

Reviewer 3

We thank Reviewer 3 for the very positive assessment of our work and for the helpful suggestions, which improved both the content and the readability of the manuscript. Most importantly, following the reviewer's recommendation, we added a dedicated Discussion section. It addresses the complementarity of DISFOR with other sensors such as Sentinel-1 and Landsat, the use of known event dates to narrow timing uncertainty, and our choice of fully manual time-series segmentation. In addition, prompted by the reviewer's remarks on low-severity and non-lethal disturbances, we formalised which canopy changes are considered during re-interpretation.

This manuscript describes a reference dataset for common forest disturbances in Europe. It is an important contribution with high relevance for many researchers as it provides detailed, timely resolved information about timing of abrupt disturbance events and periods of a certain condition/process together with labels of all events and periods. This kind of reference dataset is mandatory for proper classification of disturbance events. It builds upon existing datasets that usually lack of completeness, either thematically or with respect to exact timing of events and preceding and following processes. The described reference dataset allows for the reconstruction of sequences of multiple events, regeneration and management, providing a clearer picture of possible development pathways. In addition, the dataset is designed for easy access and flexible usability, and it is accompanied by snippets of spectral data from Sentinel-2, which helps to implement the data in deep learning approaches. I thank the authors for this important work and hope it will be continued in future.

Abstract

It is concise, links to the dataset and provides a brief description. However, acronyms and abbreviations should be defined.

We have added the long forms of used acronyms in the abstract. It is to note however, that neither FORWIND nor DISFOR are abbreviations. They do not have an official long form.

Introduction

Acronyms and abbreviations should be defined at first use, also in the main part of the manuscript. The introduction provides a good record of existing work, identifies gaps and explains the advantages of the introduced DISFOR dataset. One of the few key papers not yet included is by Stahl et al. 2023 (<https://doi.org/10.1016/j.rse.2022.113416>).

We have added abbreviations. Thank you also for bringing Stahl et al. 2023 to our attention. In the meantime, we have also considered this paper and used it to further support our three-tiered classification hierarchy that Stahl et al. 2023 similarly propose. Stahl et al. 2023 further support our claims that detailed agent attribution is not solved and availability of training data is a major hurdle in the field.

Line 64-66: I am not sure if “accuracy” is the right term. I suggest “resolution” instead.

Indeed, resolution is the right fit here. We have changed this.

The listed research gaps are reasonable. I was wondering if droughts are seen as short-duration events.

Based on the review by reviewer 2, we have formulated our definition of what we consider a forest disturbance in section 2. For drought in particular, we have changed the class name to “Drought induced leaf senescence”, because this is what we were able to observe in the satellite time-series.

This type of event is often limited to the end of the growing period which we would characterize as short-term. We have also replaced “short-term” in our research gaps with “short-lived” and explained better what we mean by that:

“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.”

Methodology

The first sentence should be moved to the introduction section.

We have removed this sentence from the methodology and added an equivalent sentence to the end of the introduction (see above).

Figure 1: The remote sensing backbone of the DISFOR dataset is a Sentinel-2 time series, mainly used for the timing of disturbances. It would be helpful if the authors could comment on the complementarity to other time series data, e.g. Sentinel-1 or Landsat – not necessarily in the methods section, but this is an important question not yet addressed.

We have added a section on this topic to 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. This interval can be narrowed if first-level data provide exact event dates, as is often the case for wildfire or windthrow. For Campaign 3, storm dates are already encoded in the column `cluster_id`, and future releases of the DISFOR dataset could record known event dates consistently for all samples. This makes the dataset usable with data which provides denser temporal resolution like Sentinel-1 or PlanetScope. However, applying DISFOR to datasets with coarser spatial resolution (like Landsat) must be carefully considered. In 30 m Landsat pixels, edge samples (Fig. 4a and b) may exhibit an even more mixed signal and low-severity disturbances such as thinning may not be detectable.”

Line 95-96: The authors state that timing is derived from Sentinel-2 time series. I was wondering if known disturbance dates were better suited, e.g. considering large storm events or fire where timing is exactly known.

Tracking the exact timing where known would indeed be advantageous. In the case of campaign three, this is done through the cluster_id field in the dataset (It encodes the country as ISO2 + the date of the windthrow event. E.g. PL20170817, for a windthrow in Poland on the 17th August 2017). We added a section in the discussion that the dataset could be consistently extended with this information where it is available:

“This interval can be narrowed if first-level data provide exact event dates, as is often the case for wildfire or windthrow. For Campaign 3, storm dates are already encoded in the column cluster_id, and future releases of the DISFOR dataset could record known event dates consistently for all samples.”

Line 99: What are low-severity disturbances? With respect to satellite data this term is scale-dependent, e.g. a removed tree (=total damage) at the edge of a pixel might have no impact visible to an interpreter of the satellite image. Partial crown defoliation in response of drought at larger scales will be visible in the satellite image but is not necessarily fatal. Just to think about it.

This is a good point. We have formalized our definition of which disturbances we consider during re-interpretation:

“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.”

Line 126: The paper of the European Forest Disturbance Atlas is now published (Viana-Soto & Senf 2025, <https://doi.org/10.5194/essd-17-2373-2025>).

Thank you, we have updated the reference.

Figure 2: This figure is quite crowded. Please increase font size

We have increased the font size and simplified the figure.

Line 154: Band 8 seems to be among the selection of bands since it is displayed in Figure2.

Indeed, we have added Band 8 to the list of used bands.

Line 158-159: I was wondering if the time series segmentation is entirely manual. BFAST, LandTrendR and others could be considered as well.

The time series segmentation was done entirely manual, we did not use any assisting algorithms to not get influenced by model classifications, which might not be accurate. We have added a short paragraph on this in the discussion for possible further exploration:

“We also explicitly did not add initial segmentation through algorithms like LandTrendr to avoid anchoring interpreters to possibly inaccurate detections. While we think this was the right decision, automatic assistance might be explored in the future to possibly speed up labelling.”

Table 1: This is a comprehensive logic. Do you think it is possible to differentiate between bark beetle-infested pixels that are salvage-logged and those in the vicinity of bark beetle-infested pixels that were sanitary-logged without necessarily being infested? This would be important to map an event and subsequent management properly.

Using satellite data alone it is not possible to tell if a tree was infected and in the green attack stage or healthy previous to sanitation logging. Because of this we have renamed class 220 to Salvage/sanitation.

Is it possible to differentiate between insect infestations and drought? If not, how is this reflected in the dataset?

Without ground truth reference data it is not possible to differentiate conclusively. In cases like this, a best guess for the class is given by the interpreter, the labelling confidence is set to medium and a comment is added to the sample, explaining the confusion. We have tried to clarify this in section 3.1: “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.”

Line 169: “Is is” should be “It is”

We fixed this.

Line 202 (and line 214, 223, 244 similarly): Please refer to the table/figure, e.g., “The table is structured as follows (Table 2).”

We have added references to the tables/figures in all places.

Dataset

Table 2: under “confidence” the sentence should read “Confidence of labelling: high where both timing and agent are confident, medium **where** only the timing is confident”.

Thank you, we fixed this.

Figure 3: I suggest plotting events as vertical bars since these are supposed to be discrete points in time rather than multiple observations (e.g., thinning). If multiple detections are required to confirm events this should be stated elsewhere.

We have changed this in the figure. Segments are now shown as colored boxes and events are shown as vertical dashed bars.

Line 231-236: Not clear to me what this paragraph tells the reader.

Since this dataset is intended to be used as training data for machine learning models, we provide pre-defined splits to derive model accuracies on data which was not seen by the model during the training. This is provided by us, so that researchers using the DISFOR data to train models have a

common baseline to validate against. We have added an introductory sentence to the paragraph to make this clear:

“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).”

Line 278: sentence should be changed to “Counting the number of unique labelled segments and events (Fig. 5)...”

Thank you, fixed.

A discussion section is missing. Authors may consider one to elaborate more on application scenarios and a vision of future work, collaboration opportunities etc.

We have added a discussion section, to also discuss the process of labelling and how DISFOR could be combined with other data.