Reviewer 1

I appreciate the authors' careful revision and response to my comments. I am satisfied with most changes/explanations, but I would suggest the authors to expand the discussion on the practical implications of their results. In particular, in light of the results here, how people should analyze resilience using remote sensing data? Or they should stop such kind of analyses, e.g., in certain context? To this end, I still feel that the robustness of composite EWS to SNR should be evaluated. I understand that the theoretical link between EWS and SNR is less clear than the individual metric of EWS. But I believe that the objective of the current study is not to illustrate theoretical relationships, but instead assessing the robustness of resilience metrics in practical uses.

Thank you for your comments, and for taking time to re-review the Manuscript.

The main practical outcome of our work is the insight that multi-satellite (or any multi-sensor data, e.g., paleo-climate data from multiple proxies) has certain potential limitations that need to be taken into consideration when examining system resilience based on such data. Namely, averaging resilience indicators from a large number of different observational time series can serve to enhance measurement issues that aren't dominant in the dynamics of individual time series. Given reasonable signal-to-noise ratios (i.e., those found on modern satellite platforms), individual pixel (location) time series are still likely to yield reliable resilience estimates. We provide a rough proxy for that reliability – the correlation between variance and autocorrelation through time. The more positive this correlation, the smaller the unwanted influence from combining different sensors and vice versa. Indeed, we show that the correlation increases as signal-to-noise ratios improve (Figure 4).

What should hence in general not be done is averaging large ensembles of resilience estimates based on multi-satellite data, as that will serve to enhance spurious signals from changing instrumentation. This is a key insight that can be applied in future studies which look at large-scale, regional, or global aggregations of resilience estimates through time.

Composite resilience metrics preclude the possibility of comparing metrics that should (but don't always) agree. In particular, the theory of critical slowing down demands that individual changes in variance and lag-1 autocorrelation have to be consistent in order to be interpretable in terms of resilience changes; by combining them, it would not be possible to check this anymore. Indeed, we strongly focus on comparing the behavior of variance and lag-1 autocorrelation (via correlation coefficients) in our paper to tell apart actual resilience changes from spurious influences stemming from combining different sensors.

We have updated our Discussion to include an additional explicit note on the practical implications of our results.

Reviewer 2

Smith et al. investigated the reliability of resilience metric with a focus on the non-stationary characteristics of the measurement process. This arouses rethinking of the reliability of resilience conclusions derived from satellite-based vegetation products in current studies. I agree with the point that higher-order moment should be considered in resilience/sensitivity studies.

Thank you very much for taking the time to review our Manuscript. Addressing your comments has helped us further improve the presentation of our results. We will respond to your individual points below, together with references of what we changed in accordance. For clarity, we have broken some of your paragraphs into individual questions to make our response easier to follow.

However, I have several questions for the construction of the synthetic datasets as described by the authors. Also, I strongly recommend the authors to add associated backgrounds/descriptions/explanations in the introduction/methods sections, e.g., why the authors calculate correlation between AR1 and variance suddenly?

The theory of Critical Slowing Down, upon which we base our analysis, stipulates that variance and autocorrelation should be correlated through time, as they follow and respond to the same underlying process (see e.g. Boers, 2021 or Smith et al., 2022, and references therein). If the two variables are not correlated, the theory does not hold – we cannot say whether a state transition is approaching or more generally, if the system resilience is declining. On the other hand, increasing measurement noise leads to increasing variance but at the same time to decreasing lag-1 autocorrelation and vice versa. Hence, negative correlations between AC1 and variance suggest that their changes are caused by measurement / sensor issues rather than resilience changes and call for caution when using such data for resilience estimation. This has been clarified in the Introduction in Lines 31-39.

Why we should care about the aggregation process?

Aggregation destroys fine-scale spatial/temporal variability that is used to monitor changes in the higher-order signal dynamics (e.g., AC1, variance). Different aggregation processes will have different impacts, depending on how exactly the aggregation is performed. In general, aggregation will suppress the fine-scale variability that we are interested in for resilience estimation. This is mentioned in Line 36.

What the aggregation means, temporally or spatially?

In this paper, we use synthetic surrogates (e.g., 1000 simulations, cf. Figures 2, 4) to represent a spatial field of data. This is an idealized simulation – we don't, for example, model spatial autocorrelation between simulations – to focus on the influence of aggregating multiple data sets with similar measurement properties. We then examine how that aggregation serves to enhance or suppress different signals in our resilience proxies AC1 and variance.

Temporal averaging is performed in a few overlapping ways aimed at mimicking the construction of multi-sensor satellite vegetation data. We start in all cases with daily data from sensors with overlapping time periods. We then average these multiple satellite time series to daily temporal resolution, and then again to a bi-weekly median (for VOD) or to a bi-weekly maximum (for AVHRR) to follow how this data has been pre-processed in previous publications (Pinzon and Tucker, 2014; Smith et al., 2022; Boulton et

al., 2022). We do not aim here to determine the 'best' means of data fusion, but rather to highlight how such data aggregation in general can introduce spurious signals to resilience estimates.

The authors did not show related studies and I got lost when I read here. Please make sure interdisciplinary readers can also understand what you want to convey.

The issue of multi-satellite data and resilience estimation has not really been discussed in the literature so far – this is the gap that our paper aims to address (Lines 50-52). While some studies – for example, the publication presenting the VODCA data used here (Moesinger et al. 2020) – consider the autocorrelation of the data as a reliability metric, they do not explicitly investigate the impacts of data fusion on the signals used for resilience estimation. This is a key gap in the literature, as several studies have used multi-sensor satellite data (Feng et al., 2021; Smith et al. 2022; Boulton et al., 2022) without fully considering the implications of data fusion. We have clarified our introduction to add more context to this discussion.

In addition, the aggregation issue pointed out by the authors are a common sense in the remote sensing and GIS discipline. For a pixel (a mixed pixel) featured with strong spatial heterogeneity, using one value to represent the resilience for the whole area may lead to biased conclusions. I do not understand why the authors try to aggregate the resilience values since they already have resilience values with a high spatial resolution. Maybe the authors just want to reveal something using synthetic data. However, in practice, I would not "first calculating AR1 and variance and then averaging those metrics over a region". That makes no sense. The authors may add several studies that did this process to support your experiment design. Otherwise, will this conclusion benefit other studies, e.g., resilience estimates, in practical?

Many studies which have examined vegetation resilience over large areas present data not only in map view (e.g., one value for a pixel), but also in chart view (e.g., one value per time step). It is this secondary case where issues arise, and this has certainly been done repeatedly in the literature (e.g., Forzieri et al., 2022; Boulton et al., 2022; Smith et al., 2022; Feng et al., 2021 and references therein). Such aggregations are commonly used to illustrate differences between regions (e.g., changes in resilience in North vs South America, Amazon vs Congo), time periods (e.g., Amazon changes in the 90s vs 2000s), or land cover types (e.g., Savanna vs Forest changes and their relative strengths over time). Our results indicate that these spatial aggregations – when done by first calculating resilience estimates and then averaging – have the potential to introduce biases (cf. Figure 3 of the MS). On the other hand, first averaging the raw data (e.g., create a single time series for the whole Amazon), and then calculating changes in resilience through time, has a lower potential to introduce bias, with the important caveat that such averaging removes a lot of actual variability that might be relevant for resilience estimates. Moreover, the averaging should only be done over similar pixel time series. That is, don't aggregate forest and savanna together into a single time series, as this would create mixed-pixel signals which are difficult to interpret and might lead to strong biases. Understanding how averaging different signals can lead to biases is the main motivation for our study, and is applicable to any study that deals with sets of time series based on multi-instrument measurements. We have clarified the point of averaging our synthetic data in the Introduction (Lines 53-55). Several papers which have followed similar steps are also discussed in the Discussion (Lines 231-238).

The authors recommend to use single instrument record. Could I interpret this point of view as current data fusion studies are not reliable or not necessary? I found the authors did not clearly show their setups for the synthetic data, e.g., what is the justice to setup the error variance value? Arbitrary or have referred to associated references?

We do not think data fusion studies are irrelevant – indeed, constantly improving data fusion methods may help ameliorate some of the issues we found here. However, most data fusion studies are geared towards maintaining a continuous mean state or preserving underlying trends – not maintaining higher-order statistics, which would be crucial for estimating resilience reliably. Hence, resilience studies – which rely on those higher order statistics – should use caution when assessing multi-instrument data. If single-instrument data is available, this would eliminate one possible source of bias in interpretations of changing system resilience.

The chosen data noise values in our synthetic time series are arbitrary, and are used to make the point that signals that might be interpreted as a change in resilience (e.g., Figure 3) can be reproduced by time-variable measurement noise (see Lines 114-115). We could vary these noise levels to make the same point about anti-correlated variance/AC1, but would not be able to visually/qualitatively show how well our synthetic experiment can reproduce the findings of a real data set (i.e., Figure 3).

Will different value change the results and conclusions?

Different noise values for the synthetic data would not change our results or conclusions; however, the synthetic data would no longer match up as well to the VOD/AVHRR global composites we present (Figure 3). As we want to show the practical value of our results, we chose noise levels that allowed us to visually present our findings about time-variable noise and potential biases in resilience estimates. We also show a broad range of signal-to-noise ratios to make our results general; also note that these biases are smaller for high-fidelity data, but do not disappear.

Data quality control and integration methods (rescaling technique, weighted average method etc.) are important issues in data fusion studies […]

We fully agree that data fusion studies are reliable in many contexts, but are not necessarily geared towards preserving higher-order statistics as needed for the resilience measures we focus on. The amount of noise removed by various data fusion techniques is not necessarily the issue here, but rather that the *relative amount of measurement noise* can change throughout the time series when multiple overlapping instruments are used. That is exactly the change that might be erroneously interpreted as a change in system resilience, but is rather due to changes in data amount, quality, or sampling through time. This paper aims to illustrate this potential issue, and explore ways to mitigate the problem.

[…] and the fusion results typically have better quality than single satellite-based product (reduced random errors). I did not find new things here.

In general, if signals from different sensors are combined, we have more reliable data by removing random errors. However, the fusion needs to guarantee stationarity in higher-order statistics, which is not common practice; mean-adjustments are much more common. Thus, common data-fusion methods in practice remove or bias exactly those signals required for resilience estimation, while improving signals needed for investigations of the mean-state or mean-shifts (which are not our focus here); reduced random errors will suppress variance and increase autocorrelation during times where there

are more overlapping satellites without changing the underlying signal – this is exactly the kind of change that might be erroneously interpreted as a change in system resilience (Lines 34-38). We do not believe that data fusion is unnecessary or somehow problematic – it is an extremely useful technique for many contexts. However, fusion techniques that are optimal for some contexts can introduce biases in different settings.

Specific comments: [L. denotes Line]

L. 74 The authors integrated the five sensors by taking an average of their values. That is, the authors regarded the weight given to each sensor is equal. However, given that the authors have defined/setup the relative reliability for each sensor in Eq. (2), it is more reasonable to weighted average the five sensors using the weights derived from their relative reliability (this process is a fundamental concept in data fusion study). In addition, the authors should clarify the justice of the magnitude of the error variance for each sensor. Did the authors setup a value arbitrarily or look up associated references? The resulting synthetic data may not be like the VODCA data which the authors are trying to simulate. Please classify.

Thank you for this comment. Our goal in this work was to showcase what can happen if measurement noise levels are not constant during the construction of multi-satellite data. Hence, we created our synthetic data in such a way as to illustrate potential pitfalls, and to see how well we could match global-scale patterns in vegetation resilience presented in previous work (e.g. Boulton et al., 2022; Smith et al., 2022). Our results are to some degree independent of the averaging scheme – whether data are merged via an average, weighted average, maximum value, or CDF-matching approach, underlying changes in noise levels will still be expressed. We chose to present a simplistic case (simple averages) with our synthetic data to make this point.

To address your concern, we have added a weighted-average version of our results as Figure 1 of this Reply. In essence, this is the same analysis as Figure 2 of the MS, but instead of taking an average for overlapping time periods, we take the weighted average, using weights defined by relative sensor reliability.

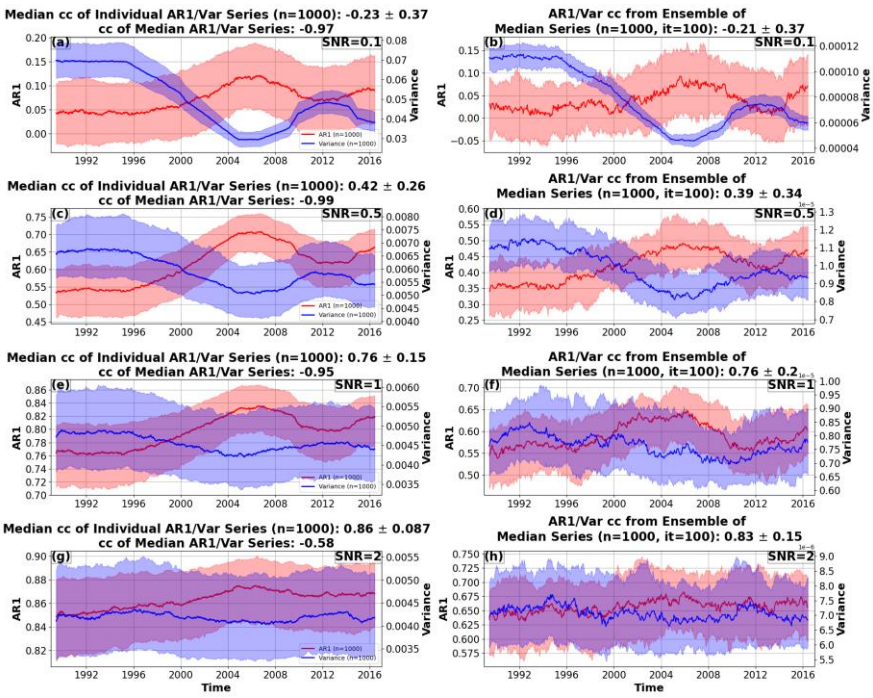

Figure 1: Replicate of Figure 2 from the MS using a weighted average – rather than simple average – to mix multiple sensors together.

As can be seen in Figure 1 of this Reply, the patterns we reveal do not alter – the effect we are trying to explore is not so much due to the averaging scheme chosen, but to the *fact of averaging disparate data itself* – changing the relative noise levels throughout the time series is what introduces biases, not the method of averaging data. There remain only slight variations in our overall statistics between the two averaging schemes.

Our quantitative choices for the noise levels (and averaging scheme) are thus not the main point of our study, and the chosen noise levels are less important in absolute terms than relative. We aimed to show how mixing data together can leave traces in resilience signals, which could be misinterpreted. We further aimed to match our 'jumps' in the resilience signals to the global-scale ones found in previous work (Smith et al., 2022). The similarity we find between aggregated synthetic and global VODCA indicates qualitatively that we construct synthetic data that illustrates the potential issues facing VODCA (or AVHRR).

In addition, why not create a bi-weekly synthetic data here to match the temporal resolution of the NDVI data? The temporal aggregation would add extra error into the final synthetic dataset.

We chose to first create daily data and then aggregate it to better match previous processing approaches. For example, AVHRR data is nominally daily data, but is time-aggregated to be more reliable against clouds and other errors. In previous work, VOD data has been similarly time-aggregated to match the temporal resolution of AVHRR data (Smith et al., 2022; Boulton et al., 2022). As we aimed to give insights that could be applied to real data, we chose to match the time-aggregation schemes of commonly used data sets, rather than construct temporally sparse data from scratch. In our tests, the

inferences were identical whether we applied our methods directly to the daily data or to the time-aggregated data; the main difference is the absolute – but not relative – values of variance and autocorrelation.

L. 82 Similar question with the VODCA simulation given that the authors are trying create a dataset to mimic the GIMMS3g NDVI.

Again, we chose to create daily data to mimic the underlying raw remote sensing data, and then time-aggregate following the methods used in previous papers.

L. 90 How did the authors to define the “changing land cover”? I understand that the authors are trying to get rid of anthropogenic effect. But all the land covers are affected by human. Please elaborate this sentence.

We defined changing land cover as any land cover that had changed classification over the period 2000-2020. For example, a pixel that had changed from Forest to Agriculture then back to Forest would be removed from our analysis, even if it was Forest at both the beginning and end of the study period. In this way we are conservative in only analyzing pixels with stable land cover. We have updated our description in the Methods.

L. 93 What resample method? The Maximum Value Composite or just averaging them?

We resample land cover to match the spatial resolution of the other data by taking the mode. We have clarified this in-text.

L. 107 The effects of combining multiple signals may further be muted if the authors considering weighted average (as I commented in L. 74).

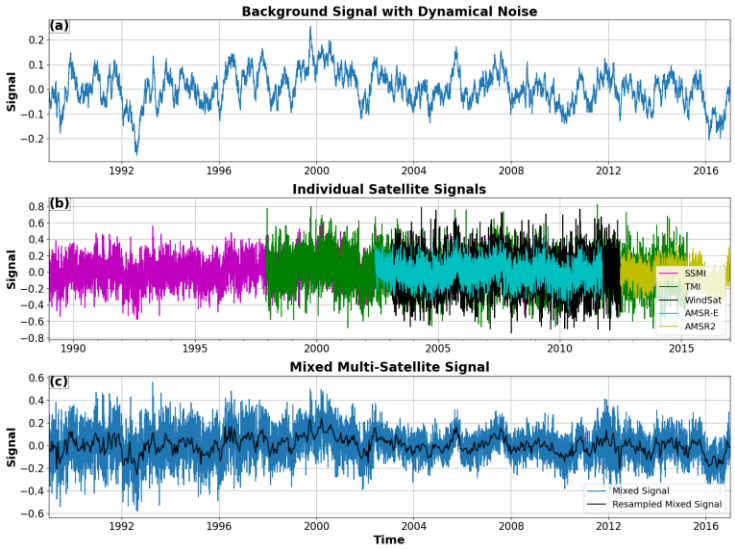

Figure 2: Replicate of Figure 1 of the MS, with weighted rather than simple averaging.

There are indeed slight changes in resultant averaged time series when an alternative mixing scheme based on weighted averages is used. However, these differences are minor, and do not change the expression of dynamic noise levels on resilience indicators (Figure 1 of this Reply), which are sensitive to the relative strength of changing noise levels through time, not to the absolute values of those noise levels.

Caption of Fig.1. Why 0.44? Why not 0.3 or 0.2? Please clarify. This sounds like the authors have considered weighted average in the synthetic data construction. This confused me.

The reliability measures used here are arbitrary, and meant to mimic the global-scale patterns seen in the VOD data (cf. Figure 3). Our aim in this study was to see if we could replicate global-scale resilience signals (i.e., their time dynamics) in real multi-satellite data with a synthetic experiment; hence, noise values for each satellite are tuned towards that goal. We had not considered weighted averaging in our synthetic data construction, but have now added this as an additional check as Figure 1 of this Reply. While weighted averaging of course changes the *absolute* amount of noise variability, it does not change the *relative influence* of changing measurement noise through time (and hence the (anti)correlation between AC1 and variance). We maintain that the influence of time-variant noise levels will be expressed in multi-sensor data regardless of the data fusion scheme chosen. It is likely that a data fusion methodology could be created to mitigate this error, but the development of such a method is outside the scope of our study.

Fig. 2. Plot titles covered the words in (d), (f), and (h).

We are not sure which words you are referring to here – the plot titles do not cover any text. It is possible that you meant the x-axis label for (d), (f), (h)? Since those subplots all share an x-axis, we only labeled it on the bottom panels.

Fig. 2 and Fig. 3. Why use a five-year rolling window? Why not a longer or shorter window? Will the change of moving-time-window have an impact on your conclusions?

A five-year rolling window is fairly standard processing for time series of this length (Boulton et al. 2022; Smith et al. 2022) and since we investigate potential effects of the satellite composition on these previous works we decided to take the same window length. A longer or shorter window will not influence our results – the key point here is the time-variant noise levels. Whether or not those signals are found over shorter or longer windows, they will still influence inferred changes in resilience. We have added a version of Figure 2 of the MS with a 3- and 7-year rolling window to this Reply to illustrate this.

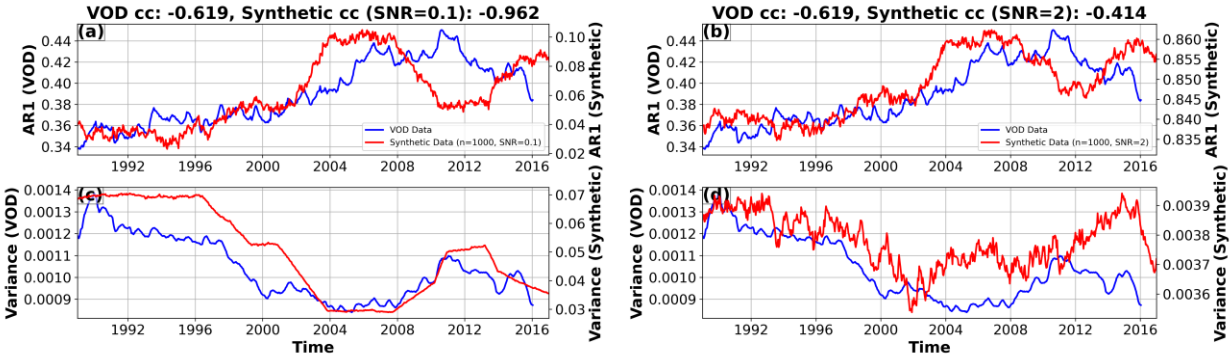

Figure 3: Same is MS Figure 3, but with a 3-year rolling window.

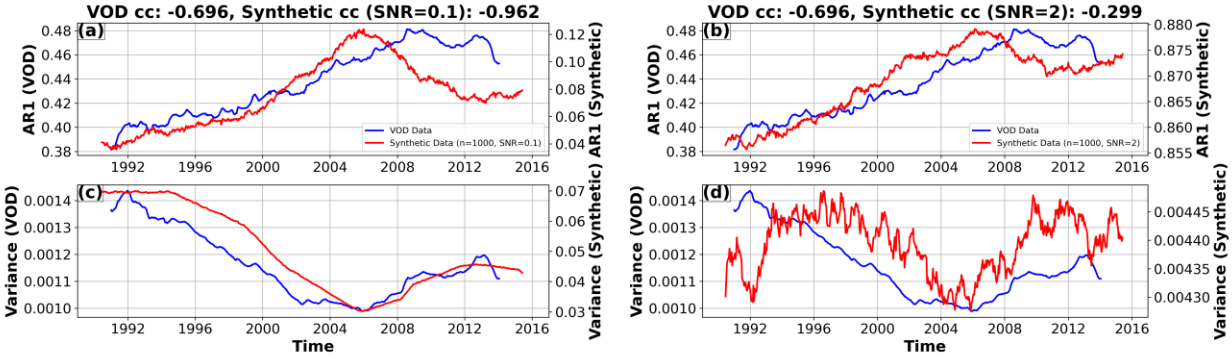

Figure 4: Same is MS Figure 3, but with a 7-year rolling window.

References for this Reply

Boers, Niklas. "Observation-based early-warning signals for a collapse of the Atlantic Meridional Overturning Circulation." Nature Climate Change 11.8 (2021): 680-688.

Boulton, C. A., Lenton, T. M. & Boers, N. Pronounced loss of amazon rainforest resilience since the early 2000s. Nature Climate Change 12, 271–278 (2022).

Feng, Y., Su, H., Tang, Z., Wang, S., Zhao, X., Zhang, H., Ji, C., Zhu, J., Xie, P., and Fang, J.: Reduced resilience of terrestrial ecosystems locally is not reflected on a global scale, Communications Earth & Environment, 2, 1–11, 2021.

Moesinger, L., Dorigo, W., de Jeu, R., van der Schalie, R., Scanlon, T., Teubner, I., and Forkel, M.: The global long-term microwave Vegetation Optical Depth Climate Archive (VODCA), Earth Syst. Sci. Data, 12, 177–196, https://doi.org/10.5194/essd-12-177-2020, 2020.

Pinzon, J. E. and Tucker, C. J.: A non-stationary 1981–2012 AVHRR NDVI3g time series, Remote sensing, 6, 6929–6960, 2014.

Smith, T., Traxl, D., and Boers, N.: Empirical evidence for recent global shifts in vegetation resilience, Nature Climate Change, 12, 477–484, 2022.