Remote Sensing of Environment 315 (2024) 114475

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

# Forest disturbance detection in Central Europe using transformers and Sentinel-2 time series

Christopher Schiller [a,*], Jonathan Költzow [a], Selina Schwarz [b], Felix Schiefer [c], Fabian Ewald Fassnacht [a]

[a] Institute of Geographical Sciences, Remote Sensing and Geoinformatics, Freie Universität Berlin, Malteserstraße 74-100, 12249 Berlin, Germany
[b] Institute of Meteorology and Climate Research - Atmospheric Environmental Research, (IMK-IFU), Karlsruhe Institute of Technology (KIT), 82467 Garmisch-Partenkirchen, Germany
[c] Institute of Geography and Geoecology, Karlsruhe Institute of Technology, Kaiserstraße 12, 76131 Karlsruhe, Germany

ARTICLE INFO

Edited by Marie Weiss

*Keywords:*
Monitoring
Sentinel-2
Forest disturbance
Deep learning
Transformer

ABSTRACT

Forests provide important ecosystem functions such as carbon sequestration and climate regulation, particularly in countries with high forest cover. Climate change-induced extreme weather events have a negative impact on many forest ecosystems. In Germany, for instance, the drought of the years 2018 until 2020 resulted in signs of damage in almost 80% of trees. This decline in forest vitality has additionally led to severe bark beetle infestations and widespread tree mortality, posing significant challenges to forest managers to obtain a complete picture of the state of their forests. Since a completely ground-based monitoring of forest condition is not feasible due to the forests' vast extent, remote sensing and particularly multispectral satellite image time series (SITS) analysis were suggested as efficient alternatives. Transformers, a state-of-the-art Deep Learning (DL) architecture, have shown promising results in the classification of multivariate SITS for other applications. Here, we use Transformers in combination with Sentinel-2 (S2) time series data to test if they can improve forest disturbance detection capabilities in comparison to conventional methods by automatically extracting relevant information from background variability throughout the whole time series. To match the large training data needs of Transformers, we use a two-step approach including pre-training and finetuning. During pre-training, we use outputs of earlier presented SITS approaches, while during finetuning, we use detailed reference data of known disturbances covering between 10 and 100% of a Sentinel-2 pixel as extracted from aerial images. We test three setups: *DL base* using ten S2 bands, *DL IND* using ten vegetation indices (VIs), and *DL +IND* utilising both as model input. F1-scores across all of our six study sites range between approx. 0.65 (DL +IND) and 0.72 (DL base) in a binary classification (undisturbed vs. disturbed) when considering both full and partial disturbances. DL base outperforms the other setups in forest disturbance detection, and detects disturbance extents as small as 40 $m^2$ within pixels of 100 $m^2$ size. Given the best performance of DL base, handcrafted vegetation indices (VIs) do not improve the model. Our model is competitive with existing approaches and slightly outperforms most earlier reported results, even though a direct comparison is challenging. Considering the option to further refine our trained model if additional reference data becomes available over time, we conclude that a combination of Transformers and Sentinel-2 time series can be developed into an effective tool for forest disturbance monitoring of Central European forests at fine spatial grain.

## 1. Introduction

Forests cover more than 30% of Germany's (Bösch et al., 2018; Holzwarth et al., 2023; Holzwarth et al., 2020) and Luxembourg's land surface (Schwarz et al., 2023), providing many ecosystem services such as carbon sequestration, climate regulation and space for recreation (Bösch et al., 2018; Senf and Seidl, 2020; Thom et al., 2017). Moreover, forests are an important economic factor, providing more than 1 million jobs and a turnover of billions of Euros in Germany (Holzwarth et al., 2020).

In the last decades, extreme climatic events such as heat waves, droughts and storms negatively impacted the vitality of German forests

* Corresponding author.
*E-mail address:* christopher.schiller@fu-berlin.de (C. Schiller).

https://doi.org/10.1016/j.rse.2024.114475
Received 23 July 2024; Received in revised form 23 September 2024; Accepted 14 October 2024
Available online 24 October 2024

(Holzwarth et al., 2020). As a result, about 80% of trees in Germany have been reported to show signs of damage in 2023 (Bundesministerium für Ernährung und Landwirtschaft (BMEL), 2023). Driven by climate change (Ionita and Nagavciuc, 2021), the frequency of extreme weather events is expected to increase even more in the future (García-Herrera et al., 2019; Grillakis, 2019; Ionita and Nagavciuc, 2021), presumably leading to a higher frequency of forest disturbances (Bréda et al., 2006; Gazol and Camarero, 2022; Holzwarth et al., 2020; Seidl et al., 2017; Senf and Seidl, 2021). Climate-induced tree stress and increased deadwood volumes as well as generally increased temperatures also benefit European bark beetles (*Ips typographus* L. and *Pityogenes chalcographus* L.) which have infested wide parts of forests in Germany and central Europe over the last years (Latifi et al., 2014; Netherer et al., 2019; Patacca et al., 2022; Senf et al., 2017).

A timely and area-wide monitoring of forest condition is urgently needed, for example to prevent bark beetles from spreading by conducting sanitation fellings (Dobor et al., 2020; Kautz et al., 2023), or to assess drought-induced forest dieback and plan preventive measures. For Germany's large forest areas of more than 11 million ha (Bundesministerium für Ernährung und Landwirtschaft (BMEL), 2023), an efficient and timely area-wide monitoring of forest condition on the ground is impossible (Holzwarth et al., 2020). Hence, remote sensing-based approaches have been discussed frequently over the last years. Remote sensing has shown potential for forest monitoring in numerous earlier studies (Abdullah et al., 2019; Grabska et al., 2020; Latifi et al., 2014; Senf and Seidl, 2020; Thonfeld et al., 2022). Many of these studies show that particularly the multispectral Sentinel-2 (S2) satellite system bears great potential due to its high spatial resolution of up to 10 m and its short revisit time of approx. 5 days in Europe. The multispectral sensor on board of S2 captures spectral data in the visible to shortwave infrared range, which enables it to capture information related to forest condition (Dutrieux et al., 2021a; Grabska et al., 2020; Mouret et al., 2024; Verbesselt et al., 2010; Zhu and Woodcock, 2014) and make it suitable as input for satellite image time series (SITS)-based forest disturbance monitoring approaches.

The analysis of SITS allows for the detection of changes in forest cover and state over time. Major challenges of SITS approaches include the creation of homogeneous time series and extracting disturbance-related changes from the diverse range of natural fluctuations in the spectral characteristics of temperate forests. The latter are caused, for instance, by the complex phenology of temperate forests with inter-annual variations (Puhm et al., 2020), varied species compositions and gradients in structural complexity (e.g. different stand ages, multi- or single-layered stands, canopy gaps, various topographic situations affecting sun-sensor geometries).

These challenges have been addressed in at least three existing SITS analysis approaches: near real-time (NRT) monitoring, temporal segmentation and time series classification. Near real-time monitoring algorithms utilise a training period of presumably undisturbed forest to model a baseline and detect deviations from forecasts of this baseline in new observations. Examples for NRT monitoring algorithms are BFAST Monitor (Verbesselt et al., 2012), CCDC/COLD (Zhu et al., 2020; Zhu and Woodcock, 2014) and FORDEAD (Dutrieux et al., 2021b). The second approach is referred to as temporal segmentation approach and includes, for instance, BFAST (Verbesselt et al., 2010). Temporal segmentation approaches decompose the SITS signal into a seasonal component, a trend component and residuals to subsequently identify breakpoints in the spectral time series. Depending on the strength of signal changes at the breakpoints and the direction of the change, a forest disturbance is then assumed. Features extracted from temporal segmentation approaches can also be used as input to time series classification methods. This approach is usually based on machine learning and can also be applied on the time series itself instead of the features engineered by temporal segmentation (Du et al., 2023; Perbet et al., 2024).

Deep Learning (DL), a sub-class of machine learning, has evolved as a promising tool in remote sensing. Numerous successful applications such as the mapping of standing deadwood (Schiefer et al., 2023), crop classification (Yuan and Lin, 2021), wildfire detection (Kong et al., 2018) and forest disturbance agent classification (Du et al., 2023) have been presented. *Transformers* (Vaswani et al., 2017), a Deep Learning architecture that revolutionized Natural Language Processing (NLP) (Ahmed et al., 2023), e.g. in Neural Machine Translation tasks (Vaswani et al., 2017), have been adapted to cope with a variety of time series data (Ahmed et al., 2023; Yuan and Lin, 2021) including SITS. The main innovation of Transformers is the *self-attention* algorithm, which enables the model to learn (the strength of) both long- and short-term dependencies between different observations.

Transformers exhibit at least three potential advantages over the aforementioned approaches for forest disturbance detection using SITS: (1) Transformers can consider single decisive observations at any position within the time series explicitly and capture their links to subsequent events. This could be advantageous with respect to events that trigger tree mortality after a certain time lag (Bigler et al., 2007) such as forest dieback occurring more than a year after a drought at certain sites (Haberstroh et al., 2022). Even an increased photosynthetic activity observed at the beginning of a drought (caused by a lot of photosynthetic active radiation hours) could be a subtle marker of a disturbance event following much later (Reinermann et al., 2019). Related to that, Transformers seem to offer an improved ability to detect small and gradual disturbance signals which are hard to disentangle from spectral variation caused by natural processes (Coops et al., 2020; Rodman et al., 2021; Ye et al., 2021). Recent studies suggest that Transformers may be able to capture such subtle disturbance signals in SITS (Perbet et al., 2024).

(2) Many established methods only operate on univariate time series. They either use satellite bands or vegetation indices (VIs) that are sensitive to disturbance (e.g. Dutrieux et al., 2021b; Verbesselt et al., 2010), or the algorithm is applied to multiple VIs/bands separately, followed by a cumulative sum of detected anomalies to decide about the final disturbance detection (e.g Puhm et al., 2020). Additionally, some of the established methods demand averaging or interpolation of the SITS as they require regular (e.g. 5 or 7 days) time series as input (Verbesselt et al., 2010). Transformers (or adaptations thereof), on the other hand, can take as input non-interpolated, multivariate time series (Yuan and Lin, 2021; Zhang et al., 2024). This is favorable, since interpolation leads to information loss (Zhang et al., 2024), and using univariate time series as input does not exploit all of the available information.

(3) Some of the established methods need parameter tuning to work for different regions and time frames (Pasquarella et al., 2022). In ideal case, a Deep Learning model trained on a very large amount of data has been exposed to many disturbed and undisturbed spectral time series, which are representative for the vast majority of the existing variability in spectral properties of disturbances of forests. If this is the case, it can be assumed that Deep Learning models can be readily applied across wider areas without further tuning.

Given these potential advantages, Transformers have been successfully used in classification of SITS for crop (Yuan and Lin, 2021) and land cover classification (Zhang et al., 2024) as well as tree species mapping (Mu et al., 2024). In the context of Deep Learning-based forest disturbance detection, Schiefer et al. (2023) use Long Short Term Memory Networks (LSTMs) to estimate the fraction of standing deadwood given S2 pixel time series in Germany. Wittich et al. (2022) use single Sentinel-2 scenes to estimate the time of upcoming disturbances using a Convolutional Neural Network at a study site in Germany. However, only few studies using Transformers for SITS have been published, yet. Du et al. (2023) and Mullissa et al. (2023) detect stand-replacing disturbances using Landsat and Sentinel-1/2, respectively. Perbet et al. (2024) successfully detect stand-replacing as well as partial disturbances using annual composites of SITS of the Landsat satellites. To the best of our knowledge, however, there is no study so far that uses Transformers on SITS of raw Sentinel-2 data without temporal compositing to detect

forest disturbance in temperate forests.

To fill this gap, we test if Transformers are capable of extracting relevant information from complex and dense multi-year time series of S2 observations to identify forest disturbances across a wide and varied geographic area covering Germany and Luxembourg. We implement the study as a binary classification task (undisturbed vs. disturbed forest). We train on stand-replacing as well as partial disturbances of multiple disturbance agents (incl. logging) in order to enable the detection of disturbances at sub-pixel size as well. We hypothesize that this approach might empower the model to detect early stages of gradual disturbances such as bark beetle infestations. Since VIs have been widely and successfully utilised in spaceborne forest disturbance detection tasks (Abdullah et al., 2019; Bandyopadhyay et al., 2017; Cohen et al., 2010; Mandl and Lang, 2023; Mouret et al., 2024), we examine three setups: 1) *DL base* using only the ten 10 to 20 m S2 bands, 2) *DL +IND* using both the ten S2 bands and ten commonly used VIs, 3) *DL IND* using only the ten VIs. In short, the goal of this study is providing an advanced forest disturbance monitoring tool capable of identifying spatially small forest disturbances applicable to all forest pixels in Germany and Luxembourg at any time of the year using state-of-the-art transformer models to support forest management. Such a tool has been called for by numerous stakeholders in the forest sector in the last years (Holzwarth et al., 2023; Holzwarth et al., 2020).

This results in three research questions:

- To what degree can Transformers accurately detect forest disturbance on formerly unseen S2 time series?
- What is the smallest disturbance extent that can be detected by the proposed method?
- Do Transformers need vegetation/disturbance indices for accurate predictions?

## 2. Methods

We use Sentinel-2 multispectral time series and forest disturbance information derived from published (mostly satellite-based) datasets (pre-training step) and high-resolution aerial imagery (finetuning step) to train a transformer model, a state-of-the-art architecture of Deep Learning (DL) models for time series classification (Ahmed et al., 2023; Du et al., 2023; Yuan and Lin, 2021). We implement the model for a binary classification task (undisturbed vs. disturbed forest). In the following, we describe in detail the steps to process input data (Sections 2.1 and 2.2), train the models (Sections 2.3 and 2.4) and compare the results (Section 2.5), as summarized in Fig. 1.

### 2.1. Sentinel-2 data preprocessing

Sentinel-2 (S2) data for Germany and Luxembourg were processed using the Framework for Operational Radiometric Correction for Environmental Monitoring (FORCE) (Frantz, 2019), including co-registration, atmospheric and topographic correction and cloud masking, among others. In the resulting datacube, S2 tiles with a cloud cover of more than 70% as stated in the metadata of the respective S2 scene or 90% as determined by the FORCE cloud masking algorithm were discarded. The level 2 datacube of bottom of atmosphere (BOA) reflectances for Germany was readily available from the EO-Lab platform (EO-Lab, 2023), while the FORCE datacube for Luxembourg was processed on the high-performance cluster of Free University Berlin (CURTA) (Bennett et al., 2020) using all available S2 data and the same FORCE parameters as the aforementioned datacube. We used the following ten spectral bands of S2, which were upsampled to 10 m spatial resolution using the ImproPhe algorithm implemented in FORCE: Red (RED), Green (GRN), Blue (BLU), Red Edge 1 (RE1), Red Edge 2 (RE2), Red Edge 3 (RE3), Near Infrared (NIR), Broad Near Infrared (BNR), Shortwave Infrared 1 and 2 (SWIR 1, SWIR 2). Additionally, we computed the ten following vegetation and disturbance indices (VI's) from the aforementioned ten spectral bands, which are commonly used for disturbance analyses: Continuum Removal Shortwave Infrared (CRSWIR) (Dutrieux et al., 2021a; Mouret et al., 2024), Normalized Burn Ratio (NBR) (García and Caselles, 1991), Tasselled-Cap Wetness (TCW) (Crist and Cicone, 1984), Tasselled-Cap Disturbance (TCD) (Healey et al., 2005), Normalized Difference Vegetation Index (NDVI) (Tucker, 1979), Normalized Difference Water Index (NDWI) (Gao, 1996), Normalized Difference Moisture Index (NDMI) (Gao, 1996), Leaf Area Index (LAI) (Boegh et al., 2002), Moisture Stress Index (MSI) (Rock, 1985) and Normalized Difference Red Edge (NDRE) (Gitelson and

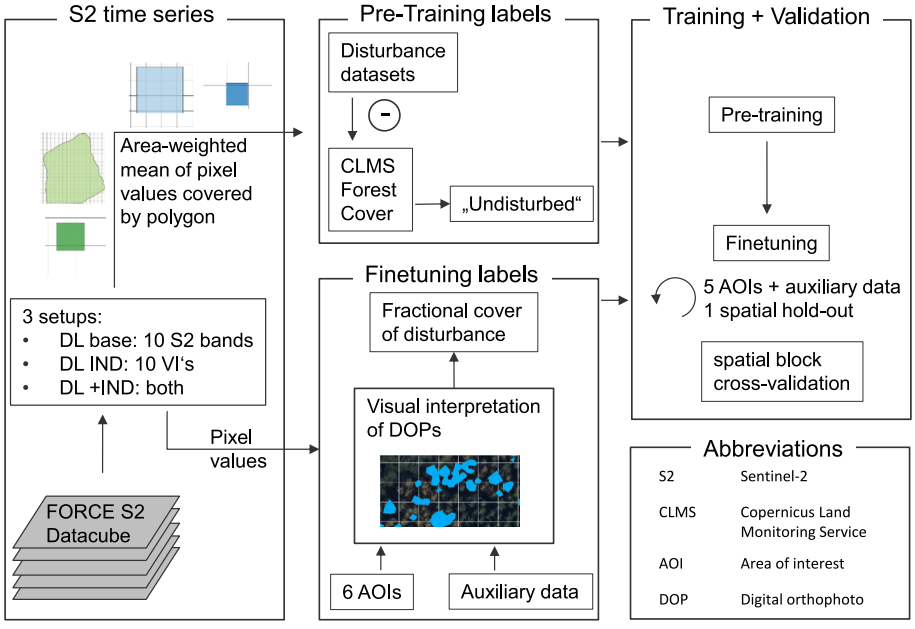


**Fig. 1.** Workflow of the study. After Sentinel-2 time series preprocessing (left), ten vegetation and disturbance indices were computed (left). For the three setups, pre-training was conducted after sampling labels from published remote sensing-based disturbance datasets and Copernicus Land Monitoring Service (CLMS; top center and top right). The pre-trained Deep Learning models were then finetuned using labels from six visually interpreted study sites and auxiliary data (bottom center and right). A spatial block cross-validation on AOI level was conducted as validation (bottom right).

Merzlyak, 1998) (see Supplementary Table 1 for details). For the pre-training dataset, we computed the time series from the described S2 datacube as the area-weighted mean of polygons described in Section 2.2.1 with end and start dates being selected according to the detection date of each disturbance (for details see below). Regarding the fine-tuning dataset, we computed the disturbed area (dieback + logging) for each S2 pixel, using each pixel's fractional cover of disturbance as label (see Section 2.2.2). We extracted the S2 time series for each pixel individually and the end date of the time series was selected to match the acquisition date of the DOPs from which the detailed reference data were collected.

### 2.2. Labels and validation datasets

#### 2.2.1. Pre-training

For pre-training the transformer models, we acquired three spatially explicit datasets containing forest disturbances and their approximate time of occurrence in Germany. In the following, we refer to these datasets as Senf (Senf and Seidl, 2020), Thonfeld (Thonfeld et al., 2022) and Forwind (Forzieri et al., 2020) (see Table 1 for details). Forwind contains windthrow disturbances following major storms (Forzieri et al., 2020) with exact dates, while Thonfeld mainly contains stand-replacing disturbances in and after the drought years of 2018–2020 in monthly temporal resolution (Thonfeld et al., 2022). Senf contains forest disturbance areas originating from a variety of forest disturbance agents (Senf and Seidl, 2020), among them harvest, bark beetle, drought, windthrow and fire, with a yearly temporal resolution. Additionally, it provides the disturbance severity of detected disturbance events as a percentage of lost canopy cover within a 30 m pixel. In the disturbance datasets, we filtered for disturbance events that occurred after July 2019 to ensure that S2 time series with a minimum duration of 4 years could be sampled (S2 data is available from July 2015). We avoided spatial autocorrelation by removing Forwind disturbances from Thonfeld and Senf, and Thonfeld disturbances from Senf, applying a buffer of 30 m to each disturbance event. We gave priority to Forwind and Thonfeld, since the former is mostly based on aerial imagery, and the latter is derived from S2 as opposed to Senf's Landsat approach, meaning that Forwind and Thonfeld are spatially higher resolved and able to reflect forest disturbances at smaller extents. Additionally, both commission and omission errors are smaller in Thonfeld according to their own validation (Senf and Seidl, 2020; Thonfeld et al., 2022). The samples derived from these three datasets were assigned to the disturbed class.

We then removed all the forest disturbance areas of Forwind, Senf and Thonfeld including a buffer of 30 m from a forest mask of 2018 acquired from European Union's Copernicus Land Monitoring Service (CLMS) and randomly sampled S2 forest pixels from the remaining forest patches. We made sure to have a buffer of at least 30 m in between two sampled pixels to avoid spatial autocorrelation in training caused, for instance, by an individual tree spanning over more than one sample.

**Table 1**
Table characterising the four datasets Senf (Senf and Seidl, 2020), Thonfeld (Thonfeld et al., 2022), Forwind (Forzieri et al., 2020) and Undisturbed in the pretraining step of the study. $N_{training}$: amount of training samples drawn for pre-training.

| Dataset | Disturbance agents | Temporal resolution of disturbance detection | Sensor | $N_{training}$ |
|---|---|---|---|---|
| Senf | harvest, biotic/abiotic and others | year | Landsat 4/5/7/8 | 41,832 |
| Thonfeld | mainly clear-cuts after drought/bark beetle damage | month | Sentinel-2, Landsat 8 | 87,665 |
| Forwind | windthrow | day | aerial imagery | 377 |
| Undisturbed | – | – | Sentinel-2 | 927,005 |

These samples were assigned to the 'undisturbed' class. Since the Senf method relies on medoid composites of the growing season ending in September 30th of each year, we determined the end date of the corresponding time series labelled with Senf data as Octobre 1st to make sure it contains the disturbance. For Thonfeld, we used the 1st day of the month after the disturbance detection. Next, we randomly assigned a date between the aforementioned end date and the date up to 6 months later to ensure that 1) the full time series actually contains the reference disturbance time (possible uncertainties in time of detection and due to temporal resolution of disturbance datasets) and 2) to support generalization of the model by preventing it from learning artefacts. Potential artefacts occur, for instance, if time series labelled as disturbed always end in September. In this case, the model might learn that an end date in September qualifies for a disturbed prediction even in the absence of disturbance (overfitting to artefacts). Moreover, the model is meant to enable forest disturbance detection at any date in a year, meaning that it needs to be trained on time series ending on an arbitrary date. The start date of each time series was fixed four years before the chosen end date. Hence, the resulting dataset contained time series of four years, the reference disturbance detection occurring approximately, but not exactly at the end of the time series and containing end dates between July 1st, 2019 and May 1st, 2021 (last date of detected disturbances in Thonfeld). Accordingly, the Undisturbed dataset was assigned an end date of the time series between July 1st, 2019, and May 1st, 2021. Hence, we assume a forest pixel of the undisturbed class to be undisturbed (or at least resistant) during this period, meaning that omissions in the disturbance datasets have led to time series labelled as undisturbed despite its actual disturbance in our pre-training data. It also implies that we expect that a pixel determined as forest by CLMS in 2018 has been forest back in July 2015 already. According to Senf and Seidl (2020)'s validation, their dataset exhibits rather small commission errors (14.6 ± 1.8%). Still, we tried to prevent the models from adopting these mistakes of the pre-training datasets by removing disturbances below 50% severity. This also guarantees a clear distinction between undisturbed and disturbed class during pre-training. The data from these four datasets (Forwind, Thonfeld, Senf and Undisturbed) cover the whole of Germany (Supplementary Fig. 1) and were used for pre-training the DL models.

#### 2.2.2. Finetuning

For the finetuning step, we used data from one forested area for Luxembourg (LUX) (from Schwarz et al., 2023) and for five federal states of Germany with significant forest cover and openly available digital orthophotos (DOPs) to visually delineate forest disturbances ourselves. The five federal states included Brandenburg (BB), Saxony (SAX), Thuringia (THU), Rhineland-Palatinate (RLP) and Northrhine-Westphalia (NRW). We made sure to cover areas representative for a variety of environmental and forest characteristics such as open (BB) and dense forest stands (RLP, LUX, NRW), steep slopes (RLP, NRW), heavily (NRW), medium (THU, RLP) and moderately disturbed (LUX) forest. These six areas of interest (AOIs) were complemented by forest disturbance data from (Schiefer et al., 2023) (in the following: Schiefer) originating from (mainly southwestern) Germany, from the FNEWS project in the states of Baden-Württemberg, Lower Saxony and Saxony (Langner et al., 2023), and (Schwarz et al., 2023) (in the following: Schwarz) in Luxembourg (Fig. 2). In the six AOIs where we delineated the forest disturbances ourselves, we considered all Sentinel-2 pixels containing forest (after removal of natural forest openings, arable land, urban areas, gravel pits, etc. by visual interpretation) excluding a 15 m buffer off the forest edges to prevent edge effects for our analyses. We manually delineated forest disturbance areas incl. logging and standing deadwood using the very high resolution aerial imagery from image flights from Luxembourg and Germany (between 10 and 40 cm spatial resolution, see Table 2) and, in case of Schiefer, uncrewed aerial vehicles (UAV). We compared the respective DOP with the preceding DOPs of each AOI to enable a distinction between older logging areas and natural

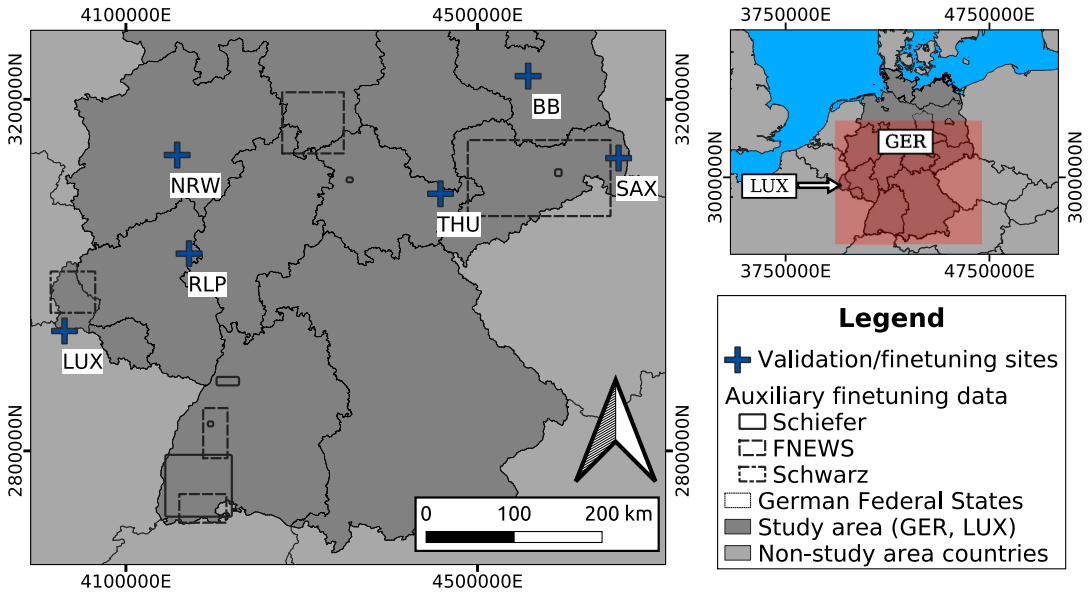


**Fig. 2.** Study site for finetuning and validation in Germany and Luxembourg. Blue markers represent the six AOIs of the study, while black rectangles are the bounding boxes of the regions in which some auxiliary data from Schiefer et al. (2023), FNEWS (Langner et al., 2023) and Schwarz et al. (2023) were available. Coordinate reference system: EPSG:3035. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

**Table 2**
Information on finetuning datasets from Brandenburg (BB), Saxony (SAX), Thuringia (THU), Northrhine-Westphalia (NRW), Rhineland-Palatinate (RLP), Luxembourg (LUX; Schwarz et al., 2023), Schiefer (Schiefer et al., 2023), Schwarz ((Schwarz et al., 2023)) and FNEWS (Langner et al., 2023). DOP: digital orthophoto, number indicating spatial resolution; RGB: red-green-blue; RGBI: reg-green-blue-infrared; extent: refers to spatial extent of studied scene (only provided for the six AOIs); UAV: uncrewed aerial vehicle; N: number of samples used in validation (Schiefer, Schwarz and FNEWS: only used in training phase, number approximate due to multiple training runs with random sampling). DOPs are subject to courtesy of GeoBasis (licence: dl-de/by-2-0), open.NRW (dl-zero-de/2.0), Landesamt für Vermessung und Geobasisinformation Rheinland-Pfalz (dl-de/by-2-0), Landesamt für Geobasisinformation Sachsen (GeoSN) (dl-de/by-2-0), Landesamt für Bodenmanagement und Geoinformation Thüringen (dl-de/by-2-0), Administration du Cadastre et de la Topographie (ACT), Grand-Duchy of Luxembourg (CC0).

| Dataset | Product | Acquisition date | Location | Extent | *N* | $N_{broeadleaved}$; $N_{coniferous}$ | Source |
|---|---|---|---|---|---|---|---|
| BB | DOP20 RGBI | 2022–06-18 (2019-04-19, 2016-04-21) | Brandenburg (GER) | 2*2 km | 29,879 | 3779; 26,100 | own work |
| SAX | DOP20 RGBI | 2022–06-03 (2020-07-30, 2017-08-05) | Saxony (GER) | 2*2 km | 32,940 | 3865; 29,075 | own work |
| THU | DOP20 RGBI | 2022–06-19/15 (2021-04-28, 2020-03-15, 2019-04-17, 2018-04-07, 2016-05-06) | Thuringia (GER) | 2*2 km | 26,880 | 5599; 21,281 | own work |
| NRW | DOP10 RGBI | 2021-06-14 (2018-04-07, 2015-06-05) | Northrhine-Westphalia (GER) | 2*2 km | 19,995 | 13,989; 6006 | own work |
| RLP | DOP40 RGB | 2021-09-06/07 (2019-07-24, 2018-07-02, 2015-06-05) | Rhineland-Palatinate (GER) | 4*4 km | 49,503 | 45,752; 3751 | own work |
| LUX | DOP10 RGBI | 2019-08-22 (2018-07-02, 2017-06-14, summer 2016) | Luxembourg | 1.75*1.3 km | 19,683 | 17,731; 1952 | Schwarz et al., 2023; own work |
| Schiefer | DOP RGB (UAV) | 2019/2021–09/10 | Black Forest (GER), Dresden Heath (GER), Karlsruhe-Bretten (GER), Hainich National Park (GER) | – | ~ 1819 | ~ 560; ~ 1259 | Schiefer et al., 2023 |
| Schwarz | DOP10 RGBI | all DOPs 2016–2020 | Luxembourg | – | ~ 5872 | ~ 4884; ~ 988 | Schwarz et al., 2023 |
| FNEWS | DOP | different dates | Baden-Württemberg, Lower Saxony, Saxony | – | ~ 502,789 | ~ 188,104; ~ 314,685 | Langner et al., 2023 |

forest openings, since the latter are neither undisturbed forest nor forest disturbance and, thus, had to be excluded from analyses. All apparent loss of canopy cover between the DOPs without any sign of dieback (e.g. standing deadwood) was labelled as logging. When in doubt if the forest disturbance happened within a maximum of 3.5 years preceding the DOP of interest, we excluded these pixels from analysis. Forest disturbance labels were then acquired by computing the area percentage of the delineated forest disturbances for each S2 pixel covered by forest. Note that although we distinguished between logging and dieback in visual interpretation of the DOPs, we combined these two disturbance categories into a single 'disturbed' class during classification, since their disturbance cover distribution and number of samples did not allow for a multi-class approach. We, however, made use of the information about the two disturbance categories during the validation step.

### 2.3. *Transformer model and setups*

Deep Learning (DL) models apply transformations to the input data (here: S2 time series) in order to produce an output (here: binary class predictions) that is as close to the labels (here: undisturbed vs. disturbed pixels) as possible. These transformations are stacked upon each other as so-called *layers*. The difference between predictions and labels

(reference data), which is called *loss*, is described by a *loss function* (here: binary cross-entropy loss). The *Stochastic Gradient Descent* (SGD) algorithm is then applied to compute the gradient of the loss function. Afterwards, the *weights*, which are the parameters of the layers defining the data transformations, are adjusted such that the loss approaches the minimum of the loss function. The *learning rate* determines how much the weights are changed in one step, i.e. after one batch of data has been input to the model.

*Transformers*, a state-of-the-art DL model architecture originating from Natural Language Processing (NLP; Vaswani et al., 2017), are becoming increasingly popular in predictive time series analysis tasks (Ahmed et al., 2023) and have been adapted to specifically suit the challenges accompanying satellite image time series (SITS; e.g. irregular time series) as well (Yuan and Lin, 2021). The main innovation of Transformers are the *self-attention* layers (also called *self-attention heads*), within which *attention scores* are computed. These attention scores inform the model which and to what extent observations of the time series (in NLP: words of a sequence) are linked (Vaswani et al., 2017). Self-attention layers can be run in parallel to learn different representations (Vaswani et al., 2017). In the encoder, self-attention heads are followed by a dense layer that can be found in any DL model containing weights to transform the input of the preceding self-attention layers. The self-attention layers and succeeding dense layers together form the *encoder*. Encoder blocks can be stacked to allow the model to learn from different levels of abstraction of the time series (e.g. long-term periodicity vs. short-term events) (Jawahar et al., 2019; Yuan and Lin, 2021). The succeeding encoder block gets as input the output of the preceding encoder block, i.e. the weights of the dense layer following each self-attention block.

The self-attention algorithm does not know the order of the observations, however. Therefore, a *Positional Encoding* (PE) is added to the input data, which informs the model about the notion of order (here: time) (Ahmed et al., 2023; Vaswani et al., 2017).

Here, we deploy SITS-BERT (Yuan and Lin, 2021), which is a Transformer architecture specifically designed to approach the challenges usually faced in satellite image time series (SITS), e.g. irregular time series due to cloud-masked observations. Other than in the original Transformer (Vaswani et al., 2017), the PEs are not summed to the input data, but concatenated (i.e., processed separately). This prevents the model from confusing the representations of time with the actual observations (Yuan and Lin, 2021). In fact, in SITS other than in NLP, there is a very important seasonal pattern in the dataset, so the dates should be modelled distinctly (Yuan and Lin, 2021). This also implies that SITS-BERT does not need equidistant time steps in the sequences (i.e., no gap filling, interpolation or averaging of multiple observations) (Yuan and Lin, 2021). This was considered important since interpolation and averaging always induces additional noise into the sequence, deviating from real observations (Zhang et al., 2024).

Furthermore, *embedding layers* precede the encoder blocks. In NLP, they are used to encode words into integer values and reduce the dimensionality of the vocabulary, since DL models can only process numbers, not words. In time series analysis, they serve a different purpose: they provide higher-dimensional representations of the input features (S2 bands and VIs). In simple terms, this can be interpreted as computing its own VIs from combinations of the input features in a dynamical (and less rigid) way. For instance, the embedding layer might combine the RED and NIR bands to model the difference between those bands.

On top of the encoder blocks, a classifier is stacked, which consists of a dense layer and provides a single output for each input time series. This represents the confidence score for the disturbed class (range: [0,1]), which can then be translated into binary classes by a threshold (here: 0.5). In DL training, the data is usually split into training, validation and test datasets. The model weights are altered in *steps*, which consist of *batches* of training data, while the validation data is used to monitor the training progress. Note that in Deep Learning, the so-called validation dataset is not used for validation of the model, but for validation of the progress of the model training. Validation of the model is done using the test dataset. The loss and accuracy were computed on the validation data after the complete set of training batches was fed to the model (which is called an *epoch*). Afterwards, the training samples were shuffled and input to the model in the next epoch. This was repeated until a stop criterion was reached (see below).

For the DL base setup, we used the ten S2 bands with 10 and 20 m spatial resolution, namely BLU, GRN, RED, RE1, RE2, RE3, BNR, NIR, SWIR1 and SWIR2. Since VIs are widely used in remote sensing to detect changes in vegetation (e.g. (Bandyopadhyay et al., 2017; Cohen et al., 2010; Grabska et al., 2020; Kennedy et al., 2010; Mandl and Lang, 2023)), we test two more setups, taking as input only the ten VIs (DL IND) or both the ten VIs and the ten S2 bands (DL +IND).

### 2.4. Deep Learning training procedure

The training of DL models can be conducted in multiple steps. Here, we conduct a *pre-training* on a large amount of medium quality data (pre-training datasets), followed by a *finetuning* on a medium amount of high quality data (finetuning datasets). The pre-training dataset is considered less accurate, as the reflectance values acquired from it are area-weighted means of the pixels using coverage fractions of disturbance polygons (polygon-based approach), while we use pixel-based disturbance labels and no averaging of reflectances in finetuning. Additionally, most of the pre-training datasets are satellite-derived itself and thus contain errors. The idea of this two-step approach is that the model gets a general notion of the task to solve in pre-training, and improves on that in finetuning. Thus, the general dependence on Big Data is met and at the same time, data with higher quality is used for training the best model for the task. We added the day of the year (DOY) of each observation as additional input for the Positional Encoding. The DOY increased with every year, i.e. in the second year of each time series, DOYs were computed by adding 365, and so on. This was done to make sure that the model does not consider two observations with the same DOY from different years as the same time step, since the order and timeliness of the observations is relevant for classification.

We used a single classification layer with one output unit for this binary classification task, and Binary Cross-Entropy loss as loss function. The learning rate was set to 0.0001 in pre-training and 0.00001 in finetuning. The learning rate was one magnitude smaller in finetuning in order not to erase the learned representations from pre-training. Training was conducted with a batch size of 128. The embedding size was 128, where 64 units were reserved for the SITS embeddings, and another 64 units modelled the observation dates. We utilised three encoder blocks with eight attention heads each. The complete model contained 594,753 trainable weights. For all training runs, we utilised early stopping with a patience of 10, meaning that training was conducted until the validation loss did not improve for ten consecutive epochs. The best-performing model was determined by the best accuracy on the validation dataset.

The forest types in both pre-training and finetuning dataset were strongly skewed towards coniferous trees in the disturbed class and broadleaved trees in the undisturbed class. In pre-training, we undersampled the majority forest type in both the undisturbed and disturbed class to prevent the model from adopting this bias. Since the resulting dataset contained many more samples in the undisturbed than in the disturbed class, we used class weights in order to prevent the model from majority votes on the undisturbed class. The class weight was determined by the ratio of amount of undisturbed to amount of disturbed samples divided by 3. In finetuning, the data distribution allowed for applying a different sampling strategy in the training phase. In the disturbed class, we oversampled the minority forest type by a factor of 2, followed by undersampling the samples of the majority forest type to yield the same amount of samples as the doubled minority class. In the undisturbed class, we undersampled the samples of the majority forest

type to the equal amount of data of the minority forest type. The resulting dataset contained the same amount of deciduous and coniferous forests' pixels in both classes. Resampling to overcome challenges associated with class imbalance has shown to be effective before (Wittich et al., 2022). Applying class weights as in the pre-training step was not necessary in finetuning, since the resulting dataset was less imbalanced. This sampling strategy for pre-training and finetuning was confirmed by the best performance in preliminary model runs.

In finetuning, we discarded all samples with disturbance labels greater than 0% and smaller than 10% in order to make a distinction between undisturbed and disturbed time series during model training and hence enable it to differentiate between the two classes. The threshold of 10% in finetuning differs from the threshold of 50% in pre-training, since the labels of the latter contain commission errors, which are more likely in case of low disturbance severities. In finetuning, however, the labels are sufficiently accurate to allow for a lower threshold. A threshold was necessary, since we cannot expect the model to be able to differentiate between, for instance, 0% (healthy) and 1% (disturbed class) disturbance fraction. Taking all disturbance labels as training data would have led to an insufficient separability of the two classes.

In the validation phase (i.e. the testing in the finetuning step), we did not discard very low disturbance pixels (smaller than 10%) from analysis to avoid an overestimation of model performance (see also Section 4.1).

For pre-training, the dataset comprised of Forwind, Thonfeld, Senf and Undisturbed dataset (see Section 2.2) was split into training and validation datasets randomly by sampling 20% of the samples as validation and 80% as training dataset. The validation dataset was used to monitor the increase of model accuracy during each training epoch. The test data was used to verify the model's ability to reproduce the labels from the pre-training datasets after completing the training procedure (for results see Supplementary Table 2).

In the finetuning step, we conducted a spatial block cross-validation on AOI level with three repetitions. Hence, we used each of the six AOIs as spatial hold-out once, performing the training on the time series of the five remaining AOIs and the auxiliary finetuning data (Schiefer, FNEWS and Schwarz). We used three different random samples of training and validation datasets (with a ratio of 4:1) in the cross-validation steps (hence the three repetitions), while the test dataset remained stable (all forest pixels of the corresponding AOI). Using 3 repetitions ensured that the model predictions did not depend on the split into training and validation dataset. Note that the term *validation dataset* refers to the Deep Learning training process here (see Section 2.3), and not to the spatial hold-out for testing. The described procedure guarantees that each model is neither exposed to the test dataset's time series nor to their site conditions during training. Thus, the validation procedure contains independent test data (spatial hold-out AOI in each training run) while at the same time, the number of data and exposure to different site conditions is maximized during training. We used the same pre-trained model for all of the finetuning setups.

#### 2.4.1. Time series augmentation during training

To reduce overfitting and improve generalization of the model (Iwana and Uchida, 2020), we applied the following *time series augmentation* (TSA) procedures during training:

window slicing, window warping, and adding random noise (also called *jitter*) in satellite signal and DOY of observations to the training process.

Window warping was implemented as in Iwana and Uchida (2020) by randomly picking a window within the time series and either stretching it by 2 or contracting it by 0.5. This method has been shown to improve accuracy and generalization (Iwana and Uchida, 2020). Jitter was added to the time series by adding or subtracting up to 5 days randomly to the DOY of the observation, and adding random values of mean 0 and standard deviation 0.05 to the satellite values. The former allowed for more generalization concerning phenological events (inter-annual variations (Puhm et al., 2020)). The satellite signal jitter was meant to account for the noise inherent to satellite data due to topographic and atmospheric effects, undetected clouds and cloud/tree shadows, etc. (Yuan and Lin, 2021).

Window slicing consisted of dropping the beginning of the sequence to yield a time series between two and four years (pre-training), or 3.5 and 4 years (finetuning) randomly. This guaranteed a minimum time series length of 2 (pretraining) and 3.5 (finetuning) years while making sure that the disturbance event is still part of the time series (temporal resolution of as low as approx. 3 years in DOP acquisition dates used for labelling).

### 2.5. Validation

We performed a number of analyses to validate and compare the models with respect to performance and plausibility of the detected disturbance patterns, accuracy, minimum detectable damage extent and generalisability (i.e., predictive performance across forest types broadleaved and coniferous).

Firstly, we mapped the confidence of the models with respect to the disturbed class (i.e., before binarization) to investigate if detected patterns match with the damage observations based on aerial imagery in the six AOIs. We provide error matrices and performance metrics for the three setups. Afterwards, we analysed the capability of the models to detect small disturbance areas by stratifying the fraction of disturbance (dieback and logging combined) per pixel into 10 $m^2$ categories, e.g. 10 to 20 $m^2$, 20 to 30 $m^2$, and so on. Note that in case of our 100 $m^2$ S2 pixels, $m^2$ equals percentage. Only the 0 $m^2$ damage category was considered undisturbed forest. We show the models' Producer's Accuracies for each of these strata. Afterwards, we further stratified these results by forest type, i.e. broadleaved and coniferous forest. The forest type information was extracted from European Union's Copernicus Land Monitoring Service (CLMS).

Preprocessing of all datasets and retrieval of S2 time series from the FORCE datacube was conducted in R (R Core Team, 2023) (v4.3.1) using terra (Hijmans, 2023) (v1.7–39), sf (Pebesma, 2018) (v1.0–14) and exactextractr (Baston, 2022) (v.9.1) packages, while model training and validation was conducted in Python 3 (Van Rossum and Drake, 2009) (v3.8) using PyTorch (Paszke et al., 2019) (v1.13.1), and captum (Kokhlikyan et al., 2020) (v.6.0).

## 3. Results

### 3.1. General performance metrics

False positives across all AOIs range between 0.01 ± 0.001 (DL +IND) and 0.027 ± 0.004 (DL base) in the relative error matrices (Tables 3, 4 and 5), while the false negatives yield between 0.055 ± 0.002 (DL base) and 0.078 ± 0.001 (DL +IND). Commission errors (omission errors) are highest (lowest) in DL base with 0.032 ± 0.005 (0.34 ± 0.014; Table 6). The overall accuracy reaches above 0.91 for all setups. The highest (lowest) f1-score is achieved by DL base (DL +IND) with 0.722 ± 0.001 (0.654 ± 0.002). When considering all disturbances (i.e., all disturbances >0% of a pixel), omission errors yield up to 0.485 ± 0.004 (DL +IND).

**Table 3**
Relative error matrix of all predictions and reference data across all of the 6 AOIs for DL base setup expressed as mean and standard deviation across three repetitions of spatial block cross-validation.

| | | Reference | | Total |
|---|---|---|---|---|
| | | Undisturbed | Disturbed | |
| Predictions | Undisturbed | 0.812 ± 0.004 | 0.055 ± 0.002 | 0.867 ± 0.007 |
| | Disturbed | 0.027 ± 0.004 | 0.106 ± 0.002 | 0.133 ± 0.007 |
| | Total | 0.839 ± 0 | 0.161 ± 0 | 1 |

**Table 4**
Relative error matrix of all predictions and reference data across all of the 6 AOIs for DL IND setup expressed as mean and standard deviation across three repetitions of spatial block cross-validation.

| | | Reference | | |
|---|---|---|---|---|
| | | Undisturbed | Disturbed | Total |
| Predictions | Undisturbed | 0.82 ± 0.002 | 0.064 ± 0.003 | 0.884 ± 0.005 |
| | Disturbed | 0.019 ± 0.002 | 0.097 ± 0.003 | 0.116 ± 0.005 |
| | Total | 0.839 ± 0 | 0.161 ± 0 | 1 |

**Table 5**
Relative error matrix of all predictions and reference data across all of the 6 AOIs for DL +IND setup expressed as mean and standard deviation across three repetitions of spatial block cross-validation.

| | | Reference | | |
|---|---|---|---|---|
| | | Undisturbed | Disturbed | Total |
| Predictions | Undisturbed | 0.829 ± 0.001 | 0.078 ± 0.001 | 0.907 ± 0.001 |
| | Disturbed | 0.01 ± 0.001 | 0.083 ± 0.001 | 0.093 ± 0.001 |
| | Total | 0.839 ± 0 | 0.161 ± 0 | 1 |

When looking at the single AOI's results, omission errors range between 0.2 ± 0.027 (DL Base) and 0.249 ± 0.008 (DL +IND) in NRW (Supplementary Table 3). When only considering larger disturbance extents within a pixel (> 50%), the omission errors in NRW are reduced to 0.029 ± 0.011 (DL base) and 0.052 ± 0.005 (DL +IND). As opposed to that, omission errors reach up to 0.85 ± 0 (DL +IND) in THU. Commission errors are lower than 0.1 in almost every AOI and setup, but are highest in NRW and RLP, reaching up to 0.103 ± 0.034 (DL IND) and 0.036 ± 0.024 (DL base), respectively.

### 3.2. Disturbance maps, detectable minimal disturbance extent

Concerning the disturbance area of a pixel, all methods' Producer's Accuracies increase with larger damage proportions: DL base shows Producer's Accuracies higher than 50% (i.e., better than random) for disturbance areas of approx. 40 $m^2$ (= 40%, as pixel area is 100 $m^2$), while DL +IND and DL IND reach more than 50% Producer's Accuracy at approx. 60 $m^2$ disturbance area (Fig. 3). In the following, we focus on THU, NRW and RLP for brevity. In THU, most disturbed pixels are missed by all methods except DL base and DL IND with approx. 76% and 65% Producer's Accuracies in the disturbance stratum of >90% (Supplementary Fig. 2). Fig. 4, however, shows that the large disturbance areas (especially logging) in THU are captured completely by DL base,

**Table 6**
Performance metrics across all AOIs for the three setups. OE: omission error (>50%: only disturbances larger than 50% of pixel), CE: commission error, BA: balanced accuracy, OA: overall accuracy.

| Setup | CE | OE | F1 | BA | OA |
|---|---|---|---|---|---|
| DL base | 0.032 ± 0.005 | 0.34 ± 0.014 (0.204 ± 0.011) | 0.722 ± 0.001 | 0.814 ± 0.005 | 0.918 ± 0.002 |
| DL IND | 0.023 ± 0.003 | 0.4 ± 0.018 (0.27 ± 0.018) | 0.697 ± 0.008 | 0.788 ± 0.008 | 0.916 ± 0.001 |
| DL +IND | 0.012 ± 0.001 | 0.485 ± 0.004 (0.368 ± 0.003) | 0.654 ± 0.002 | 0.752 ± 0.002 | 0.912 ± 0 |

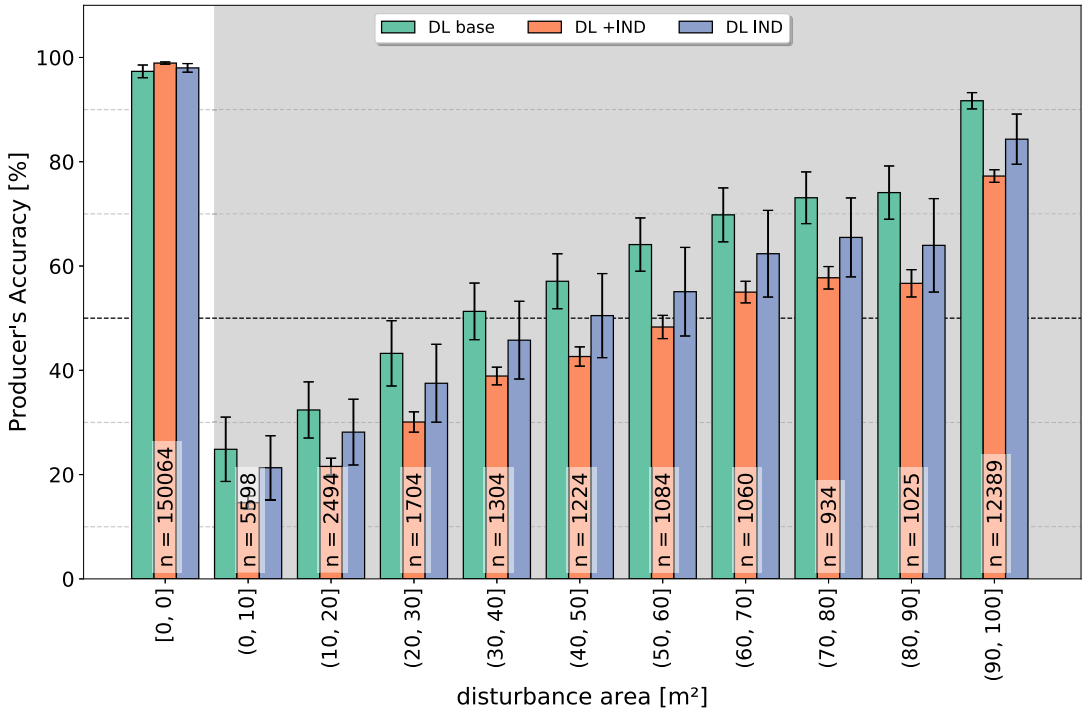


**Fig. 3.** Barplot showing the Producer's Accuracies of the three studied methods (DL base, DL +IND, DL IND) for undisturbed class (left group of bars; white background) and disturbed class (dieback and logging pixels) stratified by disturbance area in 10 $m^2$ strata (other ten groups of bars, gray background) of spatial block cross-validation with three repetitions across all of the six AOIs. Numbers indicate amount of pixels in the respective class/stratum. Error bars indicate standard deviation.

and partially by DL IND/+IND.

The small dieback areas in the northeast of THU are not captured by any method. As opposed to that, all methods show accuracies greater than 50% for disturbances between approx. 20–30 $m^2$ in NRW (Supplementary Fig. 3). In NRW, small disturbances are partly detected by all setups, except the very small dieback areas (mainly single dead trees) in the southwest, south and southeast of the AOI (Fig. 5). Finally, Producer's Accuracies of the disturbed class pixels in RLP are greater than 50% from damage areas of between approx. 40 $m^2$ in DL base, while the other two setups reach more than 50% Producer's Accuracy only in the disturbance stratum of >90% (Supplementary Fig. 4). In RLP, however, DL base reaches a lower Producer's Accuracy in the undisturbed class (approx. 96%), while the other setup's Producer's Accuracies are approx. 99%. This is reflected in patches of false positives by the DL base setup, which are not present in DL +IND and DL IND (Fig. 6). Results for SAX, LUX and BB are shown in Supplementary Figs. 2, 3, and 4.

## 4. Discussion

In this study, we use data from (mostly public) forest disturbance datasets and six visually interpreted AOIs to train transformer models to detect forest disturbances using satellite image time series (SITS) in a two-step approach (pre-training and finetuning). We include samples with small disturbed pixel fractions in training in order to test if these models, using either ten S2 bands, 10 VIs or both, are capable of detecting forest disturbances of sub-pixel size.

In the following, we discuss the three models' performances and challenges (Section 4.1), assess the minimal detectable disturbance extent of our method (Section 4.2) and discuss the performance of the models depending on the input of VIs (Section 4.3). We discuss the generalisability of the models in Section 4.4. Section 5 draws conclusions from the study.

### 4.1. Research question 1: to what degree can Transformers accurately detect forest disturbance on formerly unseen S2 time series?

Several SITS-based forest disturbance monitoring products on Central European forests have been published so far, e.g. Puhm et al. (2020), Senf and Seidl (2020); Thonfeld et al. (2022) and Dutrieux et al. (2021a).

Senf and Seidl (2020) use Landsat SITS and the LandTrendr segmentation algorithm (Kennedy et al., 2010) to detect trend changes in medoid composites of each year's growing season, yielding commission and omission errors of 0.17 and 0.37, respectively. Thonfeld et al. (2022) define a threshold Tasselled-Cap Disturbance Index value using S2 and Landsat 8 in the presumably undisturbed year of 2017 (before the hot drought years) and use this value as threshold for disturbance detection afterwards. Their commission and omission errors are 0.29 and 0.08, respectively. Two approaches using fitted harmonic models of undisturbed time series of each pixel, and detecting deviations of their consecutive predictions from new observations, have been developed by Puhm et al. (2020) and Dutrieux et al. (2021a). They achieve average commission errors of 0.23 and 0.15 as well as average omission errors of 0.2 and 0.1, respectively. Our validation indicates a smaller commission error than the mentioned methods of between approx. 0.012 (DL +IND) and 0.032 (DL base). Regarding the omission error, most of the mentioned methods achieve better performance than ours, which yield

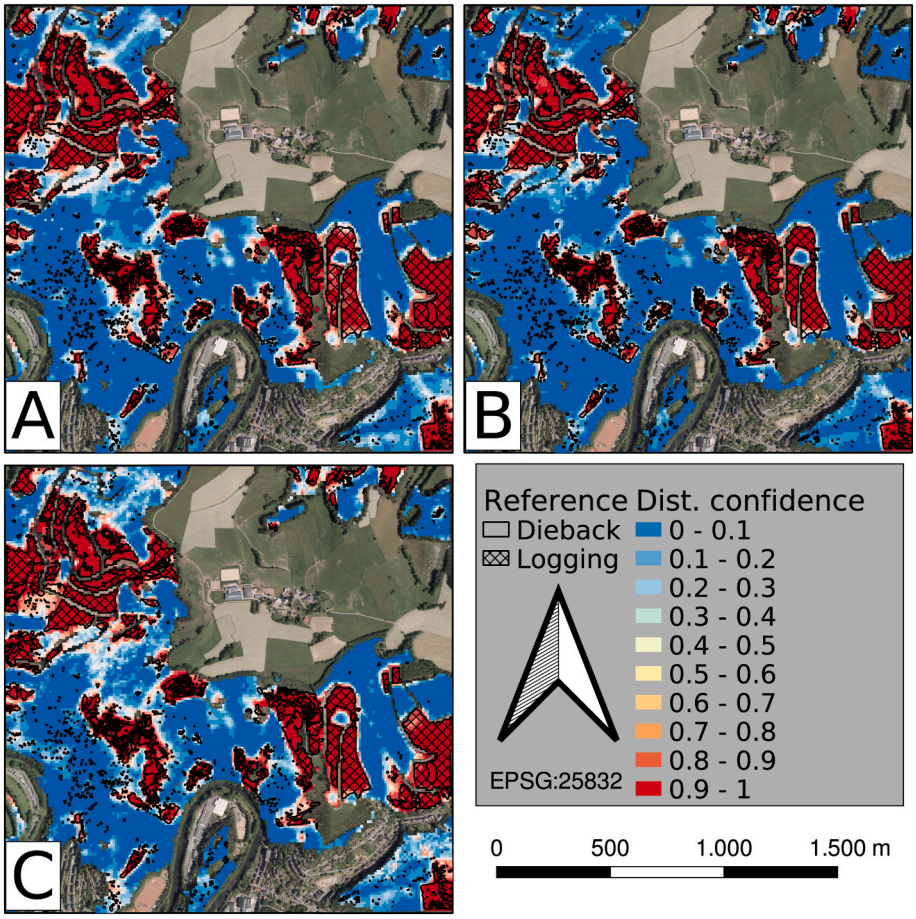


**Fig. 4.** Disturbance maps of NRW for A) DL base (top left), B) DL +IND (top right) and C) (bottom left) of the best seed of spatial block cross-validation with three repetitions. Predictions (disturbance confidences) greater than 0.5 belong to disturbed class, while predictions lower than 0.5 belong to undisturbed class. The higher the predicted value, the more confident the model is about a disturbance. DOPs courtesy of open.NRW (dl-zero-de/2.0).

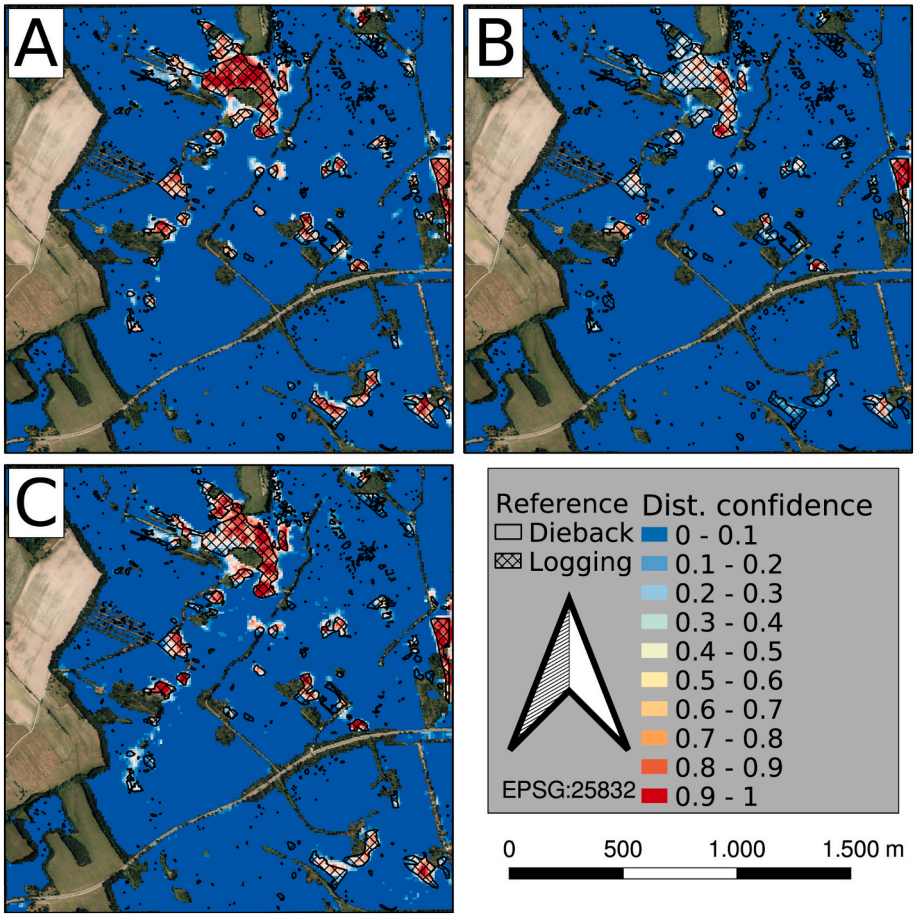


**Fig. 5.** Disturbance maps of THU for A) DL base (top left), B) DL +IND (top right) and C) (bottom left) of the best seed of spatial block cross-validation with three repetitions. Predictions (disturbance confidences) greater than 0.5 belong to disturbed class, while predictions lower than 0.5 belong to undisturbed class. The higher the predicted value, the more confident the model is about a disturbance. DOPs courtesy of Landesamt für Bodenmanagement und Geoinformation Thüringen (dl-de/by-2-0).

between 0.34 (DL base) and 0.485 (DL +IND). The overall omission error in our study, however, is strongly influenced by the very high omission errors in the small disturbance extent strata (cp. Table 6: omission error when only considering disturbances greater than 50%). Note that most of the mentioned publications do not consider partial disturbances (Dutrieux et al., 2021a; Puhm et al., 2020) or explicitly remove disturbances smaller than three pixels from their validation (Thonfeld et al., 2022). When considering only large disturbance extents of at least 90% of a pixel, our methods omit between approximately 0.08 (DL base) and 0.23 (DL +IND) of the disturbances (cp. Fig. 3), thus improving upon most of the aforementioned methods in this disturbance stratum w.r.t. omission error as well.

Additionally, most of the earlier studies (Puhm et al., 2020; Senf and Seidl, 2020; Thonfeld et al., 2022) have in common a validation using random sampling points, which is not directly comparable to ours, which takes into account all forest pixels of our AOIs and focuses on all disturbance extents in the disturbed class. Recently, Perbet et al. (2024) examined the capability of Transformers to identify disturbances on sub-pixel level (called "partial disturbances", here: partial harvest and partial windthrow, defined as between 25% and 75% of the pixel covered by a disturbed area) on Landsat yearly composites of boreal forest in Canada. They report omission errors of 0.155 (partial windthrow) and 0.205 (partial harvest) with a similar validation approach. While their reported results clearly outperform ours, note that these metrics are strongly dependent on the label distribution (e.g. if most of the partial harvest was actually 25% harvested or rather 75% harvested), which is not reported in the study. Additionally, our method has to cope with the more complex phenology and species composition of deciduous, coniferous and potentially mixed forest types in temperate forest, which imposes many challenges on SITS-based forest disturbance monitoring. Broadleaved and mixed forests exhibit more fluctuations in the spectral signal of the time series, for instance caused by a more pronounced phenological cycle and understory vegetation during leaf-off periods. The DL model architecture used in this study is specifically designed to address these challenges, as half of the weights in the encoder of the models have the sole purpose of modelling the understanding of time and seasonality (Yuan and Lin, 2021). Accordingly, the periodicity of broadleaved forest is almost never confused with a disturbance signal, as shown in the stable and low commission errors across forest types in Supplementary Table 4. While the omission errors are higher in broadleaved forest, Producer's Accuracies are still reasonably high.

Another key challenge of SITS-based forest disturbance detection models is the occurrence of gradual and partial disturbances and co-morbidities. Biotic disturbance agents, for instance, often lead to gradual or partial crown dieback with widely differing durations (Senf et al., 2017), e.g. drought dieback within years (Bigler et al., 2007; Haberstroh et al., 2022) and mortality from bark beetle infestations within weeks (Šťursová et al., 2014). Additionally, disturbance agents differ in their intra-annual timing (e.g. windthrow in winter, bark beetle in spring/summer) (Senf et al., 2017) and co-morbidities occur frequently (Seidl et al., 2017). All of this may lead to mixed spectral responses. To cope with this challenge, different levels of aggregation are considered by Transformers by stacking encoder blocks upon each other (Jawahar et al., 2019; Yuan and Lin, 2021). Hence, the overall

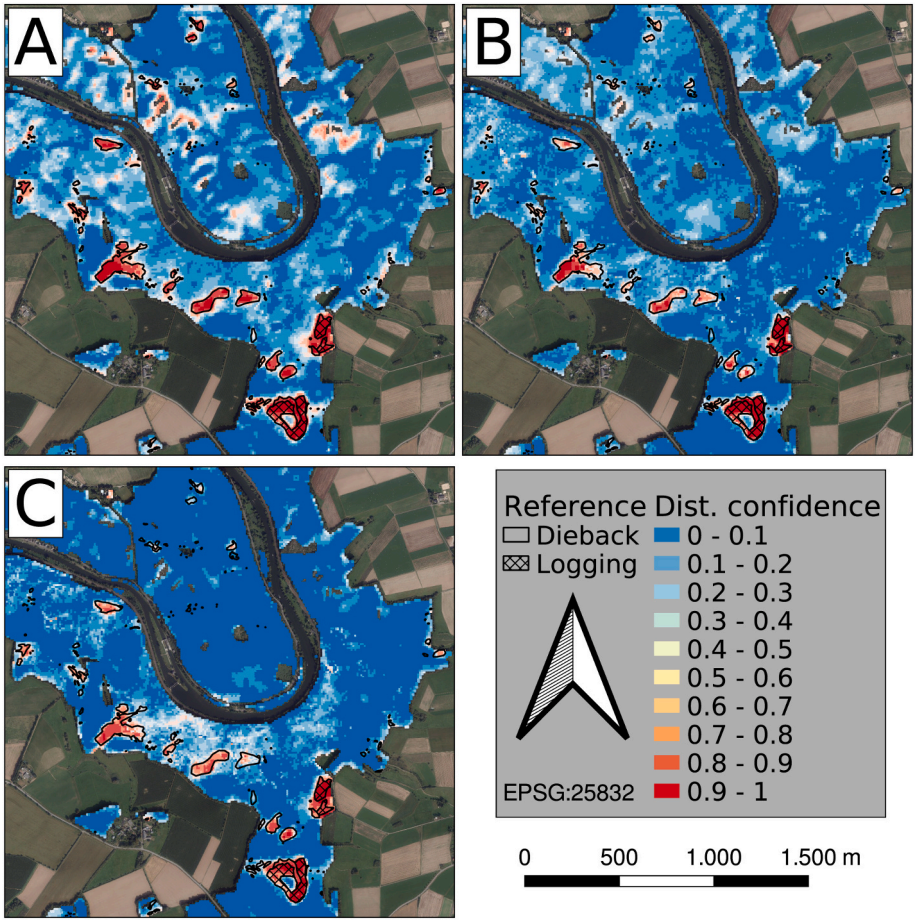


**Fig. 6.** Disturbance maps of RLP for A) DL base (top left), B) DL +IND (top right) and C) (bottom left) of the best seed of spatial block cross-validation with three repetitions. Predictions (disturbance confidences) greater than 0.5 belong to disturbed class, while predictions lower than 0.5 belong to undisturbed class. The higher the predicted value, the more confident the model is about a disturbance. DOPs courtesy of Landesamt für Vermessung und Geobasisinformation Rheinland-Pfalz (dl-de/by-2-0).

periodicity across years can be modelled as well as short-term deviations thereof. As indicated by an additional analysis using explainable AI (Supplementary Fig. 11; Sundararajan et al., 2017), our models seem to take into account temporal cues throughout the whole time series, while at the same time putting emphasis on specific observations. This suggests that sudden forest disturbances such as windthrows and harvest can likely be detected by Transformers as well as gradual drought effects, as confirmed by Perbet et al. (2024) and Du et al. (2023). During training, our models have been exposed to a very large number of time series of multiple disturbance agents' responses (Senf and Seidl, 2020; Thonfeld et al., 2022), potentially enabling them to internalise a large variety of spectral trajectories under different conditions. Due to co-morbidities and the lack of reliable reference data for disturbance agents, we could unfortunately not test if this is really the case.

While all of the models are able to detect most of the disturbance patches across all AOIs (except DL +IND in some cases, e.g. Fig. 5), the performance metrics of our validation reveal large differences in the predictive performance among the six AOIs (cp. Supplementary Table 3, Figs. 3, 4, 5, Supplementary Figs. 2, 3, 4). F1-scores in THU and BB are especially low across all setups. In BB, this might be attributed to the influence of soil background on the spectral trajectory shining through the sparse canopy. BB is the only AOI with large areas of open *Pinus sylvestris* L. stands with a corresponding notable amount of subpixel bare soil fractions.

Given that such stands only occur in BB, our spatial block cross-validation approach may lead to pessimistic results (Kattenborn et al., 2022), as the models have to cope with this feature of the time series unknown from training when trying to distinguish between undisturbed and disturbed pixels. This explanation is supported by the observation that in the southwestern part of BB, the dense forest stands are largely classified correctly, while in the open stands in the northern part, misclassifications occur (especially DL base and DL IND, cp. Supplementary Fig. 7). Secondly, in open forest stands on sandy soils as in BB, the spectral signal of soil likely superimposes the overall spectral trajectory of the forest canopy. This might also be the case for open bedrock in the steep slopes in our AOI in RLP, which might explain the overpredictions of disturbance in the DL base model (cp. Fig. 6).

Another reason for misclassifications in open stands and steep slopes might be a strong impact of the understory on the spectral signal (Haberstroh et al., 2022). This might be a particular problem in case of open forest stands (Eriksson et al., 2006). Since grasslands' photosynthetic activity and the corresponding spectral signal change more rapidly than forests under adverse conditions (Nicolai-Shaw et al., 2017), it is possible that the understory of open forest stands appears stressed more rapidly in drought years, while the response from the trees is delayed (Haberstroh et al., 2022; Nicolai-Shaw et al., 2017; Reinermann et al., 2019). The models then might perceive the degradation of the undergrowth as dieback of the canopy trees in BB.

In THU, the effect of undergrowth on the high omission error might be attributed to natural succession and remaining undergrowth after logging. Here, the disturbances (mostly logging) mainly occurred around 2019/2020 as confirmed by visual interpretation of historical DOPs, while the time series for validation ended in June 2022. Thus, many of the disturbances happened early in the time series, and

vegetation recovery became apparent afterwards. As indicated in Supplementary Fig. 11, all setups tend to regard the most recent observations of the time series as most important, meaning that re-growth and natural succession have a large impact on the predictions. Consequently, vegetation (but not necessarily tree) recovery seems to lead to undisturbed predictions in DL +IND and DL IND, and less pronounced in DL base. This is supported by the observation that recent clear-cuts revealing bare soil as shown in NRW (Fig. 4) are detected much more accurately than the large logging activities in THU. Consequently, the performance on logging areas will likely improve if running the models operationally on each new S2 observation and keeping record of disturbed predictions. Such an operational mode is fully feasible for the current version of the model. This would also enable the retrieval of a disturbance date, which the model is currently not capable of providing for a single prediction. To show the mentioned effect, we made predictions on THU for each year between 2019 and 2022 using DL base, revealing that the disturbances have been captured earlier, but succession led the model to revert to undisturbed predictions afterwards (e.g. in the north- and southeast in Supplementary Fig. 8).

Since dieback patches are often logged afterwards, the dieback apparent in our DOPs likely happened closer to the DOP acquisition date than many of the logging activities (if they happened earlier, they would appear as logging). Observing that recent disturbances are detected more accurately by our models (see above), it is plausible that less omissions occur when considering only dieback and excluding logging from analyses (Supplementary Fig. 9). Here, Producer's Accuracies increase to approx. 59% in the 40–50 $m^2$ and 94% in the >90 $m^2$ disturbance extent strata (both DL base).

### 4.2. Research question 2: what is the smallest disturbance extent that can be detected by the proposed method?

While large disturbances above 90% disturbance extent of a pixel are detected reliably by all setups (e.g. >90% Producer's Accuracy in DL base, see Section 4.1), a drop of Producer's Accuracies is visible in the 60–90% disturbance extents. This might indicate that even large partial disturbances are much more difficult to detect than stand-replacing disturbances, as they still show a spectral trajectory resembling forest. Additionally, these pixels are usually located at the edge of disturbance areas, which are difficult to delineate. Hence, inaccuracies in the geolocation of the reference data compared to the S2 data or imprecisely labelled disturbances could also lead to this effect. Additionally, very small and isolated forest disturbance patches are mostly omitted by our transformer models. Examples are the single isolated trees and small dieback patches scattered around SAX (Supplementary Fig. 5), in the southwest of NRW (Fig. 4) and in the northwest of BB (Supplementary Fig. 7), among others. One reason that single isolated or small groups of trees cannot be detected is that we exclude pixels with disturbance fractions of less than 10% from training. This, however, was a necessary step, since the model needs to be able to distinguish between the two classes in the first place. We decreased this threshold as much as possible in order to allow for the detection of disturbances that are as small as possible. Investigating the distribution of disturbance confidences among the disturbance extent strata (severities; Supplementary Fig. 10) indicates that the models are 1) able to identify undisturbed and clearly disturbed (large disturbance fractions) pixels with high confidence, and 2) are increasingly confident about disturbance predictions with increasing disturbance size. Obviously, there is greater uncertainty in the smaller disturbance area strata (Supplementary Fig. 10), as the model has to distinguish between phenological fluctuations of partly intact forest, the potentially diverse disturbance signals superimposed by phenology (Perbet et al., 2024), spectral influence of the understory (see Section 4.1), and noise inherent to satellite data (atmospheric and topographic effects, etc.). Still, the models achieve a Producer's Accuracy greater than 50% for disturbance fractions as small as approx. 40 $m^2$ (DL base) and 60 $m^2$ (DL +IND, DL IND). When regarding coniferous trees only, the minimum disturbance extents with Producer's Accuracies greater than 50% are even smaller (e.g. 30–40 $m^2$ stratum for DL base, Supplementary Table 4). In coniferous forest, detecting small disturbances is more important, as it may be beneficial for the detection of the beginning of bark beetle infestations and other pests (Kautz et al., 2023). This might enable early warnings in future applications. The capability of Transformers to detect disturbances on sub-pixel level using SITS has been confirmed by Perbet et al. (2024) even on yearly composites. Yet, the disturbance signal of disturbances smaller than 30–40% of a pixel (in coniferous forest; about 50% in deciduous forest, Supplementary Table 4) might be too subtle and dominated by noise and phenological variability for a reliable detection.

### 4.3. Research question 3: do Transformers need vegetation/disturbance indices for accurate predictions?

Although the three DL setups are exposed to roughly the same information, they express considerable differences in their predictions (Fig. 3). This is surprising, since DL +IND should be able to revert to the S2 bands as in DL base and the VIs as in DL IND. The VIs of DL IND and DL +IND, in turn, resort to the S2 bands exclusively. Therefore, one might expect the models to draw similar conclusions from the SITS, and the DL +IND setup to be best, as it contains all the information. The VIs, however, have been handcrafted for a specific purpose, meaning that they condense information to be sensitive for a specific vegetation feature. Some of the applied VIs, for instance, are sensitive to changes in chlorophyll content or cell structure (e.g. NDVI, NDRE, TCG), while others focus on water stress (MSI, TCW, NDMI, NDWI, CRSWIR). This also imposes prior knowledge on the DL IND/DL +IND setups, which might limit DL IND's predictive capability (so-called *hypothesis space*) to specific target features. Although not restricted to specific features as DL IND, we also include a high redundancy of information in DL +IND, which might confuse the model. Zhu et al. (2020) also confirmed that inputting more than the essential bands to their model does not lead to further improvement of model performance. In our case, it even results in DL +IND performing worse than the other two methods.

As opposed to that, DL base can combine the S2 bands differently and more dynamically, not being limited to prior knowledge constraints. Since DL base is able to detect more (and smaller) disturbances, while appearing to wrongly classify more undisturbed pixels, it seems to be more sensitive to subtle deviations of the satellite signal from the undisturbed class. This is also indicated by the higher confusion between confidence scores regarding small dieback and undisturbed forest in DL base compared to DL +IND/IND (Supplementary Fig. 10). Also, DL base appears to overpredict the disturbed class in case of steep terrain, as visible in the center and north of RLP (Fig. 6). Steep terrain usually has shallower soils, sometimes even exposing bedrock, which desiccates more rapidly. Thus, a higher sensitivity to small disturbances inevitably leads to commission errors regarding forests with less stable spectral conditions. Indeed, a certain amount of false positives is likely unavoidable in the attempt of detecting small forest disturbance. The output of the DL models, however, is not a binary class label in the first place, but rather a confidence score for the disturbed class. Therefore, it is possible to shift the threshold for binarization to, for instance, 0.8 rather than the commonly used 0.5 to address these false positives. When doing so, commission errors are further reduced and these potential systematic errors are avoided (cp. Fig. 6). We conclude that our models do not rely on VIs for accurate predictions.

### 4.4. Generalisability

In this study, we tried to cover many different environmental and forest conditions in our six validation AOIs. Yet, the number of pixels and AOIs as well as their distribution was limited due to limited human resources for labeling and availability of DOPs across Germany. Thus, large parts of Germany such as the north and the southeast, including

whole ecoregions such as the German Alps, have been completely omitted in validation (cp. Fig. 2). While the absence of comprehensive validation data prevents an accuracy assessment for these areas, we provide predictions throughout Germany for time series ending in September 2021 (after the end of the major drought from 2018 to 2020) for a general assessment in Supplementary Fig. 12. These results show major disturbances in regions that are known to be affected strongly by the drought, e.g. eastern Northrhine-Westphalia (A in Supplementary Fig. 12), the Harz Mountains (B) as well as the National Park Saxony Switzerland (C). On the contrary, the south and north of Germany appear less affected by disturbances. All in all, this result shows that the model provides plausible predictions in areas without finetuning data as well. Yet, without comprehensive validation data, it remains unclear if the model is ready to be employed throughout Germany.

## 5. Conclusions and outlook

Here, we present a forest disturbance monitoring method using Transformers, a state-of-the-art Deep Learning (DL) architecture, using dense Sentinel-2 time series. Our models have been trained on a rich input dataset (more than 1 million Sentinel-2 pixels in pre-training, more than 500k in finetuning) consisting of different forest types (coniferous, broadleaved), environmental conditions (steep vs. shallow slopes), forest characteristics (open forest stands with large amount of bare ground in BB vs. dense forest cover in the other AOIs) and disturbance agents (windthrow, bark beetle, drought damage, etc.). Our transformer models do not need compositing or gap filling and can cope with multivariate time series as input. The results are stable across different random sets of training data (three repetitions in spatial block cross-validation) and number of observations in the time series (average number of observations: 62 in LUX vs. 180 in THU). Since DL base detects more (and smaller) disturbances than the setups using VIs, handcrafted vegetation and disturbance indices seem not to be required by Deep Learning-based remote sensing of forest disturbance. The rather small performance gains compared to existing methods, however, are traded off by a large amount of training data needed and the complexity of the model.

Since disturbances that happened early in the classified time series tend to be omitted, an operational monitoring system should encompass a record of former detections. Since we vary the end dates of the input time series throughout training, we expect the models to be applicable at any time of the year. Given that our DL base model is able to capture subpixel disturbances starting at about 40 $m^2$ and larger, this might enable a near real-time monitoring by invoking a new model inference every time a new S2 acquisition is available, and comparing the change of its predictions afterwards. However, to what extent our models are capable of timely detections of emerging disturbances such as early-stage bark beetle infestations has yet to be investigated in further studies.

Some steps can be undertaken to further enhance the model's performance and generalisability. The data coverage in finetuning must be improved, e.g. by incorporating data from southeastern and northern Germany. We hope that with future advancements in open data policies, this can be realized in the near future. Limited access to remote sensing data (e.g. aerial images) collected by federal administrations currently prevents an operational nationwide application of the model in Germany. Furthermore, the spatial context of time series also contains valuable information about disturbance regimes and their growth, e.g. bark beetles spreading out to infest neighboring trees (Kautz et al., 2023). Recently developed adaptations of transformer models can also incorporate the spatial context (Yuan et al., 2022), which could be beneficial for enhancing detection capabilities. Such an approach could also enable a model to distinguish between standing deadwood and logging, which would be valuable information for many stakeholders in forestry (Holzwarth et al., 2023).

## Funding information

This work was funded by German Federal Ministry for Environment, Nature Conservation, Nuclear Safety and Consumer Protection in the Future Forest project under grant number 67KI21002C.

## CRediT authorship contribution statement

**Christopher Schiller:** Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Jonathan Költzow:** Writing – review & editing. **Selina Schwarz:** Writing – review & editing, Data curation. **Felix Schiefer:** Writing – review & editing, Data curation. **Fabian Ewald Fassnacht:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

## Declaration of competing interest

The authors declare no competing interests.

## Data availability

The code to reproduce the findings and continue with the trained model in further studies will be available upon acceptance of this manuscript under https://github.com/ChrSchiller/dl_forest_disturbance. Interactive versions of the maps and barplots shown in this manuscript are available under: https://futureforest.eu.pythonanywhere.com/servedownload/dl_forest_decline. The five disturbance datasets of the AOIs from Germany are available upon request under https://zenodo.org/records/8397141.

## Acknowledgments

The authors would like to thank the HPC Service of ZEDAT (*CURTA*), Freie Universität Berlin, for computing power and time. Especially, the first author wants to thank Loris Benett and Bernd Melchers from CURTA for their support, ideas and - above all - their patience and forbearance. We also want to thank the providers of the eo-lab platform for their magnificent service. We are greatly thankful for the open data services of Luxembourg and the German federal states of Brandenburg, Saxony, Thuringia, Northrhine-Westphalia and Rhineland-Palatinate.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rse.2024.114475.

## References

Abdullah, H., Darvishzadeh, R., Skidmore, A., Heurich, M., 2019. Sensitivity of Landsat-8 OLI and TIRS data to foliar properties of early stage bark beetle (Ips typographus, L.). Infestat. Remote Sens. 11, 398. https://doi.org/10.3390/rs11040398.

Ahmed, S., Nielsen, I.E., Tripathi, A., Siddiqui, S., Ramachandran, R.P., Rasool, G., 2023. Transformers in time-series analysis: a tutorial. Circuits Syst. Signal Proc. 42, 7433–7466. https://doi.org/10.1007/s00034-023-02454-8.

Bandyopadhyay, D., Bhavsar, D., Pandey, K., Gupta, S., Roy, A., 2017. Red edge index as an Indicator of vegetation growth and vigor using hyperspectral remote sensing data. Proc. Natl. Acad. Sci., India Sect. A Phys. Sci. 87, 879–888. https://doi.org/10.1007/s40010-017-0456-4.

Baston, D., 2022. Exactextractr: Fast Extraction from Raster Datasets Using Polygons.

Bennett, L., Melchers, B., Proppe, B., 2020. Curta: A General Purpose High-Performance Computer at ZEDAT. Freie Universität, Berlin.

Bigler, C., Gavin, D.G., Gunning, C., Veblen, T.T., 2007. Drought induces lagged tree mortality in a subalpine Forest in the Rocky Mountains. Oikos 116, 1983–1994.

Boegh, E., Soegaard, H., Broge, N., Hasager, C.B., Jensen, N.O., Schelde, K., Thomsen, A., 2002. Airborne multispectral data for quantifying leaf area index, nitrogen concentration, and photosynthetic efficiency in agriculture. Remote Sens. Environ. 81, 179–193. https://doi.org/10.1016/S0034-4257(01)00342-X.

Bösch, M., Elsasser, P., Franz, K., Lorenz, M., Moning, C., Olschewski, R., Rödl, A., Schneider, H., Schröppel, B., Weller, P., 2018. Forest ecosystem services in rural areas of Germany: insights from the national TEEB study. Ecosyst. Serv. 31, 77–83. https://doi.org/10.1016/j.ecoser.2018.03.014.

Bréda, N., Huc, R., Granier, A., Dreyer, E., 2006. Temperate forest trees and stands under severe drought: a review of ecophysiological responses, adaptation processes and long-term consequences. Ann. For. Sci. 63, 625–644. https://doi.org/10.1051/forest:2006042.

Bundesministerium für Ernährung und Landwirtschaft (BMEL), 2023. Ergebnisse der Waldzustandserhebung 2022 (Results of the Forest Condition Survey 2022). Bonn, Germany.

Cohen, W.B., Yang, Z., Kennedy, R., 2010. Detecting trends in forest disturbance and recovery using yearly Landsat time series: 2. TimeSync — tools for calibration and validation. Remote Sens. Environ. 114, 2911–2924. https://doi.org/10.1016/j.rse.2010.07.010.

Coops, N.C., Shang, C., Wulder, M.A., White, J.C., Hermosilla, T., 2020. Change in forest condition: characterizing non-stand replacing disturbances using time series satellite imagery. For. Ecol. Manag. 474, 118370. https://doi.org/10.1016/j.foreco.2020.118370.

Crist, E.P., Cicone, R.C., 1984. A physically-based transformation of Thematic Mapper Data—The TM Tasseled Cap. IEEE Trans. Geosci. Remote Sens. GE-22, 256–263. https://doi.org/10.1109/tgrs.1984.350619.

Dobor, L., Hlásny, T., Rammer, W., Zimová, S., Barka, I., Seidl, R., 2020. Is salvage logging effectively dampening bark beetle outbreaks and preserving forest carbon stocks? J. Appl. Ecol. 57, 67–76. https://doi.org/10.1111/1365-2664.13518.

Du, B., Yuan, Z., Bo, Y., Zhang, Y., 2023. A combined deep learning and prior knowledge constraint approach for large-scale Forest disturbance detection using time series remote sensing data. Remote Sens. 15, 2963. https://doi.org/10.3390/rs15122963.

Dutrieux, R., Féret, J.-B., Ose, K., 2021a. Mise au point d'une méthode reproductible pour le suivi généralisé des dégats de scolytes par télédétection satellitaire. ONF Rend. Techn. ONF Rend. Techn. 37–44.

Dutrieux, R., Féret, J.-B., Ose, K., De Boissieu, F., 2021b. Package Fordead.

EO-Lab, 2023. Future Forest Projct (FF.ai).

Eriksson, H.M., Eklundh, L., Kuusk, A., Nilson, T., 2006. Impact of understory vegetation on forest canopy reflectance and remotely sensed LAI estimates. Remote Sens. Environ. 103, 408–418. https://doi.org/10.1016/j.rse.2006.04.005.

Forzieri, G., Pecchi, M., Girardello, M., Mauri, A., Klaus, M., Nikolov, C., Rüetschi, M., Gardiner, B., Tomaštík, J., Small, D., Nistor, C., Jonikavicius, D., Spinoni, J., Feyen, L., Giannetti, F., Comino, R., Wolynski, A., Pirotti, F., Maistrelli, F., Savulescu, I., Wurpillot-Lucas, S., Karlsson, S., Zieba-Kulawik, K., Strejczek-Jazwinska, P., Mokroš, M., Franz, S., Krejci, L., Haidu, I., Nilsson, M., Wezyk, P., Catani, F., Chen, Y.-Y., Luyssaert, S., Chirici, G., Cescatti, A., Beck, P.S.A., 2020. A spatially explicit database of wind disturbances in European forests over the period 2000–2018. Earth Syst. Sci. Data 12, 257–276. https://doi.org/10.5194/essd-12-257-2020.

Frantz, D., 2019. FORCE—Landsat + Sentinel-2 analysis ready data and beyond. Remote Sens. 11, 1124. https://doi.org/10.3390/rs11091124.

Gao, B., 1996. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens. Environ. 58, 257–266. https://doi.org/10.1016/S0034-4257(96)00067-3.

García, M.J.L., Caselles, V., 1991. Mapping burns and natural reforestation using thematic mapper data. Geocarto Int. 6, 31–37. https://doi.org/10.1080/10106049109354290.

García-Herrera, R., Garrido-Perez, J.M., Barriopedro, D., Ordóñez, C., Vicente-Serrano, S. M., Nieto, R., Gimeno, L., Sorí, R., Yiou, P., 2019. The European 2016/17 drought. J. Clim. 32, 3169–3187. https://doi.org/10.1175/JCLI-D-18-0331.1.

Gazol, A., Camarero, J.J., 2022. Compound climate events increase tree drought mortality across European forests. Sci. Total Environ. 816, 151604. https://doi.org/10.1016/j.scitotenv.2021.151604.

Gitelson, A., Merzlyak, M.N., 1998. Remote sensing of chlorophyll concentration in higher plant leaves. Adv. Space Res. 22, 689–692.

Grabska, E., Hawryło, P., Socha, J., 2020. Continuous detection of Small-scale changes in scots pine dominated stands using dense Sentinel-2 time series. Remote Sens. 12, 1298. https://doi.org/10.3390/rs12081298.

Grillakis, M.G., 2019. Increase in severe and extreme soil moisture droughts for Europe under climate change. Sci. Total Environ. 660, 1245–1255. https://doi.org/10.1016/j.scitotenv.2019.01.001.

Haberstroh, S., Werner, C., Grün, M., Kreuzwieser, J., Seifert, T., Schindler, D., Christen, A., 2022. Central European 2018 hot drought shifts scots pine forest to its tipping point. Plant Biol. J. 24, 1186–1197. https://doi.org/10.1111/plb.13455.

Healey, S., Cohen, W., Zhiqiang, Y., Krankina, O., 2005. Comparison of tasseled cap-based Landsat data structures for use in forest disturbance detection. Remote Sens. Environ. 97, 301–310. https://doi.org/10.1016/j.rse.2005.05.009.

Hijmans, R.J., 2023. Terra: Spatial Data Analysis.

Holzwarth, S., Thonfeld, F., Abdullahi, S., Asam, S., Da Ponte Canova, E., Gessner, U., Huth, J., Kraus, T., Leutner, B., Kuenzer, C., 2020. Earth observation based monitoring of forests in Germany: a review. Remote Sens. 12, 3570. https://doi.org/10.3390/rs12213570.

Holzwarth, S., Thonfeld, F., Kacic, P., Abdullahi, S., Asam, S., Coleman, K., Eisfelder, C., Gessner, U., Huth, J., Kraus, T., Shatto, C., Wessel, B., Kuenzer, C., 2023. Earth-observation-based monitoring of forests in Germany—recent Progress and research Frontiers: a review. Remote Sens. 15, 4234. https://doi.org/10.3390/rs15174234.

Ionita, M., Nagavciuc, V., 2021. Changes in drought features at the European level over the last 120 years. Nat. Hazards Earth Syst. Sci. 21, 1685–1701. https://doi.org/10.5194/nhess-21-1685-2021.

Iwana, B.K., Uchida, S., 2020. Time Series Data Augmentation for Neural Networks by Time Warping with a Discriminative Teacher.

Jawahar, G., Sagot, B., Seddah, D., 2019. What does BERT learn about the structure of language?. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Presented at the proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp. 3651–3657. https://doi.org/10.18653/v1/P19-1356.

Kattenborn, T., Schiefer, F., Frey, J., Feilhauer, H., Mahecha, M.D., Dormann, C.F., 2022. Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. ISPRS Open J. Photogramme. Remote Sens. 5, 100018. https://doi.org/10.1016/j.ophoto.2022.100018.

Kautz, M., Peter, F.J., Harms, L., Kammen, S., Delb, H., 2023. Patterns, drivers and detectability of infestation symptoms following attacks by the European spruce bark beetle. J. Pest. Sci. 96, 403–414. https://doi.org/10.1007/s10340-022-01490-8.

Kennedy, R.E., Yang, Z., Cohen, W.B., 2010. Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr — temporal segmentation algorithms. Remote Sens. Environ. 114, 2897–2910. https://doi.org/10.1016/j.rse.2010.07.008.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., et al., 2020. Captum: A Unified and Generic Model Interpretability Library for Pytorch.

Kong, Y.-L., Huang, Q., Wang, C., Chen, Jingbo, Chen, Jiansheng, He, D., 2018. Long short-term memory neural networks for online disturbance detection in satellite image time series. Remote Sens. 10, 452. https://doi.org/10.3390/rs10030452.

Langner, N., Puhm, M., Deutscher, J., Wimmer, A., Adler, P., Backa, J., Eisenecker, P., Reinosch, E., Wiesehahn, J., Hoffmann, K., Oehmichen, K., 2023. FNEWS-Jahresprodukte 2018 bis 2022.

Latifi, H., Fassnacht, F.E., Schumann, B., Dech, S., 2014. Object-based extraction of bark beetle (*Ips typographus* L.) infestations using multi-date LANDSAT and SPOT satellite imagery. Progr. Phys. Geogr.: Earth Environ. 38, 755–785. https://doi.org/10.1177/0309133314550670.

Mandl, L., Lang, S., 2023. Uncovering early traces of bark beetle induced Forest stress via semantically enriched Sentinel-2 data and spectral indices. PFG 91, 211–231. https://doi.org/10.1007/s41064-023-00240-4.

Mouret, F., Morin, D., Martin, H., Planells, M., Vincent-Barbaroux, C., 2024. Toward an operational monitoring of oak dieback with multispectral satellite time series: a case study in Centre-Val De Loire region of France. IEEE J. Sel. Top. Appl. Earth Observat. Remote Sens. 17, 643–659. https://doi.org/10.1109/JSTARS.2023.3332420.

Mu, Y., Guo, J., Shahzad, M., Zhu, X.X., 2024. Nationwide Tree Species Mapping in Germany with Forestformer: Assessing Forest Resilience and Unveiling Distribution Patterns. https://doi.org/10.2139/ssrn.4800294.

Mullissa, A., Reiche, J., Saatchi, S., 2023, July. Seasonal Forest Disturbance Detection Using Sentinel-1 SAR & Sentinel-2 Optical Timeseries Data and Transformers. In: *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, pp. 3122–3124.

Netherer, S., Panassiti, B., Pennerstorfer, J., Matthews, B., 2019. Acute drought is an important driver of bark beetle infestation in Austrian Norway spruce stands. Front. For. Glob. Change 2, 39. https://doi.org/10.3389/ffgc.2019.00039.

Nicolai-Shaw, N., Zscheischler, J., Hirschi, M., Gudmundsson, L., Seneviratne, S.I., 2017. A drought event composite analysis using satellite remote-sensing based soil moisture. Remote Sens. Environ. 203, 216–225. https://doi.org/10.1016/j.rse.2017.06.014.

Pasquarella, V.J., Arévalo, P., Bratley, K.H., Bullock, E.L., Gorelick, N., Yang, Z., Kennedy, R.E., 2022. Demystifying LandTrendr and CCDC temporal segmentation. Int. J. Appl. Earth Obs. Geoinf. 110, 102806. https://doi.org/10.1016/j.jag.2022.102806.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library.

Patacca, M., Lindner, M., Lucas-Borja, M.E., Cordonnier, T., Fidej, G., Gardiner, B., Hauf, Y., Jasinevičius, G., Labonne, S., Linkevičius, E., Mahnken, M., Milanovic, S., Nabuurs, G., Nagel, T.A., Nikinmaa, L., Panyatov, M., Bercak, R., Seidl, R., Ostrogović Sever, M.Z., Socha, J., Thom, D., Vuletic, D., Zudin, S., Schelhaas, M., 2022. Significant increase in natural disturbance impacts on European forests since 1950. Glob. Chang. Biol. 29, 1359–1376. https://doi.org/10.1111/gcb.16531.

Pebesma, E., 2018. Simple features for R: standardized support for spatial vector data. R J. 439–446. https://doi.org/10.32614/RJ-2018-009.

Perbet, P., Guindon, L., Côté, J.-F., Béland, M., 2024. Evaluating deep learning methods applied to Landsat time series subsequences to detect and classify boreal forest disturbances events: the challenge of partial and progressive disturbances. Remote Sens. Environ. 306, 114107. https://doi.org/10.1016/j.rse.2024.114107.

Puhm, M., Deutscher, J., Hirschmugl, M., Wimmer, A., Schmitt, U., Schardt, M., 2020. A near real-time method for Forest change detection based on a structural time series model and the Kalman filter. Remote Sens. 12, 3135. https://doi.org/10.3390/rs12193135.

R Core Team, 2023. R: A Language and Environment for Statistical Computing.

Reinermann, S., Gessner, U., Asam, S., Kuenzer, C., Dech, S., 2019. The effect of droughts on vegetation condition in Germany: an analysis based on two decades of satellite earth observation time series and crop yield statistics. Remote Sens. 11, 1783. https://doi.org/10.3390/rs11151783.

Rock, B.N., 1985. Field and airborne spectral characterization of suspected acid deposition damage in red spruce (Picea rubens) from Vermont. In: Proc. Symp. On Machine Processing Remotely Sensed Data, pp. 71–81.

Rodman, K.C., Andrus, R.A., Veblen, T.T., Hart, S.J., 2021. Disturbance detection in landsat time series is influenced by tree mortality agent and severity, not by prior disturbance. Remote Sens. Environ. 254, 112244.

Schiefer, F., Schmidtlein, S., Frick, A., Frey, J., Klinke, R., Zielewska-Büttner, K., Junttila, S., Uhl, A., Kattenborn, T., 2023. UAV-based reference data for the prediction of fractional cover of standing deadwood from sentinel time series. ISPRS

Open J. Photogramme. Remote Sens. 8, 100034. https://doi.org/10.1016/j.ophoto.2023.100034.
Schwarz, S., Werner, C., Fassnacht, F.E., Ruehr, N.K., 2023. Forest canopy mortality during the 2018-2020 summer drought years in Central Europe: the application of a deep learning approach on aerial images across Luxembourg. Forestr.: Int. J. Forest Res. 97, 376–387. https://doi.org/10.1093/forestry/cpad049.
Seidl, R., Thom, D., Kautz, M., Martin-Benito, D., Peltoniemi, M., Vacchiano, G., Wild, J., Ascoli, D., Petr, M., Honkaniemi, J., Lexer, M.J., Trotsiuk, V., Mairota, P., Svoboda, M., Fabrika, M., Nagel, T.A., Reyer, C.P.O., 2017. Forest disturbances under climate change. Nat. Clim. Chang. 7, 395–402. https://doi.org/10.1038/nclimate3303.
Senf, C., Seidl, R., 2020. Mapping the forest disturbance regimes of Europe. Nat. Sustain. 4, 63–70. https://doi.org/10.1038/s41893-020-00609-y.
Senf, C., Seidl, R., 2021. Persistent impacts of the 2018 drought on forest disturbance regimes in Europe. Biogeosciences 18, 5223–5230. https://doi.org/10.5194/bg-18-5223-2021.
Senf, C., Seidl, R., Hostert, P., 2017. Remote sensing of forest insect disturbances: current state and future directions. Int. J. Appl. Earth Obs. Geoinf. 60, 49–60. https://doi.org/10.1016/j.jag.2017.04.004.
Šturšová, M., Šnajdr, J., Cajthaml, T., Bárta, J., Šantrůčková, H., Baldrian, P., 2014. When the forest dies: the response of forest soil fungi to a bark beetle-induced tree dieback. ISME J. 8, 1920–1931. https://doi.org/10.1038/ismej.2014.37.
Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic Attribution for Deep Networks.
Thom, D., Rammer, W., Seidl, R., 2017. The impact of future forest dynamics on climate: interactive effects of changing vegetation and disturbance regimes. Ecol. Monogr. 87, 665–684. https://doi.org/10.1002/ecm.1272.
Thonfeld, F., Gessner, U., Holzwarth, S., Kriese, J., 2022. A First Assessment of Canopy Cover Loss in Germany's Forests after the 2018–2020 Drought Years, 19.
Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens. Environ. 8, 127–150.
Van Rossum, G., Drake, F.L., 2009. Python 3 Reference Manual.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention Is all you Need.
Verbesselt, J., Hyndman, R., Newnham, G., Culvenor, D., 2010. Detecting trend and seasonal changes in satellite image time series. Remote Sens. Environ. 114, 106–115. https://doi.org/10.1016/j.rse.2009.08.014.
Verbesselt, J., Zeileis, A., Herold, M., 2012. Near real-time disturbance detection using satellite image time series. Remote Sens. Environ. 123, 98–108. https://doi.org/10.1016/j.rse.2012.02.022.
Wittich, D., Rottensteiner, F., Voelsen, M., Heipke, C., Müller, S., 2022. Deep learning for the detection of early signs for Forest damage based on satellite imagery. ISPRS Ann. Photogramm. Remote Sens. Spatial. Inf. Sci. V-2–2022, 307–315. https://doi.org/10.5194/isprs-annals-V-2-2022-307-2022.
Ye, S., Rogan, J., Zhu, Z., Hawbaker, T.J., Hart, S.J., Andrus, R.A., Meddens, A.J.H., Hicke, J.A., Eastman, J.R., Kulakowski, D., 2021. Detecting subtle change from dense Landsat time series: case studies of mountain pine beetle and spruce beetle disturbance. Remote Sens. Environ. 263, 112560. https://doi.org/10.1016/j.rse.2021.112560.
Yuan, Y., Lin, L., 2021. Self-supervised Pretraining of transformers for satellite image time series classification. IEEE J. Sel. Top. Appl. Earth Observat. Remote Sens. 14, 474–487. https://doi.org/10.1109/JSTARS.2020.3036602.
Yuan, Y., Lin, L., Liu, Q., Hang, R., Zhou, Z.-G., 2022. SITS-former: a pre-trained spatio-spectral-temporal representation model for Sentinel-2 time series classification. Int. J. Appl. Earth Obs. Geoinf. 106, 102651. https://doi.org/10.1016/j.jag.2021.102651.
Zhang, H.K., Luo, D., Li, Z., 2024. Classifying raw irregular time series (CRIT) for large area land cover mapping by adapting transformer model. Sci. Remote Sens. 9, 100123. https://doi.org/10.1016/j.srs.2024.100123.
Zhu, Z., Woodcock, C.E., 2014. Continuous change detection and classification of land cover using all available Landsat data. Remote Sens. Environ. 144, 152–171. https://doi.org/10.1016/j.rse.2014.01.011.
Zhu, Z., Zhang, J., Yang, Z., Aljaddani, A.H., Cohen, W.B., Qiu, S., Zhou, C., 2020. Continuous monitoring of land disturbance based on Landsat time series. Remote Sens. Environ. 238, 111116. https://doi.org/10.1016/j.rse.2019.03.009.