Response to Reviewer 1

Dear reviewer,

Thank you for taking the time to engage with and review our manuscript and for providing detailed and constructive comments! We will address them one by one in the following. All references to line numbers in our replies refer to the track-changed manuscript. We hope that the revised manuscript addresses your concerns adequately.

Comment*: I agree that it would be great if we could predict forest mortality events. However, I wonder if the term Early Warning Sign (EWS) was optimally chosen when the scientific community agreed on using it for Critical-Slowing-Down (CSD) indicators. These indicators are – at best – statistical proxies of the recovery rate of the system, which decreases as the system approaches a fold bifurcation type tipping point. By design, they cannot warn of a tipping by themselves, and could probably be used as "EWS" only in combination with some kind of threshold. However, I am not aware of that and doubt it would be robust, as probably do Scheffer et al. (2009) as they write "We note also that most of the signals we have discussed should still be interpreted in a relative sense. For instance, although autocorrelation is predicted to approach unity at a fold bifurcation, measurement noise will tend to reduce correlations." Hence, I also doubt that Scheffer et al. (2009) claim that "EWS based on the concept of alternative stable states and the theory of CSD have been proposed as a promising alternative approach to predict forest mortality events" (l 48f). Instead, often the change of the proxies is assessed, but all that these changes can tell us is IF the system is loosing stability. I thus believe that even though termed EWS, it is not the CSD-indicators fault that they cannot predict forest mortality, but rather the unfortunate wording that lead to over-optimistic applications in a way the indicators weren't meant to be applied.*

Reply: We thank the reviewer for pointing this out and acknowledge that our framing might have been unclear in some places.

We want to highlight here that by EWS, we here do not just refer to CSD-based methods. Instead, we understand EWS to include different statistical methods that aim to detect resilience loss and the risk of an upcoming regime shift. This includes the physiological indicators that are frequently applied in the case studies analysed in our literature review (cf. lines 123ff), but also methods based on CSD, flickering, and the long-term memory of a system. In this study, we evaluate all the latter three methods (cf. lines 146f).

CSD here describes that before a bifurcation, a system's local stability, i.e. its capacity to recover after small disturbances, is eroded through a slow, gradual driver, and it takes increasingly longer for the system to return to equilibrium. This is often measured through increases in temporal autocorrelation and variance. Flickering describes a behaviour in which a system briefly explores alternative stable states and is more common in stochastic systems close to a basin boundary (Dakos et al., 2013). It can manifest in increasing autocorrelation and variance, as well as increasing skewness and kurtosis over time. Lastly, we use the fractal dimension as a measure of system memory across time scales and whether a system is exploring areas outside the main basin of attraction (Dakos et al., 2012; Gneiting et al., 2012). We added a text box with definitions in the introduction to make clearer what we define as EWS (cf. Box 1). Additionally, we added a more explicit description of our different EWS approaches to the introduction (cf. lines 74ff); as well as a clearer overview of the theoretical assumptions underpinning the different EWS in the discussion (cf. figure 6).

CSD-based indicators are hence only one sub-category of EWS, and while they are expected to appear in univariate systems exhibiting fold bifurcations, they are also known to manifest for other types of bifurcations such as continuous bifurcations or second-order transitions (Kéfi et al., 2013). At the same time, other types of bifurcations such as chaotic crises, Maxwell point transitions, or more generally rate-driven, or noise-driven tipping is not expected to be preceded by CSD

(Boettiger et al., 2013; Ritchie & Sieber, 2017), as elaborated more explicitly in the expanded discussion in lines 345ff). Despite this, CSD-based EWS have received a lot of attention across literature, and it has been suggested multiple times that these EWS could contribute to making decisions on management priorities ("*will integrating CSD-based resilience indicators in decision-making and environmental policy improve ecosystem management?*", Table 2 in Dakos et al., 2015) or should be tested with respect to their potential to serve as real warnings of changes in future systems ("*The next step is to test the real potential of early warnings as preventive and management tools in anticipating natural and human-induced changes to come*.", Dakos et al., 2024), hence we still considered it appropriate to test this.

We changed the phrasing in line 50ff (it now reads "*System-agnostic temporal EWS, often based on the concept of alternative system states, have been applied as indicators of resilience loss across a number of disciplines, and have been suggested to be a promising tool to help guide ecosystem management priorities (Dakos et al., 2015; 2024)"* to reflect that these were general recommendations that were not specifically targeted to forest mortality. Additionally, as the reviewer points out, most EWS are only relative measures of resilience changes, and we have added a clear acknowledgement and description of this to be more transparent about their expected behaviour and weakness (*"Most system-agnostic EWS were not originally developed as predictors of specific tipping points, but as indicators of changes in system properties as proxies of resilience (Scheffer et al., 2009) and thus any potential management application requires a careful evaluation of - as Dakos et al. (2024) put it - "the real potential of early warnings as preventive and management tools in anticipating natural and human-induced changes to come".*

We also changed the framing of our approach in several places across the manuscript to highlight that with our matching of true mortality locations and controls, this relative resilience loss is indeed what we aim to test the EWS against. We will elaborate on this further in the replies to the following comments.

Comment: *The sentence "we assume for this study that the forest mortality events can be considered bifurcations, i.e., significant qualitative shifts in the system state" (l 274) highlights the first two of my concerns. First, a bifurcation is induced by some change in external parameters and resembles a structural change in the stability landscape of the system. For example, it could denote the change from a system with two stable and one unstable equilibrium to a system with only one stable equilibrium. This annihilation of a stable and an unstable equilibrium is the typical fold-bifurcation. This type of bifurcation is commonly assumed in "tipping" science, and is visualized e.g. by Box 1 Figure d in Scheffer et al. (2009), where F1 and F2 are denoting the fold bifurcations.*
If the system was before in the system state associated with the now disappeared stable equilibrium, it would of course get attracted to the other, now only, stable equilibrium, resembling a qualitative shift in the system state. However, this is not the only way such a qualitative shift could happen. Depending on the distance of the unstable equilibrium to the stable equilibrium the system is currently in, more or less external energy (perturbations) is needed to "kick the system over to the other stable equilibrium".
Second, as you say, the forest mortality data includes forest mortality from droughts or heat waves, so basically, the assumption "that the forest mortality events can be considered bifurcations" seems difficult to justify. From my understanding you do not investigates bifurcation induced qualitative shifts but rather, perturbation induced ones – which do not only depend on the system state's resilience, but on how "large" the perturbation is in relation to the resilience. Regarding those, Scheffer et al (2009) already say that "perturbations will often trigger a transition well before a bifurcation point is reached. Thus, although a trend in the indicators may serve as a warning, the actual moment of a transition remains difficult to predict." If, it would probably be reasonable to try to see if the "EWS" can tell us something about how "big" of a perturbation the system can absorb.

Response: We thank the reviewer for pointing out our imprecise framing of bifurcations and the role of perturbations and have revised our wording in a number of places to make this more precise.

We reframed our conceptualization of forest mortality as bifurcations to highlight that we consider large-scale forest mortality to cause a shift into a qualitatively different self-reinforcing state, i.e., regime, but do not want to claim that this shift is purely due to slowly changing, linear drivers, as might have been implied by our framing of it as bifurcations. The revised section now reads as "*Forest mortality events are often driven by a combination of slow, long-term drivers and perturbations (Seidl et al., 2017), and can lead to lasting changes in ecosystem state and dynamics (Gonzalez et al., 2010). Long-term change, such as climate change or smaller droughts over time, can erode a forest's resilience, i.e., its capacity to cope with additional perturbations, causing vulnerability to large-scale mortality as a result of perturbations (Kannenberg et al., 2020). Such mortality events can push a forest into a different self-reinforcing state (Hammond, 2020; Gonzalez et al., 2010; Allen et al., 2010) if the ecosystem that establishes after the collapse is qualitatively different and does not perform the same functions as the original ecosystem (Scheffer et al., 2001; Candell et al., 2021)"* (lines 84ff).

We do not want to claim that the EWS tested here can give us a precise warning signal of the exact timing of upcoming mortality, and to make this clearer, we have adapted the framing with respect to this in the abstract (lines 7, 11, 17), in the introduction in several places (e.g., *"To be useful in an applied context, EWS need to be able to clearly distinguish locations that have undergone resilience loss and are therefore more vulnerable and likely to experience mortality as a result of perturbations."* (lines 136ff), as well as in lines 150f and in the research questions: *"Based on an empirical analysis, how well do commonly applied EWS based on CSD, flickering, and the fractal dimension serve as actual warning signals of resilience loss before events of forest mortality?"* (lines 146f).

We assume instead that the locations that show mortality have undergone a certain resilience loss before, such that the climate extreme event is sufficient as a perturbation to push them into a different state (as outlined in lines 39ff: "*Individual droughts or heat waves often act as shocks. If previous stressors have already reduced forest resilience, the forest may not be able to absorb additional pressures*"). Hence, we do not consider these mortality events to be purely driven by slow, external drivers, but by a combination of slow background change (climate change, or an accumulation of smaller droughts that erode resilience) and the acute climate events (drought or heat wave). We also changed the wording accordingly in multiple sections (cf. lines 84ff, 153ff, 188ff, 372ff) to reflect this more explicitly.

In our methodological setup, we try to disentangle the role of perturbations and slow background change by comparing the mortality locations to controls. As the paired mortality-control sites are within 10km of one another, we assume that the perturbation (the drought or heatwave) acts on both locations. However, in one of the locations, this perturbation leads to mortality, whereas in the other it does not, which we take to signify that there has previously been a loss of resilience at the mortality site, which we aim to detect with the EWS. Regarding this, we adjusted the framing in the introduction (lines 88ff), methods section (lines 190ff), as well as in the discussion (lines 388ff).

We also want to note that, unlike in theory, in complex real-world ecosystems, such as forests that are exposed to many biogeophysical conditions (such as soil conditions, precipitation, temperature, etc.) it can be difficult to clearly differentiate between slow external drivers, perturbations, and noise (cf. revised lines 382ff). As discussed in lines 364ff, smaller droughts, for example, are known to leave legacy effects and mostly negatively impact a forest's future response to further droughts. In that sense, they might be considered to act as both a long-term driver and a perturbation. Furthermore, as soon as multiple codimensions are acting on a system (such as temperature change and change in moisture availability), or rate- or noise-based transitions come into play, many EWS are not expected to work well anymore (cf. lines 393).

We revised the relevant sections in the discussion to disentangle the assumptions regarding driver dynamics that are included in the different EWS (cf. Figure 6 and lines 357ff): "*The EWS employed here rely on different assumptions about driver dynamics as highlighted in Figure 6. For CSD to emerge, one generally assumes a slow, gradually varying driver (Brock and Carpenter, 2012), while flickering is expected to show for systems already close to a transition point undergoing stochastic perturbations (Dakos et al., 2013). The fractal dimension does not make such explicit assumptions about the dynamics of the drivers and is expected to be sensitive to both slow and abrupt changes. In this study, as outlined in 1, we assume the droughts and heat waves to be the final perturbation pushing a specific site that has already lost resilience into mortality. We test whether the EWS can detect this preceding resilience loss that ultimately increases a specific site's vulnerability to the disturbance. By comparing the different EWS that make differing assumptions about driver dynamics, we aim to incorporate both the stochastic, perturbation-like effect of droughts and heatwaves, but also the long-term effects of gradual climate change and drought legacy effects. Of course, given the complexity of ecosystem dynamics, we cannot exclude the possibility that our assumption about gradual resilience loss in combination with a specific perturbation may be misguided. As pointed out earlier, the critical transitions we observe might also be rate-driven or noise-driven transitions (Ritchie and Sieber, 2017; Ritchie et al., 2023). Additionally, even in our matched site design, we cannot fully account for additional factors affecting drought and heat wave response, such as microclimate, different species composition, or root length adaptations (Singh et al., 2020, 2022). Overall, disentangling whether and to what extent the different drivers of mortality meet the theoretical assumptions underlying different EWS is thus inherently challenging in real-world ecosystems, as the same observed mortality pattern may arise from different underlying dynamical mechanisms operating simultaneously across scales.*"

We want to be explicit that we agree with the reviewer and do not claim our cases fulfil all the theoretical requirements of the EWS employed here, but rather that it may be difficult in real ecosystems to determine whether they are actually fulfilled. Hence, as the reviewer points out, we call for caution in applying EWS to cases that are outside the appropriate cases. As expanded on further down, we revised the discussion (lines 350ff) and the recommendations for future work (lines 514ff) to reflect this more explicitly. We also highlight that testing multiple different EWS based on different assumptions about the system in question can be helpful in such cases (cf. lines 355ff).

Comment: *Moreover, even if the "size" of the perturbation was correctly quantifiable, one would still be remained with the statistics (which is often misleadingly called "EWS") that are only proxies of the resilience, or more precisely, the recovery rate, for which also one cannot be certain how much of it resembles the vegetation's intrinsic time scales (we do not assume that boreal forest has the same speed in growing as e.g. a rainforest when both are completely healthy). Hence, I wonder if "Computing recovery rates instead of recovery time in this way allows us to compare different biomes with different ecological dynamics and prevents us from having to make assumptions about which mean vegetation index value the system has to return to" (l 188ff) should be reconsidered.*

Reply: We thank the reviewer for highlighting this and agree that, of course, vegetation dynamics and time scales are not directly comparable between different biomes. We computed recovery rates in this way, following the procedure used in Smith & Boers (2022), which they apply globally to remotely sensed time series. These recovery rates are computed on our detrended and deseasoned time series, which should have removed the largest part of ecosystem-specific dynamics and allowed for more comparability across biomes. However, there might of course still be remaining effects, and we therefore use the recovery rate together with the latitude of each mortality event in our driver models (cf. lines 242) to add an additional control variable for potentially different dynamics across biomes. We don't find a consistent effect of either recovery rates or latitude on the kendall tau values (cf. lines 311ff). However, to acknowledge the limitations of recovery rates, we adjusted the wording in the given sentence to read as *"Computing recovery rates on the preprocessed timeseries allows us to compare different biomes with different ecological dynamics and*

prevents us from having to make assumptions about which mean vegetation index value the system has to return to. As there might still be some residual effects of seasonality or different biomes, we also extract information on latitude from the forest mortality database." (lines 236ff).
Comment: *As far as I understand, the Kendall tau of the statistic (AC1/var/…) are considered "the early warning signal". However, the Kendall tau only informs us if the resilience is decreasing (positive trend), and how close this change moves the system towards its "tipping point" is highly dependent on how stable the system was in the beginning. Hence, the trend or Kendall tau is only informative of IF the system is loosing resilience, due to noise I would argue it doesn't even inform us how much, but for sure not how close to the system is getting to the bifurcation point.*

Reply: We thank the reviewer for this point. We agree that the Kendall tau coefficient measures only the trend in a statistic (e.g., AC1, var, skew) and does not directly measure resilience or proximity to a bifurcation. While many studies interpret positive trends as reflecting decreasing resilience, this is conditional on theoretical assumptions (e.g., CSD near a bifurcation) rather than an intrinsic property of the coefficient. As outlined in response to the second comment, we use the mortality-control setup to evaluate whether the true mortality points show indicators of resilience loss as assessed with the Kendall tau of the different metrics. As both mortality and control points undergo the same perturbation (heat wave or drought), we assume that the area that undergoes mortality as a result of this should already show signs of decreased resilience beforehand. As outlined earlier, we revised the framing across the paper (cf. lines 84ff, 153ff, 188ff, 372ff) to reflect this more clearly.

Comment: *In summary, I believe that the conclusion "that despite their extensive utilization and ease of application, system-agnostic Early Warning Signals based on Critical Slowing Down are ineffective and non-robust predictors of forest mortality events on a global scale" cannot be "attributed to the poor approximation of ecosystem dynamics from optical vegetation indices and the multivariate character of real ecosystems" but rather that the theory was applied beyond the conceptual scope it can reliably represent. Which is, of course, an important notion to make.*

Reply: We thank the reviewer for pointing this out and adjusted the phrasing in this statement to put more focus on the limitations of the theory in application to real systems. It now reads as "*We determine the primary reasons for this to be the complexity and multivariate character of real ecosystems and their drivers, which do not necessarily align with the narrow theoretical basis for each specific EWS, as well as the poor approximation of ecosystem dynamics using optical vegetation indices.*" (lines 551ff).

We also expanded the analysis of this in the introduction (lines 70ff). Additionally, we revised the discussion significantly to highlight the specific theoretical assumptions associated with the different EWS more clearly and discuss their applicability to our system of study (Figure 6, lines 326-356 and 368 – 384).

Moreover, in our considerations for future developments, we reflect more explicitly on the appropriateness of different system-agnostic EWS methods: *"Future applications should carefully consider how far the underlying assumptions for specific EWS are fulfilled regarding the dimensionality of the ecosystem and the characteristics of its drivers. However, as outlined earlier, it is important to note that such a theoretical evaluation might be difficult for real systems. Testing multiple alternative representations of a specific system (e.g., univariate or multivariate; driven by slowly varying drivers, noise, or rate-induced transitions) for a specific, clearly defined case, might be more useful in developing applicable EWS than relying on system-agnostic CSD-based indicators without additional validation."* (lines 514ff).

Specific comments

Comment: *I find that the term "trend", "Kendall tau", and "EWS" are often used ambiguously, e.g. what are "Kendall tau trends"?*

Reply: We thank the reviewer for pointing this out and adjusted the language throughout the document to be less ambiguous. We also added explicit definitions in Box 1.

Comment: *"At true mortality locations, only a minority of cases show a statistically significant increase in AC1 before the event (41.7% ± 5.6% as mean±standard deviation as assessed across 1, 3, 5, 10, and all years before the event for NDVI). " (l 236ff)*

This makes me wonder if I understood your approach correctly. How do you get a trend from 1 year only, if your minimum window size for calculating the EWS was 1 year?

Reply*:* We calculate the raw time series statistics (e.g. AC1, var) on bi-weekly observations at each time point t using a rolling window of length l centered at (t-l/2) so that the value at point t reflects the value of the statistic in the time leading up to it and does not conflate it with values of later time steps. Hence, for the setup of a rolling window length of 1 year and trend assessment over the 1 year before mortality, the statistics do technically include values from the two years before mortality due to the rolling window configuration. We expanded the description of this in the methods section to be more transparent (lines 219ff).

Comment: *I would wish for a bit more of an explanation what "resistance" is supposed to mean. To me, it seems to be only the difference, but then later on you call it the "resistance to climate extremes" (l 253) so I know wonder if there is more to it than I see.*

Reply: This is correct, we compute resistance as the difference between mean VI in the year of the recorded mortality and the mean of the three years preceding (cf. lines 180ff). The phrase "resistance to climate extremes" was meant to emphasize that this resistance is computed on the year of the recorded mortality, but we see that this might be misleading and changed the wording to "the resistance of the ecosystem to the drought or heat wave" (line 304f). We also added an explicit definition of resistance to the text (Box 1).

Comment: *How do you treat missing values which you might have a lot in the 250m resolution data sets?*

Reply: We filter out low-quality values based on the MODIS Summary Quality Assurance Value (cf. lines 177f) and set them to missing values. In the further processing, we follow the procedure of Smith & Boers (2022) to handle missing values: In our time-series preprocessing, we chose harmonic deseasoning instead of STL (Seasonal Trend Decomposition using LOESS), as this does not require us to fill missing values artificially. Next, we omit all missing values for the computation of our statistics over the rolling time windows, but require a minimum of 5 non-missing points to compute them and otherwise return a missing value. The Kendall tau coefficients are also computed, omitting missing values. We added more details on this in the methods section (cf. lines 215ff).

Generally, the question of cloud-obscured observations is mostly problematic in the tropics, however, as only 9% of our cases are situated in the tropics (cf. lines 169ff), we do not consider this a major problem.