Reviewer #1

General comments:

First, kudos to the authors for publishing their code. This is good practice and helps ensure reproducibility.

We thank the reviewer for this comment and think that code should always be made publicly available.

I am not asking the authors to do this, but the other equally important dimension of calculating MCWD is the estimate of ET. Obviously this can produce large differences, so I don't quite understand why this was not considered. These different combinations of precipitation and ET products could produce some very contrary estimates of MCWD. I think this would be a very citeable finding that this analysis is well suited to do, and it would be a very useful contribution to the literature. However I understand if this is infeasible, and the discussion of ET differences in the discussion is useful.

We thank the reviewer for this very good suggestion and now included both ET datasets (DOLCE and GLEAM) that the reviewer suggested. We extended our analysis of the sources of variability from precipitation datasets and drought indices to now also include the variability caused by the choice of evapotranspiration dataset. We further also compare and quantify the differences when using variable ET from DOLCE and GLEAM ET against the fixed ET= 100mm per month assumption (new Fig. 4). Thereby, we find an overestimation of drought stress for all the three drought years 2005, 2010, 2016 when using a fixed ET of 100 mm per month. This overestimation gets more pronounced the further South the drought is located. We rewrote parts of the methods, results and discussion sections in the light of these new analyses.

The calculation of relative MCWD anomalies is a bit confusing. I did not understand when the 10 year interval was used to calculate the baseline, and when 16 years was used (L165). Ten or even 16 years is short for a reference period, which is typically closer to 30 years.

We used 10 years for the 2005 and 2010 drought events and 16 years for the 2016 event. We did so, because some datasets, like TRMM v6 and GSWP3 end in 2010 and others, such as TRMM v7, only start in 1999.

The discussion around differences in precipitation products is useful (as is Figure 3), but I wonder if this analysis (or discussion) could probe deeper into why these products disagree in some areas. Is it because of differences in ground station locations used by the products? Is it because some only use infrared data, and others incorporate microwave soundings?

We added some more sentences about the differences between the precipitation datasets to the discussion (e.g. lines…) and now also mention bias-correction as another source that introduces differences across the precipitation datasets. We agree that it would be very interesting to go even deeper and explore if we can find any patterns of methodological origin that can better explain the differences between the datasets. However, because of the complexity (climate models vs. satellite observations, reanalysis, bias-correction, etc…) with which such datasets have been created we argue that this would rather require a dedicated study itself. The precipitation datasets used for this study are very independent (see Table 1) and therefore there is not a surprise that they differ substantially even at a global scale (see e.g. Figure 2.15 Gulev et al., 2021), and even less surprising on this regional scale (Doblas-Reyes et al., 2021). This is why it is so important to take into account the observational uncertainty in regional climate studies. For example, four of the products are based on different reanalysis - these are four different Global Climate Models that assimilate observed data during execution. The simulatedprecipitation fields of ERA5 are not bias-corrected while NCEP-CRU and WATCH_WFDEI are bias-corrected with the gridded product CRU while GSWP3 is corrected with the gridded product GPCC. Even CRU and GPCC can give very different results at a regional scale (see figures 10.12 and 10.13 in Doblas-Reyes et al., 2021). Similarly, the products CHIRPS and TRMM are not based on comprehensive global climate models, but on satellite data that use different instruments and retrieval models (TRMM and CHIRPS), CHIRPS is further merged with observed in-situ data.

Concerns:

My biggest concern is regarding the simplistic estimation of AGB loss. I am disappointed to see the authors did not accept my earlier recommendation to drop this. The Lewis et al., (2011) paper managed to get an estimate based on a simple one term regression with a lot of actual forest inventory data, but this does not mean this is a robust way to estimate carbon loss. It does not make sense that the same MCWD value would cause equivalent loss of AGB across Amazonia when the baseline carbon stocks are different, and the forests are adapted to different seasonal variations of MCWD (i.e. aseasonally wet northwest vs seasonally dry southeast). It is not surprising that somewhat different numbers will be generated from this (flawed) approach (Figure 4), and I worry that we will see more of this approach if I were to accept this. Again, I ask the authors to remove this part of the manuscript. I think the rest of the manuscript is acceptable, but not this estimate of AGB carbon loss.

We agree with the reviewer that it is probably not feasible to apply the MCWD-AGB from Lewis et al. 2011 to the other MCWD estimates of our study. We dropped Figure 4 and all MCWD-AGB related estimates from our study. We still want to highlight that the goal of this study was not to give better estimates of the drought impact, but rather highlight the differences that arise by purely choosing a different precipitation dataset.

The sections in the Results about the calculation of the RAI and scPDSI should include more specific details about how these indices are actually calculated. It would be more clear to list the equations. Also I don't think equation 1 is quite correct as the WD or CWD is constrained to always be ≤ 0.

The reviewer is correct. We modified equation 1 accordingly:

if (P(t) - ET(T) < 0)

WD(t) = P(t) - ET(T)

else

WD(t) = 0

The ERA5 PET has a known bug: "The Potential Evaporation field (pev, parameter Id 228251) is largely underestimated over deserts and high-forested areas. This is due to a bug in the code that does not allow transpiration to occur in the situation where there is no low vegetation." from https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation.

Even the ERA5-Land PET seems problematic. Perhaps it would be better to derive a monthly climatology from GLEAM (https://www.gleam.eu/), or use one of the recent multi-product merges such as the newer version of DOLCE (https://hess.copernicus.org/articles/22/1317/2018/). I mention using a climatology of ET instead of actual monthly estimates of ET (or PET) because it would account for seasonal variation (L325), but also because the error of any ET product is likely to be very large with potentially spurious seasonal patterns.

We thank the reviewer for this detailed critique regarding PET data sets. We removed ERA-Land PET from our study and included GLEAM and DOLCE in our study.

Figure 4: I strongly suggest removing this figure.

We removed this figure from our study and included a new figure 4 (see response to your comments above).

Figure 6: This is a useful figure but these colors (red and green) are not distinguishable by colorblind people. Yellow on white is also difficult to distinguish.

We thank the reviewer for the close look and chose a different color scheme for figure 6.

Literature

Doblas-Reyes, F. J., Sorensson, A. A., Almazroui, M., Dosio, A., Gutowski, W. J., Haarsma, R., Hamdi, R., Hewitson, B., Kwon, W.-T., Lamptey, B. L., Maraun, D., Stephenson, T. S., Takayabu, I., Terray, L., Turner, A., & Zuo, Z. (2021). *Linking global to regional climate change* (V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Pean, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekci, R. Yu, & B. Zhou, Eds.). Cambridge University Press. https://centaur.reading.ac.uk/99896/

Gulev, S., Thorne, P., Ahn, J., Dentener, F., Domingues, C., Gerland, S., Gong, D., Kaufman, D., Nnamchi, H., Quaas, J., Rivera, J., Sathyendranath, S., Smith, S., Trewin, B., Schuckmann, K., & Vose, R. (2021). *IPCC AR6 WGI Chapter 2: Changing state of the climate system*.

Reviewer #2

The revision addresses some of the major comments in the first round (e.g. questions on extrapolating the drought-mortality relationship derived from the 2005 drought to 2010 and 2015/2016) and improves the clarity and accuracy of the manuscript. However, I feel one of my major comment about why the drought intensity differs was not fully answered.

First, I was suggesting pair-wise scatter plots (or heat maps) between MCWD generated from all data sets. Such figures common for all inter-comparison studies and accompanying regression analyses can tell the spatio-temporal correlation (R2) and systematic biases (intercept and slope). In the revision, the authors present a comparison of CDFs, which are very qualitative and do not contain the spatio-temporal structure as in a scatter plot. I am still suggesting the inclusion of such pair-wise comparisons (either using scatter plots or just reporting correlation/regression statistics) instead of comparing CDFs.

We are sorry that we did not fully address the reviewers comments regarding the scatter plots appropriately. We added pairwise scatter plots for all precipitation datasets and the three drought years 2005, 2010, 2016 to our analysis (Fig. S3-5). We could not find any obvious biases in the datasets apart from some spikes in the ERA 5 and GLDAS dataset.

Second, there are several tricky steps when translating uncertainties in MCWD into uncertainties in vegetation mortality. Aside from the robustness of the drought-mortality relationship as mentioned by me and the other reviewer in the first round, another reason is that the Lewis et al. 2011 relationship was generated using a specific data set, which was then used to transform MCWD from all the other datasets in this study. Isn't it a more fair comparison to first calibrate the drought-mortality relationship in 2005 using each data set? I understand it might be not easy to get the original data and do the same analysis. However, it is easy to linearly 'project' the different data sets onto the space of the data set as used in Lewis et al. 2011 (TRMM or GPCP) in the first point I made above. For example, if MCWD_TRMM = a * MCWD_CHIRPS + b from regression analysis, we can transform CHIRPS MCWD to the TRMM space using the relationship and then calculate the carbon loss. This can help to more clearly explain some of the differences in Fig. 4.

We thank the reviewer for this comment. We like the idea of getting the specific dataset with which the relationship for the 2005 drought was derived. However, we could not get access to the dataset and could also not reach the author of the study. If the reviewer has access to this dataset we are happy to perform such analysis in a follow-up study. While we also like the linear projection idea of the MCWD datasets, we decided to remove the MCWD-AGB analysis (and also figure 4) from this study as reviewer 1 pointed out flaws of our MCWD-AGB estimation.

Instead, we included two evapotranspiration datasets – DOLCE and GLEAM – in our study. We now also investigate the influence of such variable evapotranspiration input to the drought indices and compare it to the widely used ET=100mm per month. We updated the methods, results and discussion parts of the manuscript accordingly.

Finally, as raised by the other reviewer in the first round, I am now wondering about the suggestion of using an ensemble of rainfall data sets in the last paragraph ("We therefore recommend applying several climate (precipitation) datasets as well as drought metrics to account for model uncertainty when assessing the spatial extent, duration, and location of droughts"). Ensemble arises from the climate systems being chaotic and applies mainly for future predictions. However, for the drought that has already happened, there was a real and single number of rainfall for each location. So, shouldn't a recommendation of calibrating the gridded data with more ground observations be more logical?

We are sorry that our recommendation causes confusion. We acknowledge the reviewers' conclusion leading to their recommendation of including more ground observation into the dataset. However, we still think that our recommendation using multiple datasets/datasource is valid. Recent studies assessing the impact of drought events e.g. on forests often also use only one dataset to estimate drought extent and severity for both present and past drought events. With our analysis we show that such drought impacts are very dependent on the choice of precipitation dataset, the drought indicator and the evapotranspiration estimate. We think that any study that estimates basin wide drought stress should therefore take multiple datasets, etc into account.