https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

Aggregating signals of Earth system dynamics across space, time, models, and variables

Kobe De Maeyer[1], Jakob Harteg[2, 3], Jonathan F. Donges[2, 4, 5], Ricarda Winkelmann[2, 3, 4], and Sina Loriani[2, 4]

[1]Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, Netherlands
[2]Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, Telegrafenberg A 31, 14473 Potsdam, Germany
[3]Institute for Physics and Astronomy, University of Potsdam, Karl-Liebknecht-Str. 24/25, 14476 Potsdam-Golm, Germany
[4]Integrative Earth system Science, Max Planck Institute of Geoanthropology, Jena, Germany
[5]Stockholm Resilience Centre, Stockholm University, Stockholm, Sweden

Correspondence: Kobe De Maeyer (k.h.m.demaeyer@uu.nl) and Sina Loriani (sina.loriani@pik-potsdam.de)

Abstract. Model Intercomparison Projects (MIPs) provide standardised computer simulations of the Earth system, offering unique opportunities to systematically detect and assess features and dynamics, such as abrupt shifts, across diverse models and variables. Recent advances combine time-series analysis with spatiotemporal clustering to identify dynamically connected regions within individual datasets. Yet, extending this notion of connectivity across the model and variable dimensions of MIP output remains an open challenge. Here, we present a conceptual workflow that addresses this by introducing two aggregation strategies for "detect-then-cluster" pipelines: "Detect-Cluster-Aggregate-Cluster" (DCAC) and "Detect-Aggregate-Cluster" (DAC), enabling systematic synthesis of spatiotemporal signals across multiple datasets. These aggregation algorithms are evaluated and tuned using a customisable Analytic Hierarchy Process (AHP) framework, which allows users to encode prior knowledge about dataset reliability. In anticipation of output from the Tipping Points Modelling Intercomparison Project (TIPMIP) and other MIPs within the Coupled Model Intercomparison Project (CMIP), we implement the proposed aggregation methods using the "Tipping and Other Abrupt Events Detector" (TOAD) package. To demonstrate feasibility, we apply the methods to CMIP6 simulations of Amazon rainforest dynamics, detecting and clustering abrupt vegetation shifts first across multiple variables, where a shared signal indicates a coherent ecosystem response, and then across multiple models, where a shared signal reflects model alignment. Our case study reveals that this aggregation helps distinguish such shared behaviour from dynamics that are specific to individual variables or models, patterns that typically remain obscured when datasets are analysed in isolation. These results illustrate that conclusions about abrupt dynamics depend critically on how information is synthesised across time, space, models, and variables. While showcased here in the context of tipping points, the proposed aggregation framework provides a structured and transferable foundation for multimodel and multivariate risk assessments of diverse Earth-system processes within MIPs.

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

1 Introduction

Numerical model simulations, from simple conceptual representations to complex Earth system Models (ESMs), have transformed our understanding of critical Earth system and climate processes (Edwards, 2011; Steffen et al., 2020; Intergovernmental Panel on Climate Change, 2023). By explicitly representing interacting biogeochemical and physical processes, process-based models provide a virtual laboratory to explore Earth-system dynamics and emergent behaviour under changing boundary conditions. Such model experiments allow researchers to project future changes, reconstruct past states, and analyse present-day mechanisms, complementing observational evidence.

Yet, the increasing diversity of models and experimental configurations poses challenges for consistency and comparability, making coordinated frameworks such as Model Intercomparison Projects (MIPs) indispensable cornerstones of modern Earth system and climate research. MIPs are international community efforts in which multiple modelling groups follow a common experimental protocol, such that various models perform the same set of simulations. Consequently, differences in the results can be attributed to model structure and process representation rather than experimental setup. There are many different MIP initiatives, each tailoring its streamlined protocols to specific scientific questions. Examples include the Coupled Model Intercomparison Project (CMIP; e.g. Eyring et al. (2016)), the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP; e.g. Frieler et al. (2017)), and the more recently launched Tipping Points Modelling Intercomparison Project (TIPMIP; e.g. Winkelmann et al. (2025)).

The growing volume and complexity of MIP output increasingly call for data mining approaches that can consistently identify robust patterns, assess uncertainties, and diagnose relevant Earth-system dynamics in a reproducible and automatic way. Structured approaches to MIP evaluation, such as the Earth System Model Evaluation Tool (ESMValTool), have been developed to facilitate and streamline the assessment of Earth system models (Eyring et al., 2020; Righi et al., 2020). In the context of tipping points and abrupt, nonlinear Earth-system changes, a variety of detection pipelines have been proposed in recent years. In the simplest case, abrupt shifts in time series are identified as anomalies exceeding predefined value-based thresholds, for example specifically chosen for the Amazon rainforest (Parry et al., 2022). Applicable to a wider range of systems, Drijfhout et al. (2015) developed a more general, criteria-based pipeline to compile a statistical catalogue of abrupt changes in CMIP5 simulations, which was recently updated and applied to CMIP6 output (Angevaare and Drijfhout, 2025).

An alternative line of work screens for dynamically connected regions in MIP output, explicitly accounting for spatiotemporal connectivity. Bathiany et al. (2020) configured the Canny edge-detection algorithm, originally developed for image processing, to identify abrupt changes jointly in space and time, applying it to CMIP5 simulations. This approach was later extended to CMIP6 ensembles by Terpstra et al. (2025). Along similar lines, the Tipping and Other Abrupt Events Detector (TOAD v1.0) operationalises spatiotemporal connectivity through a "detect-then-cluster" pipeline, in which abrupt changes are first evaluated in per-grid-cell time series and subsequently grouped in time and space using clustering algorithms (Harteg et al., 2026). The open-source Python package TOAD provides a modular plug-and-play framework that allows users to combine and exchange detection and clustering algorithms. Beyond tipping and abrupt shifts, "detect-then-cluster" frameworks have also been applied

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

to other Earth-system dynamics, such as the identification of climate extremes and anomalies in ClimBurst (Brouillet et al., 2025).

While recent work has made progress in identifying dynamically connected regions within individual datasets, extending this notion of connectivity across multidimensional MIP output (e.g. for different models and variables) remains limited. Put simply, if two models both detect similar dynamics in the same region at the same time, most "detect-then-cluster" methods have no structured way to automatically identify and represent such cross-dataset agreement as a coherent signal. A comprehensive aggregation framework capable of systematically capturing such alignment, both where and when it occurs across models and variables, is therefore needed. Without such synthesis, signals of major Earth-system dynamics may remain fragmented or overlooked, limiting the full scientific and policy relevance of MIP data.

In this paper, we address this gap by extending "detect-then-cluster" pipelines to enable aggregation of results across space, time, models, and variables. While exercised at the concrete example of abrupt shift detection with TOAD, the methodology is transferrable to other pipelines and evaluation frameworks that perform a per-grid-cell time series analysis followed by spatiotemporal clustering of detected events. The remainder of the paper is structured as follows:

- We introduce two different aggregation algorithms that generalise "detect-then-cluster" pipelines to enable synthesis of coinciding signals across datasets: "Detect-Cluster-Aggregate-Cluster" (DCAC) and "Detect-Aggregate-Cluster" (DAC) (Section 2).
- We present a structured method based on the Analytic Hierarchy Process (AHP) to evaluate the synthesised results and tune free parameters within the generalised pipeline, explicitly incorporating prior knowledge on dataset reliability (Section 3).
- We implement the proposed algorithms with TOAD and apply them to CMIP6 simulations of Amazon rainforest vegetation dynamics to demonstrate their functionality (Section 4).
- We compare the two aggregation approaches, interpret the case study results in the context of previous studies, and discuss the strengths, limitations, and broader applicability of the methodology beyond the present case study (Section 5).

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

2 Structured aggregation methods

Throughout this paper, we use the term *aggregation* to refer specifically to the combination of detected signals *across datasets*, that is, across models and/or variables, rather than, for example, the spatial or temporal aggregation of physical quantities within a single dataset. We introduce two complementary aggregation algorithms designed to extend "detect-then-cluster" pipelines across multiple datasets: "Detect-Cluster-Aggregate-Cluster" (DCAC) and "Detect-Aggregate-Cluster" (DAC). Both aim to integrate signals of Earth-system dynamics that consistently emerge across multiple models or variables within similar regions and time periods, but they differ in their approach. DCAC identifies coinciding patterns by performing aggregation after a first individual clustering step, whereas DAC starts by aggregating detected dynamics across datasets, followed by clustering.

Figure 1 illustrates how the different aggregation strategies are embedded within the generalised "detect-then-cluster" pipeline. We define a dataset as a spatiotemporally gridded simulation output for a single variable and a single model. For each dataset, time series analysis is first performed independently in each grid cell to detect phenomena such as an abrupt shift or the crossing of a critical value. This step yields a detection time series (dts) for each grid cell, indicating where and when the dynamics under consideration occur. The resulting spatiotemporal information can then be clustered and aggregated to identify connected regions exhibiting similar dynamics across multiple datasets. We refer to this end result as a "cluster map".

To extend this "detect-then-cluster" pipeline toward multimodel or multivariate synthesis, aggregation can be introduced at two different stages: either the spatiotemporal detection information is first aggregated and then clustered (DAC), or clustering is performed first and the resulting clusters are subsequently aggregated before a final clustering step (DCAC). Both strategies rely on an underlying clustering process. Any algorithm can be chosen, as long as it can effectively handle large spatiotemporal datasets, form clusters of arbitrary shape, and delineate regions of detected dynamics without forcing every data point into a specific cluster. A common family of clustering algorithms meeting these requirements are density-based methods (Ester et al., 1996; Jain, 2010; Harteg et al., 2026).

Below, we describe the DAC and DCAC aggregation algorithms in detail. To illustrate their behaviour conceptually, we employ a synthetic ensemble of datasets, each representing a spatiotemporally gridded output for a single variable and model, constructed around a common, overarching abrupt event with slight spatial and temporal deviations between datasets. A full description of the data generation procedure and rationale is provided in appendix A1.

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

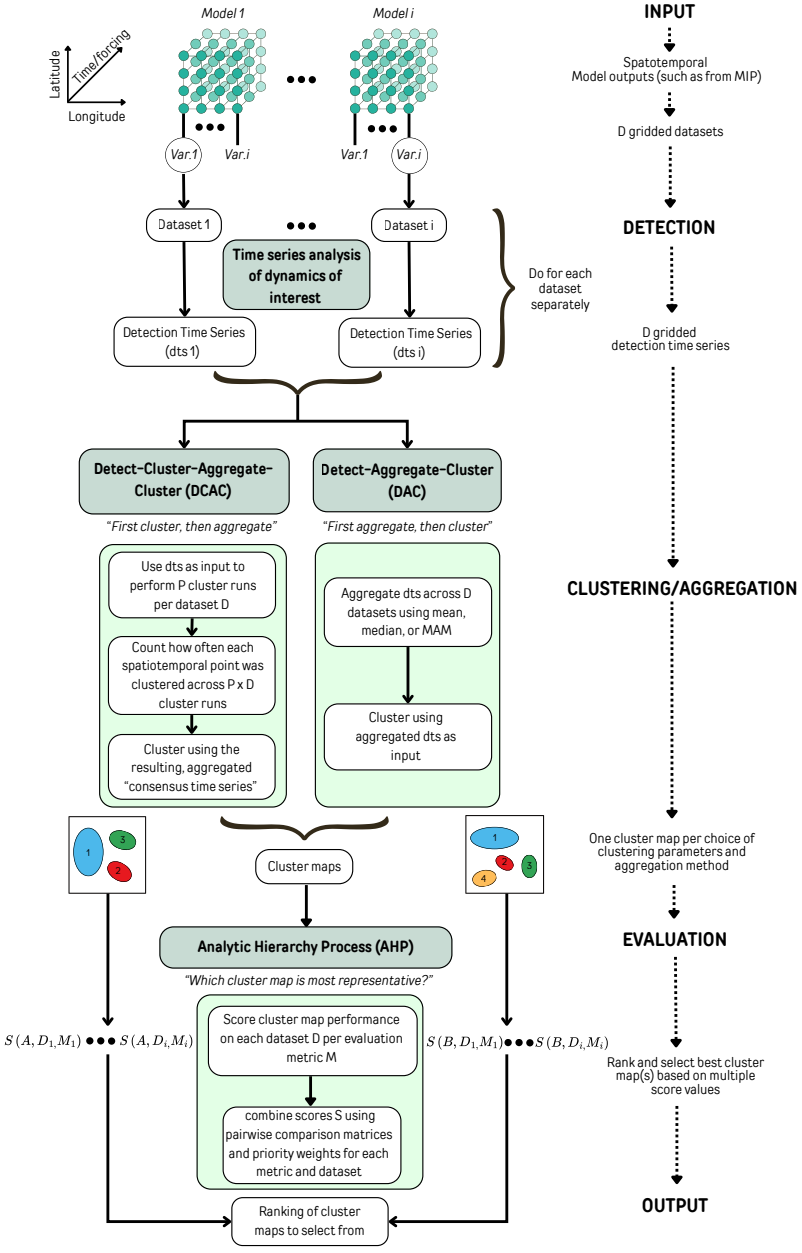

Figure 1. Extended detect-then-cluster pipeline for the aggregation of results across space, time, models, and variables. Time series analysis (detection) of Earth-system dynamics of interest (e.g. abrupt shifts) is first performed independently for each dataset on a grid level. Each dataset corresponds to a specific variable and model. This yields a set of detection time series, one per original dataset, where each value can be understood as the local likelihood of undergoing the studied dynamics. These are then used as inputs for two alternative strategies to cluster results in time and space and aggregate across datasets: (1) Detect-Aggregate-Cluster (DAC), and (2) Detect-Cluster-Aggregate-Cluster (DCAC). The resulting cluster maps are evaluated using predefined priority weights and multiple evaluation metrics measured on all original datasets following the Analtyic Hierarchy Process (AHP) framework to identify the most robust representations.

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

2.1 Detect-Aggregate-Cluster (DAC)

The Detect-Aggregate-Cluster (DAC) approach encompasses methods that initially aggregate the individual detection time series (dts_i) from all D original datasets into a single overarching dts for each grid cell. This spatiotemporal information, synthesised across datasets, can then be used as input for clustering. Given an aggregation function (see Eqs. (1-3)), we compute a single overarching dts-value from all individual dts_i for each data point in space and time. We propose three versions of DAC, each employing a different aggregation function:

$$\text{Median (after sorting values)}: \quad \mathrm{dts} = \begin{cases} \mathrm{dts}_{\left(\frac{D+1}{2}\right)}, & \text{if } D \text{ is odd} \\ \frac{1}{2}\left(\mathrm{dts}_{\left(\frac{D}{2}\right)} + \mathrm{dts}_{\left(\frac{D}{2}+1\right)}\right), & \text{if } D \text{ is even} \end{cases} \tag{1}$$

$$\text{Mean}: \quad \mathrm{dts} = \frac{\sum_{i=1}^{D} \mathrm{dts}_i}{D} \tag{2}$$

$$\text{Magnitude-Adjusted Mean (MAM)}: \quad \mathrm{dts} = \frac{\sum_{i=1}^{D} |\mathrm{dts}_i| \cdot \mathrm{dts}_i}{\sum_{i=1}^{D} |\mathrm{dts}_i|} \tag{3}$$

These aggregators are illustrated in Figure 2, which shows their application to synthetically generated ensemble data (see appendix A). In general, the MAM aggregator consistently produces high dts-values and preserves the peak signals from the individual dts_i, resulting in a conservative, coarsely aggregated dts. In contrast, the median and mean aggregators capture less of the peaks in the individual datasets, resulting in a smoother and lower aggregated dts. With only 5 datasets, the mean and MAM identify two peaks, while the median captures only the earlier peak. This may be due to the median's robustness to outliers. With increasing D, the aggregated dts from the median and mean aggregators converge. Nevertheless, the median aggregator still displays more step-like patterns and is generally lower.

The figure underscores the key differences between the various DAC aggregator functions. When it is important to cluster a small time window and spatial region where most datasets exhibit the dynamics under consideration, the smooth median or mean aggregator may be preferred. In contrast, the more conservative MAM aggregator captures a larger extent of individual peaks in each dataset's detection series, leading to a broader cluster map, which may result in a higher recall (more true signals captured) but at the cost of precision (more false signals captured). However, adjusting the hyperparameters of the underlying clustering algorithm, can also result in a narrower cluster map without the need to smooth the aggregated dts with the median or mean approach. Thus, when selecting the appropriate DAC aggregator function, it is important to also consider the corresponding clustering hyperparameter choices.

2.2 Detect-Cluster-Aggregate-Cluster (DCAC)

The second strategy for synthesising signals of Earth-system dynamics across D datasets is "Detect-Cluster-Aggregate-Cluster" (DCAC). In this approach, clustering is first performed for each dataset individually, followed by an aggregation of all these

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

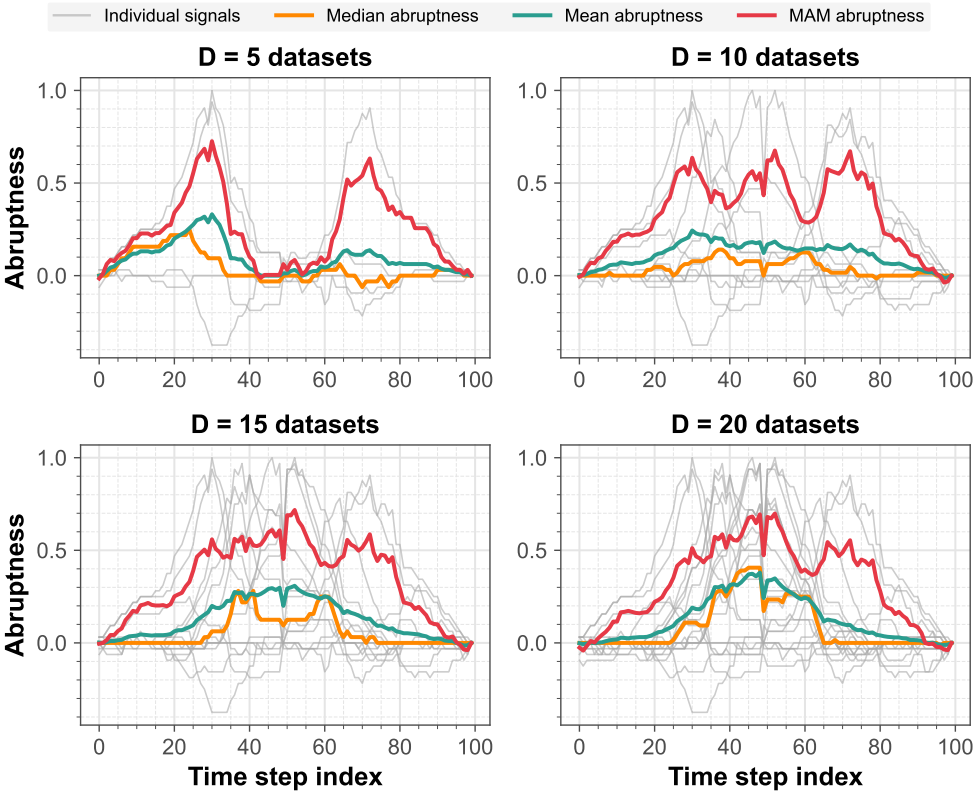

Figure 2. Illustration of the Detect-Aggregate-Cluster (DAC) approach, where a single overarching detection time series (dts) is constructed from D individual detection series. Different aggregator functions are shown for a single spatial grid cell over time, based on varying numbers of individual datasets from the synthetically derived ensemble data (see appendix A). The light gray lines indicate the original individual detection time series. We aggregate across these using three different functions: Median, Mean, and Magnitude-Adjusted Mean (MAM).

clusters across datasets. To combine the individual cluster maps into overarching clusters, DCAC evaluates each data point in time and space by counting how often it is classified as part of a cluster across all D datasets. Specifically, each dataset's spatiotemporal cluster map is binarised, assigning a value of 1 to clustered data points and 0 to unclustered ones. Summing these binarised maps across datasets yields, for each data point, the number of datasets in which it belongs to a detected cluster. Points with high counts thus indicate spatiotemporal regions where many models consistently identify dynamically coherent signals. We refer to the resulting sequence of values at each spatial grid cell as the "consensus time series" (cts), as illustrated in Figure 3.

In the next step, DCAC converts the consensus time series cts into discrete, overarching consensus clusters. This is achieved by applying a second-stage clustering procedure, using the cts as input instead of the original detection time series (dts). In this way, clustering is no longer based on individual detection events, but on the degree of agreement across datasets encoded

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

in the consensus signal. The resulting clusters therefore delineate spatiotemporal regions where multiple datasets consistently exhibit similar dynamics of interest, representing areas of relatively high inter-dataset consensus.

The description above assumes that, for each dataset, a single spatiotemporal cluster map is obtained from the first-stage clustering process and subsequently used for aggregation. In practice, however, clustering results depend on the choice of hyperparameters, and retaining only a single parameter configuration per dataset would require an arbitrary selection that may bias the outcome and propagate into the second-stage clustering. To mitigate this dependence, DCAC allows multiple cluster maps with varying clustering hyperparameters for each individual dataset. Accordingly, the cts is computed by summing the binarised cluster maps C across both parameter sets P and datasets D, as formalised in Eq. (4). Besides reducing biases, aggregating across parameter sets effectively smooths the resulting cts, which improves the robustness of the second-stage clustering, particularly when the number of available datasets is small.

$$\mathrm{cts} = \sum_{i}^{D} \sum_{p}^{P} C_{ip} \tag{4}$$

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

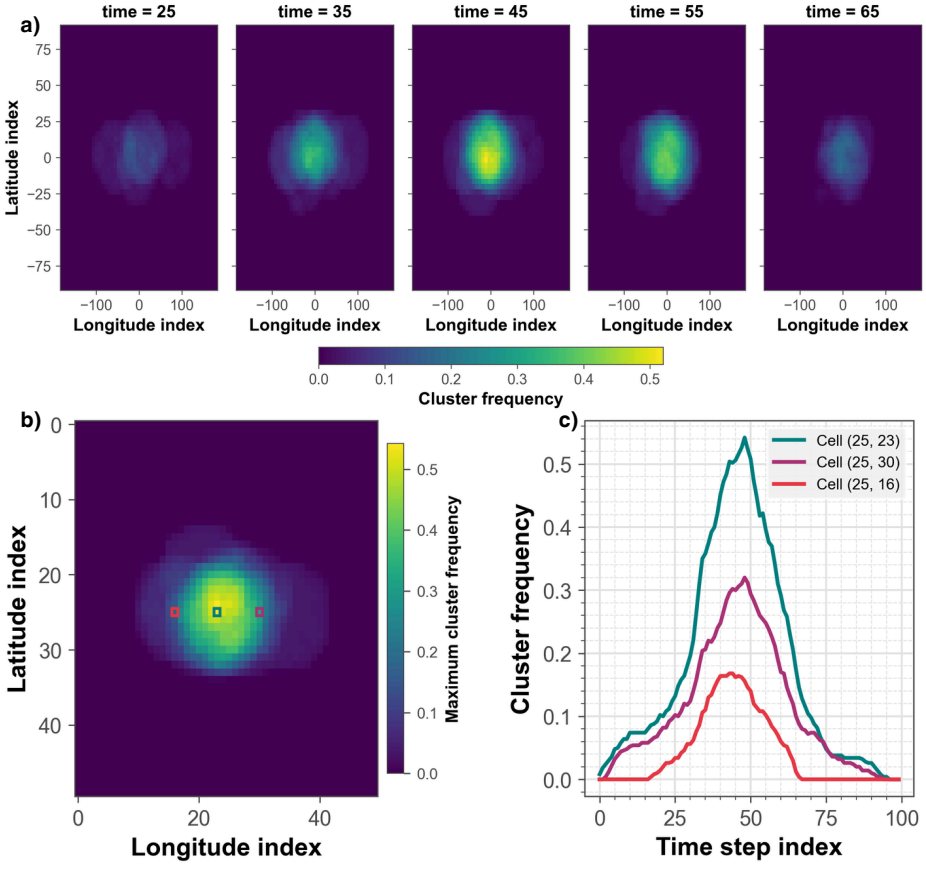

Figure 3. Illustration of the Detect-Cluster-Aggregate-Cluster (DCAC) approach, where an overarching consensus time series is constructed based on the cluster maps from each individual dataset. For each dataset (20 in total) of the synthetically derived ensemble data (see appendix A), 25 clusterings are performed using different hyperparameter settings. For each point in space and time, we then count how often it is assigned to a cluster across all resulting cluster maps (25 x 20), producing a spatiotemporal consensus time series (panel a). The maximum value of this series for each spatial grid cell is visualised in panel b. For three grid cells, each marked with a different color, the evolution of consensus values over time are shown in panel c, where the peak value corresponds to the value plotted in the heatmap.

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

3 Evaluation through Analytic Hierarchy Process (AHP)

Applying the full pipeline to multiple datasets yields an aggregated cluster map that is intended to represent the dominant spatiotemporal signals shared across the input datasets. In practice, however, different choices of detection methods, clustering algorithms, aggregation strategies, and hyperparameters for each of these result in multiple alternative cluster maps. This raises a central question: how can the quality and representativeness of an aggregated cluster map be assessed in a systematic and as objective as possible way?

An aggregated cluster map should capture the dynamics of interest consistently across the underlying datasets. To assess this, an aggregated cluster map can be compared against the original input datasets using quantitative evaluation metrics. In our implementation of the pipeline for analysing abrupt signals, for instance, we employ three complementary metrics that quantify (1) Nonlinearity (NL), (2) Cluster Consistency (CC) and (3) Cluster Spatial Autocorrelation (CSA) of each cluster, which are described in more detail in Appendix B. However, the choice of metrics is flexible and can be adapted or extended depending on the research question at hand. In our case, these metrics are designed to reward consistent clusters with high spatial autocorrelation, expressing a nonlinear mean time series behaviour representing abrupt shifts.

Applying these metrics to an aggregated cluster map yields multiple measured scores. Specifically, each metric is evaluated separately for each input dataset and for each individual cluster, since most evaluation metrics operate at the cluster level rather than on the cluster map as a whole. To enable consistent evaluation of an entire cluster map with respect to a given dataset, we therefore aggregate cluster level scores into a single score per metric at the cluster map level (see Eq. (5)). For a given metric m and dataset d, we compute a weighted average score $s_{m,d}$ from the cluster level scores $s_{k,m,d}$ across all clusters k. Each cluster is weighted by its relative spatial extent w_k, defined as its fraction of the total clustered grid cells in the map. This weighting scheme prevents small clusters from being overemphasised in cluster maps with heterogeneous cluster sizes. The resulting cluster map level score $s_{m,d}$ quantifies how well the aggregated cluster map represents the dynamics of interest in dataset d according to metric m.

$$s_{m,d} = \frac{\sum_{k=1}^{K} w_k \cdot s_{k,m,d}}{\sum_{k=1}^{K} w_k} \tag{5}$$

Having defined how to evaluate one aggregated cluster map with respect to a given dataset and metric, we can now compare and rank multiple alternative cluster maps that arise from different methodological choices and parameter settings along the pipeline. In practice, however, combining multiple datasets and evaluation metrics results in a large number of scores for each candidate cluster map (see Figure 4). To derive a meaningful ranking, these scores must be integrated in a principled and transparent manner. To this end, we use the Analytic Hierarchy Process (AHP), a multi-criteria decision analysis method based on pairwise comparisons (Saaty, 1987). AHP has previously been applied in machine learning and model-evaluation contexts, e.g. by Akogul and Erisoglu (2017) and by Peng et al. (2011). The method allows users to combine quantitative evaluation scores with explicit priority judgments across datasets and metrics, enabling transparent and reproducible selection of aggregated cluster maps.

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

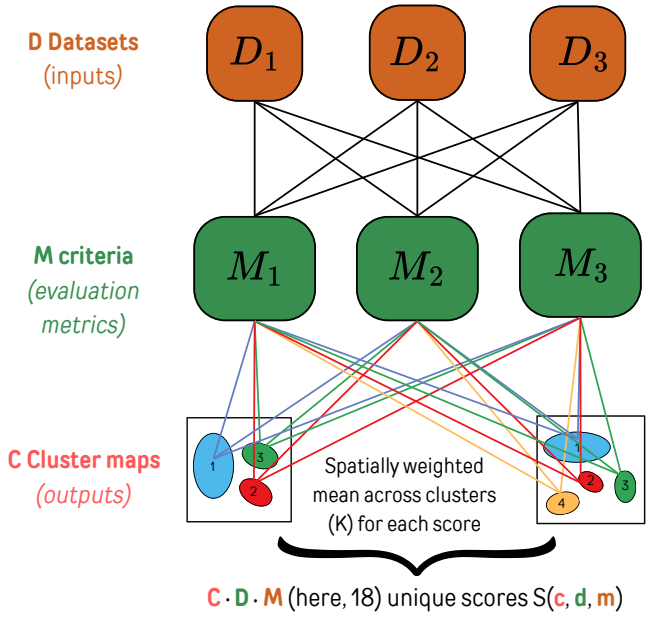

Figure 4. Diagram demonstrating how we score the ability of the final aggregated cluster maps to represent the dynamics of interest. In our evaluation approach, each cluster map is compared against the original datasets using multiple performance metrics. A spatially weighted mean then aggregates the metric values from individual clusters to the cluster-map level, yielding one score per cluster map, dataset, and metric.

In our AHP implementation, users first specify *a priori* priority weights for datasets w_d and evaluation metrics w_m using pairwise comparisons based on Saaty's fundamental scale of importance (Saaty, 1977). These priorities allow users to explicitly encode existing knowledge about the relative importance or reliability of different datasets or metrics. For example, in a Model Intercomparison Project, it may already be known from previous evaluation studies that certain models reproduce key processes or historical observations more accurately than others for a given variable (Watterson et al., 2014; Eyring et al., 2019; Baker and Spracklen, 2022; Heinicke et al., 2022; van Westen and Dijkstra, 2024). Rather than excluding less reliable models altogether, AHP allows such prior knowledge to be incorporated by assigning higher priority weights to datasets that are considered more representative of the processes under investigation. In this way, all datasets remain part of the analysis, while their influence on the final ranking is adjusted transparently. If no dataset or metric is prioritised, the weights can be set to one, ensuring equal contribution across all components.

The AHP-based ranking of cluster maps based on all the scores involved proceeds in three successive steps, moving from dataset-level comparisons to a final cluster-map–level score.

1. **Dataset-level pairwise comparisons per metric and cluster map (see Eq. (6))**

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

For a given cluster map $c \in C$ and metric $m \in M$, we construct a pairwise matrix across all datasets $d \in D$ by comparing the corresponding scores $s_{c,d,m}$. Each matrix element in row i, column j is defined as $\frac{s_{c,i,m}}{s_{c,j,m}}$, ensuring reciprocity (i.e. the $i,j.$ element equals the inverse of the j,i element). The principal eigenvector of this matrix then yields a so-called Relative Importance Vector (RIV) that expresses the relative performance of a cluster map across datasets for the given metric. Repeating this step for all cluster maps and metrics, results in $C \cdot M$ dataset level RIVs.

$$\mathbf{S} = \begin{pmatrix} \frac{s_{c,d1,m}}{s_{c,d1,m}} & \frac{s_{c,d1,m}}{s_{c,d2,m}} \\ \frac{s_{c,d2,m}}{s_{c,d1,m}} & \frac{s_{c,d2,m}}{s_{c,d2,m}} \end{pmatrix}, \quad \mathbf{s} \cdot \mathbf{RIV} = \lambda_{\max} \cdot \mathbf{RIV}, \quad \mathbf{RIV} = \begin{pmatrix} \mathrm{RIV}_{c,d1,m} \\ \mathrm{RIV}_{c,d2,m} \end{pmatrix} \tag{6}$$

2. **Aggregation across datasets per metric (see Eq. (7))**

The dataset level RIVs are then combined using the predefined dataset priority weights w_d. Each dataset's contribution is scaled by its priority, e.g. based on historical performance or reliability, and the weighted RIVs are summed to obtain a single cluster map level RIV for each metric. This step ensures that datasets deemed more important or representative *a priori* have a proportionally larger influence on the aggregated score.

$$\mathbf{RIV} = \begin{pmatrix} \mathrm{RIV}_{c1,m} \\ \mathrm{RIV}_{c2,m} \end{pmatrix} = \begin{pmatrix} \mathrm{RIV}_{c1,d1,m} \cdot w_{d1} + \mathrm{RIV}_{c1,d2,m} \cdot w_{d2} + \mathrm{RIV}_{c1,d3,m} \cdot w_{d3} \\ \mathrm{RIV}_{c2,d1,m} \cdot w_{d1} + \mathrm{RIV}_{c2,d2,m} \cdot w_{d2} + \mathrm{RIV}_{c2,d3,m} \cdot w_{d3} \end{pmatrix} \tag{7}$$

3. **Aggregation across metrics (see Eq. (8))**

Finally, the cluster map level RIVs obtained for each metric are aggregated using the metric priority weights w_m. These weights reflect *a priori* judgments about the relative importance of different metrics within the evaluation protocol. For example, if spatial consistency is considered more important than temporal coherence, this preference can be explicitly encoded through such metric priority weights. The weighted sum across metrics yields a final RIV value for each cluster map, representing a single composite score that accounts for both dataset and metric importance and representation. Based on the final RIV derived from Eq.(8), users can rank the candidate cluster maps and select which results to report or include in subsequent analyses.

$$\mathbf{RIV} = \begin{pmatrix} \mathrm{RIV}_{c1} \\ \mathrm{RIV}_{c2} \end{pmatrix} = \begin{pmatrix} \mathrm{RIV}_{c1,m1} \cdot w_{m1} + \mathrm{RIV}_{c1,m2} \cdot w_{m2} + \mathrm{RIV}_{c1,m3} \cdot w_{m3} \\ \mathrm{RIV}_{c2,m1} \cdot w_{m1} + \mathrm{RIV}_{c2,m2} \cdot w_{m2} + \mathrm{RIV}_{c2,m3} \cdot w_{m3} \end{pmatrix} \tag{8}$$

AHP differs from a simple weighted scoring approach in that it operates on relative rather than absolute scores and enforces consistency by analysing pairwise comparison matrices. This makes the ranking less sensitive to differences in scale between metrics and datasets, and allows users to combine quantitative scores with explicit *a priori* judgments about importance. In contrast, directly weighting and summing raw scores implicitly assumes commensurability across metrics and datasets, which is often difficult to justify in unsupervised, multi-metric settings. AHP therefore provides a more structured and transparent framework for selecting among alternative cluster maps.

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

4 Case study

In this section, we apply the aggregation and evaluation methods introduced above to illustrate their functionality and practical relevance. We integrate them into the pipeline shown in Figure 1 and employ the evaluation protocol from Section 3 to identify the most meaningful cluster maps produced under different parameter choices. We present two cluster maps as illustrative case studies. The first map, based on the DCAC approach, shows how signals from different variables within the same model can be aggregated (Section 4.2.1, Figure 5). The second map, based on DAC, illustrates how signals from the same variable simulated by different models can be combined (Section 4.2.2, Figure 6). Although we demonstrate the aggregation methods for two distinct applications, both approaches can in practice be applied to any cross-dataset aggregation task in the context of Model Intercomparison Projects.

For this case study, we implement the aggregation and evaluation methods using the TOAD python package (Harteg et al., 2026) and apply these methods to CMIP6 (6th phase of the Coupled Model Intercomparison Project) data to investigate climate-change-induced abrupt vegetation shifts in the Amazon rainforest. The remainder of this section is structured as follows: Section 4.1 details the implementation of the methods within an open-source "detect-then-cluster" pipeline, while Section 4.2 both introduces the CMIP6 data and explains how we applied the methods to study Amazon rainforest dieback.

4.1 Implementation in a "detect-then-cluster" pipeline

To demonstrate functionality, we implement the methods in TOAD (Tipping and Other Abrupt Events Detector), an open-source Python package developed by Harteg et al. (2026). It provides a streamlined, data-driven pipeline for detecting regions exhibiting dynamically coherent behaviour at different spatial and temporal scales by combining time series analysis (detection) and spatiotemporal clustering tools. The methodology facilitates systematic multimodel and multivariate assessments of tipping risks by automatically identifying where and when abrupt shifts occur in spatially connected regions. This methodology has been developed alongside the establishment of the TIpping Points Modelling Intercomparison Project (TIPMIP) (Winkelmann et al., 2025). The resulting clusters enable a more fine-grained analysis of large-scale climate tipping elements such as the West Antarctic Ice Sheet or Amazon rainforest than earlier assessments (Lenton et al., 2008; Armstrong McKay et al., 2022) by delineating subsystems that respond similarly to external forcing. TOAD is designed as a flexible, plug-and-play toolkit encompassing various algorithms for each step in the pipeline. For the present case study, we select one specific detection algorithm and one clustering algorithm.

To detect abrupt shifts in per-grid-cell time series we use ASDETECT (Boulton and Lenton, 2019). The algorithm identifies abrupt changes in a time series by scanning across segments of different lengths and flagging periods where the local gradient deviates strongly from the median gradient across all segments (using a MAD-based threshold). The signed, normalised frequency of such anomalous gradients across all segment lengths yields a continuous detection time series (dts) with values between –1 and +1, indicating the relative strength and direction of abrupt shifts. For detailed methodological background and validation, we refer to Boulton and Lenton (2019). ASDETECT meets the key requirement for the time series analysis step in

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

our workflow (Figure 1) by preserving the dimensions of the original data. In other words, it transforms the time series into detection time series (dts), where each value indicates whether an abrupt shift occurs at that time step in the given grid cell.

To identify cohesive spatiotemporal patterns of abrupt shifts, we cluster the dts using the DBSCAN algorithm as operationalised in TOAD (Harteg et al., 2026). DBSCAN is a density-based clustering tool that requires two hyperparameters (*eps* and *min_samples*), which together determine how the algorithm distinguishes dense regions in the data (Ester et al., 1996; Schubert et al., 2017). Within TOAD, DBSCAN is applied to the spatiotemporal information on abrupt signals derived from the detection time series (dts). Clustering is performed jointly in space and time, such that each cluster represents a group of grid cells and time points that are both spatially connected and temporally co-occurring. The dts values are used as weights in the clustering, such that spatiotemporal points exhibiting stronger abrupt signals contribute more to the formation of dense regions, while those with weak or no detected abruptness are unlikely to meet the density requirements for cluster membership. In this way, DBSCAN identifies dynamically connected regions of abrupt change.

The streamlined procedure from time series analysis to clustering in TOAD allows the implementation of both DAC and DCAC as cross-dataset aggregation methods. After running ASDETECT, DAC aggregates multiple dts resulting from various datasets into a single overarching series using either the mean, median, or MAM aggregator function which is then used as input in DBSCAN to perform clustering with varying density hyperparameters (see section 2.1). DCAC, in contrast, first generates several cluster maps for each individual dts using DBSCAN with varying density hyperparameters and creates an overarching consensus time series capturing how often each data point is clustered among the cluster maps. It then performs a second-stage DBSCAN clustering on these consensus time series to obtain an aggregated cluster map (see section 2.2).

Finally, the Analytic Hierarchy Process (AHP) framework described in section 3 systematically evaluates the cluster maps produced by varying DBSCAN parameters. For this case study, we use three complementary evaluation metrics: Nonlinearity (NL), Cluster Spatial Autocorrelation (CSA), and Cluster Consistency (CC). NL quantifies whether a cluster exhibits a synchronised, abrupt event over time relative to the surrounding unclustered grid cells. CSA measures the average dynamical similarity among grid cells within a cluster across the full time domain, capturing the degree of internal temporal coherence. CC evaluates whether a cluster forms a single dynamically coherent unit or can be decomposed into internally distinct sub-clusters. Further methodological details are provided in Appendix B.

4.2 Aggregating abrupt vegetation shifts in the Amazon rainforest

The Amazon rainforest has a vital role in regional moisture recycling (Zemp et al., 2014; Staal et al., 2018), global carbon storage (Pan et al., 2024), and biodiversity (Cardoso et al., 2017). Yet, it is increasingly threatened by deforestation and climate change, with evidence of declining carbon uptake and regional shifts to a net carbon source (Hubau et al., 2020; Gatti et al., 2021). Moreover, self-perpetuating positive feedback mechanisms such as fire-vegetation (Drüke et al., 2021) and moisture-vegetation (Zemp et al., 2017; Staal et al., 2018) interactions may amplify these human disturbances, potentially leading to irreversible and abrupt forest loss. This raises concern that under future warming and further deforestation the Amazon rainforest may at least partially tip to an alternative, degraded stable state beyond critical thresholds (Armstrong McKay et al., 2022; Flores et al., 2024). Given its complex, spatially heterogeneous nature, substantial uncertainty remains regarding the

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

existence, spatial scale and precise thresholds of such Amazon tipping (Armstrong McKay et al., 2022; Brando et al., 2025), with some Earth system models suggesting localised tipping but providing inconsistent evidence for regional or continental-scale dieback (Parry et al., 2022).

A major limitation of past studies is that they often analyse single variables or individual models in isolation, for example, focusing only on vegetation carbon (Parry et al., 2022). Yet, the state of complex ecosystems like the Amazon emerges from multiple, interacting variables and processes that can hardly be captured through a univariate lens. This multivariate complexity makes the Amazon an ideal case study for our aggregation framework, which integrates signals across both models and variables. By combining these signals, the approach can reveal more robust patterns of abrupt shifts and demonstrate the practical relevance of multivariate, multimodel analysis in assessing ecosystem transitions.

For this case study, we select six models from CMIP6 (see Table 2) that include dynamic vegetation in their land surface components, following previous studies by Parry et al. (2022) and Terpstra et al. (2025). We use the 1pctCO2 experiments, in which CO2 is increased by 1% per year over 150 years, to study climate change–induced vegetation shifts (Eyring et al., 2016). While these simulations do not provide equilibrium vegetation responses required for strict tipping point analysis, they allow examination of transient, abrupt shifts under a consistent, monotonic forcing. The models differ in precipitation response, fire representation, and moisture recycling, reflecting a diverse range of possible Amazon representations.

To enable consistent analysis across models and variables, we regrid all data to 1° resolution, clip them to the Amazon basin, aggregate monthly data to annual means, and select ten vegetation-related variables (see Table 1). To account for differences in transient climate response between the models, the model years of each individual model are mapped to Global Warming Levels (GWLs) derived from their simulated annual global mean temperatures. Since internal variability causes fluctuations in the annual GWLs, a cubic regression is applied to each model separately to smooth the mapping and ensure that each successive time step corresponds to a higher GWL than the previous one. In the following, we constrain the GWLs for each model to the 0-5°C range corresponding to the maximum interval for which all models provide data points. The resulting datasets provide the basis for applying TOAD to identify abrupt vegetation shifts as a function of GWL (rather than model time). We apply the pipeline first across multiple variables within one model and subsequently on one output variable across multiple models.

4.2.1 Aggregation across variables for one model (multivariate)

Rather than asking which individual variables undergo abrupt shifts, we ask where a shared signal emerges across them, a sign that multiple components of the Amazon ecosystem are responding coherently to forcing, rather than isolated processes changing independently. This extends the notion of spatiotemporal connectivity to the variable dimension, such that a region is identified even if only a subset of variables responds clearly or simultaneously. Figure 5 presents a cluster map obtained with TOAD using the Detect–Cluster–Aggregate-Cluster (DCAC) approach across multiple vegetation variables (Table 1) for the CMIP6 GFDL-ESM4 model. We focus on GFDL-ESM4 in this case study because it exhibits a pronounced abrupt vegetation decline, consistent with earlier findings by Parry et al. (2022). For this model, the variables *rGrowth*, *fVegLitter*, and *baresoilFrac* are not available. Since we are interested in shared spatiotemporal signals rather than the direction of change in individual variables, absolute values of the detection signals are used, ensuring that strong shifts in any variable contribute to

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

the consensus regardless of sign. Figure C1 shows the spatial distribution of the underlying consensus time series (cts) that served as input for the final clustering stage in the DCAC pipeline.

Table 1. Overview of CMIP6 output variables used in our case study to describe the vegetation of the Amazon Rainforest.

Variable	Unit	Description
cVeg	kgC/m^2	Carbon Mass in Vegetation
cSoil	kgC/m^2	Carbon Mass in Model Soil Pool
treeFrac	%	Tree Cover Percentage
grassFrac	%	Natural Grass Area Percentage
baresoilFrac	%	Bare Soil Percentage Area Coverage
fVegLitter	kgC/m^2/s	Total Carbon Mass Flux from Vegetation to Litter
fFire	kgC/m^2/s	Carbon Mass Flux into Atmosphere Due to CO2 Emission from Fire Excluding Land-Use Change
gpp	kgC/m^2/s	Carbon Mass Flux out of Atmosphere Due to Gross Primary Production on Land
lai	m^2/m^2	Leaf Area Index
rGrowth	kgC/m^2/s	Total Autotrophic Respiration on Land as Carbon Mass Flux

Using the evaluation framework described in Section 3, with priority weights for both performance metrics and vegetation variables specified in Appendix D, this cluster map was identified as the optimal DCAC result among 410 candidate cluster maps. These candidates were generated by varying the clustering parameters *epsilon* (0.15–0.24 in steps of 0.01) and *min_samples* (15–55 in steps of 1) in the final clustering stage. For comparison, we also applied the Detect–Aggregate–Cluster (DAC) approach using median, mean, and MAM aggregators across the same parameter ranges. The corresponding optimal DAC cluster maps are shown in Figure C2. However, for this application DAC exhibits high sensitivity to the aggregation choice, yielding either overly extensive spatial clusters (e.g. DAC–MAM) or overly constrained ones (e.g. DAC–mean). We therefore focus the subsequent analysis on the DCAC-derived cluster map.

The map consists of three clusters: cluster 1 in the south (1.55–2.25 °C global warming), cluster 2 in the east (0.05–1.3 °C), and cluster 3 in the central rainforest (2.10–3.0 °C). While clusters 1 and 3 indicate abrupt behaviour, cluster 2 does not show visually clear abrupt shifts. Both clusters 1 and 3 reveal abrupt declines in vegetation carbon (*cVeg*), but the patterns in the other variables differ markedly, implying that the ecosystem response and underyling dynamics are qualitatively different.

Cluster 1, in the southern Amazon (Bolivia), signals a localised abrupt transition from its prior state to a more degraded ecosystem state under increasing forcing. At around 1.5 °C of global warming, TOAD detects a drastic forest dieback: tree cover *(treeFrac)* declines from ~40% on average (up to ~75% in some grid cells) to nearly 0%, while grass cover *(grassFrac)* increases from ~40% to ~75%. Carbon pools (*cVeg and cSoil*) likewise approach zero, accompanied by a rising fire flux (*fFire*). Together, these dynamics indicate a shift toward a more savanna-like ecosystem state.

In contrast, cluster 3, reflects a large-scale disturbance or extreme event rather than a critical transition in the center of the rainforest (Brazil). Here, a pronounced drop of ~50% in vegetation carbon (*cVeg*) coincides with synchronous but transient disturbances in other variables: temporary dips in canopy density (*LAI*), ecosystem productivity (*GPP*), and vegetation cover

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

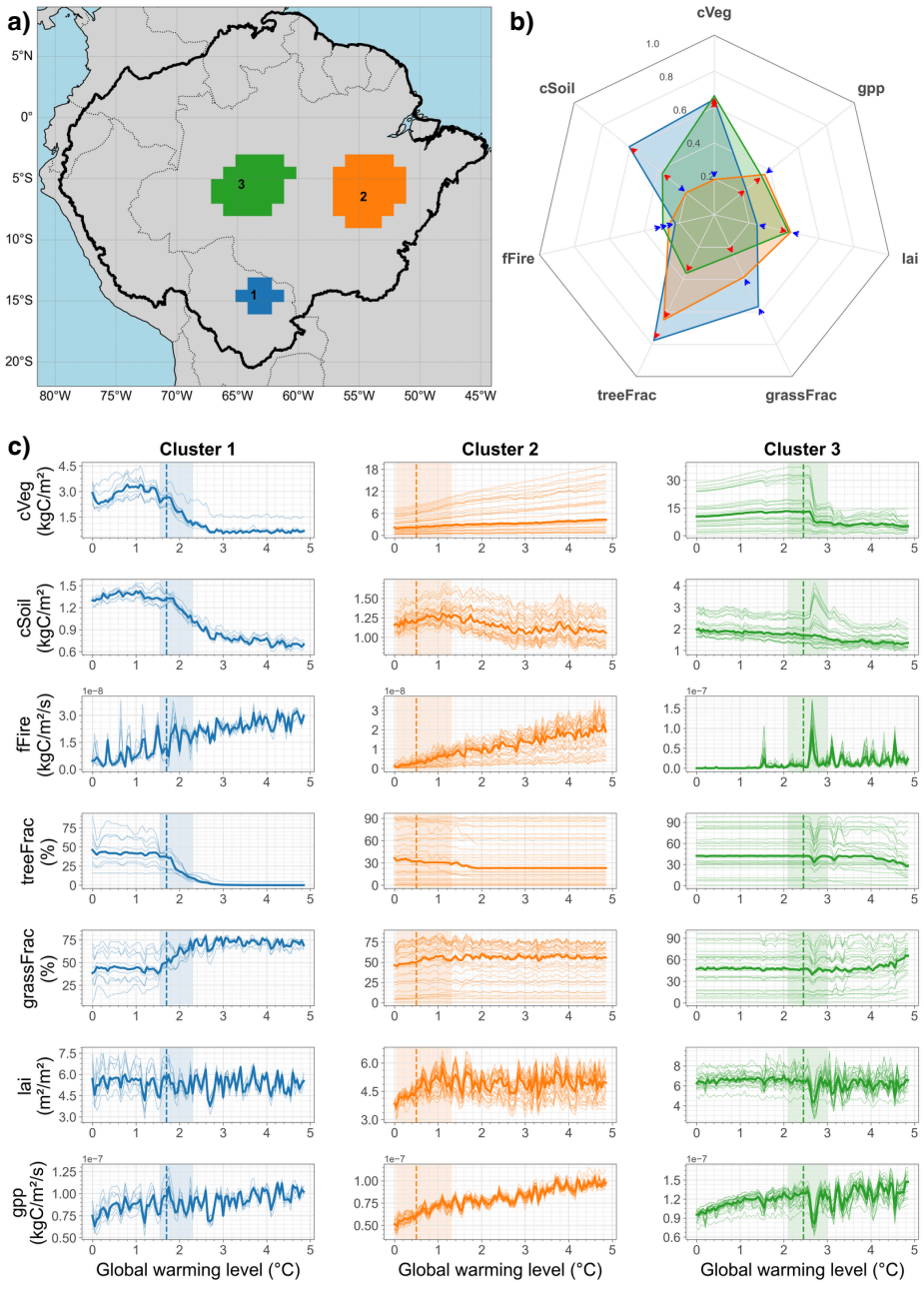

Figure 5. Aggregated TOAD results for GFDL-ESM4 using the Detect-Cluster-Aggregate-Cluster (DCAC) approach. First-stage clustering used epsilon = 0.10–0.30 (steps of 0.01) and min-samples = 10–30 (steps of 1); second-stage clustering used epsilon = 0.15 and min-samples = 24. (a) Three Amazon clusters are identified (south, central, east). (b) Mean absolute detection time series within the clustered GWL range for each contributing variable and dominant shift direction (red triangles = negative, blue = positive). (c) Original time series for grid points within each cluster; thick lines denote the spatial median. Colored bands indicate the GWL range of cluster emergence, with the dashed line marking the maximum detection signal.

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

fractions *(treeFrac, grassFrac*), along with a temporary peak in fire flux (*fFire*). Notably, similar, though less intense co-occurring disturbances appear both before and after this event, suggesting repeated perturbations rather than a permanent shift to an alternative state.

This case study demonstrates that focusing on a single variable, such as *cVeg*, risks overlooking structurally different types of ecosystem change. Assuming that different parts of the system respond uniformly based on one indicator can understate the complexity and spatial heterogeneity in feedback processes under climate forcing. A multivariate perspective thus provides a more comprehensive view of synchronous and distinct dynamics.

4.2.2 Aggregation across models for one variable (multimodel)

The aim here is to extend the notion of spatiotemporal connectivity to the model dimension, identifying regions where a shared signal emerges across models as a sign of robust, ensemble-wide evidence for abrupt vegetation change, rather than dynamics that are specific to individual models or their configurations. Figure 6 presents a cluster map generated with TOAD using the Detect-Aggregate-Cluster (DAC) approach, applied to vegetation carbon (*cVeg*) across multiple CMIP6 models (see Table 2). For this demonstrative example, *cVeg* was chosen as a single vegetation variable to enable comparison with the study from Parry et al. (2022). To focus the analysis on Amazon dieback patterns (negative *cVeg* shifts), we assign a value of zero to positive values from the aggregated detection time series (dts). Figure C1 shows the spatial distribution of the aggregated dts that served as input for the subsequent clustering in the DAC pipeline.

Table 2. Overview of CMIP6 models and their associated land surface models used in this case study.

CMIP6 Model	Land Surface Model
GFDL-ESM4	LM4.1
MPI-ESM1-2-LR	JSBACH
NorCPM1	CLM4.0
TaiESM1	CLM4.0
SAM0-UNICON	CLM4.0
UKESM1-0-LL	JULES

The cluster map presented in Figure 6 was obtained using the Magnitude-Adjusted Mean (MAM) aggregation within the DAC pipeline and selected among 779 candidate cluster maps following the evaluation protocol described in Section 3. These candidates were generated by varying the clustering parameters *epsilon* (0.17–0.35 in steps of 0.01) and *min_samples* (15–55 in steps of 1). The evaluation drew on priority weights for the performance metrics, as detailed in Appendix D, whereas equal weights were applied across the models. Since only 1 variable is considered in this case, no priority weights for variables are used in the evaluation.

For comparison, optimal cluster maps were also derived using Detect-Cluster-Aggregate-Cluster (DCAC) and DAC with mean and median aggregation (see Fig C3). While the DAC-MAM result identifies multiple spatially coherent clusters, the alternative aggregation strategies recover only subsets of these regions and do not reproduce all spatial features detected in

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

DAC-MAM simultaneously. This fragmentation across methods suggests that MAM provides a more complete and conservative representation of concurrent abrupt shifts across models. A likely explanation is that, in the presence of models exhibiting weak or no abrupt signals, mean, median and DCAC aggregation tend to dampen localised but pronounced shifts, whereas MAM preserves and highlights strong signals detected in at least a subset of the models. In the following, we therefore focus on the cluster map derived using DAC-MAM.

The cluster map is composed of three clusters. Cluster 1 stretches from the center to the east of the forest, covering a large part of Brazil. The cluster exhibits a downward *cVeg* shift from 3.4°C to 4.75°C of global warming in the TailESM1 model. To a lesser extent, we also find abrupt shifts in this clustered range for SAM0-UNICON, occurring later from 4.3°C. For the other models, we do not find consistent shifts for cluster 1.

Cluster 2 is situated in the Brazilian center of the rainforest and emerges between 2.2°C and 3.2°C. In this cluster we detect a clear, and consistently abrupt *cVeg* shift for GFDL-ESM4. Although the other models do not exhibit any downward shifts in this cluster, we do notice that for SAM0-UNICON the increasing *cVeg* reaches a plateau, and that TailESM1 shifts to higher *cVeg* from 2.8°C onward.

Cluster 3 covers the border region between Venezuela and Guyana in the North and occurs between 3.8°C and 4.6°C. Three of the six models indicate negative *cVeg* shifts for this clustered range, namely: SAM0-UNICON, NorCPM1, and TailESM1. The other models show increasing *cVeg*.

The case study cluster map, generated by combining multiple models using the DAC method with MAM aggregation, demonstrates that a structured aggregation approach can help pinpoint regions of model consensus or divergence. For example, cluster 3 exhibits a high degree of agreement among models, with approximately 50% indicating a negative shift, whereas cluster 2 displays more heterogeneous behaviour: one model shows an abrupt decline, another reaches a plateau, and yet a third exhibits an abrupt positive shift. These results highlight that, when analysing multimodel datasets, detected signals should be integrated systematically in a data-driven manner to robustly assess tendencies for models to converge or diverge.

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

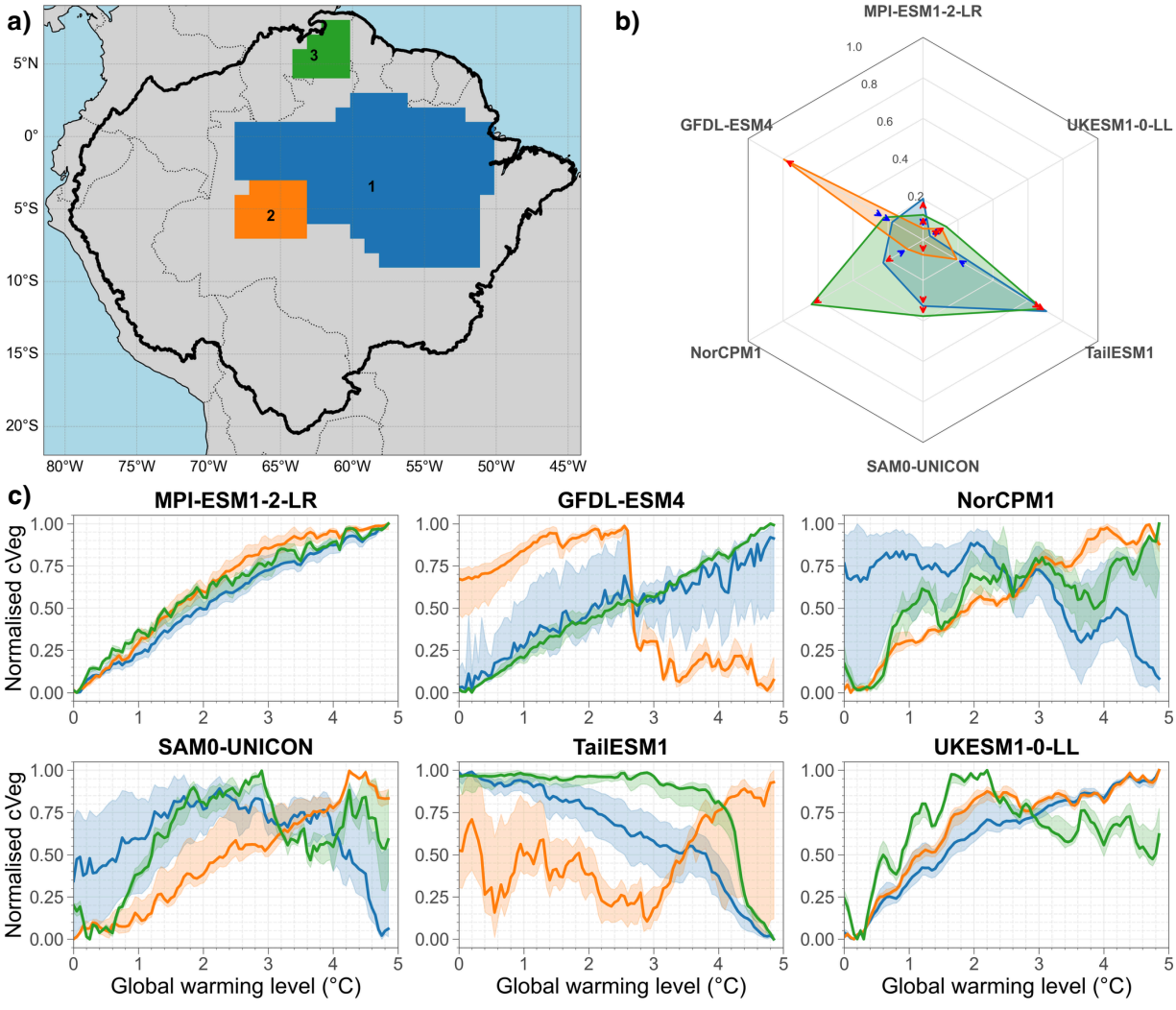

Figure 6. Aggregated TOAD results for vegetation carbon (cVeg) across models using Detect-Aggregate-Cluster (DAC) with MAM aggregation. Clustering used epsilon = 0.25 and min-samples = 46. (a) Three clusters emerge in the central, northern, and eastern Amazon basin. (b) Mean absolute detection time series (dts) within the clustered GWL range for each contributing model and dominant shift direction (red triangles = negative, blue = positive). (c) Min–max normalised multimodel dynamics of clustered grid cells across GWLs; lines denote the cluster median and shaded areas the interquartile range (25th–75th percentile).

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

5 Discussion and outlook

MIPs generate a large ensemble of datasets well suited for detecting Earth-system dynamics, such as abrupt shifts, with advanced data-mining tools, but realising their full potential calls for systematic workflows to aggregate and synthesise detected signals co-occurring across models and variables. In this paper, we addressed this by extending the spatiotemporal connectivity that recent "detect-then-cluster" approaches have operationalised within individual datasets to the model and variable dimensions, introducing two aggregation strategies: Detect-Cluster-Aggregate-Cluster (DCAC) and Detect-Aggregate-Cluster (DAC). Alongside these, we introduced a customised evaluation framework based on the Analytic Hierarchy Process, enabling systematic and transparent selection of the most representative aggregated cluster maps. We implemented these methods using the TOAD v1.0 framework (Harteg et al., 2026), forming part of its conceptual basis. Applying these methods to the case of abrupt Amazon dieback, we demonstrated that structured aggregation not only pinpoints regions where models converge or diverge, but also reveals heterogeneous ecosystem dynamics that single-variable analyses would miss. Overall, this underscores the added value of coordinated multivariate and multimodel analyses for robustly identifying and interpreting abrupt shifts or potential tipping behaviour in the Earth system.

Our illustrative case study broadly reproduces findings from earlier work scanning CMIP6 data for abrupt shifts in the Amazon rainforest (Parry et al., 2022; Terpstra et al., 2025), but the aggregation methods uncover structural patterns that were previously overlooked. For example, whereas Parry et al. (2022) reported localized dieback in the northern forest without clear, shared thresholds across models, our multimodel clustering identifies synchronous shifts in subsets of models at consistent warming levels, exposing hidden cross-model alignment. While Parry et al. (2022) illustrated the localised abrupt decline in vegetation carbon in GFDL-ESM4 using a single exemplary grid cell in the central Amazon, our approach identifies a spatially coherent cluster of grid cells undergoing the transition. Importantly, this cluster-level perspective situates the event within a broader multivariate and multimodel context: the abrupt vegetation carbon decline coincides with abrupt events in other variables (Figure 5, cluster 3) or in other models (Figure 6, cluster 2), but these accompanying signals are not uniformly persistent. Other variables exhibit only transient disturbances, and some models display contrasting behaviour (e.g., a concurrent positive shift in TailESM1), suggesting that the detected abrupt shift may reflect model-specific or variable-specific dynamics rather than a robust, system-wide regime transition. Similarly, although Terpstra et al. (2025) examined multiple variables, each was analysed independently, resulting in spatially fragmented patterns. In contrast, our multivariate clustering identifies larger, more coherent regions of ecosystem change. These comparisons suggest that structured aggregation not only recovers known signals but also clarifies their spatial coherence, cross-variable consistency, and cross-model robustness, thereby providing a more systematic basis for distinguishing convergent tipping-like behaviour from model-specific or transient dynamics.

Although we only demonstrated the case of detecting abrupt vegetation shifts, our aggregation methods could equally be applied to other "detect-then-cluster" tasks. One example is threshold-based shifts, such as identifying the point at which the cumulative net primary productivity (NPP) switches sign, marking a transition from a transient carbon sink to source (Wieder et al., 2015; Duffy et al., 2021). Another relevant application concerns the detection of extreme events, which can manifest

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

as persistent shifts when time series are temporally integrated, as discussed by Bathiany et al. (2024). A concrete example is ClimBurst, a "detect–then–cluster" framework for identifying climatological anomalies, which could be naturally extended through our aggregation approach to enable systematic multimodel or multivariate analyses (Brouillet et al., 2025). These examples illustrate that our aggregation approach is broadly applicable across different types of Earth-system dynamics, as long as its key methodological requirements are met.

Although DCAC and DAC can be used interchangeably to detect spatially coherent Earth-system dynamics across diverse datasets, they are not equivalent in practice. The key difference is their sensitivity to ensemble size: DCAC is generally more robust when the number of datasets is small, whereas DAC becomes more sensitive when there are limited datasets to aggregate. This difference arises from how aggregation is performed. In DCAC, each individual detection time series (dts) is first clustered across a wide range of hyperparameter combinations, producing multiple cluster maps per dataset. These maps are then combined into a consensus time series (cts) that reflects how frequently a given grid cell is classified as part of a cluster and is used as input for the final clustering step. Because the cts aggregates many clustering realizations, it tends to exhibit smoother and more stable spatiotemporal patterns. In contrast, DAC aggregates the individual dts directly using an aggregation function, with each dataset contributing only once before clustering is applied. As a result, when the ensemble size is small, the clustering input in DAC is comparatively coarse and more sensitive to individual-model variability. Figure C1 illustrates this distinction, showing sharper spatiotemporal contrasts in the clustering input for DAC compared to the smoother consensus patterns produced by DCAC. Nevertheless, DCAC can become computationally expensive when applied to large ensembles, as each dataset must first undergo both time series analysis and multiple clustering runs before aggregation is performed, and the procedure introduces additional parameter choices.

DAC requires the explicit selection of an aggregation function, and this choice can substantially shape the resulting cluster patterns. Different functions weigh dataset contributions differently: some emphasise widespread agreement across the ensemble, while others retain sensitivity to strong signals emerging in only a subset of models or variables. For example, aggregation schemes such as the mean or median tend to form clusters that are supported by many datasets, thereby favoring broadly agreed-upon dynamics. In contrast, approaches such as MAM remain sensitive to pronounced abrupt signals even if they occur in only a few datasets. Consequently, the aggregation step determines whether clusters primarily reflect ensemble consensus or also capture less widespread but potentially important dynamics. Although this introduces an element of methodological choice, it also represents a strength of the approach, as it allows the workflow to be aligned with specific research questions and analytical priorities.

The practical effort required to integrate these strategies into existing "detect–then–cluster" pipelines also differs. Depending on the analytical setup, one approach may be more straightforward to implement than the other. For instance, DCAC naturally extends a univariate TOAD workflow: individual detection time series and cluster maps can be generated independently for each dataset and subsequently aggregated, without modifying the internal structure of the "detect-then-cluster" pipeline. In contrast, DAC requires restructuring the workflow, as it involves constructing a combined detection time series from all individual datasets prior to clustering. This intermediate aggregation step alters the standard processing sequence and therefore demands tighter integration with the existing pipeline.

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

The evaluation framework, introduced to tune the methods in the pipeline and select the most appropriate cluster maps, also comes with advantages and disadvantages. Its key strength lies in its flexibility and generalisability: users can tailor the framework to their case by selecting multiple evaluation metrics of interest capturing different performance aspects and weighting each according to its relative importance. In addition, priority weights can be given to specific datasets, thereby allowing certain models or variables to exert greater influence on the aggregated cluster maps. This feature is particularly valuable in the context of MIPs, where model evaluation studies often reveal that some ensemble members capture specific dynamics more reliably than others when compared to observational benchmarks, especially when considering specific regions or systems (Watterson et al., 2014; Eyring et al., 2019; Baker and Spracklen, 2022; Heinicke et al., 2022; van Westen and Dijkstra, 2024). In this way, qualitative differences across datasets can be systematically translated into quantitative weights that guide the selection of the most suitable cluster map. The main limitation, however, is that the framework only ranks cluster maps generated from a prescribed, often arbitrary set of parameter ranges, rather than actively identifying an optimal configuration of methods and parameters. One way to address this limitation is to incorporate complementary hyperparameter optimisation strategies, such as random search (Bergstra and Bengio, 2012), gradient descent (Curry, 1944) or Bayesian optimisation (Wu et al., 2019) into the evaluation framework.

It is important to note that we have not yet systematically evaluated the novel aggregation methods across a broad spectrum of benchmark datasets. The present study primarily serves to introduce the approaches and illustrate their functionality and relevance. Currently, a comprehensive catalogue of artificial test datasets, systematically extending the simple synthetic data generation employed here, against which methods can be validated, is lacking. This gap is particularly relevant for multivariate and multimodel detection and clustering in ensemble datasets, since it is especially challenging in the absence of a clear data generation protocol to define what patterns across models and variables a robust method should be able to identify. We therefore argue that future work should prioritise the development of such benchmark catalogues, encompassing a wide range of dynamics and patterns representative of those encountered in MIP data settings. This would provide the means for a more thorough evaluation of current and future data-mining methods in the context of Earth system and climate science.

To conclude, this study demonstrates the value of structured approaches to aggregate results from "detect-then-cluster" pipelines across space, time, models, and variables. By implementing DAC and DCAC using the TOAD framework (Harteg et al., 2026) and illustrating their application to abrupt Amazon dieback, we show that these methods provide a systematic way to synthesise spatially coherent signals across models and variables, revealing patterns that would remain hidden in single-variable or single-model analyses. Beyond CMIP6, these approaches are broadly applicable to other MIP contexts, including initiatives such as TIPMIP (Winkelmann et al., 2025) or ISIMIP3b (Frieler et al., 2025). By providing a flexible methodological framework, we hope to inspire further methodological advances, promote systematic benchmarking, and support the development of new tools for the study of complex Earth-system dynamics.

. Code and data availability. The code used for the analysis is available from the corresponding authors upon request. Simulation data for vegetation variables are from CMIP6 (Eyring et al., 2016) and were accessed via the Earth system Grid Federation (ESGF).

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

Appendix A: Synthetic ensemble data generation

To evaluate and demonstrate the two aggregation algorithms introduced in this study under controlled conditions, we construct a set of idealised synthetic datasets. These datasets are designed to emulate a single underlying spatiotemporal abrupt shift event, while incorporating slight variations in when and where the event occurs across realizations. This design mimics the situation in Model Intercomparison Project (MIP) contexts, where individual models (or variables) represent the same physical process with small spatial and temporal discrepancies. Thus, while each synthetic dataset captures the same fundamental signal, it differs subtly in expression. The resulting data are used in Figures 2 and 3.

We first define a prototypical abrupt event that serves as the template for all realizations. This event is represented as a three-dimensional field in which the spatial pattern follows a two-dimensional Gaussian function centered at (0, 0), and the temporal evolution follows a sigmoid function centered at time = 50. The combination produces a coherent, localised increase over time that resembles a gradual approach and subsequent abrupt shift. To capture small-scale variability, Gaussian noise is added to both spatial and temporal dimensions. The point of maximum spatial intensity and midpoint of temporal transition (0, 0, 50) is referred to as the core of the underlying abrupt event.

From this base event, we generate an ensemble of synthetic realizations by applying controlled random perturbations to the spatial and temporal centers of the transition. For each realization, small random shifts are introduced to the x- and y-coordinates as well as to the timing of the transition, thereby producing fields in which the position and onset of the abrupt shift vary slightly between members. In total, 20 independent datasets are produced, each representing a spatially localised transition of the same underlying process, but displaced in space and time. Collectively, these datasets form the synthetic ensemble used to evaluate the two aggregation algorithms, providing a controlled test case for assessing their ability to recover a common signal from spatially and temporally misaligned events.

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

Appendix B: Evaluation metrics

To evaluate the quality of each cluster k in cluster map c for dataset i, we define three complementary metrics: Nonlinearity (NL), Cluster Spatial Autocorrelation (CSA), and Cluster Consistency (CC). Together, these metrics assess whether a cluster captures synchronised, spatially coherent, and internally homogeneous abrupt dynamics.

B1 Nonlinearity (NL)

The Nonlinearity (NL) metric quantifies whether a cluster exhibits a collective abrupt event over time. For each cluster, we first aggregate the time series of its grid cells into a normalised "cluster time series" (Eq. B1). Here, k denotes a cluster, c a cluster map, i a dataset, (x, y) spatial grid cells, and t time. The operators $\min_t$ and $\max_t$ denote the minimum and maximum over the full temporal domain of the aggregated cluster time series.

$$D_{k,c,i}(t) = \frac{\sum_{x,y\in k} D_{k,c,i}(x,y,t) - \min_t\left(\sum_{x,y\in k} D_{k,c,i}(x,y,t)\right)}{\max_t\left(\sum_{x,y\in k} D_{k,c,i}(x,y,t)\right) - \min_t\left(\sum_{x,y\in k} D_{k,c,i}(x,y,t)\right)}. \tag{B1}$$

A linear regression is then fitted to this aggregated series, and its root mean square error (RMSE) is computed (Eq. B2). Here, $\widehat{D}_{k,c,i}(t)$ is the fitted linear trend and T the number of time steps. A large RMSE indicates that the cluster-level dynamics deviate strongly from a linear trend, consistent with abrupt or nonlinear behaviour.

$$\mathrm{RMSE}_{k,c,i} = \sqrt{\frac{1}{T}\sum_{t}^{T}\left(\widehat{D}_{k,c,i}(t) - D_{k,c,i}(t)\right)^2}. \tag{B2}$$

Finally, to ensure that this nonlinearity is specific to the clustered region, the cluster RMSE is divided by the RMSE of the unclustered grid cells (Eq. B3). This yields a relative measure of how strongly nonlinear the cluster behaves compared to the background. Additionally, NL is multiplied by a temporal inclusion factor, which evaluates whether the most abrupt shifts in the clustered grid cells occur within the temporal window ($T_{k,c}$) assigned to the cluster. Here, $|k|$ is the number of grid cells in cluster k. This correction prevents artificially high NL values when the strongest events fall outside the clustered time range. High NL values therefore indicate: strongly nonlinear (abrupt) cluster-level dynamics, few missed abrupt events outside the cluster, temporal alignment between clustered grid cells.

$$\mathrm{NL}_{k,c,i} = \frac{\mathrm{RMSE}^{\mathrm{cluster}}_{k,c,i}}{\mathrm{RMSE}^{\mathrm{unclustered}}_{c,i}} \cdot \frac{1}{|k|}\sum_{(x,y)\in k}\max_{t\in T_{k,c}}\left(\frac{|dts_i(x,y,t)| - \min_t|dts_i(x,y,t)|}{\max_t|dts_i(x,y,t)| - \min_t|dts_i(x,y,t)|}\right). \tag{B3}$$

B2 Cluster Spatial Autocorrelation (CSA)

The Cluster Spatial Autocorrelation (CSA) metric measures the dynamical similarity among grid cells within a cluster. For each pair of grid cells in a cluster, we compute the Pearson correlation of their full time series (Eq. B4). Here, p and q are grid cells within cluster k, $X_{k,c,i}(p,t)$ the time series at grid cell p, and $\overline{X}_{k,c,i}(p)$ its temporal mean.

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

$$r_{k,c,i}(p,q) = \frac{\sum_t \left(X_{k,c,i}(p,t) - \overline{X}_{k,c,i}(p)\right)\left(X_{k,c,i}(q,t) - \overline{X}_{k,c,i}(q)\right)}{\sqrt{\sum_t \left(X_{k,c,i}(p,t) - \overline{X}_{k,c,i}(p)\right)^2}\sqrt{\sum_t \left(X_{k,c,i}(q,t) - \overline{X}_{k,c,i}(q)\right)^2}}. \tag{B4}$$

The correlations are squared to obtain coefficients of determination, which quantify the proportion of shared variance independent of sign. CSA is then defined as the mean of all pairwise coefficients of determination (Eq. B5). Here, $n = |k|$ is the number of grid cells in cluster k, and the sum is taken over all unique grid-cell pairs.

$$\mathrm{CSA}_{k,c,i} = \frac{1}{n(n-1)} \sum_{p,q} r_{k,c,i}(p,q)^2. \tag{B5}$$

CSA therefore captures the average dynamical coherence within a cluster across the entire time domain. High CSA values indicate that clustered grid cells exhibit similar temporal behaviour before, during, and after abrupt events.

B3 Cluster Consistency (CC)

The Cluster Consistency (CC) metric evaluates whether a cluster forms a single coherent unit or contains internally separable sub-clusters. While CSA measures average connectivity, CC specifically detects structural heterogeneity within the cluster.

Using the pairwise coefficient of determination matrix, we perform agglomerative hierarchical clustering with Ward's method. Similarities are converted into distances (Eq. B6) and grid cells are iteratively merged based on minimal increases in within-cluster variance. Here, $d_{k,c,i}(p,q)$ denotes the dynamical distance between grid cells p and q.

$$d_{k,c,i}(p,q) = 1 - r_{k,c,i}(p,q)^2. \tag{B6}$$

We then compute the inconsistency statistic of the final merge, which measures how exceptional the last merging step is compared to earlier ones. If the final merge requires a disproportionately large increase in within-cluster variance, this indicates that the cluster may consist of two dynamically distinct sub-groups. CC is defined as the inverse of this inconsistency statistic (Eq. B7), such that: high CC means that the cluster is internally coherent and difficult to subdivide, and low CC that the cluster likely contains separable sub-clusters. Here, h_{last} denotes the height of the final merge in the dendrogram, h_m the height of merge step m, M the total number of merges, μ_h the mean merge height, and σ_h the standard deviation of merge heights.

$$\mathrm{CC}_{k,c,i} = \left(\frac{h_{\mathrm{last}} - \mu_h}{\sigma_h}\right)^{-1}, \qquad \mu_h = \frac{1}{M}\sum_{m=1}^{M} h_m, \qquad \sigma_h = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(h_m - \mu_h)^2}. \tag{B7}$$

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

Appendix C: Supplementary figures

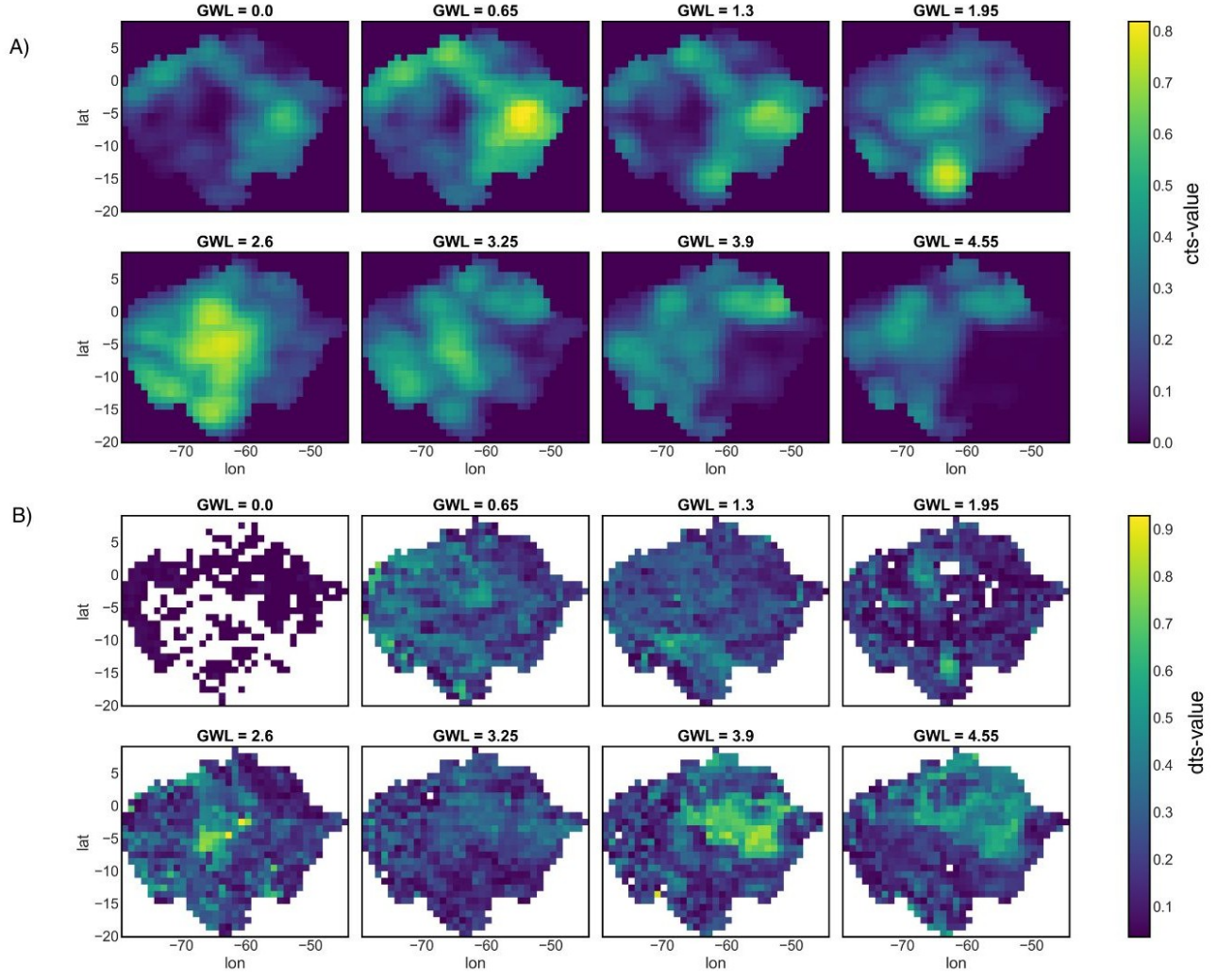

Figure C1. Clustering input plotted for different Global Warming Levels (GWLs) across space (Amazon basin). A) Consensus Time Series (cts) corresponding to the cluster map presented in Figure 5 for GFDL-ESM4 across variables following the Detect-Cluster-Aggregate-Cluster (DCAC) approach. First-stage clustering used epsilon = 0.10–0.30 (steps of 0.01) and min-samples = 10–30 (steps of 1) B) MAM-aggregated Detection Time Series (dts) corresponding to the cluster map presented in Figure 6 for cVeg across models following the Detect-Aggregate-Cluster (DAC) approach.

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

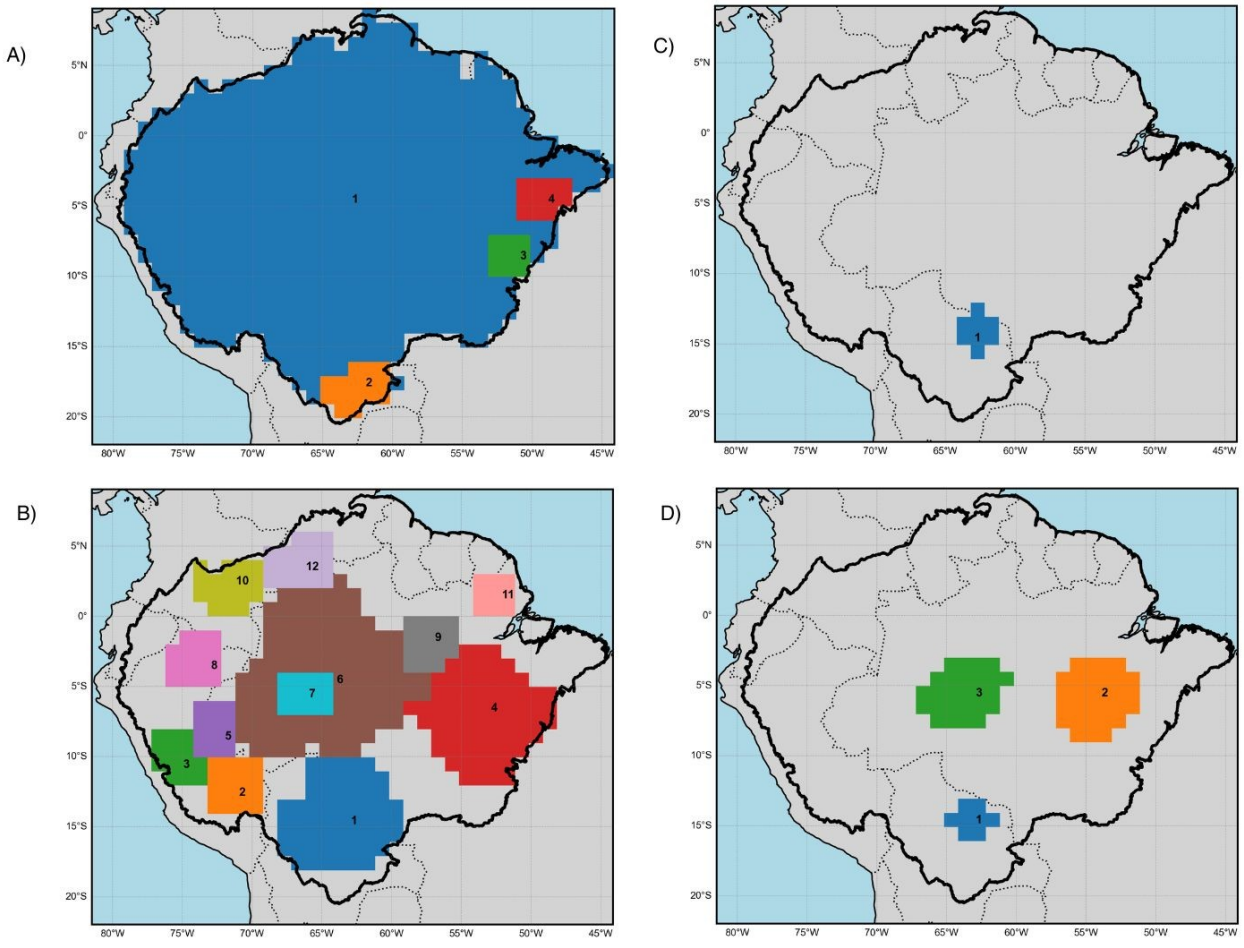

Figure C2. Spatial plots of the highest-ranked cluster maps per aggregation method for detecting abrupt shifts in GFDL-ESM4 across variables (see Figure 5). A) Cluster map for Detect-Aggregate-Cluster (DAC) using the MAM aggregation function and epsilon = 0.24 and min-samples = 36. B) Cluster map for DAC using the median aggregation function and epsilon = 0.16 and min-samples = 24. C) Cluster map for DAC using the mean aggregation function and epsilon = 0.18 and min-samples = 15. D) Cluster map for Detect-Cluster-Aggregate-Cluster (DCAC) using epsilon = 0.15 and min-samples = 24 for the final clustering step. Note that this last result is discussed in more detail in the main text (see Figure 5).

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

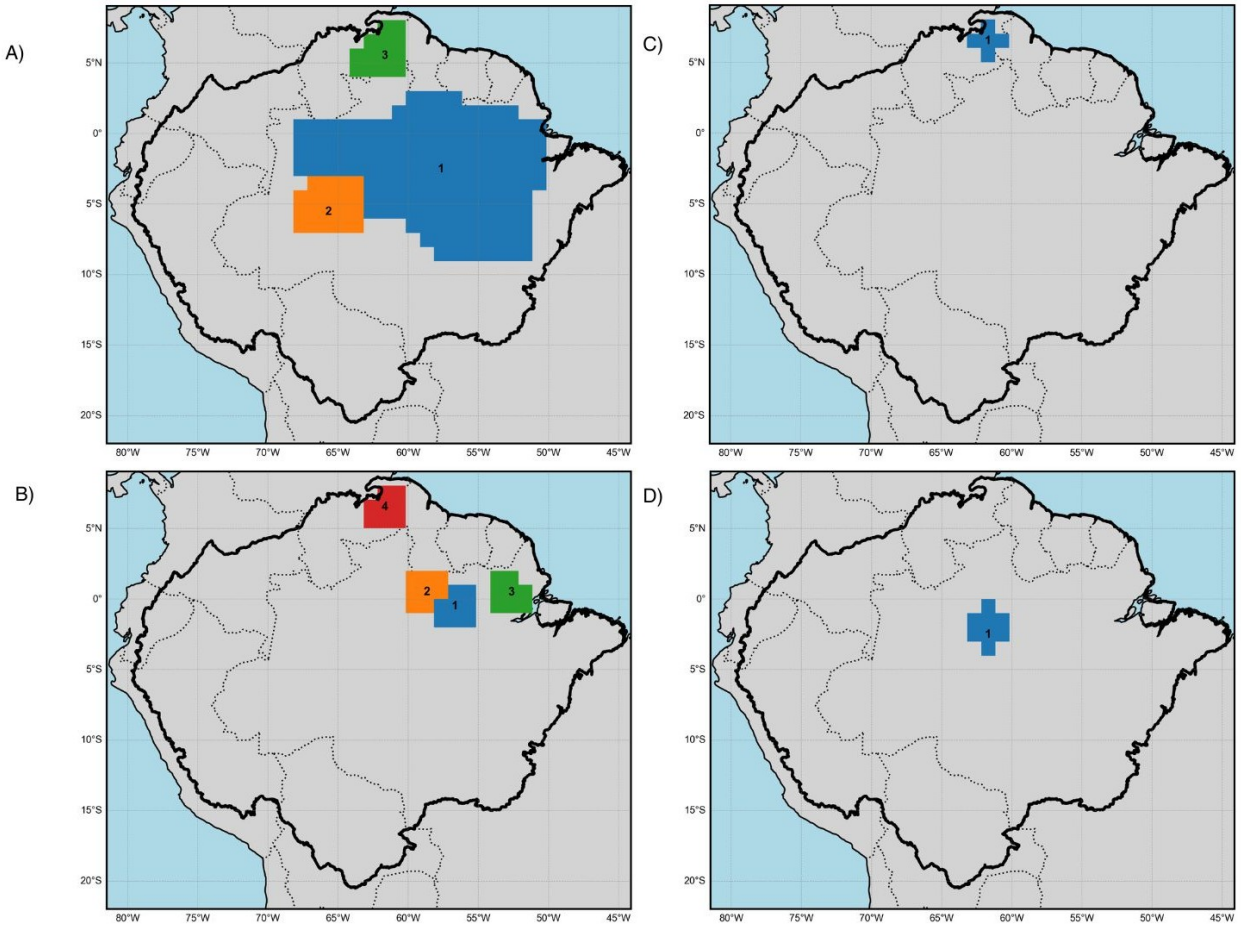

Figure C3. Spatial plots of the highest-ranked cluster maps per aggregation method for detecting abrupt shifts in cVeg across CMIP6 models (see Figure 6). A) Cluster map for Detect-Aggregate-Cluster (DAC) using the MAM aggregation function and epsilon = 0.25 and min-samples = 46. Note that this result is discussed in more detail in the main text (see Figure 6) B) Cluster map for DAC using the median aggregation function and epsilon = 0.23 and min-samples = 18. C) Cluster map for DAC using the mean aggregation function and epsilon = 0.20 and min-samples = 16. D) Cluster map for Detect-Cluster-Aggregate-Cluster (DCAC) using epsilon = 0.20 and min-samples = 24 for the final clustering step.

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

Appendix D: Supplementary tables

Table D1. Reciprocal pairwise comparison matrix for the vegetation variables based on Saaty's fundamental scale of importance, where 1 represents equal importance and 6 indicates very strong importance (Saaty, 1977). Variables are prioritised in the following order: cVeg > treeFrac > cSoil > gpp > grassFrac > baresoilFrac > fFire > lai > fVegLitter > rGrowth. Priority weights $w(d)$ are calculated using the geometric mean method and normalised to sum to 1.

Variable	cVeg	cSoil	fFire	treeFrac	grassFrac	baresoilFrac	lai	gpp	fVegLitter	rGrowth	$w(d)$
cVeg	1	2	5	2	3	3	5	2	6	6	0.257
cSoil	0.5	1	3	0.5	0.333	0.333	0.25	1	0.2	0.2	0.042
fFire	0.2	0.333	1	0.2	0.5	0.5	1	0.333	0.5	0.5	0.037
treeFrac	0.5	2	5	1	2	2	4	2	5	5	0.194
grassFrac	0.333	3	2	0.5	1	1	0.333	0.5	4	4	0.116
baresoilFrac	0.333	3	2	0.5	1	1	0.333	0.5	4	4	0.116
lai	0.2	0.25	1	0.25	0.333	0.333	1	0.333	0.5	0.5	0.034
gpp	0.5	1	3	0.5	2	2	3	1	5	5	0.146
fVegLitter	0.167	0.2	0.5	0.2	0.25	0.25	0.5	0.2	1	1	0.029
rGrowth	0.167	0.2	0.5	0.2	0.25	0.25	0.5	0.2	1	1	0.029

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

Table D2. Reciprocal pairwise comparison matrix for the three evaluation metrics and the resulting priority weights $w(m)$. The nonlinearity (NL) metric is given higher priority than cluster spatial autocorrelation (CSA) and cluster consistency (CC). Pairwise comparisons are based on Saaty's fundamental scale of importance, where 1 represents equal importance and 3 indicates moderate importance (Saaty, 1977). Priority weights are calculated using the geometric mean method and normalised to sum to 1.

Metric	NL	CSA	CC	$w(m)$
NL	1	3	3	0.6
CSA	1/3	1	1	0.2
CC	1/3	1	1	0.2

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

. Author contributions. KDM developed the methodological framework, performed the analysis, and wrote the manuscript. SL and RW provided conceptual guidance and critical feedback throughout the research process. All co-authors contributed to discussion and reviewed the manuscript.

. Competing interests. The authors declare no competing interests.

. Acknowledgements. This is ClimTip contribution #173; the ClimTip project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101137601: Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them.

This work was supported by EMBRACER (Summit Grant SUMMIT.1.034), financed by the Dutch Research Council (NWO).

We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated CMIP6. We thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the funding agencies that support CMIP6 and ESGF.

We thank Vincent Overbeeck for his earlier student work on evaluation protocols for TOAD, which provided an initial foundation for the evaluation framework further developed in this study. We are also grateful to Boris Sakschewski, Jonathan Krönke, and Lukas Röhrich for sharing data and for valuable discussions.

AI-based language tools (ChatGPT by OpenAI, and Claude by Anthropic) were used to assist with improving clarity and wording in parts of the manuscript. All scientific content, analysis, and interpretations remain the responsibility of the authors.

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

References

Akogul, S. and Erisoglu, M.: An Approach for Determining the Number of Clusters in a Model-Based Cluster Analysis, Entropy, 19, 452, https://doi.org/10.3390/e19090452, 2017.

Angevaare, J. R. and Drijfhout, S. S.: Catalogue of Strong Nonlinear Surprises in ocean, sea-ice, and atmospheric variables in CMIP6, EGUsphere, pp. 1–40, https://doi.org/10.5194/egusphere-2025-2039, publisher: Copernicus GmbH, 2025.

Armstrong McKay, D. I., Staal, A., Abrams, J. F., Winkelmann, R., Sakschewski, B., Loriani, S., Fetzer, I., Cornell, S. E., Rockström, J., and Lenton, T. M.: Exceeding 1.5°C global warming could trigger multiple climate tipping points, Science, 377, eabn7950, https://doi.org/10.1126/science.abn7950, 2022.

Baker, J. C. A. and Spracklen, D. V.: Divergent Representation of Precipitation Recycling in the Amazon and the Congo in CMIP6 Models, Geophysical Research Letters, 49, e2021GL095 136, https://doi.org/10.1029/2021GL095136, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2021GL095136, 2022.

Bathiany, S., Hidding, J., and Scheffer, M.: Edge Detection Reveals Abrupt and Extreme Climate Events, Journal of Climate, 33, 6399–6421, https://doi.org/10.1175/JCLI-D-19-0449.1, 2020.

Bathiany, S., Bastiaansen, R., Bastos, A., Blaschke, L., Lever, J., Loriani, S., De Keersmaecker, W., Dorigo, W., Milenković, M., Senf, C., Smith, T., Verbesselt, J., and Boers, N.: Ecosystem Resilience Monitoring and Early Warning Using Earth Observation Data: Challenges and Outlook, Surveys in Geophysics, https://doi.org/10.1007/s10712-024-09833-z, 2024.

Bergstra, J. and Bengio, Y.: Random search for hyper-parameter optimization, The Journal of Machine Learning Research, https://doi.org/10.5555/2188385.2188395, publisher: JMLR.orgPUB6573, 2012.

Boulton, C. A. and Lenton, T. M.: A new method for detecting abrupt shifts in time series, F1000Research, 8, 746, https://doi.org/10.12688/f1000research.19310.1, 2019.

Brando, P. M., Barlow, J., Macedo, M. N., Silvério, D. V., Ferreira, J. N., Maracahipes, L., Anderson, L., Morton, D. C., Alencar, A., Paolucci, L. N., Jacobs, S., Stouter, H., Randerson, J., Flores, B. M., Starinchak, B., Coe, M., Pires, M. M., Rattis, L., Armenteras, D., Artaxo, P., Ordway, E. M., Trumbore, S., Staver, C., Berenguer, E., Menor, I. O., Maracahipes-Santos, L., Potter, N., Spracklen, D. V., and Uribe, M.: Tipping Points of Amazonian Forests: Beyond Myths and Toward Solutions, Annual Review of Environment and Resources, 50, 97–131, https://doi.org/10.1146/annurev-environ-111522-112804, publisher: Annual Reviews, 2025.

Brouillet, A., Coulaud, G., Shasha, D., Akbarinia, R., and Masseglia, F.: ClimBurst: A Novel Method to Detect Climatological Anomalies Over Time and Space, Geophysical Research Letters, 52, e2025GL117 095, https://doi.org/10.1029/2025GL117095, _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2025GL117095, 2025.

Cardoso, D., Särkinen, T., Alexander, S., Amorim, A., Bittrich, V., Celis, M., Daly, D. C., Fiaschi, P., Funk, V. A., Giacomin, L. L., Goldenberg, R., Heiden, G., Iganci, J., Kelloff, C. L., Knapp, S., Cavalcante de Lima, H., Machado, A. F. P., dos Santos, R. M., Mello-Silva, R., Michelangeli, F., Mitchell, J., Moonlight, P., de Moraes, P., Mori, S. A., Nunes, T. S., Pennington, T. D., Pirani, J., Prance, G. T., de Queiroz, L. P., Rapini, A., Riina, R., Rincon, C. A. V., Roque, N., Shimizu, G., Sobral, M., Stehmann, J., Stevens, W. D., Taylor, C. M., Trovó, M., van den Berg, C., van der Werff, H., Viana, P. L., Zartman, C. E., and Forzza, R. C.: Amazon plant diversity revealed by a taxonomically verified species list, Proceedings of the National Academy of Sciences, 114, 10 695–10 700, https://doi.org/10.1073/pnas.1706756114, publisher: Proceedings of the National Academy of Sciences, 2017.

Curry, H. B.: The method of steepest descent for non-linear minimization problems, Quarterly of Applied Mathematics, 2, 258–261, https://doi.org/10.1090/qam/10667, 1944.

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

Drijfhout, S., Bathiany, S., Beaulieu, C., Brovkin, V., Claussen, M., Huntingford, C., Scheffer, M., Sgubin, G., and Swingedouw, D.: Catalogue of abrupt shifts in Intergovernmental Panel on Climate Change climate models, Proceedings of the National Academy of Sciences, 112, https://doi.org/10.1073/pnas.1511451112, 2015.

Drüke, M., Bloh, W. V., Sakschewski, B., Wunderling, N., Petri, S., Cardoso, M., Barbosa, H. M. J., and Thonicke, K.: Climate-induced hysteresis of the tropical forest in a fire-enabled Earth system model, The European Physical Journal Special Topics, 230, 3153–3162, https://doi.org/10.1140/epjs/s11734-021-00157-2, 2021.

Duffy, K. A., Schwalm, C. R., Arcus, V. L., Koch, G. W., Liang, L. L., and Schipper, L. A.: How close are we to the temperature tipping point of the terrestrial biosphere?, Science Advances, 7, eaay1052, https://doi.org/10.1126/sciadv.aay1052, publisher: American Association for the Advancement of Science, 2021.

Edwards, P. N.: History of climate modeling, WIREs Climate Change, 2, 128–139, https://doi.org/10.1002/wcc.95, 2011.

Ester, M., Kriegel, H. P., Sander, J., and Xiaowei, X.: A density-based algorithm for discovering clusters in large spatial databases with noise, AAAI Press, Menlo Park, CA (United States), United States, https://www.osti.gov/biblio/421283, 1996.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geoscientific Model Development, 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., Collins, W. D., Gier, B. K., Hall, A. D., Hoffman, F. M., Hurtt, G. C., Jahn, A., Jones, C. D., Klein, S. A., Krasting, J. P., Kwiatkowski, L., Lorenz, R., Maloney, E., Meehl, G. A., Pendergrass, A. G., Pincus, R., Ruane, A. C., Russell, J. L., Sanderson, B. M., Santer, B. D., Sherwood, S. C., Simpson, I. R., Stouffer, R. J., and Williamson, M. S.: Taking climate model evaluation to the next level, Nature Climate Change, 9, 102–110, https://doi.org/10.1038/s41558-018-0355-y, publisher: Nature Publishing Group, 2019.

Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., Arnone, E., Bellprat, O., Brötz, B., Caron, L.-P., Carvalhais, N., Cionni, I., Cortesi, N., Crezee, B., Davin, E. L., Davini, P., Debeire, K., de Mora, L., Deser, C., Docquier, D., Earnshaw, P., Ehbrecht, C., Gier, B. K., Gonzalez-Reviriego, N., Goodman, P., Hagemann, S., Hardiman, S., Hassler, B., Hunter, A., Kadow, C., Kindermann, S., Koirala, S., Koldunov, N., Lejeune, Q., Lembo, V., Lovato, T., Lucarini, V., Massonnet, F., Müller, B., Pandde, A., Pérez-Zanón, N., Phillips, A., Predoi, V., Russell, J., Sellar, A., Serva, F., Stacke, T., Swaminathan, R., Torralba, V., Vegas-Regidor, J., von Hardenberg, J., Weigel, K., and Zimmermann, K.: Earth System Model Evaluation Tool (ESMValTool) v2.0 – an extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP, Geoscientific Model Development, 13, 3383–3438, https://doi.org/10.5194/gmd-13-3383-2020, publisher: Copernicus GmbH, 2020.

Flores, B. M., Montoya, E., Sakschewski, B., Nascimento, N., Staal, A., Betts, R. A., Levis, C., Lapola, D. M., Esquível-Muelbert, A., Jakovac, C., Nobre, C. A., Oliveira, R. S., Borma, L. S., Nian, D., Boers, N., Hecht, S. B., ter Steege, H., Arieira, J., Lucas, I. L., Berenguer, E., Marengo, J., Gatti, L. V., Mattos, C. R. C., and Hirota, M.: Critical transitions in the Amazon forest system, Nature, 626, 555–564, https://doi.org/10.1038/s41586-023-06970-0, publisher: Nature Publishing Group, 2024.

Frieler, K., Lange, S., Piontek, F., Reyer, C. P. O., Schewe, J., Warszawski, L., Zhao, F., Chini, L., Denvil, S., Emanuel, K., Geiger, T., Halladay, K., Hurtt, G., Mengel, M., Murakami, D., Ostberg, S., Popp, A., Riva, R., Stevanovic, M., Suzuki, T., Volkholz, J., Burke, E., Ciais, P., Ebi, K., Eddy, T. D., Elliott, J., Galbraith, E., Gosling, S. N., Hattermann, F., Hickler, T., Hinkel, J., Hof, C., Huber, V., Jägermeyr, J., Krysanova, V., Marcé, R., Müller Schmied, H., Mouratiadou, I., Pierson, D., Tittensor, D. P., Vautard, R., Van Vliet, M., Biber, M. F., Betts, R. A., Bodirsky, B. L., Deryng, D., Frolking, S., Jones, C. D., Lotze, H. K., Lotze-Campen, H., Sahajpal, R., Thonicke, K., Tian, H., and Yamagata, Y.: Assessing the impacts of 1.5 °C global warming – simulation protocol of the Inter-Sectoral Impact

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

Model Intercomparison Project (ISIMIP2b), Geoscientific Model Development, 10, 4321–4345, https://doi.org/10.5194/gmd-10-4321-2017, 2017.

Frieler, K., Lange, S., Schewe, J., Mengel, M., Treu, S., Otto, C., Volkholz, J., Reyer, C. P. O., Heinicke, S., Jones, C., Blanchard, J. L., Harrison, C. S., Petrik, C. M., Eddy, T. D., Ortega-Cisneros, K., Novaglio, C., Heneghan, R., Tittensor, D. P., Maury, O., Büchner, M., Vogt, T., Quesada Chacón, D., Emanuel, K., Lee, C.-Y., Camargo, S. J., Jägermeyr, J., Rabin, S., Klar, J., Vega del Valle, I. D., Novak, L., Sauer, I. J., Lasslop, G., Chadburn, S., Burke, E., Gallego-Sala, A., Smith, N., Chang, J., Hantson, S., Burton, C., Gädeke, A., Li, F., Gosling, S. N., Müller Schmied, H., Hattermann, F., Hickler, T., Marcé, R., Pierson, D., Thiery, W., Mercado-Bettín, D., Ladwig, R., Ayala-Zamora, A. I., Forrest, M., Bechtold, M., Reinecke, R., de Graaf, I., Kaplan, J. O., Koch, A., and Lengaigne, M.: Scenario set-up and the new CMIP6-based climate-related forcings provided within the third round of the Inter-Sectoral Model Intercomparison Project (ISIMIP3b, group I and II), EGUsphere, pp. 1–70, https://doi.org/10.5194/egusphere-2025-2103, publisher: Copernicus GmbH, 2025.

Gatti, L. V., Basso, L. S., Miller, J. B., Gloor, M., Gatti Domingues, L., Cassol, H. L. G., Tejada, G., Aragão, L. E. O. C., Nobre, C., Peters, W., Marani, L., Arai, E., Sanches, A. H., Corrêa, S. M., Anderson, L., Von Randow, C., Correia, C. S. C., Crispim, S. P., and Neves, R. A. L.: Amazonia as a carbon source linked to deforestation and climate change, Nature, 595, 388–393, https://doi.org/10.1038/s41586-021-03629-6, publisher: Nature Publishing Group, 2021.

Harteg, J., Röhrich, L., De Maeyer, K., Garbe, J., Sakschewski, B., Klose, A. K., Donges, J. F., Winkelmann, R., and Loriani, S.: TOAD v1.0: A Python Framework for Detecting Abrupt Shifts and Coherent Spatial Domains in Earth-System Data, EGUsphere, pp. 1–29, https://doi.org/10.5194/egusphere-2026-356, publisher: Copernicus GmbH, 2026.

Heinicke, S., Frieler, K., Jägermeyr, J., and Mengel, M.: Global gridded crop models underestimate yield responses to droughts and heatwaves, Environmental Research Letters, 17, 044 026, https://doi.org/10.1088/1748-9326/ac592e, publisher: IOP Publishing, 2022.

Hubau, W., Lewis, S. L., Phillips, O. L., Affum-Baffoe, K., Beeckman, H., Cuní-Sanchez, A., Daniels, A. K., Ewango, C. E. N., Fauset, S., Mukinzi, J. M., Sheil, D., Sonké, B., Sullivan, M. J. P., Sunderland, T. C. H., Taedoumg, H., Thomas, S. C., White, L. J. T., Abernethy, K. A., Adu-Bredu, S., Amani, C. A., Baker, T. R., Banin, L. F., Baya, F., Begne, S. K., Bennett, A. C., Benedet, F., Bitariho, R., Bocko, Y. E., Boeckx, P., Boundja, P., Brienen, R. J. W., Brncic, T., Chezeaux, E., Chuyong, G. B., Clark, C. J., Collins, M., Comiskey, J. A., Coomes, D. A., Dargie, G. C., De Haulleville, T., Kamdem, M. N. D., Doucet, J.-L., Esquivel-Muelbert, A., Feldpausch, T. R., Fofanah, A., Foli, E. G., Gilpin, M., Gloor, E., Gonmadje, C., Gourlet-Fleury, S., Hall, J. S., Hamilton, A. C., Harris, D. J., Hart, T. B., Hockemba, M. B. N., Hladik, A., Ifo, S. A., Jeffery, K. J., Jucker, T., Yakusu, E. K., Kearsley, E., Kenfack, D., Koch, A., Leal, M. E., Levesley, A., Lindsell, J. A., Lisingo, J., Lopez-Gonzalez, G., Lovett, J. C., Makana, J.-R., Malhi, Y., Marshall, A. R., Martin, J., Martin, E. H., Mbayu, F. M., Medjibe, V. P., Mihindou, V., Mitchard, E. T. A., Moore, S., Munishi, P. K. T., Bengone, N. N., Ojo, L., Ondo, F., Peh, K. S.-H., Pickavance, G. C., Poulsen, A. D., Poulsen, J. R., Qie, L., Reitsma, J., Rovero, F., Swaine, M. D., Talbot, J., Taplin, J., Taylor, D. M., Thomas, D. W., Toirambe, B., Mukendi, J. T., Tuagben, D., Umunay, P. M., Van Der Heijden, G. M. F., Verbeeck, H., Vleminckx, J., Willcock, S., Wöll, H., Woods, J. T., and Zemagho, L.: Asynchronous carbon sink saturation in African and Amazonian tropical forests, Nature, 579, 80–87, https://doi.org/10.1038/s41586-020-2035-0, 2020.

Intergovernmental Panel on Climate Change, I.: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland., https://doi.org/https://doi.org/10.59327/IPCC/AR6- 9789291691647, 2023.

Jain, A. K.: Data clustering: 50 years beyond K-means, Pattern Recognition Letters, 31, 651–666, https://doi.org/10.1016/j.patrec.2009.09.011, 2010.

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

Lenton, T. M., Held, H., Kriegler, E., Hall, J. W., Lucht, W., Rahmstorf, S., and Schellnhuber, H. J.: Tipping elements in the Earth's climate system, Proceedings of the National Academy of Sciences, 105, 1786–1793, https://doi.org/10.1073/pnas.0705414105, publisher: Proceedings of the National Academy of Sciences, 2008.

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.

Parry, I. M., Ritchie, P. D. L., and Cox, P. M.: Evidence of localised Amazon rainforest dieback in CMIP6 models, Earth System Dynamics, 13, 1667–1675, https://doi.org/10.5194/esd-13-1667-2022, 2022.

Peng, Y., Kou, G., Wang, G., Wu, W., and Shi, Y.: ENSEMBLE OF SOFTWARE DEFECT PREDICTORS: AN AHP-BASED EVALUATION METHOD, {International Journal of Information Technology & Decision Making}, 10, 187–206, https://doi.org/10.1142/S0219622011004282, 2011.

Righi, M., Andela, B., Eyring, V., Lauer, A., Predoi, V., Schlund, M., Vegas-Regidor, J., Bock, L., Brötz, B., de Mora, L., Diblen, F., Dreyer, L., Drost, N., Earnshaw, P., Hassler, B., Koldunov, N., Little, B., Loosveldt Tomas, S., and Zimmermann, K.: Earth System Model Evaluation Tool (ESMValTool) v2.0 – technical overview, Geoscientific Model Development, 13, 1179–1199, https://doi.org/10.5194/gmd-13-1179-2020, publisher: Copernicus GmbH, 2020.

Saaty, R.: The analytic hierarchy process—what it is and how it is used, Mathematical Modelling, 9, 161–176, https://doi.org/10.1016/0270-0255(87)90473-8, 1987.

Saaty, T. L.: A scaling method for priorities in hierarchical structures, Journal of Mathematical Psychology, 15, 234–281, https://doi.org/10.1016/0022-2496(77)90033-5, 1977.

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., and Xu, X.: DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN, ACM Transactions on Database Systems, 42, 1–21, https://doi.org/10.1145/3068335, 2017.

Staal, A., Tuinenburg, O. A., Bosmans, J. H. C., Holmgren, M., Van Nes, E. H., Scheffer, M., Zemp, D. C., and Dekker, S. C.: Forest-rainfall cascades buffer against drought across the Amazon, Nature Climate Change, 8, 539–543, https://doi.org/10.1038/s41558-018-0177-y, 2018.

Steffen, W., Richardson, K., Rockström, J., Schellnhuber, H. J., Dube, O. P., Dutreuil, S., Lenton, T. M., and Lubchenco, J.: The emergence and evolution of Earth System Science, Nature Reviews Earth & Environment, 1, 54–63, https://doi.org/10.1038/s43017-019-0005-6, 2020.

Terpstra, S., Falkena, S. K. J., Bastiaansen, R., Bathiany, S., Dijkstra, H. A., and von der Heydt, A. S.: Assessment of Abrupt Shifts in CMIP6 Models Using Edge Detection, AGU Advances, 6, e2025AV001 698, https://doi.org/10.1029/2025AV001698, _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2025AV001698, 2025.

van Westen, R. and Dijkstra, H. A.: Persistent climate model biases in the Atlantic Ocean's freshwater transport, Ocean Science, 20, 549–567, https://doi.org/10.5194/os-20-549-2024, publisher: Copernicus GmbH, 2024.

Watterson, I. G., Bathols, J., and Heady, C.: What Influences the Skill of Climate Models over the Continents?, Bulletin of the American Meteorological Society, 95, 689–700, https://doi.org/10.1175/BAMS-D-12-00136.1, publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society, 2014.

Wieder, W. R., Cleveland, C. C., Smith, W. K., and Todd-Brown, K.: Future productivity and carbon storage limited by terrestrial nutrient availability, Nature Geoscience, 8, 441–444, https://doi.org/10.1038/ngeo2413, publisher: Nature Publishing Group, 2015.

https://doi.org/10.5194/egusphere-2026-3389
Preprint. Discussion started: 23 June 2026

Winkelmann, R., Dennis, D., Donges, J., Loriani, S., Sakschewski, B., and Rockström, J.: The Tipping Point Modelling Intercomparison Project (TIPMIP), https://meetingorganizer.copernicus.org/EGU24/EGU24-17399.html, dOI: 10.5194/egusphere-egu24-17399, 2025.

Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., and Deng, S.-H.: Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimizationb, Journal of Electronic Science and Technology, 17, 26–40, https://doi.org/10.11989/JEST.1674-862X.80904120, 2019.

Zemp, D. C., Schleussner, C.-F., Barbosa, H. M. J., Van Der Ent, R. J., Donges, J. F., Heinke, J., Sampaio, G., and Rammig, A.: On the importance of cascading moisture recycling in South America, Atmospheric Chemistry and Physics, 14, 13 337–13 359, https://doi.org/10.5194/acp-14-13337-2014, 2014.

Zemp, D. C., Schleussner, C.-F., Barbosa, H. M. J., Hirota, M., Montade, V., Sampaio, G., Staal, A., Wang-Erlandsson, L., and Rammig, A.: Self-amplified Amazon forest loss due to vegetation-atmosphere feedbacks, Nature Communications, 8, 14 681, https://doi.org/10.1038/ncomms14681, 2017.