* The issue of pixel aggregation at the district level needs to be addressed, it’s a common question (came up during the workshop and we’ve had some email exchanges on this before).
  + Are you simply calculating the average of the pixel values at district level? Or some other measure?
  + Pixels with in area of interest (AOI), district boundaries in this case, can be evaluated on a pixel by pixel basis in utilizing parrallel processing or summarized by the AOI's mean value for each image. In this study, district level mean values of EVI are used to represent agricultural productivity for each 8 day period.
  + Is aggregation happening before calculating the summary statistics and integration measures then used as indep vars? Or these measures are first calculated at pixel level and then aggregated? It’s probably the former but these two methods could give quite different results
  + It can be done either way, in this case we find mean values for each 8 day period then run statistics on that.
  + In any case, some robustness analyses could be done to see if and how aggregation affects results?
  + This could take a while, so I will skip unless reviewers give us trouble.
* Related to the above, are the pixels used within a district always the same across years? Say, a district encompasses pixels 1 through 10. Now, first year pixels 2 and 5 have cloud cover. Then again some of them are removed through land use information (which I assume remains constant over time?). Second year, perhaps pixels 3 and 7 have cloud cover. But then the pixels used for the aggregation are different and differences between years 1 and 2 could (at least partially) come from this variation in the sample of pixels. This is important to address.
* Masking of land use changes varies by time, same with cloud cover. So the number of and ID of pixels used will change over time. Not sure there is any good way to get around this. However considering the use of 8 day composite images, most images will be ‘cloud-free’.
* You mention spline smoothing of EVI values + outlier removal. Are these two different procedures you are conducting? Or just stating that spline smoothing is implicitly removing outliers? If the former, some explanation would be good.
* To minimize the effects of the artifacts described above we test the use of temporal smoothing splines and outlier removal . Where outliers above three standard deviations are removed before applying a cubic smoothing spline.
* Overall, I think the introductory section on setting up all the variables to be used later in the model needs more explanation and a more clear description of the steps followed.
* Ok will try to improve
* Now, to the variables and the econometrics. The discussion of percentiles wasn’t very clear to me. You mention that you use historical percentiles in order to see variation with respect to some historical benchmark. But these are then constant across all years for a given district, right? Cause in the table it somehow seems that percentiles are calculated within a single year. Perhaps these percentiles could be used differently, for example rescaling variables in different districts as a percentage of their 95th percentile before the PCA? I think this would capture this historical benchmark a bit better.
* Explanatory variables. Some of them are clear but some are not, I think a short note is needed in the table briefly explaining how these are constructed or what they represent. In particular:
  + VEG\_growing\_max\_date. Is this the date at which the max EVI is reached?
  + VEG\_growing\_v2, there is no explanation of this v2 and how it differs from “v1”.
  + “Leading”, “trailing”, “diff”, what do these mean exactly?
  + There’s 3 yield variables in the table, you mention you use the first one as dependent, but then what are the others there for? I don’t think they are being used as independent, right?
  + R\_mx\_dates, what is this?
  + Some summary statistic variables could make sense, but some others make me wonder, like for example min or mean. In a way MEAN is = to AUC/N (with N=number of 8-day periods), right? Aren’t these variables co-moving almost identically? Also, min perhaps should be used in the same way as percentile, that is, as a benchmark for that district/pixel? For example, two districts, one of them has more/less water than the other, so its EVI value is consistently higher/lower. Then shouldn’t EVI be somehow rescaled to make them comparable across districts? (all non-agricultural areas are NAed out and therefore excluded from the sample) One way of doing this would be to substract the minimum EVI value from all EVI values and then calculate all measures using this rescaled series…
  + Finally, around 40+ variables are listed in the table, but you mention you use 28 for the PC analysis, which ones are these? (text has been updated, all Xs included in PCA)
  + I include an updated table under the Variable Definitions section. I hope this improves things.
* Regressions.
  + You use the 22 principal components in the regressions. I would probably use a lot less and see if the main PCs can explain yields reasonable well. What you are getting from the last ones is probably mostly noise, as they explain very very little variation. I am not super familiar but I believe there is some sort of rule of thumb to decide up to which PC to use.
  + Will reduce number
  + It would be nice to see how the first few PCs are constructed. Perhaps there is a nice story behind them, with each representing a few main variables. This could help in teasing out what works best at explaining yield variation and it would be a nice addition to the discussion section.
  + What is the rationale for having the 1st PC lagged? That wasn’t clear to me and I wonder if doing that does not introduce some econometric issue… Also, in equation 2 you show PC#1 as lagged, and then PCs #2 to #22 as contemporaneous, but then in the R output it looks like PC#1 is both lagged and contemporaneous…
  + Yeah I dropped this not needed
* The spatial autocorrelation is interesting and probably important in our case. However, I find it odd that while Moran I’s test shows spatial autocorrelation the spatial RE regression output you have at the end shows an insignificant spatial coefficient (lambda), which would imply that there does not seem to be spatial autocorrelation! Perhaps it’s an issue of the exact spatial matrix you were using or the method itself (I understand one can introduce spatial lags at the independent variable, dependent variables, or error term levels, though I am not sure if all of them have an RE estimator version…)
* There were some issues around RE, included a spatial error lag by accident if I am remembering correctly. Using a spatial lag FE, lambda is now significant and improves model performance.
* Out-of-sample prediction would be a really nice addition, as I think many of the issues I’ve raised ultimately link to this. Two alternative procedures could be (1) a temporal out-of-sample (calibrating the model with the first, say 8, years, and forecasting the last 2 years for every district; and (2) a “spatial” out-of-sample, calibrating the model with N-5 districts and then predicting in those 5 districts for the full 10-year period. On specific metrics to assess the goodness of this forecasting I am not sure to be honest.
* I am including a leave-p-out estimates of RMSE for the moment as a metric. I however am planning to work on using this period to estimate the unbalanced portion of the record (past 2012) however this has been pretty difficult since there is no predict functions for splm. I don’t know how well it will work with unbalanced data either…

- I find it concerning that kharif indicators - which appear the most important in the first principal component - are predictive of wheat yields, even though we are controlling for the wheat NDVI measures in some of the other PCAs. (Are these kharif indicators before the rabi season of interest?) Could it be that the yields are picking up something that influences both yields and vegetation in both seasons? What is the rationale for including kharif indicators in predicting wheat yields?

- Districts must span thousands of pixels - how are you aggregating those pixels before coming to the final analyses? It wasn’t entirely clear to me. On a separate note, there might be a lot of additional data with georeferenced crop cutting experiments out now; I’d be really curious to see whether results survive when working at a smaller level than the district, because there will be much more variation to explain.

- In the section on testing for pooled, random and fixed effects models: I think in the last sentence we should say that the fixed effects estimator is used because it is the only one that is consistent (without referring to efficiency as an argument). It also feels a bit odd that this section comes first, and that later the pooled OLS results are presented nevertheless - shouldn’t this section on the tests come after describing the different models and their results? Or drop the description of the other models since you are already showing that they are inconsistent?

- For the natural cubic splines, I’m wondering what the added value is given that in the main specification they don’t seem to do a lot. Also, it is a transformation of an already transformed variable (the PCAs), so we’re losing even more insights into what type of information is going into the predictions. This is worrisome especially given that the number of observations and degrees of freedom is not large.

- Are you controlling for year effects here? Might be good, since there will be time trends. It would be good to see how much variation there is left within districts before and after controlling for year effects, both in terms of yield and NDVI variables. I’m worried that there is not much variation.

- Study area: It would be good to describe more about the agricultural setting: mostly irrigated land, hardly any rainfed agriculture (so drought is not a major shock to yields anymore); farmers have embraced modern varieties and mechanization. Punjab and Haryana are quite different from the rest of South Asia in how agriculture is practiced. There is also much less intercropping, which is important from a remote sensing perspective. Could also be added later but a discussion of this will be important.

- When introducing the methods, it would be good to intuitively explain the fixed effects approach and why we think it will be better to use fixed effects predictions as opposed to pooled models or spatial models.

- The growing seasons metrics section refers to a Figure 2 that I’m not sure is actually in the paper?

1. My worry is that it has an effect *after*already controlling for the NDVI measures from rabi (or at least we would want to properly control for NDVI measures from rabi) => are you arguing that khariv NDVI is predictive perhaps of other growth components that are not reflected in rabi NDVI? If we’d want insurance companies to adopt this, we’d need very good reasons to argue that khariv NDVI picks up something in wheat yields that rabi NDVI can’t; they might find it counterintuitive, unless we can argue that it influences growth beyond what the rabi NDVI itself can capture.

I think the issue with Kharif is that my metrics for Rabi aren't going to pick up on every nuance. So spillovers from the proceeding season might be missed, especially early season water availability for germination when things are not 'green' yet but water availability is still critical.

2)      I would just flip the order: first present the different models, how they look different, and then argue that the fixed effects estimator is preferred.

3)      Certainly, we can help. I’m caught up with a few other things now but could take a stab at it in two weeks from now.

4)      How do the splines influence the accuracy of out-of-sample predictions? I’m not surprised that within sample it does better but often when we are introducing non-linearities our out-of-sample predictions become worse, so maybe out-of-sample is a better metric to look at.