Biogeosciences Discuss.,
https://doi.org/10.5194/bg-2017-130-AC2, 2017

Biogeosciences
Discussions
Open Access

Interactive comment on "Detecting impacts of extreme events with ecological in-situ monitoring networks" *by* Miguel D. Mahecha et al.

Miguel D. Mahecha et al.

mmahecha@bgc-jena.mpg.de

Received and published: 3 August 2017

Bold text has are the comments from the reviewers. Standard text is our response. *Italics* are text elements from the revised paper.

Dear Anonymous Referee,

Many thanks for your detailed review of our manuscript, your work is highly appreciated. Please find below our detailed responses to your comments:

This study looks at the detectability of extreme events using existing ecological networks and numerical experiments. I think there is some sound science here but several issues (many are language-related) need to be addressed to make this suitable for publication in BG.

The language was also criticized by reviewer 1. Indeed, none of the authors is a native speaker. Sorry for the inconvenience! In collaboration with a native speaker we have done our very best to revise the paper (not the responses to the reviewers tough). Please note that also Biogeosciences subjects each paper to a language check after acceptance.

Specific comments follow . . .

P1L10: I'd like some other yardstick to gauge why I might care about the largest 8 vs. largest 39 events. How many sigma departures? Something more than a simple number would help the abstract. Right now the reader has little idea if these are only statistically as opposed to scientifically interesting.

The departures in term of standard deviations are happening locally (e.g. we check each voxel from it's departure from the regionally expected values). However, these "sigma departures" are independent from the total size i.e. affected area or impacts. Hence, we don't understand what the reviewer mean with "How many sigma departures?". The largest events are simply those who matter and the well known fact that extremes in the terrestrial biosphere happen to be power-law distributed naturally leads to a necessity to focus on the x larges events. The numbers 8 vs. 39 sound arbitrary, but are a direct result of comparing the 90% vs. the 50% chances.

P2L13: To a large extent is unnecessarily vague. Please add some actual information here.

We have chosen three examples studies and specified their applications. This reads as follows

Earth observations (EOs), especially satellite remote sensing data, encode relevant information on anomalous ecosystem functioning (Preifer et al. 2012; McDowell et al. 2015). Examples include the exploration of soil moisture anomalies in tandem with climate patterns to understand anomalous vegetation responses (nicolai-Shaw et *al. 2017), snow cover induced albedo anomalies with consequences for local climate /Chen et al. 2017), and the impact of weather extremes on vegetation indices to track anomalies in productivity and explain vector-borne disease outbreaks (Anyamba et al. 2014), among many others. The consistent and contiguous spatiotemporal data coverage, and, more importantly, the fact that observations of the land surface typically integrate a plethora of processes, make EO very attractive for detecting extremes affecting the land surface.*

P3L3: Please redo your objectives. It's just poorly written, I have no idea where we're going from here–until I re-read this paragraph the third time. That is twice too many. A simple declarative roadmap would be most helpful here.

We agree that this part should be rewritten. We now suggest the following text for the manuscript:

In this paper we aim to understand the potential of ecological in-situ networks of varying size for monitoring the impact of extreme events. This paper addresses this issue in three steps: 1) We propose an approach for detecting extremes that are of regional relevance. This step is important to avoid a bias toward considering extremes that take place only in high-variance regions, and may be a relevant contribution beyond our application. 2) We explore a series of random networks of varying sizes to explore the expected detection rates. We aim to understand the observed patterns using probabilistic approaches and formulate a theoretical expectation of detection probabilities of extremes. 3) We then analyze the detection probabilities in two real networks (NEON and Ameriflux) and compare these to random networks of identical size. The paper concludes with an outlook on how our remarks could lead to improvements in network design that could be implemented to improve the detection of extreme events.

P4L14: What do you mean considered? Did you use it or not? If I consider traveling to London over the weekend I may or may not go. That is not what you mean to imply here?

Ok, the paragraph is now clearly stating what we used. See also comment by and responses to reviewer 1. The new text starts as follows: *We use the geographical locations of eddy-covariance flux tower networks but to the actual measurements. Our main target is* ***FLUXNET****, a global collection of eddy covariance data collected (for in-depth descriptions see Baldocchi 2008, 2014, www.fluxdata.org) . . .*

P5Fig2: Unclear. Especially the "within each mesh cell" bit. Why are the central dots and assigned percentiles interesting?

What is interesting here is the derivation of the regional threshold. In order to also address the comments by reviewer 1 we now moved this figure to the new appendix A. This allows us to explain the approach in more detail there.

P6L1: I'm not sure I see your argument for local vs. global thresholds. Several EO products normalize by location to obviate this issue. And, of course, you do not define "implausible and undesired".

Our point here is to develop a method that is generally applicable. Of course, if a specific EO product is normalized this can be omitted. We argue, however, that many variables don't come with an intrinsic normalization. For instance, the "Normalized Difference Vegetation Index" is constrained between -1 and 1, but may always remain very low in some arid regions, or very high on others. We want to avoid that all values is either of the areas would be classified always as extreme. Hence, we seek a compromise and this compromise needs to be regional. We have now rewritten this section to motivate this approach better (and, of course, avoid writing about "implausible and undesired" effects):

The question of how to define extreme events in spatiotemporal data cubes is key to the evaluation of the suitability of ecological in-situ networks. One approach could define some global threshold and identifying values exceeding this threshold as potential extremes ("peak over threshold"). Choosing a global threshold setting is suitable when the question is how extremes add up to global anomalies (Zscheischler et al. 2014), i.e.

when one is working with the extensive data properties where the target is the integral over space and time. However, the consequence of setting a global threshold is that values that are flagged as potential extremes will occur exclusively in high variance regions. It could happen that one rather flags entiere regions that are extreme compared to others as extremes, while other regions would apparently never face such an event. An alternative would be using only local thresholds (defined over time at each spatial point x_{uv}). However, the latter would necessarily lead to an equal spatial distribution of extreme event occurrences. This is also not what we are searching for. We want to define extremes relative to regions that are characterized by a similar ecophysiology e.g. we want to compare each grid cell with grid cells that have a comparable phenology and search for extremes across these geographical locations. However, as our approach should be entirely data driven we refrain from using precomputed definitions of eco-regions.

P6L16: How this characteristic FAPAR anomaly is assigned is still a mystery... You never say how the estimation occurs.

As stated above, we have now added a new appendix A to explain the procedure in detail.

P622: If you want clusters why not cluster the MSC directly? And I'm curious how this improves upon other maps of similarity. Why not use MODIS PFTs as is. Or Koeppen-Geiger. What have you gained by this exercise?

As written in appendix A (see response to reviewer 1), we could indeed cluster the MSCs and considered this at the very beginning but we found two complications that we were able to avoid now: 1) Clustering is leading to a non-uniform partitioning of the space spanned by the MSCs and computationally more expensive. 2) it does not allow us to easily assess which are the neighbouring clusters in order to consider these values for the assignment of the thresholds. Our approach has the advantage that we can compute it on a subset of values (rendering it very efficient), and leading

to a uniform grid that allows us to efficiently include the neighboring "clusters" for the estimation of the threshold of the central cell and is controllable in terms of the variance represented and computationally extremely efficient even on this very large data set. MODIS PFTs or Koeppen-Geiger classes would be possible to use, but are very very coarse and don't reflect the details we see in Fig 1. Hence, these classifications would lead to a very coarse thresholding that is by no means comparable to the continuous threshold as shown in Fig 3 (Fig 2 in the revised version of the paper).

Given that reviewer 3 has likewise emphasized this aspect, we have also expanded the discussion to justify our approach. This paragraph read as as follows:

*More specifically, regarding the details of the chosen methodological approach, one may question why we propose simply binning the leading PCs derived from the MSC of our EO. This approach was mainly developed to effectively deal with the very high resolution of the underlying data, seeking a very efficient subgridding approach. One alternative would have been to e.g. cluster the PCs directly. However, besides the computational costs, conventional clustering methods lead to a non-uniform partitioning of the space spanned by PCs. This non-uniform partitioning makes it slightly more complicated to identify neighbouring clusters, which is necessary to stabilize the quantile-based computation of anomaly thresholds. Having an equal meshgrid over the PCs that we can also compute on a subset of MSCs renders the approach very efficient for very large data sets and is completely data adaptive. It was very important for this exercise to have many small classes, in order to compute a very well regionalized anomaly threshold (shown in Fig. **??**), which would not have been achievable using classical climate classifications of ecoregions. A more detailed follow-up study should explore the question of how the choice of the various parameters affects the event detection accuracies. A crucial question in this context will be whether one can tune these parameters such that a baseline of events is well detected.*

P6L11: How well do the top 3 work?

We guess that the reviewer is asking how much variance we explain with these components? The leading 3 PCs explain more than 70% of the variance in the data.

P7Fig3; How is "we may classify an event as extreme in one ecosystem that would be considered part of the normal variability elsewhere" interesting? This is hardly new information! And there are no percentiles here?

We see that there is an error in the caption of the figure that leads to the confusion: The caption was saying “Map of the regionally varying percentiles used for detecting extreme events.” but in fact that percentiles (or, more precisely quantiles) are constant for each region as defined in the space of the leading PCs i.e. q = 0.025. What we see here are the regional FAPAR anomaly thresholds that correspond to the quantile. We now added the quantiles to the description of the figure to clarify this. We hope this solves this misunderstanding.

P7Fig4: I like this. Good visualization of the sparse network detection problem!

Thank you.

P8L3: Would like an example of z here.

As defined in the paper, z is simply a spatial search extend in units of the data e.g. km, or degree. In section 3.1.2 we define the chosen value of z, which is 5km.

P8L1: Here the reader is confused. You spend much time on the PCA/cluster bit such that your similarity mask allows for spatially non-contiguous blocks. But here we are doing direct neighbors? While I think I understand the thought process here it needs additional detail. One might wonder why bother with PCA/cluster if you are doing direct neighbors as, for nearby cells, once can reasonably "assume a similar responsiveness to some extreme event " (from P8L11).

Sorry, if this part was not sufficiently well explained. The PCA space is only used to classify homogeneous regions in the data space. For each of these homogeneous

regions we apply the same threshold for defining what an FAPAR anomaly is. This leads to a flagging of values that we consider anomalies beyond a threshold. The result is a binary data array where values can be 1 or 0, i.e. extreme or not. What we do next is to ask if neighbors (defined by the search space) are likewise flagged as 1 or not. We have added a sentence to distinguish these two aspects.

P9L2: Did you look at this alternative? As I read on I have the impression that your workflow has so many tunable (hyper)parameters that I am already doubting the results from this one set on offer. Did you do some grid search? Across algorithms and hyperparameters? You can tweak how you arrive at the clusters, how you set thresholds (still unclear), and the z and tau params?

You are indeed right that there are many tunable parameters and we have now added a critical appraisal of this fact to the discussion. But we have actually no objective criterion to define an “optimal parameter set” except that we assume that a quantile of 0.025 is sufficiently “extreme” to identify the values we seek. Our approach detects very prominent heatwaves like the 2003 event in Europe, providing some confidence to the choice of parameters - but even there the question is what would be the correct definition. Given that we cannot optimize the parameters, all choices are inherently arbitrary. The specific resulting values will depend on the choice of parameters (e.g., a less extreme threshold will lead to larger, more connected extreme events). However, we don’t expect the specific choice of parameters to have a significant impact on the conclusions of the paper.

P9L5: Still unclear on the threshold. I like this list of settings. It’s a much needed distillation of the methods above. Your grain size concept here is more informative than the above figure.

The reviewer is right - the threshold should be defined here. We have changed the list as follows to be more precise: *In summary, in this study we used the following settings:*

- *Mean seasonal cycles computed over a time-span from 2001 to 2014.*

- *The first three PCs binned using a grain size of 4% of the range of the first PC.*
- *For each bin in the PC space and its surrounding 26 cells we estimate the quantile $= 0.025$. The FAPAR-anomaly values corresponding to this quantile are assigned as the threshold for the grid cells corresponding to this central bin.*
- *The search space for detecting extreme events is parameterized with $z = 5$ and $\tau = 1$ corresponding here to a search space of ± 5 km and ± 16 days.*

P10Fig5: Maybe it's a language thing but if b is anomaly while is it labeled as (reduction of m2 with FAPAR = 1? Fix the missing) in any event.

This is just a way to quantify the impact: As FAPAR is unitless, it's anomalies are multiplied by the areas of the associated pixels ... or one could say how much area has an FAPAR value of one. But this is indeed confusing and we omit this now in the new figure labels.

Also, why does 103 get a black line? (I read on and see this point is discussed but do put this [briefly] in the caption as well.)

An earlier critique at a conference was that people wanted a "good" number of sites on display and not just the continuum. Hence, we have picked the one which was closest to the 100 to be highlighted. In fact we also refer to this line in the abstract. The fact that it is 103 and 100 is maybe strange, but we also don't see the reason why 100 should be better.

P10L1: So Fig 8? What about Fig 6 and 7?

As explained also to reviewer 1, the reference to Fig. 8 was inherited from a very old structure of the paper where Fig. 8 came first. We have removed this reference here now as it is indeed not helpful.

P10L8: How is this different. Seems like the same message here.

Fig. A1 (in the revised version B1) is very different in that it shows the detection probability as a function of network size and not as a function of the event size. But of course this has the same message. From the discussions among coauthors we have simply seen that for some people it is more intuitive to think the one way or the other, so we believe it is worth keeping both figures in the paper - one the in the main part, one in the appendix.

Also, "that one would need to inflate in-situ networks by orders of magnitude to detect small scale events at comparable rates than large-scale extremes" is hardly new insight. We've known this for some time. If the Earth is hit by a large asteroid, a single sensor will be just fine. If several smaller objects rain down more sensors are needed to track them all. You are just restating common sense here.

We agree that in some way this is common sense - but our question is how to describe this transition? One innovation of the paper is to explain this phenomenon for extremes in the terrestrial biosphere in a very accurate way. We show that estimating the distribution of extremes helps to estimate how fast the decay rate is allowing us to quantify how many towers are need to detect an extreme of a given size x with some probability larger than x%. In view of this argument, would respectfully disagree in removing this sentence from our manuscript.

P10L14: Here you address a departure from an idealized case and invoke spatiotemporal correlation. This is fine but I would outsource this to an Appendix and state in the main text that deviations in detection probabilities relative to theoretical expectations are driven by spatiotemporal structure in the dataframe and move on. This really breaks the story as written. An alternative is to include this issue in your objectives as another question. It reads tacked on in any event as is.

We guess that the reviewer is referring here to section 4.1.2 (and not to P10L14, where we actually discuss the general case - not the departure from it). We agree that this makes the paper quite complicated and we move this part to appendix B, and only keep a one-sentence summary in section 4.1.

P13L18: Same above comment applies here. Here you might frame this as a means to more efficiently design a network. This could foreshadow your conclusions a bit better too.

This part is important for us, as it reveals the link between the power-law distribution and rank. But we now add a comment explaining the implications for the network design aspects:

In other words: gaining insights about the scaling behaviour of the extremes can be used to formulate clear expectations about event detection probabilities of a given rank and size.

P14Fig7: If caution is warranted, why show it? Also, why are the edges black. The band across Fenno-Scandinavia is odd. Does your workflow simply create such zones at the domain edge?

The method has no edge problem. The band at high latitudes comes from low-quality data that are not used in the study. We have, however removed this figure as we also don't see much added value here.

P15L5: "We therefore ranked the largest 100 events detectable in continental US by their integrated FAPAR anomalies." I still am missing some information on if these are ecologically relevant. Why should a network be configured to detect a class of anomaly (no matter how real, in a numerical sense) that has no impact on land ecosystems? I am not saying that is the case here but you never really address this point. Put another way, an FAPAR departure does not linearly map onto, say, a carbon/water/energy anomaly.

We agree that an FAPAR anomaly does not linearly map into a carbon/water/energy

anomaly - but still, land-atmosphere fluxes are tightly coupled to biophysical variables like FAPAR. And there are good reasons to at least scrutinize the in-situ observations in depth, when a strong FAPAR anomaly happens. Hence, we respectfully disagree with the reviewer here and the introduction of the paper clearly explains why anomalies in EOs are a relevant means to detect and describe extremes that affect land ecosystems. We have now added a sentence in section 2.1 explain why FAPAR is actually relevant to monitor extremes in the terrestrial biosphere. Give that FAPAR is directly (even if not linearly) coupled to GPP, we can actually use its anomalies as proxies for suspecting anomalies in GPP and other direct observations.

P16Fig8: Percentile ranges from some bootstrap scheme? How many times? I don't recall seeing that.

We have explained more explicitly in the caption of the figure that the range comes from 100 random tower placelements - each of the size of the real network to compare with.

Also, at this point in the paper I see two options. You either explicitly retool this as technical note (that means shortening the paper; this is the easier path forward). *OR* You embrace FLUXNET as is. That is, on Fig8 why not add the other regional networks, the 2015LaThuile? What level of anomaly can we detect with what we have? How does this vary across the globe? Is, say, India better covered than Africa? What can we learn about optimal designed networks and how does that vary based on the type of anomaly we wish to detect. You can do all this, it's the same workflow, FAPAR is global, and tower coordinates are public domain. You could use greenness/NDVI to get back to 1982 and address some methodological shortcomings as well. To be clear, I am not asking you to write the paper I'd maybe like to write. I dislike that in reviews and would not advocate that herein. My point is that as is you have, in the main, a technical paper. So either embrace the technical note idea (again, the straightforward path) or grow the scope and relevance (this would entail more but would likely be more

impactful on the field). As is the paper sits awkwardly between. (In all fairness, some of this thought on my part is motivated by language issues herein.)

The reviewer raises multiple aspects in this comment. The overarching question is if this paper should rather be published as “technical note” and not a full research paper. We would like to clarify that we submitted this paper to the “Ideas and perspectives” section because we think that we can offer both. But if the reviewers and/or editors of Biogeosciences consider that our paper is rather a “Technical Note” we could accept this and have no clear preference. This is simply an editorial decision. In fact, we believe that it is not very useful for the key message to extend the analysis to the entire La Thuile sites. Areas like India and Africa have so few sites that we can hardy detect any events there - except if they would be of continental relevance. And an analysis of this kind would indeed lead to another paper. Another question that comes up here is why we use FAPAR and not some data set dating back to 1982. This choice is because the respective long-term EOs originate from multiple sensors and we are still investigating if this would not lead to artifacts in the event detection. We are currently more comfortable using one consistent spatiotemporal dataset.

FigA4/5: Why does the legend obscure the lines?

Because of a weird Matlab property: If we could place the legend outside the subfigure, it would dramatically reduce the subfigure size. However, given that the lines behind the legend are – as visible in all other subfigures – just straight lines, the reader does not miss any information.

AppB: This needs to be better incorporated in the paper. Right now, it really reads as an add-on that yields but marginal insight.

We have restructured the discussion. This is now the last part, where the soil moisture case is used to substantiate yet another finding: the value of uninterrupted measurement campaigns substantially increases the event detection rates.

Language: This article suffers from several language faux pas and other overall awkwardness. For example, in P1L11 we read "These finding are". This is of course quite wrong.

Indeed there were a series of typos in the paper, but we have revised the language of the entire paper now. So we hope that this issue is solved now.

There are many others; some in specific comments. But overall there are too many and I (mostly) stopped after the Introduction. Please have this article proofread by a native (or native-level) speaker of English before resubmission.

The paper was now checked by a native speaker. Biogeosciences offers yet another proofreading service in case the paper will be accepted.

The language issues are an unneeded distraction from the actual science. One stylistic quibble is that the paper reads like a math exercise in search of a case study. I would have rather read a paper that seeks to improve our detection abilities wrt extremes in land systems.

We respectfully disagree here: We came across a few basic statistical effects that should be known to the community. We see that we offer a set of statistical tools for quantitative network design, even if we don’t actually solve the problem and offer a final solution. Network-design is still a complex optimization task and here we explain how to consider the design problem when the aim is to specifically detect extreme events.

We have carefully worked on all other editing comments. We acknowledge the reviewer for her/his efforts!

Interactive comment on Biogeosciences Discuss., https://doi.org/10.5194/bg-2017-130, 2017.