

Reviewer 2

We thank Reviewer 2 for the thorough and insightful review and for the positive evaluation of the dataset's novelty and quality. The comments helped us substantially improve the conceptual clarity of the manuscript in two respects. First, we introduced an ecologically grounded definition of forest disturbance and now explicitly mark the classes that represent non-lethal canopy cover loss rather than tree mortality. Second, we revised the classification scheme: several classes were renamed to better reflect what can be observed in the Sentinel-2 time series. A new figure with example time series now better illustrates the classes and when their segments begin and end.

The manuscript and dataset by Viehweger et al. presents a novel reference dataset on tree cover changes for Europe that includes timing and causal agents. As such, it presents an important addition to existing datasets, and I applauded the authors for investing the time to create this dataset. The novelty and quality certainly justify consideration for publications in ESSD. That said, I have several major and minor points that need to be addressed before publication.

Major points:

- First: What is the definition of a disturbance? A disturbance – in the classical sense – describes a relatively abrupt loss of biomass that allows for the establishment of new vegetation from freed resources (mostly light). This is often also described as canopy mortality or tree cover loss. As thinning, drought and defoliation are included, I wonder whether the dataset mixes tree loss events with changes in leaf area that do not lead to the death of trees. A clear definition (at best ecologically motivated) is needed.

Thank you for this valuable comment. Indeed we were missing a clear definition on what kind of events we consider for the study. We have added a definition in the Methodology:

“To homogenize datasets, we need to define which events affecting the canopy are labelled during re-interpretation. Disturbances affecting ecosystems are defined as events which "disrupt ecosystem, community, or population structure and change resources, substrate availability, or the physical environment" (White and Pickett, 1985). For forest ecosystems this usually requires the death of at least a single canopy tree (Oliver and Larson, 1996; Runkle, 1985). During re-interpretation we had to restrict this definition to only consider disturbances which affect the canopy in such a way to be visible in the Sentinel-2 time-series. This excludes low-severity selective logging or other single tree mortality which is not visible due to affecting only a small fraction of a single Sentinel-2 10x10m pixel (Frolking et al., 2009). It does however include disturbances which are also of low severity but cover large areas. This includes for example commercial thinning or dieback of spruce trees within a mixed forest stand. Additionally to these forest disturbances we also label non-lethal canopy cover loss, which are visible in the satellite time-series. This includes temporary defoliation due to insects or early leaf senescence due to drought conditions. We include this canopy cover loss because even though this loss is not necessarily fatal for trees, it can lead to vitality loss and consequently increased mortality in following years (Davidson et al., 1999; Frei et al., 2022). This additionally recorded non-lethal canopy cover loss is labelled in the same way as forest disturbances with tree mortality.”

In any case we would consider thinning a disturbance, since thinnings remove trees and are explicitly undertaken to free resources (light, water) for the remaining trees.

We also noted in Table 1 which classes do not represent tree mortality but only non-lethal canopy cover loss.

- Second: How did you interpret routine versus non-routine management interventions? How did you make sure that management interventions were not triggered by bark beetle and thus should be classified as salvage logging? Already a single bark beetle tree can force managers to cut a whole stand to prevent spread (sanitation cutting). I doubt that this can be seen in the S2 time series.

Sanitation logging due to a single bark beetle tree is indeed impossible to see in the Sentinel-2 time-series. Because of this we rely mainly on first level disturbance agent products to assess disturbance agents in the time-series. For campaign 1 we received polygons from forestry services which documented disturbance polygons and management in response to disturbance.

Where we did not have this ground truth data available, we relied on VHR imagery. If the agent could not be conclusively assessed using available data a best guess was labelled by the interpreter and medium labelling confidence was assigned to the sample. We have added a qualifying statement in chapter 3.1 to make this restriction clearer.

“A high confidence means that both timing and cause of disturbance are confidently assessed. For medium confidence the timing of the disturbance has high confidence, but the assessed disturbance agent is of low confidence. In practice this means, that for tasks classifying the disturbance agent, only samples with high confidence should be used.”

- Third: Related to my first two points, the definition and inclusion of segments and events is not properly described, nor how they were separated during interpretation. This relates mainly to Table 1, for which I have several remarks:
 - o Re-vegetation is an odd name. There will always be vegetation establishing on disturbed sites. Maybe „recovery“ or „post-disturbance“ is a more suitable term?

This is indeed a hard class to properly name and we went through different names during the preparation of the dataset. We eventually landed on revegetation because, even though they are rare, there are examples of land use change in the dataset, where no re-vegetation happens after a disturbance. We explicitly did not name this class recovery because in a previous internal review this name was flagged to imply recovery of ecosystem functions and services which were previously provided by the forest. For the re-vegetation label we explicitly look for re-establishing vegetation on the disturbed area as evident through an increase in satellite vegetation indices. If this re-vegetation signal is not present (as can be the case for land use change), delayed (in the case of severe wildfire or severe disturbance in areas with short growing seasons) this label is not set until greening is visible. We have added an example of what this label looks like in practice in Figure 3 and a more detailed explanation on when this label is set:

“Figure 3a) shows a typical pattern caused by a clear cut. The time-series shows an abrupt and large decrease of vitality, followed by slow increase in vitality as vegetation regrows. It is often not possible to tell from the time-series if revegetation after a clear cut is happening with trees (class 121) or the revegetation is only with shrubs and grasses (class 123). In these unclear cases, we assigned the broader class 120 (revegetation). For the revegetation class we explicitly look for re-establishing vegetation on the disturbed area as evident through an increase in satellite vegetation indices. If this revegetation signal is not present (as can be the case for land use change) or delayed (in the case of severe wildfire or severe disturbance in areas with short growing seasons) this label is not set until greening is visible. Revegetation segments mostly end with the end of available

data or - theoretically - with reaching a phenological stable state in the observed satellite vegetation indices, that can be similar to or different from the pre-disturbance stage. "

- o When does revegetation end? Once the canopy is closed? How is canopy closing defined?

Revegetation ends with the end of available data or with reaching a phenological stable state in the observed satellite vegetation indices. Canopy closing after a thinning was defined as the return of vegetation indices to previous levels. In this particular case this means, that the spectral signal from the ground is not present anymore in the phenology. We have added a visual example for this in Figure 3 and explained this below the figure:

"Figure 3b) shows a typical pattern caused by thinning. In this case the time-series shows a less pronounced decrease in vitality compared to a clear cut, together with fast recovery of vitality in subsequent years, as the canopy closes again. The end of the canopy closing segment after a thinning was defined as the return of vegetation indices to previous levels."

- o How did you check whether a harvest was planned or not? There could have been a single bark beetle tree that triggered harvest.

See our response to the Second point.

- o Why not use unplanned harvest, but salvage logging as name? There is more than salvage cutting, sometimes foresters do sanitation cutting to prevent spread, even though the trees are not infested yet. This is also unplanned but not salvage logging.

This is a good point, we have renamed this class to "Salvage/Sanitation".

- o Bark beetle can be relatively short (1-2 years), but decline is often used for slower processes. I would reconsider this naming choice.

This is a good point, which reviewer 1 also noted. We have renamed decline to dieback.

- o Gypsy moth often does not kill trees, as trees can resprout. Did you check this?

Yes, for the few areas with confirmed gypsy moth infestation, trees always resprouted after the end of the growing season. We have added an example in Figure 3 and made this distinction clear in the Table by marking classes which are not necessarily lethal.

- o Is drought a disturbance agent? How did you verify this? Often drought incites biotic disturbances, which cannot be ruled out and would overlap with the biotic category.

Drought indeed often only weakens trees, while tree mortality is caused by other biotic factors. In the satellite time-series, we were mostly able to identify premature leaf senescence as a direct consequence of drought. We have renamed this class to "Drought-induced leaf senescence" to be more specific to this particular effect of drought on the canopy structure. In line with our forest disturbance definition we also marked this class as a non-lethal event.

- o Overall, a bit more description would help to understand the different categories. Maybe add a description column to the table with example figures?

This is a good point which was also brought up by other reviewers. We have added Figure 3 to show examples of disturbances and their time-series.

- Fourth: Disturbance timing relates to the first time the disturbance was evident in sentinel-2, but this might be much later than the actual disturbance event. The authors acknowledge this at some point in the manuscript, but this could be highlighted a bit more.

We have now highlighted this in section 2.2:

“For the undisturbed class 110, the end of the segment is always set to the last satellite acquisition where undisturbed forest is visible. The disturbed label is set to the first acquisition with a clearly visible disturbance. The actual disturbance data therefore must have happened between the end of the undisturbed segment and the disturbed label. This is the reason in the figure for the gap between the 110 segment and the vertical yellow bar denoting the thinning. The period between these was likely snowy, or otherwise obscured a more exact date of the thinning.”

And in the discussion:

“All labels in DISFOR were timed using the Sentinel-2 time-series. A label therefore marks when a disturbance first became visible in Sentinel-2 observations, which is not necessarily when it actually occurred. For abrupt events such as windthrow or clear cuts, this lag is mainly caused by clouds, snow or topographic shadows. For insect infestations, there is an additional lag between the start of the infestation and when an infestation becomes visible in the satellite data. Sentinel-1 is independent of cloud cover and sensitive to structural changes of the canopy. It could therefore detect disturbances such as windthrow before they are visible in Sentinel-2 (Rüetschi et al., 2019). To account for this, the end of the stable segment before a disturbance was consistently set to the last acquisition on which the forest was visibly undisturbed. The actual disturbance thus lies between the end of the preceding segment and the start of the disturbance label, and multi-sensor applications can treat this interval as uncertain. “

Minor points:

L. 45: The „isn’t“ is colloquial, please use “is not”.

We changed this.

L. 51: Rephrase to: „Because of this, there is...”

We have changed this to: “Because of this, some “second-level” disturbance agent products already exist.”

L. 62: Maybe don’t name the state of vegetation here already, as it can be confusing without clear definitions.

We have removed this part.

L. 66: What are short-duration disturbances and how are they characteristic? Also unclear in the second research gap stated below. Short duration can be anything from a few seconds to 1-2 years. In the US-American literature, short duration disturbances characterized disturbances <5 years, but I believe the authors refer to shorter time frames. Clear definitions are needed.

Good point, we have now consistently renamed short-duration to short-lived to signify that we are not talking about how long the actual disturbance event is, but for how long the post-disturbed state persists. In the intensely managed European forests, it can happen that within a year a stand is thinned, then affected by windthrow and within days to months salvage logged. These are all different disturbances with different severities and effects on the ecosystem, however they never persist for a long time. These kinds of progressions would not be captured by annual labelling.

L. 67ff: Please write out the last paragraph, instead of using bullet points.

We rewrote this paragraph: "With DISFOR, we address three research gaps. First, training data for disturbance agent classification rarely provide precise intra-annual timing of disturbance events. This limits the development of models that exploit the dense temporal sampling of Sentinel-2. Second, few data sources capture short-lived disturbance events. An example would be a sequence of thinning, windthrow and salvage logging all occurring within a single year. These quick successions of disturbances are characteristic of intensely managed European forests. Third, existing first-level disturbance agent data sources are heterogeneous in format, thematic scope, spatial and temporal resolution, and quality, which hinders their combined use for model training. By re-interpreting samples from multiple first-level sources in a consistent way using the Sentinel-2 time series, DISFOR provides harmonized, temporally explicit labels that address all three gaps."

L. 94: Consider rewording into high-severity and low-severity disturbances, as stand-replacing refers to a spatial unit (a stand), but your interpretation is at the pixel level. Also, please give clear definitions of what constitutes a stand-replacing/non-stand-replacing disturbance.

We have completely changed this section to include our definitions of disturbance. See response to Point 1.

L. 124: Were the tiles selected randomly or subjectively? Any inclusion criteria (i.e. forest cover) used? From eyeballing I would think they're a bit too regularly spaced for random selection.

They were not entirely randomly selected. They were stratified by biogeographic regions and distribution of Flux towers within Europe due to other requirements in the project. We have added a qualifying statement: "In the first stage, 13 Sentinel-2 tiles covering various biogeographic regions of Europe were selected. These Sentinel-2 tiles were not randomly selected, but hand-picked to include flux towers due to other requirements not relevant to this study."

L. 126: Did you sample equalized by disturbance/non-disturbance stratum? Can you please provide the stratum weights? Those are needed for estimation. Generally, a bit more context on the sampling design would be warranted.

We have added additional information on the sampling: "This approach used information on forest disturbances between 2019 and 2024 from the European Disturbance Atlas (Viana-Soto and Senf, 2025) as strata to increase the amount of samples in disturbed forest areas. All strata were randomly sampled. The undisturbed stratum was sampled at a density of 1 point per 3,000 ha and the disturbed stratum was randomly sampled at a density of 1 point per 250 ha to increase sample density in the disturbed stratum. We additionally sampled a 100m buffered area around disturbance patches at a density of 1 point per 1,000 ha to be able to get a better separation of disturbed and undisturbed areas in disturbance edges, which the higher spatial resolution of Sentinel-2 compared to the European Forest Disturbances Atlas should allow."

However it is important to note, that the dataset was explicitly built as a training dataset and not to validate map accuracies. We have also made this more clear in the introduction "DISFOR is explicitly intended to be used as training data for multi-temporal classification models. Due to the lack of probabilistic sampling the dataset cannot be used to validate map accuracies."

Section 2.2.: Many line breaks and one-sentence paragraphs. Don't break text into too many paragraphs but find a good flow within paragraphs, with paragraphs guiding the reader from one message to the next.

Thank you for this note, we have tried to improve the flow of text in this section.

L. 160: Bark beetles can also be relatively fast. The distinction between short and long disturbances is a bit arbitrary and never really defined, nor supported by literature. Also, if the event changes the canopy, it should be considered a change event, even though it changes it over several years (like in unmanaged bark beetle).

We agree that the distinction between short and long disturbances is arbitrary. Our distinction stems from the limitation of the maximum 3 days temporal resolution the satellite data provides. We have tried to sharpen this definition in the text:

“Each sample time-series was temporally labelled using the entire cloud-free Sentinel-2 time series. For the assignment of reference labels a distinction is made between segments and events. For our study this distinction is based on how a change is observed in the Sentinel-2 time-series, not on the ecological duration of the underlying disturbance. Temporal segments are periods over which a distinct condition of the forest persists or a change progresses gradually across multiple acquisitions (Cohen et al., 2010; Kennedy et al., 2010). This includes stable undisturbed forest, revegetation or bark beetle dieback in unmanaged stands, where trees die progressively over months to years. Events are changes that occur between two consecutive cloud-free acquisitions without observable intermediate states, such as clear cuts, wildfire or windthrow. Both disturbance events and disturbance segments represent canopy change. Events can be understood as segments with identical start and end date.”

We have also reworded the references to short-duration disturbances to short-lived disturbances. See also our response to your comment on L.66.

L. 206: Please rephrase „used first level data“ into „the first level data used“; otherwise data was bought second-hand ;-)

Thanks :) Removed *used* altogether “If the first-level disturbance agent product [...]”

L. 231ff: There is some confusion between model accuracy and map accuracy. You will need an independent sample following a proper probabilistic sample design for assessing map accuracy (i.e. estimating the probability of any a pixel being correct). I suggest using the terms model and map accuracy specifically.

We have specified model and map accuracy in this paragraph to make the distinction clearer: “To enable consistent validation of model accuracies for models trained with the provided DISFOR dataset, two lists of non-overlapping sample IDs are included to optionally split the dataset into a training and validation set. For the model validation set 10% of the full dataset was held out. The dataset was constructed by drawing a random sample of the defined clusters to reduce spatio-temporal correlation between the training and test set. However, it is important to note that since all the data was labelled using the same data sources and methodology, the model validation data will still be very similar to the training data. This model validation dataset can not be used to derive map accuracies, since its samples were not collected using a probability sampling design. For a rigorous study a fully independent validation dataset collected using a probability sampling design is necessary to also assess map accuracy (Olofsson et al., 2014; Reinosch et al., 2024). ”

Figure 6: The increase in planned harvests after 2018 makes we wonder whether this class actually includes a lot of sanitation (and potentially salvage) logging. What other reason would result in an increase in harvest after one of the biggest waves of bark beetle in recent history.

We have removed this figure altogether, since it only reproduced the sampling and thus could not be interpreted in the way we did. Namely, campaign 2 is the only campaign in the dataset which samples routine management in a meaningful amount. It only sampled disturbances occurring between 2019

and 2024. This is the sampling artifact which was present in this figure causing the increase of planned harvests after 2018.