Referee #1

This manuscript addressed an interesting topic by investigating whether measurement noises can impact the inference of resilience using remote sensing data. They used a simulation approach to investigate how signal-to-noise ratio (SNR) influences the calculations of two indicators of resilience, namely lag-1 autocorrelation (AR1) and variance. Their results have implications for assessing the possible impact of measurement errors in observational data. Overall I found this study well designed and conducted. I have a few comments that may help improve the paper.

Thank you for your time with the manuscript and the helpful comments! We will address each comment individually below.

(1) the study generated simulated time series by combining a background time series with a number of random noises. I was wondering if this characterized the realistic errors introduced by changing instruments. I was not an expert in remote sensing, but I thought in some cases the change of instrument might induce sudden increase/decrease in the time series (rather than a small noise term). Such changes can have major impacts on the calculation of resilience indicators, keeping in mind that such indicators were used to detect 'sudden changes' in the time series, whether they were due to measurement errors or underlying processes?

Thank you for this comment. Sharp jumps are indeed a potential issue in time series obtained from combining signals from different sensors that are active over different time periods. There are many different studies on how and when to cross-calibrate long-term data to handle this. For example, in the VODCA data set that we attempt to model, biases between sensor absolute values are rectified with a CDF matching approach.

Given that problems arising from mean-shifts can be handled comparably easily, in our synthetic set up we address problems that go beyond a mean-shift. We assume that any composited multi-satellite data has had such shifts removed; our analysis rather focuses on what is *not* well treated in such common data compositing/calibration schemes – namely changes in higher-order statistics such as variance or autocorrelation.

(2) AR1 and variance are two important indicators of resilience, or early warning signals (EWS) for catastrophic changes, but there are more. Moreover, researchers had been developing composite EWS by combining different metrics. Given that measurement errors may influence AR1 and variance differently or in opposite directions, I was wondering if a composite EWS would be more robust to measurement errors.

You are right that some studies have proposed combinations of the Variance and AR1 coefficient, as well as other indicators such as Skewness or Kurtosis. In the present case, we focused on the Variance and Lag-1 autocorrelation as the underlying theory provides clear equations relating them to the recovery rate λ and – in this sense – to the resilience of the underlying system: $\langle x^2 \rangle = \sigma^2/(2\lambda)$ and $\alpha(1) = \exp(-\lambda \Delta t)$. We believe that in order to obtain a detailed understanding of the effects of changing sensors on these indicators, it is best to keep them separated and not combine them, at least in the context of our study. Formulating a combined resilience indicator would make it more difficult to attribute the individual effects of the changing sensor mix to changes in system parameters (e.g., system memory – autocorrelation, system variability – variance).

(3) the authors discussed about the difference between the average of variance from individual time series and AR1 of the aggregate time series, particularly their different behaviors in the presence of measurement errors. Similarly, the reference Feng et al. (2021) found different temporal trends of these two metrics. However, these two metrics represent different properties (i.e., local- vs. larger-scale resilience) and they did not necessarily exhibit different patterns, even if there was no measurement error. The problem is, the local-scale variance did not add up to give the larger-scale variance, but modulated by the synchrony between local grids. I attached a theoretical paper illustrating this:

Wang, S. & Loreau, M. Ecosystem stability in space: α, β and γ variability. Ecol. Lett. 17, 891–901 (2014).

Thank you for providing that reference – this is a very interesting piece of work! This is something that we did not really consider in our work: how does the variance of individual time series match up to the aggregate variance of many time series in the absence of problems due to combining data from different sensors. Our focus was rather on how resilience analysis has typically been presented (e.g., with regional mean time series), and what problems there might be with that approach given data with dynamic uncertainties. It is not unexpected, however, that the sum of many individual variances does not add up to a single mean-series variance – these two cases would be exposed to quite different noise levels (e.g., random fluctuations are suppressed in a multi-series mean). While we find this a very interesting problem for further exploration, we feel that multi-scale and spatially-conditioned changes in resilience are outside of the scope of this current work. We will, however, add an explanation in the revised discussion that differences between the variance/AR1 of aggregated time series and aggregated variance/AR1 time series may also be due to differences between the local- and larger-scale resilience.

(4) while the manuscript was overall well written, I had to say that I was confused by the different metrics involved in the figures, which seemed to be quite related but differ in important ways. For instance, the authors calculate resilience indicators using several approaches, e.g., deriving the numbers for an individual time series, first aggregating the time series and then calculating AR1 and variance, or first calculating AR1 and variance and then averaging them. They also calculate correlation between AR1 and variance at different levels of complexity. I would suggest to add a table to clearly define all key metrics in the figures, with explanations what a positive/negative or higher/lower value mean.

Thank you for this suggestion – we endeavored to keep our terminology as clear and explicit as possible, but it remains difficult to parse in some places! Adding a comprehensive table is an excellent suggestion, and we will do so if a revision of the paper is requested.

Specific comments:

L52: What does 'synthetic series' mean? I think it is simply a simulated time series.

Yes, we refer here to simulated time series. To keep our language specific, we refer to each individual series as 'synthetic', and when we produce multiple realizations each is referred to as a single 'simulation'. We will make sure this is clear in a revised version of the MS.

L79: How was this 'aggregating' implemented?

In each case of aggregation, we use either a time-explicit (e.g., daily) mean, or a time-explicit maximum value. To simulate VOD, we aggregate multiple instruments with a mean. To simulate AVHRR, we take the multi-week maximum, to match the method used in the original data set. For mixing multiple multi-instrument series, we use a time-explicit mean (e.g., averaging 100 synthetic series into a single averaged series). We will clarify this in a revision.

Figure 2: Not sure that I understood these figures correctly. Did the 'median corrcoef median signals' on the right represent the median of the 'corrcoeff of median signal' on the left? Why they were so difficult, even by sign?

This is admittedly a rather dense figure. We think that the suggestion of the referee to add a table will greatly help in describing the differences in the metrics and making this figure more accessible. On the left, AR1/Variance are first calculated, and the median of 1000 iterations is displayed. On the right, we take the median signal from 1000 simulations, then calculate AR1/Variance on that. We perform that calculation 100 times, and display the median of that result on the right panels. The difference between the columns comes from when the data was aggregated – aggregating many time series and then calculating AR1/Variance produces a significantly different result from calculating many AR1/Variance series and then aggregating those.

L130: "the correlation between AR1 and variance is generally positive for individual synthetic series" – any result supporting this argument?

This refers to the labels of the left hand column (Fig 2) and the histogram shown in Figure 4 – for higher SNRs, the median correlation coefficient between individual-series AR1/Variance is positive, while it is negative when those AR1/Variance series are instead aggregated. We will update the reference on this statement to be clearer.

Figure 3: I must miss something. How to determine the SNR in the real data? And did it happen to exhibit SNR = 0.1 and 2 in the empirical data?

Determining the SNR of the real data is indeed challenging – there are many factors that influence sensor noise, and most change dramatically in space and time even for individual sensors. Cloud cover, atmospheric water content, and time-of-day of satellite overpass are a few key influences on local-scale sensor noise. Since we cannot quantify sensor noise effectively in space and time, we instead aimed to construct a synthetic experiment which mimicked the global-scale patterns seen in VOD (e.g., blue lines) using only noise-level changes (red lines). SNR is then a secondary factor – this allows us to vary not only the relative noise between sensors (e.g. through time) but also the amplitude of that noise relative to the underlying signal. The values of 0.1 and 2 are thus not related to the SNR of the real data, but rather constraints we placed on the synthetic data to illustrate the potential influence of not only changes in noise through time, but also how the amplitude of that noise plays a role. Note that we deliberately chose very low values for the SNRs; in reality, a sensor with SNR equal to 0.1 will of course not be considered useful, and even SNR = 2 is still far lower than what is reported for operational sensors. MODIS aims for SNRs of individual bands between ~70-1000

(https://modis.gsfc.nasa.gov/about/specifications.php); Landsat 8 OLI aims for SNRs of at least 100 (https://landsat.gsfc.nasa.gov/satellites/landsat-8/spacecraft-instruments/operational-land-imager/oli-requirements/); Sentinel-2 aims for SNRs around ~100 (https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi/resolutions/radiometric). Our values of SNR from 0.1-2 are rather related to our synthetic experiment than trying to match real-world data.

L242: what does it mean by 'aggregated'? You had explained that first aggregating time series and then calculating AR1 and variance can remove the influence of changes in satellite instruments to some extent. So did you mean 'first calculating AR1 and variance and then taking the averaging of these metrics'?

Yes, in this case we mean that when many AR1/Variance series are averaged, changes in the noise structure through time are emphasized. Conversely, first averaging many series and then calculating AR1/Variance on the resultant averaged series reduces the impact of changes in sensor noise through time. We will clarify the use of the term "aggregation" throughout the manuscript in a revision.