Data documentation

Harmonization of Tegemeo Kenya data (1997, 2000, 2004, 2007)

and downscaled GCM future climate data

Ben Schiek

4 August 2013

Table of contents

[Introduction 3](#_Toc363349238)

[Level ID Variables 5](#_Toc363349239)

[Revenue 6](#_Toc363349240)

[Cost 7](#_Toc363349241)

[Crop and fertilizer price and quantity variables, per-acre variables (yield and fert/acre) 9](#_Toc363349242)

[Labor 10](#_Toc363349243)

[Farm demographic and infrastructure variables 13](#_Toc363349244)

[Soil 15](#_Toc363349245)

[Hybrid adoption variables 15](#_Toc363349246)

[Baseline climate variables 16](#_Toc363349247)

[Perturbed (future) climate variables – 2050s 17](#_Toc363349248)

[References 20](#_Toc363349249)

Introduction

The following new data files are generated from the raw Tegemeo farm data and downscaled future climate data downloaded from the CCAFS climate data portal website.

“farmlev07B pre-clean.dta”

“farmlev07B clean.dta”

“farmlev04B pre-clean.dta”

“farmlev04B clean.dta”

The clean files are also available in excel format:

“Kenya 2007 Clean 03-08-13.xlsx”

“Kenya 2004 Clean 03-08-13.xlsx”

These files contain farm level variables for:

* Crop and livestock revenue
* Crop and livestock cost
* Crop production and yield
* Fertilizer use
* Hired and family labor
* Farm demographic and infrastructural information
* Hybrid adoption
* Soil
* Baseline climate
* Perturbed (future) climate

The “clean” files have been purged of outliers, and are main season only. The “pre-clean” files contain data for both main and short seasons, and have not been purged of outliers.

The crop variables exist for all of the crops in [*crop list*], where [*crop list*] is the vector of crops of interest to the analyst. As a default setting, the only crop in [*crop list*] is maize. Note that “maize” combines dry and green maize, which are separate in the raw Tegemeo files, into one crop type. Green maize is equivalized to dry maize by reducing its weight by 15%.

The fertilizer variables exist for all of the fertilizers in [*fert list*], where [*fert list*] is the vector of fertilizers/inputs of interest to the analyst. As a default setting, the fertilizers in [*fert list*] are diammonium phosphate (-dap-), calcium ammonium nitrate (-can-), manure (-man-), and dap+can+man (-dcm-). These are, by far, the most widely used fertilizers in any Tegemeo survey year.

The Stata (v12) pre-clean file generator programs for each year are written so that it is easy to add more fertilizers to [*fert list*] or more crops to [*crop list*].

The crop cost and revenue variables were built up from the crop level files croplev[*yr*] and validated against the farm level variables contained in the farm level files cropinc[*yr*]. The fertilizer variables were built up from the field level files field[*yr*] and validated against farm level variables contained in the farm level files cropinc[*yr*]. The same build up was attempted for the seed and land preparation variables, but failed to validate. Seed and land preparation cost variables were taken directly from the validation files cropinc[*yr*].

Any logged variable *var* is created using a binary dummy variable *dumvar*. I.e.:

log(*var*)=log(*var*+1-*dumvar*)

Variable names generally end with the suffix “\_F” indicating farm level. If the variable ends with a different level identifier suffix such as “–vil,” “-dist,” or “-prov,” then the variable is averaged over the level indicated by the suffix.

Note that no seed cost or labor data exists in ’00, ’97. A labor wage variable can still be obtained for these years by using the “wagerate” variable in the hh[*yr*] file.

The labour[*yr*] files available in ’04, ’07 make it possible to create variables for specific types of labor activity. Taking advantage of this feature, pre- and post-harvest labor variables were created. A “main activity” variable was also created, which includes labor only for those activities where N>400. Variables that include only pre-harvest labor bear the suffix “–PRE.” Variables that include only post-harvest labor bear the suffix “–POST.” Variables that include main activities only bear the suffix “-M.” This is covered in more detail in the section on labor variables.

When devising a model for regression, there are many options for crop cost and livestock revenue and cost. The following suggestions might be helpful to get started:

(For crop revenue models there is one choice: revctot\_F)

For crop cost models:

costctotPRE\_F (if analysis is limited to ’04, ‘07) - Crop total variable cost, including pre-harvest wage labor costs (seed+lp+fert+labhwgPRE)

For livestock revenue models:

revls1\_F - Livestock revenue, including all animal sales plus the value of all livestock production except milk.

For livestock cost models:

Costls1\_F - Livestock cost, including all animal purchases and livestock variable costs (costlspurch+costlsvar1), where livestock variable costs are defined as *feed+vet services+all salaried labor spent on livestock*.

Note that any livestock revenue variables built from the vanprod[*yr*] variable in the vlivestock[*yr*]\_net files can take negative values, because the vanprod[*yr*] variables themselves take negative values. I couldn’t figure out why this is so. Pending clarification I think it is advisable to use livestock revenue variables built from the vprod\_lp variable, which does not take negative values. The livestock revenue variables built from vprod\_lp are revls1\_F and revlsprod1\_F, while those built from vanprod[*yr*] are revls2\_F and revlsprod2\_F.

Note that a milk revenue variable is available only in ’07.

Keep in mind that crop revenue refers to the value of the crop produced, regardless of whether or not it is commercialized. Since most households are net consumers of their crop product, and in many cases in order to meet their needs they must purchase more crop products over and above what they have produced, I don’t think it should come as too big a surprise if the coefficient on maize price comes out negative.

The following errata were found:

* In the ’97, ’00, ’04 Stata files acre values are displayed as strings.
* 9 duplicates exist at the seed type level in the croplev07 file. These were dropped when creating the files presented here.

Level ID Variables

Source files: Most of the ’07 files include these variables. In the survey years prior to ’07 these variables are included only in some of the files. They can always be found in the income[*yr*] file.

|  |  |
| --- | --- |
| **Table 1: ID variables** | |
| **ID designation** | **Definition** |
| Hhid | Household identifier |
| Field | Field identifier |
| Vil | Village identifier |
| Dist | District identifier |
| Province | Province identifier |
| Zone | Kenya agricultural ecological region identifier |
| AEZ | FAO universal agro-ecological zone identifier |

Revenue

Source files: croplev[*yr*], pricecrop, vlivestock[*yr*]\_net

Variable labeling convention: rev[*rev category*][*rev subcategory*]\_F

Eg.: revctot\_F, revlssoldo\_F

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 2: Revenue variables** | | | | | | |
| **Revenue category**  **design-nation** | **Revenue category definition** | **Revenue sub-category designation** | **Source file** | **Source variable(s)** | **Revenue sub-category definition** | **Notes** |
| -c- | Crop revenue | -[*crop list*]-  (-maiz-) | Croplev[*yr*], pricecrop | kgharv, pkg | Revenue from [*crop list*] | Default setting for [*crop list*] is maize. |
| -tot- | Revenue from all crops |  |
| -ls- | Livestock revenue | -1- | vlivestock[*yr*]\_net | Vprod\_lp, vsold\_ls | Livestock revenue (revlsprod1+revlssold, all animals) | Recommended over -2- because vanprod[*yr*] can take negative values. |
| -2- | Vanprod[*yr*], vsold\_ls | Livestock revenue (revlsprod2+revlssold, all animals) | Not a good revenue var since vanprod[*yr*] takes negative values. |
| -3- | vmilkprod,  Vsold\_ls | Livestock revenue (revlsmilk+revlssold, all animals) | Available only in ’07. |
| -prod1- | vprod\_lp | Value of livestock production except milk (all livestock) | Recommended over –prod2- because it is always positive. |
| -prod2- | vanprod[*yr*]\_c, vanprod[*yr*]\_o | Value of livestock production except milk (all livestock) | vanprod[*yr*] takes negative values, so this is not a good choice for a revenue variable. |
| -prodc- | Value of livestock production except milk (cattle only) | Same warning about vanprod[*yr*]. Distinction between cattle and non-cattle livestock available only in ’07. |
| -prodo- | Value of livestock production (non-cattle only) | Same warning about vanprod[*yr*]. Distinction between cattle and non-cattle livestock available only in ’07. |
| -milk- | vmilkprod | Value of milk produced. | Available only in ’07. |
| -sold- | vsold\_ls\_c,  vsold\_ls\_o | Value of livestock sales (all animals) |  |
| -soldc- | Value of livestock sales (cattle only) | Distinction between cattle and non-cattle livestock available only in ’07. |
| -soldo- | Value of livestock sales (non-cattle only) | Distinction between cattle and non-cattle livestock available only in ’07. |
| -c- | Vsold\_ls\_c, vanprod[*yr*]\_c | Livestock revenue, cattle only (revlsprodc+revlssoldc) | Distinction between cattle and non-cattle livestock available only in ’07. |

Cost

Source files: croplev[*yr*], vlivestock[*yr*]\_net, fert[*yr*], pricefert, cropinc[*yr*], field[*yr*], labour[*yr*]

Variable labeling convention: cost[cost category][cost subcategory]\_F

Eg.: costlspurch\_F, costctot\_F

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 3: Cost variables** | | | | | | |
| **Cost category**  **design-nation** | **Cost category definition** | **Cost sub-category designation** | **Source file** | **Source variable(s)** | **Cost sub-category definition** | **Notes** |
| -c- | Crop costs | -tot- | Croplev[*yr*], fert[*yr*], pricefert, cropinc[*yr*], labour[*yr*] | fertcost, sdcost, lpcost, lb01, lb02 | Total cost of crop production, including hired wage (but not contract) labor  (fertilizer cost+seed cost+land prep. cost+wage labor) |  |
| -totnolab- | fertcost, sdcost, lpcost | Total cost of crop production except hired labor  (fertilizer cost+seed cost+land prep. cost) |  |
| -totM- | fertcost, sdcost, lpcost, lb04 | Total cost of crop production, including main activity hired wage labor costs only  (fertilizer cost+seed cost+land prep. cost+main activity wage labor) | “Main activities” are those where the number of farms conducting these activities is > 400.  Available only in ’04 and ’07 |
| -totPRE- | Total cost of crop production, including pre-harvest activity hired wage labor costs only  (fertilizer cost+seed cost+land prep. cost+pre-harvest activity wage labor) | “Pre-harvest activities” are those that occur before the harvest.  Available only in ’04 and ’07. |
| -totPOST- | Total cost of crop production, including post-harvest activity hired wage labor costs only  (fertilizer cost+seed cost+land prep. cost+post-harvest activity wage labor) | “Post-harvest activities” are those occurring during and after the harvest (includes the harvest itself).  Available only in ’04 and ’07 |
| -tot2- | fertcost, sdcost, lpcost, lb04 | Total cost of crop production, including hired contract (but not wage) labor  (fertilizer cost+seed cost+land prep. cost+wage labor) | Available only in ’04 and ’07 |
| -totM2- | Total cost of crop production, including main activity hired contract labor costs only  (fertilizer cost+seed cost+land prep. cost+main activity contract labor) | See definition of “main activity” above  Available only in ’04 and ’07 |
| -totPRE2- | Total cost of crop production, including pre-harvest activity hired contract labor costs only  (fertilizer cost+seed cost+land prep. cost+pre-harvest activity contract labor) | See definition of “pre-harvest activity” above.  Available only in ’04 and ’07 |
| -totPOST2- |  | Total cost of crop production, including post-harvest activity hired contract labor costs only  (fertilizer cost+seed cost+land prep. cost+post-harvest activity contract labor) | See definition of “post-harvest activity” above.  Available only in ’04 and ’07 |
| -tot3- | fertcost, sdcost, lpcost, lb01, lb02, lb04 | Total cost of crop production, including both hired contract and wage labor  (fertilizer cost+seed cost+land prep. cost+wage and contract labor) | Available only in ’04 and ’07 |
| -totM3- | Total cost of crop production, including main activity hired wage and contract labor costs only  (fertilizer cost+seed cost+land prep. cost+main activity wage and contract labor) | See definition of “main activity” above  Available only in ’04 and ’07 |
| -totPRE3- | Total cost of crop production, including pre-harvest activity hired wage and contract labor costs only  (fertilizer cost+seed cost+land prep. cost+pre-harvest activity wage and contract labor) | See definition of “pre-harvest activity” above.  Available only in ’04 and ’07 |
| -totPOST3- | Total cost of crop production, including post-harvest activity hired wage and contract labor costs only  (fertilizer cost+seed cost+land prep. cost+post-harvest activity wage and contract labor) | See definition of “post-harvest activity” above.  Available only in ’04 and ’07 |
| -ls- | Livestock costs | -- | vlivestock[*yr*]\_net | costlspurch, costlsvar | Total cost of livestock production  (all animal purchases+variable cost) |  |
| ’07: -2-  ’97-’04: N/A | costlspurch, costlsvar2 | Total cost of livestock production  (all animal purchases+alternative variable cost) | Available only in ’07. |
| ’07: -c-  ’97-’04: N/A | costlspurchc, costlsvar | Total cost of livestock production  (cattle purchases+variable cost) | Available only in ’07. |
| ’07: -2c-  ’97-’04: N/A | costlspurchc, costlsvar2 | Total cost of livestock production  (cattle purchases+alternative variable cost) | Available only in ’07. |
| -purch- | ’97-’04: vpur\_ls  ’07: vpur\_ls\_c,  vpur\_ls\_o | Livestock purchases, all animals | In ’07 vpur\_ls has suffixes “\_c” for cattle and “\_o” for other livestock. These are summed to produce the costlspurch variable. |
| ’07: -purchc-  ’97-’04: N/A | vpur\_ls\_c | Livestock purchases, cattle only | Available only in ’07. |
| -var- | ’97-’04: vcost\_lv  ’07: vlivecost | Livestock variable cost (feed+vet services +all salaried labor spent on livestock) |  |
| ’07: -var2-  ’97-’04: N/A | ’07: tlivecost  ’97-’04: N/A | Livestock variable cost (feed + vet services + tick control + deworming + artificial insemination + natural insemination +  farm structure maintenance) | Available only in ’07. |
| -f-  -fert- | Fertilizer costs | -[*fert list*]- | Fert[*yr*], pricefert |  |  |  |
| -tot- |  |  |  |
| -sd- | Seed costs | -- | Cropinc[*yr*] | Seedcost |  | It was not possible to validate seed cost when building it up from the crop level. Therefore it was taken directly from the validation file (cropinc[*yr*]). |
| -lp- | Land prep. Costs | -- | Lpcost |  | It was not possible to validate seed cost when building it up from the field level. Therefore it was taken directly from the validation file (cropinc[*yr*]). |

Crop and fertilizer price and quantity variables, per-acre variables (yield and fert/acre)

Source files: croplev[*yr*], pricefert, pricecrop

Quantity variable labeling convention: kg[*item type*][*item*]\_F

Eg.: kgfdap\_F, kgcmaize\_F

Price variable labeling convention: pkg[*item type*][*item*]\_F

Eg.: pkgfdap\_F, pkgcmaize\_F

Fert. per acre variable labeling convention: fertac[*fert type*]\_F

Eg.: fertaccan\_F, fertacdap\_F

Yield variable labeling convention: yield[*crop*]\_F

Eg.: yieldmaiz\_F

All crop and fertilizer quantities are expressed in kgs. All prices are expressed in KES/kg.

The default setting for [*crop list*] is maize (-maiz). The default setting for [*fert list*] is: diammonium phosphate (-dap), calcium ammonium nitrate (-can), manure (-man), and dap+can+man (-dcm). These are, by far, the most widely used fertilizers in any Tegemeo survey year.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 4: Crop and fertilizer price, qty., and per-acre variables** | | | | | | |
| **Variable type** | **Variable type definition** | **Item type designation** | **Item type definition** | **Item** | **Source file** | **Definition** |
| Kg- | Quantity (kgs.) | -f- (or –fert-) | Fertilizer | -[*fert list*]- | Fert[*yr*] | Qty of fertilizer purchased (kgs.) |
| -c- | Crop | -[*crop list*]- | Croplev[*yr*] | Qty of [*crop*] produced (kgs.) |
| Pkg- | Per kilogram price | -f- (or –fert-) | Fertilizer | -[*fert list*]- | Pricefert |  |
| -c- | Crop | -[*crop list*]- | Pricecrop |  |
| Fertac- | Fertilizer per acre | -- | -- | -[*fert list*]- | Fert[yr], field[*yr*] | Fertilizer per acre |
| Yield- | Crop yield | -- | -- | -[*crop list*]- | Croplev[*yr*], field[*yr*] | Crop yield |

Labor

Source file(s): labour[*yr*], hh[*yr*]

Labeling convention: lab[*labor type*][*labor sub-type*]\_[*var type*][*labor activity type*]\_F

Eg.: labfamchilqtyPRE\_F, labhwg\_daywage\_F

The “labour[*yr*]” files exist only for ’04 and ’07. In ’97 and ’00 one can use the wage variable from the hh[*yr*] file, but cannot create any hired or family labor quantity variables.

A “non-paid salaried labor” is available in 2007.

Labor variables are defined for

* Quantity
* Wage
* Cost

The hired labor cost variables are explained under the cost section.

Labor variables are defined for all activities, main activities, pre-harvest activities, and post-harvest activities. These are explained in table 5.

|  |  |  |
| --- | --- | --- |
| **Table 5: Labor activity types (’04 and ’07 only).**  The activities corresponding to each activity code can be found in the original data documentation for the labour[*yr*] files. | | |
| Labor activity type designation | Definition | Activity code |
| (No designation) | All labor activities | (All codes) |
| M | Main labor activities (N>400) | ’04: 4, 5, 9, 11, 12, 14, 22, 23, 25, 26, 29, 35  ’07: 1, 4, 5, 6, 7, 10, 13 |
| PRE | Pre-harvest labor activities | ’04: 1-21, 32-35, 37, 42, 53, 55, 73, 76, 77, 81, 85-87, 90, 92, 100, 106  ’07: 1, 2, 3, 4, 5, 6, 7, 8, 9 |
| POST | Post-harvest activities, including harvest activities | All non pre-harvest activity codes |

The hired wage labor quantity is a straightforward calculation of the number of wage workers times the hours each worked. However, the hired contract labor quantity must be calculated differently—by dividing the contract labor cost (labour[*yr*].dta variable: lb04) by the district level wage. This should be thought of as a proxy variable, not an accurate representation of hired contract labor quantity.

For variables limited to a specific set of activities (-M,-PRE,-POST) the wage is averaged only over these activities.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 6 : Labor variables**  **lab[*labor type*][*labor sub-type*]\_[*var type*][*labor activity type*]\_F**  **Note that *labor activity type* suffixes not displayed here. These suffixes explained in table 5.** | | | | | | | | |
| **category**  **design-nation** | **category definition** | **sub-category designation** | **sub-category definition** | **Variable type designation** | **Source file** | **Source variable(s)** | **Variable definition** | **Notes** |
| -h- | Hired labor | -wg- | Wage labor | -qty- | ’04, ’07: labour[*yr*]  ’97, ’00: N/A | ’04,’07: lb03, lb04  ’97,’00: N/A | Hired labor quantity in man-days | Straightforward calculation of number of wage workers hired (lb01) times the hours each worked (lb02).  Note that units are man-days, whereas family labor qty. expressed in man-hours. |
| -daywage- | ’04, ’07: labour[*yr*]  ’97, ’00: N/A | ’04,’07: lb03  ’97,’00: N/A | Hired labor daywage | This is just lb03. Suffix –dist-, -vil-,-prov-, etc. indicates the wage is averaged over the level indicated by the suffix. |
| -daywage2- | hh[*yr*] | wagerate | Hired labor daywage (alternative) | This is the only wage variable available for all years. |
| -con- | Contract labor | -qty- | ’04, ’07: lb03, lb04  ’97, ’00: N/A |  | Hired contract labor qty. in man-hours | Number hired and hrs. worked vars not available for contract labor, so this was calculated by dividing the contract labor cost (lb04) by district level wage (lb03 averaged over district level). Do not assume this is very accurate. |
| -fam- | Family labor | -chil- | Child family labor | -qty- | ’04, ’07: lb09, lb10 labour[*yr*]  ’97, ’00: N/A |  | Child family labor qty. in man-hours | Number hired (lb09) times hours worked (lb10). Note that units are man-hours, whereas hired labor qty. expressed in man-days. |
| -fem- | Adult family labor | -qty- | ’04, ’07: lb07, lb08 labour[*yr*]  ’97, ’00: N/A |  | Adult female family labor qty. in man-hours | Number hired (lb07) times hours worked (lb08). Note that units are man-hours, whereas hired labor qty. expressed in man-days. |
| -male- | Adult male family labor | -qty- | ’04, ’07: lb05, lb06 labour[*yr*]  ’97, ’00: N/A |  | Adult male family labor qty. in man-hours | Number hired (lb05) times hours worked (lb06). Note that units are man-hours, whereas hired labor qty. expressed in man-days. |

Farm demographic and infrastructure variables

Source files:

’07: demoga07, demog07

’04: demogA\_A04, demogA04

’97, ’00: demog[*yr*]

In ’04 and ’07, the age, education, and gender demographic variables are created from two files. One file contains the demographic data for household members included in the previous survey, while the other file contains the data for new household members. In all survey years, Household members born before 1989 are not counted.

Note that in the source files education is measured differently from year to year. In ’97 and ’00 it is simply the number of years of schooling. In the ’04 and ‘07 source files new systems are introduced which codify higher education levels. In ’04, codes 0-14 cover basic schooling, where “0” means no schooling. Codes 20-24 cover higher education. In the present data set I remove the gap so that the range is 0 to 19.

In ’07, codes 0-8 cover “standard” education, where “0” means pre-school (“-9” means no schooling). Codes 9-14 cover “formal” education, codes 15-18 cover “college,” and codes 19-23 cover “university.” I modify this so that “0” means no schooling, and “1” means preschool, etc. (all other codes are bumped up by 1).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 7: Farm demographic and infrastructure variables** | | | | |
| Variable name | Definition | Source file | Source variable | Notes |
| Aehhsize | Adult equivalized number of members in the household. | ae\_hhsize\_[*yr*] | aehh[*yr*] | Children or adults present for less than half the year counted as fractions of adults. |
| Offrinc | Off farm income | Income[*yr*] | ’07: offrinc07  ’97-’04: offrin[*yr*] |  |
| Acres | Acres under cultivation | Field[*yr*] | Acres |  |
| Gend | Gender of household head is female if gend=1, male if gend=0 | ’07: demoga07, demog07  ’04: demogA\_A04, demogA04  ’97, ’00:demog[*yr*] | Ad02, da02 |  |
| Age | Age of household head | ’07: Age  ’04: ad01, da01  ’97, ’00: da01 | Includes only members born after 1989 |
| Ageavg | Average age of members in household | Includes only members born after 1989 |
| Educ | Years of education of household head | ’04, ’07: ad06, da06  ’97, ’00: da06 | Coding is different from year to year. |
| educavg | Average years of education of members in household |  |
| Dtmroad | Distance to nearest tarmac road | Hh[*yr*] | Dtmroad |  |
| Dmtroad | Distance to nearest motorable road | Dmtroad |  |
| Dextn | Distance to nearest extension advice | Dextn |  |

Soil

Source file: soil

Soil variables are the same for all survey years.

Soil variable: “Drainage”

|  |  |
| --- | --- |
| **Table 8: Soil Drainage** | |
| 1 | Very poorly drained |
| 2 | Poorly drained |
| 3 | Imperfectly to poorly drained |
| 4 | Imperfectly drained |
| 5 | Mod (sic) to imperfectly drained |
| 6 | Moderately/variable (sic) drained |
| 7 | Well to moderately drained |
| 8 | Well drained |
| 9 | Somewhat excessively to well drained |
| 10 | Somewhat excessively drained |
| 11 | Excessively drained |

|  |  |
| --- | --- |
| **Table 9: Soil Depth** | |
| 1 | Shallow |
| 2 | Shallow-mode (sic) |
| 3 | Mod/dp-varia (sic) |
| 4 | Mod deep-deep |
| 5 | Deep |
| 6 | Deep-very deep |
| 7 | Very deep |
| 8 | Extremely deep |

Hybrid adoption variables

Source files: croplev[*yr*], fert[*yr*], field[*yr*]

Hybrid adoption variables are included for hybrid crops in general, hybrid maize, and hybrid maize plus a minimum fertilizer condition. The Kenyan Agricultural research Institution (KARI) recommends a minimum application of 25 kg./acre CAN, 25 kg./acre DAP, and 4000 kg./acre of manure for optimum maize yield. The minimum fertilizer condition is set at 15 kg./acre of some combination of DAP and CAN.

|  |  |
| --- | --- |
| **Table 10: Hybrid adoption variables** | |
| **Variable name** | **Variable definition** |
| Hyb\_F | Purchased hybrid seed used for at least 1 crop (0/1) |
| Hybmz\_F | Purchased hybrid maize seed used (0/1) |
| Hybfmz\_F | Purchased hybrid maize seed used and CAN+DAP fert/acre>14kg (0/1) |

Baseline climate variables

The baseline climate variables were produced by Texas A&M’s ACT project, version 2.01, based on 30-50 years of historical climate data (see the Tegemeo 2007 original data documentation), and made available in the Tegemeo file “climate.dta.” Note that the weblink to the ACT project provided in the Tegemeo documentation is dead, and a google search failed to locate the project. The climate variable suffixes “–xt,” “–it,” and “–pre,” indicate maximum temperature, minimum temperature, and rainfall, respectively.

ACT generated wet season quarterly averages, five month optimum growing season averages, and trigger season averages. These are explained briefly:

* The wet season quarterly averages (qwet-) are averaged over the rainiest three consecutive months of the year. The starting months of these three month periods vary from district to district and are displayed in table 11.

|  |  |  |
| --- | --- | --- |
| Table 11: First month of the long rains season by district | | |
| Province | District(s) | First month of the long rains season |
| Coast | Taita-Taveta | Mar |
| Eastern | Makueni, Meru | Mar |
| Mwingi, Machakos | Oct |
| Kitui | Nov |
| Nyanza | Kisii, Siaya, Kisumu | Mar |
| Western | Bungoma | Apr |
| Kakamega | Jun |
| Vihiga | Mar |
| Central | Muranga, Nyeri | Mar |
| Rift Valley | Narok, Bomet | Mar |
| Trans Nzoia, Uasin Gishu, Nakuru | Apr |
| North Rift | Laikipia | Mar |
| Source: (Tegemeo 2007) | | |

* The five month optimum growing season averages (OPT5M-) are averaged over the five consecutive months of the year that maximize the rainfall to evapotranspiration ratio (P/PE).
* The trigger season averages (T05S1-) are averaged over the longest run of consecutive months in the year where P/PE 0.5, and thus vary in main season length.
* The prefix for annual climate variables is ann-.

The authors of the Tegemeo data documentation remark that the five month optimum averages fit the main growing seasons of arid regions well, while the trigger season averages are a better fit for the wetter regions (Tegemeo 2007). The definitions of “main season” generated by these three different methods coincide for most districts, but not all. See the 2007 Tegemeo original data documentation for further details. Note that econometric results are sensitive to which of the three above main season definitions is chosen in model specification.

Additional main and short season total rainfall data was compiled for each year of the survey by Tavneet Suri in 2007 from raw rain station and satellite data available from the National Weather Service Climate Prediction Center (as a part of the Famine Early Warning System Project). This variable is “main07.” The main and short seasons in this case were defined by Thomas Jayne at the village level and are set forth in the data documentation (Tegemeo 2007). Again, for most districts Jayne’s definitions of main season coincide with those of the other three methods outlined above. The important difference here is that whereas the ACT rainfall data serves as a generic baseline extracted from decades of data, the rainfall data compiled by Suri is the actual historical data for each particular year of the survey.

In the thesis, the main07 rainfall variable is used with the qwetxt and qwetit max and min temperature variables.

Perturbed (future) climate variables – 2050s

Source file: climatePCB

The perturbed climate file climatePCB was created by the author. Raw Global Circulation Model (GCM) data was obtained from CGIAR’s CCAFS climate data portal. The data was obtained for several GCM projections out to the 2050s, under the A2 and B2 SRES emissions scenarios, at a downscaled resolution of 5 (.0833 of a degree) and 10 arcminutes (0.166 of a degree), and matched to each observation in the 2007 Tegemeo farm sample. Since the farm sample is geo-referenced at a resolution of 0.10 of a degree, the coordinates of each farm observation were rounded to the nearest .0833 (if doing a 5 arcminute analysis) or 0.166 (if doing a 10 arcminute analysis) of a degree in order to achieve a match with the CCAFS data. Note that the 5 arcminute data is a higher resolution than the farm data, and would thus introduce a degree of spurious accuracy.

The climatePCB file does not include data from each individual GCM, but rather averages the data from various GCMs. Such averages are called GCM ensembles. Four ensembles were created for the climatePCB file based on different rationales. These are summarized in table 12.

|  |  |  |
| --- | --- | --- |
| **Table 12: GCM ensembles available in climatePCB file** | | |
| **Ensemble designation** | **GCMs** | **Notes (rationale)** |
| -E1- | cccma\_cgcm3\_1\_t63  mpi\_echam5  csiro\_mk3\_0  miub\_echo\_g  hccpr\_hadcm3 | The first four GCMs in the list are recommended for East Africa analysis by Washington and Pearce (2012). The last (hccpr\_hadcm3) is recommended for East Africa by Thornton (2010). At the time of writing, the other three GCMs recommended by Washington and Pearce are not yet available in downscaled form. |
| -E2- | bccr\_bcm2\_0  mpi\_echam5  iap\_fgoals1\_0\_g  hccpr\_hadcm3  cccma\_cgcm2  csiro\_mk2  nies99  miub\_echo\_g | These are the eight models that are currently available from CCAFS for the SRES A2 scenario at both 5 and 10 arcminute resolution. |
| -E3- | csiro\_mk2  cccma\_cgcm2  mpi\_echam5  hccpr\_hadcm3  ncar\_pcm1 | This is the ensemble used in Kabubo’s study of climate change and livestock in Kenya (Kabubo 2009) |
| -E4- | hccpr\_hadcm3  cccma\_cgcm2  nies99  csiro\_mk2 | These are the four models that are currently available from CCAFS for the SRES B2 scenario at both 5 and 10 arcminute resolution. |

The perturbed climate variable labeling convention is as follows:

[*Climate variable type*]PC [*downscaled resolution*][*emissions scenario*][*clim*. *variable*][*ensemble*]

Where:

*Climate variable type* is the baseline climate variable type (qwet, OPT5M, T05S1, or ann).

*Downscaled resolution* is either 5 or 10 arcminutes (5 or 10).

*Emissions scenario* is the SRES emissions scenario (a2a or b2a).

*Clim*. *Variable* refers to whether the variable is rainfall or max or min temperature (pre, xt, it).

*Ensemble* is the downscaled GCM ensemble over which the variable is averaged (see table 11).

For example, the variable “qwetPC10a2axtE4” means the A2 emissions scenario perturbed climate main season max temperature projected by the GCMs in ensemble 4, downscaled to 10 arcminutes, and averaged over the ensemble—where “main season” is defined as the wet season quarterly average (qwet-). This nomenclature mixes the Tegemeo convention of designating baseline climate variables with the CCAFS file naming convention.

The individual GCM future climate projections included in the ensembles above, plus others not included in these ensembles, are available in the following general compilation files (also created by the author):

CCAFS Downscaled GCM Data Matched to Tegemeo Observations\_sres\_a2a\_5min.xlsx

CCAFS Downscaled GCM Data Matched to Tegemeo Observations\_sres\_a2a\_10min.xlsx

CCAFS Downscaled GCM Data Matched to Tegemeo Observations\_sres\_b2a\_5min.xlsx

CCAFS Downscaled GCM Data Matched to Tegemeo Observations\_sres\_b2a\_10min.xlsx

Each of these compilation files contains perturbed climate data (geo-referenced to the Tegemeo farm data) for a particular downscaled resolution and SRES scenario combination, generated by each of the GCMs that are available (from the CCAFS data portal) for that combination. The resolution-SRES combination is indicated in the file extensions—eg. “\_sres\_a2a\_10min.” This data is summarized in table 12:

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 12: GCM future climate projections available in the general source files** | | | |
|  |  | **SRES scenario** | |
| **A2** | **B2** |
| **Downscaled resolution** | **10 arcminute** | File name: “CCAFS Downscaled GCM Data Matched to Tegemeo Observations\_sres\_a2a\_10min.xlsx”  GCMs:  hccpr\_hadcm3  cccma\_cgcm2  cccma\_cgcm3\_1\_t63  nies99  bccr\_bcm2\_0  mpi\_echam5  iap\_fgoals1\_0\_g  csiro\_mk3\_0  csiro\_mk2  ncar\_pcm1  miub\_echo\_g | File name: “CCAFS Downscaled GCM Data Matched to Tegemeo Observations\_sres\_b2a\_10min.xlsx”  GCMs:  hccpr\_hadcm3  cccma\_cgcm2  nies99  csiro\_mk2 |
| **5 arcminute** | File name: “CCAFS Downscaled GCM Data Matched to Tegemeo Observations\_sres\_a2a\_5min.xlsx”  GCMs:  bccr\_bcm2\_0  mpi\_echam5  iap\_fgoals1\_0\_g  hccpr\_hadcm3  cccma\_cgcm2  csiro\_mk2  nies99  miub\_echo\_g | File name: “CCAFS Downscaled GCM Data Matched to Tegemeo Observations\_sres\_b2a\_5min.xlsx”  GCMs:  hccpr\_hadcm3  cccma\_cgcm2  nies99  csiro\_mk2 |

References

Tegemeo Institute of Agricultural Policy and Development, Egerton University, MichiganState University. 2007. “Tegemeo Agricultural Policy Research Analysis (TAPRA) Project: Household survey 2007 data documentation.”

Thornton, P.K., Jones, P.G., Alagarswamy, G., Andresen, J., Herrero, M. 2010. “Adapting to climate change: Agricultural system and household impacts in East Africa.” Agricultural

Systems 103, 73-82.

Washington, Richard, Pearce, Helen. 2012. “Climate Change in East African Agriculture: Recent Trends, Current Projections, Crop-climate Suitability, and Prospects for Improved Climate Model Information.” CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS). Copenhagen, Denmark. Available online at: [www.ccafs.cgiar.org](http://www.ccafs.cgiar.org)