

Reviewer 1

We sincerely thank Loïc Dutrieux for his careful and constructive review and for his encouraging assessment of the value of DISFOR to the community. His comments led to two major improvements of the manuscript. First, we made the labelling process considerably more transparent: Section 2.2 now includes two new figures illustrating the typical spatio-temporal dynamics of the disturbance classes and common labelling challenges, and the new Discussion section reflects on the labelling process. Second, we sharpened the conceptual distinction between segments and events and clarified how both are represented in the dataset.

Very interesting data compilation and re-interpretation effort that will certainly greatly benefit the community as data with this type of thematic details are either scarce or highly heterogeneous in quality. I have a few overall comments and some more minor remarks linked to sections or statements made in the manuscript.

Coverage:

It could have included at least two more datasets:

- <https://www.nature.com/articles/s41597-026-07084-8>
- <https://zenodo.org/records/10449087>

Thank you for bringing these datasets to our attention. We have included both in the introduction in line 50 and in the discussion as options for further extension. They will be considered for later extensions of the DISFOR dataset.

Details on labeling process:

The manuscript includes little details on the labeling process. I would have like to see illustrations and textual descriptions of spectro-spatio-temporal aspects of the various dynamics considered in the dataset. Some kind of feedback on the ideal visual interpretation parameters (color composites, vegetation indices) would be a nice addition too. Finally further discussion of some specific labeling challenges would make the data creation process fully transparent.

Thank you for this comment. We have extended Section 2.2 by two figures: Figure 3 showing typical spatio-temporal dynamics of different disturbance classes and Figure 4 showing examples of typical challenges present during labelling and how they were handled. Finally we have added a section on the labelling process in the discussion.

Temporal conceptualization of classification framework:

The manuscript discusses how various dynamics are either mapped to temporal segments of abrupt events, but it's not clear how this conceptualization is later handled in the dataset. I see only temporal segment labels in the dataset fields description.

We have clarified this in Table 3 and in the following text:

“Table 3:

start	Start date of the segment/event
end	End date of the segment/event (same date as start in the case of events)

Some labels like clear cut or thinning are defined as events. For events, the start and end are on the same day since the change is not gradual (compared to gradual events like regrowth or bark beetle decline)."

Minor remarks:

Line 30: The community often makes bark beetle its own class and consider it to be the only type of biotic disturbance. It surely is by far the largest biotic agent but it's not the only one.

Indeed, this was not meant as an exhaustive list, but as examples of major disturbance agents in Europe. The classification scheme in Table 1 should make it clear that we do not consider bark beetle as the only biotic disturbance.

Line 41: metadata format heterogeneity is true and datasets cannot be ingested in their raw form in machine learning pipeline, but given data scarcity, this is hardly a blocking issue

This is true, we have removed the remark on metadata heterogeneity.

Line 42: DEFID2 is in theory broader in scope than just insects; it includes all pests and diseases

Added to the sentence. *"DEFID2 (Forzieri et al., 2023) aggregates insect, pest and disease disturbance polygons and points from multiple sources."*

Line 61-66: I believe the process you described can be characterized as augmentation

Yes, it can indeed be described as augmentation. We have added a statement on how this augmentation happened: *"These first-level disturbance agent products are augmented by deriving disturbance timing from the full Sentinel-2 time series."*

Line 69: You use "training data"; good to make it clear whether it's indeed intended for training, or if it is a more generic "reference dataset"

DISFOR is intended as training data and not as a reference dataset. We see DISFOR as taking in "reference data" (FORWIND, EFFIS, VHR, etc.) which can not directly be used as training data, and augmenting this reference data so that it can be used for training. We have added a statement on this in the previous paragraph: *"DISFOR is explicitly intended to be used as training data for multi-temporal classification models. "*

Line 96: It would be good to have some idea of the amount of sample for which the visual re-interpretation did not confirm the agent or even the disturbance. In my experience compiled datasets are widely heterogeneous in quality

This was also our impression while interpreting. We added a paragraph in Section 3.3 on our observations:

"An analysis of the labelled data for Campaign 3 confirmed the heterogeneity of these types of datasets (e.g. FORWIND, DEFID2) which are compiled from different sources without re-

interpretation. Our re-interpreted data showed that for all unique storm events analysed in Campaign 3, at least 30% of samples were not labelled as windthrow. This was especially significant in the case of the emergency mapping of storm Vaia in Austria: about 80% of the labelled samples were actually connected to a different storm event which occurred one year after Vaia, or were already disturbed previous to the storm event. Other possible explanations for the lower amount of windthrow detected during re-interpretation is that low severity breakage could have been missed in the Sentinel-2 time series; thus it would not be labelled during re-interpretation (see 2.2 Labelling interface and response design). We also frequently observed that the sampled point within the disturbance polygon is just slightly outside of the visibly affected area.”

Lines 104-110: I'm not entirely sure it is relevant to name the projects at this stage

We have shortened this paragraph and pointed the reader to the financial support section. “The labelling in the different campaigns was carried out within different research projects (see section Financial support).”

Line 134: How were local news reports discovered?

We added a line on the discovery process: “[...] and local news reports which were discovered by web search for the approximate place, time and likely agent.”

Line 138: FORWIND has not been updated in a long time and marginally overlap with the sentinel era

This is true. Because of this we filtered the FORWIND database to only the events which overlap the Sentinel 2 time-series. We have added a short statement on this: “We used the FORWIND (Forzieri et al., 2020) database which we filtered to only include windthrow events which overlap the Sentinel-2 time-series. We also included emergency mapping activations in response to windstorms in the Copernicus Emergency Management Service (EMS) (EC JRC, 2019) as first-level disturbance agent data.”

Line 147: In my opinion you're missing an introductory paragraph here detailing the general idea and key assumption behind this app (visual interpretation potential, simultaneous spatial and temporal visualization, ergonomic, etc).

We have added an introductory paragraph:

“All samples were interpreted by the same group of interpreters using the same labelling interface called tsbrowser. This labelling interface is available open-source at <https://jr-digital.github.io/tsbrowser/>. The labelling interface was explicitly designed to show both the spatial and temporal context of the location to be interpreted. In contrast to existing labelling interfaces such as TimeSync (Cohen et al., 2010), optimized for showing annual Landsat data, our tsbrowser is specifically designed for Sentinel-2 data. Both tools allow the same multi-dimensional view on a sampling point, showing both spectral trajectory through time as well as spatial context. For our use-case we designed tsbrowser to show all available cloud-free Sentinel-2 acquisitions over the selected time-period (See Fig. 2).”

Figure 2: The right most explanation bubble explains that the sample footprint is not the same. Why?

This was due to limitations in the matplotlib based interpretation app. We did not integrate a full transform from the S2-UTM grid to the EPSG:3857 transform used by the ESRI Wayback Service. The yellow rectangle was used for orientation, and the sampled point was guaranteed to be within the rectangle. Since we did not need to interpret fractions of tree cover and only used the VHR to assess disturbance agents (lying trees, dead standing trees in the neighbourhood) this was sufficient for interpretation. In a different version of the interpretation app this was fixed (github.com/jonasViehweger/cdse-tsbrowser/).

Line 160: Dieback is a common term used instead of decline

We changed this to dieback.

Line 161: Stretching the concept of segments/events, not only disturbances would fit into events, but any transition between two segments, such as a stabilization at the end of a regrowth period.

This is indeed true, we have tried to sharpen our definition of events and segments:

“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.”

Line 164: I do not understand why you'd need two labels per segment

The interpretation app only allows to set labels for single dates. We used two repeating labels on different dates with the same class to denote segments. E.g. One label with class 110 on 2015-01-01 and one on 2016-01-01 with the same class 110 would denote a segment starting on 2015-01-01 and ending on 2016-01-01. We have clarified this in the text: “In practice this distinction means that a single label is set for events, while two labels with the same class are set for segments, one for the start of the segment and one for the end.”

Line 210: I think this was already mentioned a few paragraphs earlier

We have decided to leave this repetition, since it is essential information which is relevant for the response design as well as the interpretation of the resulting table. For the part in this section we have also added a short sentence on how this confidence should be used in practice: “In practice this means, that for tasks classifying the disturbance agent, only samples with high confidence should be used.”

Line 235: Fully agree, I would even further insist on that. Because of unknown or non-zero inclusion probabilities, the dataset cannot be used for assessment of mapping accuracy. It can be used for model performance assessment though, but that does not necessarily translate to mapping accuracy.

We have added a further sentence restricting what the dataset can be used for: "This dataset can not be used to derive map accuracies, since its samples were not collected using a probability sampling design."

Line 247: OK, but which baseline was used. Differences are marginal between baselines, but still good to be exhaustive

Unfortunately, this cannot be reconstructed for each S2 acquisition. Due to the small chip size of 32x32, we used streamlined data access using Sentinel Hub APIs reading data directly from CDSE. But exact baselines were not recorded for each acquisition.

Line 248: How do you justify applying the offset but not the scaling factor?

This is the default behaviour of Sentinel Hub APIs when returning DN. See <https://documentation.dataspace.copernicus.eu/APIs/SentinelHub/Data/S2L2A.html#harmonize-values>

We plan to release a newer version of the data, which provides larger image chips. For this release we would make sure to return the new v4 baseline DN without offset and scaling factor applied.

Line 259: Not sure it makes sense to use COG for a 32 pixels chip that is probably made of a single data block. That's fine, the greater includes the lesser.

Indeed, COG layout is probably unnecessary for 32x32. However we will keep it if we eventually distribute larger chips.

Lines 303-309: These are new elements; it's uncommon to mention things that have not been discussed previously in a conclusion

As also requested by Reviewer 3, we have moved this to a separate Discussion section.

Loïc Dutrieux