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], Anna Candotti[5], Ulrich Weber[1], Ranit De[1,6], Claire Robin[1], Felix Cremer[1], Matthias Forkel[2], Markus Reichstein[1,7], Nuno Carvalhais[1,7,8]

[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] Faculty of Agricultural, Environmental and Food Sciences, Free University of Bozen-Bolzano, Italy
[6] Department of Geography, Friedrich Schiller University Jena, Germany
[7]Departamento de Ciências e Engenharia do Ambiente, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal
[8]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, size, and severity 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 disturbance rate (μ), gap-size distribution (α), disturbance severity (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 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 continuous, 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 the divergence of observations from the training set. An empirical evaluation comparing paired forest landscapes with contrasting extremes in their predicted disturbance parameters support the methodology and assumptions used to build the dataset. Our global maps

of disturbance regimes provide a novel, process-based tool for investigating the coupled dynamics of disturbance, 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. This leads to poor quantification of the resulting large-scale tree damage as well as mortality, and their effects on the carbon cycle are poorly quantified (Senf and Seidl, 2021a, b), thus, being 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).

In the context of this study, a forest disturbance is explicitly defined as any event, whether natural or anthropogenic, that results in a measurable reduction in aboveground biomass (AGB), thereby leaving a structural spatial signature on the forest landscape. Consequently, forest disturbance regimes describe the long-term spatial and temporal patterns of these biomass-reducing mortality events within a landscape, encompassing varied natural and anthropogenic processes that result in a significant loss of aboveground biomass. Efforts to characterize these regimes from Earth Observation (EO) product have largely followed two distinct approaches. The individual event-detection method applies a continuous change detection algorithm to the 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 (light detection and ranging, viz. LiDAR) data. This method heavily relied on good-quality and continuous time series that were often lacking owing to the limitations of satellite-based observations (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. Therefore, this method was not suitable to derive regime parameters that reflect disturbance dynamics over past decades (Turner, 2010).

In contrast, the second approach infers disturbance regimes from landscape-scale characteristics, building on the ecological assumption that the spatial variation of forest biomass is the long-term integral of primary productivity, background mortality, and the occurrence of episodic disturbance in space (Williams et al., 2013). This approach uses ecosystem models to inversely estimate disturbance parameters (such as disturbance rate, size distribution, and severity) 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 result in similar landscape patterns, reducing the ability of identifying different DRPs. This limitation to identifiability, or equifinality, is more likely to occur when using univariate, or single feature approaches, or when relying on coarse-resolution data (Delbart et al., 2010; Williams et al., 2013). The recent proliferation of globally available, high-resolution satellite biomass products now provides the critical observational foundation to test and apply this pattern-based framework at a global scale (Quegan et al., 2019; Le Toan et al., 2018; Reichstein and Carvalhais, 2019).

The primary objective of this study is to globally map four key forest Disturbance Regime Parameters (DRPs): (1) disturbance rate (μ [% yr^{-1}]; the mean annual fractional area affected), (2) gap-size distribution (α [-]; the scaling governing the spatial clustering of disturbance patch sizes; Fisher et al., 2008), (3) disturbance severity (β [-]; the scaling slope governing the size-dependent fraction of biomass lost; Chambers et al., 2013), and (4) background mortality (K_b [yr^{-1}]; the continuous non-episodic carbon turnover). We achieved this by inverting an adaptive forest disturbance framework, which learnt how to infer long-term DRPs directly from the emergent spatial patterns of satellite biomass data. The implementation of this framework proceeded in three main steps: (1) expanding our previously established synthetic training dataset (Wang et al., 2024) to comprehensively represent realistic disturbance regimes by increasing parameter ranges and shapes of disturbances, (2) identifying the optimal spatial aggregation scale to bridge the scale mismatch between simulations and EO product, explicitly reducing the risk of model extrapolation, i.e., applying the model to spatial feature value ranges not covered by the synthetic training data, and (3) applying a pre-trained machine learning model to derive the global disturbance regime parameters, and evaluating the out of domain-extrapolation-predictions via the Dissimilarity Index (DIK).

In the following sections, we provide further details on this comprehensive framework: Section 2 describes the input datasets and the multi-stage methodology, moving from the forward modeling and observational data processing to the calibration of spatial aggregation scales, machine learning prediction, and uncertainty quantification. Section 3 presents the resulting data products in both their native tile-level and gridded global formats. It also provides a comprehensive evaluation of spatially explicit uncertainty, an assessment of the product’s plausibility, and an in-depth discussion of the limitations and sensitivities introduced by the satellite input data. Ultimately, 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.

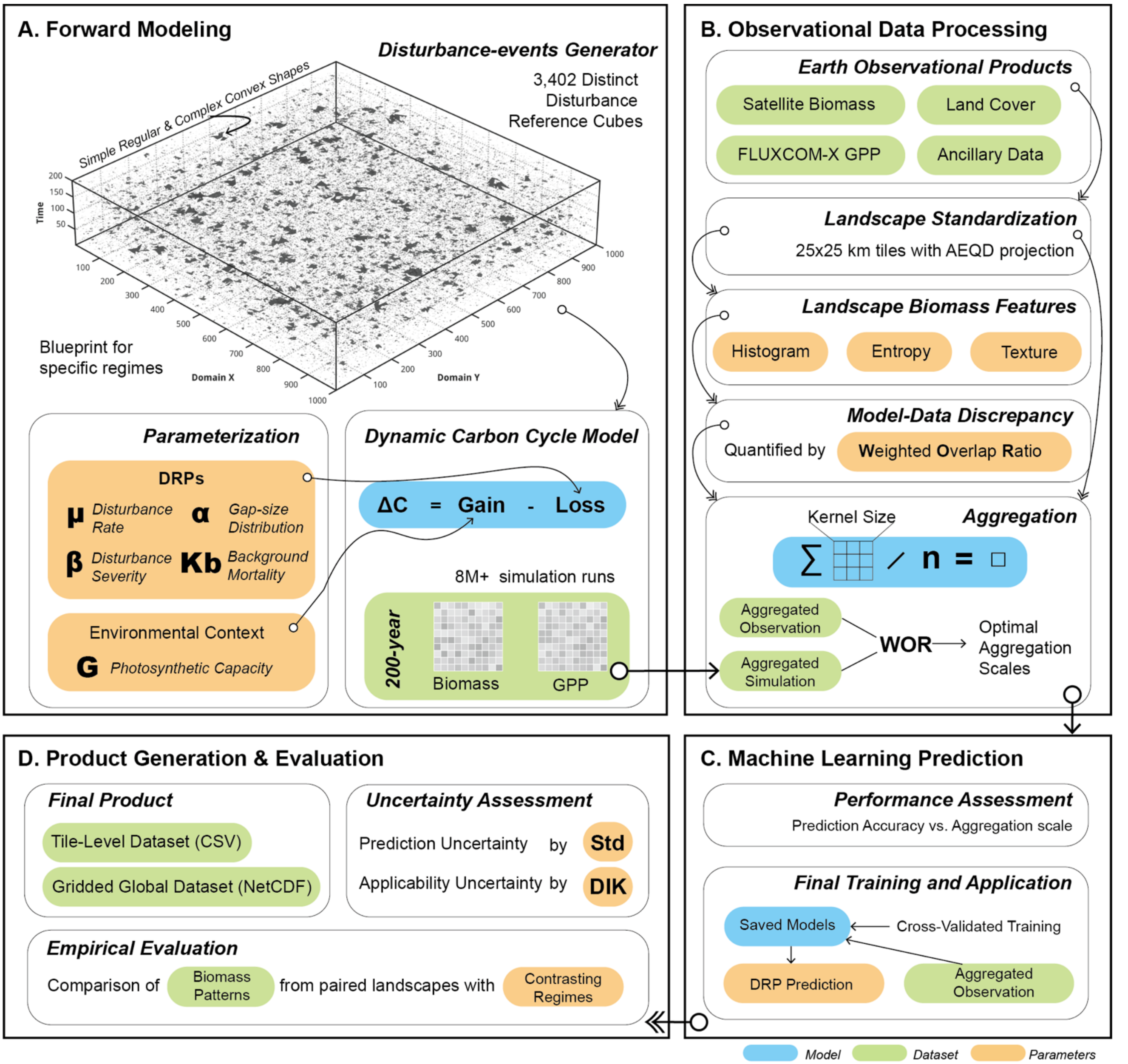

Figure 1. Conceptual workflow of the disturbance regime prediction framework. The framework is organized into four stages. (A) Forward Modeling: A synthetic dataset is created by simulating how known disturbance regimes (μ, α, β, K_b) produce unique biomass patterns. (B) Observational Data Processing & Discrepancy Analysis: Global satellite biomass product 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 tile-level and gridded global datasets are produced and then evaluated for uncertainty and plausibility.

2 Data and Methods

Our framework aims to derive long-term disturbance regimes from a single, static realization of satellite forest biomass. This approach is grounded on the concept of gap dynamics, i.e., at a sufficiently large landscape scale (e.g., 25 × 25 km in our case), a forest ecosystem is continuously shaped by the interplay between carbon loss (disturbance events and background mortality) and carbon gain (gross primary productivity). Consequently, the landscape becomes a mosaic of individual patches in various stages of successional recovery (Watt, 1947; Bray, 1956; Bormann and Likens, 1979; Shugart, 1984; Bonan, 2015). Therefore, the spatial heterogeneity across the landscape acts as a substitute for the temporal changes, capturing the region's history of disturbance and regrowth.

To quantitatively link these spatial biomass patterns to their underlying disturbance regimes, we utilized a forward-modeling approach (Wang et al., 2024), that generates the AGB patterns resulting from the modelled dynamics of forest carbon coupled to a stochastic disturbance-events generator. We used an inverse modeling approach, based on a Random Forest that determines from the emerging AGB patterns and distributions. Inferring disturbance parameters from static biomass maps is challenging because different disturbance combinations produce similar mean biomass outcomes (Williams et al., 2013). Our framework overcomes this by using second-order spatial texture features rather than solely first-order statistics. For this study, we significantly advanced our previous synthetic training library. By broadening the parameterization ranges and incorporating a diversity of non-rectangular disturbance shapes, we scaled the number of simulated regime parameter combinations from 0.85 million to over 8 million. Therefore, we generated a massive training dataset that uniquely links specific disturbance regimes to their resulting spatial biomass patterns by forward-simulation of millions of forest landscapes under diverse disturbance regime parameter combinations (Section S1 in the supplement).

To distinguish among disturbance regimes, the framework characterizes each observed satellite biomass map using first-order distribution statistics, Shannon entropy, and second-order texture metrics derived from a gray-level co-occurrence matrix (GLCM). These features are complementary in characterizing the distribution and the spatial patterns in the biomass maps. For instance, high GLCM homogeneity corresponds to large, even-aged stands created by infrequent, large-scale, stand-replacing disturbances. Conversely, high spatial variance or GLCM contrast may be consistent with a multi-aged, structurally complex forest shaped by frequent, small-scale gap dynamics. The pre-trained model subsequently integrates these observational spatial features with realistic GPP products, which constrain the post-disturbance growth and recovery rates, enabling a direct inversion of the underlying DRPs.

2.1 Observational Datasets and Post-Processing

This analysis used three primary observational datasets to derive predictive features: (1) the ESA GlobBiomass product (Santoro et al., 2021), alongside subsequent ESA CCI Biomass products (Santoro and Cartus, 2026), to characterize fine-scale biomass spatial patterns; (2) the European Space Agency Climate Change Initiative (ESA CCI) Land Cover dataset (ESA, 2017), to provide a temporally dynamic and consistent 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.

2.1.1 Biomass observation

We selected the ESA GlobBiomass product dataset (Santoro et al., 2021), as it provides a globally continuous, spatially explicit map of above-ground biomass for the year 2010 at a native resolution of approximately 25 m at the equator. The product 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. The native 25 m map served as the input to our processing workflow; however, the spatial features used for the final inversion were calculated only after 10 × 10 block-mean aggregation to an effective resolution of approximately 250 m, as described in Section 2.2. This aggregation reduces the influence of pixel-scale retrieval noise and aligns the feature distributions of the observed and simulated biomass maps.

To transform the raw GlobBiomass map into a set of predictive features, we first established a global grid of non-overlapping, true-to-area 25 × 25 km tiles, each representing a landscape-scale analysis unit. The tiles were generated using geodetic calculations on the WGS84 ellipsoid, ensuring that each tile represents a consistent surface area regardless of latitude. For each tile, the corresponding 25 m resolution AGB data were 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: (1) the data was first resampled to a 1 m resolution using cubic interpolation and (2) subsequently aggregated via pixel averaging to a final, standardized 1000 × 1000 pixel grid at a 25 m resolution. A global forest cover mask was then applied to this standardized grid, assigning non-forest pixels with a NaN value to isolate only forested areas for the 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. Following Wang et al. (2024), we used this masked AGB grid to calculate a comprehensive suite of biomass spatial statistics (Table 1), including first-order distribution metrics, Shannon entropy, and second-order texture metrics derived from GLCM. Detailed explanations for each of these features can be found in Table S3 in the supplement. To avoid texture artifacts at forest-mask boundaries, the GLCM was constructed exclusively from adjacent pairs of valid forest pixels in four directions. Pixel pairs involving non-forest or missing pixels were excluded. Instead of calculating across a standard rectangular grid, our method adjusts to the irregular boundaries of the actual forest cover. This guaranteed that we strictly measure the structure within continuous forest patches, avoiding the artificial edges and false patterns that typically occur when the algorithm paired forest pixels with non-forest gaps. For the final feature extraction at the selected aggregation scale, where each 25 × 25 km tile comprised a 100 × 100 pixel grid, GLCM texture features were computed only for tiles containing at least 100 valid forest pixels, corresponding to 1% of the aggregated grid. The resulting vector of 17 statistical features for each landscape formed the quantitative basis for the machine learning models used to predict disturbance regimes.

To evaluate framework robustness across different biomass products, we extended the biomass feature extraction pipeline to two alternative datasets derived from the 100 m native resolution ESA Climate Change Initiative (CCI) Biomass product (Santoro and Cartus, 2026): (1) the static CCI v7 baseline map representing the year 2010, and (2) a temporal multi-year mean constructed from available v7 epochs around this baseline period (spanning 2005 to 2011). To ensure exact geometric alignment with our established landscape tiles, the native 100 m AGB data within each 25 km × 25 km landscape was reprojected and aggregated via pixel averaging into a uniform 100 × 100 pixel grid (yielding an effective spatial resolution of ~250 m). This standardized resampling protocol improves the comparability of the derived spatial distribution and GLCM texture statistics across EO biomass products with different native resolutions and temporal configurations.

Table 1. Summary of the 16 spatial biomass statistical features

Feature Type	Specific Feature Names
First-order Histogram Features	Mean, Median, Variance, Standard Deviation, Coefficient of Variation, Skewness, Kurtosis, 25th Percentile, 75th Percentile, Range, Trimean
Informative Feature	Shannon Entropy
Second-order Texture Features	GLCM Contrast, Correlation, Energy, Homogeneity

2.1.2 Forest cover mask

The forest cover mask used in this study was derived from the European Space Agency Climate Change Initiative (ESA CCI) Land Cover dataset (1992–2015), which categorizes land surfaces based on the UN-FAO Land Cover Classification System (LCCS). The land cover layers were reprojected and resampled to match the spatial extent and resolution of each biomass landscape tile using nearest-neighbor interpolation. We then identified pixels belonging to tree cover categories, specifically, LCCS codes 50–90, 160, and 170, which encompass evergreen needleleaf, evergreen broadleaf, deciduous needleleaf, deciduous broadleaf, mixed, and flooded forests. The identified pixels were used to generate a binary forest mask, while excluding missing data. Crucially, this binary mask served two fundamental purposes in our framework: first, to filter out entirely non-forest landscapes (retaining only landscape tiles with a forest cover fraction greater than 0); and second, to strictly isolate forest-covered pixels within those retained landscapes, ensuring that the extraction of spatial biomass statistics and the subsequent prediction of disturbance regimes are computed exclusively using actual forest biomass values.

To align with our modeling objectives, we implemented a specific temporal matching strategy for the masks to correspond with the timeline of each input biomass dataset. For the static single-year baseline biomass analysis, we consistently applied the fixed 2010 land cover layer to match the 2010 observation snapshot. In contrast, for the 2005–2011 annual biomass time series, we dynamically aligned the land cover mask with the corresponding year of the biomass observations (e.g., pairing the 2005 biomass data with the 2005 land cover layer) to accurately account for concurrent forest transitions.

2.1.3 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 photosynthetic capacity. This data-driven product is derived using in-situ eddy covariance measurements and provides a reliable representation of spatial variation in GPP. To mitigate the uncertainties introduced by interannual climate variability, we calculated a multi-year mean of GPP spanning the decade from 2001 to 2010, rather than relying on a single year. Monthly GPP values were first aggregated into annual sums and then averaged across the 10-year period for each grid cell. For each forested landscape tile, the multi-year mean GPP of all grid cells within the bounding box was spatially averaged. This yielded a single value per landscape, representing its long-term integrated annual photosynthetic capacity (gC m^{-2} yr^{-1}), which was subsequently used as a key predictor in the machine learning models to predict the four disturbance regime parameters. In order to quantify how much a GPP estimate based on a single year could change the spatial patterns of the different DRPs, a comparison between the DRPs from the 2001-2010 mean is contrasted with a prediction based on the GPP of 2010 (Figure S9).

2.2 Calibration of Spatial Aggregation Scale for Model-Data Consistency

Prior to applying statistical features derived from EO biomass product 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 (Table S3 in the supplement). This step was carried out 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 product (Section S2.1 in the supplement). This discrepancy was particularly evident for GLCM Correlation, which was in the range of [0 – 0.75], and [0.75 – 0.98] for simulation and EO product, respectively (Figure S1). Moreover, expanding the parameter space and incorporating non-rectangular disturbance shapes did not significantly reduce this divergence (Figure S2). One other possible reason for the discrepancy in the representation of spatial patterns seats on the fact that the simulation framework treats each grid cell as an independent unit, whereas the EO AGB product exhibits spatial autocorrelation that may reflect both the actual spatial organization of forest biomass, associated with factors such as topography, soil, hydrology, community competition, and the spatially contagious spread of disturbances, and the spatial dependence introduced by the AGB retrieval and processing chain, including spatial filtering and resampling (Santoro and

Cartus, 2026). As such, spatial aggregation likely mitigates the simplicity of the approach used here for generating disturbance events, which may fall short in generating realistic spatial patterns by not resolving mechanistically disturbance spread.

To overcome this discrepancy, we implemented a spatial aggregation procedure for both simulated and observed data to adjust the value ranges of selected features, including the first-order Coefficient of Variation and the second-order GLCM Correlation. 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 smoothed the data and fundamentally altered 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. Figure S4 shows that the divergence between important features derived from simulation and EO datasets reduced substantially at coarser scales, (kernel size ranging from 5 to 40). We quantified this divergence as weighted overlap ratio (WOR; see details in Section S2.2 in the supplement).

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 a synthetic dataset aggregated to different scales using an identical cross-validation scheme with consistent train-test split. 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 (Figure 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 S2.3).

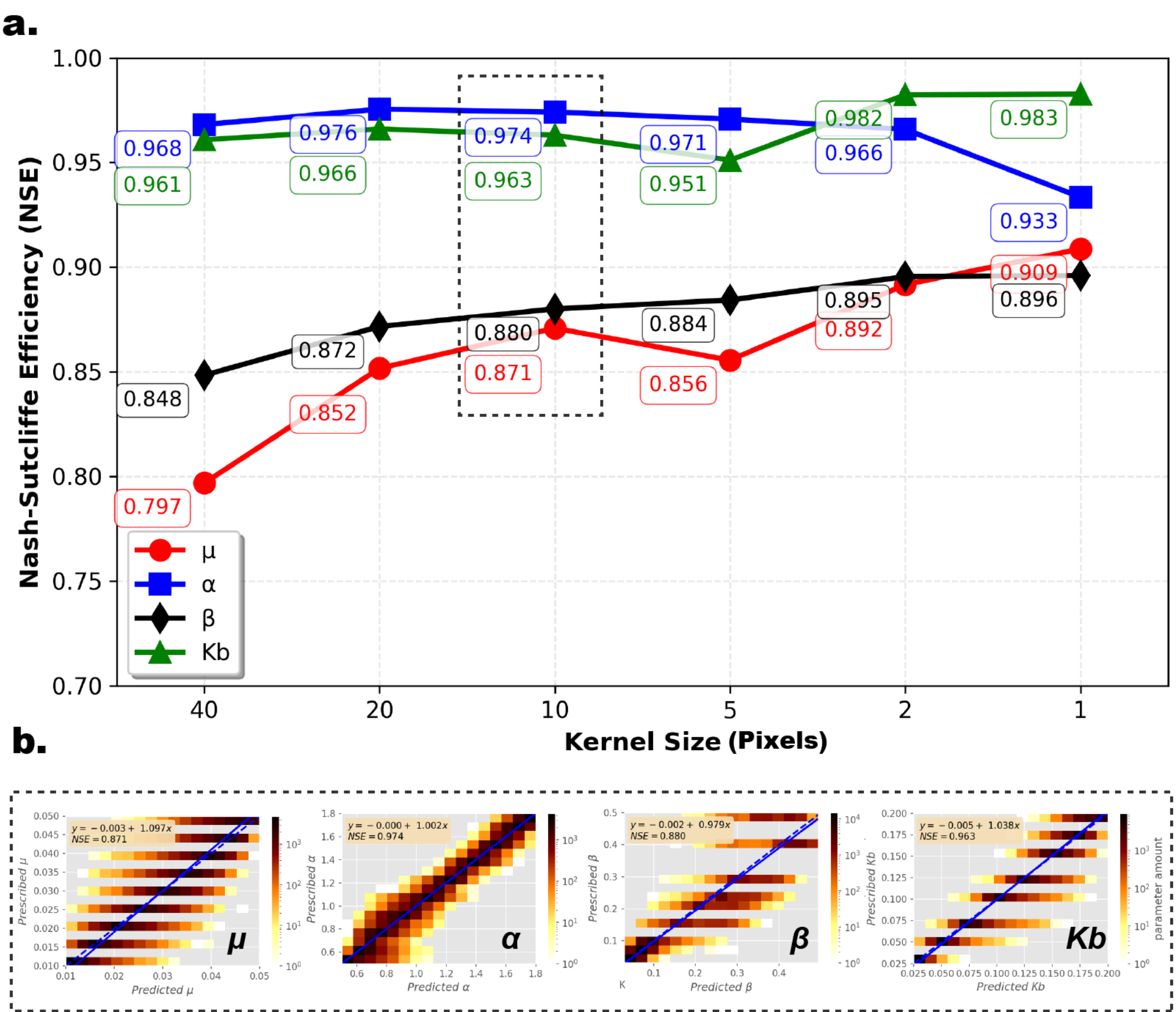

Figure 2. Model predictive performance across spatial aggregation scales. The main panel a. displays the Nash-Sutcliffe Efficiency (NSE) for predicting disturbance parameters using a 10-fold random cross-validation Random Forest model. The x-axis represents the aggregation of kernel size, where a value 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.

2.3 Machine Learning Prediction and Uncertainty Quantification

The inversion of forest disturbance regimes was formulated as a supervised regression task using the Random Forest algorithm in Python (scikit-learn). As explanatory variables, the models utilized 17 spatial-statistical metrics extracted from biomass maps (including first-order statistics, Shannon entropy, and GLCM textures) alongside FLUXCOM-X Gross Primary Production (GPP). The target variables were the four DRPs. For this production run, the models were trained on the entire synthetic dataset using all 17 features aggregated to the optimal 100 m x 100 m scale (kernel size of 10). We employed a standard ensemble configuration comprising 100 decision trees grown without depth restrictions and requiring a minimum of two samples to split an internal node. A fixed random seed was applied to guarantee reproducibility, and the training leveraged parallel processing across multi-CPU nodes on a high-performance computing (HPC) cluster. Model generalizability was robustly evaluated by performing a 10-fold cross-validation framework. 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 training set based on simulation and EO product, 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 because the observed spatial patterns fall outside the theoretical training domain. This analysis involved 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 Section S3 in the supplement.

3 The Dataset: Global Patterns of Forest Disturbance Regimes

This section presents the two distinct, but related data products derived from our modelling framework: (1) a primary Tile-Level Dataset containing predictions for each 25 × 25 km landscape and (2) a derivative 0.25° × 0.25° Gridded Global Dataset for large-scale analysis and visualization. This dual-product approach was motivated by the need to serve two distinct purposes. The tile-level data provided the native, high-resolution detail required for in-depth, local-scale investigation (Fig. 3), while the gridded product was 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 plausibility.

3.1 Data Product Description

3.1.1 Tile-Level Disturbance Regime Dataset

The Tile-Level Disturbance Regime Dataset was 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 (Figure 3).

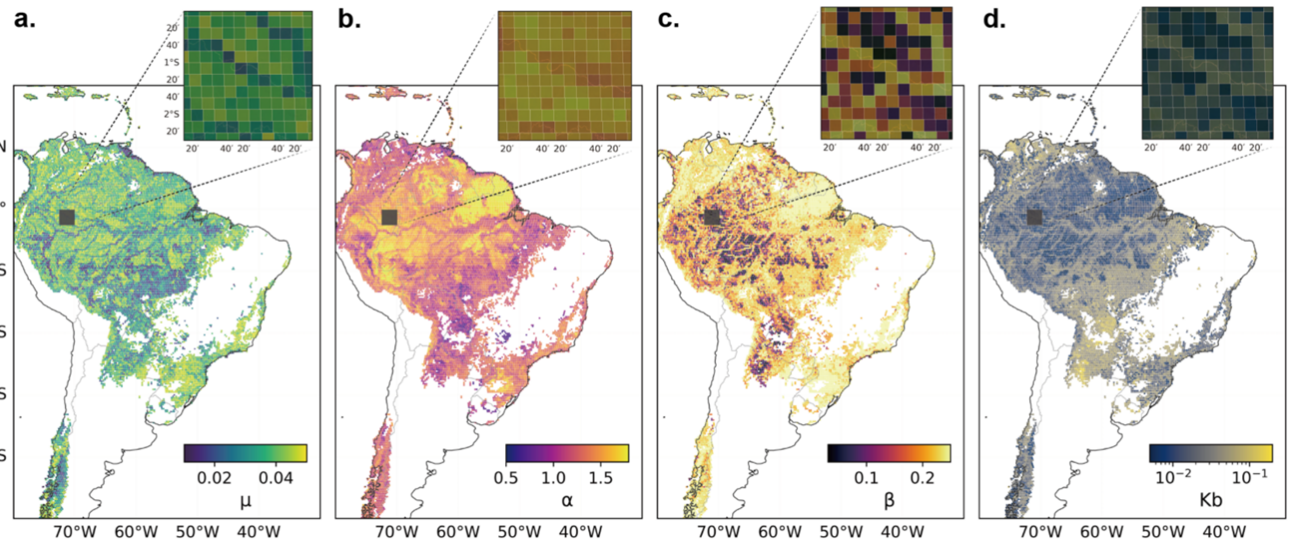

Figure 3. Example of the Tile-Level Dataset for a region in the Amazon basin. The four panels show the spatial distribution of the mean predicted (a) Disturbance Rate μ, (b) Gap-size Distribution α, (c) Disturbance Severity β, and (d) Background Mortality K_b. Each colored square represents a single 25 × 25 km landscape tile. The inset in each panel provides a magnified view of a 2.5° × 2.5° region, illustrating the spatial arrangement of multiple 25 × 25 km tiles.

Dataset Variables: The dataset is provided as four separate prediction files, one for each of the disturbance parameters (μ, α, β, K_b). 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 in the dataset include:

- Disturbance Rate (μ [% yr^{-1}]): the mean annual fractional area affected, distinguishing between highly disturbed and stable regimes.
- Gap-size Distribution (α [−]): the scaling exponent governing the spatial clustering of disturbance patch sizes, distinguishing between regimes of many small events versus few large events.
- Disturbance Severity (β [−]): the scaling slope governing the size-dependent fraction of biomass lost, distinguishing between regimes of low-severity (e.g., partial canopy thinning) and high-severity (e.g., stand-replacing).
- Background Mortality (K_b [yr^{-1}]): the baseline mortality rate from non-episodic processes of carbon turnover, such as natural decay and competition.

- Dissimilarity Index (DIK_{param}): the dissimilarity (for a given landscape) in biomass statistics from the training data domain for a specific parameter, which was being predicted. This provides a direct measure of model applicability uncertainty for each landscape tile.

Technical Specifications:

- Spatial Coverage: Global (90° N to 90° S), masked in forested areas.
- Format: Comma-Separated Values (CSV).
- 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 was defined by a bounding box in standard geographic coordinates (WGS84). To ensure analytical accuracy, the spatial statistics for each landscape tile were calculated internally using a local AEQD projection.
- Uncertainty: The dataset provides two measures of uncertainty for each landscape tile. First, the inclusion of all 10-fold predictions allows users to quantify the uncertainty of model ensemble by calculating the variance across predictions. Second, the parameter-specific DIK quantifies the models' 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 re-gridding the tile-level data onto a regular 0.25° grid. This was due to the fact that the 25 km × 25 km landscape tiles maintain a constant true surface area, while the physical footprint of the 0.25° grid cells varies by latitude, this spatial aggregation calculated the area-weighted mean value of all independent landscape tiles that overlap with a specific 0.25° grid cell. This product is ideal for large-scale analysis and integration with other global climate and ecosystem models.

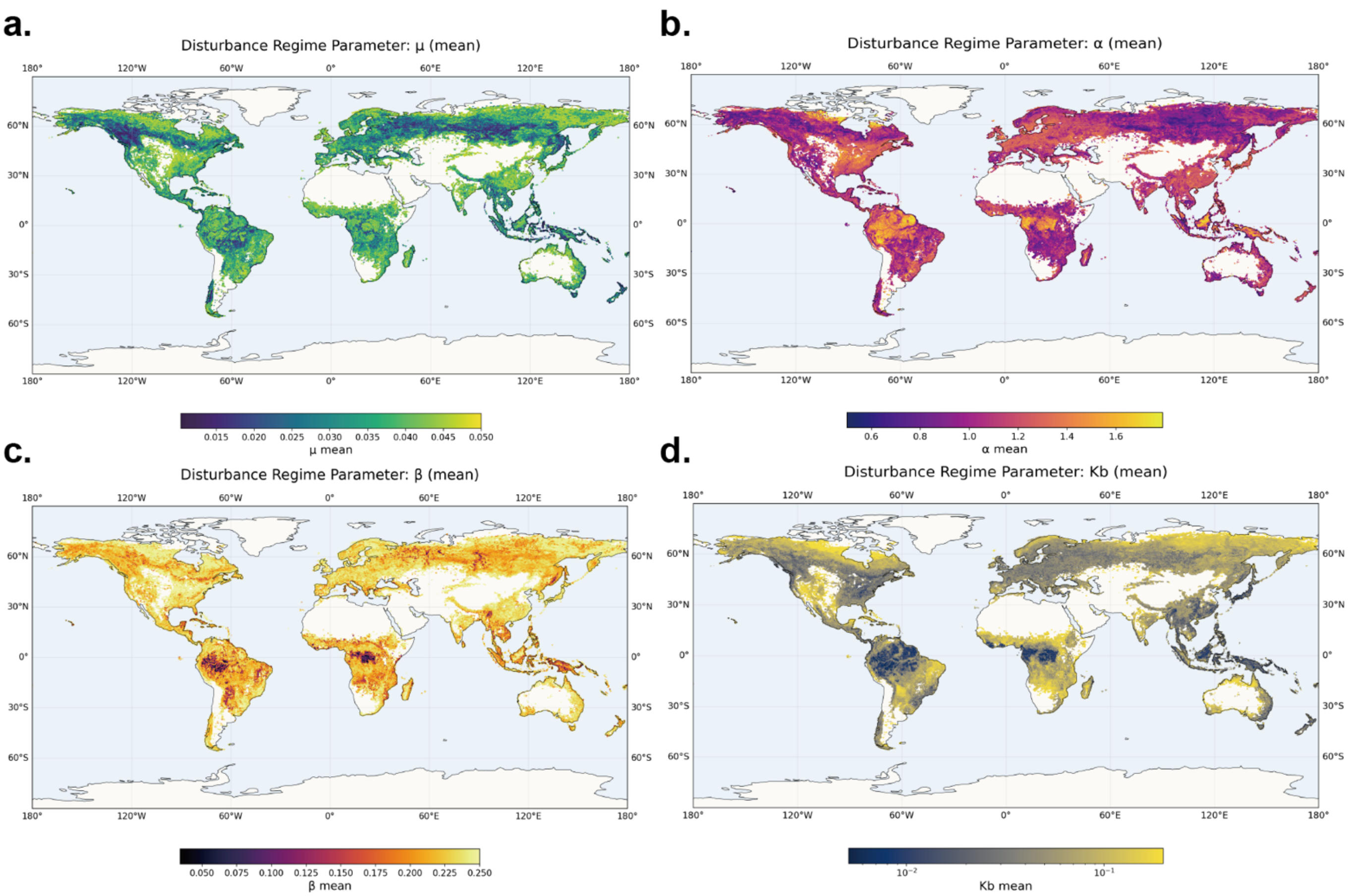

Figure 4. Global patterns of predicted disturbance regime parameters. The four panels show the grided global maps of ensemble means for (a) μ, (b) α, (c) β, and (d) K_b. Each parameter is displayed with a distinct color scale to highlight its unique spatial patterns.

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, K_b):

- Mean Parameters ({param}_mean): the mean value for each parameter within a grid cell, calculated by averaging the ensemble means of all landscape tiles that overlap with that cell (Figure 4).
- Dissimilarity Index (DIK_{param}_mean): the mean model applicability uncertainty for the grid cell, derived by averaging the DIK values of all overlapping landscape tiles (Figure 4).
- Model Prediction Uncertainty ({param}_std_all_folds): A comprehensive uncertainty metric representing the standard deviation of all individual fold predictions from all landscape tiles within a grid cell. This layer combines both the models' ensemble uncertainty and the sub-grid spatial variability (Figure 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), only covers forested areas as per the mask described in Section 2.1.2.
- 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 Model Robustness and Prediction 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. The overall model-related uncertainty for all parameters was relatively small compared to their respective parameter ranges, indicating generally robust predictions globally (Figure 5). However, regions with relatively higher uncertainty exhibit distinct regional hotspot characteristics. This was particularly pronounced for β, with higher uncertainty consistently concentrated in the humid tropics (e.g., the Amazon and Congo basins). This suggests the model had greater difficulty disentangling the disturbance severity signal from biomass patterns in these high-biomass, structurally complex ecosystems, which was also linked to the physical saturation of satellite radar signals, a limitation that is explicitly evaluated in Section 3.4.3. It is important to note that this uncertainty is simply not a function of sampling density (Figure S5 shows the landscape tile count distribution for global grid cells), as the K_b parameter uncertainty exhibits a different spatial pattern with distinct regional hotspots out of the tropical belt.

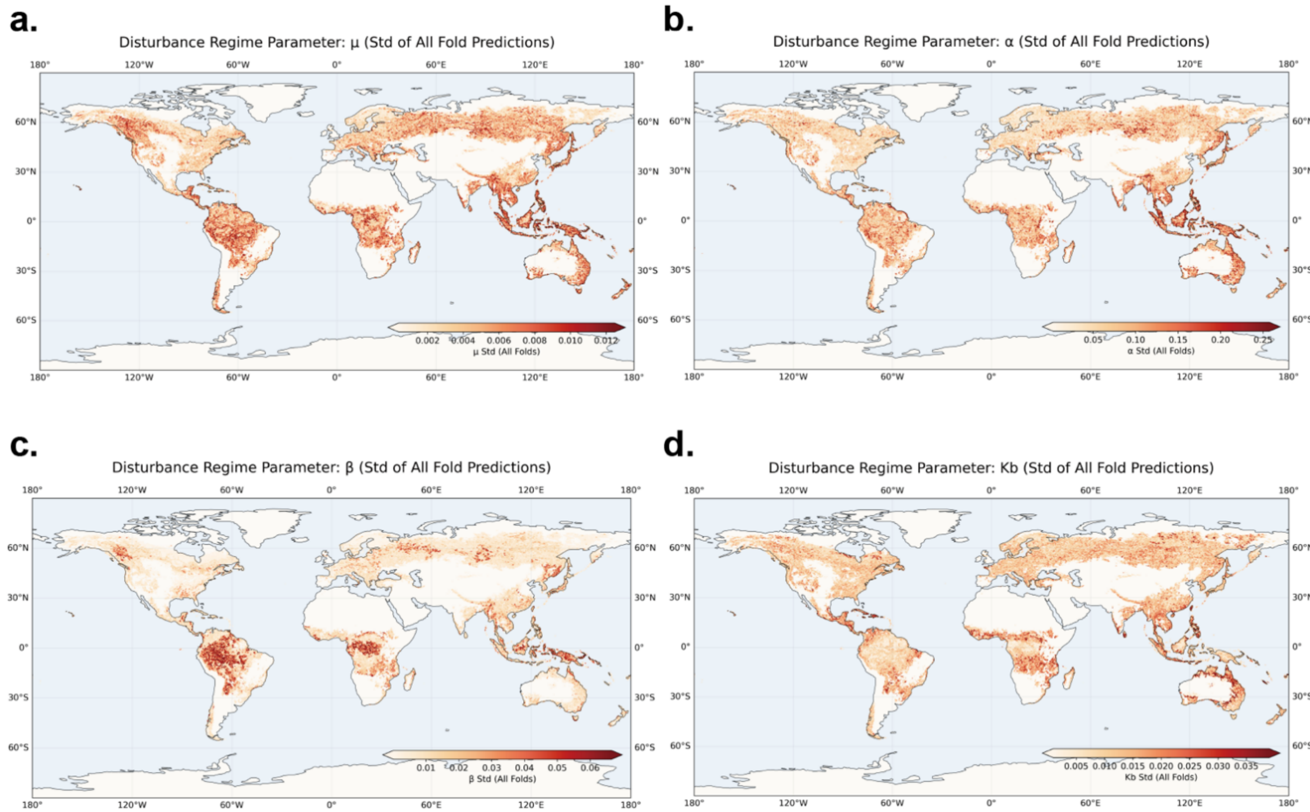

Figure 5. Comprehensive prediction of uncertainty of disturbance regime parameters. The four panels show the global gridded maps of the comprehensive prediction of uncertainty (std_all_folds) for each parameter. This metric represents the standard deviation of all cross-validation fold predictions from all landscape 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 landscape 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 Section S3 in the supplement). 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 disturbance parameters (μ, α, and β) and background mortality separately (Figure 6). The DIK maps for three of the disturbance parameters (μ, α, and β) were spatially consistent, identifying regions of high confidence (DIK close to 0, green areas in Figure 6) 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. In contrast, these maps flagged areas of potential extrapolation (DIK > 1.0, red areas in Figure 6) in landscapes where biomass patterns are dissimilar from the disturbance scenarios in our simulation library. These included regions with strong anthropogenic influence such as the United Kingdom and Western Europe, agriculture-dominated ecosystems like in India, and regions with low overall forest fraction, such as Western Australia. The DIK for background mortality (K_b) exhibited 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 K_b values in these regions, which contradicted the long carbon turnover times known to characterize these ecosystems. This suggested that in systems dominated by large, infrequent, stand-replacing disturbances like fires in boreal regions, the subtle spatial signature of background mortality was masked by the strong imprint of the dominant disturbance regime, reducing parameter identifiability and model reliability for this specific parameter.

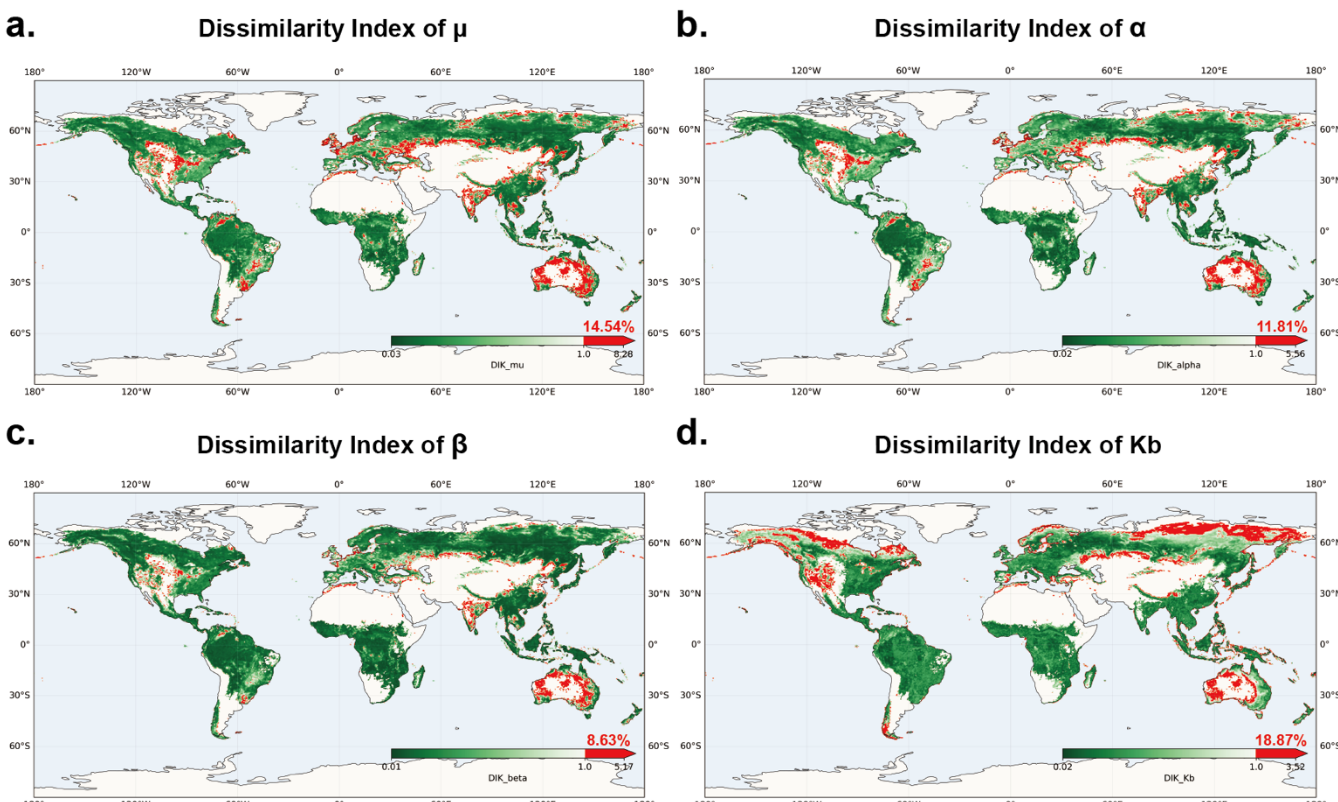

Figure 6. Model applicability uncertainty quantified by Dissimilarity Index. The four panels show the global gridded maps of the mean Dissimilarity Index (DIK) for each parameter. The DIK quantifies the novelty of observed landscapes compared to the training data. Values below 1.0 indicate that observed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

3.3 Empirical Evaluation

A direct, quantitative validation of our disturbance regimes is inherently challenging. While regional properties are documented through dendroecological studies and long-term inventories, a global dataset at the landscape scale as defined in our study is currently missing. Nevertheless, a preliminary cross-biome comparison over European forests using independent Landsat-derived records shows that our retrieved parameters fell within a highly comparable numerical range to event-based observations (Figure S10). Therefore, in this section we assessed the plausibility of our product by verifying whether real-world landscapes at the extreme tails of the global DRPs distributions (i.e., the lowest and highest predicted values) are successfully and accurately differentiated by our framework. To isolate the impact of each disturbance parameter (μ, α, β), we selected a pair of contrasting landscapes for a given parameter of interest (Figure 7c). However, these landscapes shared highly similar background conditions (GPP, K_b) as well as identical non-target disturbance parameters. This enabled us to perform a controlled visual assessment of how each parameter uniquely influenced biomass patterns (Figure 7).

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

The pairing for α contrasted a low-α site in the Amazon rainforest with a high α site in Southeast Asia (Figure 7c). The Amazonian site (low α) exhibited a coarse-grained spatial texture with a large clustered low biomass area (Figure 7b, Panel 3). The broad biomass histogram from this site indicated a landscape mix of intact forest blocks and significant clearings, which is a result of large-scale, infrequent events. Conversely, the Southeast Asian site (high α) presented a fine-grained, highly fragmented texture (Panel b4). Its biomass patterns were 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 highlighted how α modulates the spatial aggregation of disturbance. It effectively differentiated regimes dominated by a few large, contiguous events (low α) from those characterized by many small, dispersed events (high α), a finding consistent with our conceptual design (Figure 7a, Panel 3-4).

Both the low- and high- β sites were in the Amazon rainforest, and critically, the biomass histograms were highly similar. Both exhibited a bimodal distribution with no clear difference in mean biomass, which was expected as their μ and α were nearly identical. Despite the similarity in the first-order statistics, there was a stark contrast in spatial pattern between the low- and high- β sites, as low- β showed a more gradual spatial transitions in biomass while high- β featured distinct boundaries typical of severe disturbance footprints. This pairing powerfully illustrated the insufficiency of differentiating regimes solely based on first-order statistics from biomass. It underscored the necessity of using spatial-statistical features to capture more sophisticated disturbance characteristics, 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) were negatively skewed, characterized by a dominant peak at high values and a long tail extending toward low values. This feature was particularly pronounced for β, which displayed two distinct peaks for high values, with the extreme-high peak being the most dominant. In contrast, the parameter α was a unimodal distribution, approximately symmetric, with a 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 demonstrated that the predicted parameters could distinguish between remarkably different disturbance regimes that correspond to visually and ecologically distinct real-world biomass patterns. Furthermore, it highlighted the value of using richer feature sets, including second-order spatial statistics, to derive these parameters, as illustrated by the low- and high-β landscapes. The derived disturbance parameters were 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 continuous, observationally constrained representation of natural disturbance regimes.

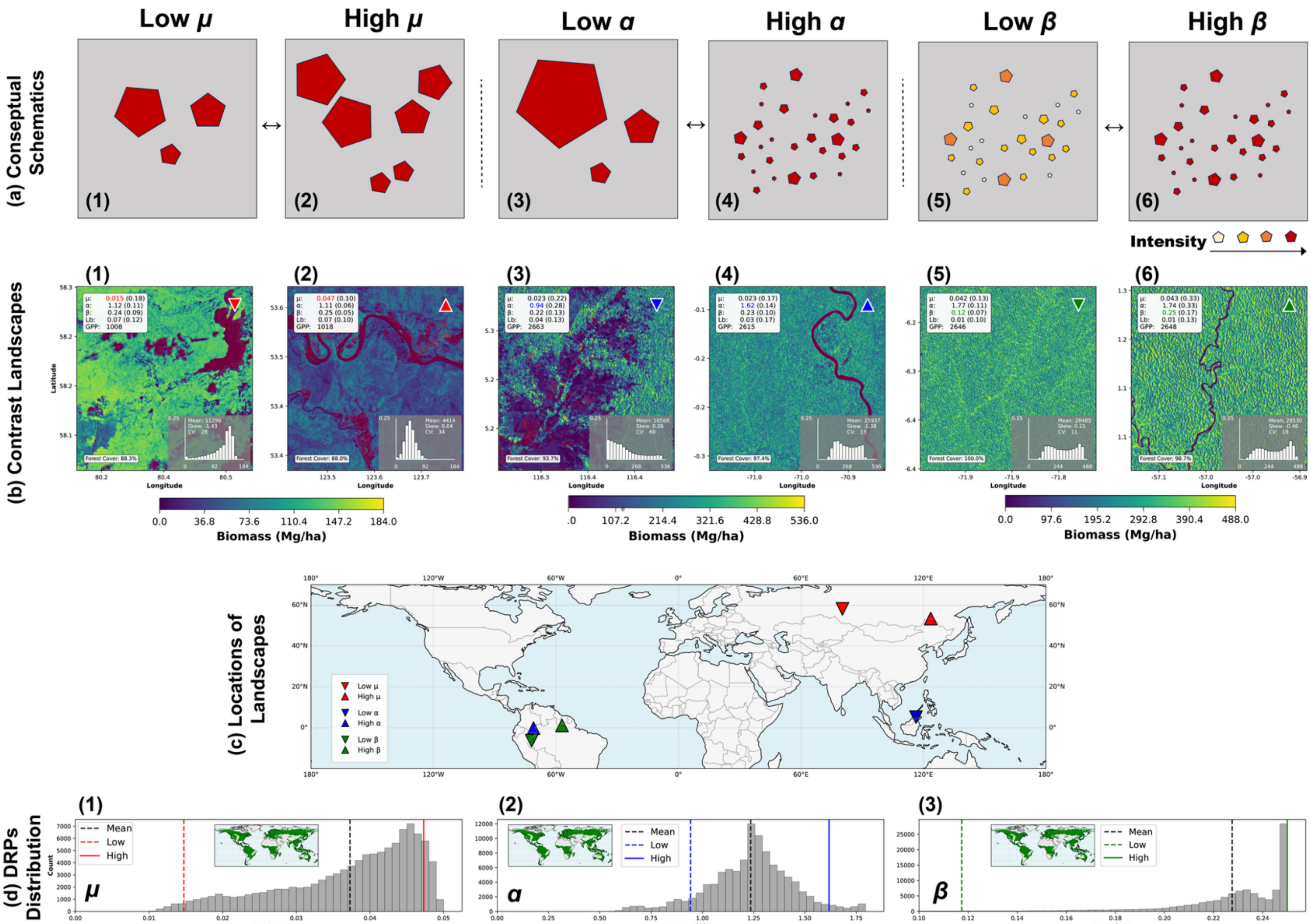

Figure 7. Correspondence between Predicted Disturbance Parameters and Empirical Biomass Patterns. (a) Conceptual Schematics: Theoretical landscape-scale patterns for six extreme cases (low vs. high) of the parameters μ, α, and β. (b) Contrasting Landscapes: Selected empirical biomass maps from the ESA GlobBiomass product. These low/high pairs were selected by filtering landscapes to ensure high similarity in photosynthesis level (GPP), background mortality (K_b), and the other two non-target DRPs, thereby isolating the contrast in the target parameter. (c) Locations of Landscapes: Geographic locations of the six selected contrasting landscapes 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.

3.4 Methodological Scope and Limitations

3.4.1 Pattern-based Inference versus Process-based Simulation

It is important to distinguish the conceptual scope of our framework from dynamic process-based models. In disturbance ecology, dynamic process-based models are designed to mechanistically simulate the physical spread of events over time, such as the step-by-step propagation of fire or insect outbreaks (e.g., Rammer and Seidl, 2022). In contrast, our approach operates within a pattern-based statistical paradigm. Consequently, our framework focuses on characterizing the spatial attributes represented in the EO AGB product, namely the size, form, and distribution of mapped biomass loss, rather than simulating the mechanistic temporal spread of disturbance processes (i.e., spatial contagion). These mapped spatial attributes may reflect both the true spatial organization of forest AGB, but also a spatial dependence in the dataset introduced by the AGB retrieval and processing chain, which cannot be partitioned using the available data.

Within this pattern-based paradigm, the stochastic disturbance event generator placed independent geometric shapes onto the simulated landscape. Crucially, while the synthetic grids have no predefined physical resolution, EO biomass products are mapped at defined native spatial resolutions. To bridge this fundamental scale mismatch and align the discrete simulated events with the continuous patterns of satellite data, spatial aggregation is utilized as a necessary scaling step to anchor the spatial statistics to the observational scale, aligning the spatial statistical properties of the simulations more closely with those of the EO product.

Furthermore, instead of explicitly simulating mechanistic contagion, our framework mathematically encapsulated the structural outcomes of complex spatio-temporal interactions, including linked disturbances and disturbance cascades, through a massive training library of over 8 million diverse regime combinations. By incorporating highly clustered gap-size distributions (α) and a wide diversity of complex, non-rectangular disturbance morphologies, the training dataset statistically approximated the diverse spatial footprints left by contagious processes. This pattern-based representation provided a highly efficient and robust means to infer generalized disturbance regime parameters from static landscape biomass patterns at a global scale.

3.4.2 Sensitivity to the Choice of Biomass Product

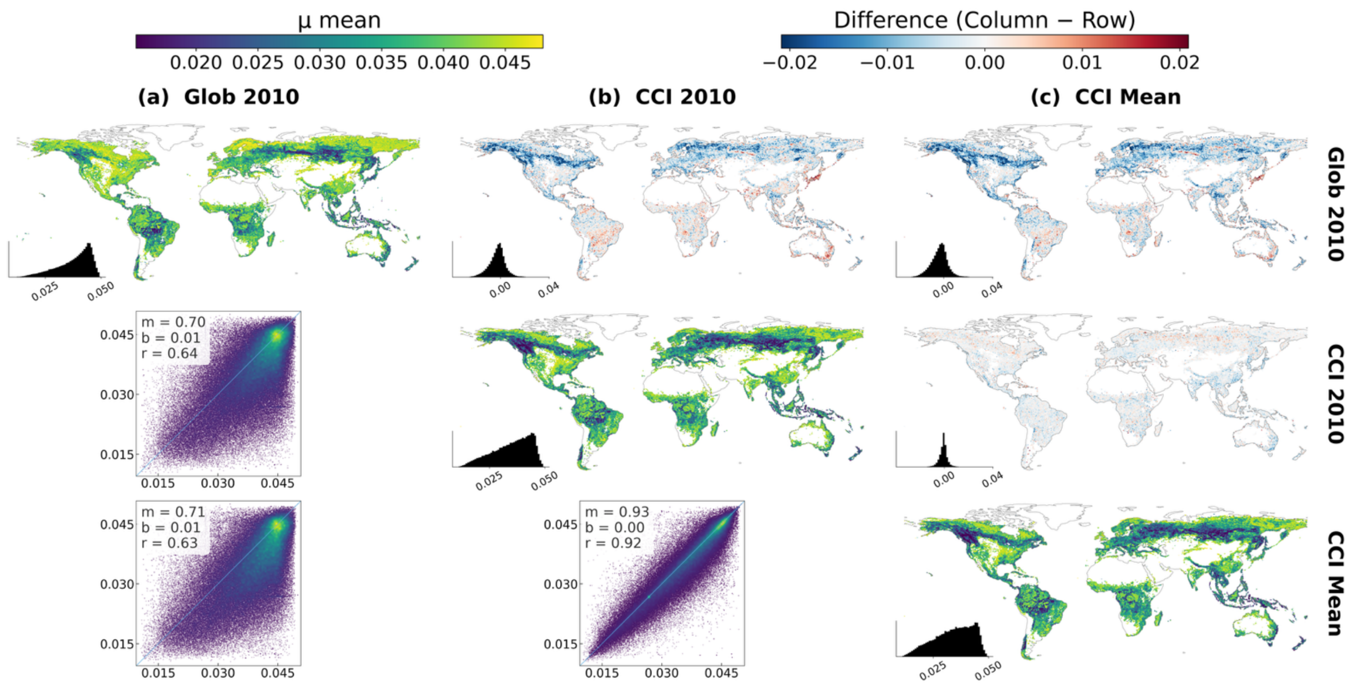

Figure 8. Spatial cross comparison of retrieved μ derived from different baseline biomass products. The diagonal panels display the global spatial distributions of μ predictions gridded at 0.25° × 0.25° resolution derived from (a) GlobBiomass (Glob 2010), (b) ESA CCI Biomass (CCI 2010), (c) the multi-year mean of ESA CCI Biomass (CCI Mean, 2005-2011), with a unified color bar. The upper off-diagonal panels display the spatial difference maps between each corresponding pair of datasets (column dataset minus row dataset) to illustrate localized discrepancies, using a separate symmetrical color bar centered at 0. The bottom off-diagonal panels present the pixel-to-pixel scatter plots.

To evaluate how baseline biomass data affects parameter retrieval, we repeated the inversion pipeline by replacing the primary GlobBiomass dataset (Glob 2010, 25 m resolution) with two alternative ESA CCI Biomass configurations (version 7) at 100 m resolution: a single-year baseline (CCI 2010) and a multi-year temporal mean (CCI Mean, averaging 2005-2011). This comparative analysis was designed to systematically assess the discrepancies arising from (1) different processing chains and native spatial resolutions (GlobBiomass versus ESA CCI, which share the same foundational SAR inputs and core algorithms), and (2) the temporal robustness of the framework against interannual variability (IAV) and short-term noise (CCI 2010 versus CCI Mean). The global inversion results showed that the spatial patterns of the four retrieved DRPs are highly consistent. The model captures the same broad geographical gradients and regional variations, confirming that the framework's performance was stable and reliable regardless of the specific spatial or temporal configuration of the input biomass products.

However, subtle differences in regional predictions do exist, which were linked to the structural characteristics of each baseline product. Figure 8 illustrates the comparison of the disturbance rate (μ) across the products. The cross-comparison, supported by pairwise scatter evaluations and spatial difference mapping, revealed that the predictions derived from the two ESA CCI configurations (CCI 2010 and CCI Mean) were highly coherent with minimal deviations between them (r = 0.92), whereas larger discrepancies occurred when comparing the CCI-based predictions against the primary GlobBiomass-based output (r ≈ 0.64). Among the four retrieved parameters, K_b exhibited the highest stability and lowest sensitivity to the choice of input data (r = 0.79), followed by α (r = 0.78) and β (r = 0.70), while the disturbance rate (μ) showed the largest variation across products (r = 0.64), even though its macro-scale spatial distributions remained consistent. The global maps and pixel-to-pixel scatter plots for the remaining three parameters are provided in the supplement (Figure S6; S7; S8).

The deviations observed between the GlobBiomass- and CCI-driven predictions were primarily attributable to their different baseline spatial resolutions and data processing lineages. Although both datasets were aggregated into a consistent 100 × 100 matrix (yielding an effective spatial resolution of 250m) within each 25 km × 25 km landscape tile to extract the spatial metrics used for inversion, their native resolution differences and specific processing chains potentially altered the local spatial heterogeneity of the biomass maps. These distinct data characteristics propagated through the biomass spatial feature extraction and lead to the variations in parameter estimation. Additionally, to support ecological modeling and enable users to account for these input-data uncertainties, we provided alternative global datasets of the four DRPs derived from both the CCI 2010 and CCI Mean pipelines. These companion products were gridded consistently at a spatial resolution of 0.25° × 0.25° as well.

3.4.3. Sensitivity of DRPs to Dampened Spatial Variance in High-Biomass Forests

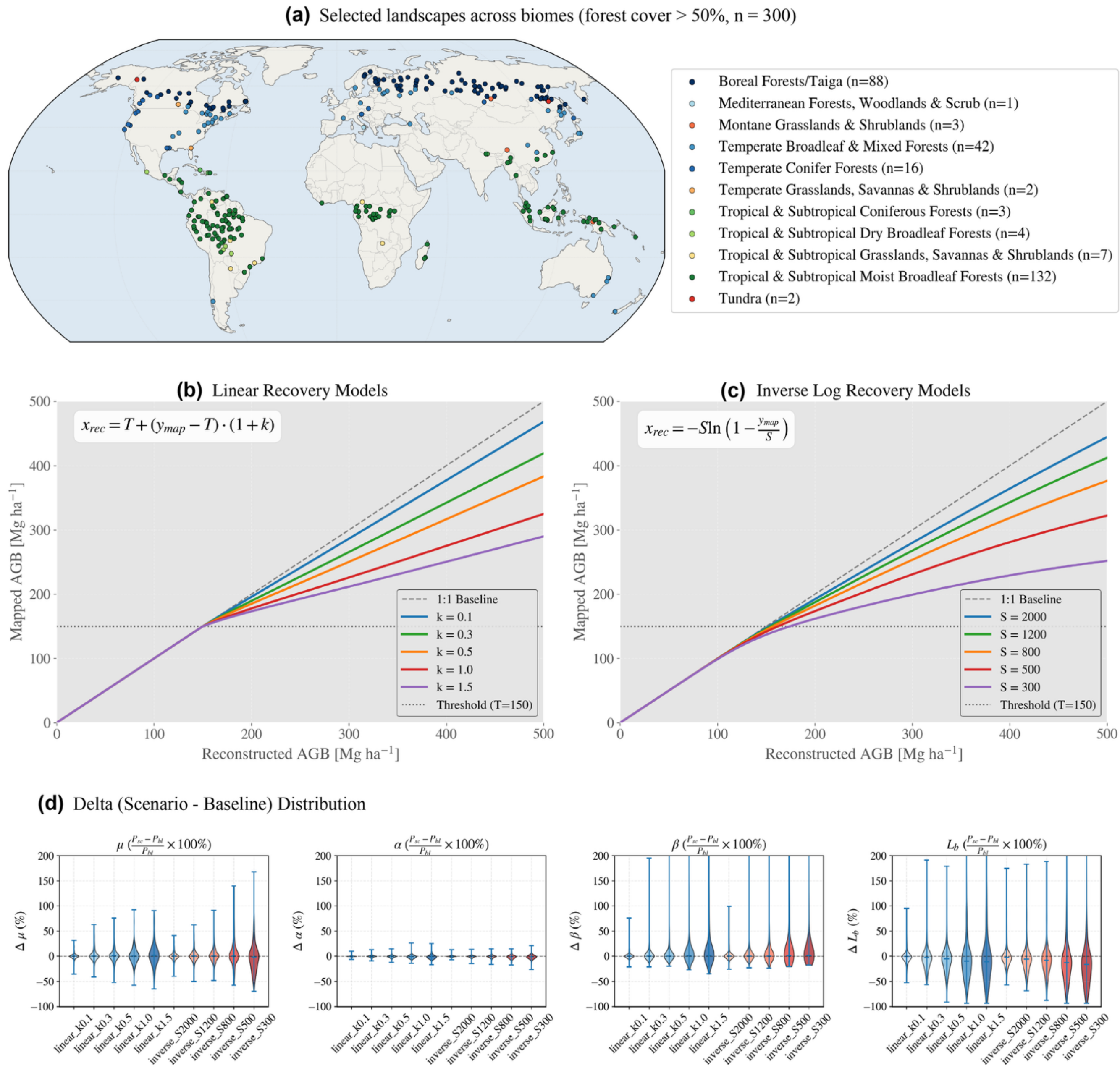

Figure 9. Sensitivity of predicted DRPs to potential dampening of spatial variance across global landscapes. (a) Geographic distribution of the 300 randomly selected baseline landscapes (forest cover > 50%) across major biomes. (b, c) Mathematical configurations for the (b) linear and (c) inverse logarithmic recovery models used to synthetically inject variance for AGB values exceeding high-biomass thresholds. The linear model uses a proportional scaling parameter $k \in [0.1, 1.5]$ while the inverse model uses a parameter $S \in [300, 2000]$. Both methods apply a continuous sigmoid weight function to smoothly blend the original and recovered above-ground biomass (AGB). (d) Multi-panel violin plots displaying the relative percentage changes ($\Delta P = \frac{P_{sc} - P_{bl}}{P_{bl}} \times 100\%$, where sc denotes the reconstructed AGB and bl denotes the mapped

baseline) of the four Disturbance Regime Parameters across the 10 reconstructed scenarios. Visually, the plots illustrate the expansion of prediction variance for μ, β, and K_b as recovery intensity increases (from left to right), the stability of α centered around 0%, and the coupled opposite shifts in the predicted means of β ($+\Delta$) and K_b ($-\Delta$).

When evaluated against ground reference data, global products such as GlobBiomass and ESA CCI Biomass exhibit a tendency for mapped biomass values to be lower than reference values for high-biomass ranges (Santoro et al., 2021; 2024). However, as detailed in Santoro et al. (2021; 2024), this product-level discrepancy cannot be attributed solely to signal saturation and may also be associated with algorithmic dependencies on other datasets, such as spaceborne LiDAR (e.g., ICESAT-2 or GEDI) or land cover masks for model calibration, or spatial filtering steps that mitigate inherent SAR sensor noise and processing artifacts, smoothing local signals and reducing the spatial variance within mature forests.

Because our framework for deriving DRPs from AGB relies on capturing localized spatial variations, such as the Coefficient of Variation, and GLCM texture features, this combined dampening effect in the high-biomass upper tail may systematically impact the DRP prediction. To explore possible implications from such potential biases, we designed a sensitivity analysis by artificially reconstructing the dynamic range of biomass values, i.e. increasing the AGB values at high biomass values, at different levels of intensity, to observe the resulting impacts in DRP predictions.

For the sensitivity analysis, we randomly selected 300 high-biomass 25 × 25 km forest tiles across major biomes globally (forest cover > 50% and AGB > 220 Mg ha^{-1}) to capture a continuous macro-ecological gradient. The experiment operated on the biomass data at its native resolution (i.e., 25 m for GlobBiomass and 100 m for CCI Biomass) prior to the spatial aggregation process. We implemented two comparable methods to reconstruct each landscape's AGB above different specified high-biomass thresholds. To comprehensively test the framework (Figure 9), we considered a range of high-biomass thresholds ($T \in \{150,200,250,300\}$ Mg ha^{-1}). We further applied two transformations to AGB values above each threshold: (1) a linear transformation controlled by the scaling parameter k ($k \in [0.1,1.5]$), with larger k values producing stronger increases in AGB; and (2) an asymptotic inverse-log transformation controlled by the scale parameter S ($S \in [300,2000]$ Mg ha^{-1}), with smaller S values producing stronger increases in AGB. In both approaches, a soft sigmoid weighting function at the threshold T was applied to smoothly blend the original and reconstructed AGB, preventing artificial edge artifacts (Supplementary S5.1). The resulting reconstructed biomass maps were systematically aggregated to standardized coarse-resolution pixels (100 × 100 matrix of 250 × 250 m^2 resolution), following the Area-Weighted Spatial Averaging (for CCI) and Block-Mean Averaging. Finally, the 16 biomass spatial statistics were extracted from the aggregated pixels, which, alongside the GPP predictor, were used to invert the four DRPs using our pre-trained, multi-output Random Forest model.

We quantified the sensitivity of the DRPs to the changes in biomass by calculating the difference (Δ) between the predicted DRPs under the different reconstructed landscapes and the baseline. As the difference between the baseline and the reconstructed landscapes increase (e.g., k from 0.1 to 1.5, or S decreasing from 2000 to 300) so does the variance in the ΔDRPs (Figure 9d). We observe a degradation of R^2 values across landscapes for all four parameters proportional to the fractions of grid cells beyond the high-biomass thresholds and to the underestimation strength (Figure S14). Ultimately, the magnitude of these deviations is tightly controlled by the choice of the high-biomass threshold (T); at higher thresholds, fewer forest pixels are affected (Figure S12; S13), leading to smaller deviations in the final DRP predictions (Figure S15). The large fraction of mapped AGB values above the selected high-biomass thresholds in both tropical and extratropical regions (Figure S12) reflect the presence of a broad upper tail, indicating that the biomass product does not impose a universal hard truncation at these values. However, it is not clear to which extent these distributions reflect true high biomass estimates from variations cause by the retrieval-model assumptions, constraints or noise, and should therefore be interpreted with caution.

Further to note that the different DRPs -μ, β and K_b- exhibit different sensitivities to dampened retrievals. The parameter α demonstrates a strong resistance to the possible compression of the dynamic range, maintaining structural alignment with the baseline even under the most extreme reconstruction scenarios. Conversely, at larger differences between reconstructed and baseline AGB, systematic shifts emerge for the predicted means of β - that shift upward - and for the predicted means of K_b - that shift downward. These differences reflect that an artificial expansion of the dynamic range on high biomass result in a higher AGB retrieval, which, for the same GPP, would imply lower baseline mortality rates (lower K_b) but also an amplification in the spatial contrast, implying more severe disturbance footprints which then drives higher β retrievals. This experimental protocol was replicated using the ESA CCI Biomass v7 (single-year) and the CCI 7-year temporal mean datasets. We observed consistent results across the two alternative CCI biomass datasets, with α remaining stable and β and K_b showing the same directional shifts (Figure S16).

Overall, these experiments illustrate how the different DRPs respond to compression of the upper end of the biomass distribution. The stability of α across all scenarios indicates that the inferred spatial clustering frequency of disturbances is comparatively robust to upper-tail dampening. In contrast, μ, β, and K_b are sensitive to reductions in spatial variance and to underestimation in high-biomass forests. This sensitivity represents an intrinsic limitation of the framework because suppressed biomass variability, together with retrieval noise, can propagate into the inferred disturbance parameters. As such, the magnitude of the resulting deviations increases with both the fraction of affected pixels and the strength of the imposed reconstruction. Despite differences in spatial resolution and temporal configuration, the spatial harmonization protocols produced consistent sensitivity bounds across all evaluated products. To further overcome the dampened retrievals in high-biomass forests, the emergence of data streams from novel missions, such as the P-band BIOMASS, may play a significant role in diagnosing global disturbance regime parameters.

4 Data Availability

The two data products described in this paper, the Tile-Level Dataset (CSV) and the Gridded Global Dataset (NetCDF), are publicly available in Edmond Repository at https://edmond.mpg.de/previewurl.xhtml?token=9dea6a9e-f4da-4d83-8bf8-de46d7fcfd4a (Wang et al., 2025). The datasets contain the global disturbance regime parameters (μ, α, β, K_b) and the associated uncertainty layers detailed in this manuscript.

5 Code Availability

The source code used to process the input data, train the models, and generate the final dataset is available in Edmond Repository at https://edmond.mpg.de/previewurl.xhtml?token=c4883cbb-7292-4d9c-82b7-570c015ca741 (Wang et al., 2025).

6 Conclusion

This study presents a novel global dataset of forest disturbance regimes, comprising disturbance rate, gap-size distribution, disturbance severity, and background mortality, derived from high-resolution satellite-based biomass observations. Grounded on the concept of gap dynamics, this framework was enabled by a massive synthetic training dataset and high-performance computing. We quantified two key sources of uncertainty: one inherent to the machine learning predictions and another related to model applicability (extrapolation risk). Together, these uncertainty measures indicate reliable predictions across approximately 90% of global forest regions. Furthermore, cross-comparisons across alternative input data sources demonstrate that the retrieved spatial distributions of disturbance regime parameters remain highly robust. In addition, evaluations using paired contrasting case studies confirm that the derived parameters correspond to ecologically distinct real-world biomass patterns, confirming the plausibility of inferring long-term disturbance regimes from spatial structural features.

Nevertheless, several challenges remain, including the reduced sensitivity of spaceborne radar signals to local spatial variance in dense forests, the model assumption of pixel independence that omits explicit transient contagion dynamics, and the current lack of extensive, independent global ground-truth validation. Addressing these gaps, particularly by integrating and comparing our dataset with regional observations, represents critical directions for future research. Ultimately, this dataset offers an important 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 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

Bonan, G.: Ecological Climatology: Concepts and Applications, 3rd ed., Cambridge University Press, Cambridge, https://doi.org/10.1017/CBO9781107339200, 2015.

Bormann, F. H. and Likens, G. E.: Pattern and Process in a Forested Ecosystem: Disturbance, Development and the Steady State Based on the Hubbard Brook Ecosystem Study, Springer Science & Business Media, 265 pp., 2012.

Bray, J. R.: Gap Phase Replacement in a Maple-Basswood Forest, Ecology, 37, 598–600, https://doi.org/10.2307/1930185, 1956.

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.

ESA: Land Cover CCI Product User Guide Version 2, European Space Agency, 2017.

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., 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.

Le Toan, T., 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, IGARSS 2018 - 2018 IEEE

International Geoscience and Remote Sensing Symposium, 8563–8566, https://doi.org/10.1109/IGARSS.2018.8518491, 2018.

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.

Rammer, W. and Seidl, R.: Box 17.1: Disturbance Modelling, in: Disturbance Ecology, Springer International Publishing, 399–401, 2022.

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.

Santoro, M. and Cartus, O.: ESA Biomass Climate Change Initiative (Biomass_cci): Global datasets of forest above-ground biomass for the years 2005-2012 and 2015-2024, v7.0, https://doi.org/10.5285/6429D1AAFE1E43B9B414E4A5A7F8B903, 2026.

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.

Shugart, H. H.: A Theory of Forest Dynamics: The Ecological Implications of Forest Succession Models, Springer New York, 308 pp., 1984.

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.

Wang, S.: Codes for manuscript of A Global Dataset of Forest Disturbance Regimes Derived from Satellite Biomass Observations (DRAFT VERSION), https://doi.org/doi:10.17617/3.DBIFDT, 2025.

Wang, S., Yang, H., Koirala, S., Santoro, M., Weber, U., Robin, C., Cremer, F., Forkel, M., Reichstein, M., and Carvalhais, N.: A Global Dataset of Forest Disturbance Regimes Derived from Satellite Biomass Observations (DRAFT VERSION), https://doi.org/doi:10.17617/3.EF4QGR, 2025.

Watt, A. S.: Pattern and Process in the Plant Community, Journal of Ecology, 35, 1–22, https://doi.org/10.2307/2256497, 1947.

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.