Biogeosciences Discuss.,
https://doi.org/10.5194/bg-2020-374-RC1, 2020


Biogeosciences
Discussions

# *Interactive comment on* "Functional convergence of biosphere–atmosphere interactions in response to meteorology" *by* Christopher Krich et al.

**Anonymous Referee #1**

Received and published: 27 October 2020

Review of Biogeosciences 2020-374

Title: Functional convergence of biosphere-atmosphere interactions in response to meteorology

General Comments:

The manuscript, "Functional convergence of biosphere-atmosphere interactions in response to meteorology," investigates a number of variables and their connections from publicly-available FLUXNET datasets using a relatively novel causal analysis method called "Peter Clark Momentary Conditional Independence" (PCMCI), in conjunction with a dimensionality reduction technique called "t-distributed stochastic neighbor embedding" (t-SNE) and a subsequent clustering algorithm called "Ordering Points To



Identify the Clustering Structure" (OPTICS). The specific research questions motivating the study are not clearly stated; the general motivation provided is, "to investigate how biosphere-atmosphere interactions vary across vegetation types and climate zones." This the manuscript accomplishes through a notion of linkages between biosphere and atmospheric variables, with the primary units of analysis being 1) network representations of those variables and their causal interactions over three-month windows at daily scale, 2) a two-dimensional representation of the structure of those high-dimensional networks, and 3) clusters formed within that 2d space of high dimensional networks.

The methods will likely be unfamiliar to most readers, and the units of analysis are quite abstract and require considerable explanation for readers to fully grasp the results being presented: this explanation is not currently sufficient in the manuscript.

Broad discussion focuses on some very interesting topics, such as 1) the universality or functional convergence of biosphere/atmosphere processes, 2) trajectories of ecosystems through a 2d space of land surface "network" states, including seasonal cycles and deviations due to extreme events, and 3) linkages between biosphere and atmospheric variables, and how their causal relationships could be represented as clinal processes along some continuum from linked to unlinked. Ultimately though, this discussion turns back to separating water/energy/radiation/temperature limitations on ecosystem productivity from land-atmosphere feedbacks (both of which are areas of deep physical research), which leads the reader to ask what the analysis gains from combining them in the first place.

While providing an interesting lens for looking at highly complex interactions between the biosphere and atmosphere across time and space, I found that this study failed to specify its intents and rather motivated too much using the tools (which instead should be motivated as useful for answering the question at hand). This led to results for which I am hard-pressed to find applications. I am not convinced that sufficiently substantial conclusions have been reached. I recommend a major structural overhaul of the paper, driven by specific, answerable scientific questions. At the same time–and

this is the difficulty in a study with such “boutique” methodology (with no judgement passed on that label)–the readers will still need *more* description of what is being shown in the analysis, all leading back to the primary research questions. I have tried to provide as specific guidance as I can in the following comments.

Specific Comments: What is/are the primary research question(s) being asked here? What is the knowledge gap?

The fundamental units of analysis need to be very clearly specified, as readers will be unfamiliar. Each point in Figure 1 is a network of connections between a bunch of variables at daily scale, but each representing three months of data, including some lagged effects. This network is the primary unit of analysis. It would really help to show an example of one of these networks at one location for three months before jumping into Figure 1, even though the authors have written papers on these networks before.

The authors need to discuss seasonality at some point earlier in the analysis to let readers know that all seasons will be studied, and that points in Figure 1 will represent different locations and different times of year. Line 80: (Probably an easy, but major point) “A comprehensive description from theoretical assumptions. . .” These assumptions should be stated clearly here, as the method is not well-known. As with any paper using basic regression analysis, a statement of the ways in which the analysis meets basic methodological assumptions is necessary. Rationale/justification for using the method when assumptions are not met are necessary as well. Some of this discussion can happen in supplementary materials if it is particularly involved, but the assumptions and their validity should probably be stated in the main text.

Line 96: “Unobserved common drivers can still render links as spurious.” How do the critical *non-stationary* variables (at the time scale of your analyses) of biomass and phenostage influence 1) the validity of the estimation of your networks, and 2) the structure of the 2d space in Figs 1 and 2?

Line 105: “subtracted a smoothed seasonal mean from each variable. . .” I agree that



this needs to be done to remove non-stationarity which can cause spurious correlations. At the same time, subtraction of a Fourier series from a time series could either solve that problem or partially solve the problem while introducing new non-stationarities. How robust is the de-seasoning technique? Do the results change if you use other filtering methods?

Use of PCMCI, t-SNE, and OPTICS really makes this difficult for readers to follow the methodology. I would guess almost no one (particularly outside the author’s list) is familiar with all of these. The authors need to motivate why they are using these methods with respect to some research question, and not just because causal tools exist.

Line 156: “The strongest gradients measured via distance correlations. . .” As the manuscript stands, I don’t think even the most careful methodologically-focused readers are going to know how to interpret these results. It took me a lot of re-reading to get the idea that the distance correlations represent the spatial (in this 2d space) coherence of the link strengths. The link strengths themselves should be more clearly explained and motivated, probably in a preliminary figure showing an example network. The meaning of the link strength should be clarified (does link strength 1 mean fully causal? Completely dependent? One-to-one?)

Figure 1: As the manuscript stands, I do not think readers are equipped to understand what is being shown in this figure, which needs to change. While some methods may be dense and opaque, results need to be comprehensible to readers in the field, even if they are not close enough to the sub-field method specifics.

While Szekely et al. 2007 is highly-cited in the statistics literature, it is unlikely that many of your readers in biogeosciences will be familiar. How do we interpret these distance correlations? Having referred to Szekely myself, I can see that the correlations are metrics of dependence between random vectors, but can you clarify what are the vectors in question here (say for NEE-LE)? What is their dimensionality, what are the

constituent dimensional components? Are they across space and time (I think so) and season of the year (I don't think so), and if so, how do these constituent components combine to give a single number (rho=0.75)? Does this represent something like a fraction of explained variance, and if so across what conditions? Can I compare the rho for NEE-LE and the rho for T-H to infer something about bivariate coupling? What time-scale should I think about these metrics representing? Mostly daily? Does Rg-> H mean that Rg almost always causes H (with positive partial correlation)? Does T-VPD being red mean that T causes VPD with positive correlation or that VPD causes T with negative correlation? Your readers need their hands held through all of this to interpret your results and see the patterns you are seeing in your analysis.

Figure 1 caption: "As Rg can only be a cause. . ." Is this true? I'd imagine that LE -> Rg often if Rg is measured at the tower (as opposed to top of atmosphere). There are certainly LE -> humidity -> cloud formation processes at the local scale in many locations, aren't there?

Line 156: "The colouring reveals that the link strengths are ordered along gradients." This sounds like A Finding, but of course the world is spatially autocorrelated and neighborhoods are similar. What is it explicitly that is interesting about this? Is it expected or unexpected (I would expect it) and why?

Figure 2: This figure could in theory be used to add interpretability to Figure 1, but otherwise the main take-away is that GPP, NEE, and LE are fairly correlated, as are Rg and T. I enjoy looking at this for patterns, and I can imagine spending time in a study looking for structure and emergent relationships in this data projection, but I am not sure what "results" it represents. How could I as a biogeoscientist make use of the information in this Figure? What question is this helping to answer?

Line 165: "The results show that a high dimensional space encompassing more than 10000 ecosystem networks representing the states of biosphere–atmosphere interactions from ecosystems of various geographic origins can be reduced to a compact two-



dimensional manifold characterized by four edges and gradients of biosphere and atmosphere conditions." Maybe I'm missing a key piece of nuance here. It is by definition true that applying a dimensionality-reduction algorithm to high-dimensional data will yield a lower dimensional representation. Are you claiming the positive (and sufficiently large absolute values) of the distance correlation metric imply something more significant about biosphere-atmosphere interactions and coupling? Isn't the t-SNE method designed to do something *close* to maximizing these distance correlations? And didn't you select your dimensionality reduction method to basically do that (maybe not with the explicit cost function of maximal distance correlations, but with local and global neighborhood coherence maximization)? I'm either 1) not seeing how this is a finding rather than the necessary outcome of your approach, or 2) not seeing how significant these specific metrics are relative to what I should be expecting (maybe the rho values would for some reason be expected to all be less than 0.1 for some reason?). I think it is well-known that there are continua of all of these variables (ranges of GPP, ranges of LE, etc.) and that stepping from one location to another nearby location (in space, time, or say VPD space) will lead to small changes in the biospheric and atmospheric states. This is not surprising. If you can quantify or qualify something ABOUT those gradients, it would be very interesting because that science is wide-open, but I don't see how Figure 2 is doing that. I see that you are suggesting that this is not obvious in the statement, "While gradients in MCI partial correlation strength are expected as they were used as features in the dimensionality reduction, gradients in climatic and biospheric conditions were not." But I don't see that as actually surprisingâĂŤthere are entire disciplines focused on the biogeographical structure of ecosystems and their gradients. How could we not expect a clinal change in LE and GPP to be related to a clinal change in LE-GPP coupling?

Line 183: "LE and NEE are weakly, not, or even negatively connected" This is interesting because it is commonly thought that arid/semi-arid locations have the highest coupling between LE (or Bowen ratio) and NEE because of omnipresent water limitation (e.g., in references below). Are these networks so arid as to not have vegetation?

Dirmeyer, P. A., F. J. Zeng, A. Ducharne, J. C. Morrill, and R. D. Koster, 2000: The Sensitivity of Surface Fluxes to Soil Water Content in Three Land Surface Schemes. J. Hydrometeor., 1, 121–134, https://doi.org/10.1175/1525-7541(2000)001<0121:TSOSFT>2.0.CO;2. Williams, I. N., Lu, Y., Kueppers, L. M., Riley, W. J., Biraud, S. C., Bagley, J. E., and Torn, M. S. (2016), LandâĂŘatmosphere coupling and climate prediction over the U.S. Southern Great Plains, J. Geophys. Res. Atmos., 121, 12,125– 12,144, doi:10.1002/2016JD025223. Short Gianotti, D. J., Rigden, A. J., Salvucci, G. D., & Entekhabi, D. (2019). Satellite and station observations demonstrate water availability's effect on continentalâĂŘscale evaporative and photosynthetic land surface dynamics. Water Resources Research, 55, 540– 554. https://doi.org/10.1029/2018WR023726. Crow, W. T., and Coauthors, 2020: Soil Moisture–Evapotranspiration Overcoupling and L-Band Brightness Temperature Assimilation: Sources and Forecast Implications. J. Hydrometeor., 21, 2359–2374, https://doi.org/10.1175/JHM-D-20-0088.1.

Figure 3: How were these archetypal clusters determined, by eye? It's a little weird to define 17 clusters algorithmically and then define 4 clusters of clusters by hand. Do the 4 types fall out on their own if restricted to 4 clusters? Type 2 looks like barren, arid landscapes. Type 3 is the mid-latitudes growing season. Type 1 is winter. Type 4 is interesting in that I wouldn't expect strong coupling in the tropics between T and anything or NEE and anything, since coupling (and even causality) is generally thought of as related to bottleneck variables. What leads to this full connectedness in physical terms do you think?

Line 149: "The monthly median network is the *average* of the networks. . ." The mixing of average and median here is complicating an already complex processing step. Are these averages being taken in the 2d reduced space axes?

Line 216: "for a given month" One month or three month window, shifting by one month at a time? This is confusing throughout. If using overlapping data (single-month network definition, but using sliding three-month windows), I think that will cause some



real problems in discussions of the inter-connectedness of your neighborhoods. Your rho values in Figs 1 and 2 will be artificially high by triply counting your data. I am sure you don't want this, but I think you need to either switch to 1-month windows or remove any networks with overlapping data windows (which will reduce your data points by a factor of 3 if I am understanding the method correctly). You can't talk about how nice and smooth the 2d space is when the analysis units are all 2/3 the same data as other neighboring units.

Figure 4: Why aren't Bowen ratios defined for the whole year for any of the sites except US-SRM? Aren't those variable observed? Don't you need them to map the seasonal trajectories for the plot on the left? Maybe not, and a lot of the network points are fit with partial data? Is that a problem in terms of robustness of the 2d space, the clusters, or the link strengths? You need to clarify how you deal with missing data throughout.

Lines 229-230: The fact that you are post-hoc trying to talk about this in terms of water/energy/temperature limitation on ecosystem productivity, but then calling out a separate, loosely connected concept of atmospheric interaction covariance highlights a general weakness in your storyline. There are physical concepts that are well-understood here: water, energy, and radiation (and temperature) can all act as limitation on photosynthetic activity in ways we largely understand (at the plot scale). At the same time, the land surface and atmosphere feed back on one another. I respect and am intrigued by the way in which you are attempting to link those two, but the question remains: what are we learning about how the land surface works by doing so? This is a major issue to be resolved for this manuscript.

Figure 5: I like the idea of this figure, and think it is a compelling way to look at extreme events. At the same time, it is worth asking whether this 2d space is good at representing extreme events. Does it make sense to think that a tropical rainforest undergoing an extreme drought is really just suddenly (and temporarily) turning into a system akin to a woody savanna, with all of the accompanying *causal* land-atmosphere feedbacks and carbon-energy-water coupling? I wouldn't assume so. That doesn't mean that this isn't

a fine first-order way to think about extreme events from a new analytical framework, but I would not a priori think that this is physically representative in any way beyond very coarse correlational descriptions. Presumably extreme events are another suite of dimensions that could be characterized, except for their lack of statistical representation in any data set (by definition). This warrants an explanation and some discussion of limitations.

Technical Corrections:

Line 29: "only consider two variables. . ." Granger Causality and Transfer Entropy at this point are only reasonably considered bivariate if you state "bivariate Granger Causality" as the authors do. This bivariance by necessity stance is a false position to take here. I don't know as much about CCM (although a quick search turns up a few recent multivariate extensions), but no econometricians think of VAR-based GC as bivariate, and there have been wikipedia articles about multivariate mutual information for more than a decade. I don't know that you need to have this discussion, depending on how you re-frame your research questions and motivation, but you can't publish this sort of claim.

Line 38-42: A nice synthesizing motivation. The motivation for the tie-in to *extremes* is not very clear though. Are you going to be looking at just biosphere-atmosphere interactions under meteorologically-extreme conditions? Or across the whole range of observed conditions?

Line 59: Strange citation format for Nelson.

Line 87-95: In the partial correlations, are the correlations controlling for (multiple) lags of the X and Y variables as well, or just other variables?

Line 116: And SWC?

Line 147: "non-intercepting convex hulls. . ." Even as a very methodological reader I am completely lost here. Is this a typo? What does non-intercepting mean? Intersecting



maybe?

Line 180: "Leave" -> leaf

Line 226: month-> months

Interactive comment on Biogeosciences Discuss., https://doi.org/10.5194/bg-2020-374, 2020.