Reviewer #1

The manuscript presents maize yield prediction results based on a model that combines hydrological, meteorological, and remote sensing features in a random forest regression. The authors performed feature importance and sensitivity analyses to determine which features influenced maize yield predictions the most and which types of features contributed most to yield prediction accuracy. Overall, the paper is well written and presents useful findings for future studies.

We thank Dr. Kerner for her thorough assessment and helpful comments.

My main criticism for this manuscript is that the random 80%/20% train/test split of yield observations and use of the test set in model optimization are likely overestimating the performance of the test set compared to the performance that could be expected in practice. The 80%/20% random split across all available yield observations across 70 districts and 8 years does not ensure the test set is independent from the training set. For example, observations of the same district in different years probably have high correlation, as are observations in the same year but different districts. The authors are also optimizing the model for their test set by performing RFE using the test set (they should instead use a third validation set or cross-validation with the training set as was done in 2.3.2). The goal (I assume) of this study is to present a method that can be used to predict maize yields in future years – for the test set to be representative of the performance in this setting, there should be no overlap in years present in the training and test sets (e.g., the training set could include observations from 6 of the 8 years and the test set include observations from the other two years).

We agree with the reviewer's comment that testing/training sampling split per year can provide more robust evaluation statistics. In the revised manuscript, we will include updated statistics using the proposed year-based testing/training sampling approach. We also agree with the reviewer's comment regarding the RFE, in the revised manuscript we will include cross-validation on the RFE feature selection using the year-based testing/training scheme as proposed by the reviewer.

Additional comments:

- Why did the authors use 250m MODIS instead of 30m Landsat-resolution NDVI? The latter is much closer to the field scales observed in Zambia and would have the same resolution as the HydroBlocks simulations.

We agree that Landsat is a better sensor for this purpose, however, MODIS was used instead of Landsat mainly because of the high cloud coverage and a long revisit time of Landsat in Zambia, especially from January to March. Our estimates assessed that cloud coverage between December to February was ~50% in Landsat and ~30% in MODIS (2014-2017). This difference was particularly relevant when calculating the NDVI integrals over the season, in which missing half of the observations could significantly change the integral results. We will add this information to the revised manuscript.

- What computational, time, or cost resources would be required to use the HydroBlocks model operationally to predict maize yields in all of Zambia in future years? Is this feasible to do operationally? (Also, note that sometimes the authors write “Hydroblocks” and sometimes “HydroBlocks”.)

At this first run, the HydroBlocks simulations (1981-2017) took roughly two weeks wall clock time to complete using 200 cores on the Princeton University supercomputing facility (10 nodes, with 20 cores and 128GB per node, in a 2.6 GHz Haswell processor). However, recent developments in the model now allow for faster computations. Although we have not explored this potential yet, as the purpose of these simulations was to understand the historical relationships between yields and droughts, we estimate that ensemble future prediction for 1-5 years could be possible within 1-3 days using the same resources. Therefore, although very powerful, the large computational and big data storage requirements probably wouldn’t allow local managers to apply this modeling framework directly, but it can provide a workflow basis for future research development, as highlighted in the conclusion section.

We have replaced “Hydroblocks” with “HydroBlocks” throughout the manuscript. Thanks for the comment.

- The ESA-CCI 2016 land cover map is used as a cropland mask, and the authors assume all cropland is maize. How valid is this assumption (i.e., what percentage of crops grown in Zambia are typically maize?)? What is the accuracy of this land cover map across Zambia? (I have not seen promising results for this map in Africa.) This could affect the authors’ interpretation of shrubland percentage as an important indicator of maize yields.

We relied on the assumption that all cropland is maize because of the lack of ground truth for developing a crop type model that would allow us to distinguish maize from non-maize croplands. This assumption is of course incorrect, but maize is by far the dominant crop by planted area in Zambia. According to Zambia’s Post Harvest Survey data from 2014-2015, planted maize averaged 60% of the total planted area across all of Zambia’s provinces. It’s lowest share was 28.4% of planted area in Luapula Province, but the maize share fell below 50% of the area in only three provinces that together accounted for 28% of Zambia’s 2014-2015 planted area. In the remaining 72%, the maize share was 69% of planted area.

The odds are thus fairly high that a randomly selected field in Zambia’s cropland will be growing maize. Previous studies that used remote sensing to estimate maize yield in Zambia also relied on this same assumption (Azzari et al, 2017). In terms of the impact that this assumption could have on our results, they would likely reduce the accuracy of our yield predictions relative to the PHS dataset, particularly in those districts where the maize share falls below 50%. Mitigating that loss of accuracy, however, is the fact that our model’s predictions are influenced only to a small degree by crop-specific data--since LAI/NDVI makes relatively minor contributions to our model, LAI/NDVI collected from a non-maize pixel will have introduce only small error into the predicted yield. In this sense, our model is essentially predicting for each cropland pixel what the maize yield would be if maize was growing in that pixel.

In terms of accuracy of the cropland mask derived from ESA CCI, we assessed it using an independent accuracy assessment conducted under a separate project. The accuracy assessment was based on a reference sample collected using visual interpretation by three separate raters of both high resolution and Landsat imagery available in CollectEarth, and resulted in 608 validation points where all three raters agreed (444 non-cropland, 164 cropland). Using this sample, the ESA CCI map over Zambia was shown to have user’s accuracies of 59% (cropland) and 89% (non-cropland), producer’s accuracies of 71% (cropland) and 82% (non-cropland), and overall accuracy of 79%. Since the largest source of error in this map was commission error (41%) by the cropland class, this means that a substantial number of non-cropland areas was predicted to have a maize yield.

Maize yields were calculated at the 250-m pixels for where percentage exceeded 50%. Although our results indicate that shrubland percentage is an important indicator for maize yields, we would like to highlight that, as shown in Figure 5S, high shrubland percentages are associated with low yields (Figure 5S). Thus, despite the low accuracy in land cover classification at the 20-m resolution, the shrubland percentage at 250-m resolution may be accounting for lower productivity at sites that are cropland/shrubland mosaics--which are more likely to be lower-yielding smallholders’ fields, and at the same time actual shrublands/cropland mosaics may be more likely to be misclassified as pure cropland (a commission error), and thus explain the negative influence that shrubland percent has on yield. We will add discussion of these important points to the revised paper.

- The colormap in Figure 9 (right) is missing a title.

Corrected.

- The Figure number is not shown on line 434.

Corrected.

- “Unknowing” should be “Not knowing” on line 60.

Corrected.

- "its" should be "their" on line 2.

Corrected.

- Lines 251-252: the second set of MAE and R-squared values should say they are for the training set (only the test set is mentioned).

Corrected.