Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-266-RC1, 2017

Geoscientific Model Development Discussions

Interactive comment on "A non-linear Granger causality framework to investigate climate–vegetation dynamics" *by* Christina Papagiannopoulou et al.

Anonymous Referee #1

Received and published: 30 January 2017

A non-linear Granger causality framework to investigate climate-vegetation dynamics

by

Christina Papagiannopoulou, Diego G. Miralles, Niko E. C. Verhoest, Wouter A. Dorigo, and Willem Waegeman

---

The manuscript introduces a Granger causal inference approach to investigate climate-vegetation dynamics. A great effort in collecting a representative enough dataset has

been pursued to study such dependencies. The authors put emphasis on the non-linearity of the approach since the VAR method typically used in the canonical Granger approach is here replaced by a nonlinear regression tool, the random forests method. Authors claim that the causal patterns are more clearly identifiable than with traditional linear models. Overall, I think this is a very nice piece of work that worths publishing after some clarifications and addressing some problems. Below authors will find a long list of minor and major comments that I hope they can address.

- abstract:

3: unravel the influence... : this looks like an ambitious goal that I'm not sure authors finally managed to address 4: existing statistical methods: do authors refer to linear ones only, right? 8: (also in the title) the word 'framework' looks too ambitious. In the end, authors only proposed to follow the Granger approach with a different feature selection and regression method. Does this qualify to call it framework?

p2.29: y alludes to the NDVI time series: shouldn't be the IAV of NDVI thereof?

p3.7: for me, describing the R^2 is too verbose and useless in a scientific journal nowadays

p3.eq2-3: the \approx symbol is meaningless here. I'd suggest to include the signal model here (y = \hat y + e), and describe the assumptions about the noise model (Gaussian, uncorrelated?). Also, I don't find natural that both eqs. have the same model coefficients \beta_11p.

p3.27: authors should clarify the sentence "neither variables nor observational ... and errors are ...". Independent of what? each other? independent noise? Please be explicit and consistent in the use of the terms 'error', 'noise', 'residuals'.

p4.10, eq4: describe the meaning of \beta_13 and all terms involved in the equation

p4.26: Maybe I'm missing something but if you split the data this way, aren't you discarding long-term correlations. Also, by simple xval, results depend to a large extent of

the selected data splits. To avoid this, why not LOO?

p5.10: the same comment about the \approx symbol before: please include the signal model equations here too.

p5.15: formally it is straightforward, but not computationally or for decision making which may be an infeasible problem.

p6.1-3: if you want to keep this statement, please discuss about the theoretical implications, and cite other nonlinear Granger causality methods (a simple search in Google will return you several dozens of works in machine learning, kernel methods, time series forecasting, econometrics and finance).

p6.1-14: verbose, remove or summarize a lot.

p9.eq: the upperscript T may confuse as in standard algebra that symbol stands for transpose.

p9.3: obvious non-stationary: sometimes it is not that obvious.

p12.6: a sentence does not conform a paragraph. And by the way... is 1° enough resolution to claim something about causation? do the expected relations occur at such broad scale?

p12.13: please avoid overoptimistic phrases like "our nonlinear random forestS".

p12.17: "simple correlations" should be "spurious correlations"? in any case this sentences deserves more clarification and be more explicit

Fig4: some discussions and words of caution should be given about deriving conclusions out of R2$\sim$0.4. By the way, why the maximum in the scale is not explicit for R2 and you select that threshold in 0.4? Why not using the statistical significance of the correlation rather than the R2 score? Can authors include and discuss the maps of R p-values?

Fig4 caption: 'with respect to a the' to be corrected

p13.3: 'our'?

p14.3: what are these patterns of the explained variance? some clarification is needed here? I guess authors refer to spatial patterns of variation? If that is the case, it looks not really obvious to talk about spatial relations when no such relations are considered to build up the regression models.

p14.7: unambiguous? some more comments are needed, and if possible supported by numerical scores.

p13-14: as a reader I'd prefer to have in the same figure panel the current figures 4 and 5 so I could directly compare results in one shot.

p15.3: what do authors mean by 'higher-lever variables'? are you thinking of higher-order statistical relations between variables? this is absolutely confusing.

p15.5: please provide a copy of the (Papagiannopolou et al, in review) so reviewers can appreciate differences in approaches and results. Alternative, cite an accessible work to support the claims in this paper.

p16.3-18: please clarify these paragraphs in several ways: 1) the spatial encoding is not at all clear since typically the input (feature) space is augmented with the neighbors which are then used to predict on the central pixel (the length of the observation variable does not change), which seems not to be the case here. 2) it is weird that the spatial info didn't improve the results: I'd thank the authors to include such 'negative results' but then some comments and clarifications are needed (e.g. 1° is already integrating too much info, or spatial encoding was not taking into account pixel spatio-temporal variances?)

p17.9: as said before I feel claiming a 'novel framework' is far too much for this contribution.

p17.15-20: some claims are contained here without empirical justification. I think that authors lost a nice opportunity here to explain the causal relations. For example, to me it seems ad hoc to justify results with a simple 'the predictive power of the model is especially high in water-limited regions'. Probably this is true but some numbers are needed to support it. I suggest to include a summarizing feature ranking of the LR vs RFs (e.g. permutation analysis, and surrogate analysis). Also, summarize results per regions and biomes would help discussing the results more profoundly, elevating the debate. Of course, these two issues may require some more work, but I sincerely think they are mandatory to make a sound publication.

p18.8: reproducibility is not possible as data is not available yet. do authors plan to make these data available to the community?

Interactive comment on Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-266, 2016.