https://doi.org/10.5194/essd-2025-670
Preprint. Discussion started: 24 November 2025

Open Access Earth System Science Data Discussions

A Global Dataset of Forest Disturbance Regimes Derived from Satellite Biomass Observations

Siyuan Wang[1,2], Hui Yang[3], Sujan Koirala[1], Maurizio Santoro[4], Ulrich Weber[1], Claire Robin[1], Felix Cremer[1], Matthias Forkel[2], Markus Reichstein[1,5], Nuno Carvalhais[1,5,6]

[1]Max-Planck Institute for Biogeochemistry, Jena, Germany
[2]TU Dresden, Institute of Photogrammetry and Remote Sensing, Dresden, Germany
[3]Peking University, College of Urban and Environmental Sciences, Beijing, China
[4]Gamma Remote Sensing, Gümligen, Switzerland
[5]Departamento de Ciências e Engenharia do Ambiente, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal
[6]ELLIS Unit Jena, Jena, Germany

Correspondence to:
Siyuan Wang (siyuan.wang@bgc-jena.mpg.de)
Nuno Carvalhais (nuno.carvalhais@bgc-jena.mpg.de)

Abstract. Forests play a central role in the global carbon cycle by serving as critical carbon sinks for atmospheric CO_2. Yet, the stability and continued capacity of these sinks are increasingly threatened by a growing number of disturbances. Accurately representing the stochastic nature of disturbance remains a major challenge and a key source of uncertainty in our understanding of carbon cycle dynamics. This study presents a novel framework for deriving disturbance regimes characterized by extent (μ), frequency (α), intensity (β), as well as background mortality (K_b) directly from landscape features of high-resolution satellite biomass data. These regimes reflect the characteristics of long term disturbances at the landscape scale rather than the properties of any single event. Our analysis inverts the forward model framework developed by Wang et al. (2024), which used a machine learning model trained on a massive synthetic dataset of over 8 million forward model simulations to link known disturbance regimes to spatial biomass patterns. Instead of predicting patterns from regimes, we use observed satellite biomass patterns to infer the underlying disturbance regimes. To ensure robustness, we first identified the optimal spatial resolution for aggregating both simulation and satellite data, minimizing discrepancies in feature value ranges and reducing extrapolation risk. Using this framework, we produced the first globally consistent, observationally constrained dataset of forest disturbance regime parameters and their associated uncertainties, provided at both a 25x25 km^2 tile level and as a gridded 0.25° global product. Additionally, we used a Dissimilarity Index (DIK) to quantify prediction uncertainty and identify potential extrapolation by measuring observations' divergence from the training set. An empirical evaluation of borderline disturbance regimes supports the assumptions and methodological approach used to build the dataset. Our global maps of disturbance regimes provide a novel, process-based tool for investigating the coupled dynamics of disturbance,

https://doi.org/10.5194/essd-2025-670
Preprint. Discussion started: 24 November 2025

vegetation, and the carbon cycle, with potential applications for improving the representation of stochastic disturbances in large-scale ecosystem models.

1 Introduction

Global forests serve as significant carbon sinks, playing a vital role in mitigating climate change through sequestering atmospheric carbon dioxide derived from anthropogenic fossil fuel burning and land use change (Reichstein and Carvalhais, 2019). The mean carbon sink attributed to forests has remained steady at around 3.6 Pg C yr^{-1} since the 1990s, surpassing the ocean sink at around 2.3 Pg C yr^{-1} (Friedlingstein et al., 2022; Pan et al., 2024). However, the persistence of this sink is increasingly threatened by a wide array of natural and anthropogenic disturbances, including fires, droughts, insect outbreaks, windthrow, and land use change (Kulakowski et al., 2017; Wohlgemuth et al., 2022). The frequency, temporal duration, and spatial extent of these disturbances remain highly unknown; as a consequence, the resulting large-scale tree damage and mortality and their effects on the carbon cycle are poorly quantified (Senf and Seidl, 2021a, b), creating a primary source of uncertainty in the projection of future carbon cycle dynamics within Earth System Models (ESMs) (Friend et al., 2014; Seidl et al., 2014). A key limitation of current ESMs is their overly simplistic representation of forest disturbance dynamics (Seidl et al., 2011; Wohlgemuth et al., 2022), or in some cases their complete omission, due to limited understanding of the spatiotemporal regimes that dictate the long-term impact of these events on forest carbon cycling dynamics (Turner, 2010; Turner and Seidl, 2023).

Efforts to characterize forest disturbance regimes, including their frequency, intensity, and extent, from Earth observation have largely followed two distinct approaches. The individual event-detection method, which applied a continuous detection change algorithm to time series of satellite imagery to identify and map individual disturbance events (Kennedy et al., 2010; Senf and Seidl, 2021a, b). This approach provides a detailed historical record of disturbances and benefits from a growing diversity of data sources, including optical, microwave (radar), and laser-based (lidar) data . However, it is difficult to use this method to derive regime parameters that reflect disturbance dynamics over past decades (Turner, 2010), because it heavily relied on good-quality and continuous time series that were often lacking (Fisher et al., 2008; Chambers et al., 2013). This is especially true for high-spatial-resolution data, which have only become available relatively recently and cover short temporal period. On the other hand, the second approach infers disturbance regimes from landscape-scale characteristics such as spatial patterns of forest biomass derived from Earth observation data (Williams et al., 2013), then uses ecosystem models to inversely estimate disturbance parameters that synthesize historical disturbance dynamics. A key limitation, however, is that different combinations of disturbances (e.g., long duration and weak drought vs. rapid severe heatwave) can produce similar landscape outcomes; the use of coarse-resolution data often led to equifinality (Delbart et al., 2010; Williams et al., 2013). The recent proliferation of globally consistent, high-resolution satellite biomass products now provides the critical observational

https://doi.org/10.5194/essd-2025-670
Preprint. Discussion started: 24 November 2025


Open Access Earth System Science Data Discussions

foundation to test and apply this pattern-based framework at a global scale (Quegan et al., 2019; Toan et al., 2018; Reichstein and Carvalhais, 2019).

This study was built upon our previous framework (Wang et al., 2024), which first used a forward-modeling approach to create a synthetic forest regimes dataset (8 million regime parameter combinations) that linked to unique biomass patterns and landscape-level photosynthetic capacity, using a machine learning model. The primary objective of this study is to globally map forest disturbance regime parameters (extent μ [%], frequency α [-], intensity β [-], and background mortality K_b [$year^{-1}$]) by inverting this advanced forest disturbance framework against biomass landscape features derived from 25 m resolution ESA CCI biomass data (Santoro et al., 2021) and landscape-level photosynthetic capacity from FLUXCOM data (Nelson et al., 2024). To achieve this, we first extended our synthetic forest regimes dataset by broadening the parameterization range and incorporating a diversity of non-rectangular disturbance shapes, increasing the number of regime parameter combinations from 0.85M to 8M (see supplementary S1). Second, we identified the optimal spatial resolution for aggregating simulation and EO data to minimize discrepancies in biomass feature value ranges between simulations and observations, thereby reducing extrapolation risk. Third, we applied landscape features calculated from EO data into the pre-trained machine learning model to derive forest disturbance regime parameters. Additionally, we developed a scalable Dissimilarity Index (DIK) to provide a spatially explicit measure of model applicability uncertainty for the final predictions (the entire workflow see Figure 1).

In the following sections, we detail this comprehensive framework: Section 2 describes the input datasets and multi-stage methodology, from the forward modeling and observational data processing to the machine learning prediction and uncertainty quantification. Section 3 presents the data products in both their native tile-level and gridded global formats and provides a comprehensive evaluation of spatially explicit uncertainty and scientific plausibility. This dataset provides a novel, observationally constrained tool for improving the representation of stochastic disturbances, with the potential to reduce a key uncertainty in future carbon cycle projections.

https://doi.org/10.5194/essd-2025-670
Preprint. Discussion started: 24 November 2025

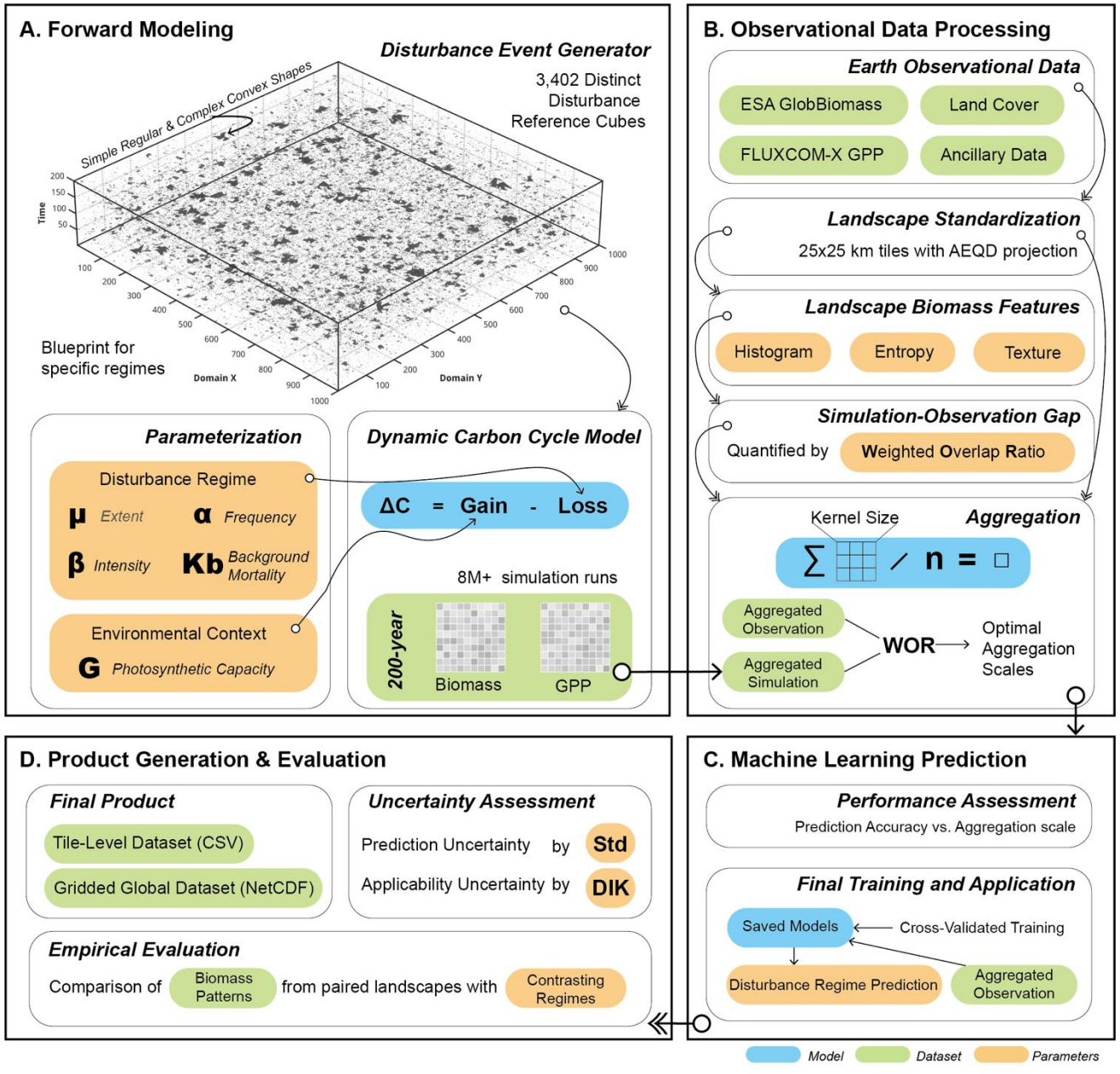


Figure 1. Conceptual workflow of the disturbance regime prediction framework. The framework is organized into four stages. (A) Forward Modeling: A synthetic dataset is created by simulating how known disturbance regimes (μ, α, β, Kb) produce unique biomass patterns. (B) Observational Data Processing & Gap Analysis: Global satellite biomass data is processed into standardized tiles, and the discrepancy between the simulated and observed data is resolved using a spatial aggregation strategy. (C) Machine Learning Prediction: A Random Forest model is trained on the aligned synthetic data and applied to the observed data to predict disturbance regimes. (D) Product Generation & Evaluation: The final tile-level and gridded global datasets are produced and then evaluated for uncertainty and scientific plausibility.

https://doi.org/10.5194/essd-2025-670
Preprint. Discussion started: 24 November 2025

Open Access Earth System Science Data Discussions

2 Data and Methods

2.1 Observational Datasets and Post-Processing

This analysis uses three primary observational datasets to derive predictive features: (1) the ESA GlobBiomass product (Santoro et al., 2021), to characterize fine-scale biomass spatial patterns; (2) the Copernicus land cover dataset (Buchhorn et al., 2020), to provide a high-resolution forest mask; and (3) the FLUXCOM-X GPP product (Nelson et al., 2024), to represent landscape-level photosynthetic capacity. The specifics of each dataset and its subsequent processing are detailed below.

Biomass observation

We selected the ESA GlobBiomass product dataset, as it provides a globally consistent, spatially explicit map of above-ground biomass for the year 2010 at a native resolution of approximately 25 meters at the equator, the highest publicly available spatial resolution. Its high resolution is critical for resolving the fine-scale spatial heterogeneity of forest distribution required for our analysis. The dataset was generated primarily from a fusion of Synthetic Aperture Radar (SAR) backscatter observations, including L-band data from ALOS PALSAR and C-band data from ENVISAT ASAR. The retrieval algorithm first estimates Growing Stock Volume (GSV) and subsequently converts it to AGB using spatially explicit layers of wood density and biomass expansion factors, resulting in a product that captures key structural attributes of forests across the globe.

To transform the raw GlobBiomass map into a set of predictive features, we first established a global grid of analysis domains and then systematically characterized the biomass structure within each. A global grid of non-overlapping, true-to-area 25 km × 25 km landscapes was generated using geodetic calculations on the WGS84 ellipsoid, ensuring that each landscape unit represents a consistent surface area regardless of latitude. For each landscape, the corresponding 25 m resolution AGB data was extracted and then reprojected to a local Azimuthal Equidistant (AEQD) projection to standardize its internal geometry. The crucial step involved a two-stage resampling process to ensure accurate pixel alignment: the data was first resampled to a 1 m resolution using cubic interpolation and subsequently aggregated via pixel averaging to a final, standardized 1000 × 1000 pixel grid at a 25 m resolution. A global forest cover mask derived from the latest Copernicus land cover dataset was then applied to this standardized grid, assigning non-forest pixels with a NaN value to isolate only forested areas for analysis. Prior to statistical derivation, AGB values were converted from tons per hectare (t/ha) to grams per square meter (g/m^2). From this masked AGB grid, we derived a comprehensive suite of biomass spatial statistics (Supplementary Table S3), including first-order distribution metrics and second-order texture metrics from a Gray-Level Co-occurrence Matrix (GLCM). Critically, to ensure ecological integrity, it was constructed exclusively from adjacent pairs of valid forest pixels in four directions, an adaptive approach that captures the true spatial arrangement of forest structure without introducing artifacts from interpolating across non-forested gaps. For methodological robustness, texture features were only computed for landscapes containing at

https://doi.org/10.5194/essd-2025-670
Preprint. Discussion started: 24 November 2025

least 100 valid forest pixels. The resulting vector of 17 statistical features for each landscape formed the quantitative basis for the machine learning models used to predict disturbance regimes.

Forest cover mask

The forest cover mask used in this study was derived from the Copernicus land cover dataset, which provides the latest global coverage at 100 meter spatial resolution for the year 2019 (Buchhorn et al., 2020). This product includes 23 discrete land cover classes aligned with the UN-FAO Land Cover Classification System. The forest type layer was reprojected and resampled to match the spatial extent and resolution of each biomass tile using nearest-neighbor interpolation. Then, we identified the pixel belonging to any forest categories (e.g., evergreen needleleaf, evergreen broadleaf, deciduous needleleaf, deciduous broadleaf, and mixed forest) and generated a binary forest mask. We used the resulting masks to select all landscapes with a forest cover ratio greater than 0. This process filtered out non-forest areas, ensuring that the biomass statistics extraction and disturbance regime prediction were performed only on forested landscapes.

GPP dataset

We used the GPP product from the FLUXCOM-X (X-BASE) dataset (Nelson et al., 2024) at a spatial resolution of 0.05° to calculate landscape-level photosynthesis capacity. This product is a data-driven product constrained by in-situ eddy covariance observations with superior skill in spatial variation in GPP. We used GPP data from the year 2010, and the monthly GPP values were converted into a total annual sum for each grid cell. For each of the selected forested landscape domains, the annual GPP totals of all grid cells within the landscape's bounding box were spatially averaged into a single value. As a result, for each landscape domain, we obtained one single value, representing the landscape's integrated annual photosynthetic capacity (g C m^{-2} yr^{-1}), which was subsequently used as a key predictor in the machine learning models to predict disturbance regime.

2.2 Calibration of Spatial Aggregation Scale for Model-Data Consistency

Prior to applying statistical features derived from EO biomass data to the machine learning model, we evaluated the value ranges of features identified as highly important for predicting disturbance regimes, including GLCM Correlation and Coefficient of Variation (Supplementary Table S3). This is to ensure these features fall within the range encountered during model training, thereby minimizing extrapolation risk. However, a substantial mismatch was observed between the feature value ranges from the synthetic forest regimes dataset generated by Wang et al. (2024) and those computed from the actual EO biomass data (Supplementary S2.1). This discrepancy was particularly evident for GLCM Correlation, which was [0 – 0.75] for simulation but [0.75 – 0.98] for EO data (Figure S2.1 c). Moreover, expanding the parameter space and incorporating non-rectangular disturbance shapes did not significantly reduce this divergence (Fig. S2.1 b). The primary source of inconsistency was identified in the representation of spatial patterns: the simulation framework treats each grid cell as an

https://doi.org/10.5194/essd-2025-670
Preprint. Discussion started: 24 November 2025

independent unit, whereas real-world landscapes exhibit strong spatial autocorrelation arising from complex interaction among topography, soil, hydrology, and community competition.

To overcome this, we implemented a spatial aggregation procedure for both the simulated and observed data to adjust value ranges of spatial autocorrelation-related features such as GLCM Correlation and Coefficient of Variation. This approach used a moving window (kernel) to systematically aggregate the original 1000 × 1000 pixel biomass maps from both simulations and EO observations to coarser resolutions. By calculating the mean value within the kernel (e.g., a 2 × 2 kernel aggregates four pixels into one), this process smooths the data and fundamentally alters the texture statistics, and it applies a range of distinct aggregation scales to generate a suite of down-sampled biomass maps and their corresponding features, allowing us to find a common scale where the statistical domains of simulation and observation align. As shown in Figure S1, the divergence (quantified by a metric called weighted overlap ratio, WOR; see details in supplementary S2.2) for important features between simulation and EO observation was substantially reduced across coarser scales, i.e., at kernel size ranging from 5 to 40.

A primary concern regarding spatial aggregation for simulation data is the potential reduction in the accuracy of machine learning model prediction due to the loss of fine-scale details. To evaluate this, we assessed whether model prediction accuracy change across spatial aggregation scales. We trained a series of Random Forest models on the synthetic dataset aggregated to different scales using an identical cross-validation scheme with consistent training-testing splitting. The predictive accuracy of the machine learning model on the test folds was quantified using the Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe, 1970), where a value of 1 signifies perfect model-data correspondence (D. N. Moriasi et al., 2007). The results show that predictive accuracy remained high across a wide range of aggregation scales (Fig. 2), dropping only when the kernel size was 40 (at very coarse spatial resolution). In summary, a kernel size of 10 was selected as the optimal balance between prediction accuracy (all NSE values > 0.85 for 4 parameters, Figure 2) and consistency between aggregated simulated and observed biomass features (most of WOR values > 0.9; Supplementary Figure S2.4).

https://doi.org/10.5194/essd-2025-670
Preprint. Discussion started: 24 November 2025

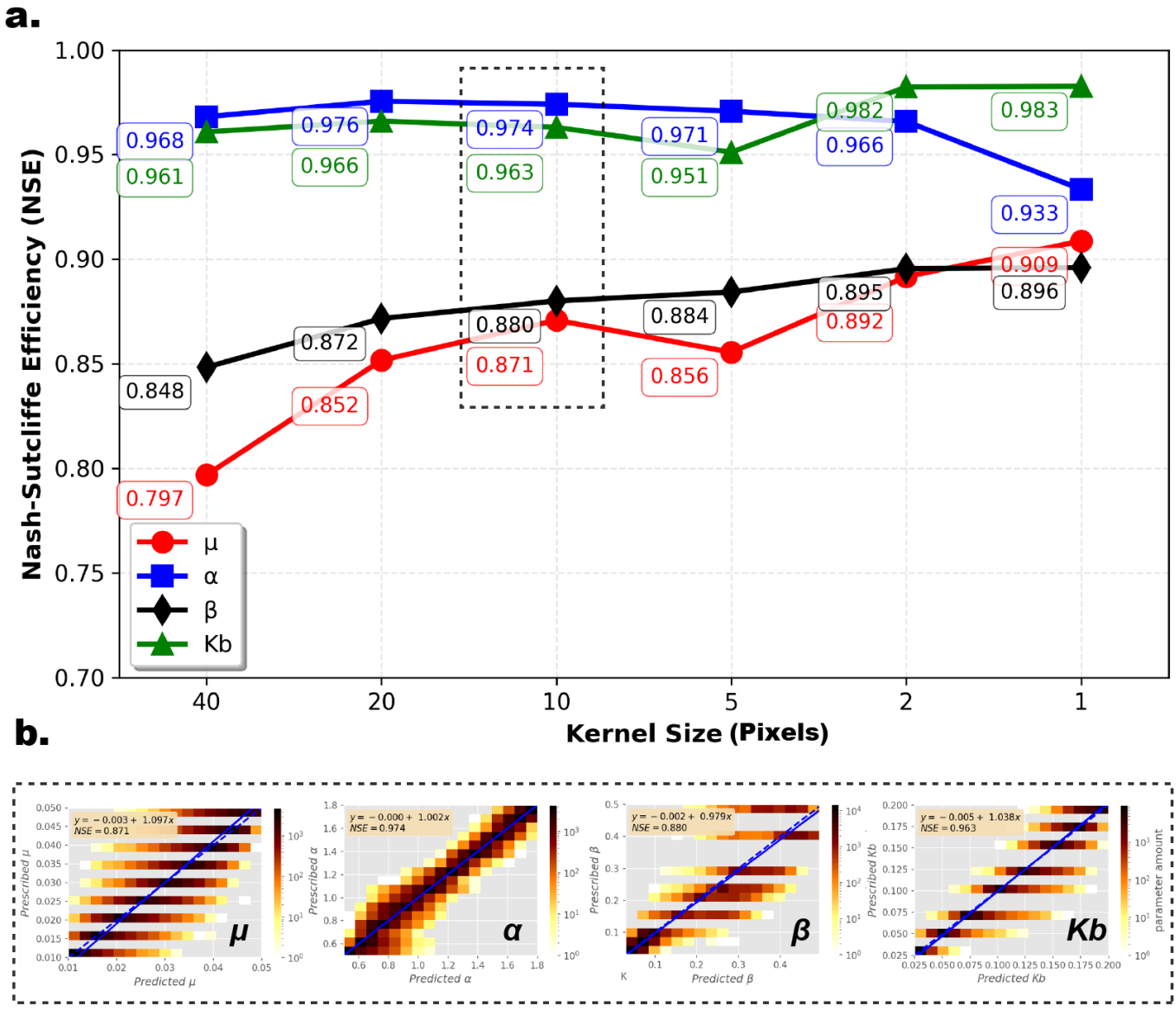


Figure 2. Model predictive performance across spatial aggregation scales. The main panel a. displays the Nash-Sutcliffe Efficiency (NSE) for predicting disturbance parameters using a 10-fold random cross-validation Random Forest model. The x-axis represents the aggregation of kernel size, where a value of 1 corresponds to the original resolution, and coarser scales are to the left. The panel b shows density scatter plots for the optimal aggregation scale (kernel size = 10) at the global level, comparing model predictions (x-axis) to the prescribed parameters (y-axis). The color scale indicates the density of samples, with darker colors representing a higher concentration of points. The consistently high NSE values across scales demonstrate that predictive power is maintained even at coarser resolutions.

https://doi.org/10.5194/essd-2025-670
Preprint. Discussion started: 24 November 2025

2.3 Machine Learning Prediction and Uncertainty Quantification

For this production run, the Random Forest models were trained on the entire synthetic dataset, using all 17 predictive features aggregated to the optimal kernel size of 10, i.e., 100 m x 100 m scale. The final trained models from this specific run were then saved and applied to the globally aggregated satellite biomass statistics to generate the disturbance regime dataset presented in Section 3.

Although the spatial aggregation process has effectively reduced the discrepancy between the simulation training set and EO data, we further used the Dissimilarity Index (DIK) from Meyer and Pebesma (2021) to quantify the extrapolation-related uncertainty of each pixel-level prediction of three disturbance regimes and background mortality parameters. The DIK measures how different a prediction sample is from the training data in the model's feature space. DIK values below 1.0 suggest the landscape is well-represented, whereas values significantly greater than 1.0 serve as a flag for potential extrapolation, indicating that predictions for that landscape are less reliable. The calculation procedure involves three key steps: (1) standardizing each feature individually, (2) pre-calculating a baseline average dissimilarity from the training data, and (3) computing the final DIK for each new prediction. The detailed theoretical formulation and scalable implementation are provided in the Supplementary S3.

3 The Dataset: Global Patterns of Forest Disturbance Regimes

This section presents the two distinct but related data products derived from our modelling framework: a primary Tile-Level Dataset containing predictions for each 25 × 25 km landscape and a derivative 0.25° × 0.25° Gridded Global Dataset for large-scale analysis and visualization. This dual-product approach is motivated by the need to serve two distinct purposes: the tile-level data provides the native, high-resolution detail required for in-depth, local-scale investigation (Fig. 3), while the gridded product is aggregated for large-scale analysis, visualization of broad biogeographic patterns (Fig. 4), and integration with global ecosystem models. We first provide a comprehensive description of the data product, including variables and technical specifications, to facilitate user understanding and application. Subsequently, we detail the comprehensive, multi-faceted evaluation undertaken to assess the dataset's quality, uncertainty, and scientific plausibility.

3.1 Data Product Description

3.1.1 Tile-Level Disturbance Regime Dataset

The Tile-Level Disturbance Regime Dataset is the primary, high-resolution output of our prediction workflow. It provides disturbance regime parameters for each of the individual 25 km x 25 km forested landscape tiles analyzed globally. An example of this tile-level data is visualized in Figure 3.

https://doi.org/10.5194/essd-2025-670
Preprint. Discussion started: 24 November 2025

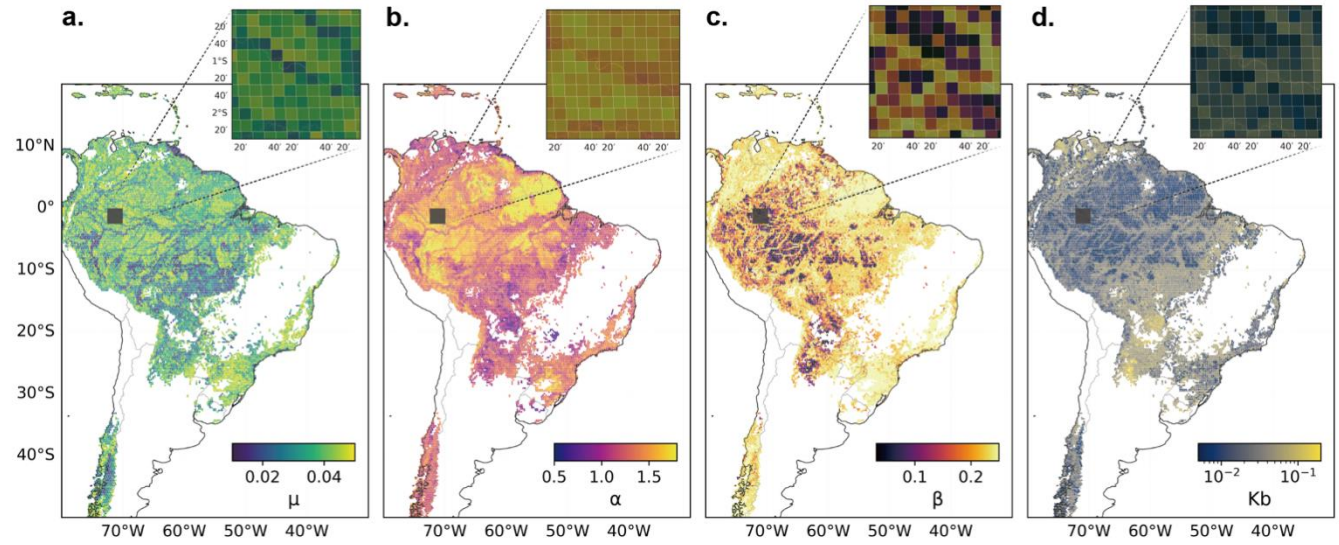


Figure 3. Example of the Tile-Level Dataset for a region in the Amazon basin. The four panels show the spatial distribution of the mean predicted (a) Disturbance Extent (μ), (b) Disturbance Frequency (α), (c) Disturbance Intensity (β), and (d) Background Mortality (Kb). Each colored square represents a single 25 × 25 km landscape tile. The inset in each panel provides a magnified view of a 0.25° x 0.25° area, illustrating how multiple high-resolution tiles are nested within the area of a single grid cell from the gridded global product.

Dataset Variables: The dataset is provided as four separate prediction files, one for each disturbance parameter (μ, α, β, Kb). Each file contains the unique identifier, the raw prediction from each of the 10 randomly cross-validation folds (fold 0 to fold 9), the final ensemble mean prediction, and the corresponding parameter-specific Dissimilarity Index (e.g., DIK_mu). Key variables include:

- Disturbance Extent (μ [%]): the characteristic spatial extent of total annually disturbed area.
- Disturbance Frequency (α [–]): the spatial pattern of disturbance, distinguishing between regimes of many small events versus few large events.
- Disturbance Intensity (β [–]): the severity of disturbance, representing the relative fraction of biomass lost during an event.
- Background Mortality (Kb [$year^{-1}$]): the baseline mortality rate from non-catastrophic processes like natural decay and competition.
- DIK: the dissimilarity of a given landscape's biomass statistics from the training data domain from the specific parameter being predicted, providing a direct measure of model applicability uncertainty for each tile.

Technical Specifications:

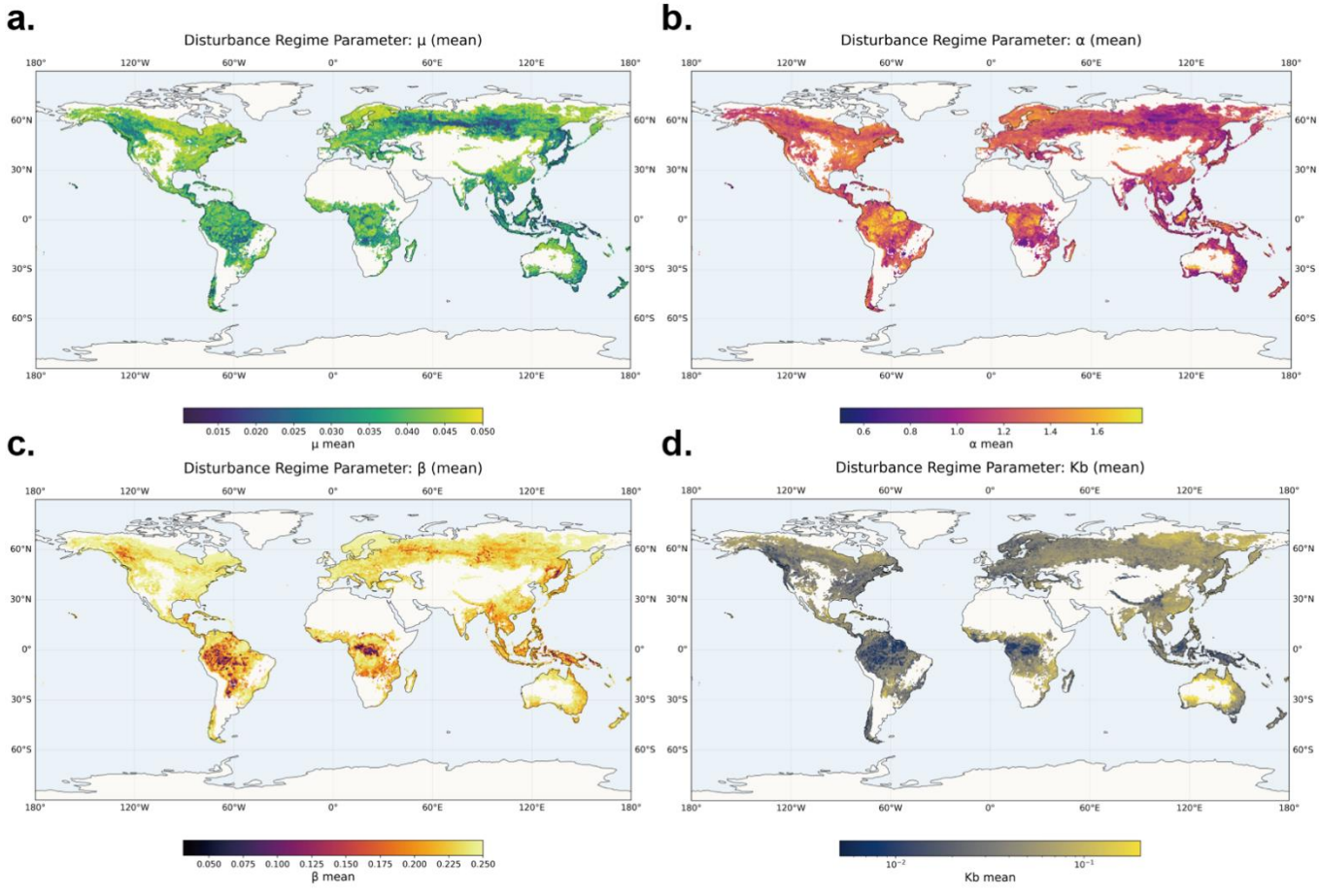
- Spatial Coverage: Global (90° N to 90° S), masked in forested areas.
- Format: Comma-Separated Values (CSV).

https://doi.org/10.5194/essd-2025-670
Preprint. Discussion started: 24 November 2025

- Spatial Representation: The dataset is structured in a vector format, with each record representing a 25 × 25 km landscape tile. For compatibility, each tile location is defined by a bounding box in standard geographic coordinates (WGS84). To ensure analytical accuracy, the spatial statistics for each tile were calculated internally using a local AEQD projection.
- Uncertainty: The dataset provides two measures of uncertainty for each tile. First, the inclusion of all 10-fold predictions allows users to quantify the model of ensemble uncertainty by calculating the variance across predictions. Second, the parameter-specific DIK quantifies the model's applicability uncertainty, indicating how similar the landscape is to the training data.

3.1.2 Gridded Global Disturbance Regime Dataset

This dataset provides continuous global maps of the disturbance regime parameters and multiple associated uncertainty and variability layers, created by aggregating the tile-level data onto a regular 0.25° grid. This product is ideal for large-scale analysis and integration with other global climate and ecosystem models.



https://doi.org/10.5194/essd-2025-670
Preprint. Discussion started: 24 November 2025

Figure 4. Global patterns of predicted disturbance regime parameters. The four panels show the grided global maps of ensemble means for Disturbance Extent (μ), (b) Disturbance Frequency (α), (c) Disturbance Intensity (β), and (d) Background Mortality (Kb). Each parameter is displayed with a distinct color scale to highlight its unique spatial patterns.

Dataset Variables: The dataset is a single NetCDF file containing multiple layers derived from the tile-level data for each of four parameters (μ, α, β, Kb):

- Mean Parameters ({param}_mean): the mean value for each parameter within a grid cell, calculated by averaging the ensemble means of all tiles that overlap with that cell (Fig.4).
- Dissimilarity Index (DIK_{param}_mean): the mean model applicability uncertainty for the grid cell, derived by averaging the DIK values of all overlapping tiles (Fig.4).
- Model Prediction Uncertainty ({param}_std_all_folds): A comprehensive uncertainty metric representing the standard deviation of all individual fold predictions from all tiles within a grid cell. This layer combines both the model's ensemble uncertainty and the sub-grid spatial variability (Fig. 5).
- Sub-grid Spatial variability ({param}_std): The standard deviation of tile-level ensemble means within a grid cell. This metric isolates the spatial heterogeneity of the disturbance regime within the 0.25° cell.
- Tile Count (tile_count): A data density layer indicating the number of landscape tiles used to calculate the value for each grid cell.

Technical Specifications:

- Spatial Resolution: 0.25° × 0.25°
- Spatial Coverage: Global (90° N to 90° S), masked to forested areas.
- Format: NetCDF-4
- Coordinate System: Geographic, WGS84.
- Uncertainty: the dataset provides multiple layers to characterize uncertainty. The DIK_{param}_mean layers quantify model applicability uncertainty. The {param}_std_all_folds layers provide a comprehensive measure of prediction uncertainty, combining model ensemble variance and sub-grid heterogeneity. Additionally, the {param}_std layers isolate the sub-grid spatial variability.

3.2 Data Quality and Uncertainty

To provide a comprehensive assessment of prediction confidence, the dataset includes two distinct uncertainty layers. The first layer reflects the machine learning prediction uncertainty (std_all_folds), quantified as the standard deviation across the different cross-validation folds. As shown in Fig. 5, this model-related uncertainty is low for forested regions, indicating a generally robust prediction. However, for most parameters, particularly β, relatively high uncertainty is consistently concentrated in the humid tropics—most notably the Amazon and Congo basins. This suggests the model has greater difficulty

https://doi.org/10.5194/essd-2025-670
Preprint. Discussion started: 24 November 2025

Open Access Earth System Science Data Discussions

disentangling the disturbance severity signal from biomass patterns in these high-biomass, structurally complex ecosystems. It is important to note that this uncertainty is not simply a function of sampling density (Supplementary Fig. S2 shows the tile count distribution for global grid cells), as the Kb parameter exhibits a different spatial pattern with distinct regional hotspots out of the tropical belt.

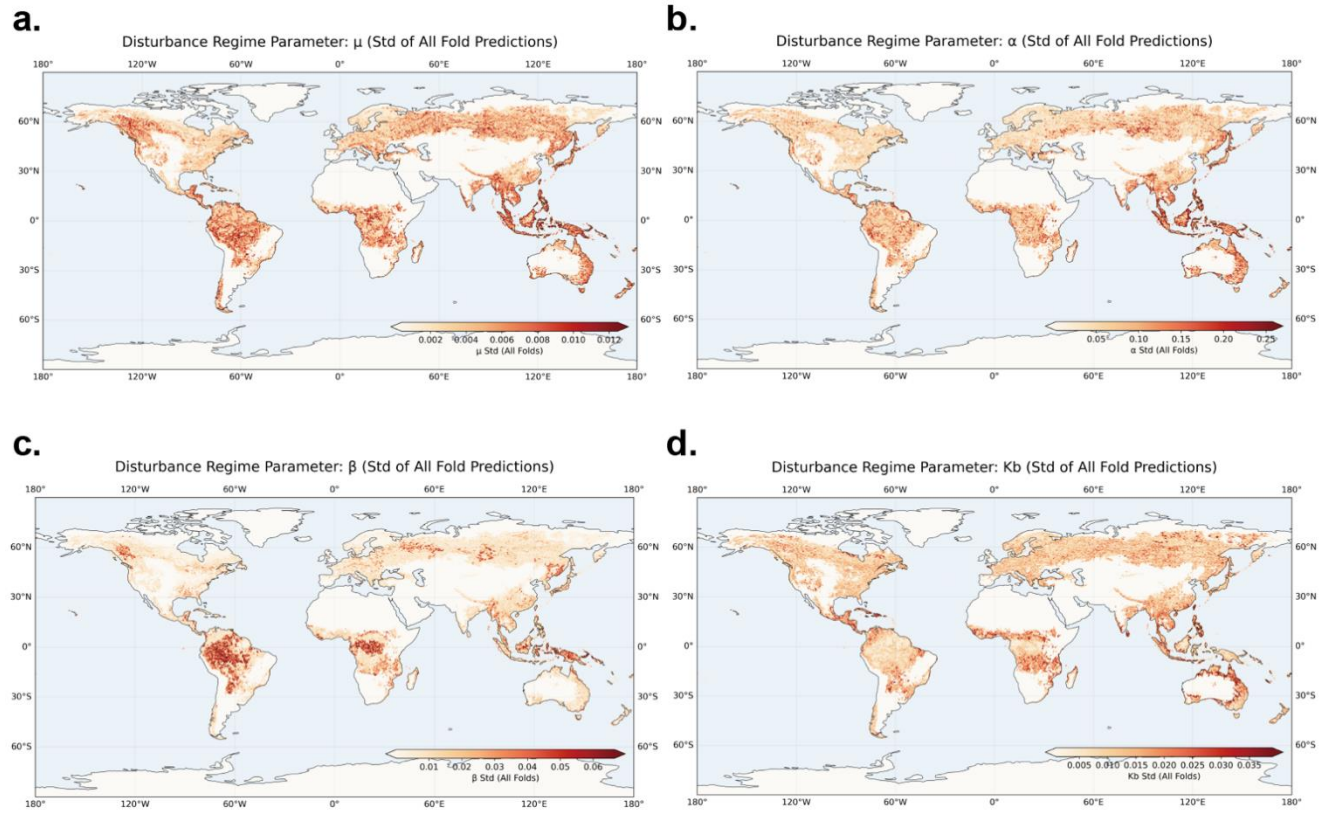


Figure 5. Comprehensive prediction of uncertainty of disturbance regime parameters. The four panels show the global gridded maps of the comprehensive prediction of uncertainty (std_all_folds) for each parameter. This metric represents the standard deviation of all cross-validation fold predictions from all tiles within each 0.25° grid cell, thereby integrating both model ensemble uncertainty and sub-grid spatial heterogeneity. Higher values, indicated by warmer colors on a shared color scale, represent greater overall uncertainty. The number of tiles contributing to each grid cell is provided in the tile_count layer of the dataset (see Supplementary Information).

The second uncertainty layer represents model applicability, quantified by the Dissimilarity Index (DIK; see Supplementary S3). The DIK measures the dissimilarity in the key landscape features of biomass patterns between satellite observations and our synthetic training data. Because the relative importance of these landscape features may differ for each parameter, we generated DIK maps for four disturbance parameters (μ, α, and β) and background mortality separately (Figure 6). The DIK maps for the disturbance parameters (μ, α, and β) are spatially consistent, identifying regions of high confidence (DIK close

https://doi.org/10.5194/essd-2025-670
Preprint. Discussion started: 24 November 2025

to 0, dark blue areas) across major intact forest ecosystems, including the Amazon and Congo Basin rainforests, the Pacific temperate rainforests of North America, the forests of insular Southeast Asia, and large tracts of the Eurasian boreal forest. Conversely, these maps flag areas of potential extrapolation (DIK > 1.0, purple areas) in landscapes where biomass patterns are dissimilar from the disturbance scenarios in our simulation library. These include regions with strong anthropogenic influence such as the United Kingdom and Western Europe, agriculture-dominated ecosystems like India, and regions with low overall forest fraction, such as Western Australia. The DIK for background mortality (Kb) exhibits a strikingly different pattern, indicating high uncertainty and potential extrapolation across nearly the entire boreal forest biome of North America and Eurasia. This high uncertainty aligns with the model's prediction of unexpectedly high Kb values in these regions, which contradicts the long carbon turnover times known to characterize these ecosystems. This suggests that in systems dominated by large, infrequent, stand-replacing disturbances like fires in boreal regions, the subtle spatial signature of background mortality is masked by the strong imprint of the dominant disturbance regime, leading to equifinality and reduced model reliability for this specific parameter.

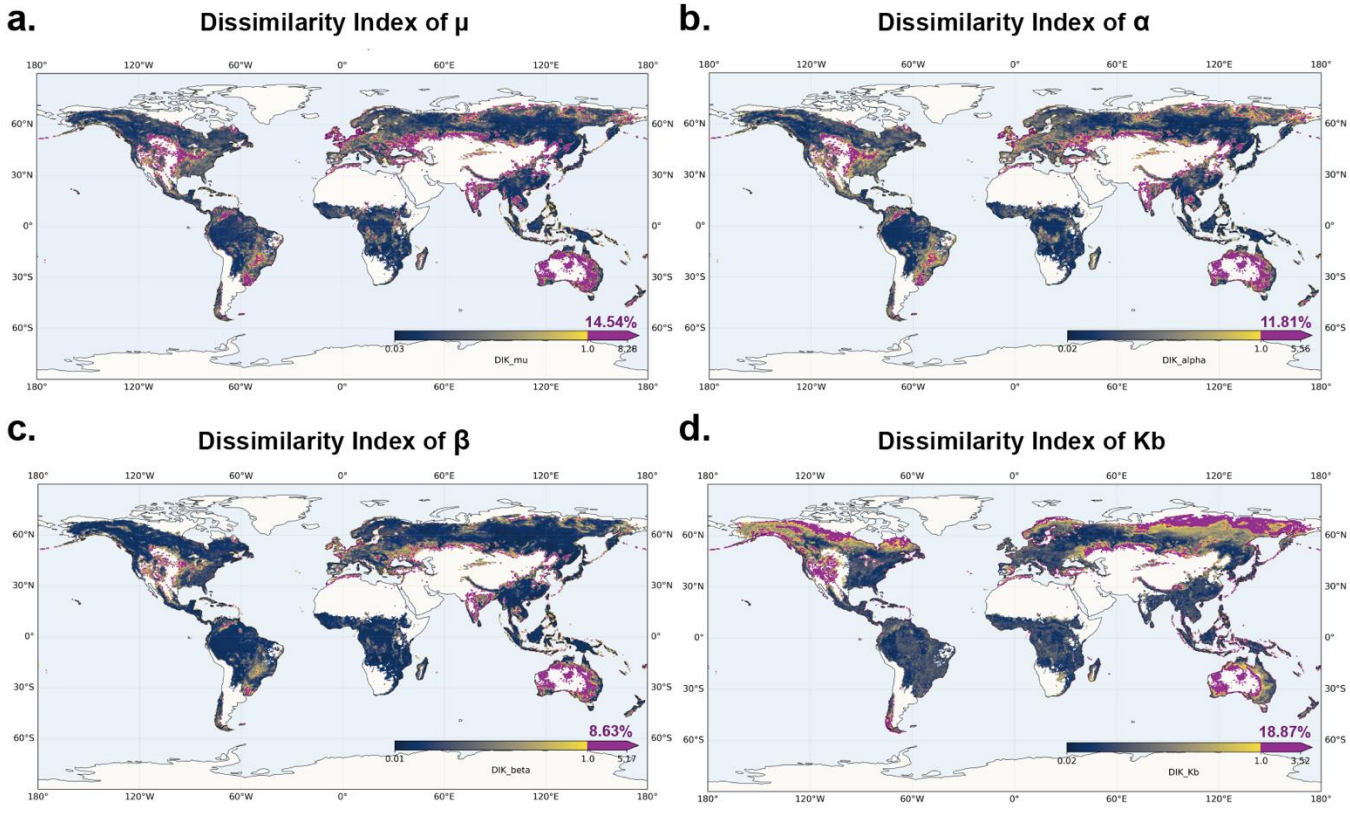


Figure 6. Model applicability uncertainty quantified by Dissimilarity Index. The four panels show the global gridded maps of the mean Dissimilarity Index (DIK) for each parameter. The DIK quantifies the novelty of observed landscapes compared to the training data. Values below 1.0 indicate that observed biomass patterns are well represented within the model's

https://doi.org/10.5194/essd-2025-670
Preprint. Discussion started: 24 November 2025

training domain. Values exceeding this threshold (labelled as purple) serve as a flag for potential extrapolation, indicating that predictions in these areas have higher uncertainty because the model is encountering landscape patterns not seen during training.

330

3.3 Empirical Evaluation

A direct, quantitative validation of our disturbance regimes is inherently challenging, as no existing dataset captures the long-term, landscape-scale characteristics we aim to represent. Therefore, in this section we assessed the scientific plausibility of our product by evaluating the correspondence between predicted disturbance regime parameters and biomass patterns in the 335 extreme high/low scenarios. To isolate the impact of each parameter (μ, α, β), we selected extreme high/low study sites (Figure 7c) that were the most similar in all other variables (DPRs, Kb, GPP), allowing a controlled visual assessment of how each parameter uniquely influences biomass patterns (Figure 7).

The low- and high-μ sites are both located in boreal forests (Figure 7c). The low-μ site exhibits a high-biomass, intact forest 340 canopy with a heavily skewed biomass histogram, confirming its mature, undisturbed state. In contrast, the high-μ site shows more areas with low-biomass patches (Figure 7b, Panel 2), reflected in its left-skewed histogram with lower mean biomass values. This observed pairing demonstrates how a significant difference in μ alone, as other variables are similar, drives a clear divergence in landscape patterns, especially for the overall mean value. This pattern aligns with the conceptual definition (Figure 7a, Panel 1-2) that μ governs the total area affected by disturbance.

345

The pairing for α contrasts a low-α site in the Amazon rainforest with a high α site in Southeast Asia (Figure 7c). The Amazonian site (low α) exhibits a coarse-grained spatial texture with large clustered low biomass area (Figure 7b, Panel 3), where its broad biomass histogram indicates a landscape mix of intact forest blocks and significant clearings, indicative of large-scale, infrequent events. Conversely, the Southeast Asian site (high α) presents a fine-grained, highly fragmented texture 350 (Panel b4). Its biomass patterns are small and intermixed, lacking the large, consolidated clearings of the low-α site and suggesting a regime driven by small-scale, scattered events. This observed contrast highlights how α modulates the spatial aggregation of disturbance. It effectively differentiates regimes dominated by a few large, contiguous events (low α) from those characterized by many small, dispersed events (high α), a finding consistent with the conceptual design (Figure 7a, Panel 3-4).

355

Both the low- and high- β are in the Amazon rainforest, and critically, the biomass histograms are highly similar. Both exhibit a bimodal distribution with no clear difference in mean biomass, which is expected as their μ and α are nearly identical. Despite the similarity in the first dimensional statistics, a stark contrast in spatial pattern between the low- and high- β landscapes, as low- β shows a more transition while high- β has more obvious footprints of disturbance. This pairing powerfully illustrates

https://doi.org/10.5194/essd-2025-670
Preprint. Discussion started: 24 November 2025

the insufficiency of differentiating regimes solely based on first-order statistics from biomass. It underscores the necessity of using spatial-statistical features to capture more sophisticated disturbance scenarios, validating our model's approach.

The global distributions for the three parameters are presented in Figure 7d. Both μ (global mean value of 0.035) and β (global mean value of 0.25) are negatively skewed, characterized by a dominant peak at high values and a long tail extending toward low values. This feature is particularly pronounced for β, which displays two distinct peaks for high values, with the extreme-high peak being the most dominant. In contrast, the parameter α is more central distributed, with the global mean value of 1.25. The colored vertical lines in these histograms confirm that the selected case studies are representative of the extremes of the global distributions.

This analysis demonstrates that the predicted parameters can distinguish between remarkably different disturbance regimes that correspond to visually and ecologically distinct real-world biomass patterns. Furthermore, it highlights the value of using high-dimensional, spatial-statistical information to derive these parameters (example of low- and high- β landscapes). The derived disturbance parameters are suited for implementing stochastic disturbance modules for within-forest perturbations in Dynamic Global Vegetation Models (DGVMs) and Earth System Models (ESMs), which currently rely on more simplistic schemes. This application has the potential to reduce a key uncertainty in carbon cycle projections by providing a globally consistent, observationally constrained representation of natural disturbance regimes.

https://doi.org/10.5194/essd-2025-670
Preprint. Discussion started: 24 November 2025

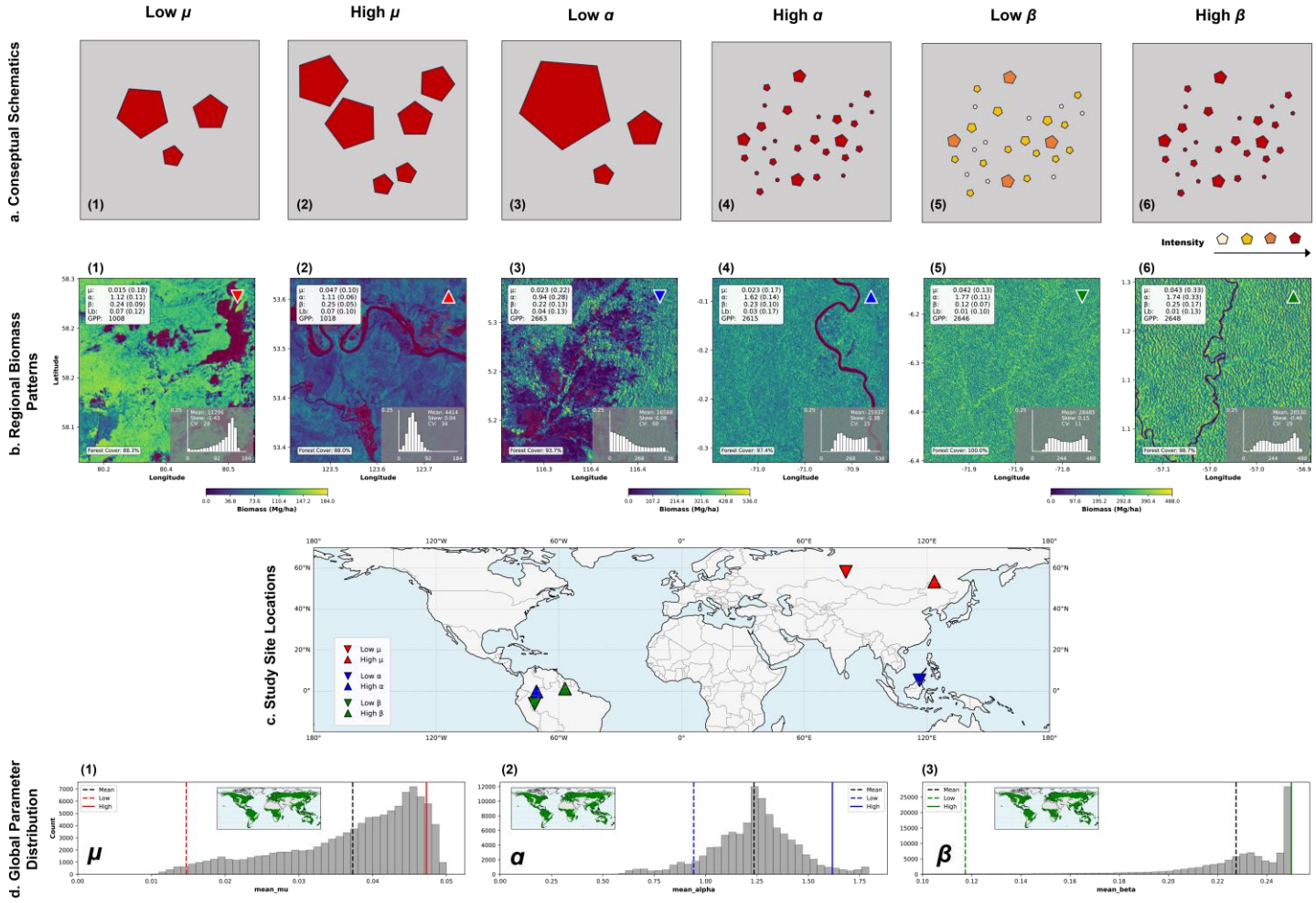


Figure 7. Correspondence between Predicted Disturbance Parameters and Empirical Biomass Patterns. (a) Conceptual Schematics: Illustrates the theoretical landscape-scale patterns for six extreme cases (low vs. high) of the parameters μ, α, and β. (b) Regional Biomass Patterns: Shows corresponding empirical biomass maps from satellite observations. These low/high pairs were selected by filtering landscapes to ensure high similarity in photosynthesis level (GPP), background mortality (Kb) and the other two disturbance parameters, thereby isolating the observable impact of the target parameter on biomass pattern. (c) Study Site Locations: Geographic locations of the six regional case studies shown in panel b. (d) Global Parameter Distributions: Histograms showing the full global distribution of predicted μ, α, and β. The colored vertical lines indicate the values for the selected low and high case studies, placing these examples within their global context.

4 Data Availability

The two data products described in this paper, the Tile-Level Dataset (CSV) and the Gridded Global Dataset (NetCDF), will be publicly available in the Edmond Repository after peer review. The datasets contain the global disturbance regime parameters (μ, α, β, Kb) and the associated uncertainty layers detailed in this manuscript.

https://doi.org/10.5194/essd-2025-670
Preprint. Discussion started: 24 November 2025


Open Access Earth System Science Data Discussions

5 Code Availability

The source code used to process the input data, train the models, and generate the final dataset will be available in Edmond Repository after peer review.

6 Conclusion

This study presents a novel global dataset of forest disturbance regimes, including extent, frequency, intensity, and background mortality, derived from high-resolution satellite-based biomass observations. This process-based product was enabled by a massive synthetic training dataset and high-performance computing. We quantified two key sources of uncertainty for these derived disturbance regime parameters and background mortality: one inherent to the machine learning predictions and another related to model applicability (extrapolation risk). Together, these uncertainties show low uncertainty, i.e., reliable prediction across widespread ~90% forest regions. In addition, an empirical evaluation using paired contrasting landscapes from the distributional extremes confirms that the derived parameters correspond to ecologically distinct real-world biomass patterns, qualitatively confirming the scientific plausibility of inferring disturbance regimes from biomass landscape features. This dataset offers a critical new resource, providing not only a means to implement more realistic, stochastic disturbance modules in Earth System Models, but also a novel pathway to investigate the coupled dynamics of disturbance, vegetation, and the carbon cycle, with the potential to reduce a key uncertainty in future carbon cycle projections.

Author Contributions

S.W. and N.C. conceptualized the study. S.W. curated the data and performed the formal analysis. U.W. processed the datasets. M.R. and N.C. acquired the funding. S.W., H.Y., and N.C. conducted the investigation. The methodology was developed by S.W., H.Y., S.K., M.F., and N.C. M.S. provided and contributed to the GlobBiomass dataset. All authors contributed to the writing of the manuscript.

Competing Interests

The authors declare that they have no conflict of interest.

Acknowledgements

The authors acknowledge the Max Planck Computing and Data Facility (MPCDF) for providing the high-performance computing resources necessary for this study. We thank the FLUXCOM team for providing the GPP data. SW acknowledges support from the International Max Planck Research School for Biogeochemical Cycles (IMPRS-gBGC), Project Office

https://doi.org/10.5194/essd-2025-670
Preprint. Discussion started: 24 November 2025

BIOMASS, Project NextGenCarbon, and the German Federal Ministry of Economics and Technology (50EE1904). Open Access funding was enabled and organized by Projekt DEAL.

Financial Support

This research is supported by the International Max Planck Research School for Biogeochemical Cycles (IMPRS-gBGC), Project Office BIOMASS (EEBIOMASS), and Project NextGenCarbon (grant No.101184989) .

References

Buchhorn, M., Smets, B., Bertels, L., Roo, B. D., Lesiv, M., Tsendbazar, N.-E., Herold, M., and Fritz, S.: Copernicus Global Land Service: Land Cover 100m: collection 3: epoch 2019: Globe (V3.0.1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3939050.

Chambers, J. Q., Negron-Juarez, R. I., Marra, D. M., Vittorio, A. D., Tews, J., Roberts, D., Ribeiro, G. H. P. M., Trumbore, S. E., and Higuchi, N.: The steady-state mosaic of disturbance and succession across an old-growth Central Amazon forest landscape, PNAS, 110, 3949–3954, https://doi.org/10.1073/pnas.1202894110, 2013.

Cohen, W. B., Healey, S. P., Yang, Z., Stehman, S. V., Brewer, C. K., Brooks, E. B., Gorelick, N., Huang, C., Hughes, M. J., Kennedy, R. E., Loveland, T. R., Moisen, G. G., Schroeder, T. A., Vogelmann, J. E., Woodcock, C. E., Yang, L., and Zhu, Z.: How Similar Are Forest Disturbance Maps Derived from Different Landsat Time Series Algorithms?, Forests, 8, 98, https://doi.org/10.3390/f8040098, 2017.

D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith: Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations, Transactions of the ASABE, 50, 885–900, https://doi.org/10.13031/2013.23153, 2007.

Delbart, N., Ciais, P., Chave, J., Viovy, N., Malhi, Y., and Le Toan, T.: Mortality as a key driver of the spatial distribution of aboveground biomass in Amazonian forest: results from a dynamic vegetation model, Biogeosciences, 7, 3027–3039, https://doi.org/10.5194/bg-7-3027-2010, 2010.

Fisher, J. I., Hurtt, G. C., Thomas, R. Q., and Chambers, J. Q.: Clustered disturbances lead to bias in large-scale estimates based on forest sample plots, Ecology Letters, 11, 554–563, https://doi.org/10.1111/j.1461-0248.2008.01169.x, 2008.

Friedlingstein, P., Jones, M. W., O'Sullivan, M., Andrew, R. M., Bakker, D. C. E., Hauck, J., Le Quéré, C., Peters, G. P., Peters, W., Pongratz, J., Sitch, S., Canadell, J. G., Ciais, P., Jackson, R. B., Alin, S. R., Anthoni, P., Bates, N. R., Becker, M., Bellouin, N., Bopp, L., Chau, T. T. T., Chevallier, F., Chini, L. P., Cronin, M., Currie, K. I., Decharme, B., Djeutchouang, L. M., Dou, X., Evans, W., Feely, R. A., Feng, L., Gasser, T., Gilfillan, D., Gkritzalis, T., Grassi, G., Gregor, L., Gruber, N., Gürses, Ö., Harris, I., Houghton, R. A., Hurtt, G. C., Iida, Y., Ilyina, T., Luijkx, I. T., Jain, A., Jones, S. D., Kato, E., Kennedy, D., Klein Goldewijk, K., Knauer, J., Korsbakken, J. I., Körtzinger, A., Landschützer, P., Lauvset, S. K., Lefèvre, N., Lienert, S., Liu, J., Marland, G., McGuire, P. C., Melton, J. R., Munro, D. R., Nabel, J. E. M. S., Nakaoka, S.-I., Niwa, Y., Ono, T., Pierrot, D., Poulter, B., Rehder, G., Resplandy, L., Robertson, E., Rödenbeck, C., Rosan, T. M., Schwinger, J., Schwingshackl, C., Séférian, R., Sutton, A. J., Sweeney, C., Tanhua, T., Tans, P. P., Tian, H., Tilbrook, B., Tubiello, F., van der Werf, G. R., Vuichard, N., Wada, C., Wanninkhof, R., Watson, A. J., Willis, D., Wiltshire, A. J., Yuan, W., Yue, C., Yue, X., Zaehle, S.,

https://doi.org/10.5194/essd-2025-670
Preprint. Discussion started: 24 November 2025

and Zeng, J.: Global Carbon Budget 2021, Earth System Science Data, 14, 1917–2005, https://doi.org/10.5194/essd-14-1917-2022, 2022.

Friend, A. D., Lucht, W., Rademacher, T. T., Keribin, R., Betts, R., Cadule, P., Ciais, P., Clark, D. B., Dankers, R., Falloon, P. D., Ito, A., Kahana, R., Kleidon, A., Lomas, M. R., Nishina, K., Ostberg, S., Pavlick, R., Peylin, P., Schaphoff, S., Vuichard, N., Warszawski, L., Wiltshire, A., and Woodward, F. I.: Carbon residence time dominates uncertainty in terrestrial vegetation responses to future climate and atmospheric CO2, Proceedings of the National Academy of Sciences, 111, 3280–3285, https://doi.org/10.1073/pnas.1222477110, 2014.

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

Kulakowski, D., Seidl, R., Holeksa, J., Kuuluvainen, T., Nagel, T. A., Panayotov, M., Svoboda, M., Thorn, S., Vacchiano, G., Whitlock, C., Wohlgemuth, T., and Bebi, P.: A walk on the wild side: Disturbance dynamics and the conservation and management of European mountain forest ecosystems, Forest Ecology and Management, 388, 120–131, https://doi.org/10.1016/j.foreco.2016.07.037, 2017.

Meyer, H. and Pebesma, E.: Predicting into unknown space? Estimating the area of applicability of spatial prediction models, Methods in Ecology and Evolution, 12, 1620–1633, https://doi.org/10.1111/2041-210X.13650, 2021.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I — A discussion of principles, Journal of Hydrology, 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Nelson, J. A., Walther, S., Gans, F., Kraft, B., Weber, U., Novick, K., Buchmann, N., Migliavacca, M., Wohlfahrt, G., Šigut, L., Ibrom, A., Papale, D., Göckede, M., Duveiller, G., Knohl, A., Hörtnagl, L., Scott, R. L., Zhang, W., Hamdi, Z. M., Reichstein, M., Aranda-Barranco, S., Ardö, J., Op de Beeck, M., Billdesbach, D., Bowling, D., Bracho, R., Brümmer, C., Camps-Valls, G., Chen, S., Cleverly, J. R., Desai, A., Dong, G., El-Madany, T. S., Euskirchen, E. S., Feigenwinter, I., Galvagno, M., Gerosa, G., Gielen, B., Goded, I., Goslee, S., Gough, C. M., Heinesch, B., Ichii, K., Jackowicz-Korczynski, M. A., Klosterhalfen, A., Knox, S., Kobayashi, H., Kohonen, K.-M., Korkiakoski, M., Mammarella, I., Mana, G., Marzuoli, R., Matamala, R., Metzger, S., Montagnani, L., Nicolini, G., O'Halloran, T., Ourcival, J.-M., Peichl, M., Pendall, E., Ruiz Reverter, B., Roland, M., Sabbatini, S., Sachs, T., Schmidt, M., Schwalm, C. R., Shekhar, A., Silberstein, R., Silveira, M. L., Spano, D., Tagesson, T., Tramontana, G., Trotta, C., Turco, F., Vesala, T., Vincke, C., Vitale, D., Vivoni, E. R., Wang, Y., Woodgate, W., Yepez, E. A., Zhang, J., Zona, D., and Jung, M.: X-BASE: the first terrestrial carbon and water flux products from an extended data-driven scaling framework, FLUXCOM-X, EGUsphere, 1–51, https://doi.org/10.5194/egusphere-2024-165, 2024.

Pan, Y., Birdsey, R. A., Phillips, O. L., Houghton, R. A., Fang, J., Kauppi, P. E., Keith, H., Kurz, W. A., Ito, A., Lewis, S. L., Nabuurs, G.-J., Shvidenko, A., Hashimoto, S., Lerink, B., Schepaschenko, D., Castanho, A., and Murdiyarso, D.: The enduring world forest carbon sink, Nature, 631, 563–569, https://doi.org/10.1038/s41586-024-07602-x, 2024.

Quegan, S., Le Toan, T., Chave, J., Dall, J., Exbrayat, J.-F., Minh, D. H. T., Lomas, M., D'Alessandro, M. M., Paillou, P., Papathanassiou, K., Rocca, F., Saatchi, S., Scipal, K., Shugart, H., Smallman, T. L., Soja, M. J., Tebaldini, S., Ulander, L., Villard, L., and Williams, M.: The European Space Agency BIOMASS mission: Measuring forest above-ground biomass from space, Remote Sensing of Environment, 227, 44–60, https://doi.org/10.1016/j.rse.2019.03.032, 2019.

Reichstein, M. and Carvalhais, N.: Aspects of Forest Biomass in the Earth System: Its Role and Major Unknowns, Surv Geophys, 40, 693–707, https://doi.org/10.1007/s10712-019-09551-x, 2019.

https://doi.org/10.5194/essd-2025-670
Preprint. Discussion started: 24 November 2025

Santoro, M., Cartus, O., Carvalhais, N., Rozendaal, D. M. A., Avitabile, V., Araza, A., de Bruin, S., Herold, M., Quegan, S., Rodríguez-Veiga, P., Balzter, H., Carreiras, J., Schepaschenko, D., Korets, M., Shimada, M., Itoh, T., Moreno Martínez, Á., Cavlovic, J., Cazzolla Gatti, R., da Conceição Bispo, P., Dewnath, N., Labrière, N., Liang, J., Lindsell, J., Mitchard, E. T. A., Morel, A., Pacheco Pascagaza, A. M., Ryan, C. M., Slik, F., Vaglio Laurin, G., Verbeeck, H., Wijaya, A., and Willcock, S.: The global forest above-ground biomass pool for 2010 estimated from high-resolution satellite observations, Earth System Science Data, 13, 3927–3950, https://doi.org/10.5194/essd-13-3927-2021, 2021.

Seidl, R., Fernandes, P. M., Fonseca, T. F., Gillet, F., Jönsson, A. M., Merganičová, K., Netherer, S., Arpaci, A., Bontemps, J.-D., Bugmann, H., González-Olabarria, J. R., Lasch, P., Meredieu, C., Moreira, F., Schelhaas, M.-J., and Mohren, F.: Modelling natural disturbances in forest ecosystems: a review, Ecological Modelling, 222, 903–924, https://doi.org/10.1016/j.ecolmodel.2010.09.040, 2011.

Seidl, R., Schelhaas, M.-J., Rammer, W., and Verkerk, P. J.: Increasing forest disturbances in Europe and their impact on carbon storage, Nature Clim Change, 4, 806–810, https://doi.org/10.1038/nclimate2318, 2014.

Senf, C. and Seidl, R.: Mapping the forest disturbance regimes of Europe, Nat Sustain, 4, 63–70, https://doi.org/10.1038/s41893-020-00609-y, 2021a.

Senf, C. and Seidl, R.: Storm and fire disturbances in Europe: Distribution and trends, Global Change Biology, 27, 3605–3619, https://doi.org/10.1111/gcb.15679, 2021b.

Toan, T. L., Chave, J., Dall, J., Papathanassiou, K., Paillou, P., Rechstein, M., Quegan, S., Saatchi, S., Seipel, K., Shugart, H., Tebaldini, S., Ulander, L., and Williams, M.: The Biomass Mission: Objectives and Requirements, in: IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, 3 citations (Crossref) [2023-06-01], 8563–8566, https://doi.org/10.1109/IGARSS.2018.8518491, 2018.

Turner, M. G.: Disturbance and landscape dynamics in a changing world, Ecology, 91, 2833–2849, https://doi.org/10.1890/10-0097.1, 2010.

Turner, M. G. and Seidl, R.: Novel Disturbance Regimes and Ecological Responses, Annual Review of Ecology, Evolution, and Systematics, 54, 63–83, https://doi.org/10.1146/annurev-ecolsys-110421-101120, 2023.

Wang, S., Yang, H., Koirala, S., Forkel, M., Reichstein, M., and Carvalhais, N.: Understanding Disturbance Regimes From Patterns in Modeled Forest Biomass, Journal of Advances in Modeling Earth Systems, 16, e2023MS004099, https://doi.org/10.1029/2023MS004099, 2024.

Williams, M., Hill, T. C., and Ryan, C. M.: Using biomass distributions to determine probability and intensity of tropical forest disturbance, Plant Ecology & Diversity, 6, 87–99, https://doi.org/10.1080/17550874.2012.692404, 2013.

Wohlgemuth, T., Jentsch, A., and Seidl, R. (Eds.): Disturbance Ecology, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-030-98756-5, 2022.