General comments:

The manuscript explores the biosphere-climate interactions at global scale. The method, based on a Granger Causality framework, quantifies the climate impact on vegetation and the vegetation feedback on climate using satellite observations. The same approach is then applied to four ESMs and differences between data and model results are discussed. The study is well written and potentially interesting as – to my knowledge – is the first work aimed to isolate the climate-vegetation interactions analytically using observations and can help the modelling community to improve ESMs. However, I have some major concerns that need to be carefully addressed before publication.

We thank the reviewer for the insightful comments and we hope that we have addressed them all adequately.

Specific comments:

1. The study is based on a limited set of observational datasets: only one product per variable. In particular, LAI and precipitation data show large discrepancies and inconsistencies across products (Jiang et al., 2017). Results, based on a so limited set of products, may be largely affected by specific product uncertainties. The analysis should be replicated by using an ensemble of different products for LAI, P and possibly T and RN. Results based on an ensemble of combinations would be much more robust. Comparison of results obtained from different combinations of products would also enable you to assess the validity of your approach and the consistency of your results. Jiang, C. et al. Inconsistencies of interannual variability and trends in longterm satellite leaf area index products. Glob. Change Biol. 23, 4133–4146 (2017).

We thank the reviewer for this comment. We agree that by creating an ensemble of LAI, P, T and Rn, we could significantly improve the robustness of the results. We are currently searching and processing additional datasets. We will update all figures accordingly.

2. Spatial patterns shown in figures (e.g., figs. 2, 3 and appendices) are very jeopardized and – a part of the radiation control patterns – are not very credible. There is a huge spatial heterogeneity even in regions characterized by the same environmental conditions. I'm wondering, if such spatial variability reflects some problems of stability in the algorithm or noise in the modelled signal. These strange patterns emerge particularly at longer time scales (seasonal, interannual) maybe because the sample size is more limited (?). I really find difficult to believe in such patterns and authors should make an extra-effort to improve or at least understand such spatial variability. In my opinion, such spatial variability could originate from the native time series (possible uncertainties in the signal) and the processing of the signal, as I do not see any patterns that can be easily related to physical conditions. Maybe, the use of ensemble of different observational products (see comment 1) may help to retrieve a more robust signal.

We are aware of the heterogeneity at longer timescales, and also reviewer #1 pointed to this issue. The problem is partly due to the parameterisation of the frequency parameter of the wavelet, which provides a trade-off between temporal and spectral resolution. As mentioned in the response to referee #1, fine tuning the parameter to provide a better temporal resolution can improve the inter-annual patterns. However, the noisy patterns can also occur due to noise in the input data, as mentioned in the first comment. By creating an ensemble of datasets, we will reduce the impact of errors in the forcing.

3. The benchmark of ESMs is very useful and interesting. However, the authors should try to identify potential areas of model improvements. This exercise should be aimed to clearly understand what are the strengths and deficiencies of each single model with respect to the data-model comparison performed. A table to synthesize areas of improvements could help to convey the key information to modelers.

Although we agree this might be of interest to the modelling community, we aimed not to single out individual models in this manuscript due to length restrictions. Moreover, we believe that by focussing on model differences, or even specific model parameterisations, we might dilute the main findings that relate to the

whole range of models. Therefore, we believe this feels outside the scope for this study, and hope the reviewer may agree with this rationale.

4. Remote sensing LAI data in winter season are affected by snow cover conditions. I'm wondering how you have addressed this issue. If you did not account for this, I think your results may be strongly affected by this bias.

We acknowledge that snow cover might affect the LAI in the high northern latitudes, especially at the seasonal and monthly scales. At the moment, we do not address this issue. As pointed in the response document to referee #1, preliminary explorations indicate that the strength of the signal during the growing season strongly dominates average temporal patterns. We are also confident that the adoption of an ensemble approach will dampen the sensitivity to errors in the individual LAI data products during winter time, being however aware of the fact that these errors are likely systematic and shared by all data products. Moreover, as pointed in the response to referee #1 as well, in the present working on adapting the SCGC algorithm to explicitly resolve different time scales, which would in the future allow to resolve the causal relationships in time and mask out periods of poor data quality. This however requires an in-depth adaptation of the method. In the revised version, this issue will be explored, palliated by the use of ensembles, and discussed more explicitely.

5. The relevance of the multi-temporal scale needs to be clarified, what is the added value of a such analysis compared to previous studies focusing only on monthly scale?

We feel that the fact that our results differ for different frequencies (time-scales) highlights by itself the need to consider these frequencies separately to better understand the driving role of climate in ecosystem dynamics. However, we will add a clarification on why, conceptually, phenology scales and inter-annual variability also need to be considered separately when models are evaluated.

Minor comments

We thank the author for the thorough list of minor comments and will try to address all of them. Underneath follows a selection of minor comments that deserve a short reply, which aren't addressed above:

- Page 1, Line 15: It is not clear to what phenology refers to.

  We will define what we mean by "phenology" briefly. Here we have adopted it as synonym of seasonal-scale vegetation (LAI) cycle.

- Page 4, Line 17: Not sure this is correct. You basically used two different periods of analysis for observations and models: 1981-2015 (ca 35 years) for observations; 1956-2005 (50 years) for models. A part of the temporal shift between the two experiments, I would suggest to verify that the different length in the time series do not introduce a systematic bias between observational- and model-based results. Why you decided to start from 1956 for models? To me it would be more logic at least th preserve the same length of observations (35 years). Please, check this and clarify your choices.

  In the updated manuscript we will add a supplementary figure comparing the runs over the overlap analysis period, i.e. 1982–2005, and discuss the results in the main text.

- Page 5, Line 24: In the presented formulation of GC, the temporal lag m is implicitly assumed the same for all predictors. In practice, I expected that the legacy effects may differ depending on the predictor. Can this be included in the formulation? Please, discuss the implications.

  To clarify, the formula on page 5 refers to traditional Granger causality. (Conditional) Spectral Granger causality as calculated by Dhamala et al. (2008) is non-parametric and consequently no longer requires prescribing a specific lag; the dominant lag is in fact resolved by the formulation. This will be clarified explicitely.

- Page 11, Line 28: To me the comparison performed only on these numbers is misleading because they refer to the relative contribution to the total explained variance. Therefore, ESMs could be in principle represent well the variability of the T control on vegetation in absolute terms, but could overestimate the P control on vegetation in absolute terms. This would lead to an underestimation of the T control in relative terms over the globe ... again not because they fail to represent the T control but because they fail the P or RN controls. The analyses should be complemented with the comparison in absolute terms.

  We acknowledge that well-modelled interactions can be masked by an overestimation of the importance of another variable. However, within a pixel, the 3 drivers are rescaled using an identical factor. This results in 'brighter' figures, but the fractions of each color remain the same as for the absolute values. Showing the maps in absolute terms does not resolve the problem in case of overestimation of a variable. For clarification, the latitudinal plots are shown in absolute values.

  _____

  We also thank the reviewer for the multiple editorial suggestions and minor comments, which will all be integrated in the revised version of the manuscript.