Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-266-AC1, 2017


Geoscientific Model Development Discussions

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

**Christina Papagiannopoulou et al.**

christina.papagiannopoulou@ugent.be

Received and published: 3 February 2017

**Response to Anonymous Referee #1**

**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 is worth publishing after some clarifications and addressing some problems.**

We would like to thank the reviewer for his/her appreciation of the manuscript, for the constructive feedback, and the thorough assessment of the manuscript. Below we provide a point-to-point response to each comment.

**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**

The sentence reads 'Data of this kind [Earth observations] provide new means to unravel the influence of climate on vegetation dynamics'. We are sure the reviewer would agree that they do, so we do not really see an issue regarding this sentence.

**- 4: existing statistical methods: do authors refer to linear ones only, right?**

'Existing' will be replaced by 'commonly used'. We believe this addresses the issue the reviewer refers to.

**- 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?**

We understand the concern, yet 'framework' can be defined as 'a basic conceptional structure (as of ideas)' (Merrian-Webster dictionary). We chose to use the word 'framework', because we believe it reflects well the conceptional structure followed here, which consists of several sequential steps. First, multiple datasets of the most important climatic variables have been collected and converted into a common temporal and spatial resolution (a multidimensional data-cube). Second, by applying feature extraction techniques and domain knowledge, predictor variables have been constructed. Third, a nonlinear machine learning algorithm has been designed and applied. Forth, causality has been assessed based on Granger causality tests. As such, we think that we propose a complete framework that can be used for knowledge discovery in climate sciences, and which in this paper is applied to unveil climate-vegetation dynamics, but that could be used to detect other causal patterns in the climate system. Moreover, by using nonlinear models and feature extraction techniques, we argue that this framework substantially differs from methodologies that are common practice in the field.

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

Yes, it is true that we finally model the IAV of NDVI, this is just a starting point of explaining the basic model. We will clarify it in the revised manuscript.

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

The $R^2$ is indeed a well-known performance measure, we wrote down the formula just to avoid confusion: for linear models one often computes the $R^2$ as a correlation coefficient, while for nonlinear models one has to compute the $R^2$ in a different way, using the formula in the manuscript. This is perhaps obvious, but it might be useful to have the formula in the paper for readers that are less familiar with the different definitions of $R^2$.

**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 $\beta_{11p}$.**

We agree. We will incorporate these changes.

**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'.**

We meant independent from each other. We will clarify this in the revised version.

**p4.10, eq4: describe the meaning of $\beta_{13}$ and all terms involved in the equation**

We believed this was already clear, but perhaps it is not the case; we will include a more extensive explanation of the tri-variate extension of Granger causality in the revised manuscript. In the tri-variate case (as in the case of the bivariate Granger causality) we examine if the time series x Granger-causes the time series y. The time series w might act as a confounder in deciding whether x Granger-causes y; for this reason it is included in both models (baseline and full). That way the method can cope with cross-correlations between climatic drivers of vegetation. The $\beta_{13}$ are the coefficients of the time series w while the rest of the coefficients have the same meaning as in the bivariate case.

**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?**

We are not sure whether we understand this comment. Our motivation for doing 5-fold cross-validation instead of leave-one-out was mainly motivated by computational reasons. LOO takes a long time to compute and is generally not the recommended method when analyzing large datasets. As we are working with an extremely large dataset here, computational efficiency is always the first criterion to look for.

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

Thanks. We will revise this as well.

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

Of course. We only want to convey here that the formal definition of Granger causality does not change in the case of more than three time series.

**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).**

We agree that there are previous works on nonlinear Granger causality. Those methods typically assume stationarity of the time series, and they are hence not immediately applicable for climatic time series. We will extend the paragraph with a more thorough discussion of related work and new references to these articles. We should also clarify that we have not introduced nonlinear Granger causality. But these



more complex methods which use Granger causality have not been widely applied in the field of climate sciences.

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

This might be obvious for a well-informed reader, but we believe that an explanation of that kind is needed for readers that are less familiar with Granger causality and time series forecasting. We will nonetheless try to condense these sentences without losing information.

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

We will replace the symbol T with Tr in order to make it clearer in the revised manuscript.

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

We propose to delete the word ‘obvious’, indeed.

**p12.6: a sentence does not conform a paragraph.**

We disagree, a paragraph is ‘a subdivision of a written composition that consists of one or more sentences’ (Merrian-Webster dictionary, but also any other...). Yet, we will consider merging it with the previous paragraph, although the decision to make it a stand-alone sentence was to highlight it.

**And by the way... is 1 degree enough resolution to claim something about causation? do the expected relations occur at such broad scale?**

Most atmospheric variables change consistently at spatial resolutions that are even coarser than 1 degree; in fact current climate models resolve the land–atmospheric interactions while working at coarser resolutions. Nonetheless, in heterogeneous environments this would be a limitation, and this issue should be acknowledged in the revised version of the manuscript. We also note that there is a trade-off between spatial resolution and time span. The 1-degree resolution is a characteristic of the datasets we are working with, and if we wanted to focus on finer resolutions, we would need to incorporate datasets from sensors covering more recent years only. The 1-degree resolution, in addition, still also allows us to perform our calculations in a reasonable amount of time.

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

Thanks, this will be rephrased to "the nonlinear random forests models".

**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 $R^2 \sim 0.4$. By the way, why the maximum in the scale is not explicit for $R^2$ and you select that threshold in 0.4? Why not using the statistical significance of the correlation rather than the $R^2$ score? Can authors include and discuss the maps of R p-values?**

This is a remark that we expected. As mentioned in the manuscript, the assumptions of common statistical tests are violated due to the non-stationarity of the data and the nonlinearity of the proposed model. Developing a statistical test that is able to handle non-stationary time series and nonlinear models is not a trivial task.



As far as we know, no such test exists. Therefore, we decided to focus on expressing Granger causality in a quantitative way instead of a qualitative way, and stress the gained improvement with the use of a nonlinear model. We are currently developing a statistical testing procedure, but this is work in progress and will be the subject of future contributions. We will also include the relevant references and a more thorough discussion about existing statistical tests in the revised manuscript.

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

Thanks.

**p13.3: 'our'?**

Yes, thanks.

**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.**

In fact this section does not refer to spatial patterns, but to a general improvement of the full model versus the restricted model. We will update the text accordingly to clarify this.

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

With unambiguous we just mean that the improvement is clearly visible here (in the order of 20 to 60%). This claim is supported by Figure 5b.

**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.**

We understand this concern and understand that a 3 $\times$ 2 figure would be more convenient for the reader. However, we chose to have Figure 5 in a separate panel because this is the main figure of the paper, and unless the reviewer is insisting on this issue we would rather keep it this way.

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

With the term 'higher-level variables' we refer to past cumulative climate and climate extreme indices which are in the dataset as predictor variables. We will clarify this in the revised manuscript.

**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.**

The referred article will be enclosed to the resubmission of the revised paper, as long as it can be made available to the editorial and reviewers only.

**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 degree is already integrating too much info, or spatial encoding was not taking into account pixel spatiotemporal variances?)**

Yes, the approach we followed is as described by the reviewer. The feature space of one pixel is augmented with the features of the 8 neighboring pixels. We also expected to see a substantial improvement using spatial information but this was not the case. We think that this is due to the huge dimension of the feature space which may include redundant information, in combination with the low number of observations per pixel. We will extend the discussion in that direction. Let us also stress that we are currently working on (more complex) spatial models that could overcome this kind of issues.

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

See above response.

**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.**

Actually, we have performed this kind of analyses, taking feature rankings using RFs. However, it has been shown that those rankings become unstable due to highly-correlated predictors. A specialized approach would be needed here, in which groups of features are ranked instead of individual features. This makes the rankings more stable and improves the interpretability. It is exactly what we do in the complementary paper (Papagiannopoulou et al., under review).
We agree with the reviewer that a stratification of the results according to regions/biomes is a relevant addition to the paper. The revised version will provide the results stratified according to IGBP vegetation classes for both the baseline and full random forest model. These new results will be discussed in the relevant sections.

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

All codes will be freely available and documented on GitHub and will comply with the Copernicus data policy. On the other hand, the database is formed by a collection of datasets that are all publicly available and that, due to copyright conflicts, we cannot openly distribute. Nonetheless, we have decided to add the link to these datasets below Table 1 in the revised version.