Dear Editor,
we thank both reviewers for their considerations on our manuscript. We provide below our answers to all comments of the referees. Please contact me if you need any further information.

The referee comments are given in normal fonts.
Our point-to-point revisions are given in bold letters.

Anonymous Referee #1

This study links TRW and LPJ NPP in Europe using coincidences as an analytical framework. I must say I am confused on the prominence of the "benchmarks" idea. I don't really see any benchmarking here (sensu Luo et al., 2012, which you cite). It feels oversold, especially in the title. I would think it might be a useful add to the conclusion (that TRW can serve as a reference, with the caveats you cite).
We thank the reviewer for this cautionary remark. We had some discussions amongst coauthors and we agree that we should frame the paper more carefully. As a consequence, we will change the title to "Coincidences of climate extremes and anomalous vegetation responses: Comparing tree-ring patterns to simulated productivity". In addition we will also adapt the introduction and conclusion accordingly to clarify our idea. However, we also would like to emphasize that we developed the manuscript inspired by Luo et al. (2012). Hence, we are indeed interested in exploring the value of the presented coincidence analysis as an additional benchmark metric in a generic system as proposed by Luo et al. Given that our coincidence analysis meets the criteria they suggest "… objectivity, effectiveness, and reliability for evaluating model performance", we think that this is a valid idea. In their table 1 Luo et al. (2012) state that "functional patterns" emerging from data such as responses to precipitation could play an important role in the "evaluation of environmental scalars and response functions". To our mind, the proposed coincidence metric based on long-term tree ring values offers a perspective of this kind. Please note that the paper by Luo et al. (2012) emerged from an iLAMB project meeting where it was discussed that it is time to systematically evaluate land model benchmarks of different kind to allow the community to converge to a consensus of modeling skill. However, given that this was apparently not very clear we will revise the manuscript accordingly and we will explain this background based on a more profound reflection of the latest literature.

Also, you state (2548) "Hence, an in-depth investigation as to how these differences can be attributed and to what extent TRW can be used to benchmark dynamic vegetation models for responses to extreme events is necessary". I thought that was a goal of the study given the two questions you pose (2542) and the use of "benchmark" in framing the Introduction?
This formulation was misleading. We will change the text accordingly: "Hence, in the following, we carry out an in-depth investigation as to how these differences can be attributed."

I am also quite curious about coincidences between TRW, P and T (all three at once). Did you investigate that at all, apart from Fig. 3?
Thank you for this suggestion. We now performed additional analyses on coincidences between NPP and simultaneous P and T extremes and we will include the new results in the revised manuscript. A characteristic feature is that these triple coincidence rates are generally lower because of the rareness of these compound events. The impacts however are slightly larger during these simultaneous events (if requested, the new figures can be provided to the reviewer).

Also, and acknowledging the effort to run an LSM, did you ever think to corroborate LPJ with another LSM. LSMs have, generally speaking, low skill, but also exhibit large gradients in skill. It would be enlightening to know whether the coincidences you've found here are model-specific.
We agree with the reviewer that a model intercomparison using the coincidence analysis would be very interesting but this is clearly out of the scope of the present paper. Our study is an early attempt to develop a method where tree-ring data can be used to evaluate the simulated productivity using non-scale free metrics. Given both the success and caveats (e.g., the desirability to test non-scale free metrics as indicated by reviewer 2) we hope that this work will i) catalyze multi-model comparisons and ii) advances in the types of metrics (both dimensional and dimensionless) where tree-rings can be used to benchmark LSMs.

Apart from this I have a few minor/specific comments below. Many relate to language. I would strongly encourage an overall tightening of the language.
We will revise the manuscript accordingly and improve the language.

2539/4: Try "reductions of net primary productivity"
We will change the text accordingly.

2540/4: No comma before which
We will change the text accordingly.

2540/7: Try "response, e.g., to drought events (Schwalm et al., 2012), may" Comma issues throughout (I've highlighted two here). Minor, yes, but I would encourage a native or native-like speaker to proofread the text.
We will change the text accordingly.

2541/9: Try "alone as a proxy"
We will change the text accordingly.

2544/2: Are the TRW available online? I would encourage the authors to add this as a supplement to the study.
Yes, the tree-ring data are available online in the supplementary material of Babst et al. 2013, GEB. See http://onlinelibrary.wiley.com/doi/10.1111/geb.12023/suppinfo
We will clarify this in the revised manuscript.

Section 2.1.4: This reads very ad hoc. Why/how were the thresholds chosen? MODIS goes back to 2000 but you speak of multi-decadal time scales? The 1901-2001 reference for example. What are the "the connected phases of GSob"? I feel this section needs to be better substantiated.

We agree that the current wording was not fully satisfactory. These definitions are based on our inspection of the observations, and represent a compromise between site-specific definitions of growing season length and comparability across the network. Please note that the underlying FAPAR data have been derived based on albedo and are very generic, i.e. they are not plagued by vegetation-type specific parameterizations. We will clarify these issues in the revised manuscript.

2549/8: What are "climatically attributable extreme years"?

We agree with the reviewer that the formulation "climatically attributable extreme years" was misleading. We will change the text to "…we quantified TRW and NPP anomalies in years with extreme climate conditions."

Also I am unclear what was done here? Did you take extreme years, z-score them, and then do a histogram? So a histogram of values in the most extreme deciles that were then z-scored? I think so but I am unclear why you did this? This might be semantics on my part but I think I'm struggling with z-scoring extremes and calling these values extremes? Some values will always be below your sigma thresholds when you z-score?

We agree with the reviewer that a more detailed explanation of what was done here is necessary. For the analysis of down-regulation, we scaled all TRW and NPP values (zero mean and unit variance, i.e. computing z-scores) and then selected the values for which we found coincidences between T, P and TRW and NPP extremes. The z-scoring was done in order to have a comparable scale that allows us to compare the NPP responses to the tree-ring response. We will add a paragraph in the methods section to explain this better in the revised manuscript. Please note that figure 2 will change in the revised manuscript as we found a time-shift error in this very plotting function and because we will include the triple-coincidences. Additionally, we performed a more systematic assessment of the impacts of extremes on TRW and simulated NPP, and we will report the results in the revised manuscript.

2549/20: What is a carbon storage product NPP? I think you mean that modeled NPP behaves differently than TRW based on some aspect of how LPJ is put together, how LPJ simulates NPP?

This formulation was indeed misleading. We will change this in the manuscript: "In contrast to that, simulated NPP responds rather instantly to extreme events."

2549/24: Try "reducing" That or comma placement?

We will change the text accordingly.

2552/1: Does it not also highlight the difficulties in matching a point to a pixel? I think that's what you are saying but my concern here is from the "other " side. I don't want downscaled

data per se. Rather, I want to understand how scale-mismatch degrades coincidence (or any validation study where a point is matched to a modeled pixel).
The problem of representativeness of coarse grids for local conditions has been a matter of debate for a very long time, and the coincidence analysis will reflect this scale-mismatch accordingly. To investigate how scale-mismatch degrades coincidences would be interesting but is impossible due to the lack of station data (i.e. point data), as we point out in 2551/25.

Figure 6: I find this hard to decipher. The transparency is quite subtle, and the double legend seems off. Also, you mention zonal patterns in the text. But that is more wrt latitude as opposed to climatic space. Also, what is n in each bin?
We will improve Figure 6. The legend will be changed to "TRW" and "simulated NPP at TRW sites". The x-axis label will be changed to "Temperature in 2.5°C temperature bins". We will indicate the n for each bin in the figure and also include the triple coincidences as outlined above.

Figure 7: I find the discussion of this in the text very hard to follow. For example, 2553/1 "Hence,..." implies some 3D surface that is never displayed. The last sentence in the caption is also confusing. I had thought that you had the same time scale for both TRW and NPP? That you pulled the matching pixel and time span from your transient LPJ run from 1901 to present?
Our text was apparently misleading. We will change 2553/1 to "Hence, we can confirm that negative precipitation anomalies and positive temperature extremes lead to reduced growth not only in the current but also in the subsequent year (Figure 7).
The time scale for TRW and NPP was not the same. We selected pixels at tree ring sites but did not adjust the time series length. We will add this explanation in the methods section to make this clearer in the revised manuscript. For more clarity, we will also improve the caption of Figure 7.

2553/11: I think you need to be careful here. TRW is, trivially, based on NPP. I think you mean that the way NPP is modeled is less cumulative than TRW. That is, the model is instantaneous and summed to get a GS value whereas TRW has lags naturally embedded?
We agree with the reviewer - indeed we have to be more careful here, especially when explaining what happens in the model. We will carefully revise the argumentation in the paper by changing the text to: "Constant or less accumulated biomass then leads to reduced simulated NPP during the following year. Because simulated NPP represents a rather short-term measure of carbon use compared to observed TRW, it thus responds more instantaneously to changes in photosynthesis and respiration during extreme events. In contrast, observed TRW integrates carbon accumulation and growth over a whole growing season, relies in part on stored carbohydrates, and may even be influence by longer-term response to canopy and root architecture. These considerations may explain why observed TRW may therefore not react in a similar way as simulated NPP to extreme events."

2553/25: The final sentence of this paragraph really does not make sense to me.

We will remove the final sentence from the paragraph.

2554/7: "triggers" Not sure I buy this. Triggers implies a certain degree of causality that I do not find here.
Correlative findings never allow us to say something about causation, but still allow us to formulate hypothesis on how a system works. Here we are discussing the hypothesis that we can formulate based on our coincidence analysis i.e. "the probability that a climate extreme triggers an extreme reaction". We will reword the sentence in the revised manuscript.

Last para in Conclusions: Your a, b and c. Did we not already know that? I don't think this study is the first to show this. That's not bad, I'm all for contextualization and linking back to established findings. But you might add what specifically your study tells us that we did not know beforehand.
We agree that the presentation of these three aspects is not very good here, as it indeed is repetitive to previous studies. Hence, we will improve the conclusion emphasizing the methodological perspectives offered by our study: "Our study has shown the potential to use tree-ring data in a scale-free metric that should be used to evaluate the abilities of DGVMs to simulate growth responses to climate extremes. Current model evaluation studies are lacking this type of analysis. As climate extremes can have long-lasting impacts, DGVMs need to be able to simulate such effects as well as to capture processes being responsible for multi-year lagged effects. The combination of improved DGVMs and the method of coincidence analysis can then be applied to quantify the impacts of extreme events, e.g. on the long-term fate of the European carbon balance."

Anonymous Referee #2

The authors conclude that "using radial tree growth is a good basis for generic model-benchmarks if the data are analyzed by scale-free measures such as coincidence analysis." I fundamentally disagree with the statement.
We will revise the paper in order to clarify our point of view, according to which scale-free measures are very helpful for model-evaluation purposes. Our rationale is that scale-free measures are opening a path to capitalize on long-term monitoring observations that capture processes that are not explicitly represented by terrestrial biosphere models. In cases of this kind, we have to abstract from the measurement to a "pattern" that we can then compare against an analogous pattern in the model output. Pattern oriented model evaluation is anyway a very common approach, for instance when remote sensing observations are compared to modeled vegetation phenology. We will elaborate the text further in order to clarify our perspective more strongly.

Studies focusing on the environmental sensitivity of tree productivity should work with the absolute growth rate of biomass. Working with standardized radial increment continues a bad precedent because (1) the age-based standardization does not adequately distinguish geometric, ontogenetic, and extrinsic effects on tree growth and (2) sensitivity of radial increment to extrinsic factors is not comparable with productivity, as universally defined by a flux density of mass or energy. The most likely explanation for the reported discrepancy between sensitivities of model net primary production and standardized ring width index to extreme climatic events is that the latter variable is only a subdimension of the system of interest.

Conceptually we agree that tree ring patterns are only representing one part of the total productivity and we still are in the earlier stages of knowing how changes in radial growth are either amplified or buffered by changes in e.g., leaf production, root growth, changes in wood density etc. However, the different endogenous and extrinsic influences are intrinsically interlinked and constitute the systems behavior that should be matched by the model. Hence, we are less pessimistic and would not agree that we cannot use TRW at all in the context of models that do not represent plants at the individual level (which would be the logical consequence of the reviewers view). But please note that also the model contains some of the features, especially if we consider NPP as reference where also e.g. mortality effects and carbon loss from autotrophic respiration are implicitly considered.

Expressing the results in relative terms (or scale-free terms) does not circumvent this problem and the development of new statistical methods as a means to deal with inadequacies in the type of observation being collected is not the right direction.

If qualities of the cores used in this study disallow the authors from transforming ring width increment into absolute growth rate of biomass, then they will be of little value in advancing our understanding of the environmental sensitivity of tree growth.

We appreciate the reviewers' desire for absolute biomass comparisons of model results and data. Yet, we disagree with a possible implied message from the reviewers' criticism: namely, it may be better to do nothing than compare in scale free terms. We assume that both models and observations are incomplete in many ways. Hence, our objective is to find a level of abstraction where we can compare patterns that carry comparable signs of response and then evaluate "functional patterns" (cf. our first response to reviewer 1).

At the time of the study, it may have been possible to compare absolute woody biomass increment for a few sites across Europe (see Babst et al. 2014 New Phytologist) with model results. We are convinced that our benchmarking of the coincidence of extreme growth and climate events for several hundred sites across Europe is a notable accomplishment, and does represent a new and positive precedent. As mentioned above and in the revised text, we hope that our work will contribute to the advancement of tree-ring networks that can be used for model benchmarking of absolute biomass and growth rate.