methods, differences in spatial and temporal scales, and varying levels of detail in the disturbance records of each dataset. Below, we discuss how the uncertainties underlying the different datasets explain these mismatches, and provide guidance for users on how these differences might affect analyses and interpretations.*

L. 604: Yes, but how? How can we improve the surveys to reduce those shortcomings? Potential pre-disturbance state records? Tighter timing in doing the surveys? Including other data layers for disturbance mapping (remote-sensing based indices changes as base layer to support mapping?)

We added some sentences on this point:

> *L. 774: Increasing survey revisit frequency could allow for better detection of slow declines or compound events, which are often missed with the current 5–10 year interval (Cohen et al., 2016; Schroeder et al., 2014).*
>
> *L. 652: In situations where accessibility constrains field surveys, remote sensing data can help bridge this gap by providing consistent observations even in remote areas.*
>
> *For IDS L. 772: However, given potential uncertainties from manual interpretation and limited revisit frequency, IDS is best used in combination with other datasets for robust temporal analyses or validation.*
>
> *For GFC L. 783: However, given that it does not include information about disturbance agents, it needs to be combined with other datasets, preferentially ground-based, e.g., as reference data. In that case though, careful consideration of the uncertainties underlying such additional datasets.*
>
> *For all datasets L. 792: In principle, increasing the temporal resolution of the datasets, e.g. with revisits for different phenological stages, could support earlier and more timely detection of disturbances.*
>
> *L. 797-799: Satellite-based datasets could bridge this gap. While trend break detection algorithms based on Landsat are likely to be limited to annual scale, new sensors such as the ones on board of the Sentinel constellation, with both high spatial and frequent revisit times, as well as global coverage, might allow for sub-annual disturbance detection.*

L. 621: yes, but this threshold is also a strength, as it standardizes the recording of the disturbances, as you need to reach a certain extent and progress in mortality, to be recognized. The problem here is of course the miss of small-scale and low severity disturbances, but also problems across forest ecosystem types, which can vary in their spectral expressions. So, a disturbance detected in costal rainforests might be differently severe disturbed than one in evergreen forests in the Rockies.

Thank you for pointing that out. We incorporated that in the section.

> *L. 583-586: For disturbances to be detected, the signal must reach a detectable threshold for satellites to identify and quantify the affected area,* ***which standardizes detection but also makes disturbance severity a critical factor in satellite-based mapping*** *(Masek et al., 2013; McDowell et al., 2015).*
>
> ***Moreover, similar disturbances can manifest differently across ecosystem types, resulting in varying detection outcomes (Cohen et al., 2017).***

L. 626-627: I would argue that the missing disturbance agent attribution is one shortcoming, but calling this “simplicity” feels a bit off, considering the pioneering work Hansen did to finally consistently map disturbances in space and for some time. Also, I would point out other challenges with the GFC: While the GFC product enables globally consistent monitoring, its reliance on spectral change detection means that similar declines in NDVI or tree cover can have different ecological meanings across forest types. Also, the reliance on

single-date maximum declines in tree cover and NDVI can misclassify gradual or low-severity events, underestimate partial canopy disturbances, and confuse temporary changes (e.g., phenology, noise) with actual loss.

We agree that the wording was unfortunate as indeed the fact that GFC is still the longest globally consistent map of forest disturbances was the reason we chose it for our analysis. We added some more limitation/ explanation to the paragraph:

> *L. 587-590: The approaches used to determine forest disturbances differ notably between GFC and NAFD. In GFC, using the maximum annual NDVI decline (Section 2.5) can result in gradual or low-severity disturbances being missed or misclassified, leading to underestimation of forest loss (McDowell et al., 2015). Conversely, short-term fluctuations caused by phenological changes or sensor noise can be erroneously identified as loss events, even though they might be only temporary.*

L. 629: Well but the Hansen team used a decision tree model, which has been trained with training data as well, which has been equally collected by humans (when I recall this correctly). Human labelling of training and reference data is always flawed, but I do not know any way to get around this. So, this is something you will inherently find in most remote sensing-based mapping methods and downplays the NAFD method unjustified.

It is true that human labeling is still the common method for training data. The statement is mainly there to inform about the fact, but we rephrased the discussion about the NAFD methodology. It now reads:

> *L. 596-601: In contrast, the NAFD algorithm uses a more complex modeling approach by applying Random Forest models (Section 2.6), which are well suited for capturing complex relationships between spectral features and disturbance types (Prasad et al., 2006). The combination of multiple decision trees increases accuracy and reduces overfitting as the ensemble grows (Prasad et al., 2006). In our temporal comparison, NAFD detects disturbances earlier than the other datasets – on average by about half a year (Figures 4, B1). This could reflect the ability of the framework to capture subtle spectral changes preceding the disturbance detection by other approaches.*

L. 631: And here you point out that the team used different datasets to validate their product, which is very good practice!

Indeed.

L. 635: I do not really understand why the “black box” argument is one at this place. Random forests are pretty straight forward (a combination of many decision trees) and outperform single decision trees products in all cases which are known to me in accuracy and precision. Variable important plots give a good idea about which variables are most important to the random forest models to classify. This of course is problematic when you want to infer causality of f.e. what drives a disturbance, but for the purpose of mapping and prediction, this is definitely the preferred method compared to the GFC single decision tree.

We rephrased this part. It now reads:

> *L. 604: In contrast, the NAFD algorithm uses a more complex modeling approach by applying Random Forest models (Section 2.6), which are well suited for capturing complex relationships between spectral features and disturbance types (Prasad et al., 2006).*

L. 644-647: You tend to repeat at each section what you have done. It is sufficient to state that once clearly in the methods and state your results and lead into the discussion without repetitions. This can considerably shorten your (pretty long) manuscript.
Thank you for the advice, we removed result repetitions and tried to shorten the discussion to focus on the main interpretation.

L. 657-658: Yes, but what would you trust more, the polygons spanning large areas or the pixels identified by the RS products? Did you manually check some disturbance incidents with independent data (such as high resolution google earth images), to get an intuition which data products capture the disturbances in which way? I can imagine IDS can overestimate (surveyors being not so neat and precise in delineating some areas), and remote sensing products over and underestimate disturbances, depending on the areas and disturbance types.
We have not checked the delineation of disturbances ourselves, but we agree with the reviewer statements, that IDS might overestimate, while the remote sensing product underestimates the disturbance area. A comparison of IDS, Sentinel-1, and high-resolution PlanetScope data is provided in Müller et al. (2025). In their study, Figure 7 illustrates differences in disturbance delineation, and an additional figure (Figure R1.3) highlights uncertainties in IDS polygons, including cases where disturbances are mapped over urban areas.

Figure R1.3: Disturbance patches from a 2020 wind event (left) and 2020 bark beetle events (middle and right) in the southeastern United States, overlaid on PlanetScope imagery. The wind-related polygon extends across an urban area, while the middle bark beetle polygon is placed in a meadow rather than within forested land.

In the manuscript, we added a short part referring to that:

> *L. 627: IDS may overestimate disturbance extent due to surveyor bias and manually drawn delineations, whereas remote sensing products may underestimate it because of their reliance on spectral signals and algorithmic limitations, as discussed above.*

L. 660: I would be interested in where the two remote sensing products agree and disagree and for which agents (classified by NAFD). This is indeed a very low agreement.
We agree that such an analysis would be valuable. However, since the GFC dataset does not include information on disturbance agents, we were unable to assess agreement in that regard. To provide additional spatial context, we have added a figure in the appendix showing the total affected area per state for the spatially explicit datasets (see Figure A2 in the manuscript).

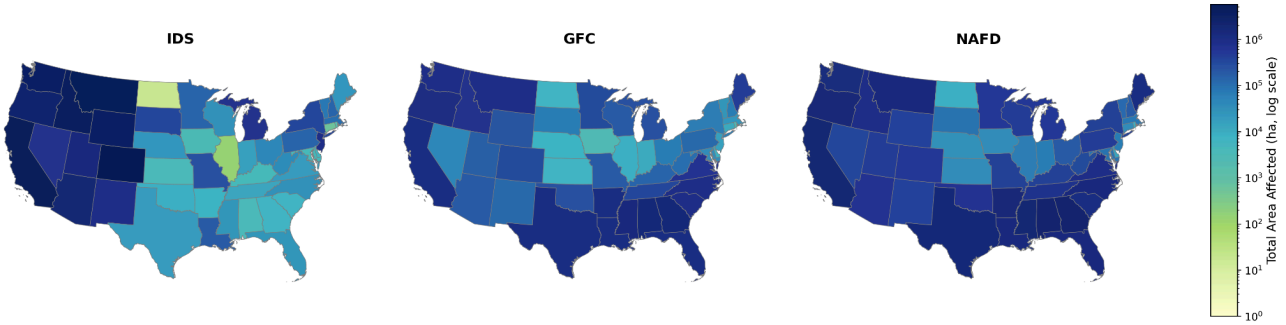

Figure R1.4: Total area affected per state of IDS, GFC and NAFD.

We also tested and visualized agreement at the state level (Figure R1.5, below) and tested the agreement per agent (based on NAFD original agents) and per state (Figure R1.6). Figure R1.6 is added in the Appendix as Figure A15 to highlight the differences between the two datasets.

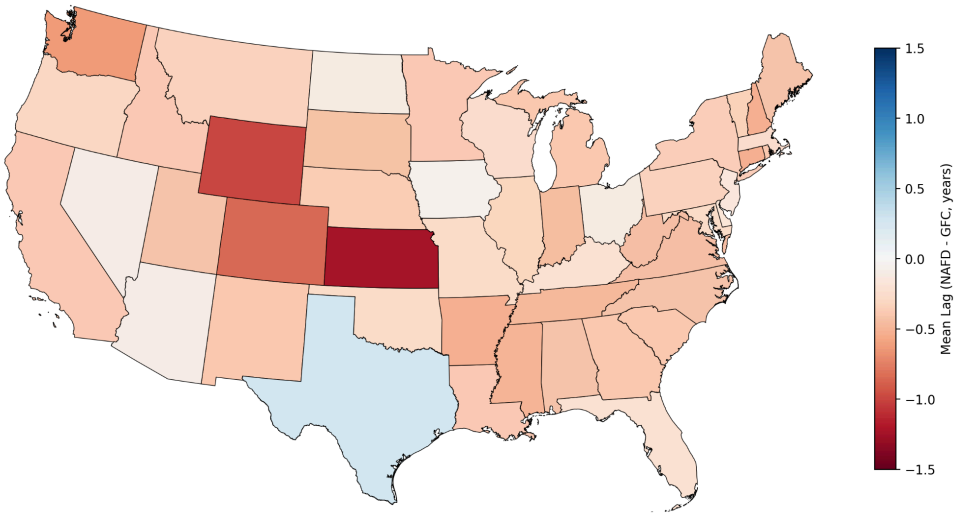

Figure R1.5: Mean lag of overlapping disturbance events in NAFD and GFC. Positive values indicate an earlier detection by GFC and negative values an earlier detection by NAFD.

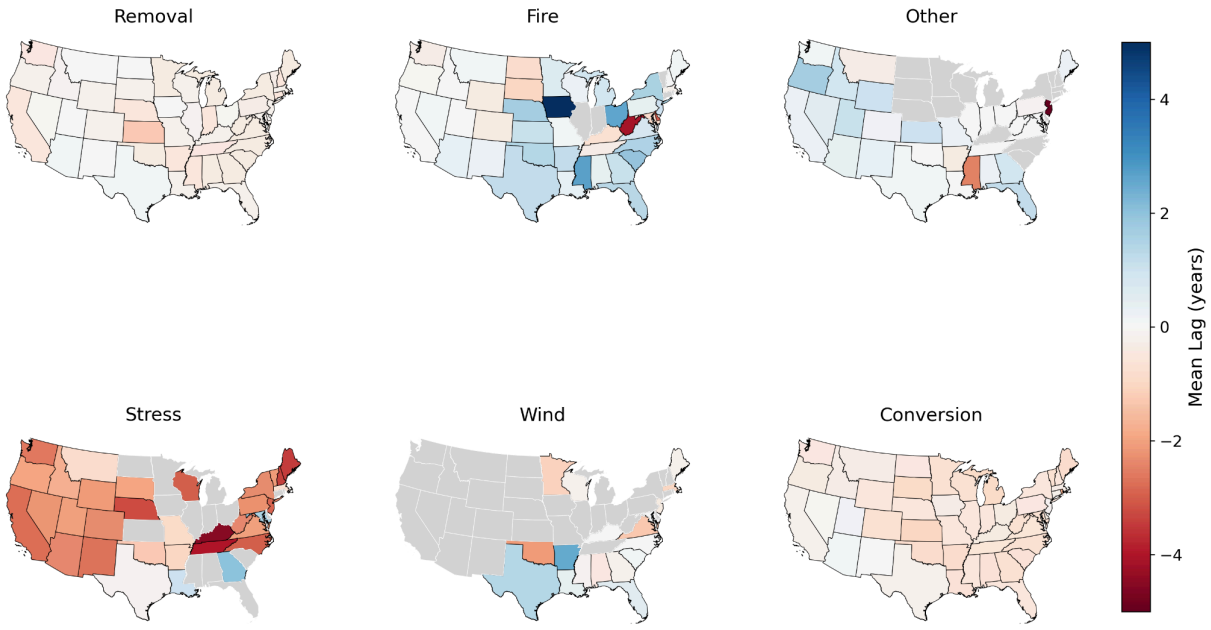

Figure R1.6: Mean lag of overlapping disturbance events in NAFD and GFC per state and original NAFD disturbance agent. Positive values indicate an earlier detection by GFC. Negative values show that NAFD reports earlier.

L. 702: Have you taken a look into high resolution images for the NAFD disturbance pixels what happened after the disturbances, so the NDVI increased that sharp?

No we have not done that, but we removed the sections about NDVI, so this no longer applies, see also R1C1.

L 720: one “and” too much

Thank you, we removed it.

L. 727: Do you have an idea or any indication, why there is zero overlap in the IDS Wind and NAFD wind agent classification?

NAFD only reports wind events in the East of the US - in total 11495 wind events in the 10 years. The maps and the heatmap are from the perspective of IDS, showing how often NAFD agrees with IDS. Further, the percentages are normalized to the IDS data.

Below we show the heatmap from the NAFD perspective (Figure R1.7), which indicates that 56% of events labeled as *Wind* in NAFD are also mapped as *Wind* in IDS. However, examining the absolute counts in Table R1.2 provides important context: although NAFD wind pixels most frequently correspond to IDS wind (239 events), this represents only a very small share of all IDS wind detections. From the IDS perspective, wind events mainly overlap with other NAFD classes, e.g. with NAFD class *Other* (62560 events).

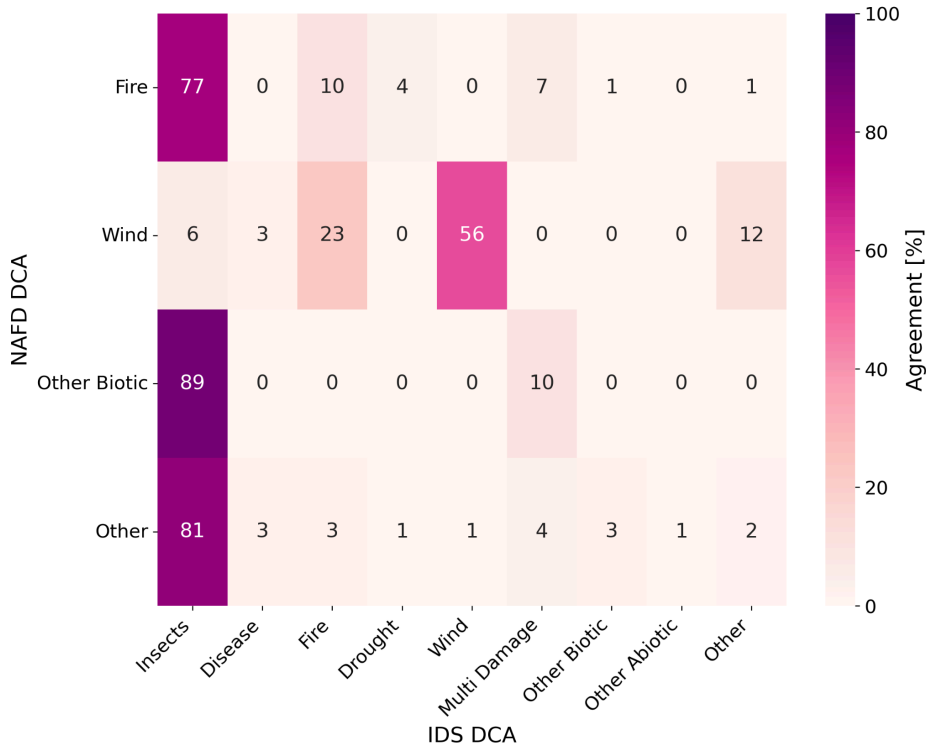

Figure R1.7: Comparison of disturbance agent agreement between NAFD and IDS shown as heatmap of the confusion matrix. It shows how often each NAFD-assigned disturbance agent category corresponds to classifications in IDS.

Table R1.2: Confusion matrix showing the absolute counts of overlapping events per disturbance agent between IDS (rows) and NAFD (columns). Wind events are in italics.

DCA NAFD -> DCA IDS	**Fire**	**Wind**	**Other abiotic**	**Other**
Insects	1406243	*25*	13962470	3972784
Disease	6106	*11*	70932	129455
Fire	172969	*99*	55924	170033
Drought	70609	*0*	27218	55416
Wind	*944*	*239*	*3107*	*62560*
Multi damage	120384	*0*	1488481	195622
Other biotic	25164	*0*	30898	153319
Other abiotic	383	*0*	1019	32129
Other	12721	*52*	1418	117073

L. 756-757: This is super interesting and important to know!
Thank you for the positive feedback.

L. 776: What are your suggestions in how to address these limitations and alignment problems?

To address the limitations and alignment challenges in agent attribution, users could consider integrating NAFD with complementary datasets that provide higher-resolution or inventory-based agent information, such as IDS or FIA. Combining multiple sources can improve confidence in both the spatial and causal characterization of disturbances.

We added:

> *L. 789-791: To improve agent attribution, NAFD could be integrated with complementary ground-based or higher-resolution datasets, such as PlanetScope or Google Earth, that provide more precise information on disturbance type. Structural information from sensors like Sentinel-1 may further aid in distinguishing certain disturbance types (Müller et al., 2025).*

L. 773- 774: Which criteria do they use to categorize more ambiguous classes like human activity or stress?

Stress is defined as “any event resulting in slow gradual loss of forest canopy, including insect damage, drought and disease, which often co-occur” (Schleeweis et al., 2020). Stress shows a dispersed spatial pattern.

We added the definition of stress in Section 2.6 Line 222.

Regarding human activity, they define the groups: removal (mechanical removal of trees) and conversion (land use land cover change). In the methods section, the authors write: “Plots originally labeled “Mechanical” change processes and which occurred in forest cover were relabeled to “removals” or “conversion,” depending on the land cover/use and timing observations in the plot records. At the end of the imagery time series, it is harder to determine with confidence whether a forest clearing activity was related to conversion or removals. We cautiously labeled all mechanical clearings as removals, if they occurred within six years of the end of the time-series stack, unless evidence of specific land use change was recorded.” (Schleeweis et al. (2020))

L. 779: You here touch on human activities aka human disturbances. These do not seem to be classified in the other datasets, which surprises me a lot as I assume that harvesting plays an important role in forest (disturbance) dynamics in the USA. I would appreciate if you can discuss a bit further how the datasets account for or treat human disturbances, if they do so. And if they don’t, to elaborate the implications of that in the data product comparison (as remote sensing product map all disturbances, including harvesting).

The inventory based datasets actually report human activities. In FIA the category is *Silvicultural or Land cleaning activity*. In IDS it is grouped under *Human activities*, with several sub-groups about the actual human activity. In this study, we grouped under *Other,* because our focus was on natural forest disturbances which tend to be harder to map and detect.

L. 791: Well, depending on what data I train a model, point-based information is actually a pretty good input, when I can trust that the point-based information has a good quality. Can you elaborate for what kind of modelling approaches this would be problematic? (as I see no problems here for training models, which classify pixels of remote sensing images, as I can assign a point to a pixel)

The issue is that point-based data always correspond to a given footprint, in this case a circular area of 7m diameter, and the data used to train the model often has a different footprint, for example 30x30m for Landsat. Therefore, in most cases there is a scale mismatch that needs to be considered, even if one is measuring exactly the same quantity in the point-based and the remote-sensing observations (which often we are not).

L. 834: You should at least mention the potential to prolong the dataset, as Landsat imagery is yearly updated, so with using the same methodology, one can extend the dataset until this/last year.

Thank you for pointing that out, we agree and added a sentence:

> *L. 788: In principle, the dataset could be extended to the present using annually updated Landsat imagery.*

L 835: I would disagree, the datasets do diverge immensely form each other, especially in the spatial coverage, which is a key feature for any analysis with disturbances. This is an interesting insight in and of itself, but shows how bad we do as a community in defining, identifying and mapping disturbances.

Indeed, the spatial agreement is low. The statement in Line 835 only referred to the temporal consistency, apart from spatial agreement.

We added two sentences addressing the spatial coverage. The paragraph now reads:

> *L. 792-796: Generally, the datasets show good agreement in disturbance timing. In principle, increasing the temporal resolution of the datasets, e.g. with revisits for different phenological stages, could support earlier and more timely detection of disturbances. However, despite the temporal agreement, spatial overlaps among datasets remain low, highlighting substantial divergence in the location and extent of disturbances. This discrepancy underscores the challenges in defining, detecting, and mapping forest disturbances consistently across datasets.*

L. 840: Yes, that is true, but what about better high-quality training datasets? How to account for low severity disturbances? What about multiple or secondary disturbances (humans doing salvage logging after an initial natural disturbance)? What about including more environmental information in the mapping process, to improve that? There is so much more to improving remotely sensed disturbance maps than sub-annual mapping strategies, which should be elaborated on here.

Indeed, there are more possibilities of improving disturbance maps. We added:

> *L. 801-804: Additional information about disturbance severity and specific impacts (e.g., leaf discoloration, legacy pattern, percent affected) are available for IDS and could in principle be added to regular forest inventories. This would allow to identify the most appropriate satellite dataset (optical, radar, etc) for each type of disturbance and associated impact, and possibly improve their detection regarding timing and spatial features.*

Conclusion:

L. 842: Technically you did not demonstrate it, as this would need one or several case studies applying the datasets to answer (research) questions and to demonstrate how different or similar the datasets would answer these questions.

We agree with the poor wording. We rephrased the paragraph, it now reads:

L. 806-809: In this study, we assessed the consistency of five forest disturbance datasets across the conterminous United States, highlighting the challenges of comparing and interpreting these data. Our results reveal varying levels of agreement, with the remote-sensing datasets generally reporting disturbances earlier than the others, underscoring the need for careful consideration of dataset differences when analyzing forest disturbance patterns.

L. 845: You focus a lot on the temporal agreement, while I would argue, that the spatial agreement is more important to a lot of research and practical questions. What is your argument for this focus and what is your take on the low spatial agreements and what does this apply for the use of these products in analysis?

In the revised manuscript, we added more information and discussion about the spatial agreement in Figure A2 showing the affected area per state in the spatially explicit datasets, in Figures A13 and A14 showing the trend in area in the eastern and western U.S., and in Sections 5.2 discussing the spatial uncertainty.

L. 853: Yes, and the multiple agents also make it difficult to identify one disturbance year, as the forest decline can stretch across several years.

That is true, we added that part at the end of the sentence, it now reads:

L. 815-818: These discrepancies stem from varying levels of detail in the datasets and the subjective determination of agents, ***as well as the prevalence of compound disturbance events, where multiple interacting stressors – such as drought followed by insect outbreaks – complicate clear attribution to a single cause and a single mortality year.***

Response to Reviewer #2

This paper attempts to quantify and describe the level of agreement between disturbance datasets that cover CONUS. This paper quantifies the level of temporal, spatial, event-count, and agent detection agreement between datasets, and discusses potential sources of error. It represents a considerable and worthy effort to corral different datasource describing a nebulous concept "Disturbance" into a unified framework so that different datasets could be evaluated. I have a few minor concerns in the framing of this work.
Thank you for acknowledging the relevance of our effort.

R2C1: The motivation for this comparison is presumably to make a predictive model. This motivation is stated in the hypothetical, but not explicitly used as a framing mechanism for the paper. If that is the motivation, I think it would add clarity to the paper and the methods for that to be stated directly.
We agree that our motivation could be laid out more clearly. Our motivation is to better understand the underlying uncertainties of each dataset and to raise awareness of these limitations among users. Recognizing and accounting for such uncertainties - for example, by incorporating confidence ranges for mortality years or combining multiple datasets to enhance spatial accuracy - can improve the robustness of disturbance analyses. Furthermore, since different datasets emphasize different disturbance agents, it is important for users to consider these differences when interpreting results or comparing data sources.
We rephrased the motivation in the introduction, it now reads:

> *L. 125-131: By comparing these five forest disturbance datasets, we aim to highlight the variability and underlying uncertainties associated with different observation systems. Our goal is to provide a systematic assessment of their consistency in disturbance extent, timing, and agent attribution. Through this comparison, we seek to identify dataset-specific limitations and offer guidance on their use, for instance, by incorporating uncertainty ranges for disturbance timing or by combining complementary datasets to improve spatial reliability. Although our analysis focuses on the conterminous United States, the approach is transferable to other regions and can inform the design of more robust, uncertainty-aware forest monitoring and classification frameworks (European Commission: Directorate-General for Environment et al., 2020).*

R2C2: A "Disturbance" is a concept that spans multiple fields, ecosystems, and meanings. From the methods, I inferred that the authors focused specifically on mechanisms involved in forest mortality. The definition of "disturbance" for the purposed of this analysis needs to be stated in the introduction.
Indeed, there are different definitions of disturbance, therefore we use the definition by White and Pickett (1985) (see below). We clarified that in this study, we consider only disturbances associated with tree mortality. It is found in:

> *L. 38-43: 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". In the inventory datasets used here, disturbances encompass not only tree death, but also early indicators such as*

discoloration and crown dieback. ***For comparability across datasets, here, we consider disturbance as any event that causes tree mortality***

R2C3: I would appreciate more discussion of the interpretation of dependent datasets. For example, I was unsure how to interpret the temporal differences between different datasets mutually based on Landsat. I was also unsure how to interpret Figure 7's NAFD vs NDVI time series, given the level of connection between the datasets.

We agree that the comparison of NDVI with the datasets was not independent, especially comparing NDVI with GFC and NAFD. We therefore removed this section and replaced it with a trend analysis of the West and East U.S. (see also R1C1 and R1C2)

Minor language edits:

Line 24: "end to show an advanced detection of disturbance events," confusing wording as "advanced" can mean "before" or "better resolution/ quality" in this case.

Thank you for pointing that out. We changed it to earlier, as they detect before the other datasets. It now reads:

> *L. 23: The satellite-based datasets tend to show an* ***earlier*** *detection of disturbance events, compared to the other datasets, possibly due to the inconsistent revisiting times of the inventory datasets (FIA and IDS).*

Table 1: consider adding an explicit row for sampling period. The descriptor "annually" appears in different rows for different datasets

Thank you for the suggestion, we updated the table by replacing the *Years* row with two rows: *Temporal coverage* and *Sampling rates*.

Line 453: "but still within their temporal resolution" — could you clarify if you mean annually here?

We agree that this does not sound clear. What we meant is that the mean lags are smaller than/ comparable to their typical temporal resolution of one year, since the temporal granularity is one year. Due to updates in the code and the results, the mean lag between IDS and NAFD increased to 1.8 years. We therefore rephrased the sentence, removing the clause. It now reads:

> *L. 414: Among the spatially explicit datasets, IDS generally records disturbances later than the satellite-based datasets, with average delays of 0.5 years compared to GFC and 1.9 years compared to NAFD.*

475: missing space

Thank you, we corrected that.

References

S. Archibald, C.E.R. Lehmann, J.L. Gómez-Dans, & R.A. Bradstock, Defining pyromes and global syndromes of fire regimes, Proc. Natl. Acad. Sci. U.S.A. 110 (16) 6442-6447, https://doi.org/10.1073/pnas.1211466110 (2013).

Bastos, A., Orth, R., Reichstein, M., Ciais, P., Viovy, N., Zaehle, S., ... & Sitch, S. (2021). Vulnerability of European ecosystems to two compound dry and hot summers in 2018 and 2019. *Earth system dynamics*, *12*(4), 1015-1035.

Bastos, A., Sippel, S., Frank, D., Mahecha, M. D., Zaehle, S., Zscheischler, J., & Reichstein, M. (2023). A joint framework for studying compound ecoclimatic events. *Nature Reviews Earth & Environment*, *4*(5), 333-350.

Bowman, D. M., Balch, J. K., Artaxo, P., Bond, W. J., Carlson, J. M., Cochrane, M. A., ... & Pyne, S. J. (2009). Fire in the Earth system. *science*, *324*(5926), 481-484.

Clark, J. S., Iverson, L., Woodall, C. W., Allen, C. D., Bell, D. M., Bragg, D. C., ... & Zimmermann, N. E. (2016). The impacts of increasing drought on forest dynamics, structure, and biodiversity in the United States. *Global change biology*, *22*(7), 2329-2352.

Cohen, W. B., Yang, Z., Stehman, S. V., Schroeder, T. A., Bell, D. M., Masek, J. G., ... & Meigs, G. W. (2016). Forest disturbance across the conterminous United States from 1985–2012: The emerging dominance of forest decline. *Forest Ecology and Management*, *360*, 242-252.

Cohen, W. B., Healey, S. P., Yang, Z., Stehman, S. V., Brewer, C. K., Brooks, E. B., ... & Zhu, Z. (2017). How similar are forest disturbance maps derived from different Landsat time series algorithms?. *Forests*, *8*(4), 98.

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.

Curtis, P. G., Slay, C. M., Harris, N. L., Tyukavina, A., & Hansen, M. C. (2018). Classifying drivers of global forest loss. *Science*, *361*(6407), 1108-1111.

DeFries, R. S., Rudel, T., Uriarte, M., & Hansen, M. (2010). Deforestation driven by urban population growth and agricultural trade in the twenty-first century. *Nature geoscience*, *3*(3), 178-181.

Emmett, K. D., Renwick, K. M., & Poulter, B. (2019). Disentangling climate and disturbance effects on regional vegetation greening trends. *Ecosystems*, *22*(4), 873-891.

Forzieri, G., Pecchi, M., Girardello, M., Mauri, A., Klaus, M., Nikolov, C., ... & Beck, P. S. (2020). A spatially explicit database of wind disturbances in European forests over the period 2000–2018. *Earth System Science Data*, *12*(1), 257-276.

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.

Forzieri, G., Dutrieux, L. P., Elia, A., Eckhardt, B., Caudullo, G., Taboada, F. Á., ... & Beck, P. S. (2023). The database of European forest insect and disease disturbances: DEFID2. *Global change biology*, *29*(21), 6040-6065.

Gao, Y., Skutsch, M., Paneque-Gálvez, J., & Ghilardi, A. (2020). Remote sensing of forest degradation: a review. *Environmental Research Letters*, *15*(10), 103001.

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.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., ... & Townshend, J. R. (2013). High-resolution global maps of 21st-century forest cover change. *science*, *342*(6160), 850-853.

Hicke, J. A., Johnson, M. C., Hayes, J. L., & Preisler, H. K. (2012). Effects of bark beetle-caused tree mortality on wildfire. *Forest Ecology and Management*, *271*, 81-90.

Hicke, J. A., Xu, B., Meddens, A. J., & Egan, J. M. (2020). Characterizing recent bark beetle-caused tree mortality in the western United States from aerial surveys. *Forest Ecology and Management*, *475*, 118402.

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.

Kurz, W. A., Dymond, C. C., Stinson, G., Rampley, G. J., Neilson, E. T., Carroll, A. L., ... & Safranyik, L. (2008). Mountain pine beetle and forest carbon feedback to climate change. *Nature*, *452*(7190), 987-990.

Masek, J. G., Goward, S. N., Kennedy, R. E., Cohen, W. B., Moisen, G. G., Schleeweis, K., & Huang, C. (2013). United States forest disturbance trends observed using Landsat time series. *Ecosystems*, *16*(6), 1087-1104.

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.

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.

Müller, F., Eifler, L., Cremer, F., Beck, P., Camps-Valls, G., & Bastos, A. (2025). Hybrid forest disturbance classification using Sentinel-1 and inventory data: a case-study for Southeastern USA. *EGUsphere*, *2025*, 1-37.

Patacca, M., Lindner, M., Lucas‐Borja, M. E., Cordonnier, T., Fidej, G., Gardiner, B., ... & Schelhaas, M. J. (2023). Significant increase in natural disturbance impacts on European forests since 1950. *Global change biology*, *29*(5), 1359-1376.

Potapov, P., Hansen, M. C., Stehman, S. V., Loveland, T. R., & Pittman, K. (2008). Combining MODIS and Landsat imagery to estimate and map boreal forest cover loss. *Remote sensing of environment*, *112*(9), 3708-3719.

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, *9*(2), 181-199.

Schleeweis, K. G., Moisen, G. G., Schroeder, T. A., Toney, C., Freeman, E. A., Goward, S. N., ... & Dungan, J. L. (2020). US national maps attributing forest change: 1986–2010. *Forests*, *11*(6), 653.

Schroeder, T. A., Healey, S. P., Moisen, G. G., Frescino, T. S., Cohen, W. B., Huang, C., ... & Yang, Z. (2014). Improving estimates of forest disturbance by combining observations from Landsat time series with US Forest Service Forest Inventory and Analysis data. *Remote Sensing of Environment*, *154*, 61-73.

Seidl, R., Schelhaas, M. J., & Lexer, M. J. (2011). Unraveling the drivers of intensifying forest disturbance regimes in Europe. *Global Change Biology*, *17*(9), 2842-2852.

Senf, C., Pflugmacher, D., Wulder, M. A., & Hostert, P. (2015). Characterizing spectral–temporal patterns of defoliator and bark beetle disturbances using Landsat time series. *Remote Sensing of Environment*, *170*, 166-177.

Sommerfeld, A., Senf, C., Buma, B., D'Amato, A. W., Després, T., Díaz-Hormazábal, I., ... & Seidl, R. (2018). Patterns and drivers of recent disturbances across the temperate forest biome. *Nature communications*, *9*(1), 4355.

Thom, D., & Seidl, R. (2016). Natural disturbance impacts on ecosystem services and biodiversity in temperate and boreal forests. *Biological Reviews*, *91*(3), 760-781.

White, P. S. and Pickett, S. T. A.: Chapter 1 - Natural Disturbance and Patch Dynamics: An Introduction, in: The Ecology of Natural Disturbance and Patch Dynamics, edited by Pickett, S. T. A. and White, P. S., pp. 3–13, Academic Press, San Diego, https://doi.org/10.1016/B978-0-12-554520-4.50006-X, 1985.