*To begin with, estimate each of the three equations below separately. If time, then estimate them together as a system of equations (SEM)*

*September/October – writing.*

*Most likely, I will not have time to do all tasks listed below or answer all questions, but this can serve as a guideline.*

# Equation 1 = Production function

**Prod = f(Temp, Precipitation, other climate data)**

# Equation 2 = Inverse demand function

**Price = f(production, international price, trade)**

* Match the price data to counties (they are per market place, not for counties)
  + - Do I have a data point for each county?
    - To begin with, I can do this just googling the locations of the market places. If there is time, I can use more sophisticated method of interpolation (i.e. polygons, see e.g. Erin’s Lentz paper)
* Get the international trade and price data
* What unit of analysis will I use?
  + Problem – Price monthly and county level, Production – yearly and county level, Trade – just national level, also International prices not at county level
  + What frequency of international prices and trade data will I be able to find? (monthly? Yearly?)
  + If I don’t find a good data or a way how to deal with the data incompatibility, I may have to abandon estimating of the Equation..?
* Idea: first estimate Equation 1 using aggregated (average) z-scores per year. Then use the estimated coefficient to get fitted values of production per month and use these fitted values as inputs into this equation…
  + Does this way of interpolation make sense?
* For the crops data I need to match the county IDs

# Equation 3 = Social indicators (e.g. MUAC)

**MUAC (or other indicators) = f(price,….,)**

* Shall I only use price in this equation or also test for the climate data? More importantly, shall I use yearly aggregates of the climate (as Erin) or monthly aggregate?
  + If yearly aggregate-> more climate data needed. Climate data currently available for period 1999-2015 while NDMA phases available for 2014-2017-> not enough overlap
* Finish the replication of Erin’s paper
* **Do we actually want to use the NDMA phase classification, given that we know exact rule for change from Normal to Alert, and the rule is not always obeyed? There may be political factors behind decisions about the phases, which I believe would be difficult to model using climate and weather.**
  + **If we decide to use the NDMA phase classification anyways, what about using is as two values response? I.e. just model the counties for example in Alert and Alarm or Merge Alert and Alarm (as there is only 10% of Alarms). I believe that using probit/logit model in panel context would be easier to apply than ordinal probit (multiple response)**
* As for the data which I got from the NDMA reports (i.e. the EW phases, MUAC and CSI) - they are only available for the 23 ASAL counties. Is there a way to get them for the other (non ASAL) counties?
  + If not, how can I deal with this? If I decide to only estimate the equation for the ASAL counties, I am not sure if it will be possible to estimate the 3 equations simultaneously. Maybe I would need to restrict all 3 equations for the ASAL counties.
    - What would this mean for the models and for their estimation?
* If we decide to model the phase classifications then I need to study carefully:
  + (Ordinal) Logit in panel data context – There is a whole chapter on section on binomial response models in panel data context it in a Woodridge textbook
  + How the decisions about the phases are made. I believe that for this the best would be reading the report ‘A broad scale summary of existing early warning systems in Kenya and regional initiatives in the Greater Horn of Africa’ (see email from Pedram 31. January – WP1.1\_Analysis\_EWS 2901118.docx). Maybe also reading description/explanation in a couple of the NDMA reports would be helpful

# All 3:

* **Panel regression models:**

**Find the best specification as follows. I can estimate many possible specifications and do various test (see plm package in R). Study more about the specification / residual testing as I go. Possibly, even use some selection operator (lasso). This task can take at least a week or more for each equation. (and if a new data or information appears after I have done this, I may need to do this again)**

* Do I actually need the fixed effects variables in the panel regression (e.g. education, life expectancy, mortality = the data from KNBS, surveys) which are only available for one or two years? Or do the county level dummy variables (fixed effects) account for these?
  + If these variables are useful, it would be needed to digitalize them from the pdf files on the KNBS webpage, potentially get more detailed data from Kenya
* Maybe I should at least use a proxy for GDP (or not needed?)
  + If we decide yes for this – I should ask Alex Moradi if he could provide the historical railway data as a proxy of GDP
* Study once more carefully how VCI and SPEI for various periods (number of months) are constructed **(for VCI read carefully the paper of Klisch and Atzberger 2016**, see email from Pedram 4. April 2018)
* It would be good to get more county level climate data if possible. Currently I have tem for period 1999-2015. Maybe I could try to search google if I can find some county level aggregates of the climate data?
* Study spatial models in panel data context and apply them. Particularly for **equations 2 and 3** as Annemie said that for yields there are probably not much of the spatial effects
* Study more about SEM and possible combination of panel data structure and SEM
  + Good literature sources:
    - Wooldridge econometric textbook
    - Paper ‘systemfit: A Package for Estimating Systems of Simultaneous Equations in R’ by Henningsen and Hamann (2014) https://www.researchgate.net/publication/26538601\_systemfit\_A\_Package\_for\_Estimating\_Systems\_of\_Simultaneous\_Equations\_in\_R

* + Is it possible to have different numbers of observations and different time frequencies in each equation?
  + Which estimation methods would be the best? (e.g. Instrumental variables, SUR,…)
* Study about (S)ARIMA class of models
  + Could I apply (S)ARIMA relatively easily in my case? Or at least test for lagged dependent variables?
  + Or can the panel data structure capture the effects of lagged dependent variables? (if present)
  + Study about Kalman filter and its applicability in my case (paper(s) of Molly Brown (2015)
* Maybe read again papers of Molly Green and Greenstone. Time series (Kalman filter approach). But they use climate data in general not drought (this is what Pedram once pointed out, I believe)
  + Can I apply something like this in my case?