Contrasting drought legacy effects on gross primary productivity in a mixed versus pure beech forest

Xin Yu, René Orth, Markus Reichstein, Michael Bahn, Anne Klosterhalfen, Alexander Knohl, Franziska Koebsch, Mirco Migliavacca, Martina Mund, Jacob A. Nelson, Benjamin D. Stocker, Sophia Walther, and Ana Bastos. *Biogeosciences Discussion*

Response to Reviewer #1

Major comments:

R1C1: In this manuscript, Yu et al. investigate drought legacy effects in GPP at two contrasting forest types in Germany. This manuscript represents several notable advances, including: 1) direct observation of GPP legacy effects, 2) a method to quantify sub-annual legacies, 3) incorporating uncertainty in legacy effect calculations, and 4) a neat idea to get at the mechanism behind GPP legacies. In addition, the manuscript is written very clearly and is quite compelling to read. This is a great contribution to the literature and only have a few suggestions.

We thank the reviewer for the overall positive evaluation of our study and for the constructive comments. Below, we provide a point-by-point reply to the reviewer comments.

R1C2: The approach to calculate a "tree ring width" based on dendrometer bands is interesting. However, due to bark shrinkage and expansion, these processes aren't exactly analogous. I think there needs to be an acknowledgement of this and a discussion of how these biases might play out.

Thanks for pointing this out. The method we used has - compared to tree cores where TRW is directly measured - three central advantages: permanent, long-term and non-invasive/non-destructive measurement of the same trees, measurement of fresh wood increment, and measurement of the entire stem at breast height (not only of a very small part). The latter two advantages are important for an upscaling to total wood growth at stand level. The main disadvantage of our method - the inclusion of shrinkage and swelling of the bark - is an "accepted" uncertainty. This uncertainty can be accepted (1) because we used only the annual increment (not as in Mund et al 2010 where monthly growth rates were investigated), (2) because the dominant species is beech that has only a thin bark, (3) because we recorded the final stem diameter of each year in winter, when the water status of the xylem and the bark is relatively constant, and when stem wood or the bark are not affected by frost or late/early growth or water uptake, (4) because in this study we were interested only in the interannual variability of stem growth, which is less affected by shrinkage and swelling at the described temporal scale than absolute growth rates. Thus, in our case the bark-associated uncertainty results mainly from annual bark increment that is small compared to stem growth in the studied species (no pine or oak). Nevertheless, to be clear and to avoid any misunderstanding, we have changed the term 'tree ring width' to 'radial increment', and have added the description in line 106 of section 2.3:

'Annual radial increment (RI) was calculated from permanent band dendrometers which measures change in stem girth (or circumference) over bark. The effect due to the inclusion of shrinkage and swelling of the bark is a negligible uncertainty for four reasons: 1) we used only the annual increment, 2) the dominant species is beech that has only a thin bark, 3) we recorded the final stem diameter of each year in winter, when the water status of the xylem and the bark is relatively constant, and when stem wood or the bark are not affected by frost or late/early growth or water uptake, and 4) in this study we were interested only in the interannual variability of stem growth,

<u>which is less affected by shrinkage and swelling at the described temporal scale than absolute growth rates.</u>'

R1C3: How well does the RF model predict GPP during drought years, if trained on non-drought data? Or, just trained on a subset of droughts and used to predict other droughts? The answer to this question has implications for the interpretation of the legacy effect calculation.

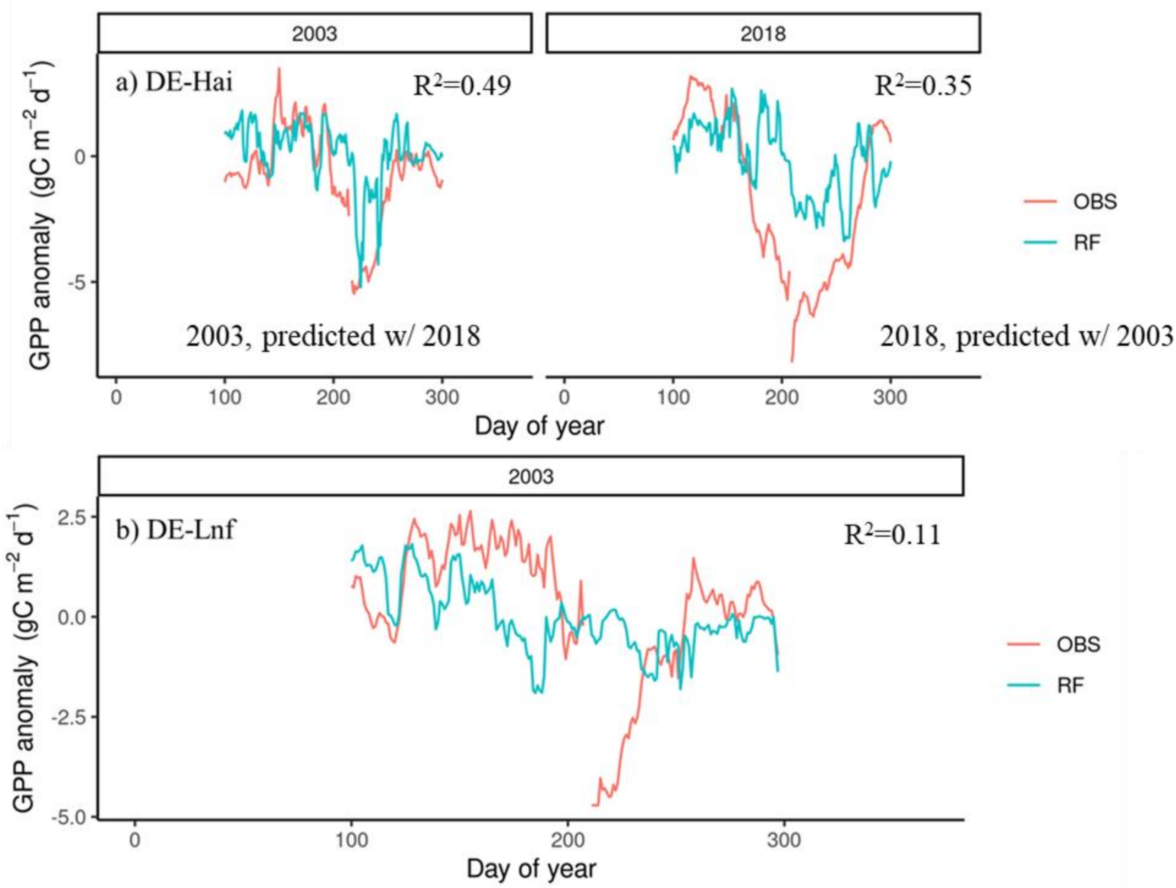

Figure R1.1. Observed (OBS) and predicted (RF) GPP anomalies from trained models using subsets of the training dataset a) at DE-Hai in 2003 and 2018, b) at DE-Lnf in 2003.

We trained RF models using a subset of training dataset (sub) to test the model performance during drought years. For example, in order to test the model performance in 2003, the data in 2003 was excluded from the training dataset and the trained model was used to predict GPP anomalies in 2003 given meteorological variables. The same strategy was applied to 2018 at DE-Hai and 2003 at DE-Lnf. We compared observed (OBS) and predicted (RF) GPP anomalies from subset-trained RF models a) in 2003 and 2018 at DE-Hai and b) in 2003 at DE-Lnf. We found that RF matches OBS GPP relatively well at DE-Hai during the growing seasons of 2003 and 2018 (R2 of 0.49 and 0.35, respectively) and captures the GPP decrease during the drought period of 2003, but estimates a weaker decrease in the 2018 drought. This is thanks to the fact that the 2018 drought is more extreme than the 2003 drought, and the RF model of which training dataset including the 2018 drought is expected to capture the 2003 drought which is within the range of training data. However, the prediction skills of the RF model in 2018 at DE-Hai and 2003 at DE-Lnf are limited, because of the limited range of training data. Therefore, in the manuscript, to avoid extrapolation beyond the range of training data, we included drought and non-drought years in our training dataset, which maximizes the range of training data. The meteorological conditions in legacy years are within the range of training data, and the trained model is expected to be able to predict reasonable potential GPP anomalies in legacy years. Furthermore, such RF prediction inaccuracies are included in our uncertainty calculation, and the respective uncertainty ranges have been shown in the main figures.

R1C4: Along those lines, there also needs to be some information regarding model fit, predictive ability, variable importance, etc. in the methods or results. Does model fit differ across sites, years, etc? It seems like a model with a lot of uncertainty at one site or one year may drastically alter the legacy effect calculation. What variables are most important for predicting fluxes at these sites?

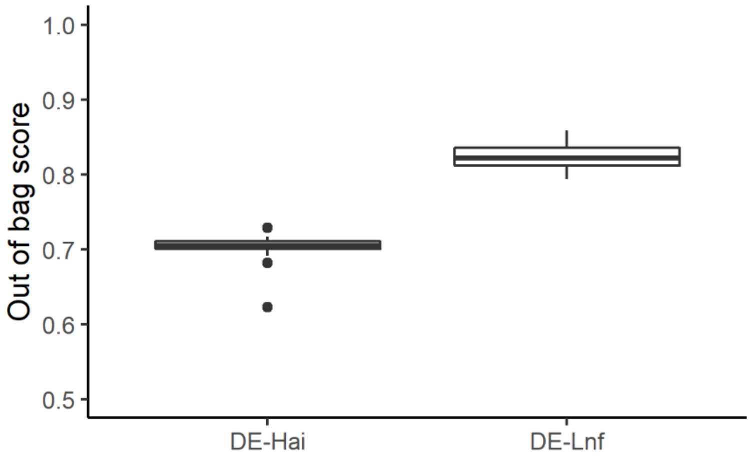

Figure R1.2. Out of bag (OOB) scores of RF models at DE-Hai and DE-Lnf.

We reported here the Out of bag (OOB) scores, which indicate the model prediction ability (closer to 1 is better), of RF models at DE-Hai and DE-Lnf (Figure R1.2). Since using leave-one-out strategy (see Section 3.4), each RF model for a resulting time series has its own OOB score. The medians of all these OOB scores are ~0.7 and ~0.8 at DE-Hai and DE-Lnf, respectively, which are good enough to allow predicting reasonable potential GPP anomalies in legacy years. We agree that the uncertainty from one year could alter the legacy effects calculation, but the difference is not strong enough to change the main legacy signals. This can be seen in the different quantiles of legacy effects in Figure 3 based on the ensemble runs (see Section 4.2), and the band range is relatively narrow.

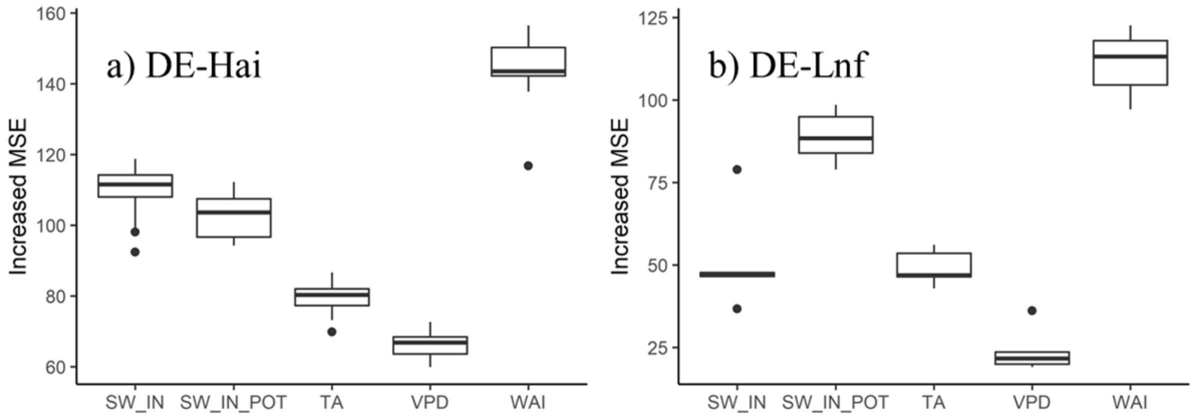

Figure R1.3. Variable importance, indicated by increased MSE, of RF models at DE-Hai and DE-Lnf.

Here, we also showed the increased MSE, indicating variable importance, from ensemble model runs at a) DE-Hai and b) DE-Lnf (Figure R1.3). The water availability index is the most important explanatory factor at both sites, followed by incoming radiation at DE-Hai and the phenological stage (given by potential radiation) at DE-Lnf.

We have added Figure R1.2 and R1.3 to supplementary materials as Figure S2 and S3, adjusted the order of other figures in the supplement, and added the description in line 179 of section 3.4:

'*First, all daily data in non-legacy years were used as input for the RF model to determine the relationships between anomalies of GPP (GPP_{anom}) and anomalies of hydro-meteorological variables (SW_IN_{anom}, TA_{anom}, VPD_{anom}, and WAI_{anom}) along with absolute values of SW_IN_POT to capture seasonal variations in the response of ecosystems to hydro-meteorological conditions. These relationships represented long-term controls of climate on GPP, including drought events and near-average or wet conditions. <u>The Out of bag (OOB) of scores indicating the prediction ability of RF models were ~0.7 and ~0.8 (where zero indicates no skill and 1 denotes perfect performance) at DE-Hai and DE-Lnf, respectively (Fig. S2). WAI_{anom} is the most important explanatory factor at both sites, followed by SW_IN_{anom} at DE-Hai and the phenological stage (given by SW_IN_POT) at DE-Lnf (Fig. S3).</u>*'

R1C5: It doesn't seem like there is any mention in the methods regarding how the length and size of legacies were calculated. It is implied that GPP recovers when it hits the uncertainty boundary, but not explicitly stated.

We briefly mentioned how the length of legacies was selected in Section 3.3 (Lines 153-158 of the original manuscript).

But following the reviewer's comment, we realize that this statement is not explicit enough, therefore we have rephrased it in line 156 as follows:

'we selected a legacy period of two years and this choice was justified by the fact that GPP anomalies residuals returned to the range of model uncertainties (i.e. 25^{th}-75^{th} percentiles of model residuals), <u>which is considered as the point when GPP recovers,</u> in 2005 (see Section 4.3) following the 2003 drought at both sites and, for 2018 at DE-Hai, data was only available up to 2020.'

Minor comments:

L118: What constitutes "good" gapfilling?

It refers to the standard eddy-covariance gap filling algorithm (Pastorello et al., 2020; Reichstein et al., 2005). At the half-hourly (or hourly) scale, there are three different reliable degrees of gap filled data, which are good quality gapfill, medium, and poor. The reliability of gap fill data depends on the availability of data and the size of the time window. For more details, please check Appendix A in Reichstein et al., 2005.

L133: I might be missing something, but this doesn't seem to mention how WAI is calculated. WAIt depends on the calculation of WAIt-1, which is undefined. So, the definition seems circular.

We have added more details in lines 133-136 of section 3.2 about WAI calculation.

'Therefore, we used a bucket model approach based on observed evapotranspiration and precipitation to estimate a vegetation water availability index, WAI (Tramontana et al., 2016), calculated as:

$$\underline{WAI_0 = WAI_{wam\text{-}up}} \quad (1)$$
$$\underline{WAI_t = min(WAI_{max}, WAI_{t-1} + P_t - ET_t)} \quad (2)$$

<u>where WAI_0 is the initial value of the water availability index (WAI), $WAI_{warm\text{-}up}$ is the end value of WAI from the warm-up of the bucket model (Eq. 1). To warm up the bucket model, we ran it 5 times through the first year before starting the actual computation across all considered years.</u> WAI_{t-1} (mm) and WAI_t (mm) were WAI at time step t-1 and t, respectively, P_t (mm) and ET_t (mm) were precipitation and evapotranspiration at time step t (Eq. 2). We set the bucket size (i.e. WAI_{max}) as the maximum cumulative water deficit (CWD) at each site. The estimated bucket sizes were 205mm and 191mm at DE-Hai and DE-Lnf, respectively.

Additionally, we calculated the CWD, which was estimated from cumulative differences between observed evapotranspiration and precipitation over periods where cumulative net water loss from the soil (Σ (ET-P)) is positive.'

L197-199: Detrended how?

We have added one sentence to describe how to detrend in line 199 of section 3.5.

'We detrended the time series of all variables by removing any significant long-term linear trend detected using the Mann-Kendall test (Kendall, 1948).'

L198: Is this annual, or a mean from the growing season? The latter would probably be more relevant.

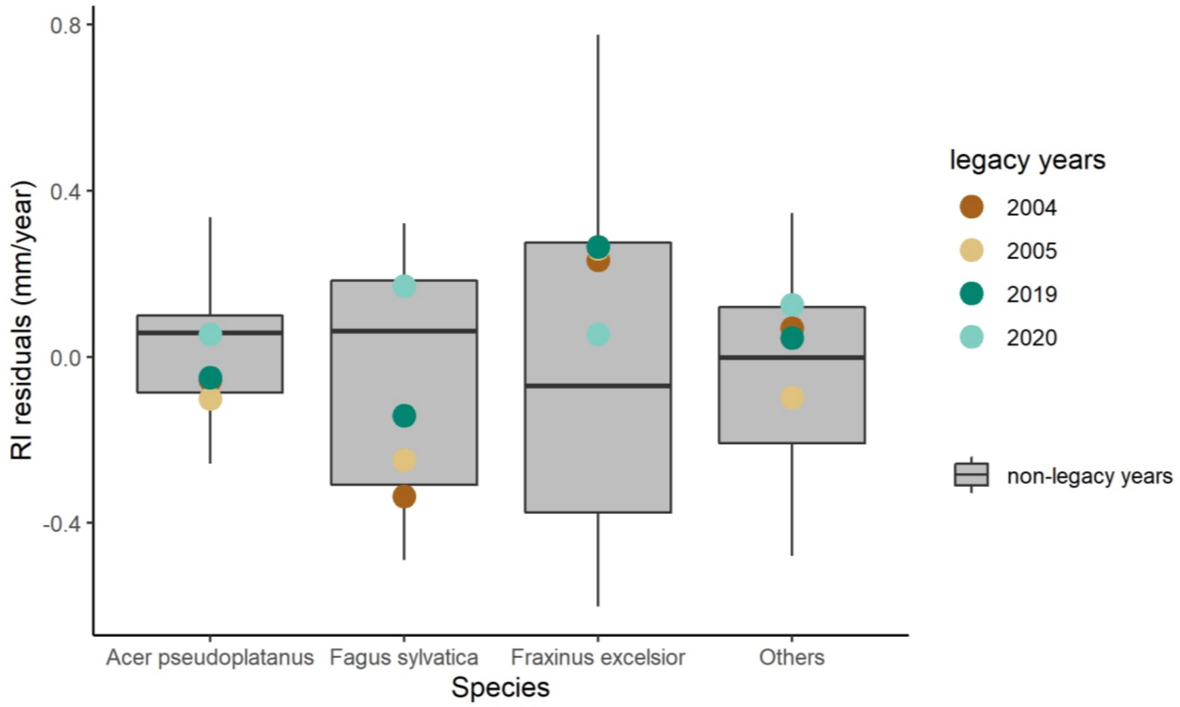

Figure R1.4. Residuals of radial increment (RI) in legacy years at DE-Hai across species.

We agree and have changed to growing-season mean values, but still, only negative legacy effects in 2004 were found on RI of *Fagus sylvatica*. We found only slightly positive or no legacy effects on RI of *Fagus sylvatica* in 2020, which is different from the original results using annual values where RI of *Fagus sylvatica* in 2020 showed strong positive legacy effects. We have replaced the original Figure 6 in the manuscript as Figure R1.4 and have corrected the relevant description in line 198 of section 3.5:

'We used the following explanatory variables: detrended growing-season mean WAI, detrended growing-season mean VPD, detrended growing-season mean SW_IN, and detrended growing-season mean TA for each species.'

in line 306 of section 4.5:

'To complement the analysis of the legacy effects on GPP at the seasonal and annual scales, we also evaluated legacy effects on tree growth at the annual scale. RI of Fagus sylvatica was below the 25th percentile of model residuals in the post-drought year 2004 and returned to the 25th-75th percentiles of model residuals in 2005. For species of Acer pseudoplatanus, Fraxnius excelsior, and others, residuals of RI were almost within 25th-75th percentiles of model residuals in 2004 and 2005. After the 2018 drought, RI of all species for 2019 and 2020 were almost within or close to 25th-75th percentiles of model residuals.'

and in line 405 of section 5.2:

'Stronger negative legacy effects on GPP in 2020 than those in other legacy years were found at DE-Hai in the seasonal and annual scales. This might be associated with significant tree mortality

in the period 2018-2020 (about 6% year^{-1} between 2017 and 2020 compared to less than 1% year^{-1} between 2005 and 2017) mainly caused by the storm Friedrike in January 2018 and the heat and/or drought in summer 2018 and 2019 (unpublished data). RI of Fagus sylvatica in 2020 showed <u>*slightly positive legacy effects*</u> *in growth, since only living trees were sampled. This might be explained by the favorable weather conditions in winter/spring 2019/2020 associated with high mineralization rates and reduced competition for nutrients, light and water of the surviving trees (Grossiord, 2020). The RI data reflected mean growth signals from individual surviving trees, while the GPP data reflected mean carbon assimilation at stand level, including positive, negative or absent legacy effects at individual tree level as well as the reduction of assimilating individuals due to higher tree mortality.'*

L443: Some of your citations do exactly this, so perhaps cut this sentence.

Thanks for pointing this out, we have deleted it.

Reference

Mund, M., Kutsch, W. L., Wirth, C., Kahl, T., Knohl, A., Skomarkova, M. V., and Schulze, E.-D.: The influence of climate and fructification on the inter-annual variability of stem growth and net primary productivity in an old-growth, mixed beech forest, Tree Physiology, 30, 689–704, https://doi.org/10.1093/treephys/tpq027, 2010.

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., Poindexter, C., Chen, J., Elbashandy, A., Humphrey, M., Isaac, P., Polidori, D., Reichstein, M., Ribeca, A., van Ingen, C., Vuichard, N., Zhang, L., Amiro, B., Ammann, C., Arain, M. A., Ardö, J., Arkebauer, T., Arndt, S. K., Arriga, N., Aubinet, M., Aurela, M., Baldocchi, D., Barr, A., Beamesderfer, E., Marchesini, L. B., Bergeron, O., Beringer, J., Bernhofer, C., Berveiller, D., Billesbach, D., Black, T. A., Blanken, P. D., Bohrer, G., Boike, J., Bolstad, P. V., Bonal, D., Bonnefond, J.-M., Bowling, D. R., Bracho, R., Brodeur, J., Brümmer, C., Buchmann, N., Burban, B., Burns, S. P., Buysse, P., Cale, P., Cavagna, M., Cellier, P., Chen, S., Chini, I., Christensen, T. R., Cleverly, J., Collalti, A., Consalvo, C., Cook, B. D., Cook, D., Coursolle, C., Cremonese, E., Curtis, P. S., D'Andrea, E., da Rocha, H., Dai, X., Davis, K. J., Cinti, B. D., Grandcourt, A. de, Ligne, A. D., De Oliveira, R. C., Delpierre, N., Desai, A. R., Di Bella, C. M., Tommasi, P. di, Dolman, H., Domingo, F., Dong, G., Dore, S., Duce, P., Dufrêne, E., Dunn, A., Dušek, J., Eamus, D., Eichelmann, U., ElKhidir, H. A. M., Eugster, W., Ewenz, C. M., Ewers, B., Famulari, D., Fares, S., Feigenwinter, I., Feitz, A., Fensholt, R., Filippa, G., Fischer, M., Frank, J., Galvagno, M., et al.: The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data, Sci Data, 7, 225, https://doi.org/10.1038/s41597-020-0534-3, 2020.

Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., Bernhofer, C., Buchmann, N., Gilmanov, T., Granier, A., Grünwald, T., Havránková, K., Ilvesniemi, H., Janous, D., Knohl, A., Laurila, T., Lohila, A., Loustau, D., Matteucci, G., Meyers, T., Miglietta, F., Ourcival, J.-M., Pumpanen, J., Rambal, S., Rotenberg, E., Sanz, M., Tenhunen, J., Seufert, G., Vaccari, F., Vesala, T., Yakir, D., and Valentini, R.: On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm, 11, 1424–1439, https://doi.org/10.1111/j.1365-2486.2005.001002.x, 2005.

Rocha, A. V. and Goulden, M. L.: Drought legacies influence the long-term carbon balance of a freshwater marsh, 115, https://doi.org/10.1029/2009JG001215, 2010.