Code Appendix

Contents

1 Pre-processing reference cropland data layer

The following describes the initial pre-processing of landcover datasets that was undertaken for an analysis of error in landcover data.

1.1 Gridding the SA landcover data

Much of the original analysis was undertaken on the Mapping Africa server, using postgis/postgres as well as pprepair to do much of the cleaning and conversion of the GTI (reference) dataset into a 1 km resolution cropland percentage cover dataset. The following steps were taken:

1.1.1 Cleaning up fields in GTI data

This was done for both the 2007 and 2011 datasets, using a script that pulled out the fields, found the bad ones, read them into an R data object, ran pprepair over it, and then stuck the cleaned geometries back into a postgis table holding the full dataset. The intersection was then performed.

```r
source("R_ext/grid-fields-intersect-postgis.R")  # run on mapper.princeton.edu
```

1.1.2 Calculating field areas

The next step was running the following script to calculate the area for different cropcover classes in each 1 km^2 pixel. This created two core gridded datasets, cover2007.tif and cover2011.tif.

```r
source("R_ext/grid-fields-intersect-analyze.R")  # run on mapper.princeton.edu
```

2 Pre-processing of landcover datasets

Fetching and pre-processing of global landcover datasets to be analyzed relative to the SA dataset: the MODIS 2011 landcover product for South Africa was downloaded, the GLC-SHARE dataset, the 2009 GlobCover product, South Africa's 2009 landcover product, and geo-wiki's cropland fusion map. These were converted to 1 km^2 percentage cropland estimates after making several versions of MODIS and GlobCover products to account for mixed cropland classes. A further processing was applied to the two GTI derived cover images.

2.1 Analyses

2.1.1 Set-up

```
library(SAcropland)
library(gdalUtils)
library(RCurl)
library(RPostgreSQL)
library(rgdal)
library(rgeos)
library(raster)
library(lmisc)
```

Paths

```
#setwd("/u/lestes/analyses/SAcropland/")
p_root <- proj_root("SAcropland")
p_fig <- full_path(p_root, "SAcropland/external/figures/")
p_data <- full_path(p_root, "SAcropland/external/ext_data/")
```

Background data, projections, database connections, and SA grid

```
# Get SA shape
if(!file.exists(full_path(p_data, "MZshapes.Rdata"))) {
  system("wget http://dl.dropboxusercontent.com/u/31890709/MZshapes.Rdata")
}
load(full_path(p_data, "MZshapes.Rdata"))
sa_buf <- gBuffer(sa.shp, width = 3000)

# Connection
drv <- dbDriver("PostgreSQL")
con <- dbConnect(drv, dbname = "SouthAfricaSandbox", user = "postgis",
                 password = "P0stG1S")

# projections, etc
prjsrid <- 97490  # Albers
prj_sql <- paste0("select proj4text from spatial_ref_sys where srid=", prjsrid)
prjstr <- dbGetQuery(con, prj_sql)$proj4text
gcsid <- 4326  # GCS
gcs_sql <- paste0("select proj4text from spatial_ref_sys where srid=", gcsid)
gcsstr <- dbGetQuery(con, gcs_sql)$proj4text

# Convert SA to GCS for cropping extent
```

```
sa_buf_gcs <- spTransform(sa_buf, CRSobj = CRS(gcsstr))

# Create 1 km grid from sa1kilo to rectify grids against that
if(!file.exists(full_path(p_data, "sagrid.tif"))) {  # already made?
  sql <- paste("SELECT gid, id, ST_AsText(ST_Centroid(geom)) as center",
               "FROM sa1kilo")
  geom.tab <- dbGetQuery(con, sql)
  coord.mat <- do.call(rbind, lapply(1:nrow(geom.tab), function(y) {
    strip <- gsub("POINT|\\(|\\)", "", geom.tab[y, 3])
    coord.mat <- do.call(rbind, strsplit(strsplit(strip, ",")[[1]], " "))
  }))
  class(coord.mat) <- "numeric"
  point.tab <- cbind(geom.tab[, 1:2], coord.mat)
  colnames(point.tab)[3:4] <- c("x", "y")
  pointsXYZ <- point.tab[, c(3:4, 2)]
  sa_r <- rasterFromXYZ(pointsXYZ)
  projection(sa_r) <- sa.buf@proj4string
  sa_r <- writeRaster(sa_r, filename = full_path(p_data, "sagrid.tif"),
                      overwrite = TRUE)
} else {
  sa_r <- raster(full_path(p_data, "sagrid.tif"))
}
```

2.1.2 Bring in landcover datasets

Let's start with the improved/fusion datasets, GLC share and geo-wiki's

```
if(!file.exists(full_path(p_data, "glcsa_masked.tif"))) {  # see if file exists,
  # to avoid redoing step
  path <- "/u/lestes/spatial_data/glc_share/"
  url <- "http://www.fao.org/geonetwork/srv/en/resources.get?id=47948&fname=GlcShare_v10_02.zip&acces=p
  download.file(url, method = "auto", destfile = "glc_share.zip")
  unzip("glc_share.zip", exdir = "glc_share")
  glc <- raster(full_path(path, "glc_shv10_02.Tif"))
  projection(glc) <- gcsstr
  glcSA <- crop(glc, y=sa_buf_gcs, file = full_path(path, "glcSA.tif"),
                overwrite = TRUE)  # crop to SA
  gdalwarp(srcfile = glcSA@file@name,
           dstfile = full_path(path, "glcSA_alb.tif"), t_srs = prjstr,
           tr = c(1000, 1000), r = "bilinear")
  glcSA_alb <- raster(full_path(path, "glcSA_alb.tif"))

  # Warp to SA grid
  glcsa <- resample(glcSA_alb, sa_r, method = "bilinear",
                    filename = full_path(path, "glcSA_alb_rect.tif"))
  glcsa_m <- mask(glcsa, sa_r, file = full_path(p_data, "glcsa_masked.tif"),
                  overwrite = TRUE)
} else {
  glcsa_m <- raster(full_path(p_data, "glcsa_masked.tif"))
}

if(!file.exists(full_path(p_data, "geowikisa_masked.tif"))) {  # see if file exists, to avoid redoing s
   # first downloaded manually from password protected geo-wiki site
  path <- "/u/lestes/spatial_data/geowiki/"
  unzip("cropland_hybrid_14052014v8.zip", exdir= "geowiki")
  geowiki <- raster(full_path(path, "Hybrid_14052014V8.img"))
  projection(geowiki) <- gcsstr
  geowikiSA <- crop(geowiki, y=sa_buf_gcs,
                    file = full_path(path, "geowikiSA.tif"))  # crop to SA
  gdalwarp(srcfile = geowikiSA@file@name,
           dstfile = full_path(path, "geowikiSA_alb.tif"), t_srs = prjstr,
           tr = c(1000, 1000), r = "bilinear")
  geowikisa_alb <- raster(full_path(path, "geowikiSA_alb.tif"))
  geowikisa <- resample(geowikisa_alb, sa_r, method = "bilinear", # Warp to SA
                        filename = full_path(path, "geowikisa_alb_rect.tif"))
  geowikisa_m <- mask(glcsa, sa_r,
                      file = full_path(p_data, "geowikisa_masked.tif"),
                      overwrite = TRUE)
} else {
  geowikisa_m <- raster(full_path(p_data, "geowikisa_masked.tif"))
}
```

And the standard products, Globcover 2009 and MODIS

```
path <- "/u/lestes/spatial_data/globcover_2009/"
if(!file.exists(full_path(path, "globSA_alb.tif"))) {
  url <- "http://due.esrin.esa.int/globcover/LandCover2009/Globcover2009_V2.3_Global_.zip"
  download.file(url, method="auto", destfile = "globcover_2009.zip")
  unzip("globcover_2009.zip", exdir= "globcover_2009")
  glob <- raster(full_path(path, "GLOBCOVER_L4_200901_200912_V2.3.tif"))
  globSA <- crop(glob, y=sa_buf_gcs)  # crop to extent of SA
  gdalwarp(srcfile = globSA@file@name,
           dstfile = full_path(path, "globSA_alb.tif"), t_srs = prjstr,
           tr = c(300, 300))
  globSA_alb <- raster(full_path(path, "globSA_alb.tif"))
} else {
  globSA_alb <- raster(full_path(path, "globSA_alb.tif"))
}

# MODIS landcover
moddir <- "/u/lestes/spatial_data/MCD12Q1/"
if(!file.exists(full_path(moddir, "modis_type_1_alb.tif"))) {
  tiles <- expand.grid("h" = c(19, 20), "v" = c(11, 12))
  tiles <- apply(tiles, 1, function(x) paste("h", x[1], "v", x[2], sep = ""))
  url <- "http://e4ftl01.cr.usgs.gov/MOTA/MCD12Q1.051/2011.01.01/"
  tile.names <- getURL(url, verbose=TRUE, dirlistonly = TRUE)
  tile.names1 <- strsplit(tile.names, "alt")[[1]]
  tile.names2 <- unique(substr(tile.names1, 20, 64))
  modnames <- tile.names2[sapply(tiles, function(x) grep(x, tile.names2))]
  lapply(modnames, function(x) {
    url <- paste0("http://e4ftl01.cr.usgs.gov/MOTA/MCD12Q1.051/2011.01.01/", x)
    download.file(url, method="auto", destfile = x)
  })
  dir.create("MCD12Q1")
  file.copy(from = dir(pattern = "hdf"),
            to = paste(getwd(), "MCD12Q1", dir(pattern = "hdf"), sep = "/"))
  file.remove(dir(pattern = "hdf"))

  modis <- dir(moddir)

  # Translate to geotiff
  lapply(1:5, function(x) {
    batch_gdal_translate(infiles = modis, outdir = moddir,
                         outsuffix = paste("_LC_", x, ".tif", sep = ""),
                         sd_index = x)
  })
  sdnames <- lapply(1:5, function(x) dir(moddir,
                                         pattern = paste0("LC_", x, ".tif")))

  # warp to albers, mosaic, and clip to SA extent
  lapply(1:5, function(x) {
    print(paste("processing type", x))
    gdalwarp(srcfile = sdnames[[x]],
             dstfile = paste0("modis_type_", x, "_alb.tif"),
             t_srs = prjstr, tr = res(r), te = bbox(sa.buf)[1:4])
  })
  mod_1 <- raster("modis_type1_alb.tif")
} else {
  mod_1 <- raster(full_path(moddir, "modis_type_1_alb.tif"))
}
```

And then South Africa's landcover database (30 m)

```
path <- "/u/lestes/spatial_data/sa_lc_2009/"
setwd("/u/lestes/spatial_data/")
lcnm <- c("sa_ag.tif", "sa_af.tif")
if(length(dir(p_data, pattern = "sa_.*._masked.tif")) != 2) {
  url <- "http://planet.uwc.ac.za/BGISdownloads/landcover_2009.zip"
  download.file(url, method="auto", destfile = "sa_landcover_2009.zip")
  unzip("landcover_2009.zip", exdir = "sa_lc_2009")

  setwd(path)
  gdal_translate(src_dataset = "landcover", dst_dataset = "sa_lc_2009.tif",
                 of = "GTiff", ot = "BYTE")
  #system(paste("rm -r", full_path(path, "landcover")))
  #file.remove("../landcover_2009.zip")

  # process masks--extract extract landcover class 2, cultivation, and class 6,
  # plantations, resample, mask with gdal tools
  av <- c(2, 6)
  a <- "sa_lc_2009.tif"
  dang <- Sys.time()
  salc_l <- lapply(1:2, function(j) {
    print(paste("extracting cover for", lcnm[j]))
    gdal_calc(cstr = paste("A==", av[j], sep = ""), x = list("A" = a),
              type = "Byte", filename = lcnm[j])
    nm <- full_path(p_data,
                    paste(gsub("\\.tif", "", lcnm[j]), "_masked.tif", sep = ""))
    ext <- sapply(c("xmin", "ymin", "xmax", "ymax"), function(x) {
      slot(extent(sa_r), x)
    })
    print(paste("warping and masking", gsub(p_data, "", nm)))
    gdalwarp(srcfile = lcnm[j], t_srs = projection(sa_r), dstfile = nm,
             r = "average", ot = "Float32", te = ext, srcnodata = 255,
             dstnodata = 255, tr = c(1000, 1000), of = "GTiff", verbose = TRUE,
             overwrite = TRUE)
    file.remove(lcnm[j])
    r <- raster(nm)
  })
  Sys.time() - dang

} else {
  lcmnm <- paste(gsub("\\.tif", "", lcnm), "_masked.tif", sep = "")
  salc_l <- lapply(lcmnm, function(x) raster(full_path(p_data, x)))
}
setwd(p_data)
```

2.1.3 Mask out sugarcance

```
if(!file.exists(full_path(p_data, "kzn_cane_masked.tif"))) {
  path <- "/u/lestes/spatial_data/kzn_landcover/"
  setwd(kznpath)
  gdal_translate(src_dataset = "kznlc11v1w31", dst_dataset= "kznlandcover.tif",
                 of = "GTiff")
  system(paste("rm -r", full_path(path, "kznlc11v1w31")))
  kznlc <- raster(full_path(npath, "kznlandcover.tif"))

  # system call to gdal_calc.py to extract landcover class 2, cultivation
  kznlconm <- "kzn_cane.tif"  # file.remove(kznlconm)
  ss <- strsplit(kznlc@file@name, "/")
  kzninnm <- ss[[1]][length(ss[[1]])]
  # kzninnm <- strsplit(kznlc.ss@file@name, "/")[[1]][length(strsplit(kznlc.ss@file@name, "/")[[1]])]
  pcalc <- paste("gdal_calc.py -A ", kzninnm, " --outfile=", kznlconm,
                 " --overwrite --calc='(A==9)|(A==10)'", sep = "")
  system.time(system(pcalc))  # 49 seconds
  kzn_cane <- raster(full_path(path, kznlconm))
  # cext <- extent(c(20000, 40000, -3220000, -3200000))  # crops to small extent for examination
  # kznlc.ss <- crop(kznlc, cext, filename = full_path(kznpath, "kznlandcover_crop.tif"), overwrite = T
  # plot(kznlc.ss)
  # scane1 <- (kznlc.ss == 9); scane2 <- (kznlc.ss == 10)
  # plot(scane1, add = TRUE, col = c("transparent", "red"))
  # plot(scane2, add = TRUE, col = c("transparent", "purple"))
  # chk <- crop(kzn_cane, cext)
  # plot(chk); plot(scane1, add = TRUE, col = c("transparent", "red"), legend = FALSE)
  # plot(scane2, add = TRUE, col = c("transparent", "red"), legend = FALSE)
  # file.remove("kznlandcover_crop.tif")
  # rm(chk, kznlc.ss, scane1, scane2)
  file.remove("kznlandcover.tif")

  # Warp and resample to SA grid
  setwd(p_data)
  gdalwarp(srcfile = kzn_cane@file@name, t_srs = projection(sa_r),
           dstfile = "kzn_cane_1km.tif", r = "average", ot = "Float32",
           srcnodata = 255, tr = c(1000, 1000), of = "GTiff",
           verbose = TRUE, overwrite = TRUE)
  kzn_1km <- raster(full_path(p_data, "kzn_cane_1km.tif"))
  kzn_1km_r <- resample(kzn_1km, sa_r, method = "bilinear")
  kzn_1km_m <- mask(kzn_1km_r, sa_r,
                    filename = full_path(p_data, "kzn_cane_masked.tif"),
                    overwrite = TRUE)
} else {
  kzn_1km_m <- raster(full_path(p_data, "kzn_cane_masked.tif"))
}
```

2.1.4 MODIS cropland variants

Different variants of cropland fractions from MODIS classes 12 and 14 (from the IGBP-DIS scheme). Class 14 is a cropland mosaic class, with minimum of 10% and a max of 60%, so create variants of 10%, 35% (mean) and 60%.

```
if(any(!file.exists(full_path(p_data,
                              dir(p_data,
                                  pattern = "mod_1*.*1kmrect_sum.tif"))))) {
  # MODIS, using main classification scheme (IGBP-DIS)
  mod_1_crop <- stack(mod_1 == 12, mod_1 == 14)
  mod_1_cropb <- brick(mod_1_crop)
  names(mod_1_crop) <- c("cropland", "mosaic")

  # assign percentages
  minset <- c(100, 10)  # assume 10% is min crop cover for MODIS mosaic class
  meanset <- c(100, 35)  # assume 35% is mean crop cover
  maxset <- c(100, 60)  # assume 60% is mean crop cover
  modsa_min <- assignLCPct(mod_1_cropb, minset,
                           fname = full_path(p_data, "mod_1_min.tif"))
  modsa_mu <- assignLCPct(mod_1_cropb, meanset,
                          fname = full_path(p_data, "mod_1_mu.tif"))
  modsa_max <- assignLCPct(mod_1_cropb, maxset,
                           fname = full_path(p_data, "mod_1_max.tif"))

  # aggregate, resample, mask to SA grid
  modsa_list <- lapply(list(modsa_min, modsa_mu, modsa_max), function(x) {
    nm <- gsub("//.tif", "", x@file@name)
    print(paste("Processing", nm))
    print("...aggregating")
    agg <- aggregate(x, fact = 2, na.rm = TRUE)
    print("...resampling")
    crop1km <- resample(agg, sa_r, method = "bilinear")
    print("...masking")
    crop1km <- mask(crop1km, mask = sa_r, filename = paste0(nm, "_1kmrect.tif"),
                    overwrite = TRUE)
    return(crop1km)
  })
  # And them create single raster with summed proportions
  modsa_list_sum <- lapply(modsa_list, function(x) {
    nm <- paste(gsub("\\.tif", "", x@file@name), "_sum.tif", sep = "")
    paste(nm)
    calc(x, sum, filename = nm, overwrite = TRUE)
  })
} else {
  modsa_list_sum <- lapply(full_path(p_data, dir(p_data, "mod_1*.*1kmrect_sum.tif")[c(2:3, 1)]), raster
}
setwd(p_data)
```

2.1.5 GlobCover cropland variants

And the same for GlobCover. We want Globcover class 11 (irrigated cropland), class 14 (rainfed), 20 (50-70% cropland), and 30 (20-50% cropland). For the latter two classes we create variants assumung the minimum, mean, and max of earh class

```
if(any(!file.exists(full_path(p_data, dir(p_data, "globSA*.*1kmrect_sum.tif"))))) {
  l <- lapply(c(11, 14, 20, 30), function(x) r <- globSA_alb == x)  # pull out classes
  s <- stack(l)
  globSA_cropb <- brick(full_path(p_data, "globSA_crop.grd"))
  rm(s, l)

  # assign percentages
  minset <- c(100, 100, 50, 20)
  maxset <- c(100, 100, 70, 50)
  meanset <- c(100, 100, 60, 35)
  globsa_min <- assignLCPct(globSA_cropb, minset, fname = full_path(p_data, "globSA_300_min.tif"))
  globsa_mu <- assignLCPct(globSA_cropb, meanset, fname = full_path(p_data, "globSA_300_mu.tif"))
  globsa_max <- assignLCPct(globSA_cropb, maxset, fname = full_path(p_data, "globSA_300_max.tif"))
  glob300list <- lapply(dir(pattern = "globSA_300+.+tif")[c(2:3, 1)], raster)

  # Aggregate them and resample and mask to SA grid
  globsa_list <- lapply(list(globsa_min, globsa_mu, globsa_max), function(x) {
    nm <- gsub("_300|.tif", "", x@file@name)
    print(paste("Processing", nm))
    print("...aggregating")
    agg <- aggregate(x, fact = 3, na.rm = TRUE)  # equivalent to summing, dividing by 900
    print("...resampling")
    crop1km <- resample(agg, sa_r, method = "bilinear")
    print("...masking")
    crop1km <- mask(crop1km, mask = sa_r, filename = paste(nm, "_1kmrect.tif", sep = ""),
                    overwrite = TRUE)
    return(crop1km)
  })

  # And them create single raster with summed proportions
  globsa_list_sum <- lapply(globsa_list, function(x) {
    nm <- paste(gsub("\\.tif", "", x@file@name), "_sum.tif", sep = "")
    paste(nm)
    calc(x, sum, filename = nm)
  })
} else {
  globsa_list_sum <- lapply(full_path(p_data, dir(p_data, "globSA*.*1kmrect_sum.tif")[c(2:3, 1)]), raste
}
```

2.1.6 Further masking of reference maps

```
# Bring in SA landcover sets and mask
cover2007 <- brick(full_path(p_data, "cover2007.tif"))
cover2011 <- brick(full_path(p_data, "cover2011.tif"))

# sum these also with and without horticulture
sumfun <- function(x) sum(x, na.rm = TRUE)
gti <- lapply(list(cover2007, cover2011), function(y) {
  r <- calc(y[[-6]], sumfun)
  r[r > 100] <- 100
  writeRaster(r, file = gsub("\\.tif", "sum.tif", y@file@name), overwrite = TRUE)
})

# Leave commented out in case want to bring back horticulture class
# gti_hort <- lapply(list(cover2007, cover2011), function(y) {
#   r <- calc(y, sumfun)
#   r[r > 100] <- 100
#   writeRaster(r, file = gsub("\\.tif", "sum_h.tif", y@file@name), overwrite = TRUE)
# })

# Remove communal farmlands
sust_mask <- !is.na(cover2007[[2]]) | !is.na(cover2011[[2]])  # remove communal areas from both dates
sust_mask[sust_mask > 0] <- NA

# Set up SA mask including sugarcane and forestry -- add in Mpumalanga sugar cane when it becomes avail
m1 <- kzn_1km_m
m1[is.na(m1)] <- 0
m1[m1 > 0] <- NA
m2 <- salc_l[[2]]
m2[is.na(m2)] <- 0
m2[m2 > 0] <- NA
sa_masks <- sa_r > 0
sa_masks <- m1 + m2 + sa_masks
rm(m1, m2)
full_mask <- writeRaster(sa_masks + sust_mask, file = full_path(p_data, "mask.tif"), overwrite = TRUE)

# Apply masks
lapply(gti, function(x) {
  out_name <- paste(gsub("\\.tif", "", x@file@name), "_mask.tif", sep = "")
  mask(x, full_mask, file = out_name, overwrite = TRUE)
})
```

3 Cropland errors, part 1

This section of the analysis calculates the differences between the reference cropland percentages and those from the various landcover products. It draws from results generated by the cropland pre-processing. Bias statistics calculated here are primarily for the supplementals.

3.1 Analyses

```
library(lmisc)
library(raster)
library(SAcropland)
p_root <- proj_root("SAcropland")
p_data <- full_path(p_root, "SAcropland/external/ext_data/")
```

3.1.1 Load cropland test grids

Derived from global landcover products (note: it turns out GLC-Share and geowiki are the same thing, so GLC-Share dropped)

```
salc <- raster(full_path(p_data, "sa_ag_masked.tif")) * 100
ds <- full_path(p_data, dir(p_data, "globSA*.*1kmrect_sum.tif")[c(2:3, 1)])
globsa_list_sum <- lapply(ds, raster)
fnames <- full_path(p_data, dir(p_data, "mod_1*.*1kmrect_sum.tif")[c(2:3, 1)])
modsa_list_sum <- lapply(fnames, raster)
# glcsa <- raster(full_path(p_data, "glcsa_masked.tif"))
geow <- raster(full_path(p_data, "geowikisa_masked.tif"))
lclist <- c(salc, globsa_list_sum, modsa_list_sum, geow)  # into list
names(lclist) <- c("sa30", "globmin", "globmu", "globmax", "modmin", "modmu",
                   "modmax", "geow")
```

```
load(full_path(p_data, "MZshapes.Rdata"))
paths <- full_path(p_data, dir(p_data, paste0("cover*.*sum_mask.tif$")))
gti <- lapply(paths, raster)
names(gti) <- c("g2007", "g2011")
```

3.1.1.1 Reference data

```
sumna <- function(x) sum(x, na.rm = FALSE)
namask <- calc(stack(stack(gti), stack(lclist)), sumna)
namask[namask >= 0] <- 1
namask2 <- !is.na(namask)  # set NAs to zero
# cellStats(namask2, sum)
namask[namask == 0] <- NA
# cellStats(namask, sum)
writeRaster(namask, filename = full_path(p_data, "namask.tif"),
            overwrite = TRUE)

# mask out all joint NAs
gti <- lapply(gti, function(x) mask(x, namask))
lclist <- lapply(lclist, function(x) mask(x, namask))
```

3.1.1.2 Apply NA masks

3.1.2 Calculate differences at different levels of aggregation

This step draws on functions I wrote to calculate the actual and absolute differences between a given “master” grid and lists of other grids. This comparison is done at multiple aggregations, ranging from 1-600 km resolution cropland averages. The lists have two levels of nesting, here the upper provides the resolution, the inner provides the different landcover datasets.

```r
# Aggregate rasters
fact <- c(5, 10, 25, 50, 100)
#fact <- 25

area_wgts <- aggregate_rast_list(fact, list(namask2), fun = sum)  # cells/pixel
lc_agg <- aggregate_rast_list(fact, lclist)   # landcover rasters
gti_agg <- aggregate_rast_list(fact, gti)  # GTI rasters

# Calculate differences
# actual
dlist_act <- rast_list_math(list(names(lclist), names(lc_agg), names(gti)),
                            gti_agg, lc_agg, "a - b", silent = FALSE)
# absolute
dlist_abs <- rast_list_math(list(names(lclist), names(lc_agg), names(gti)),
                            gti_agg, lc_agg, "abs(a - b)", silent = FALSE)
dlist_gti <- rast_list_math(list("g2007", names(gti_agg), "g2011"),
                            gti_agg, gti_agg, "a - b", silent = FALSE)

save(dlist_act, file = full_path(p_data, "d_grid_act.rda"))
save(dlist_abs, file = full_path(p_data, "d_grid_abs.rda"))

# Write out the gti cover raster and the actual bias rasters for running impact of bias examples in lan
# surface model (25 km resolution)
r <- gti_agg$f25$g2011
projection(r) <- projection(salc)
writeRaster(r, filename = full_path(p_data, "gti_2011_25km.tif"))
b <- brick(stack(lapply(c("sa30", "globmu", "modmu", "geow"), function(x) {
  dlist_act[[x]]$f25$g2011
})))
projection(b) <- projection(salc)
writeRaster(b, filename = full_path(p_data, "lc_bias_25km.tif"))

# r <- raster(full_path(p_data, "gti_2011_25km.tif"))
# lcb <- brick(full_path(p_data, "lc_bias_25km.tif"))
# plot(r - lcb, axes = FALSE, box = FALSE)
# plot(lcb)
# plot(round(lc_agg$f25$geow - (r - lcb$lc_bias_25km.4), 7))
```

3.1.3 Error versus cropland density

3.1.3.1 Weighting rasters Based on number of non-NA cells in each aggregated pixel, and then “bin” rasters, where each category represents a different cropland fractional cover range (0-5%, 5-10%, etc)

```
# cropland cover bins
binv <- seq(0, 100, 5)
gti_bins <- lapply(gti_agg, function(x) {
  lapply(x, function(y) cut(y, breaks = binv, include.lowest = TRUE))
})
```

3.1.3.2 Calculate bias/MAE, pooling MODIS and GlobCover across all versions

Step 1. Reshape lists, so that all MODIS and GlobCover variants are pooled by scale

```
# Set up indexing parameters
fact_all <- c(1, fact)# full aggregation levels
lev_vec <- paste("f", fact_all, sep = "")
nm_rt <- c("sa30", "glob", "mod", "geow")

# Reshape the list of actual landcover differences
dlists <- lapply(list(dlist_act, dlist_abs), function(i) {
  resh <- lapply(nm_rt, function(x) {  # select sensor
    l2 <- lapply(lev_vec, function(y) {  # isolate by scale
      l3 <- lapply(i[grep(x, names(i))], function(z) z[[y]])  # sensor by scale
        named_out(l3, names(i)[grep(x, names(i))])
      })
    named_out(l2, lev_vec)
  })
  named_out(resh, nm_rt)
})
names(dlists) <- c("act", "abs")
rm(dlist_act, dlist_abs)  # remove original to make space
```

3.1.3.3 Then calculate bias/MAE across datasets

Note: this uses an older and slower version of the code. Subsequent analyses use the faster version based on data.tables.

```
dang <- Sys.time()
# i <- dlists[[1]]; x <- names(i)[1]; y <- lev_vec[5]; j <- 14; z <- 1; k <- 1
dstats_all <- lapply(dlists, function(i) {
  print("processing 1 of 2 main bias lists")
  dstats <- lapply(names(i), function(x) {
    print(paste("..", x))
    l1 <- lapply(lev_vec, function(y) {
      print(paste("....", y))
      d1 <- do.call(rbind, lapply(1:(length(binv) - 1), function(j) {
        print(paste("......bin", j))
        d2 <- do.call(rbind, lapply(1:length(i[[x]][[y]]), function(z) {
          print(paste("........variant", z))
          l3 <- i[[x]][[y]][[z]]
          d3 <- do.call(rbind, lapply(1:length(l3), function(k) {
            rs <- gti_bins[[y]][[k]] == j
            rs[rs == 0] <- NA
            o <- rs * l3[[k]]
            w <- area_wgts[[y]][[1]] * rs
            oo <- cbind.data.frame("lc" = rep(names(i)[z], length(w)),
                                   getValues(o), getValues(w))
          }))
          d3f <- d3[!is.na(d3[, 2]), ]
        }))
        if(nrow(d2) > (length(i[[x]][[y]]) * 10)) {
          qs <- c(nrow(d2), box_stats(d2[, 2], weighted = TRUE,
                                      weight.opts = list("weights" = d2[, 3])))
        } else if((nrow(d2) > 0) & (nrow(d2) < length(i[[x]][[y]]) * 20)) {
          qs <- c(nrow(d2), rep(NA, 5),
                  Hmisc::wtd.mean(d2[, 2], weights = d2[, 3], na.rm = FALSE))
        } else {
          qs <- rep(NA, 7)
        }
      }))
      cbind(binv[-1], d1)
    })
    named_out(l1, lev_vec)
  })
  named_out(dstats, names(i))
})
dut <- Sys.time() - dang  # 33 minutes
save(dstats_all, file = full_path(p_data, "bias_stats.rda"))
```

3.1.4 Differences between 2007 and 2011 reference

```
dstats_gti <- lapply(lev_vec, function(x) {
    #x <- lev_vec[[2]]
    i <-  dlist_gti$g2007[[x]][[1]]
    print(paste("..", x))
    d1 <- do.call(rbind, lapply(1:(length(binv) - 1), function(y) {
      #y <- 1
      print(paste("....", y))
      rs <- gti_bins[[x]][[2]] == y
      rs[rs == 0] <- NA
      o <- rs * i
      w <- area_wgts[[x]][[1]] * rs
      oo <- cbind.data.frame(rep(y, length(w)), getValues(o), getValues(w))
      oof <- oo[!is.na(oo[, 2]), ]
      if(nrow(oof) > 10) {
        qs <- c(nrow(oof), box_stats(oof[, 2], weighted = TRUE,
                                     weight.opts = list("weights" = oof[, 3])))
      } else if((nrow(oof) > 0) & (nrow(oof) < 10)) {
        qs <- c(nrow(oof), rep(NA, 5),
                Hmisc::wtd.mean(oof[, 2], weights = oof[, 3], na.rm = FALSE))
      } else {
        qs <- rep(NA, 7)
      }
    }))
    cbind(binv[-1], d1)
})
save(dstats_gti, file = full_path(p_data, "bias_stats_gti.rda"))
```

Checks: compared number of non-NA cells coming out of each GTI version at each level of aggregation (fine); compared cell counts between previous committed version and this one (difference exists, but due to fact that gtimu is no longer considered; checked that previous versions of difference rasters and reshapes of them were referencing the correct GTI-landcover bias rasters (i.e. to make sure rasters weren't incorrectly indexed: all fine)); box_stats with weighted quantile option compared to previous commit where wtd.quantile was used separately (fine). Compared results from original roughed out version of approach for calculating bias statistics with results from new outputs list (using geowiki at 20 km aggregation as the comparison)

3.1.5 2011 reference-test bias/accuracy

Whole country and agricultural areas only

```r
load(full_path(p_data, "d_grid_act.rda"))  # actual diffence grids

# Subset difference grids, for 2011
snms <- c("sa30", "globmu", "modmu", "geow")
lc_pct_diff <- lapply(dlist_act[snms], function(x) {
  sapply(x, function(y) list(y$g2011))
})

# construct agricultural area union mask
lev <- names(lc_pct_diff[[1]])
lcu <- lapply(lev, function(x) {
  print(paste(".", x))
  lcb <- lapply(snms, function(y) {
    print(paste("...", y))
    gti_gt0 <- Which(gti_agg[[x]][[1]] > 0)
    lc_gt0 <- Which(lc_agg[[x]][[y]] > 0)
    all_gt0 <- gti_gt0 + lc_gt0
    all_gt0[all_gt0 > 0] <- 1
    all_gt0
  })
  named_out(lcb, snms)
})
names(lcu) <- lev

bins <- lapply(gti_bins, function(x) x$g2011)
awgts <- lapply(area_wgts, function(x) x[[1]])

# calculate means for 2011
wm <- function(x, w) stats::weighted.mean(x, w)
wma <- function(x, w) stats::weighted.mean(abs(x), w)
a <- bias_stats_list(bins, awgts, lcu, lc_pct_diff, wm, "mu", "bias", TRUE)
b <- bias_stats_list(bins, awgts, lcu, lc_pct_diff, wma, "mua", "bias", TRUE)
mu_dt <- rbind(a, b)
mu <- extract_stat(lev, snms, "all", "mu", "bias", mu_dt)  # ag area means
mua <- extract_stat(lev, snms, "all", "mua", "bias", mu_dt)  # ag area abs mean
mu_nm <- extract_stat(lev, snms, "all", "mu.nm", "bias", mu_dt)  # country mean
mua_nm <- extract_stat(lev, snms, "all", "mua.nm", "bias", mu_dt)  # country abs
# across bin mean and abs mean
mu_bin <- mu_dt[!like(bin, "all") & bvals == "mu", mean(bias), by = .(ol, il)]
setnames(mu_bin, "V1", "bias")
mua_bin <- mu_dt[!like(bin, "all") & bvals == "mua", mean(bias), by = .(ol, il)]
setnames(mua_bin, "V1", "bias")
# mu_dt[il == "sa30" & bin != "all" & ol == "f1" & bvals == "mua", mean(bias)]
# check
# Older variant of code from compare-landcover.Rmd to evaluate whether newer DT
# version is finding correct results
bv <- "mua"
for(chk in list(c("modmu", "f25"), c("geow", "f10"), c("sa30", "f50"),
                c("globmu", "f1"))) {
  x <- chk[1]
  y <- chk[2]
  print(paste(".", x, "..", y))
  rs <- lcu[[y]][[x]]
  rs[rs == 0] <- NA
  l3 <- abs(lc_pct_diff[[x]][[y]])
  o <- rs * l3
  w <- awgts[[y]][[1]] * rs
  wmu <- weighted.mean(getValues(o), getValues(w), na.rm = TRUE)
  print(b[ol == y & il == x & bvals == bv & bin == "all", bias] ==
          round(wmu, 2))
}  # check

svec <- c("mu_nm", "mua_nm", "mu", "mua", "mu_bin", "mua_bin")
aa <- c(rep("Country", 2), rep("Agricultural", 2), rep("Density independent", 2))
bb <- c("Bias", "MAE", "Bias", "MAE", "Bias", "MAE")
mcap <- c("SA-LC", "GlobCover", "MODIS", "GeoWiki")

lcf_out11 <- do.call(rbind, lapply(1:length(svec), function(i) {
  ao <- do.call(rbind, lapply(snms, function(x) {
    ai <- sapply(lev, function(y) {
      get(svec[i])[ol == y & il == x, bias]
    })
    named_out(ai, paste(c(1, fact), "km"))
  }))
  #cbind.data.frame("Region" = aa[i], "Map" = mcap, a)
  cbind.data.frame(Region = aa[i], Metric = bb[i], "Map" = mcap, ao)
}))

caption <- paste("Biases and mean absolute errors (MAE) in cropland maps",
                 "relative to the 2011 reference map for each aggregation",
                 "scale calculated over the entire country, for the union of",
                 "agricultural regions (cropland $>$ 0), and as density",
                 "independent means, wherein the mean bias/MAE values",
                 "for each of 20 cropland cover classes (representing",
                 "5\\% increments of cover 0\\% to 100\\% defined",
                 "by the reference map) were calculated and then averaged.")
lcfout_xtab <- xtable(lcf_out11, digits = 1, caption = caption)
print(lcfout_xtab, type = "latex", file = "paper/figures/cropland2011-bias.tex",
      tabular.environment = "longtable", floating = FALSE,
      caption.placement = "top", include.rownames = FALSE)

# do.call(rbind, lapply(snms, function(i) {
#   mu_dt[!like(bin, "all") & ol == "f1" & il == i & bvals == "mu.nm", N]
# }))
mubias_2011 <- mu_dt
save(lcf_out11, mubias_2011, file = full_path(p_data, "lcf_out2011.rda"))
```

3.1.6 2007 reference-test bias/accuracy

Whole country and agricultural areas only

```r
# Subset difference grids, for 2011
snms <- c("sa30", "globmu", "modmu", "geow")
lc_pct_diff <- lapply(dlist_act[snms], function(x) {
  sapply(x, function(y) list(y$g2007))
})

# construct agricultural area union mask
lev <- names(lc_list[[1]])
lcu <- lapply(lev, function(x) {
  print(paste(".", x))
  lcb <- lapply(snms, function(y) {
    print(paste("...", y))
    gti_gt0 <- Which(gti_agg[[x]][[1]] > 0)
    lc_gt0 <- Which(lc_agg[[x]][[y]] > 0)
    all_gt0 <- gti_gt0 + lc_gt0
    all_gt0[all_gt0 > 0] <- 1
    all_gt0
  })
  named_out(lcb, snms)
})
names(lcu) <- lev

bins <- lapply(gti_bins, function(x) x$g2007)
awgts <- lapply(area_wgts, function(x) x[[1]])

# calculate means for 2007
wm <- function(x, w) stats::weighted.mean(x, w)
wma <- function(x, w) stats::weighted.mean(abs(x), w)
a <- bias_stats_list(bins, awgts, lcu, lc_pct_diff, wm, "mu", "bias", TRUE)
b <- bias_stats_list(bins, awgts, lcu, lc_pct_diff, wma, "mua", "bias", TRUE)
mu_dt <- rbind(a, b)
mu <- extract_stat(lev, snms, "all", "mu", "bias", mu_dt)
mua <- extract_stat(lev, snms, "all", "mua", "bias", mu_dt)
mu_nm <- extract_stat(lev, snms, "all", "mu.nm", "bias", mu_dt)
mua_nm <- extract_stat(lev, snms, "all", "mua.nm", "bias", mu_dt)
# across bin mean and abs mean
mu_bin <- mu_dt[!like(bin, "all") & bvals == "mu", mean(bias), by = .(ol, il)]
setnames(mu_bin, "V1", "bias")
mua_bin <- mu_dt[!like(bin, "all") & bvals == "mua", mean(bias), by = .(ol, il)]
setnames(mua_bin, "V1", "bias")

# check
# Older variant of code from compare-landcover.Rmd to evaluate whether newer DT
# version is finding correct results
bv <- "mua"
for(chk in list(c("modmu", "f25"), c("geow", "f10"), c("sa30", "f50"),
                c("globmu", "f1"))) {
  x <- chk[1]
  y <- chk[2]
  print(paste(".", x, "..", y))
  rs <- lcu[[y]][[x]]
  rs[rs == 0] <- NA
  l3 <- abs(lc_pct_diff[[x]][[y]])
  o <- rs * l3
  w <- awgts[[y]][[1]] * rs
  wmu <- weighted.mean(getValues(o), getValues(w), na.rm = TRUE)
  print(b[ol == y & il == x & bvals == bv & bin == "all", bias] ==
          round(wmu, 2))
}  # check

svec <- c("mu_nm", "mua_nm", "mu", "mua", "mu_bin", "mua_bin")
aa <- c(rep("Country", 2), rep("Agricultural", 2), rep("Density independent", 2))
bb <- c("Bias", "MAE", "Bias", "MAE", "Bias", "MAE")
mcap <- c("SA-LC", "GlobCover", "MODIS", "GeoWiki")

lcf_out07 <- do.call(rbind, lapply(1:length(svec), function(i) {
  ao <- do.call(rbind, lapply(snms, function(x) {
    ai <- sapply(lev, function(y) {
      get(svec[i])[ol == y & il == x, bias]
    })
    named_out(ai, paste(c(1, fact), "km"))
  }))
  #cbind.data.frame("Region" = aa[i], "Map" = mcap, a)
  cbind.data.frame(Region = aa[i], Metric = bb[i], "Map" = mcap, ao)
}))

caption <- paste("Biases and mean absolute errors (MAE) in cropland maps",
                 "relative to the 2007 reference map for each aggregation scale",
                 "calculated over the entire country, for the union of",
                 "agricultural regions, and as density",
                 "independent means, wherein the mean bias/MAE values",
                 "for each of 20 cropland cover classes",
                 "(representing 5\\% increments of cover 0\\% to 100\\% defined",
                 "by the reference map) were calculated and then averaged.")
lcfout_xtab <- xtable(lcf_out07, digits = 1, caption = caption)
print(lcfout_xtab, type = "latex", file = "paper/figures/cropland2007-bias.tex",
      tabular.environment = "longtable", floating = FALSE,
      caption.placement = "top", include.rownames = FALSE)
```

4 Cropland errors, part 2

The following provide further analyses and plots of bias and accuracy metrics calculated from spatial error maps created in compare-landcover (which includes some supplementary metrics),

4.1 Spatial biases

4.1.1 Datasets

```
library(raster)
library(lmisc)
library(R.utils)
library(RColorBrewer)

# Paths
p_root <- proj_root("SAcropland")
p_fig <- full_path(p_root, "SAcropland/paper/figures/")
p_data <- full_path(p_root, "SAcropland/external/ext_data/")

# Load in datasets
load(full_path(p_data, "MZshapes.Rdata"))  # SA shape
load(full_path(p_data, "d_grid_act.rda"))  # actual diffence grids
load(full_path(p_data, "d_grid_abs.rda"))  # actual diffence grids
load(full_path(p_data, "bias_stats.rda"))  # bias statistics
load(full_path(p_data, "bias_stats_gti.rda"))  # bias statistics
# cellStats(!is.na(dlist_act$modmu$f1$g2011), sum)
```

4.2 Error maps

Plot maps of biases at different resolutions. To avoid the visual distortion of NA areas filled in by larger levels of aggregation, each aggregated difference must be disaggregated first and masked before plotting.

```
namask <- raster(full_path(p_data, "namask.tif"))
# cellStats(namask, sum)
snms <- c("sa30", "globmu", "modmu", "geow")  # all lc data (no MOD/GC extremes)
lev <- c("f1", "f25", "f50", "f100")  # just show four levels

# disggregate selected rasters at selected levels
# x <- snms[3]; y <- lev[3]
disagg <- lapply(snms, function(x) {
  l1 <- lapply(lev, function(y) {
    if(y == "f1") {
      r <- dlist_act[[x]][[y]]$g2011
    } else {
      r <- disaggregate(dlist_act[[x]][[y]]$g2011,
                        fact = as.numeric(gsub("f", "", y)))
      r <- raster::mask(crop(r, namask), namask)
    }
  })
  named_out(l1, lev)
})
names(disagg) <- snms

# writeRaster(disagg$globmu$f100,
#             filename = "external/ext_data/test/disagg100km.tif")
```

4.2.1 Plot of selected rasters and levels

```
# Plotting colors
# cols <- colorRampPalette(c("red", "grey80", "blue4"))
brks <- c(-100.1, -75, -50, -25, -10, -5, -1, 1, 5, 10, 25, 50, 75, 100.1)
```

```
cols <- c(rev(brewer.pal(length(brks) / 2, "Reds")[-1]), "grey80",
          brewer.pal(length(brks) / 2, "Blues")[-1])
brklen <- length(brks) - 1
# cols <- cols(brklen)
# cols[c(1, length(cols))] <- c("darkred", "purple3")
legtext <- "% Difference"
cx <- 1.4
lcol <- "black"
mcap <- c("SA-LC", "GlobCover", "MODIS", "GeoWiki")
lev2 <- c("1 km", "25 km", "50 km", "100 km")
pdf(full_path(p_fig, "bias_map.pdf"), height = 7, width = 7)
par(mfrow = c(4, 4), mar = c(0, 0, 0, 0), oma = c(3.7, 1.5, 1.5, 0))
for(i in 1:length(snms)) {
  print(snms[i])
  for(j in 1:length(lev)) {
    print(lev[j])
    plot(sa.shp, lty = 0)
    plot(disagg[[snms[i]]][[lev[j]]], add = TRUE, col = cols, breaks = brks,
         legend = FALSE)
    if(j == 1) mtext(mcap[i], side = 2, line = -0.25, cex = cx)
    if(i == 1) mtext(lev2[j], side = 3, line = -0.5, cex = cx)
  }
}
flex_legend(ncuts = brklen, legend.text = legtext,
            legend.vals = round(brks), longdims = c(0.2, 0.8),
            shortdims = c(0.05, 0.01), colvec = cols, srt = c(270, 0),
            horiz = TRUE, textside = "bottom", legend.pos = c(4, 7),
            leg.adj = list(c(0.2, 0.35), c(0.5, -0.5)), cex.val = cx,
            textcol = lcol, bordercol = lcol)
dev.off()

lev <- names(dlist_abs$sa30)
```

The map below was created by the code above, showing biases at 1 km, 25 km, 50 km, and 100 km from South Africa's landcover product, the mean estimates from GlobCover and MODIS, and Geowiki/GLC-SHARE.

4.3 Bias/MAE in relation to field density

4.3.1 1 km Bias/MAE with boxplots

For supplemental.

```
#btype <- c("bias_act.pdf", "bias_abs.pdf")
# plot matrix
nc <- 4
m <- matrix(rep(c(rep(1, 8), rep(2, 8), rep(3, 8), rep(4, 8)), 2), nrow = 6,
            ncol = 32, byrow = TRUE)
cols <- c("red", "orange3", "green4", "blue", "yellow")
bcol <- c("grey50", "grey50", "grey50", "grey")
lcnms <- c("SA LC", "GlobCover", "MODIS", "Geowiki")
fcol <- "grey"
lev <- names(dstats_all$act$sa30)
```

```r
lvec <- seq(5, 100, 5)
pp <- list("x" = c(-2, 21), "y" = c(0, 0),
           "x21" = list(c(0, 9), c(9, 21)),
           "y2" = list(c(seq(-100, -10, 10), seq(10, 100, 10)),
                       seq(10, 100, 10)),
           "y1lwd" = list(2, 1),
           "y2lwd" = 0.5, #c(1, 0.5, 1, 0.5, 0.5, 0.5, 0.5, 1, 0.5, 1),
           "y2lc" = list(c("black", rep("grey80", 4), "black", rep("grey80", 4),
                           rep("grey80", 4), "black", rep("grey80", 4),"black"),
                         rep("grey80", 10)),
           "yl" = list(c(-100, 100), c(0, 100)),
           "yl2" = c(-20, 60), "xl" = c(1, 20),
           "xaxl" = unlist(lapply(seq(5, 100, 10), function(x) c(x, " "))),
           "xaxl2" = ifelse(seq(5, 100, 5) %in% c(25, 50, 75), seq(5, 100, 5),
                            " "),
           "yax1" = list(seq(-100, 100, 50), seq(0, 100, 20)),
           "ycol" = c("grey70"))  # plot parameters

pdf(full_path(p_fig, "biases_1km.pdf"), width = 7, height = 5)
lmat <- rbind(m, m + max(m))
l <- layout(lmat)
par(mar = c(3, 0, 0, 0), oma = c(1, 4, 2, 2))
for(bt in 1:2) {
  snms <- names(dstats_all[[bt]])[names(dstats_all[[bt]]) != "glcsh"]
  dset <- dstats_all[[bt]][snms]
  for(i in 1:length(snms)) { i
    d <- dset[[snms[i]]]$f1
    n <- nrow(d)
    plot(pp$x, pp$y, ylim = pp$yl[[bt]], xlim = pp$xl, type = "l",
         lwd = pp$y1lwd[[bt]], axes = FALSE,
         xlab = "", ylab = "")
    if(i == 1) axis(2, at = pp$yax1[[bt]], labels = pp$yax1[[bt]], las = 2,
                    mgp = c(1, 0.6, 0))
    axis(1, at = 1:20, labels = pp$xaxl, las = 2, mgp = c(1, 0.4, 0),
         tcl = -0.3, pos = pp$yaxp[[bt]], cex = 0.8)
    for(z in 1:length(pp$y2[[bt]])) {
      lines(pp$x, rep(pp$y2[[bt]][z], 2), lwd =pp$y2lwd, col = pp$y2lc[[bt]][z])
    }
    for(k in 1:n) {
      a <- seq(grep("2.5", colnames(d)), by = 1, length.out = 6)
      if(!is.na(d[k, 2])) {
        boxplot_v(k, d[k, a], n, 150, 10, bcol, fcol, pcex = 0.5,
                  lwd = c(1, 0.5, 2))
      }
    }
    lines(1:n, d[, "mu"], pch = 20, col = cols[i], lwd = 3)
    if(i == 1) mtext(ifelse(bt == 1, "Bias (%)", "Mean absolute error (%)"),
                     side = 2, line = 2.2, cex = 0.8)
    if(bt == 1) mtext(lcnms[i], side = 3, line = 0, col = cols[i])
  }
}
mtext("% Cropland", side = 1, outer = TRUE, line = -1, cex = 0.8)
dev.off()
```

4.3.2 Bias/MAE with scale

Note: the mean values presented in the figure here is a straight mean across 5 percentile bins, which gives bias independent of cropland density. Supplementary tables report the bin-weighted biases.

```
#lev <- lev[!lev %in% c("f5", "f300")]
levnms <- paste(gsub("f", "", lev), "km")
m2 <- unlist(lapply(seq(1, by = 1, length = length(lev) * 4), function(x) {
  rep(x, nc)
}))
m3 <- do.call(rbind, lapply(seq(1, length(m2), length(lev) * nc), function(x) {
  do.call(rbind, rep(list(m2[x:(x + length(lev) * nc - 1)]), nc))
}))

yl <- seq(pp$yl2[1], pp$yl2[2], 10)
ylcol <- c("black", "grey", "black", rep("grey", 5), "black")
lt <- c(5, 1)
ylwd <- c(1.5, 2)

wm <- function(x, w) stats::weighted.mean(x, w, na.rm = TRUE)
wma <- function(x, w) stats::weighted.mean(abs(x), w)

pdf(full_path(p_fig, "biases_1-100km.pdf"), width = 7.5, height = 7)
l <- layout(m3)
# layout.show(l)
# i <- 1; j <- 1; bt <- 1
par(mar = c(0.5, 0, 1, 0.3), oma = c(5, 5, 2, 2))
for(i in 1:length(snms)) {
  highcounter <- rep(0, length(lev))
  for(j in 1:length(lev)) {
    plot(pp$x, pp$y, ylim = pp$yl2, xlim = pp$xl, type = "l", lwd = 1,
         axes = FALSE, xlab = "", ylab = "", xaxs = "i")
    for(z in 1:length(yl)) {
      lines(pp$x, rep(yl[z], 2), lwd = pp$y2lwd, col = ylcol[z])
    }
    muv <- rep(NA, 2)
    muv2 <- rep(NA, 2)
    for(bt in 1:2) {
      d <- dstats_all[[bt]][[i]][[j]]
      n <- nrow(d)
      lines(1:n, d[, "mu"], pch = 20, col = cols[i], lwd = ylwd[bt],
            lty = lt[bt])
      muv[bt] <- round(mean(d[, 8], na.rm = TRUE)) # unweighted means
      muv2[bt] <- round(wm(d[, 8], d[, 2]))  # weighted means
      toohigh <- which(d[, "mu"] > pp$yl2[2])
      if(length(toohigh) > 0) {
        highcounter[j] <- 1
        xi <- toohigh[which.max(d[toohigh, "mu"])]
        text(xi, y = pp$yl2[2] + 4, labels = round(d[xi, "mu"]), cex = 1.1,
             col = "grey50", xpd = NA)
      }
    }
    text(5, pp$yl2[2] - 5, labels = paste(muv, collapse = " / "), cex = 1.1,
         col = "grey10", font = 4)
    text(5, pp$yl2[2] - 16, labels = paste(muv2, collapse = " / "), cex = 1.1,
         col = "grey10")
    if(j == 1) axis(2, at = yl, labels = yl, las = 2, mgp = c(1, 0.6, 0))
    if(j == 1) mtext(lcnms[i], side = 2, line = 3, cex = 1, col = cols[i])
    if(i == 1) mtext(levnms[j], side = 3, line = -0.4, cex = 1.2)
    axis(1, at = c(1, 5, 10, 15, 20), labels = rep("", 5), mgp = c(1, 1, 0),
         tcl = -0.7, lwd.ticks = 1.2, pos = pp$yl2[1])
    axis(1, at = 1:20, labels = rep("", 20), las = 2, mgp = c(1, 0.4, 0),
         tcl = -0.3, pos = pp$yl2[1])
    if(i == length(snms)) {
      axis(1, at = 1:20, labels = pp$xaxl2, las = 2, mgp = c(1, 0.8, 0),
           tcl = -0.3, pos = pp$yl2[1])
    }
  }
}
mtext("% Bias/MAE", side = 2, line = 2, cex = 0.8, outer = TRUE, adj = 0.5)
mtext("% Cropland", side = 1, outer = TRUE, line = 1.5, cex = 0.8)
par(xpd = NA)
x <- grconvertX(0.75, from = "ndc", to = "user")
y <- grconvertY(0.07, from = "ndc", to = "user")
legend(x = x, y = y, legend = c("Bias", "Mean absolute error"), lty = lt,
       lwd = ylwd, bty = "n", cex = 1.2)
dev.off()
```

4.3.3 Total error

Calculate how much cropland estimates differ between datasets

```
paths <- full_path(p_data, dir(p_data, paste0("cover*.*sum_mask.tif$")))
gti <- lapply(paths, raster)
names(gti) <- c("g2007", "g2011")
gti <- lapply(gti, function(x) mask(x, namask))  # apply namask
snms <- c("sa30", "globmu", "modmu", "geow")  # all lc data (no MOD/GC extremes)

# recreate cropland percentages
lc2007 <- lapply(snms, function(x) gti$g2007 - dlist_act[[x]]$f1$g2007)
lc2011 <- lapply(snms, function(x) gti$g2011 - dlist_act[[x]]$f1$g2011)
# plot(lc2007[[2]] - lc2011[[2]])  # same thing, as it should be

# calculate km2 cropland in LC datasets
lckm2 <- sapply(lc2011, function(x) cellStats(x / 100, sum))
refkm2 <- sapply(gti, function(x) cellStats(x / 100, sum))
refkm2[2] / refkm2[1]  # 3.6 percent more in 2011 than 2007

# bias in total cropland relative to GTI
stats2007 <- round((refkm2[2] - lckm2) / refkm2[2] * 100, 1)
stats2011 <- round((refkm2[1] - lckm2) / refkm2[1] * 100, 1)

# N km^2 in country
#sapply(lc2007, function(x) cellStats(!is.na(x), sum))
# sapply(lc2011, function(x) cellStats(!is.na(x) , sum))
# sapply(gti, function(x) cellStats(!is.na(x) , sum))
totkm2 <- cellStats(!is.na(gti$g2011) , sum)
```

```r
save(stats2007, stats2011, totkm2, refkm2,
     file = full_path(p_data, "total-cropland.rda"))
```

Total square kilometers of South Africa being evaluated

```
## [1] 1079793
```

Total square kilometers of cropland according to 2007 and 2011 GTI data

```
##    g2007    g2011
## 100624.4 104303.7
```

Percent bias in total cropland area estimates relative to 2007 GTI data

```
## [1] -26.0  21.0  26.1  -5.7
```

And relative to 2011 GTI data

```
## [1] -30.6  18.2  23.4  -9.6
```

4.4 Cropland bias/MAE plots

4.4.1 Primary method

Weighted by **actual** cropland percentage, and then, with aggregation, by number of pixels being aggregated.

```r
# Create weights maks
fact <- c(5, 10, 25, 50, 100)
namask <- raster(full_path(p_data, "namask.tif"))  # NA mask
# cellStats(!is.na(gti$g2011), sum)
# cellStats(namask, sum)

namask2 <- !is.na(namask)
awgts <- aggregate_rast_list(fact, list(namask2), fun = sum)
awgts <- lapply(awgts, function(x) x[[1]])

# Subset difference grids, for 2011
snms <- c("sa30", "globmu", "modmu", "geow")
err2011 <- lapply(lev, function(x) {
  sapply(dlist_act[snms], function(y) list(y[[x]]$g2011))
})
names(err2011) <- lev

gti_agg <- aggregate_rast_list(fact, gti)  # GTI rasters
gti_agg11 <- lapply(gti_agg, function(x) x$g2011)

# calculate bias/MAE for data
a <- bias_statsw_list(gti_agg11, awgts, err2011, snms, wm, "mu")
b <- bias_statsw_list(gti_agg11, awgts, err2011, snms, wma, "mua")
bacc2011 <- rbind(a, b)
```

```r
# check error stats - do they match alternate approaches?
for(i in c("f1", "f25", "f10")) {
  for(j in c("sa30", "modmu", "geow")) {
    print(paste("cross-checking calculations in", i, j))
    a1 <- bias_statsw(gti_agg11[[i]], awgts[[i]], err2011[[i]], snms, wm,
                      "mu", aweight = FALSE)[, j, with = FALSE]
    a2 <- bias_statsw(gti_agg11[[i]], awgts[[i]], err2011[[i]], snms, wm,
                      "mu", rweight = FALSE, aweight = FALSE)[,j, with=FALSE]
    a3 <- bias_statsw(gti_agg11[[i]], awgts[[i]], err2011[[i]], snms, wm,
                      "mu")[, j, with = FALSE]
    c1 <- getValues(err2011[[i]][[j]])
    w1 <- getValues(gti_agg11[[i]])
    w2 <- getValues(awgts[[i]])
    print("non-area weighted mean matches?")
    print(a1 == round(weighted.mean(c1, w1, na.rm = TRUE), 2))
    print("totally unweighted mean matches?")
    print(a2 == round(cellStats(err2011[[i]][[j]], mean), 2))
    print("double weighted mean matches?")
    print(a3 == round(weighted.mean(c1, w1 * w2, na.rm = TRUE), 2))
  }
}

# Subset difference grids, for 2007
snms <- c("sa30", "globmu", "modmu", "geow")
err2007 <- lapply(lev, function(x) {
  sapply(dlist_act[snms], function(y) list(y[[x]]$g2007))
})
names(err2007) <- lev
gti_agg07 <- lapply(gti_agg, function(x) x$g2007)

a <- bias_statsw_list(gti_agg07, awgts, err2007, snms, wm, "mu")
b <- bias_statsw_list(gti_agg07, awgts, err2007, snms, wma, "mua")
bacc2007 <- rbind(a, b)

fact <- c(1, 5, 10, 25, 50, 100)
alph <- c(225, 40)
x <- c(0, 3, 6, 9, 12, 15)
w <- 3 / 8
xo <- (cumsum(rep(w, 8)) - w / 2)[-c(2, 4, 6, 8)]
o <- c(0, w)
xa <- sapply(x, function(x) x + xo)
cx <- c(1.25, 1, 1)
g1 <- "grey90"
reg <- c("Country", "Agricultural", "Density independent")
lcnms <- c("SA-LC", "GlobCover", "MODIS", "GeoWiki")
mutype <- c("mua", "mu")
# svec <- paste(gsub("f", "", bacc2011[stnm == "mu", ol]), "km")
svec <- bacc2011[stnm == "mu", ol]
xl <- c(-0.5, 18)
yl <- c(-30, 50)
shd <- c(4.5, 10.5, 16.5)
cols <- c("red", "orange3", "green4", "blue")
colnms <- snms
```

```
pchs <- c("+", "o", "x")
yax <- seq(yl[1], yl[2], 5)
xax <- seq(1, 6, 5 / (length(yax) - 1))

pdf(full_path(p_fig, "cropland_bias_main.pdf"), height = 7, width = 7)
par(mar = rep(1, 4), oma = c(2, 2, 0, 0), mgp = c(1, 0.5, 0), tcl = -0.3)
plot(xl, yl, pch = "", yaxt = "n", xaxt = "n", xaxs = "i", yaxs = "i",
     ylab = "", xlab = "")#,
for(i in shd) polyfunc2(x = i, y = yl, w = 3, col = g1, bcol = g1, lwd = 1)
abline(h = yax, v = NULL, col = "grey80", lty = 1)
polyfunc2(x = 8.75, y = yl, w = 18.5, col = "transparent", bcol = "black")
lines(c(-1, 18), c(0, 0), lwd = 2, col = "grey80")
# i <- 1; j <- 1; k <- 1
for(i in 1:length(svec)) {
  lv <- svec[i]
  for(j in 1:length(snms)) {
    # pchs1 <- pchs
    nm <- snms[j]
    for(k in 1:length(mutype)) {
      dat <- bacc2011[ol == lv & stnm == mutype[k], nm, with = FALSE][[1]]
      pcol <- makeTransparent(cols[j], alpha = alph[k])
      polyfunc2(xa[j, i] + o[k], c(0, dat), w = w, col = pcol, bcol = pcol)
    }
  }
}
axis(1, at = seq(1.5, 18.5, 3), labels = fact)
axis(2, at = yax, labels = yax, las = 2)
mtext(side = 1, text = "Resolution (km)", outer = TRUE, line = 0.5)
mtext(side = 2, text = "Bias/MAE (%)", outer = TRUE, line = 1)
legend(x = 12.4, y = -19.9, legend = lcnms, pch = 15, col = cols, adj = 0,
       pt.cex = 1.5, bty = "n", cex = 0.8, x.intersp = 0.5)
legend(x = 11.9, y = -19.9, legend = rep("", 4), pch = 15, adj = 0,
       col = makeTransparent(cols, alpha = alph[2]), pt.cex = 1.5, bty = "n",
       cex = 0.8)
text(x = 12.7, y = -20.5, labels = "MAE", srt = 45, adj = c(0, 0), cex= 0.8)
text(x = 12.2, y = -20.5, labels = "Bias", srt = 45, adj = c(0, 0), cex = 0.8)
dev.off()

# Difference between 2007 and 2011 bias and accuracy statistics
baccdiff <- bacc2011[, snms, with = FALSE] -  bacc2007[, snms, with = FALSE]
setnames(baccdiff, colnames(baccdiff), mcap)

# baccdiff <- round(rbind(baccdiff, baccdiff[, lapply(.SD, mean)]), 2)
baccdiff <- cbind("Resolution" = rep(paste(fact, "km"), 2), baccdiff)
xtable::xtable(baccdiff)
```

4.4.2 Other calculations methods

Across country, agricultural region only, or bin-wise (cropland density classes) mean.

```
load(full_path(p_data, "lcf_out2011.rda"))
lcf <- data.table(lcf_out11)
```

```r
fact <- c(1, 5, 10, 25, 50, 100)
alph <- c(225, 40)
x <- c(0, 3, 6, 9, 12, 15)
w <- 3 / 8
xo <- (cumsum(rep(w, 8)) - w / 2)[-c(2, 4, 6, 8)]
o <- c(0, w)
xa <- sapply(x, function(x) x + xo)
cx <- c(1.25, 1, 1)
g1 <- "grey90"
reg <- c("Country", "Agricultural", "Density independent")
lcnms <- c("SA-LC", "GlobCover", "MODIS", "GeoWiki")
mutype <- c("MAE", "Bias")
svec <- colnames(lcf)[grep("km", colnames(lcf))]
xl <- c(-0.5, 18)
yl <- c(-40, 40)
shd <- c(4.5, 10.5, 16.5)
cols <- c("red", "orange3", "green4", "blue")
colnms <- names(lcf)[4:length(names(lcf))]
pchs <- c("+", "o", "x")
yax <- seq(yl[1], yl[2], 5)
xax <- seq(1, 6, 5 / (length(yax) - 1))

pdf(full_path(p_fig, "cropland_bias_region.pdf"), height = 7, width = 7)
par(mar = rep(1, 4), oma = c(2, 2, 0, 0), mgp = c(1, 0.5, 0), tcl = -0.3)
plot(xl, yl, pch = "", yaxt = "n", xaxt = "n", xaxs = "i", yaxs = "i",
     ylab = "", xlab = "")#,
for(i in shd) polyfunc2(x = i, y = yl, w = 3, col = g1, bcol = g1, lwd = 1)
abline(h = yax, v = NULL, col = "grey80", lty = 1)
polyfunc2(x = 8.75, y = yl, w = 18.5, col = "transparent", bcol = "black")
lines(c(-1, 18), c(0, 0), lwd = 2, col = "grey80")
# i <- 1; j <- 1; k <- 1
for(i in 1:length(svec)) {
  lv <- colnms[i]
  for(j in 1:length(snms)) {
    pchs1 <- pchs
    nm <- lcnms[j]
    for(k in 1:length(mutype)) {
      #dat <- get(mutype[k])[V1 == snms[j] & V2 %in% LC2, get(lev_vec[i])]
      dat <- lcf[Metric == mutype[k] & Map == nm, lv, with = FALSE]
      pcol <- makeTransparent(cols[j], alpha = alph[k])
      polyfunc2(xa[j, i] + o[k], range(dat), w = w, col = pcol, bcol = pcol)
      for(z in 1:length(reg)) {
        lpfunc(x[i] + xo[j] + o[k], dat[z], col = "black", size = cx[z],
               pch = pchs[z], type = "pt")
      }
    }
  }
}

axis(1, at = seq(1.5, 18.5, 3), labels = fact)
axis(2, at = yax, labels = yax, las = 2)
mtext(side = 1, text = "Resolution (km)", outer = TRUE, line = 0.5)
mtext(side = 2, text = "Bias/MAE (%)", outer = TRUE, line = 1)
```

```r
legend(x = 12.4, y = -19.9, legend = lcnms, pch = 15, col = cols, adj = 0,
       pt.cex = 1.5, bty = "n", cex = 0.8, x.intersp = 0.5)
legend(x = 11.9, y = -19.9, legend = rep("", 4), pch = 15, adj = 0,
       col = makeTransparent(cols, alpha = alph[2]), pt.cex = 1.5, bty = "n",
       cex = 0.8)
text(x = 12.7, y = -20.5, labels = "MAE", srt = 45, adj = c(0, 0), cex= 0.8)
text(x = 12.2, y = -20.5, labels = "Bias", srt = 45, adj = c(0, 0), cex = 0.8)
legend(x = 12.1, y = 40,
       legend = c("Country", "Agricultural", "Density independent"), pch = pchs,
       pt.cex = c(1.5, 1.5), bty = "n", cex = 0.8)
dev.off()
```

4.4.3 Bin density plots

To see how many values contribute to bin-wise means.

```r
# Using the union of agricultural areas, rather than whole country
bin_size <- do.call(rbind, lapply(lev, function(i) {
  # i <- "f50"
  st <- mubias_2011[bin != "all" & ol == i & il == "sa30" & bvals == "mu",
                    .(bin, N)]
  st[, bin := as.integer(bin)]  # convert bin to integers
  setkey(st, bin)
  st[data.table(bin = 1:20), N]  # merge to bin key to allow NAs
}))

pdf(full_path(p_fig, "cropland_bins.pdf"), height = 7, width = 7)
par(mfrow = c(3, 2), mar = c(2, 2, 2, 2), oma = c(3, 3, 0, 0),
    mgp = c(1, 0.25, 0), tcl = -0.1)
for(i in 1:nrow(bin_size)) {
  barplot(bin_size[i, ], las = 2, col = "grey70", names.arg = seq(5, 100, 5))
  # axis(1, at = 1:20, labels = seq(5, 100, 5), las = 2)
  mtext(side = 3, text = paste(fact[i], "km"))
}
mtext(side = 1, text = "% Cropland", outer = TRUE)
mtext(side = 2, text = "Bin count", outer = TRUE, line = 1)
dev.off()

# check bin frequencies starting a bit higher: 2.5% cover
binv <- seq(2.5, 97.5, 5)
gti_bins <- lapply(gti_agg, function(x) {
  lapply(x, function(y) cut(y, breaks = binv, include.lowest = TRUE))
})
bins_25 <- lapply(gti_bins, function(x) freq(x$g2011))
barplot(bins_25$f1[1:19, 2])
```

5 Accuracy in relation to typical crop cover

5.1 Data

```
library(raster)
library(lmisc)
library(nlme)
library(mgcv)
library(ape)
library(fitdistrplus)
library(SAcropland)

# Paths
p_root <- proj_root("SAcropland")
p_fig <- full_path(p_root, "SAcropland/paper/figures/")
p_data <- full_path(p_root, "SAcropland/external/ext_data/")
p_data2 <- full_path(p_root, "SAcropland/data/")

# Load in datasets
load(full_path(p_data, "MZshapes.Rdata"))  # SA shape
load(full_path(p_data2, "ZAF_adm2.Rdata"))  # magisterial districts
load(full_path(p_data, "d_grid_act.rda"))  # actual diffence grids
load(full_path(p_data, "d_grid_abs.rda"))  # actual diffence grids
paths <- full_path(p_data, dir(p_data, paste0("cover*.*11sum_mask.tif$")))
gti <- raster(paths)  # gti 2011
gti@crs <- sa.shp@proj4string
namask <- raster(full_path(p_data, "namask.tif"))  # NA mask
gti <- mask(gti, namask)  # apply mask to GTI
# cellStats(!is.na(gti), sum)

# Transform magisterial districts, remove Prince Edward Islands
md <- spTransform(gadm, CRSobj = sa.shp@proj4string)
md <- md[md$ID_2 != 313, ]
```

Mean area of magisterial district.

```
round(mean(rgeos::gArea(md, byid = TRUE) / 1000000))
```

```
## [1] 3445
```

5.2 Analyses

5.2.1 Extract data for magisterial districts

Both cropland percentages and absolute errors.

```
# GTI data
# mean cropland cover
gti_md <- extract(x = gti, y = md, progress = "text")
save(gti_md, file = "external/ext_data/gti_md.rda")
```

```r
# load("external/ext_data/gti_md.rda")
gti_mu <- t(sapply(gti_md, function(x) {
  c(length(x), length(which(is.na(x))), round(mean(x, na.rm = TRUE), 1))
}))

# mean no zeros
gti_mu0 <- t(sapply(gti_md, function(x) {
  c(length(x), length(which(is.na(x))), round(mean(x[x > 0], na.rm = TRUE), 1))
}))

# how many contribute when non-cropland removed, and which are non-zero
nonzero <- sapply(gti_md, function(x) length(x[!is.na(x) & (x > 0)]))
whichnonzero <- sapply(gti_md, function(x) which(!is.na(x) & (x > 0)))

# median
gti_median <- t(sapply(gti_md, function(x) {
  c(length(x), length(which(is.na(x))), round(median(x, na.rm = TRUE), 1))
}))

gti_median0 <- t(sapply(gti_md, function(x) {
  c(length(x), length(which(is.na(x))),
    round(median(x[x > 0], na.rm = TRUE), 1))
}))

par(mar = c(0, 0, 0, 0))
plot(md[323, ])
plot(gti, add = TRUE)
plot(md[323, ], add = TRUE)

# Absolute biases at 1 km
snms <- c("sa30", "globmu", "modmu", "geow")
dlist_abs1km <- stack(lapply(dlist_abs[snms], function(x) x$f1$g2011))
# cellStats(!is.na(dlist_abs1km), sum)
dlist_abs1km@crs <- sa.shp@proj4string
dlist_abs1km_md <- extract(x = dlist_abs1km, y = md, progress = "text")
save(dlist_abs1km_md, file = "external/ext_data/dlist_abs1km_md.rda")
# load("external/ext_data/dlist_abs1km_md.rda")

# check counts, primarily of NA values for each dataset
dlist_counts <- t(sapply(dlist_abs1km_md, function(x) {
  apply(x, 2, function(y) {
    c(length(y), length(which(is.na(y))))#, round(mean(y, na.rm = TRUE), 1))
  })
}))

chkfun <- function(x, tol = 0.001) {
  return(apply(x, 1, function(x) abs(max(x) - min(x)) < tol))
}
a <- seq(1, 7, 2)
b <- seq(2, 8, 2)
dlist_counts[which(!chkfun(dlist_counts[, a])), a]
dlist_counts[which(!chkfun(dlist_counts[, b])), b]
```

```r
# bind stats
# MD
# some have more NAs than others, so select which LC set has max NAs
nas <- apply(dlist_counts[, seq(2, 8, 2)], 1, max)
all(gti_mu[, 1] == dlist_counts[, 1])

# then calculate mean per LC set per district
# mean
dlist_mu <- cbind(nas, t(sapply(dlist_abs1km_md, function(x) {
  round(colMeans(x, na.rm = TRUE), 1)
})))

# mean no zero
# dlist_mu0 <- cbind(nas, t(sapply(dlist_abs1km_md, function(x) {
#   tab <- round(x)
#   round(apply(tab, 2, function(x) mean(x[x > 0], na.rm = TRUE)), 1)
# })))
dlist_mu0 <- cbind(nas, t(sapply(1:length(dlist_abs1km_md), function(x) {
  tab <- round(dlist_abs1km_md[[x]])
  sel <- whichnonzero[[x]]  # select out reference cropland pixels only
  round(apply(tab, 2, function(x) mean(x, na.rm = TRUE)), 1)
})))

# plot(dlist_mu0[, 2], dlist_mu02[, 2])
# i <- 2
# plot(gti_mu0[, 3], dlist_mu02[, i])
# plot(gti_mu0[, 3], dlist_mu0[, i])

# median
dlist_median <- cbind(nas, t(sapply(dlist_abs1km_md, function(x) {
  round(matrixStats::colMedians(x, na.rm = TRUE), 1)
})))
colnames(dlist_median) <- c("nas", snms)

# dlist_median0 <- cbind(nas, t(sapply(dlist_abs1km_md, function(x) {
#   tab <- round(x)
#   round(apply(tab, 2, function(x) median(x[x > 0], na.rm = TRUE)), 1)
# })))
dlist_median0 <- cbind(nas, t(sapply(1:length(dlist_abs1km_md), function(x) {
  tab <- round(dlist_abs1km_md[[x]])
  sel <- whichnonzero[[x]]  # select out reference cropland pixels only
  round(apply(tab, 2, function(x) median(x, na.rm = TRUE)), 1)
})))


# bind function
bind_stats <- function(x, y, coords = xy,
                       snms = c("sa30", "globmu", "modmu", "geow")) {
  dat <- cbind.data.frame(x, y)
  dat$nd <- apply(dat[, c(2, 4)], 1, max)  # max of GTI and LC NAs
  colnames(dat)[c(1, 3)] <- c("N", "gti")
  dat <- dat[, -2]
  dat <- cbind(coords, dat)
  dat$dmu4 <- round(rowMeans(dat[, snms]), 1)  # mean across all 4
  dat$dmu3 <- round(rowMeans(dat[, snms[snms != "globmu"]]), 1) # mu - globcov
  dat
}

# centroids for districts
xy <- as.data.frame(rgeos::gCentroid(md, byid = TRUE))  # get centroid coords

# bind to gti for comparisons
mu <- bind_stats(x = gti_mu, y = dlist_mu)
mu$w <- mu$N - mu$nd  # weight - N contributing pixels
mu0 <- bind_stats(x = gti_mu0, y = dlist_mu0)
mu0$w <- nonzero  # weight - N contributing pixels (non-zeros)
med <- bind_stats(x = gti_median, y = dlist_median)
med$w <- med$N - med$nd  # weight - N contributing pixels
med0 <- bind_stats(x = gti_median0, y = dlist_median0)
med0$w <- nonzero  # weight - N contributing pixels (non-zeros)
```

5.2.2 Run regressions and check for spatial autocorrelation

```
# check the polygons and rows line up for coordinates
# i <- 300
# a <- extract(x = gti, y = md[i, ], progress = "text")
# length(a[[1]])
# length(which(is.na(a[[1]])))
# mean(a[[1]], na.rm = TRUE)
# mu[i, ]

proc_df <- function(x, nm) {
  nms <- which(colnames(x) %in% c("x", "y", "gti", nm, "w"))
  dat <- x[, nms]
  colnames(dat)[3:4] <- c("pv", "dv")
  return(dat)
}

nms <- colnames(mu)[match(c(snms, "dmu4", "dmu3"), colnames(mu))]
lm_mu <- lapply(nms, function(x) {
  dat <- proc_df(x = mu, nm = x)
  dlm <- lm(dv ~ poly(pv, degree = 2), weights = w, data = dat)
  return(list("model" = dlm, "dat" = dat))
})

lm_mu0 <- lapply(nms, function(x) {
  dat <- proc_df(x = mu0, nm = x)
  dat <- dat[!is.na(dat$pv), ]
  dlm <- lm(dv ~ poly(pv, degree = 2), weights = w, data = dat)
  return(list("model" = dlm, "dat" = dat))
})

lm_med <- lapply(nms, function(x) {
  dat <- proc_df(x = med, nm = x)
  dlm <- lm(dv ~ poly(pv, degree = 2), weights = w, data = dat)
  return(list("model" = dlm, "dat" = dat))
})

lm_med0 <- lapply(nms, function(x) {
  dat <- proc_df(x = med0, nm = x)
  dat <- dat[!is.na(dat$pv), ]
  dlm <- lm(dv ~ poly(pv, degree = 2), weights = w, data = dat)
  return(list("model" = dlm, "dat" = dat))
})
```

5.2.3 Plot the fitted LM models

```
mcap <- c("SA-LC", "GlobCover", "MODIS", "GeoWiki")
cols <- c("red", "orange3", "green4", "blue", "black", "grey50")
lw <- c(rep(1, 4), 3, 3)
lt <- c(rep(2, 4), 1, 1)
xl <- c(0, 100)
yl <- c(0, round(max(mu[, snms] / 10 )) * 10)

# mean cropland in districts
pdf(full_path(p_fig, "biases_magdist.pdf"), width = 7, height = 7)
par(oma = c(5, 0, 0, 0), mar = c(5, 3, 2, 3))
plot(xl, yl, pch = "", ylab = "% bias", xlab = "% cropland", xaxs = "i",
    mgp = c(1.5, 0.5, 0), yaxs = "i")
for(j in 1:length(snms)) {
  d <- mu[, c("gti", snms[j])]
  points(d[, 1], d[, 2], pch = 20, col = cols[j], cex = 0.5)
}
for(j in 1:length(lm_mu)) {
  dlm <- lm_mu[[j]]$model
  lines(predict(dlm, newdata = data.frame("pv" = 1:110)), col = cols[j],
        lwd = lw[j], lty = lt[j])
}
legend(x = 0, y = -5, mcap, pch = 20, col = cols, bty = "n", xpd = NA,
       title = "District means")
legend(x = 80, y = -5, c(mcap, "All 4", "No GlobCover"), lty = lt, lwd = lw,
       col = cols, bty = "n", xpd = NA, title = "Fit")
dev.off()

# mean non-zero cropland
yl <- c(0, round(max(mu0[, snms] / 10, na.rm = TRUE)) * 10)
pdf(full_path(p_fig, "biases_magdist_no0.pdf"), width = 7, height = 7)
par(oma = c(5, 0, 0, 0), mar = c(5, 3, 2, 3))
plot(xl, yl, pch = "", ylab = "% bias", xlab = "% cropland", xaxs = "i",
    mgp = c(1.5, 0.5, 0), yaxs = "i")
for(j in 1:length(snms)) {
  d <- mu0[, c("gti", snms[j])]
  points(d[, 1], d[, 2], pch = 20, col = cols[j], cex = 0.5)
}
for(j in 1:length(lm_mu0)) {
  dlm <- lm_mu0[[j]]$model
  lines(predict(dlm, newdata = data.frame("pv" = 1:110)), col = cols[j],
        lwd = lw[j], lty = lt[j])
}
legend(x = 0, y = -5, mcap, pch = 20, col = cols, bty = "n", xpd = NA,
       title = "District means")
legend(x = 80, y = -5, c(mcap, "All 4", "No GlobCover"), lty = lt, lwd = lw,
       col = cols, bty = "n", xpd = NA, title = "Fit")
dev.off()

# median cropland
yl <- c(0, round(max(med[, snms] / 10, na.rm = TRUE)) * 10)
pdf(full_path(p_fig, "biases_magdist_med.pdf"), width = 7, height = 7)
par(oma = c(5, 0, 0, 0), mar = c(5, 3, 2, 3))
plot(xl, yl, pch = "", ylab = "% bias", xlab = "% cropland", xaxs = "i",
    mgp = c(1.5, 0.5, 0), yaxs = "i")
for(j in 1:length(snms)) {
  d <- med[, c("gti", snms[j])]
  points(d[, 1], d[, 2], pch = 20, col = cols[j], cex = 0.5)
}
for(j in 1:length(lm_med)) {
  dlm <- lm_med[[j]]$model
  lines(predict(dlm, newdata = data.frame("pv" = 1:110)), col = cols[j],
        lwd = lw[j], lty = lt[j])
}
legend(x = 0, y = -5, mcap, pch = 20, col = cols, bty = "n", xpd = NA,
       title = "District means")
legend(x = 80, y = -5, c(mcap, "All 4", "No GlobCover"), lty = lt, lwd = lw,
       col = cols, bty = "n", xpd = NA, title = "Fit")
dev.off()

# median of non-zero cropland
yl <- c(0, round(max(med0[, snms] / 10, na.rm = TRUE)) * 10)
pdf(full_path(p_fig, "biases_magdist_med0.pdf"), width = 7, height = 7)
par(oma = c(5, 0, 0, 0), mar = c(5, 3, 2, 3))
plot(xl, yl, pch = "", ylab = "% bias", xlab = "% cropland", xaxs = "i",
    mgp = c(1.5, 0.5, 0), yaxs = "i")
for(j in 1:length(snms)) {
  d <- med0[, c("gti", snms[j])]
  points(d[, 1], d[, 2], pch = 20, col = cols[j], cex = 0.5)
}
for(j in 1:length(lm_med0)) {
  dlm <- lm_med0[[j]]$model
  lines(predict(dlm, newdata = data.frame("pv" = 1:110)), col = cols[j],
        lwd = lw[j], lty = lt[j])
}
legend(x = 0, y = -5, mcap, pch = 20, col = cols, bty = "n", xpd = NA,
       title = "District means")
legend(x = 80, y = -5, c(mcap, "All 4", "No GlobCover"), lty = lt, lwd = lw,
       col = cols, bty = "n", xpd = NA, title = "Fit")
dev.off()
```

5.2.4 Check for spatial autocorrelation in residuals

```
dists <- as.matrix(dist(mu[, c("x", "y")]))
dists_inv <- 1 / dists
diag(dists_inv) <- 0
for(i in 1:length(snms)) {
  print(Moran.I(residuals(lm_mu[[i]]$model), dists_inv)$p.value)
}

dists <- as.matrix(dist(lm_mu0[[1]]$dat[, c("x", "y")]))
dists_inv <- 1 / dists
diag(dists_inv) <- 0
for(i in 1:length(snms)) {
  print(Moran.I(residuals(lm_mu0[[i]]$model), dists_inv)$p.value)
}
```

Present and significant. To deal with it, first try to use:

5.2.5 A generalized least squares (GLS) with spatial autocovariance terms.

```
# mean of all data
i <- "geow"  # geowiki
dat <- proc_df(x = mu, nm = i)
dgls <- gls(dv ~ poly(pv, 2), weights = ~1 / w, data = dat)  # orthogonal polys
dgls_exp <- update(dgls, correlation = corExp(form = ~x + y, nugget = TRUE))
plot(Variogram(dgls, form = ~x+y, resType = "n"))
plot(Variogram(dgls_exp, form = ~x+y, resType = "n"))  # funny shape remains

# cropland only mean
dat <- proc_df(x = mu0, nm = i)
dat <- dat[!is.na(dat$pv), ]
dgls <- gls(dv ~ poly(pv, 2), weights = ~1 / w, data = dat)  # orthogonal polys
dgls_exp <- update(dgls, correlation = corExp(form = ~x + y, nugget = TRUE))
plot(Variogram(dgls, form = ~x+y, resType = "n"))
plot(Variogram(dgls_exp, form = ~x+y, resType = "n"))  # funny shape remains
```

That seemed inadequate for removing autocorrelation in residuals. The GLS did not seem to properly deal with residual autocorrelation, probably because the spatial structure is complex, given the pattern of rainfall and cropland in it. A better approach is to include coordinates directly in the model, leveraging the 2-D smoothing capabilities of the GAM model in the mgcv library.

5.2.6 Generalized additive model

First check fit of model family, and whether assumptions of normally distributed errors are met.

```
# stats.stackexchange.com/questions/99425/distribution-for-percentage-data
# quartz()
# dat <- mu0[, "sa30"]
dat <- mu[, "sa30"]
dat <- dat[!is.na(dat)]
```

```r
dat[dat == 0] <- dat[dat == 0] + 0.1
descdist(dat, discrete = FALSE)  # check potential distributions

# Check what distributions work best on data (non-cropland removed)
dists <- c("norm", "lnorm", "exp", "cauchy", "gamma", "logis", "beta","weibull")
fit_distrs <- lapply(snms, function(x) {
  dat <- mu0[, x]
  dat <- dat[!is.na(dat)] / 100
  dat[dat == 0] <- dat[dat == 0] + 0.1
  fits <- lapply(dists, function(x) fitdist(dat, x))
  names(fits) <- dists
  fits
})
names(fit_distrs) <- snms
# lapply(fit_distrs, function(x) sort(sapply(x, function(y) y$aic)))
plot(fit_distrs[[2]]$lnorm)  # lognorm shows up as best fit

# Matters more on residuals from model fit, however.
# dataset
dat <- proc_df(x = mu, nm = snms[4])  # interactively change response
# dat <- proc_df(x = med0, nm = snms[1])  # and input data
dat <- dat[!is.na(dat$pv), ]
dat[, 4] <- dat[, 4] + 0.1  # add 1 to allow for log transform

# GAM assuming Gaussian
dgam <- gam(dv ~ poly(pv, 2) + s(x, y), weights = w, data = dat,
            method = "ML")
summary(dgam)
AIC(dgam)
gam.check(dgam)
summary(fitdist(residuals(dgam), "norm"))
plot(fitdist(residuals(dgam), "norm"))

# GAM Gaussian with log link
dgam2 <- gam(dv ~ poly(pv, 2) + s(x, y), weights = w, data = dat,
            family = gaussian(link = "log"), method = "ML")
summary(dgam2)
AIC(dgam2)
gam.check(dgam2)
summary(fitdist(residuals(dgam2), "norm"))
plot(fitdist(residuals(dgam2), "norm"))

# GAM w/Gamma
dgam3 <- gam(dv ~ poly(pv, 2) + s(x, y), weights = w, data = dat,
            family = Gamma(link = "identity"), method = "ML")
summary(dgam3)
AIC(dgam3)  # bizarre large AIC
gam.check(dgam3)  # crazy deviance residuals
summary(fitdist(residuals(dgam3), "norm"))
plot(fitdist(residuals(dgam3), "norm"))

# GAM w/log
dgam4 <- gam(log(dv) ~ poly(pv, 2) + s(x, y), weights = w, data = dat,
             method = "ML")
summary(dgam4)
AIC(dgam4)
gam.check(dgam4)
summary(fitdist(residuals(dgam4), "norm"))
plot(fitdist(residuals(dgam4), "norm"))
# termplot(dgam4, terms = "poly(pv, 2)", ask = FALSE)
# pred <- predict(dgam4, newdata = newdat, type = "response", se.fit = TRUE)

# above, with smooth fit instead of quadratic
dgam4b <- gam(log(dv + 0.0001) ~ s(pv, k = 3) + s(x, y), weights = w,
              data = dat, method = "ML")
plot(dat$pv, dat$dv)
summary(dgam4b)
AIC(dgam4b)
gam.check(dgam4)
summary(fitdist(residuals(dgam4b), "norm"))
plot(fitdist(residuals(dgam4b), "norm"))

plot.gam(dgam4b, residuals = TRUE, select = 1, se = TRUE, all.terms = TRUE,
         pages = 0)
```

Testing each of the main landcover sets for model shape, the best fit in terms of AIC and assumption of normality of residuals is from a log-normal model.

5.2.7 Plot final results

Helper functions

```
predfunc <- function(model, newdat) {
  pred <- predict(model, newdata = newdat, type = "response", se.fit = TRUE)
  pdat <- transform(newdat, fit = pred$fit)
  pdat <- transform(pdat, up = fit + (1.96 * pred$se.fit),
                    lo = fit - (1.96 * pred$se.fit))
  return(pdat)
}

termfunc <- function(model, newdat) {
  pred <- predict(model, newdata = newdat, type = "terms", se.fit = TRUE)
  pdat <- transform(newdat, fit = pred$fit[, 1] + coef(model)[1])
  pdat <- transform(pdat, up = fit + (1.96 * pred$se.fit[, 1]),
                    lo = fit - (1.96 * pred$se.fit[, 1]))
  return(pdat)
}

polyfunc <- function(pdat) {
  coord <- cbind.data.frame("x" = c(rev(pdat$pv), pdat$pv),
                            "y" = c(rev(pdat$lo), pdat$up))
}
```

```
gam_mods <- lapply(nms, function(i) {
  dat1 <- proc_df(x = mu0, nm = i)
  dat1 <- dat1[!is.na(dat1$pv), ]
  dat1$dv <- dat1$dv + 0.1  # adding 0.1 to allow log transform
  dat2 <- proc_df(x = mu, nm = i)
  dat2 <- dat2[!is.na(dat2$pv), ]
  dat2$dv <- dat2$dv + 0.1  # adding 0.1 to allow log transform
  dgam1 <- gam(log(dv) ~ poly(pv, 2) + s(x, y), weights = w, data = dat1,
               method = "ML")
  dgam2 <- gam(log(dv) ~ poly(pv, 2) + s(x, y), weights = w, data = dat2,
               method = "ML")
  newdat <- with(dat1,
                 data.frame(pv = 0:100, "x" = mean(dat1$x), "y" = mean(dat1$y)))
  pdat1 <- predfunc(dgam1, newdat)
  pdat2 <- predfunc(dgam2, newdat)
  # pdat1 <- termfunc(dgam1, newdat)
  # pdat2 <- termfunc(dgam2, newdat)
  return(list("mumod" = dgam1, "medmod" = dgam2, "pdat1" = pdat1,
              "pdat2" = pdat2,  "data1" = dat1, "data2" = dat2))
})
```

5.2.7.1 GAM fits

```
# plot(xl, yl2, pch = "", ylab = "% bias)", xlab = "% cropland", xaxs = "i",
#      yaxs = "i", yaxt = "n")
# round(log(seq(1, 90, 10)), 2)
# exp(seq(0, 4.5, 0.5))

ptcols <- sapply(cols, function(x) makeTransparent(x, 80))  # CI colors
lw <- c(rep(2, 4), 3, 3)
yl1 <- c(0, round(max(sapply(gam_mods, function(x) max(x$pdat1$fit) * 1.1)), 1))
yl2 <- c(0, round(max(sapply(gam_mods, function(x) max(x$pdat2$fit) * 1.1)), 1))
pdf(full_path(p_fig, "biases_md_lnorm_gam_mu0.pdf"), width = 7, height = 4.5)
par(oma = c(0, 0, 0, 0), mar = c(3, 3, 2, 2), mgp = c(1.5, 0.5, 0))
plot(xl, yl1, pch = "", ylab = "Mean Absolute Error (%)",
     xlab = "% Cropland", xaxs = "i", mgp = c(1.5, 0.5, 0), yaxs = "i",
     yaxt = "n")
axis(2, at = 0:4, round(exp(c(-10, 1:4))))
# for(i in 1:length(gam_mods)) {
for(i in 1:4) {
  m <- gam_mods[[i]]
  p <- polyfunc(m$pdat1)
  polygon(p$x, p$y, col = ptcols[i], border = NA)
  lines(fit ~ pv, data = m$pdat1, col = cols[i], lwd = 1.5, lty = 1)
}
# legend("topleft", c(mcap, "All 4", "No GlobCover"), lty = 1, lwd = 1.5,
#        col = cols, bty = "n", xpd = NA)#, title = "Fit")
legend("topleft", mcap, lty = 1, lwd = 1.5,
       col = cols, bty = "n", xpd = NA)#, title = "Fit")
```

```r
dev.off()

pdf(full_path(p_fig, "biases_md_lnorm_gam_mu.pdf"), width = 7, height = 4.5)
par(oma = c(0, 0, 0, 0), mar = c(3, 3, 2, 2), mgp = c(1.5, 0.5, 0))
plot(xl, yl2, pch = "", ylab = "Mean Absolute Error (%)",
     xlab = "% Cropland", xaxs = "i", mgp = c(1.5, 0.5, 0), yaxs = "i",
     yaxt = "n")
axis(2, at = 0:4, round(exp(c(-10, 1:4))))
#for(i in 1:length(gam_mods)) {
for(i in 1:4) {
  m <- gam_mods[[i]]
  p <- polyfunc(m$pdat2)
  polygon(p$x, p$y, col = ptcols[i], border = NA)
  lines(fit ~ pv, data = m$pdat2, col = cols[i], lwd = 1.5, lty = 1)
}
# legend("topleft", c(mcap, "All 4", "No GlobCover"), lty = 1, lwd = 1.5,
#        col = cols, bty = "n", xpd = NA)#, title = "Fit")
legend("topleft", mcap, lty = 1, lwd = 1.5,
       col = cols, bty = "n", xpd = NA)#, title = "Fit")
dev.off()

# Check p-values for all model terms
lapply(gam_mods, function(x) {
  m1 <- summary(x$mumod)
  m2 <- summary(x$medmod)
  rbind("mu" = round(c(m1$p.table[, 4], m1$s.pv), 3),
        "med" = round(c(m2$p.table[, 4], m2$s.pv), 3))
})  # all terms significant at p < 0.001 except 2nd order term on GlobCov med
```

5.2.7.2 GAM plots

5.2.7.3 Final plot of magisterial districts

For supplemental.

```r
provs <- unique(md$ID_1)
greys <- paste0("grey", seq(15, 95, 10))
pdf(full_path(p_fig, "md_map.pdf"), width = 7, height = 7, bg = "transparent")
par(mar = rep(0, 4))
plot(sa.shp, lty = 0)
for(i in 1:length(provs)) {
  plot(md[md$ID_1 == provs[i], ], col = greys[i], #border = "transparent",
       add = TRUE, lwd = 1)
}
dev.off()
```

6 Error, bias, and accuracy in carbon stock estimates

Based on the Ruesch & Gibbs (2008) approach for estimating carbon density.

6.1 Data

```r
library(raster)
library(rgdal)
library(lmisc)
library(data.table)
library(SAcropland)
library(RColorBrewer)
library(xtable)

# Paths
p_root <- proj_root("SAcropland")
p_fig <- full_path(p_root, "SAcropland/paper/figures/")
p_data <- full_path(p_root, "SAcropland/external/ext_data/")
p_data2 <- full_path(p_root, "SAcropland/data/")
p_carb <- full_path(p_root, "SAcropland/external/ext_data/carbon/")
```

6.1.1 Carbon values

Starting with the carbon look-up tables used for Africa by Ruesch & Gibbs (2008), inputs downloaded from here.

```r
cfiles <- dir(p_carb, pattern = "m6", full.names = TRUE)
nms <- gsub("*.*m6|\\.txt", "", cfiles)
cval_list <- lapply(1:length(cfiles), function(x) {
  tab <- read.table(cfiles[x])
  tab <- tab[ c(1, 3)]
  colnames(tab) <- c("CL", nms[x])
  tab
})

# merge into single carbon table
mergefun <- function(x, y) merge(x, y, by = "CL", all.x = TRUE, all.y = TRUE)
ctab <- Reduce(mergefun, cval_list)

# Read in lookup table key
key <- readLines(dir(p_carb, pattern = "key", full.names = TRUE))[25:44]
key2 <- readLines(dir(p_carb, pattern = "key", full.names = TRUE))[47:57]
mtch <- sapply(gregexpr("[0-9]", key2), max)
lcs <- sapply(1:length(key2), function(x) substr(key2[x], 1, mtch[x]))
lcs <- gsub(" & ", ",", gsub("-", ":", lcs))

# Reshape and reduce according to carbon classes
ctab2 <- t(data.frame(sapply(lcs, function(x) {
  ind <- eval(parse(text = paste0("c(", x, ")")))
  as.numeric(as.vector(
    round(colMeans(ctab[ctab$CL %in% ind, -1], na.rm = TRUE))))
})))
rownames(ctab2) <- 1:nrow(ctab2)
colnames(ctab2) <- nms
ctab2 <- ctab2 * 0.01  # convert to tons/ha from 1000 kg/ha
```

```
ctab2 <- cbind(transform(lcs), ctab2)
colnames(ctab2)[1] <- "class"
```

6.1.2 Further compress classes that have the same carbon value.

- 1:3, 6:8 => 1 (broadleaf and mixed forests)
- 4:5 => Drop
- 9, 10; 17 => 2 (Secondary forests, forest/cropland mosaic)
- 11, 12, 15 => 3 (shrublands)
- 20:23 => drop (water, snow, artificial surfaces)
- 19 => drop (bare areas)

```
ctabf <- round(rbind(colMeans(ctab2[c(1, 3), 2:ncol(ctab2)], na.rm = TRUE),
                     colMeans(ctab2[10:11, 2:ncol(ctab2)], na.rm = TRUE),
                     ctab2[4:6, 2:ncol(ctab2)]))

# Read in ecofloristic regions and calculate their areas for Africa
ecoflora <- readOGR("external/ext_data/africa_ecofloristic_zones.sqlite",
                    layer = "africa_ecofloristic_zones")
ecoflora@proj4string <- CRS("+proj=longlat +datum=WGS84 +no_defs")
afalb <- paste0("+proj=aea +lat_1=20 +lat_2=-23 +lat_0=0 +lon_0=25 +x_0=0 ",
                "+y_0=0 +ellps=WGS84 +towgs84=0,0,0,0,0,0,0 +units=m +no_defs")
ecoflora_alb <- spTransform(ecoflora, CRS = CRS(afalb))
ecoflora_num <- cbind.data.frame(unique(ecoflora_alb@data),
                                 "ID" = c("06", "08", "09", "15", "NA", "18",
                                          "20", "16", "17", "WA", "19", "07"),
                                 stringsAsFactors = FALSE)
ecoflora_alb$ID <- rep("NA", nrow(ecoflora_alb))
head(ecoflora_alb)
ecoflora_alb$ID <- ecoflora_num[match(ecoflora_alb$gez_term,
                                      ecoflora_num$gez_term), "ID"]
# ecoflora_alb@data <- sp::merge(ecoflora_alb@data, ecoflora_num,
#                                by = "gez_term", sort = FALSE, keep = TRUE)
ecoflora_alb@data[sample(1:nrow(ecoflora_alb@data), 2), ]
ecoflora_alb$area <- round(rgeos::gArea(ecoflora_alb, byid = TRUE) / 1000000, 1)
ecoareas <- sapply(ecoflora_num$ID, function(x) {
  sum(ecoflora_alb@data[ecoflora_alb$ID == x, "area"])
})
ecoareas <- ecoareas[!names(ecoareas) %in% c("WA", "NA")]
ecowgts <- ecoareas / sum(ecoareas)
ecowgts <- ecowgts[sort(names(ecowgts))]

# par(mar = rep(0, 4))
# plot(ecoflora_alb)
# plot(ecoflora_alb[ecoflora_alb$ID == "17", ], col = "red", add = TRUE)

i <- 1:ncol(ctabf)
ctabf$mu <- round(sapply(1:nrow(ctabf), function(x) {
  sum(ctabf[x, ] * ecowgts, na.rm = TRUE)
}), 1)
ctabf <- cbind(ctabf, t(apply(ctabf[, i], 1, function(x) range(x, na.rm=TRUE))))
colnames(ctabf)[(ncol(ctabf) - 1):ncol(ctabf)] <- c("min", "max")
```

```r
rownames(ctabf) <- 1:nrow(ctabf)
LC <- c("forest", "second", "shrub", "grass", "sparse")
ctabo <- data.frame(t(ctabf[, c("mu", "min", "max")]))  # reclass table
colnames(ctabo) <- LC
```

6.2 Analysis

6.2.1 Prepare raster data

```r
load(full_path(p_data, "d_grid_act.rda"))  # actual diffence grids
paths <- full_path(p_data, dir(p_data, paste0("cover*.*11sum_mask.tif$")))
gti <- raster(paths) / 100  # gti 2011
namask <- raster(full_path(p_data, "namask.tif"))  # NA mask
gti <- mask(gti, namask)  # apply mask to GTI
# cellStats(!is.na(gti), sum)

# Reconstruct original landcover estimates
snms <- c("sa30", "globmu", "modmu", "geow")
dlist_1km <- lapply(dlist_act[snms], function(x) x$f1$g2011)
lc_list <- lapply(dlist_1km, function(x) gti - x / 100)
plot(lc_list[[1]])
# cellStats(!is.na(lc_list[[1]]), sum)
plot(gti)
# tst <- raster("external/ext_data/geowikisa_masked.tif")  # check
# plot(round(tst / 100 - lc_list[[4]], 4)) # okay

# aggregate rasters
fact <- c(5, 10, 25, 50, 100)
lc_agg <- aggregate_rast_list(fact, lc_list)   # landcover rasters
gti_agg <- aggregate_rast_list(fact, list("gti" = gti))  # GTI rasters

# need NA mask for weighting of aggregated value
# sumna <- function(x) sum(x, na.rm = FALSE)
# namask <- calc(stack(stack(gti), stack(lclist)), sumna)
# namask[namask > 0] <- 1
namask2 <- !is.na(namask)
# namask <- raster(full_path(p_data, "namask.tif")) # load in NA mask
# namask2 <- !is.na(namask)  # set NAs to zero
area_wgts <- aggregate_rast_list(fact, list(namask2), fun = sum)

# cropland cover bins, for looking at error as a function of cover
binv <- seq(0, 1, 0.05)
gti_bins <- lapply(gti_agg, function(x) {
  cut(x$gti, breaks = binv, include.lowest = TRUE)
})
```

6.2.2 Create cropland carbon estimates

```r
# Apply carbon estimates
# Function for pixel-wise carbon density from crop & non-crop fractions
```

```r
carbon <- function(fcrop, cropC, noncropC) {
  carb <- fcrop * cropC + noncropC * (1 - fcrop)
  return(round(carb, 2))
}

cc <- ctab2[ctab2$class == 16, 2]
gti_c <- lapply(gti_agg, function(x) {
  r <- x$gti
  s <- stack(lapply(1:ncol(ctabo), function(y) {
    carbon(r, cc, ctabo[1, y])
  }))
  names(s) <- colnames(ctabo)
  s
})  # gti

lc_c <- lapply(lc_agg, function(x) {
  lc <- lapply(x, function(j) {
    s <- stack(lapply(1:ncol(ctabo), function(y) {
      carbon(j, cc, ctabo[1, y])
    }))
    names(s) <- colnames(ctabo)
    s
  })
  names(lc) <- names(x)
  lc
})  # landcover datasets

# checks--right rasters being referenced?
tst <- cbind(sample(1:5, 10, replace = TRUE),
             sample(1:4, 10, replace = TRUE),
             sample(1:5, 10, replace = TRUE))
sapply(1:nrow(tst), function(i) {
  cellStats(lc_c[[tst[i, 1]]][[tst[i, 2]]][[tst[i, 3]]] -
            carbon(lc_agg[[tst[i, 1]]][[tst[i, 2]]], cc, ctabo[1, tst[i, 3]]),
            sum)
}) # should be all zeroes
```

6.2.3 Difference the carbon datasets

```r
# percent difference
pct_diff <- function(x, y) (x - y) / x * 100
c_pct_diff <- lapply(1:length(gti_c), function(x) {
  dif <- lapply(1:length(lc_c[[x]]), function(y) {
    s <- stack(lapply(1:nlayers(lc_c[[x]][[y]]), function(z) {
      p <- pct_diff(gti_c[[x]][[z]], lc_c[[x]][[y]][[z]])
    }))
    names(s) <- colnames(ctabo)
    s
  })
  names(dif) <- names(lc_c[[x]])
  dif
})
```

```
names(c_pct_diff) <- names(lc_c)

# checks - right rasters being referenced
tst <- cbind(sample(1:5, 10, replace = TRUE),
             sample(1:4, 10, replace = TRUE),
             sample(1:5, 10, replace = TRUE))
sapply(1:nrow(tst), function(i) {
  x <- gti_c[[tst[i, 1]]][[tst[i, 3]]]
  y <- lc_c[[tst[i, 1]]][[tst[i, 2]]][[tst[i, 3]]]
  z <- c_pct_diff[[tst[i, 1]]][[tst[i, 2]]][[tst[i, 3]]]
  cellStats(z - ((x - y) / x * 100), sum)
})  # zeroes

# Then for a map plot figure, calculate the mean pixel-wise percent difference
c_pct_diff_mu <- lapply(c_pct_diff, function(x) {
  lapply(x, function(y) calc(y, mean))
})

# # disggregate selected rasters at selected levels for plotting
# namask <- raster(full_path(p_data, "namask.tif")) # load in NA mask
lev <- names(c_pct_diff_mu)
disagg <- lapply(snms, function(x) {
  l1 <- lapply(lev, function(y) {
    if(y == "f1") {
      r <- c_pct_diff_mu[[y]][[x]]
    } else {
      r <- raster::disaggregate(c_pct_diff_mu[[y]][[x]],
                                fact = as.numeric(gsub("f", "", y)))
      r <- raster::mask(crop(r, namask), namask)
    }
  })
  named_out(l1, lev)
})
names(disagg) <- snms

stats <- lapply(disagg, function(x) {
  sapply(x, function(y) {
    c(cellStats(y, mean), quantile(y, seq(0, 1, 0.05)))
  })
})
```

6.3 Outputs

6.3.1 Carbon error maps

```
load(full_path(p_data, "MZshapes.Rdata"))  # SA shape

# Plotting colors
lims <- c(ceiling(min(sapply(stats, function(x) x[3, ]))),
          floor(max(sapply(stats, function(x) x[21, ]))))
rng <- range(sapply(stats, range))
```

```
brks <- c(rng[1], lims[1], -45, -20, -10, -5, -1, 1, 5, 10, lims[2], rng[2])
#n_cols <- length(brks) - 1
colsall <- brewer.pal(n = 11, "Spectral")
cols <- c(colsall[1:6], "grey80", colsall[c(7, 9, 10, 11)])
#colorRampPalette(c("red", "tan", "grey80"))
#cols2 <- colorRampPalette(c("grey80", "green4", "blue"))
#cols <- c(cols1(8), c(cols2(5)))

# brks <- c(-100.1, -75, -50, -25, -10, 0, 10, 25, 50, 75, 100.1)
# n_cols <- length(brks) - 1
legtext <- "% Difference"
cx <- 1.4
lcol <- "black"
mcap <- c("SA-LC", "GlobCover", "MODIS", "GeoWiki")
lev <- names(disagg[[1]])[-c(2:3)]
lev2 <- c("1 km", "25 km", "50 km", "100 km")
pdf(full_path(p_fig, "carbon_bias_map.pdf"), height = 6, width = 7)
par(mfrow = c(4, 4), mar = c(0, 0, 0, 0), oma = c(5, 5, 2, 0))
for(i in 1:length(snms)) {
  print(snms[i])
  for(j in 1:length(lev)) {
    print(lev[j])
    plot(sa.shp, lty = 0)
    plot(disagg[[snms[i]]][[lev[j]]], add = TRUE, col = cols,
         breaks = brks, legend = FALSE)
  if(j == 1) mtext(mcap[i], side = 2, line = 1, cex = cx)
  if(i == 1) mtext(lev2[j], side = 3, line = 0, cex = cx)
  }
}
flex_legend(ncuts = length(brks) - 1, legend.text = legtext,
            legend.vals = round(brks),
            longdims = c(0.2, 0.8), shortdims = c(0.06, 0.01),
            colvec = cols, #(length(brks) - 1),
            srt = c(270, 0), horiz = TRUE, textside = "bottom",
            legend.pos = c(4, 5), leg.adj = list(c(0.25, 0), c(0, -0.5)),
            cex.val = cx, textcol = lcol, bordercol = lcol)
dev.off()
```

6.3.2 Carbon bias/accuracy

```
lev_vec <- names(c_pct_diff)
# created mask for non-cropland areas, unioning GTI and each LC, filtering out
# areas of no-cropland (<1/2% total cover)
lc_union <- lapply(lev_vec, function(x) {
  lcb <- lapply(snms, function(y) {
    gti_gt0 <- Which(round(gti_agg[[x]]$gti * 100) > 0)
    lc_gt0 <- Which(round(lc_agg[[x]][[y]] * 100) > 0)
    all_gt0 <- gti_gt0 + lc_gt0
    all_gt0[all_gt0 > 0] <- 1
    all_gt0
  })
  named_out(lcb, snms)
})
names(lc_union) <- lev_vec
# plot(lc_union$f50$globmu)
# plot(gti_agg$f50$gti)

# calculate for whole country
tareas <- lapply(area_wgts, function(x) x[[1]] * 100)
gti_ctot <- sapply(1:length(gti_c), function(x) {
  cellStats(gti_c[[x]] * tareas[[x]], sum)
})  # gti total carbon stock
lc_c2 <- lapply(snms, function(x) {
  sapply(names(lc_c), function(y) lc_c[[y]][[x]])
})  # reshape lc_c2 -> landcover in outer list
names(lc_c2) <- snms
lc_ctot <- lapply(lc_c2, function(x) {
  sapply(1:length(x), function(y) {
    cellStats(x[[y]] * tareas[[y]], sum)
  })
})  # landcover total carbon stocks
totc_cntry <- lapply(lc_ctot, function(x) {
  stats <- (gti_ctot - x) / gti_ctot * 100
  colnames(stats) <- names(lc_c2$sa30)
  rownames(stats) <- names(lc_c2$sa30$f1)
  stats
})  # country-level percent differences: gti versus landcover
names(totc_cntry) <- snms

# for cropped areas only - note here we are masking on cropland fraction only,
# not on union of gti and each landcover map, which is needed below for mean
# bias estimates
tot_area <- sum(freq(namask)[1, 2])  # sum of just the non-NA area
crop_areas <- lapply(lc_union$f1, freq)
sapply(crop_areas, function(x) x[2, 2] / tot_area)  # 29, 53, 33, 31

lc_ag2 <- lapply(snms, function(x) {
  sapply(names(lc_agg), function(y) lc_agg[[y]][[x]])
})  # reshape lc_ag2 -> landcover fractions in outer list
gti_ctot2 <- sapply(1:length(gti_c), function(x) {
  msk <- gti_agg[[x]]$gti > 0.005  # farmland > 0.05% mask
  # msk <- gti_agg[[x]]$gti > 0.05  # farmland > 0.05% mask
  msked <- mask(gti_c[[x]] * tareas[[x]], msk, maskvalue = 0)
  cellStats(msked, sum)
})  # gti
lc_ctot2 <- lapply(1:length(lc_c2), function(x) {
  xx <- lc_c2[[x]]  # recycle reshaped lc_c2 list
  jj <- lc_ag2[[x]]  # recycle reshaped lc_c2 list
  sapply(1:length(xx), function(y) {
    msk <- jj[[y]] > 0.005  # farmland > 0.05% mask
    # msk <- jj[[y]] > 0  # farmland > 0.05% mask
    msked <- mask(xx[[y]] * tareas[[y]], msk, maskvalue = 0)
    cellStats(msked, sum)
  })
```

```r
})  # landcover carbon estimates

# but discrepancy will only be relevant at 1 km scale, because the carbon total
# keeps increasing when cropland areas are the only ones being considered
totc_crop <- lapply(lc_ctot2, function(x) {
  stats <- (gti_ctot2 - x) / gti_ctot2 * 100
  colnames(stats) <- names(lc_c2$sa30)
  rownames(stats) <- names(lc_c2$sa30$f1)
  stats
})
names(totc_crop) <- snms

# Combine tables for output to supplementals, 1 km % differences for country
# and agricultural levels
totc_out <- rbind(t(sapply(totc_cntry, function(x) x[, 1])),
                  t(sapply(totc_crop, function(x) x[, 1])))
pnms <- c("Forest", "Secondary", "Shrubland", "Grassland", "Sparse")
knms <- rep(mcap, 2)
totc_out <- cbind.data.frame(knms, round(unname(totc_out), 2))
colnames(totc_out) <- c("Map", pnms)
totc_out <- cbind(Region = c(rep("Country", 4), rep("Agricultural", 4)),
                  totc_out)
caption <- paste("Percent differences in total carbon stock estimates",
                 "calculated from the reference maps and from each of the four",
                 "cropland maps. Differences are evaluated for total carbon",
                 "estimates either at the country scale or over just the",
                 "agricultural regions (cropland $>$0.05\\%), using",
                 "the carbon densities of 5 different cover types to provide",
                 "the values for the non-agricultural portions of each pixel",
                 "(cover types indicated by column names).")
totc_xtab <- xtable(totc_out, caption = caption, digits = 1)
print(totc_xtab, type = "latex", caption.placement = "top",
      file = "paper/figures/totC-bias.tex", include.rownames = FALSE)
```

6.3.2.1 Calculate how much country-level carbon estimates differ between datasets

6.3.3 Calculate bias/MAE statistics

6.3.3.1 Primary method Weighted by **actual** cropland percentage, and then, with aggregation, by number of pixels being aggregated.

```r
# Helper functions to pass into data.table
# not used - switch on if quantiles needs
# bfn <- function(x, y) {
#   box_stats(x, weighted = TRUE, weight.opts = list("weights" = y))
# }
# bfna <- function(x, y) {
#   box_stats(abs(x), weighted = TRUE, weight.opts = list("weights" = y))
# }

wm <- function(x, w) stats::weighted.mean(x, w)
wma <- function(x, w) stats::weighted.mean(abs(x), w)
```

```r
# Bias
cb_statsw_mu <- rbindlist(lapply(lev_vec, function(x) {
  il <- rbindlist(lapply(snms, function(y) {
    ref <- gti_agg[[x]][[1]]
    awgts <- area_wgts[[x]][[1]]
    rerror <- c_pct_diff[[x]][[y]]
    bstats <- bias_statsw(ref, awgts, rerror, LC, wm, "Bias", rnd = 2,
                          rweight = TRUE, aweight = TRUE, trim_wgt = TRUE)
    cbind("map" = y, bstats)
  }))
  # named_out(ol, snms)
  cbind("ol" = x, il)
}))

# calculate mean across cover types
cb_statsw_mu <- cbind(cb_statsw_mu,
                      "All" = apply(cb_statsw_mu[, LC, with = FALSE], 1, mean))

# Accuracy
cb_statsw_mua <- rbindlist(lapply(lev_vec, function(x) { # x <- "f1"
  il <- rbindlist(lapply(snms, function(y) { # y <- "sa30"
    ref <- gti_agg[[x]][[1]]
    awgts <- area_wgts[[x]][[1]]
    rerror <- c_pct_diff[[x]][[y]]
    bstats <- bias_statsw(ref, awgts, rerror, LC, wma, "MAE", rnd = 2,
                          rweight = TRUE, aweight = TRUE, trim_wgt = TRUE)
    cbind("map" = y, bstats)
  }))
  # named_out(ol, snms)
  cbind("ol" = x, il)
}))

# calculate mean across cover types
cb_statsw_mua <- cbind(cb_statsw_mua,
                       "All" = apply(cb_statsw_mua[, LC, with=FALSE], 1, mean))
# rowMeans(as.data.frame(cb_statsw_mua[1, ])[, LC])

# check error stats - do they match alternate approaches?
for(i in c("f1", "f25", "f10")) { # i <- "f1"
  for(j in c("sa30", "modmu", "geow")) { # j <- "modmu"
    for(k in c("forest", "shrub", "sparse", "second")) {  # k <- "second"
      print(paste("cross-checking calculations in", i, j, k))
      ref <- gti_agg[[i]][[1]]
      awgts <- area_wgts[[i]][[1]]
      rerror <- c_pct_diff[[i]][[j]]
      a1 <- bias_statsw(ref, awgts, rerror, LC, wm, "mu",
                        aweight = FALSE)[, get(k)]
      a2 <- bias_statsw(ref, awgts, rerror, LC, wm, "mu",
                        rweight = FALSE, aweight = FALSE)[, get(k)]
      a3 <- bias_statsw(ref, awgts, rerror, LC, wm, "mu")[, get(k)]
      c1 <- getValues(c_pct_diff[[i]][[j]][[k]])
      w1 <- getValues(gti_agg[[i]][[1]])
      w2 <- getValues(area_wgts[[i]][[1]])
      print("...non-area weighted mean matches?")
      print(a1 == round(weighted.mean(c1, w1, na.rm = TRUE), 2))
      print("...totally unweighted mean matches?")
      print(a2 == round(cellStats(rerror[[k]], mean), 2))
      print("...double weighted mean matches?")
      print(a3 == round(weighted.mean(c1, w1 * w2, na.rm = TRUE), 2))
    }
  }
}
```

```
# selection variables
nms1 <- c(colnames(ctabo), "wgt")
nms2 <- colnames(ctabo)
# bvals <- names(box_stats(sample(1:100, 200, replace = TRUE)))
bvals <- "mu"

# check constancy of < 1/2% cropland being excluded
areas <- sapply(lc_agg, function(x) res(x[[1]])[1]^2 / 10000)
plot(areas, (areas * 0.005))  # scales
(areas * 0.005)  # 1/2 to 5000 ha

dang <- Sys.time()
  #print(paste("..", x))
cb_stats <- lapply(lev_vec, function(x) {  # level
  #x <- lev_vec[4]
  print(paste(".", x))
  #levr <- names(c_pct_diff[[x]])
  l1 <- lapply(names(c_pct_diff[[x]]), function(y) {
    #y <- levr[1]
    print(paste("...", y))
    lc <- c_pct_diff[[x]][[y]]
    lcmask <- lc_union[[x]][[y]]

    # stack raster bins, cropland bins, and landcover set
    s <- stack(list("bin" = gti_bins[[x]], "wgt" = area_wgts[[x]][[1]],
                    "mask" = lc_union[[x]][[y]], c_pct_diff[[x]][[y]]))
    DT <- na.omit(as.data.table.raster(s))
    setkey(DT, "bin")

    # Potentially useful material deleted here: check repo prior to 17/10 if
    # needed

    fr1 <- data.table("bvals" = c("m", "m0", "ma", "ma0"))
    fr2 <- data.table("bin" = "all", "N" = nrow(DT))
    binl <- DT[, .N, by = bin]  # n obs per bin
    a <- round(rbind(DT[mask == 1, lapply(.SD, wm, wgt), .SDcols = nms1],
                     DT[, lapply(.SD, wm, wgt), .SDcols = nms1],
                     DT[mask == 1, lapply(.SD, wma, wgt), .SDcols = nms1],
                     DT[, lapply(.SD, wma, wgt), .SDcols = nms1]), 2)
    dtl <- list(DT[mask == 1, lapply(.SD, wm, wgt), by = bin, .SDcols = nms1],
                DT[, lapply(.SD, wm, wgt), by = bin, .SDcols = nms1],
                DT[mask == 1, lapply(.SD, wma, wgt), by = bin, .SDcols = nms1],
                DT[, lapply(.SD, wma, wgt), by = bin, .SDcols = nms1])
    b <- rbindlist(lapply(1:4, function(x) cbind(fr1[x], round(dtl[[x]], 2))))
    setkey(b, "bin")
    odt <- rbind(cbind(fr1, rbind(cbind(fr2, a))), binl[b])
  })
  named_out(l1, names(c_pct_diff[[x]]))
})
dut <- Sys.time() - dang  # 3.65 minutes (vs 33 in earlier incarnation),
# 24 sec for means only
names(cb_stats) <- lev_vec
#save(cb_stats, file = "external/ext_data/carbon_bias_tables.rda")
save(cb_stats, file = "external/ext_data/carbon_bias_tables_mus.rda")

namevec <- LC
stats <- cb_stats
i1 <- "all"
i2 <- "m"

cb_stats$f25$geow[bvals == "m" & bin != "all", lapply(.SD, mean), .SDcols = LC]
cb_stats$f25$geow[bvals == "m", ]
cb_stats$f25$geow[bvals == "m" & bin != "all", lapply(.SD, mean),
                  .SDcols = "forest"]

extract_stat0 <- function(namevec, stats, i1, i2, type = "density") {
  estats <- do.call(rbind, lapply(namevec, function(i) {
      dat <- sapply(stats, function(x) {
      vals <- unlist(sapply(x, function(y) {
        if(type == "density") {
          v <- y[bin == i1 & bvals == i2, i, with = FALSE]
        } else if(type == "nodensity") {
          v <- y[bin != i1 & bvals == i2, lapply(.SD, mean), .SDcols = i]
        }
        if(nrow(v) == 0) v <- rbindlist(list(v, as.list(NA)))
        v
      }))
    })#)
  }))
  out <- cbind(do.call(rbind, strsplit(rownames(estats), "\\.")),
               data.table(round(estats, 2)))
  #setnames(out, old = names(mu), new = "")
  return(out)
}
N <- function(namevec, stats, i1, i2) {
  estats <- data.table(do.call(rbind, lapply(namevec, function(i) {
    o <- sapply(stats, function(x) {
      sapply(x, function(y) y[bin == i1 & bvals == i2, N][1])
    })
  })))
  return(estats)
}
```

6.3.3.2 Secondary methods

6.3.3.3 Extract statistics for secondary methods

For supplementals

```
# cropland area only
a <- extract_stat0(LC, cb_stats, "all", "m")  # mean cropland only
b <- cbind(V2 = "All", a[, lapply(.SD, mean), by = V1, .SDcols = lev_vec])
setcolorder(a, c(2:1, 3:ncol(a)))
mu <- rbind(a, b)  # bias

a <- extract_stat0(LC, cb_stats, "all", "ma")  # mean abs cropland only
b <- cbind(V2 = "All", a[, lapply(.SD, mean), by = V1, .SDcols = lev_vec])
setcolorder(b, c(2:1, 3:ncol(b)))
mua <- rbind(a, b)  # MAE

# whole country
a <- extract_stat0(LC, cb_stats, "all", "m0")  # mean whole country
b <- cbind(V2 = "All", a[, lapply(.SD, mean), by = V1, .SDcols = lev_vec])
setcolorder(b, c(2:1, 3:ncol(b)))
mu0 <- rbind(a, b)  # bias

a <- extract_stat0(LC, cb_stats, "all", "ma0")  # mean abs whole country
b <- cbind(V2 = "All", a[, lapply(.SD, mean), by = V1, .SDcols = lev_vec])
setcolorder(b, c(2:1, 3:ncol(b)))
mu0a <- rbind(a, b)  # MAE

# density independent
a <- extract_stat0(LC, cb_stats, "all", "m0", type = "nodensity")
b <- cbind(V2 = "All", a[, lapply(.SD, mean), by = V1, .SDcols = lev_vec])
setcolorder(a, c(2:1, 3:ncol(a)))
mund <- rbind(a, b)  # bias

# check
all(round(sapply(lev_vec, function(x) {
  mean(cb_stats[[x]]$modmu[bvals == "m0" & bin != "all", forest])
}), 2) == mund[V2 == "forest" & V1 == "modmu", lev_vec, with = FALSE])

a <- extract_stat0(LC, cb_stats, "all", "ma0", type = "nodensity")
b <- cbind(V2 = "All", a[, lapply(.SD, mean), by = V1, .SDcols = lev_vec])
setcolorder(b, c(2:1, 3:ncol(b)))
munda <- rbind(a, b)  # MAE

# check
all(round(sapply(lev_vec, function(x) {
  mean(cb_stats[[x]]$modmu[bvals == "ma0" & bin != "all", forest])
}), 2) == munda[V2 == "forest" & V1 == "modmu", lev_vec, with = FALSE])
```

6.3.4 Plot mean carbon bias/MAE against scale

```


cols <- c("red", "orange3", "green4", "blue")
lcnms <- c("SA-LC", "GlobCover", "MODIS", "Geowiki")
pnms <- c("Forest", "Secondary", "Shrubland", "Grassland", "Sparse")
lw <- 1.5
```

```
# yl <- range(cb_statsw_mu[, LC, with = FALSE],
#              cb_statsw_mua[, LC, with = FALSE])
# yl <- round(yl / 10) * 10
yl <- c(-150, 150)
yax <- seq(yl[1], yl[2], 10)
xax <- seq(1, 6, 5 / (length(yax) - 1))

alph <- c(225, 40)
LC2 <- c(LC[-grep("second|grass", LC)], "All")
x <- c(0, 3, 6, 9, 12, 15)
w <- 3 / 8
xo <- (cumsum(rep(w, 8)) - w / 2)[-c(2, 4, 6, 8)]
xa <- sapply(x, function(x) x + xo)
mutype <- c("cb_statsw_mua", "cb_statsw_mu")

o <- c(0, w)
pchs <- c("*", "+", "o", "-")
cx <- c(2, 1.25, 1, 1)
g1 <- "grey90"

# ctabo
xl <- c(-0.5, 18)
shd <- c(4.5, 10.5, 16.5)

pdf(full_path(p_fig, "carbon_veg_scalew.pdf"), height = 7, width = 7)
par(mar = rep(1, 4), oma = c(2, 2, 0, 0), mgp = c(1, 0.5, 0), tcl = -0.3)
plot(xl, yl, pch = "", yaxt = "n", xaxt = "n", xaxs = "i", yaxs = "i",
     ylab = "", xlab = "")#,
for(i in shd) polyfunc2(x = i, y = yl, w = 3, col = g1, bcol = g1, lwd = 1)
abline(h = yax, v = NULL, col = "grey80", lty = 1)
polyfunc2(x = 8.75, y = yl, w = 18.5, col = "transparent", bcol = "black")
lines(c(-1, 18), c(0, 0), lwd = 2, col = "grey50")
for(i in 1:length(lev_vec)) { # i <- 1; j <- 1; k <- 1
  for(j in 1:length(snms)) {
    for(k in 1:length(mutype)) {
      dat <- get(mutype[k])[ol == lev_vec[i] & map == snms[j], LC2, with=FALSE]
      pcol <- makeTransparent(cols[j], alpha = alph[k])
      polyfunc2(xa[j, i] + o[k], range(dat), w = w, col = pcol, bcol = pcol)
      for(z in 1:length(LC2)) {
        lpfunc(x[i] + xo[j] + o[k], dat[, z, with = FALSE], col = "black",
               size = cx[z], pch = pchs[z], type = "pt")
      }
    }
  }
}
axis(1, at = seq(1.5, 18.5, 3), labels = c(1, fact))
axis(2, at = yax, labels = yax, las = 2)
mtext(side = 1, text = "Resolution (km)", outer = TRUE, line = 0.5)
mtext(side = 2, text = "Bias/MAE (%)", outer = TRUE, line = 1)
legend(x = 12.4, y = -110, legend = lcnms, pch = 15, col = cols, adj = 0,
       pt.cex = 1.5, bty = "n", cex = 0.8, x.intersp = 0.5)
legend(x = 11.9, y = -110, legend = rep("", 4), pch = 15, adj = 0,
       col = makeTransparent(cols, alpha = alph[2]), pt.cex = 1.5, bty = "n",
       cex = 0.8)
text(x = 12.7, y = -115, labels = "MAE", srt = 45, adj = c(0, 0), cex = 0.8)
text(x = 12.2, y = -115, labels = "Bias", srt = 45, adj = c(0, 0), cex = 0.8)
legend(x = 12.1, y = 120, legend = c("Forest", "Shrubland", "Sparse", "Mean"),
       pch = pchs, pt.cex = c(2, 1.5, 1.5, 2), bty = "n", cex = 0.8)
dev.off()
```

6.3.4.1 Density-weighted bias and accuracy

6.3.5 Supplementary bias/accuracy tables

```
# Statistics from cropland area-weighted measures, reshaped
cb_stats_s <- lapply(rev(mutype), function(i) {
  i0 <- do.call(rbind, lapply(snms, function(x) {
    i1 <- do.call(rbind, lapply(c(LC, "All"), function(y) {
      i2 <- do.call(cbind, lapply(lev_vec, function(z) {
        get(i)[ol == z & map == x, get(y)]
      }))
      cbind.data.frame(y, i2)
    }))
    i1 <- cbind.data.frame(x, i1)
    colnames(i1) <- c("Map", "Cover", lev_vec)
    data.table(i1)
  }))
})
names(cb_stats_s) <- c("Bias", "Accuracy")
muws <- copy(rbind(cbind(Metric = "Bias", cb_stats_s$Bias),
                   cbind(Metric = "MAE", cb_stats_s$Accuracy)))
setnames(muws, lev_vec, paste(c(1, fact), "km"))
setkeyv(muws, c("Map", "Cover"))
for(i in 1:length(LC)) muws[Cover == LC[i], Cover := pnms[i]]
setkey(muws, "Map")
for(i in 1:length(snms)) muws[Map == snms[i], Map := mcap[i]]


# Statistics from non-croplands areas included in means, for supplementals
# mu0[, c(lev_vec) := lapply(.SD, round), .SDcols = lev_vec]
# mu0a[, c(lev_vec) := lapply(.SD, round), .SDcols = lev_vec]
mu0s <- copy(rbind(cbind(Metric = "Bias", mu0),
                   cbind(Metric = "MAE", mu0a)))
setnames(mu0s, c("V1", "V2", lev_vec),
         c("Map", "Cover", paste(c(1, fact), "km")))
setkeyv(mu0s, c("Map", "Cover"))
for(i in 1:length(LC)) mu0s[Cover == LC[i], Cover := pnms[i]]
setkey(mu0s, "Map")
for(i in 1:length(snms)) mu0s[Map == snms[i], Map := mcap[i]]

# Statistics from agricultural areas
mus <- copy(rbind(cbind(Metric = "Bias", mu),
                  cbind(Metric = "MAE", mua)))
setnames(mus, c("V1", "V2", lev_vec),
         c("Map", "Cover", paste(c(1, fact), "km")))
```

```
setkeyv(mus, c("Map", "Cover"))
for(i in 1:length(LC)) mus[Cover == LC[i], Cover := pnms[i]]
setkey(mus, "Map")
for(i in 1:length(snms)) mus[Map == snms[i], Map := mcap[i]]

# Bind all three tables together
fulls <- muws
# fulls <- rbind(cbind("Region" = "Density", muws),
#                cbind("Region" = "Country", mu0s),
#                cbind("Region" = "Agricultural", mus))


caption <- paste("Biases and mean absolute errors, weighted by reference",
                 "cropland density, for each of the test",
                 "maps across aggregation scales and each possible landcover",
                 "type sharing the pixel with cropland.")
# fulls_xtab <- xtable(fulls[order(Region, Metric, Cover)], digits = 1,
fulls_xtab <- xtable(fulls[order(Metric, Cover)], digits = 1,
                     caption = caption)
print(fulls_xtab, type = "latex", file = "paper/figures/C-bias-accuracy.tex",
      tabular.environment = "longtable", floating = FALSE,
      caption.placement = "top", include.rownames = FALSE)
#       add.to.row = list(pos = list(0),
#                         command = "\\hline \\endhead"))
```

7 Error, bias, and accuracy in evapotranspiration estimates

Bias in PET estimates as calculated using different version of the cropland datasets to determine the vegetation properties in VIC.

7.1 Analysis

7.1.1 Prepare datasets

Bring in cropland datasets, ET grids, reproject the latter to SA Albers, mask, convert to monthly total ET, etc.

```
library(SAcropland)
library(RColorBrewer)
p_root <- full_path(proj_root("SAcropland"), "SAcropland")
p_fig <- full_path(p_root, "paper/figures/")
p_data <- full_path(p_root, "external/ext_data/")
p_data2 <- full_path(p_root, "data/")
p_et <- full_path(p_root, "external/ext_data/vic/")

# cropland data
load(full_path(p_data, "MZshapes.Rdata"))  # SA shape
load(full_path(p_data, "d_grid_act.rda"))  # actual diffence grids
paths <- full_path(p_data, dir(p_data, paste0("cover*.*11sum_mask.tif$")))
gti <- raster(paths) / 100  # gti 2011
namask <- raster(full_path(p_data, "namask.tif"))  # NA mask
```

```r
gti <- mask(gti, namask)  # apply mask to GTI
# cellStats(!is.na(gti), sum)

# Reconstruct original landcover estimates
snms <- c("sa30", "globmu", "modmu", "geow")
anms <- c("gti", snms)
dlist_1km <- lapply(dlist_act[snms], function(x) x$f1$g2011)
lc_list <- lapply(dlist_1km, function(x) gti - x / 100)
# cellStats(!is.na(lc_list$sa30), sum)

# check
# plot(lc_list[[1]])
# plot(gti)
# tst <- raster("external/ext_data/geowikisa_masked.tif")  # check
# plot(round(tst / 100 - lc_list[[4]], 4)) # okay

# aggregate rasters
fact <- 25
lc_agg <- aggregate_rast_list(fact, lc_list)   # landcover rasters
gti_agg <- aggregate_rast_list(fact, list("gti" = gti))  # GTI rasters

# Create namask to remove all NA areas across datasets
# sumna <- function(x) sum(x, na.rm = FALSE)
# namask <- calc(stack(gti, stack(lc_list)), sumna)
namask2 <- !is.na(namask)  # set NAs to zero
# cellStats(namask2, sum)
# # namaskt <- raster(full_path(p_data, "namask.tif")) # load in NA mask
# # namaskt2 <- !is.na(namaskt)  # set NAs to zero
# cellStats(is.na(gti), sum); cellStats(namask2, sum); cellStats(namask2t, sum)
# cellStats(namaskt, sum)
awgts <- aggregate_rast_list(fact, list(namask2), fun = sum)
awgts <- lapply(awgts$f25, function(x) {
  r <- x[[1]]
  projection(r) <- sa.shp@proj4string
  r
})  # reshape a bit into list for subsequent use
names(awgts) <- "f25"
mask25k <- awgts$f25 > 0  # mask for et datasets

# read in ET estimates, but drop first year because of VIC problem in January
etf <- dir(p_et, pattern = "nc", full.names = TRUE)
etb <- lapply(etf, function(x) brick(x))
etb <- lapply(etb, function(x) dropLayer(x, i = 1:12))

# reproject them to Albers
etba <- lapply(etb, function(x) {
  b <- projectRaster(x, awgts$f25)
})

# rename layers and convert to total month mm
dts <- seq.Date(as.Date("1981/1/1"), as.Date("2008/12/31"), "days")
allmos <- substr(dts, 6, 7)
allyrs <- substr(dts, 1, 4)
```

```
ndays <- unlist(lapply(unique(allyrs), function(x) {
  unname(sapply(unique(allmos), function(y) {
    length(which(allyrs == x & allmos == y))
  }))
}))  # n days in each month in time series

mos <- format(ISOdatetime(2000,1:12,1,0,0,0),"%b")
lnames <- sapply(1981:2008, function(x) paste0(mos, "_", substr(x, 3, 4)))
lnames <- lnames[1:length(lnames)]
# length(lnames) == nlayers(etba[[1]])
etba <- lapply(etba, function(x) {
  b <- mask(x, mask25k, maskvalue = 0)
  names(b) <- lnames
  b
})
names(etba) <- anms

# check that mm/month multiplies through correctly
tst <- etba[[1]] * ndays
plot(tst[[15]] - etba[[1]][[15]] * ndays[15])

# convert to monthly total mm
etba <- lapply(etba, function(x) x * ndays)
# plot(etba[[1]][[1]])

# date indices
yrs <- gsub("*.*_", "", lnames)
mos <- gsub("_.*", "", lnames)
yi <- t(sapply(unique(yrs), function(x) range(which(yrs == x))))  # year index
mi <- sapply(unique(mos), function(x) which(mos == x))  # month index
```

7.1.2 Process ET time series

Figure out which months produce maximum ET values, calculate monthly means, overall country-wide time series, etc.

```
# Calculate mean, median, mode of dates in time series when PET max occurs
et_max <- lapply(etba, function(x) {
  bi <- stack(lapply(1:nrow(yi), function(y) {
    ind <- yi[y, 1]:yi[y, 2]
    which.max(x[[ind]])
  }))
  s <- stack(calc(bi, max), calc(bi, modal), calc(bi, mean), calc(bi, median))
  names(s) <- c("maxmax", "mode", "mean", "median")
  s
})

# plot to see which statistic picks out PET peak date most effectively
plot(et_max$gti)  # all the same, not coherent
plot(et_max$gti %in% 6:9)  # median looks most sensible and coherent
plot(et_max$gti - et_max$sa30)  # some differences in date of max
plot(et_max$gti - et_max$globmu)
plot(et_max$gti - et_max$modmu)
```

```r
plot(et_max$gti - et_max$geow)

# Monthly mean values
et_momu <- lapply(etba, function(x) {
  bi <- stack(lapply(1:ncol(mi), function(y) {
    ind <- mi[, y]
    calc(x[[ind]], mean)
  }))
  names(bi) <- unique(mos)
  bi
})

# do a focal max to separate out noisy areas
et_medsm <- focal(et_max$gti$median, w = matrix(1, 3, 3), fun = modal, pad = 3)
buf <- (!is.na(et_max$gti$median) - !is.na(et_medsm)) * et_max$gti$median
et_medsm[is.na(et_medsm)] <- 0
et_medsm <- buf + et_medsm  # fill back in areas edged out by moving window
# plot((et_medsm > 0) - (et_momu$gti[[1]] > 0))

# 1 month before and after peak
et_meds_m1 <- et_medsm - 1
# plot(et_medsm - 1)  # a few zero areas
et_meds_m1[et_meds_m1 < 1] <- 12  # set to 12 (December) if median - 1 = 0
et_meds_p1 <- et_medsm + 1
# plot(et_medsm + 1)  # nothing greater than 12

# calculate 3 month peak ET
gs_peak <- lapply(et_momu, function(x) {
  et1 <- stackSelect(x, et_meds_m1)
  et2 <- stackSelect(x, et_medsm)
  et3 <- stackSelect(x, et_meds_p1)
  stack(et1, et2, et3)
})

# Annual mean values
et_amu <- lapply(etba, function(x) {
  bi <- stack(lapply(1:nrow(yi), function(y) {
    ind <- yi[y, 1]:yi[y, 2]
    calc(x[[ind]], sum)
  }))
  et_ts <- cellStats(bi, mean)
  names(et_ts) <- 1981:2008
  list("gridmu" = calc(bi, mean), "cntry_ts" = et_ts)
})

# plot annual mean values
rbcol <- rainbow(4)
plot(1981:2008, et_amu$gti$cntry_ts, type = "l", las = 2)
for(i in 2:5) lines(1981:2008, et_amu[[i]]$cntry_ts, col = rbcol[i])

# figure out min_max years
mm_yr <- sapply(list(which.min, which.max), function(x) x(et_amu$gti$cntry_ts))
```

```
# annual totals
et_mutot <- lapply(etba, function(x) {
  bi <- stack(lapply(1:nrow(yi), function(y) {
    ind <- yi[y, 1]:yi[y, 2]
    calc(x[[ind]], sum)
  }))
})
et_mm <- lapply(et_mutot, function(x) x[[mm_yr]])
```

7.1.3 Calculate ET differences

```
# difference rasters and convert to percent
# between time series annual mean
ilist <- lapply(et_amu[2:5], function(x) x[[1]])
a <- et_amu$gti$gridmu
amu_diff <- lapply(ilist, function(b) {
  (a - b) / a * 100
})

# between time series mins
ilist <- lapply(et_mm[2:5], function(x) x[[1]])
a <- et_mm$gti[[1]]
mumin_diff <- lapply(ilist, function(b) {
  (a - b) / a * 100
})

# between time series maxs
ilist <- lapply(et_mm[2:5], function(x) x[[2]])
a <- et_mm$gti[[2]]
mumax_diff <- lapply(ilist, function(b) {
  (a - b) / a * 100
})

# between mean peak growing season ET (3 month total)
ilist <- lapply(gs_peak[2:5], function(x) calc(x, sum))
a <- calc(gs_peak$gti, sum)
gs3_diff <- lapply(ilist, function(b) {
  (a - b) / a * 100
})

# between median max growing season ET (3 month total)
ilist <- lapply(gs_peak[2:5], function(x) x[[2]])
a <- gs_peak$gti[[2]]
gsmx_diff <- lapply(ilist, function(b) {
  bl <- (a - b) / a * 100
})
```

7.1.4 Plot difference maps

```r
# function to reshape lists for use in disaggregate_rast_list
disaggl <- function(rl, lev = "f25") {
  bo <- lapply(rl, function(x) {
    b <- list(x)
    names(b) <- lev
    b
  })
  bo
}

# reshape in list
# l2dag <- list(amu_diff, mumax_diff, mumin_diff, gs3_diff, gsmx_diff)
# names(l2dag) <- c("mu", "mumx", "mumn", "gs", "gspk")
l2dag <- list(amu_diff, mumax_diff, mumin_diff, gs3_diff)
names(l2dag) <- c("mu", "mumx", "mumn", "gs")
l2dag <- lapply(l2dag, disaggl)

# disaggregate
l2dagd <- lapply(l2dag, function(x) {
  disaggregate_rast_list(snms, "f25", x, namask)
})

# stats for plotting
dstats <- function(dlist) {
  statsd <- lapply(dlist, function(x) {
    sapply(x, function(y) {
      c(cellStats(y, mean), quantile(y, seq(0, 1, 0.05)))
    })
  })
}
```

7.1.4.1 Reshape and disaggregate lists

```r
# calculate single set of breakranges for all sets
statsl <- lapply(l2dag, dstats)
rng <- do.call(cbind, lapply(statsl, function(x) {
  ilims <- c(ceiling(min(sapply(x, function(y) y[3, ]))),
             floor(max(sapply(x, function(y) y[21, ]))))
  olims <- range(sapply(x, function(y) range(y)))
  c(olims[1], ilims, olims[2])
}))
rngdt <- data.table(t(rng))
rng <- unname(unlist(rngdt[, lapply(.SD, function(x) x[which.max(abs(x))])]))

fnm <- "et_bias_map4.pdf"
legtext <- rep("% Difference", 5)
cx <- 1.4
lcol <- "black"
mcap <- c("SA-LC", "GlobCover", "MODIS", "GeoWiki")
```

```
# ecap <- c("Annual mean", "TS max", "TS min", "Peak 3", "Peak")
ecap <- c("Annual mean", "TS max", "TS min", "Peak")

rnder <- function(x) {
  dig <- nchar(max(round(x)))
  rmat <- cbind(1:8, c(1, 5, 10, 20, 50, 100, 100, 1000))
  rnd <- rmat[dig, 2]
  round(c(floor(x[1] / rnd), ceiling(x[2] / rnd))) * rnd
}

# For Supplementals
# brks <- c(seq(floor(rng[1]), -5, by = 5), -1, 1, seq(5, ceiling(rng[4]), 5))
brks <- c(-25, -15, -10, -5, -3, -1, 1, 3, 5, 10, 15, 25)
# cvec <- c("red", "orange", "grey80", "green", "blue4")
# cols <- colorRampPalette(cvec)(length(brks) - 1)
# cols[c(1, length(cols))] <- c("darkred", "purple3")
cols <- brewer.pal(length(brks) - 1, name = "RdBu")
cols[6] <- "grey80"
pdf(full_path(p_fig, fnm), height = 6, width = 7)
par(mfcol = c(4, 4), mar = c(0, 0, 0, 0), oma = c(5, 5, 2, 0))
for(j in 1:length(l2dagd)) {
  print(names(l2dagd)[j])
  for(k in 1:length(snms)) {
    rl <- l2dagd[[j]][[k]]$f25
    print(paste("...", snms[k]))
    plot(sa.shp, lty = 0)
    image(rl, add = TRUE, col = cols, breaks = brks, legend = FALSE)
    if(j == 1) mtext(mcap[k], side = 2, line = 1, cex = cx)
    if(k == 1) mtext(ecap[j], side = 3, line = 0, cex = cx)
  }
}
flex_legend(ncuts = length(brks) - 1, legend.text = legtext[i],
            legend.vals = brks, longdims = c(0.2, 0.8),
            shortdims = c(0.06, 0.01),
            colvec = cols, #(length(brks) - 1),
            srt = c(270, 0), horiz = TRUE, textside = "bottom",
            legend.pos = c(4, 4), leg.adj = list(c(0.1, 0.3), c(-0.25, -0.5)),
            cex.val = cx - 0.4, textcol = lcol, bordercol = lcol)
dev.off()


fnm <- "et_bias_map.pdf"
pdf(full_path(p_fig, fnm), height = 6, width = 7)
par(mfcol = c(2, 2), mar = c(0, 0, 0, 0), oma = c(4, 0, 1, 0))
for(j in 1) {
  print(names(l2dagd)[j])
  for(k in 1:length(snms)) {
    rl <- l2dagd[[j]][[k]]$f25
    print(paste("...", snms[k]))
    plot(sa.shp, lty = 0)
    image(rl, add = TRUE, col = cols, breaks = brks, legend = FALSE)
    mtext(mcap[k], side = 3, line = -1, cex = cx)
  }
}
flex_legend(ncuts = length(brks) - 1, legend.text = legtext[i],
            legend.vals = brks, longdims = c(0.2, 0.8),
            shortdims = c(0.07, 0.01),
            colvec = cols, #(length(brks) - 1),
            srt = c(270, 0), horiz = TRUE, textside = "bottom",
            legend.pos = c(4, 4), leg.adj = list(c(0.3, 0.3), c(-0.25, -0.75)),
            cex.val = cx - 0.4, textcol = lcol, bordercol = lcol)
dev.off()
```

7.1.4.2 Plot maps

7.1.5 Calculate bias/accuracy statistics

7.1.5.1 Primary method

Density dependent method

```
# weighted mean functions
wm <- function(x, w) stats::weighted.mean(x, w)
wma <- function(x, w) stats::weighted.mean(abs(x), w)
snms <- c("sa30", "globmu", "modmu", "geow")

# Quick look at density dependent
# reshape for ET mu
err2011 <- lapply("f25", function(x) {
  sapply(l2dag$mu, function(y) list(y[[x]]))
})
names(err2011) <- "f25"

# calculate bias/MAE for data
a <- bias_statsw(gti_agg$f25, awgts, err2011$f25, snms, wm, "mu")
b <- bias_statsw(gti_agg$f25, awgts, err2011$f25, snms, wma, "mua")
et_muw <- rbind(a, b)

# reshape for ET mu
err2011 <- lapply("f25", function(x) {
  sapply(l2dag$gs, function(y) list(y[[x]]))
})
names(err2011) <- "f25"

# calculate bias/MAE for data
a <- bias_statsw(gti_agg$f25, awgts, err2011$f25, snms, wm, "mu")
b <- bias_statsw(gti_agg$f25, awgts, err2011$f25, snms, wma, "mua")
et_gsw <- rbind(a, b)
```

Agricultural area method

```
lev_vec <- "f25"
# create mask for non-cropland areas, unioning GTI and each LC, filtering out
# areas of no-cropland (<1/2% total cover)
lcu <- lapply(lev_vec, function(x) {
  lcb <- lapply(snms, function(y) {
    gti_gt0 <- Which(round(gti_agg[[x]]$gti * 100) > 0)
```

```r
    lc_gt0 <- Which(round(lc_agg[[x]][[y]] * 100) > 0)
    all_gt0 <- gti_gt0 + lc_gt0
    all_gt0[all_gt0 > 0] <- 1
    all_gt0
  })
  named_out(lcb, snms)
})
names(lcu) <- lev_vec

# cropland cover bins, for looking at error as a function of cover
binv <- seq(0, 1, 0.05)
bins <- lapply(gti_agg[[2]], function(x) {
  cut(x[[1]], breaks = binv, include.lowest = TRUE)
})
names(bins) <- lev_vec
# bins <- cut(gti_agg$f25$gti, breaks = binv, include.lowest = TRUE)

# calculate bias stats
vnms <- names(l2dag)
bstats <- lapply(vnms, function(x) {
  a <- bias_stats_list(bins, awgts, lcu, l2dag[[x]], wm, "mu", "bias", TRUE)
  b <- bias_stats_list(bins, awgts, lcu, l2dag[[x]], wma, "mua", "bias", TRUE)
  stat <- rbind(a, b)
})
names(bstats) <- vnms

# checks to see if stats correct
# bias_stats* functions correct
chk <- bias_stats(bins$f25, awgts$f25, lcu$f25$globmu, l2dag$mu$globmu$f25, wm,
                  "mu", "bias", TRUE)
all(chk[bvals == "mu", bias] == bstats$mu[il == "globmu" & bvals == "mu", bias])
# compared to raster
for(i in snms) {
  msk <- lcu$f25[[i]]
  msk[msk == 0] <- NA
  a <- l2dag$mu[[i]]$f25 * msk
  b <- awgts$f25 * msk
  d <- weighted.mean(values(a), values(b), na.rm = TRUE)
  print(round(d, 2) == bstats$mu[il == i & bvals == "mu" & bin == "all", bias])
}  # check

# extract stats
mu <- lapply(bstats, function(x) {
  extract_stat(lev_vec, snms, "all", "mu", "bias", x)
})
mua <- lapply(bstats, function(x) {
  extract_stat(lev_vec, snms, "all", "mua", "bias", x)
})

# reshape
mus <- rbindlist(lapply(names(mu), function(x) cbind(x, mu[[x]])))
mus[, ol := NULL]
setnames(mus, c("x", "il", "bias"), c("Variable", "Map", "Bias"))
```

```
setkeyv(mus, c("Map", "Variable"))
muas <- rbindlist(lapply(names(mua), function(x) cbind(x, mua[[x]])))
muas[, ol := NULL]
setnames(muas, c("x", "il", "bias"), c("Variable", "Map", "MAE"))
setkeyv(muas, c("Map", "Variable"))
mutab <- mus[muas]
mutab[, lapply(.SD, function(x) max(abs(x))), .SDcols = 3:4]
setkey(mutab, "Variable")
vnmr <- c("Annual Mean", "29-year Max", "29-year Min", "Peak")
for(i in 1:length(vnms)) mutab[Variable == vnms[i], Variable := vnmr[i]]
setkey(mutab, "Map")
for(i in 1:length(snms)) mutab[Map == snms[i], Map := mcap[i]]
mutab[order(-rank(Map))]

caption <- paste("Biases and mean absolute errors (as \\%) for",
                 "evapotranspiration variables derived from a 29-year time",
                 "series calculated by the VIC model, including the average",
                 "total ET for the",
                 "3 months of the year when ET is highest, the annual mean",
                 "and the minimum and maximum annual ETs in the time series.")
mutab_xtab <- xtable::xtable(mutab, digits = 1, caption = caption)
print(mutab_xtab, type = "latex", file = "paper/figures/et-bias.tex",
      tabular.environment = "longtable", floating = FALSE,
      caption.placement = "top", include.rownames = FALSE)
```

The largest (in an absolute sense) mean actual bias is -0.6%, while the largest mean absolute bias is 1.25%, so not much here.

7.1.6 Bias in relation to rainfall

We'll end by having a look at the relationship between bias and rainfall

```
# rainfall layer prepared for SA crop subsidy analysis
load("external/ext_data/climate/mag_dist_clim.rda")
pre <- calc(pre2, mean)  # 31 year mean rainfall for Africa
p4s <- projection(raster(nrow = 10, ncol = 10)) # GCS
presa <- crop(pre, spTransform(sa.shp, p4s))  # crop to SA
projection(presa) <- p4s
presa <- projectRaster(presa, mask25k)  # project to Albers
presa <- mask(presa, awgts$f25, maskvalue = 0)  # mask

plot(presa, breaks = seq(0, 1000, 100), col = topo.colors(9))
plot(abs(l2dag$mu$globmu$f25))
plot(presa, abs(l2dag$mu$globmu$f25))
```

8 Error, bias, and accuracy of gridded crop yield & production data

Based on Ramankutty et al's (2008) procedure for calculating cropland area, followed by Monfreda et al's (2008) to disaggregate yields and harvested areas.

8.1 Data

Use SA provinces with GTI to create the amount of agricultural land estimates per province, akin to provincial statistics used by Ramankutty et al (2008). Note: they mentioned SA has 11 administrative units they used, but the publications they cited here has just 9 provinces, so maybe they mistakenly counted Lesotho and Swaziland?

```
library(SAcropland)
library(xtable)
library(RColorBrewer)
p_root <- full_path(proj_root("SAcropland"), "SAcropland")
p_fig <- full_path(p_root, "paper/figures/")
p_data <- full_path(p_root, "external/ext_data/")
p_data2 <- full_path(p_root, "data/")
#p_carb <- full_path(p_root, "SAcropland/external/ext_data/carbon/")
# load(full_path(p_data, "yield-bias.rda"))

# SA provinces
prov <- readOGR("external/ext_data/provinces.sqlite", layer = "provinces")
prov <- prov[prov$id_1 != 9, ]  # remove Princeton Edward Islands
prov[prov$id_1 == 10, "id_1"] <- 9 # reset WC to id 10
load("external/ext_data/MZshapes.Rdata")

# cropland data
load(full_path(p_data, "d_grid_act.rda"))  # actual diffence grids
paths <- full_path(p_data, dir(p_data, paste0("cover*.*11sum_mask.tif$")))
gti <- raster(paths) / 100  # gti 2011
namask <- raster(full_path(p_data, "namask.tif"))  # NA mask
gti <- mask(gti, namask)  # apply mask to GTI
# cellStats(!is.na(gti), sum)

# Reconstruct original landcover estimates
snms <- c("sa30", "globmu", "modmu", "geow")
dlist_1km <- lapply(dlist_act[snms], function(x) x$f1$g2011)
lc_list <- lapply(dlist_1km, function(x) gti - x / 100)
# cellStats(!is.na(lc_list$sa30), sum)

# plot(lc_list[[1]])
# plot(gti)
# tst <- raster("external/ext_data/geowikisa_masked.tif")  # check
# plot(round(tst / 100 - lc_list[[4]], 4)) # okay

# Create namask to remove all NA areas across datasets
sumna <- function(x) sum(x, na.rm = FALSE)
# namask <- calc(stack(stack(gti), stack(lclist)), sumna)
# namask[namask >= 0] <- 1
namask2 <- !is.na(namask)  # set NAs to zero
# cellStats(!is.na(gti), sum); cellStats(namask2, sum)
fact <- c(5, 10, 25, 50, 100)
```

8.2 Analyses

8.2.1 Cropland fractions

```r
# rasterize provinces and stack with gti raster, convert to data.table
# provr <- rasterize(prov, y = gti, field = "id_1",
#                    filename = "external/ext_data/provinces.tif")
provr <- raster("external/ext_data/provinces.tif")
provr <- mask(provr, namask)
gtis <- stack(list("prov" = provr, "f" = gti))

# calculate fractions using raster::extract with provinces, to compare speed
# prov_est <- extract(gti, prov, progress = "text")
# prov_estNA <- lapply(prov_est, function(x) x[!is.na(x)])
# cftest <- cbind("carea" = sapply(prov_estNA, sum),
#                 "parea" = sapply(prov_estNA, length),
#                 "cf" = sapply(prov_estNA, sum) / sapply(prov_estNA, length))

# data.table calculations
gti_dt <- na.omit(as.data.table.raster(gtis, xy = TRUE))
setkey(gti_dt, prov)
cf <- gti_dt[, list("carea" = sum(f), "parea" = length(f),
                    "cf" = sum(f) / length(f)), by = prov]
# cftest - cf[, 2:4, with = FALSE]  # close, but diff caused by few NAs in provr

# data.table approach is much more efficient
```

8.2.1.1 Get provincial cf factors from 1 km GTI dataset

8.2.1.2 Calculate cropland fractions from 1 km landcover sets

Derived at 1 km resolution rather 10 km (as in Ramankutty et al, 2008), using data.tables to calculate provincial-level correction factors for each

```r
lcs <- stack(list("prov" = provr, stack(lc_list)))
lcs_dt <- na.omit(as.data.table.raster(lcs, xy = TRUE))
setkey(lcs_dt, prov)
fc <- function(x) sum(x) / length(x)  # fraction functions
lc_cf <- lapply(list(sum, length, fc), function(x) {
  lcs_dt[, lapply(.SD, x), by = prov, .SDcols = snms]
})
names(lc_cf) <- colnames(cf)[-1]
# lcs_dt[, .N, by = prov] == gti_dt[, .N, by = prov]

# correction factors
pcf <- cbind("prov" = prov$id_1,
             sapply(snms, function(x) cf$cf / lc_cf$cf[, get(x)]))
pcf <- data.table(pcf)

# tst1 <- disaggregate(tst, fact = 10)
# namask <- raster("external/ext_data/namask.tif")
# tst1m <- mask(crop(tst1, namask), mask = namask)
# plot(tst1m - gti$cover2011sum_mask)
# multiply each lc fraction value by correction factor for the particular
# province
lcs_dta <- copy(lcs_dt)
for(j in snms) {
  for(g in pcf$prov) {
    lcs_dta[prov == g, j := (get(j) * pcf[prov == g, get(j)]), with = FALSE]
  }
}

# check to see if DT syntax correct
for(i in 1:9) {
  for(j in snms) {
    print(all(lcs_dta[prov == i, get(j)][1:10] ==
                (lcs_dt[prov == i, get(j)] * pcf[i, get(j)])[1:10]))
  }
}

# check to see if the factors resulted in same total area per province
fchk <- lcs_dta[, lapply(.SD, sum), by = prov, .SDcols = snms]
par(mfrow = c(2, 2), mar = rep(1, 4))
for(i in snms) plot(cf[, carea], fchk[, get(i)])  # yup

# how many pixels have fraction > 1
for(i in 1:9) {
  for(j in snms) {
    print(paste(j, ":", i, ":",
                lcs_dta[prov == i, length(which(get(j) > 1))] /
                  lcs_dta[prov == i, .N]))
  }
}  # several in globcover, modis, and geow

# Adjust these to equal 1
lcs_dta2 <- copy(lcs_dta)

# function to adjust fractions to fall within 1 while maintaining statistics
force1 <- function(x) {
  a <- x
  while(any(a > 1)) {
    reall <- sum(a[a > 1] - 1)  # total of parts of x > 1, to reallocate
    ind <- which(a < 1)  # parts of x less than 1
    a[a > 1] <- 1
    af <- (reall + sum(a[a < 1])) / sum(a[a < 1])
    a[a < 1] <- a[a < 1] * af
  }
  return(a)
}
# check if works on modis set
hist(lcs_dta[prov == 8, modmu])
hist(force1(lcs_dta[prov == 8, modmu]))
round(sum(force1(lcs_dta[prov == 8, modmu])), 2) ==
  round(sum(lcs_dta[prov == 8, modmu]), 2)
```

```
for(j in snms) {
  for(g in pcf$prov) {
    lcs_dta2[prov == g, j := force1(get(j)), with = FALSE]
  }
}

# Check to see if all fractions <= 1 and that provincial sums are equal
for(i in 1:9) {
  for(j in snms) {
    print(paste(j, ":", i, ":",
                lcs_dta2[prov == i, length(which(get(j) > 1))] /
                  lcs_dta[prov == i, .N], ":",
                round(lcs_dta2[prov == i, sum(get(j))], 4) ==
                  round(lcs_dta[prov == i, sum(get(j))], 4)))
  }
}  # check

# convert back to rasters
lc_adj <- dt_to_raster(lcs_dta2, CRSobj = sa.shp@proj4string)
lc_adj <- dropLayer(lc_adj, i = 1)
```

8.2.1.3 Re-scale landcover cropland fraction using cf factors

```
tl <- lapply(1:4, function(x) lc_adj[[x]])
names(tl) <- snms
lc_agg <- aggregate_rast_list(fact, tl) # landcover rasters
rm(tl)
gtic <- crop(gti, lc_adj$sa30)
gti_agg <- aggregate_rast_list(fact, list("gti" = gtic))  # GTI rasters

# set up masks and weights for assessing output bias and absolute errors
namask <- calc(stack(gtic, lc_agg$f1), sumna)
# cellStats(!is.na(gtic), sum); cellStats(!is.na(lc_agg$f1$sa30), sum)
namask[namask >= 0] <- 1
namask2 <- !is.na(namask)  # set NAs to zero
# cellStats(!is.na(namask), sum); cellStats(namask2, sum)
fact <- c(5, 10, 25, 50, 100)
awgts <- aggregate_rast_list(fact, list(namask2), fun = sum)
awgts <- lapply(awgts, function(x) x[[1]])  # unlist on inner loop

# compare extents to make sure no offsets - look at in QGIS
# writeRaster(gti_agg$f1$gti, filename = "external/ext_data/test/gti_crop.tif")
# writeRaster(lc_agg$f1$sa30, filename = "external/ext_data/test/sa30_crop.tif")
# looks fine

# tst <- gti_agg$f5$gti - lc_agg$f5$sa30
# plot(tst * 100)
#pct_diff <- function(x, y) (x - y) / x * 100

# calculate differences between adjusted fraction and reference
```

```r
cf_pct_diff <- lapply(snms, function(x) {
  dif <- lapply(names(lc_agg), function(y) {
    d <- gti_agg[[y]][[1]] - lc_agg[[y]][[x]]
  })
  named_out(dif, names(lc_agg))
})
names(cf_pct_diff) <- snms

# same, but as percent
cf_pct_diff2 <- lapply(snms, function(x) {
  dif <- lapply(names(lc_agg), function(y) {
    d <- (gti_agg[[y]][[1]] - lc_agg[[y]][[x]]) * 100
  })
  named_out(dif, names(lc_agg))
})
names(cf_pct_diff2) <- snms

# cropland cover bins, for looking at error as a function of cover, based on
# original extent of gti data, for ease
binv <- seq(0, 1, 0.05)
bins <- lapply(gti_agg, function(x) {
  cut(x$gti, breaks = binv, include.lowest = TRUE)
})

lev <- names(cf_pct_diff[[1]])
lcu <- lapply(lev, function(x) {
  lcb <- lapply(snms, function(y) {
    gti_gt0 <- Which(gti_agg[[x]][[1]] > 0)
    lc_gt0 <- Which(lc_agg[[x]][[y]] > 0)
    all_gt0 <- gti_gt0 + lc_gt0
    all_gt0[all_gt0 > 0] <- 1
    all_gt0
  })
  named_out(lcb, snms)
})
names(lcu) <- lev

# calculate bias/MAE for data
# calculate means
wm <- function(x, w) stats::weighted.mean(x, w)
wma <- function(x, w) stats::weighted.mean(abs(x), w)

# reshape error raster for density-weighted bias/accuracy calculations
cf_pct_resh <- lapply(lev, function(x) {
  sapply(cf_pct_diff2, function(y) list(y[[x]]))
})
names(cf_pct_resh) <- lev

# bias/accuracy
a <- bias_statsw_list(gti_agg, awgts, cf_pct_resh, snms, wm, "mu")
b <- bias_statsw_list(gti_agg, awgts, cf_pct_resh, snms, wma, "mua")
lcf_err <- rbind(a, b)
```

```r
# check
# Older variant of code from compare-landcover.Rmd to evaluate whether newer DT
# version is finding correct results
# check error stats - do they match alternate approaches?
for(i in c("f1", "f25", "f10")) {
  for(j in c("sa30", "modmu", "geow")) {
    print(paste("cross-checking calculations in", i, j))
    a1 <- bias_statsw(gti_agg[[i]]$gti, awgts[[i]], cf_pct_resh[[i]], snms, wm,
                      "mu", aweight = FALSE)[, j, with = FALSE]
    a2 <- bias_statsw(gti_agg[[i]]$gti, awgts[[i]], cf_pct_resh[[i]], snms, wm,
                      "mu", rweight = FALSE, aweight = FALSE)[,j, with=FALSE]
    a3 <- bias_statsw(gti_agg[[i]]$gti, awgts[[i]], cf_pct_resh[[i]], snms, wm,
                      "mu")[, j, with = FALSE]
    c1 <- getValues(cf_pct_resh[[i]][[j]])
    w1 <- getValues(gti_agg[[i]]$gti)
    w2 <- getValues(awgts[[i]])
    print("non-area weighted mean matches?")
    print(a1 == round(weighted.mean(c1, w1, na.rm = TRUE), 2))
    print("totally unweighted mean matches?")
    print(a2 == round(cellStats(cf_pct_resh[[i]][[j]], mean), 2))
    print("double weighted mean matches?")
    print(a3 == round(weighted.mean(c1, w1 * w2, na.rm = TRUE), 2))
  }
}

# Removed whole-country and agricultural area accuracy measures. See repo prior
# to 18/10 if needed.
svec <- c("mu", "mua")
aa <- c("Bias", "Accuracy")
mcap <- c("SA-LC", "GlobCover", "MODIS", "GeoWiki")

# Reshape for output table
lcf_out <- do.call(rbind, lapply(1:length(svec), function(i) {
  a <- do.call(rbind.data.frame, lapply(snms, function(x) {
    aa <- sapply(lev, function(y) {
      lcf_err[ol == y & stnm == svec[i], x, with = FALSE]
    })
    named_out(c(x, aa), c("Map", paste(c(1, fact), "km")))
  }))
  named_out(cbind.data.frame(aa[i], a),
            c("Metric", "Map", paste(c(1, fact), "km")))
}))

# rename landcover sets for output
lcf_out <- as.data.table(lcf_out)
setkey(lcf_out, "Map")
for(i in 1:length(snms)) lcf_out[Map == snms[i], Map := mcap[i]]

# Output table
caption <- paste("Bias and mean absolute errors (MAE) in statistically",
                 "constrained cropland maps across aggregation scales,",
                 "weighted by density of cropland",
                 "cover in the reference map. ")
```

```r
lcfout_xtab <- xtable(lcf_out[order(Metric)], digits = 1, caption = caption)
print(lcfout_xtab, type = "latex",
      file = "paper/figures/cropadj-bias-accuracy.tex",
      tabular.environment = "longtable", floating = FALSE,
      caption.placement = "top", include.rownames = FALSE)

# disaggregate for plotting
# namaskc <- crop(namask, gtic)
disagg <- disaggregate_rast_list(snms, lev, cf_pct_diff, namask)

stats <- lapply(disagg, function(x) {
  sapply(x, function(y) {
    c(cellStats(y * 100, mean), quantile(y * 100, seq(0, 1, 0.05)))
  })
})
```

8.2.1.4 Aggregate adjusted rasters, compare cropland fractions

```r
lims <- c(ceiling(min(sapply(stats, function(x) x[3, ]))),
          floor(max(sapply(stats, function(x) x[21, ]))))
rng <- range(sapply(stats, range))
brks <- c(rng[1], -50, -30, -20, -10, -5, -1, 1, 5, 10, 20, 30, 50, rng[2])
cols <- colorRampPalette(c("red", "grey80", "blue4"))(length(brks) - 1)

legtext <- "% Difference"
cx <- 1.4
lcol <- "black"
mcap <- c("SA-LC", "GlobCover", "MODIS", "GeoWiki")
lev <- names(disagg[[1]])[-c(2:3)]
lev2 <- c("1 km", "25 km", "50 km", "100 km")
pdf(full_path(p_fig, "cropland_adj_bias_map2.pdf"), height = 6, width = 7)
par(mfrow = c(4, 4), mar = c(0, 0, 0, 0), oma = c(5, 5, 2, 0))
for(i in 1:length(snms)) {
  print(snms[i])
  for(j in 1:length(lev)) {
    print(lev[j])
    plot(sa.shp, lty = 0)
    plot(disagg[[snms[i]]][[lev[j]]] * 100, add = TRUE, col = cols,
         breaks = brks, legend = FALSE)
  if(j == 1) mtext(mcap[i], side = 2, line = 1, cex = cx)
  if(i == 1) mtext(lev2[j], side = 3, line = 0, cex = cx)
  }
}
flex_legend(ncuts = length(brks) - 1, legend.text = legtext,
            legend.vals = round(brks),
            longdims = c(0.2, 0.8), shortdims = c(0.06, 0.01),
            colvec = cols, #(length(brks) - 1),
            srt = c(270, 0), horiz = TRUE, textside = "bottom",
            legend.pos = c(4, 5), leg.adj = list(c(0.25, 0), c(0, -0.5)),
            cex.val = cx, textcol = lcol, bordercol = lcol)
dev.off()
```

8.2.1.5 Plot for supplementary data

8.2.2 Yield disaggregation

Following Monfreda et al's (2008) methods. They appear to have used the 2002 district census, so we will use the more recent 2007 census.

```
# Load in yield dataset and magisterial district polygons
load(full_path(p_data2, "ZAF_adm2.Rdata"))  # magisterial districts
yld <- fread(full_path(p_data, "maize_wheat_2007.csv"))

# Transform magisterial districts, remove Prince Edward Islands
md <- spTransform(gadm, CRSobj = sa.shp@proj4string)
md <- md[md$ID_2 != 313, ]
md@data <- md@data[, c("PID", "ID_2", "NAME_1", "NAME_2")]  # trim down columns
# fix a few names that occur more than once to make unique for merging properly
md@data[md$ID_2 == 43, "NAME_2"] <- "MiddelburgEC"
md@data[md$ID_2 == 143, "NAME_2"] <- "RichmondKZ"
md@data[md$ID_2 == 86, "NAME_2"] <- "HeidelbergG"

# mean MD area
round(mean(rgeos::gArea(md, byid = TRUE) / 1000000))  # 3445 km square

# check names in two datasets
# match(md@data$NAME_2, yld$district)
# match(yld$district, md@data$NAME_2)
# yld$district[which(!yld$district %in% md@data$NAME_2)]
# sort(md@data$NAME_2[which(!md@data$NAME_2 %in% yld$district)])

# some offline fixing of names ensues
# write.csv(md@data, file = "external/ext_data/mdrect.csv")
# writeOGR(md, dsn = "external/ext_data/mdcheck.sqlite", layer = "mdcheck",
#          driver = "SQLite")

# reread fixed data, merge a few districts' data where there are synonyms or 1
# district subsumed by another--sum yields, because these districts were
# probably broken into smaller pieces
# fix same names for merging properly
yld[district == "Middelburg" & province == "Eastern Cape",
    district := "MiddelburgEC"]
yld[district == "Richmond" & province == "KwaZulu-Natal",
    district := "RichmondKZ"]
yld[district == "Heidelberg" & province == "Gauteng",
    district := "HeidelbergG"]

# sumna <- function(x) sum(x, na.rm = TRUE)
mrg <- rbindlist(lapply(unique(yld$merge)[-1], function(x) {
  yld[merge == x, lapply(.SD, sumna), .SDcols = grep("mz|wh", names(yld))]
}))
nms <- sapply(unique(yld$merge)[-1], function(x) {
  nm <- yld[merge == x, district]
})
mnms <- unname(sapply(nms, function(x) x[x %in% md$NAME_2][1]))

keep <- names(yld)[-c(2:5)]
yld2 <- rbind(yld[!district %in% unname(unlist(nms)), keep, with = FALSE],
              cbind("district" = mnms, mrg))  # adjusted yield dataset

# merge with magisterial district data
mdyld <- sp::merge(md@data, yld2, by.x = "NAME_2", by.y = "district",
                   all.x = TRUE)  # merge
mdyld <- mdyld[match(mdyld$NAME_2, md@data$NAME_2), ]  # reorder rows correctly
nrow(mdyld) == nrow(md@data)  # check

# calculate yields and total planted ha
mdyld$mzha <- rowSums(mdyld[, grep("mz_ha", colnames(mdyld))])
mdyld$mzyld <- round(rowSums(mdyld[, grep("mz_pr", colnames(mdyld))]) /
                       mdyld$mzha, 1)
mdyld$whha <- rowSums(mdyld[, grep("wh_ha", colnames(mdyld))])
mdyld$whyld <- round(rowSums(mdyld[, grep("wh_pr", colnames(mdyld))]) /
                       mdyld$whha, 1)

# join to md data
m <- mdyld[match(md$ID_2, mdyld$ID_2),
           c("ID_2", "NAME_2", "mzha", "mzyld", "whha", "whyld")]
colnames(m)[1:2] <- c("id", "name")
md@data <- cbind(md@data, m)
all(md$NAME_2 == md$name); all(md$ID_2 == md$id)
md@data <- md@data[, -c(2:4)]

# plot to make sure districts didn't get reordered at all
par(mar = rep(0, 4))
plot(sa.shp)
plot(md, add = TRUE)
plot(md[md$id == 214, ], col = "red", add = TRUE)
plot(md[md$id == 327, ], col = "red", add = TRUE)
plot(md[md$id == 263, ], col = "red", add = TRUE)
plot(md[md$name == "Barberton", ], col = "red", add = TRUE)
```

8.2.2.1 Process magisterial districts and census data

8.2.2.2 Calculate individual crop fractions Rasterize MDs and calculate individual crop fractions (for maize), for both GTI and adjusted LC cropland fractions

```
# mdr <- rasterize(md, provr, field = "id",
#                   filename = "external/ext_data/mag_dist.tif", overwrite = TRUE)
mdr <- raster("external/ext_data/mag_dist.tif")
mds <- stack(list("md" = crop(mdr, gti_agg$f1$gti), "gti" = gti_agg$f1$gti,
                  lc_adj))  # use gti and lc_adj
# v <- values((lc_agg$f1$modmu > 1) * 1); sum(v[!is.na(v)])

# data.table calculations
md_dt <- as.data.table.raster(mds, xy = TRUE)
```

```r
setkey(md_dt, md)
mdarea <- md_dt[, .N, by = md]  # how many km2 in each MD
setnames(mdarea, "N", "mdkm2")
mdarea <- mdarea[!is.na(md)]  # remove NAs
setkey(mdarea, md)
md_dt <- na.omit(md_dt)  # remove NAs from fraction data.tables
varea <- md_dt[, .N, by = md]  # area of non-NA data in MDs
setnames(varea, "N", "vkm2")
setkey(varea, md)
# mdcf <- md_dt[, list("carea" = round(sum(f), 2), "parea" = length(f),
#                          "cf" = round(sum(f) / length(f), 4)), by = md]
mdva <- mdarea[varea][, fmd:= round(vkm2 / mdkm2, 2)]  # md valid area
# mdcf[, fmd := mdarea[varea][, round(vkm2 / mdkm2, 2)]]  # fraction non-NA in md

# yield data for MDs
ydt <- data.table(md@data[, c("id", "mzha", "mzyld", "whha", "whyld")])
setnames(ydt, "id", "md")
setkey(ydt, "md")
mdvay <- mdva[ydt][is.na(mzha), c("mzha", "mzyld") := 0]
mdvay[is.na(whha), c("whha", "whyld") := 0]
mdvay[, c("mdkm2", "vkm2") := NULL]
# mdcfy[, mfrac := round(((mzha * fmd) / 100) / carea, 4)]  # maize fraction

# calculate crop fractions for each dataset
md_dta <- copy(md_dt)
# md_dt[, lapply(.SD, mean), by = md, .SDcols = c("gti", snms)]
md_lcl <- lapply(c("gti", snms), function(j) {
  DT <- md_dta[, list("carea" = round(sum(get(j)), 2),  # crop area
                      "parea" = length(get(j)),  # non-NA area in MD
                      "cf" = round(sum(get(j)) / length(get(j)), 4)),
               by = md]
  # calculate crop fraction, adjusting ha first by how much non-NA area there
  # is in MD
  DTy <- DT[mdvay][, mfrac := round(((mzha * fmd) / 100) / carea, 4)]
  DTy[is.na(mfrac), mfrac := 0]  # set mfrac to 0 in 0 cropland areas
  # print(length(which(DTy$mfrac > 1)))  # 5, 4, 15, 9, 4
  #DTy[mfrac > 1, mfrac := 1]  # set to 1 - mostly missing data causing > 1
  DTy
})
names(md_lcl) <- c("gti", snms)
# lapply(md_lcl, function(x) x[, sum(mzha)])
#length(which(is.na(md_lcl$modmu$mfrac)))
#length(which(md_lcl$sa30$mfrac > 1))

par(mfrow = c(2, 3))
for(i in 1:5) hist(md_lcl[[i]]$mfrac)
for(i in 1:5) hist(md_lcl[[i]]$carea)
plot(md_lcl[[1]]$carea, md_lcl[[3]]$carea)
for(i in 1:5) print(length(which(md_lcl[[i]]$mfrac > 1)))
```

Note: when assigning crop fractions, previously we set all fractions > 1 to 1, but now turned this off because Monfreda et al (2010) allow for double-cropping. This is likely not correct, but we are doing it here to be consistent.

```
# first in data.tables
asnms <- c("gti", snms)
crop_ay <- lapply(c("gti", snms), function(j) {
  #j <- snms[3]
  DTya <- copy(md_dt)[, list(x, y, md, "f" = get(j))]
  DTya <- DTya[md_lcl[[j]][, list(md, fmd, mfrac, mzyld)]]
  DTya[, fa := f * mfrac]  # this is Monfreda et al equation (pg. 10)
  #print(DTya[which(round(DTya$fa, 4) > 1)[1:4], ])  # check line
  DTya[, mzya := mzyld]  # adjusted yield variable
  DTya[fa == 0, mzya := 0] # set to 0 in pixels having no crop
  DTya
})
names(crop_ay) <- asnms
sapply(crop_ay, function(x) x[, round(sum(fa), 4), by = md][, V1])

# then back to rasters for aggregation
crop_ayr <- lapply(crop_ay, function(x) {
  dt_to_raster(x[, list(x, y, fa, mzya)], CRSobj = sa.shp@proj4string)
})

# redo NA mask because of some slight shifts
namask <- calc(stack(lapply(crop_ayr, function(x) x$fa)), sumna)
namask[namask >= 0] <- 1
namask2 <- !is.na(namask)  # set NAs to zero
# cellStats(!is.na(gti), sum); cellStats(namask2, sum)
awgts <- aggregate_rast_list(fact, list(namask2), fun = sum)
awgts <- lapply(awgts, function(x) x[[1]])  # unlist on inner loop
# sumna <- function(x, na.rm = na.rm) sum(x, na.rm = na.rm)
# amsk <- aggregate_rast_list(fact, list(namask2), fun = sumna)
# amsk <- lapply(amsk, function(x) x[[1]])  # unlist on inner loop
```

8.2.2.3 Assign back crop fraction and yield to grid

```
# Yield aggregation - aggregating with weighting by crop fraction
# using a custom data.table function to allow this
lev <- names(disagg[[1]])
wmfun <- parse(text = "sum((mzya * fa) / sum(fa, na.rm = TRUE), na.rm = TRUE)")
sr <- sa.shp@proj4string
d <- c(dim(gti_agg[[1]]$gti)[1:2], xres(gti_agg[[1]]$gti))  # dimensions
cyld <- lapply(lev[-1], function(i) {
  print(i)
  f <- as.integer(gsub("f", "", i))
  il <- lapply(c("gti", snms), function(j) {
    print(paste("...", j))
    rdt <- as.data.table(crop_ayr[[j]], xy = TRUE)
    o <- dt_aggregate(rdt, d[1], d[2], f, wmfun, d[3], "yw")
  })
  named_out(il, asnms)
})
```

```
names(cyld) <- lev[-1]

# convert back to rasters
cyldr <- lapply(cyld, function(i) lapply(i, function(j) dt_to_raster(j, sr)))
cyldf1 <- lapply("f1", function(i) {
  f1yld <- lapply(asnms, function(j) crop_ayr[[j]]$mzya)
  named_out(f1yld, asnms)
})
names(cyldf1) <- lev[1]
cyld_agg <- c(cyldf1, cyldr)
# plot(cyld_agg$f10$gti)

# check Monfreda's yield data to see how it deals with null areas
monf <- full_path("external/ext_data/monfreda_data/maize_HarvAreaYield_NetCDF",
                  "maize_AreaYieldProduction.nc")
monfmz <- brick(monf, level = 2)
plot(crop(monfmz, spTransform(sa.shp, monfmz@crs)), col = bpy.colors(20))
plot(crop(monfmz, spTransform(sa.shp, monfmz@crs)) > 0)  # almost all SA > 0

# Removed other versions of yield aggregation. Refer to repo prior to 18/10 to
# to recover

## crop areas
carea <- lapply(1:5, function(x) crop_ayr[[x]]$fa)
names(carea) <- c("gti", snms)

# sumha <- function(x, na.rm) sum(x, na.rm)
carea_agg <- aggregate_rast_list(fact, carea) # crop area, mean aggregation
#carea_agg2 <- aggregate_rast_list(fact, carea, sum) # crop area, sum agg
# plot(carea_agg$f25$globmu)
# plot(carea_agg2$f25$globmu)

# area of each pixel at each aggregation scale
agg_res <- sapply(carea_agg, function(x) res(x$gti)[1]^2 / 10000)
tareas <- lapply(awgts, function(x) x[[1]] * 100)

# plot(tareas$f50)
#plot((carea_agg$f50$gti * tareas$f50) - (carea_agg2$f50$gti * 100))  # equiv
# ta <- rep(1, 50 * 50)
# a <- runif(50 * 50, 0, 1)
# (sum(a) * 100) / (sum(ta) * 100)
# mean(a) * (sum(ta) * 100)

# Removed yat check. Refer to repo prior to 18/10 to recover

# check aggregated maize fractions
for(i in lev) {
  for(j in snms) {
    ck <- crop_ay[[j]][, sum(fa)] * 100
    d <- round((ck - cellStats(tareas[[i]] * carea_agg[[i]][[j]], sum)) / ck,2)
    print(d)
  }
}  # all equal
```

8.2.2.4 Aggregate yields and crop area to coarser resolutions

8.2.2.5 Crop production estimates at different aggregation scales

Including differences relative to reference versions.

```
# Type A aggregation: multiplying aggregated yields by aggregated fractions
# with crop-fraction weighted aggregate yields
cpagg1 <- lapply(1:length(cyld_agg), function(x) {
  p <- lapply(1:length(cyld_agg[[x]]), function(y) {
    (tareas[[x]] * carea_agg[[x]][[y]]) * cyld_agg[[x]][[y]]
  })
  named_out(p, asnms)
})
names(cpagg1) <- names(agg_res)

# Removed production aggregation with 0-removed yieldsother. Refer to repo prior
# to 18/10 to recover

# Type B (formerly II) aggregation: crop production differences when aggregated
# from 1 km calculate production at 1 km
cprod <- lapply(1:length(cyldf1$f1), function(x) {
  cyldf1$f1[[x]] * (carea[[x]] * 100)
})
names(cprod) <- asnms
cpaggII <- aggregate_rast_list(fact, cprod, "sum") # aggregate production

# check production estimates for consistency
# pat <- list(cpagg1, cpagg3, cpaggII)
# ptype <- c("1", "3", "II")
pat <- list(cpagg1, cpaggII)
ptype <- c("1", "II")

# check production estimates
for(i in lev[-1]) {
  print(paste("factor", i))
  for(j in asnms) {
    print(paste("...", j))
    for(k in 1:length(ptype)) {
      print(paste("...... type", ptype[k], "agg"))
      ck <- cellStats(cprod[[j]], sum)  # weighted by fraction
      ck2 <- cellStats(cprod$gti, sum) # weighted by fraction
      v <- cellStats(pat[[k]][[i]][[j]], sum)
      d <- round((ck - v) / ck, 2)
      d2 <- round((ck2 - v) / ck2, 2)
      print(c(d, d2))
    }
  }  # all types consistent across scales
}

# sapply(cprod, function(x) cellStats(x, sum))
# sapply(carea, function(x) cellStats(x, sum))

DT <- na.omit(as.data.table.raster(s, xy = TRUE))
```

```
DT[, wgt := ref * awgts]  # calculate weights from ref and area weights

# Mean production in producing cells at different levels of aggregation.
# No need to weight by cropland density, because production value already factors
# in area - just mask out zero production areas
mup <- sapply(lev, function(i) {
  msk <- cpagg1[[i]]$gti > 0
  msk[msk == 0] <- NA
  cpmsk <- msk * cpagg1[[i]]$gti
  cellStats(cpmsk, mean)
})

# mean yield in producing cells at different levels of aggregation, use density
# weighting here
muy <- sapply(lev, function(i) {
  # i <- lev[2]; print(i)
  s <- stack(cyld_agg[[i]]$gti, carea_agg[[i]]$gti, awgts[[i]])
  names(s) <- c("yld", "ref", "awgts")
  DT <- na.omit(as.data.table.raster(s, xy = TRUE))
  DT[, wgt := ref * awgts]  # calculate weights from ref and area weights
  DT[, weighted.mean(yld, wgt)]
})  # yield is 3.361 kg/ha at all scales

# Removed original mup and muy. Refer to repo prior to 18/10 to recover

# calculate differences between them
# check first that cpagg1 and II are equivalent, make cpaggII master
for(i in lev) {
  for(j in snms) {
    a <- (cpagg1[[i]]$gti - cpagg1[[i]][[j]]) -
      (cpaggII[[i]]$gti - cpagg1[[i]][[j]])
    print(round(cellStats(a, sum), 5))
  }
}

# Production differences
# first standardize
# std_raster <- function(x) {
#   (x - cellStats(x, min)) / diff(cellStats(x, range))
# }
# cpagg_std <- lapply(cpagg1, function(x) lapply(x, function(y) std_raster(y)))
ilist <- list(names(cpagg1[[1]])[-1], names(cpagg1), "gti")
# list1 <- lapply(cpagg1, function(x) x[1])
# list2 <- lapply(cpagg1, function(x) x[-1])
list1 <- lapply(cpagg1, function(x) x[1])
list2 <- lapply(cpagg1, function(x) x[-1])
p_diff1 <- rast_list_math(ilist, list1, list2, expr = "a - b")  # type 1 agg

# Removed original p_diff2/3. Refer to repo prior to 18/10 to recover

# yield differences
list1 <- lapply(cyld_agg, function(x) x[1])
list2 <- lapply(cyld_agg, function(x) x[-1])
```

```
y_diff1 <- rast_list_math(ilist, list1, list2, expr = "a - b")  # type 1

# check grids (output in QGIS and examine for sensible results--they are)
# for(i in lev) {
#   writeRaster(y_diff1$globmu[[i]]$gti,
#               file = full_path(p_data, paste0("test/globdiff", i, ".tif")))
# }
# for(i in lev) {
#   writeRaster(list1[[i]]$gti,
#               file = full_path(p_data, paste0("test/gtiyldtst", i, ".tif")))
# }
# for(i in lev) {
#   writeRaster(list2[[i]]$globmu,
#               file = full_path(p_data, paste0("test/globyld2tst", i, ".tif")))
# }

# Removed original y_diff2. Refer to repo prior to 18/10 to recover

# convert them to percent differences relative to mean pixel-wise production
# here only for map plotting.

pdiffs <- lapply(list(p_diff1), function(x) {
  l1 <- lapply(snms, function(y) {
    l2 <- lapply(lev, function(z) {
      x[[y]][[z]]$gti / mup[z] * 100
    })
    named_out(l2, lev)
  })
  named_out(l1, snms)
})
names(pdiffs) <- "pd1"

ydiffs <- lapply(list(y_diff1), function(x) {
  l1 <- lapply(snms, function(y) {
    l2 <- lapply(lev, function(z) {
      x[[y]][[z]]$gti / muy[z] * 100
    })
    named_out(l2, lev)
  })
  named_out(l1, snms)
})
names(ydiffs) <- "yd1"

# disaggregate for plotting
lev <- names(y_diff1[[1]])
p_diff1d <- disaggregate_rast_list(snms, lev, pdiffs$pd1, namask)
y_diff1d <- disaggregate_rast_list(snms, lev, ydiffs$yd1, namask)

# stats for plotting
dstats <- function(dlist) {
  statsd <- lapply(dlist, function(x) {
    sapply(x, function(y) {
      c(cellStats(y, mean), quantile(y, seq(0, 1, 0.05)))
    })
  })
}
stats_pd1 <- dstats(p_diff1d)
stats_yd1 <- dstats(y_diff1d)
```

8.2.3 Plot difference maps

```
dnms <- ls()[grep("[y_|p_]diff*.*d", ls())]
stnms <- ls()[grep("stats_", ls())]
disaggl <- lapply(dnms, function(x) get(x))
statsl <- lapply(stnms, function(x) get(x))
rng <- lapply(statsl, function(x) {
  ilims <- c(ceiling(min(sapply(x, function(y) y[3, ]))),
             floor(max(sapply(x, function(y) y[21, ]))))
  olims <- range(sapply(x, function(y) range(y)))
  c(olims[1], ilims, olims[2])
})

fnms <- c("prod_bias_map.pdf", "yld_bias_map.pdf")  # output names
legtext <- rep("% Difference", 5)
cx <- 1.4
lcol <- "black"
mcap <- c("SA-LC", "GlobCover", "MODIS", "GeoWiki")
lev <- names(p_diff1d[[1]])[-c(2:3)]
lev2 <- c("1 km", "25 km", "50 km", "100 km")

rnder <- function(x) {
  dig <- nchar(max(round(x)))
  rmat <- cbind(1:8, c(1, 5, 10, 20, 50, 100, 100, 1000))
  rnd <- rmat[dig, 2]
  round(c(floor(x[1] / rnd), ceiling(x[2] / rnd))) * rnd
}
div_breaks <- function(olim, ilim, ibrk) {
  il <- rnder(ilim)
  ol <- c(floor(olim[1]), ceiling(olim[2]))
  brks <- c(ol[1], seq(il[1], -ibrk, abs(floor(il[1]) - -ibrk) / 4),
            seq(ibrk, il[2], abs(ceiling(il[2]) - ibrk) / 4), ol[2])
  brks
}

rng2 <- lapply(rng, function(x) {
  x[2:3] <- c(-100, 100)
  x
})
brksl <- lapply(rng2, function(x) {
  c(floor(x[1:2]), c(-50, -25, -10, -5, -1, 1, 5, 10, 25, 50), ceiling(x[3:4]))
})
# brksl <- list(c(-1, -0.75, -0.5, -0.25, -0.1, -0.05, -0.01, 0.01, 0.05, 0.1,
#                 0.25, 0.5, 0.75, 1),
#               c(-10.7, -7, -5, -3, -2, -1, -0.1, 0.1, 1, 2, 3, 5, 7, 10.7))
```

```
for(i in 1:length(statsl)) {
  # i <- 2
  print(paste("figure", i, "of", length(statsl)))
  brks <- brksl[[i]]
  cols <- c(rev(brewer.pal(length(brks) / 2, "Reds")[-1]), "grey80",
            brewer.pal(length(brks) / 2, "Blues")[-1])
  brklen <- length(brks) - 1
  pdf(full_path(p_fig, fnms[i]), height = 6, width = 7)
  par(mfrow = c(4, 4), mar = c(0, 0, 0, 0), oma = c(5, 5, 2, 0))
  for(j in 1:length(snms)) {
    print(snms[j])
    for(k in 1:length(lev)) {
      rl <- disaggl[[i]]
      print(lev[k])
      plot(sa.shp, lty = 0)
      plot(rl[[snms[j]]][[lev[k]]], add = TRUE, col = cols,
           breaks = brks, legend = FALSE)
      if(k == 1) mtext(mcap[j], side = 2, line = 1, cex = cx)
      if(j == 1) mtext(lev2[k], side = 3, line = 0, cex = cx)
    }
  }
  flex_legend(ncuts = length(brks) - 1, legend.text = legtext[i],
              legend.vals = brks, longdims = c(0.15, 0.8),
              shortdims = c(0.07, 0.01),
              colvec = cols, #(length(brks) - 1),
              srt = c(270, 0), horiz = TRUE, textside = "bottom",
              legend.pos = c(4, 5), leg.adj = list(c(0.25, 0), c(0, -0.5)),
              cex.val = cx, textcol = lcol, bordercol = lcol)
  dev.off()
}

lev <- names(p_diff1d[[1]])  # reset levs
```

Several versions of yield aggregation were tested in this analysis, each of which has a different impact on yield and production impacts. There are three types of yield aggregation that can be done (most commented out now after initial checking):

1. **Type 1**: weighted mean averaging, with the weights provided by crop fractions. This produces correct country-level average yields (<1-2% different) across scales and between datasets, provided that the averaging is weighted by the cropland fractions that were aggregated to the level from which the country-level yield estimate is being calculated

2. **Type 2**: Straight averaging, which results in diluted country-level yield estimates. Incorrect, and should be obvious that this wouldn't be done, so we haven't pursued it here.

3. **Type 3**: or zero-removed averaging, where only areas having fraction of that crop contribute to the aggregated average. As with **Type 1**, this produces correct country level estimates when weighting by cropland fraction, but it does not produce correct production estimates, probably because it inflates the value of high-yielding irrigated areas out towards the dry west, where the crop fractions are small. (see below)

There are also two ways of aggregating production estimates:

1. **Type A**: Aggregating yield and crop area separately, then multiplying. This is correct with **Type 1** yield aggregation, and producing identical production estimates to **Type II**, but not with the other two types of yield aggregation.
2. **Type B**: Calculating production at the base resolution, then aggregating by sum. This should be the correct way of doing it.

Why would someone do a yield aggregation, when you have the gridded estimates? To calculate values and compare them to a coarser resolution product, for instance, or to get a country-level average.

Causes for bias concern to examine:

- Bias in yield gap estimates, and how much closing that could affect contribution of closing gaps to boosting production.
- Country-level statistics might cancel out bias, but if bias of one type of sign is spatially correlated with the driver of values in yield gaps, then that would bias larger-scale estimates of gap closure potential.
- These concerns apply to all resolutions.

Type 2 and 3 yield aggregation are incorrect methods, and are not retained here (but see earlier commits of code, specifically prior to 18/10/2015. The other variants were used for comparison and for initial methods checks only.

8.3 Bias/MAE statistics

8.3.1 Primary method

Density weighted, following methods developed in cropland-bias, but here metrics are based on residuals rather than percent errors, because of infinite values (harvested area/yield disaggregation methods mean that there are gridded yields/production estimates in some areas where no reference value exists, leading to infinite values if converting to percent)

```
# yield bias/accuracy
# reshape
yerr <- lapply(lev, function(x) {
  sapply(y_diff1, function(y) list(y[[x]]$gti))
})
names(yerr) <- lev

# carea_gti <- lapply(carea_agg, function(x) x$gti)
a <- bias_statsw_list(gti_agg, awgts, yerr, snms, wm, "mu")
b <- bias_statsw_list(gti_agg, awgts, yerr, snms, wma, "mua")
yld_bacc <- rbind(a, b)

# production bias/accuracy
# reshape
perr <- lapply(lev, function(x) {
  sapply(p_diff1, function(y) list(y[[x]]$gti))
})
names(perr) <- lev

# carea_gti <- lapply(carea_agg, function(x) x$gti)
a <- bias_statsw_list(gti_agg, awgts, perr, snms, wm, "mu")
```

```
b <- bias_statsw_list(gti_agg, awgts, perr, snms, wma, "mua")
prod_bacc <- rbind(a, b)

# check error stats - do they match alternate approach?
ref <- lapply(gti_agg, function(x) x$gti)
test_err <- "perr" #"yerr"
evst <- "wma" # "wm"
for(i in c("f1", "f25", "f10")) {
  for(j in c("sa30", "modmu", "geow")) {
    print(paste("cross-checking calculations in", i, j))
    a1 <- bias_statsw(ref[[i]], awgts[[i]], get(test_err)[[i]], snms,
                      get(evst), aweight = FALSE)[, j, with = FALSE]
    a2 <- bias_statsw(ref[[i]], awgts[[i]], get(test_err)[[i]], snms,
                      get(evst), rweight = FALSE,
                      aweight = FALSE)[, j, with = FALSE]
    a3 <- bias_statsw(ref[[i]], awgts[[i]], get(test_err)[[i]], snms,
                      get(evst))[, j, with = FALSE]
    s <- stack(get(test_err)[[i]][[j]], ref[[i]], awgts[[i]])
    sv <- getValues(s)
    sv <- sv[which(!is.na(rowSums(sv))), ]  # remove NAs
    if(evst == "wm") {
      print(a1 == round(weighted.mean(sv[, 1], sv[, 2], na.rm = TRUE), 2))
      print(a2 == round(mean(sv[, 1]), 2))
      print(a3 == round(weighted.mean(sv[, 1], sv[, 2] * sv[, 3]), 2))
    } else if(evst == "wma"){
      print(a1 == round(weighted.mean(abs(sv[, 1]), sv[, 2], na.rm = TRUE), 2))
      print(a2 == round(mean(abs(sv[, 1])), 2))
      print(a3 == round(weighted.mean(abs(sv[, 1]), sv[, 2] * sv[, 3]), 2))
    }
  }
}
```

8.3.2 Secondary method

Bias and accuracy within agricultural areas

```
# create mask for non-cropland areas, unioning GTI and LC areas with maize
# estimates -- probably not needed if doing production estimates.
lev <- names(y_diff1[[1]])

# first check aggregation to make yield aggregation variants have common area,
# so that just one lc_union is needed
# for(i in lev) {
#   for(j in snms) {
#     print(cellStats(cyld_agg[[i]][[j]] > 0 & cyld_agg3[[i]][[j]] == 0, sum))
#     #print(cellStats(carea_agg[[i]][[j]] > 0 & cyld_agg[[i]][[j]] == 0, sum))
#   }
# }  # all zeros

snms <- names(p_diff1)
lcu <- lapply(lev, function(x) {
  lcb <- lapply(snms, function(y) {
    gti_gt0 <- Which(carea_agg[[x]][[1]] > 0)
    lc_gt0 <- Which(carea_agg[[x]][[y]] > 0)
    all_gt0 <- gti_gt0 + lc_gt0
    all_gt0[all_gt0 > 0] <- 1
    all_gt0
  })
  named_out(lcb, snms)
})
names(lcu) <- lev

# cropland cover bins, for looking at error as a function of cover
binv <- seq(0, 1, 0.05)
bins <- lapply(gti_agg, function(x) {
  cut(x$gti, breaks = binv, include.lowest = TRUE)
})

# Many commented out checks removed here. Look at pre 18/10 commits to restore

# Spatial bias stats calculation on yield and production differences, not on
# percentages
# Many stats for alternate yield/production metrics removed here.
# Look at pre 18/10 commits to restore
a <- bias_stats_list(bins, awgts, lcu, p_diff1, wm, "mu", "bias", TRUE)
b <- bias_stats_list(bins, awgts, lcu, p_diff1, wma, "mua", "bias", TRUE)
d <- bias_stats_list(bins, awgts, lcu, p_diff1, sum, "sum", "bias", FALSE)
pd1_st <- rbind(a, b, d)

a <- bias_stats_list(bins, awgts, lcu, y_diff1, wm, "mu", "bias", TRUE)
b <- bias_stats_list(bins, awgts, lcu, y_diff1, wma, "mua", "bias", TRUE)
yd1_st <- rbind(a, b)

# Output statistics
p1mu <- extract_stat(lev, snms, "all", "mu", "bias", pd1_st)  # identical to p3
p1ma <- extract_stat(lev, snms, "all", "mua", "bias", pd1_st)  # ident to p3
y1mu <- extract_stat(lev, snms, "all", "mu", "bias", yd1_st)
y1ma <- extract_stat(lev, snms, "all", "mua", "bias", yd1_st)

# Older variant of code from compare-landcover.Rmd to evaluate whether newer DT
# version is finding correct results
i <- y_diff1 #pdiff_3 # p_diff1
jdt <- yd1_st # pd3_st # pd1_st
bv <- "mua" # bv <- "mu"
for(chk in list(c("modmu", "f25"), c("geow", "f10"), c("sa30", "f50"),
                c("globmu", "f1"))) {
  x <- chk[1]
  y <- chk[2]
  print(paste(".", x, "..", y))
  rs <- lcu[[y]][[x]]
  rs[rs == 0] <- NA
  l3 <- abs(i[[x]][[y]][[1]])
  # l3 <- i[[x]][[y]][[1]]
  o <- rs * l3
  w <- awgts[[y]][[1]] * rs
  wmu <- weighted.mean(getValues(o), getValues(w), na.rm = TRUE)
  print(jdt[ol == y & il == x & bvals == bv & bin == "all", bias] ==
          round(wmu, 2))
}  # check
```

8.3.3 Bias/MAE plots

```
alph <- c(225, 60)
x <- c(0, 3, 6, 9, 12, 15)
w <- 3 / 8
xo <- (cumsum(rep(w, 8)) - w / 2)[-c(2, 4, 6, 8)]
o <- c(0, w)
xa <- sapply(x, function(x) x + xo)
cx <- c(1.25, 1)
g1 <- "grey90"

# get stats data into shape
mulw <- list(prod_bacc[stnm == "mu"], yld_bacc[stnm == "mu"])
malw <- list(prod_bacc[stnm == "mua"], yld_bacc[stnm == "mua"])
malmulw <- list(malw, mulw)  # density-weighted stats, reordered
pyl <- c("mup", "muy")

xl <- c(-0.5, 18)
yl <- c(-30, 110)  # try it out, play with it
shd <- c(4.5, 10.5, 16.5)
cols <- c("red", "orange3", "green4", "blue")
lcnms <- c("SA LC", "GlobCover", "MODIS", "Geowiki")
pchs <- c("+", "o")

yax <- seq(yl[1], yl[2], 10)
xax <- seq(1, 6, 5 / (length(yax) - 1))

pdf(full_path(p_fig, "yield_prod_bias.pdf"), height = 7, width = 7)
par(mar = rep(1, 4), oma = c(2, 2, 0, 0), mgp = c(1, 0.5, 0), tcl = -0.3)
plot(xl, yl, pch = "", yaxt = "n", xaxt = "n", xaxs = "i", yaxs = "i",
     ylab = "", xlab = "")#,
for(i in shd) polyfunc2(x = i, y = yl, w = 3, col = g1, bcol = g1, lwd = 1)
abline(h = yax, v = NULL, col = "grey80", lty = 1)
polyfunc2(x = 8.75, y = yl, w = 18.5, col = "transparent", bcol = "black")
lines(c(-1, 18), c(0, 0), lwd = 2, col = "grey80")
for(ii in 1:length(lev)) { # ii <- 1; i <- 1; k <- 1
  lv <- lev[ii]
  for(i in 1:length(snms)) {
    pchs1 <- pchs
    nm <- snms[i]
    for(k in 1:length(malmulw)) {
      mm <- malmulw[[k]]
      # fetch bias stats, convert to percentage relative to mean
      # e.g. GTI mean yield
      v <- sapply(1:length(mm), function(j) {  # j <- 1
        round(mm[[j]][ol == lv, get(nm)] / get(pyl[j])[lv] * 100, 1)
      })
      pcol <- makeTransparent(cols[i], alpha = alph[k])
      polyfunc2(xa[i, ii] + o[k], range(v), w = w, col = pcol, bcol = pcol)
      xx <- rep(x[ii] + xo[i] + o[k], length(v))
      points(xx, v, pch = pchs, cex = cx)
    }
  }
}
axis(1, at = seq(1.5, 18.5, 3), labels = c(1, fact))
axis(2, at = yax, labels = yax, las = 2)
mtext(side = 1, text = "Resolution (km)", outer = TRUE, line = 0.5)
mtext(side = 2, text = "Bias/MAE (%)", outer = TRUE, line = 1)
legend(x = 12.4, y = -10, legend = lcnms, pch = 15, col = cols, adj = 0,
       pt.cex = 1.5, bty = "n", cex = 0.8, x.intersp = 0.5)
legend(x = 11.9, y = -10, legend = rep("", 4), pch = 15, adj = 0,
       col = makeTransparent(cols, alpha = alph[2]), pt.cex = 1.5, bty = "n",
       cex = 0.8)
text(x = 12.7, y = -12, labels = "MAE", srt = 45, adj = c(0, 0), cex= 0.8)
text(x = 12.2, y = -12, labels = "Bias", srt = 45, adj = c(0, 0), cex = 0.8)
legend(x = 12.1, y = 100, legend = c("Production", "Yield"), pch = pchs,
       pt.cex = c(1.5, 1.5), bty = "n", cex = 0.8)
dev.off()
```

8.3.3.1 Density-weighted

8.3.3.2 Agricultural area Supplemental

```
alph <- c(225, 60)

mula <- c("p1mu", "y1mu")
mala <- c("p1ma", "y1ma")
malmula <- list(mala, mula)
pyl <- c("mup", "muy")

yl <- c(-70, 80)
yax <- seq(yl[1], yl[2], 10)

pdf(full_path(p_fig, "yield_prod_bias_agric.pdf"), height = 7, width = 7)
par(mar = rep(1, 4), oma = c(2, 2, 0, 0), mgp = c(1, 0.5, 0), tcl = -0.3)
plot(xl, yl, pch = "", yaxt = "n", xaxt = "n", xaxs = "i", yaxs = "i",
     ylab = "", xlab = "")
for(i in shd) polyfunc2(x = i, y = yl, w = 3, col = g1, bcol = g1, lwd = 1)
abline(h = yax, v = NULL, col = "grey80", lty = 1)
polyfunc2(x = 8.75, y = yl, w = 18.5, col = "transparent", bcol = "black")
lines(c(-1, 18), c(0, 0), lwd = 2, col = "grey80")
for(ii in 1:length(lev)) {
  lv <- lev[ii]
  for(i in 1:length(snms)) {
    pchs1 <- pchs
    nm <- snms[i]
    for(k in 1:length(malmula)) {
      mm <- malmula[[k]]
      # fetch bias stats, convert to percentage relative to mean
      # e.g. GTI mean yield
```

```r
      v <- sapply(1:length(mm), function(j) {
        round(get(mm[j])[ol == lv & il == nm,  bias] / get(pyl[j])[lv] * 100,1)
      })
      pcol <- makeTransparent(cols[i], alpha = alph[k])
      polyfunc2(xa[i, ii] + o[k], range(v), w = w, col = pcol, bcol = pcol)
      xx <- rep(x[ii] + xo[i] + o[k], length(v))
      points(xx, v, pch = pchs, cex = cx)
    }
  }
}
axis(1, at = seq(1.5, 18.5, 3), labels = c(1, fact))
axis(2, at = yax, labels = yax, las = 2)
mtext(side = 1, text = "Resolution (km)", outer = TRUE, line = 0.5)
mtext(side = 2, text = "Bias/MAE (%)", outer = TRUE, line = 1)
legend(x = 12.4, y = -40, legend = lcnms, pch = 15, col = cols, adj = 0,
       pt.cex = 1.5, bty = "n", cex = 0.8, x.intersp = 0.5)
legend(x = 11.9, y = -40, legend = rep("", 4), pch = 15, adj = 0,
       col = makeTransparent(cols, alpha = alph[2]), pt.cex = 1.5, bty = "n",
       cex = 0.8)
text(x = 12.7, y = -42, labels = "MAE", srt = 45, adj = c(0, 0), cex= 0.8)
text(x = 12.2, y = -42, labels = "Bias", srt = 45, adj = c(0, 0), cex = 0.8)
legend(x = 12.1, y = 65, legend = c("Production", "Yield"), pch = pchs,
       pt.cex = c(1.5, 1.5), bty = "n", cex = 0.8)

dev.off()

lapply(1:length(lev), function(ii) {
  lv <- lev[ii]
  lapply(1:length(snms), function(i) {
    nm <- snms[i]
    lapply(1:length(malmula), function(k) {
      mm <- malmula[[k]]
      v <- sapply(1:length(mm), function(j) {
        round(get(mm[j])[ol == lv & il == nm,  bias] / get(pyl[j])[lv] * 100,1)
      })
    })
  })
})

do.call(cbind, lapply(lev, function(x) {
  rbind(cbind.data.frame("Map" = y1mu[ol == x, il],
                         "v" = round(y1mu[ol == x, bias] / muy[x] * 100, 1)),
        cbind.data.frame("Map" = p1mu[ol == x, il],
                         "v" = round(p1mu[ol == x, bias] / mup[x] * 100, 1)))
}))
```

8.3.4 Supplemental tables

```r
vnms <- c(rep("Yield", length(snms)), rep("Production", length(snms)))
aa <- c("Bias", "MAE")

# Density-weighted
```

```
out_tabw <- do.call(rbind, lapply(1:2, function(x) {  # x <- 1
  out <- do.call(cbind, lapply(1:length(lev), function(ii) {  # ii <- 1
    lv <- lev[ii]
    mm <- malmulw[2:1][[x]]
    v <- do.call(rbind, lapply(length(mm):1, function(j) {  # j <- 1
      v2 <- do.call(rbind, lapply(1:length(snms), function(i) { # i <- 1
        nm <- snms[i]
        round(mm[[j]][ol == lv, get(nm)] / get(pyl[j])[lv] * 100, 1)
      }))
      rownames(v2) <- snms
      v2
    }))
  }))
  out <- cbind.data.frame(aa[x], mcap, "Variable" = vnms, unname(out))
  colnames(out) <- c("Metric", "Map", "Variable", paste(c(1, fact), "km"))
  out
}))

# Agricultural area
out_taba <- do.call(rbind, lapply(1:2, function(x) {
  out <- do.call(cbind, lapply(1:length(lev), function(ii) {
    lv <- lev[ii]
    mm <- malmula[2:1][[x]]
    v <- do.call(rbind, lapply(length(mm):1, function(j) {
      v2 <- do.call(rbind, lapply(1:length(snms), function(i) {
        nm <- snms[i]
        round(get(mm[j])[ol == lv & il == nm,  bias] / get(pyl[j])[lv] * 100,1)
      }))
      rownames(v2) <- snms
      v2
    }))
  }))
  out <- cbind.data.frame(aa[x], mcap, "Variable" = vnms, unname(out))
  colnames(out) <- c("Metric", "Map", "Variable", paste(c(1, fact), "km"))
  out
}))

# Join them
out_tab <- rbind(cbind("Region" = "Density", out_tabw),
                 cbind("Region" = "Agricultural", out_taba))

caption <- paste("Biases and mean absolute errors (MAE) in disaggregated maize",
                 "yield and production (calculated from disaggregated yield",
                 "and harvested area estimates)",
                 "maps. Values for both density-weighted and agricultural",
                 "areas bias and accuracy are presennted.",
                 "Bias and MAE were normalized to their",
                 "respective mean values calculated from reference maps.")
out_xtab <- xtable(out_tab, digits = 1, caption = caption)
print(out_xtab, type = "latex", file = "paper/figures/yldprod-bias.tex",
      tabular.environment = "longtable", floating = FALSE,
      caption.placement = "top", include.rownames = FALSE)
```

8.3.5 The impacts of spatial patterns in bias

(Supplmentary analysis not presented). Using yield gap estimates as an indicator. First we'll bring in mean rainfall as a means of calculating a spatial correlation in yield gap estimates. Using data from the Princeton Global Forcing dataset, downloaded from the Africa Flood and Drought Monitor. Note: currently this is only explored here.

```r
# rainfall layer prepared for SA crop subsidy analysis
load("external/ext_data/climate/mag_dist_clim.rda")
pre <- calc(pre2, mean)  # 31 year mean rainfall for Africa
p4s <- projection(raster(nrow = 10, ncol = 10)) # GCS
presa <- crop(pre, spTransform(sa.shp, p4s))  # crop to SA
projection(presa) <- p4s
presa <- projectRaster(presa, lcu$f25$sa30)  # project to Albers
presa <- mask(presa, awgts$f25, maskvalue = 0)  # mask
presa1k <- disaggregate(presa, fact = 25)

# reclassify rainfalls to nearest 100 mm, tweak to catch lower and upper bound
rclmat <- cbind(seq(50, 750, 100), seq(150, 850, 100), seq(100, 800, 100))
rclmat[1] <- 15
rclmat[8, 2] <- 875
prebin <- crop(reclassify(presa1k, rclmat), namask)  # rainfall bins

# Assume a yield gap over-estimated by 100% in a particular location, say the
# 300-400 mm isohyet
prez <- (prebin == 300) | (prebin == 400)
notprez <- (prebin < 300) | (prebin > 400)
g1 <- 0
g2 <- 1

ygap_bias <- sapply(lev, function(x) {
  if(x != "f1") {
    preagg <- aggregate(prebin, fact = as.numeric(gsub("f", "", x)),
                        fun = modal)
  } else {
    preagg <- prebin
  }
  prez <- (preagg == 300) | (preagg == 400)
  notprez <- (preagg < 300) | (preagg > 400)
  sapply(snms, function(y) {
    yg <- (g1 * cyld_agg[[x]]$gti * notprez) + (g2 * cyld_agg[[x]]$gti * prez)
    yg2 <- (g1 * cyld_agg[[x]][[y]] * notprez) + (g2*cyld_agg[[x]][[y]] * prez)
    pg <- tareas[[x]] * carea_agg[[x]]$gti * yg
    pg2 <- tareas[[x]] * carea_agg[[x]][[y]] * yg2
    cellStats(pg - pg2, sum) / cellStats(pg, sum) * 100
  })
})

# Assume a yield gap over-estimated by 100% in 400 mm isohyet
ygap_bias2 <- sapply(lev, function(x) {
  if(x != "f1") {
    preagg <- aggregate(prebin, fact = as.numeric(gsub("f", "", x)), fun=modal)
  } else {
    preagg <- prebin
  }
  prez <- preagg == 400
  notprez <- preagg != 400
  sapply(snms, function(y) {
    yg <- (g1 * cyld_agg[[x]]$gti * notprez) + (g2 * cyld_agg[[x]]$gti * prez)
    yg2 <- (g1 * cyld_agg[[x]][[y]] * notprez) + (g2 * cyld_agg[[x]][[y]]*prez)
    pg <- tareas[[x]] * carea_agg[[x]]$gti * yg
    pg2 <- tareas[[x]] * carea_agg[[x]][[y]] * yg2
    cellStats(pg - pg2, sum) / cellStats(pg, sum) * 100
  })
})

# Assume a yield gap over-estimated by 100% in 500 mm isohyet
ygap_bias3 <- sapply(lev, function(x) {
  if(x != "f1") {
    preagg <- aggregate(prebin, fact = as.numeric(gsub("f", "", x)), fun=modal)
  } else {
    preagg <- prebin
  }
  prez <- preagg == 500
  notprez <- preagg != 500
  sapply(snms, function(y) {
    yg <- (g1 * cyld_agg[[x]]$gti * notprez) + (g2 * cyld_agg[[x]]$gti * prez)
    yg2 <- (g1 * cyld_agg[[x]][[y]] * notprez) + (g2 * cyld_agg[[x]][[y]]*prez)
    pg <- tareas[[x]] * carea_agg[[x]]$gti * yg
    pg2 <- tareas[[x]] * carea_agg[[x]][[y]] * yg2
    cellStats(pg - pg2, sum) / cellStats(pg, sum) * 100
  })
})

# Print out combined results to latex table
scen <- c(rep("300-400 mm", nrow(ygap_bias)), rep("400 mm", nrow(ygap_bias)),
          rep("500 mm", nrow(ygap_bias)))
ygaps <- cbind.data.frame(scen, "sensor" = rep(snms, 3),
                          rbind(ygap_bias, ygap_bias2, ygap_bias3))
ygap_xtab <- xtable::xtable(ygaps, digits = 1)
print(ygap_xtab, type = "latex",
      file = "paper/figures/ygap-spat-bias.tex")

save(list = ls(), file = full_path(p_data, "yield-bias.rda"))
```

The actual impact of the spatial bias is fairly low, being largely cancelled out by biases in other direction in other areas. This illustrates the advantage of a statistically constrained approach to crop area estimates.

9 Agent-based model sensitivity to map error

Selection of magisterial districts and associated landcover sets for use in ABM example, and analysis of bias after initial allocation of agents performed based on available farmland estimated from 100 m downscaled versions of the landcover data.

9.1 Data

```r
library(SAcropland)
library(raster)
library(rgdal)
library(lmisc)
library(readxl)
library(rgeos)

# Paths
p_root <- proj_root("SAcropland")
p_fig <- full_path(p_root, "SAcropland/paper/figures/")
p_data <- full_path(p_root, "SAcropland/external/ext_data/")
p_data2 <- full_path(p_root, "SAcropland/data/")

# landcover data
load(full_path(p_data, "d_grid_act.rda"))  # actual diffence grids
paths <- full_path(p_data, dir(p_data, paste0("cover*.*sum_mask.tif$")))
gti <- raster(paths[2])  # gti 2011
namask <- raster(full_path(p_data, "namask.tif"))  # NA mask
gti <- mask(gti, namask)  # apply mask to GTI
# cellStats(!is.na(gti), sum)

# Reconstruct original landcover estimates
snms <- c("sa30", "globmu", "modmu", "geow")
dlist_1km <- lapply(dlist_act[snms], function(x) x$f1$g2011)
# dlist_1km <- stack(lapply(dlist_act[snms], function(x) x$f1$g2011))
lc_list <- lapply(dlist_1km, function(x) gti - x)
# lc_list <- gti - dlist_1km
# cellStats(!is.na(lc_list$sa30), sum)
lc_list <- stack(lc_list)

# magisterial district data cropland statistics (2011, but only to look at for
# selection
load("external/ext_data/gti_md.rda")
mu <- data.frame(t(sapply(gti_md, function(x) {
  c(length(x), length(which(is.na(x))), round(mean(x, na.rm = TRUE), 1))
})))
colnames(mu) <- c("N", "na", "f")

# district shapes, filtered by N missing data, minimum fraction, and area
md <- readOGR("external/ext_data/mag_dists.sqlite", layer = "mag_dists")
mdfilt <- mu[mu$na < 20 & mu$f > 5 & mu$N < 1500, ]
mdsel <- as.numeric(row.names(mdfilt))

# Examine, including in QGIS
# plot(md)
# plot(md[mdsel[mdsel < 314], ], col = "red", add = TRUE)  # filter out Cape ones
writeOGR(md[mdsel[mdsel < 314], ], dsn = "external/ext_data/md_abms.sqlite",
         layer = "md_abms", driver = "SQLite")
```

9.1.1 Magisterial District selection

Looking at the data in QGIS, including some rainfall maps from earlier papers and the inter-tubes, it looks like the following districts are a good match:

- Clocolan (269)
- Ficksburg (274)
- Fouriesburg (275)
- Marquard (288)

```
abm_md <- md[mdsel[mdsel %in% c(269, 274:275, 288)], ]
writeOGR(abm_md, dsn = "external/ext_data/md4.sqlite", layer = "md4",
         driver = "SQLite")

# Deprecated: earlier used when running 2007 rather than 2011
# Let's compare the fractions between the GTI 2007 and GTI 2011 datasets for
# interest's sake
# gti_abm <- extract(gti, abm_md)
# abm_mu <- data.frame(t(sapply(gti_abm, function(x) {
#   c(length(x), length(which(is.na(x))), round(mean(x, na.rm = TRUE), 1))
# })))
# colnames(abm_mu) <- c("N", "na", "f")
# abm_mu$f - mu[mdsel[mdsel %in% c(269, 274:275, 288)], "f"]
 # -1.0  0.0 -0.4 -0.2  # tinf differences!
rgeos::gArea(abm_md, byid = TRUE) / 10000 / 100
```

9.1.2 Extract cropland percentage from selected districts

```
gti4 <- crop(gti, extent(abm_md))
lc4 <- crop(lc_list, extent(abm_md))
cb <- brick(stack(gti4, lc4), filename = "external/ext_data/cropland-md4.tif",
            overwrite = TRUE)
cpct <- raster::extract(gti4, abm_md)
dcrop_pct <- round(sapply(cpct, sum, na.rm = TRUE) / sapply(cpct, length), 1)
```

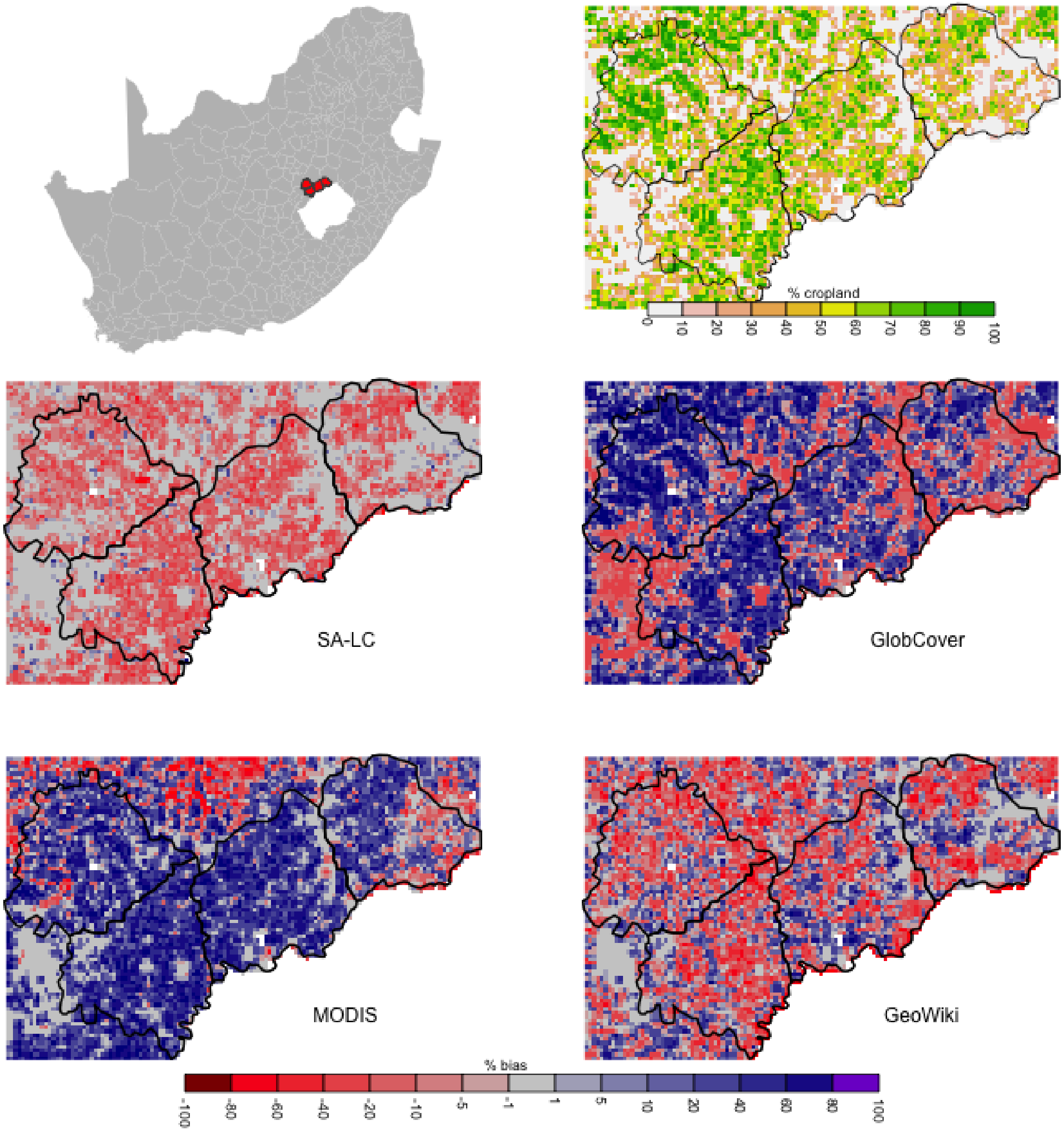
Cropland percentages:

- Clocolan (269) - 45%
- Ficksburg (274) - 39%
- Fouriesburg (275) - 29%
- Marquard (288) - 44%

9.1.3 Map selected districts, cropland percentage, and error

```
brks1 <- seq(0, 100, 10)
cols1 <- rev(terrain.colors(length(brks1) - 1))
brks2 <- c(-100, -80, -60, -40, -20, -10, -5, -1)
brks2 <- c(brks2, rev(abs(brks2)))
```

```r
cols2 <- colorRampPalette(c("red", "grey80", "blue4"))(length(brks2) - 1)
cols2[c(1, length(cols2))] <- c("darkred", "purple3")
cx <- 1.1
lcol <- "black"
mcap <- c("SA-LC", "GlobCover", "MODIS", "GeoWiki")

# layout.show()
pdf(full_path(p_fig, "abm-selected-districts.pdf"), height = 7, width = 7)
# png(full_path(p_fig, "abm-selected-districts.png"), height = 700, width = 700)
par(mfrow = c(3, 2), mar = c(0, 0, 0, 0), oma = c(2, 0, 0, 0))
plot(md, col = "grey", border = "transparent")
plot(abm_md, col = "red", border = "grey30", add = TRUE)
par(mar = c(3, 0, 0, 0))
plot(abm_md)
plot(cb[[1]], axes = FALSE, box = FALSE, breaks = brks1, col = cols1,
     add = TRUE, legend = FALSE)
plot(abm_md, add = TRUE)
polygonsLabel(abm_md, labels = paste0(abm_md$name_2, " (", 1:4, ")"),
              method = "buffer", cex = 0.9)
for(i in 2:5) {
  plot(abm_md)
  plot(cb[[1]] - cb[[i]], axes = FALSE, box = FALSE, breaks = brks2,
       col = cols2, add = TRUE, legend = FALSE)
  plot(abm_md, add = TRUE, lwd = 2)

  mtext(mcap[i - 1], side = 1, line = -4, adj = 0.7)
}
nct <- length(brks1) - 1
flex_legend(ncuts = nct, legend.text = "% cropland", legend.vals = rep("", nct),
            longdims = c(0.6, 0.9), shortdims = c(0.72, 0.012), colvec = cols1,
            srt = c(270, 0), horiz = TRUE, textside = "bottom",
            legend.pos = c(4, 3), leg.adj = list(c(4.2, 0), c(-0.5, -0.5)),
            cex.val = 1.1, textcol = lcol, bordercol = lcol)
tc <- rect_coords(0.6, 0.9, nct, constEWorNS = 0.72, resEWorNS = 0.02)
xs <- c(tc[[1]][1, 1], sapply(tc, function(x) x[2, 1]))
ys <- sapply(tc, function(x) x[1, 2])
text(xs, ys, labels = brks1, srt = 270, adj = c(-0.15, 0.5), cex = 1.1)
nct <- length(brks2) - 1
flex_legend(ncuts = nct, legend.text = "% bias", legend.vals = rep("", nct),
            longdims = c(0.2, 0.8), shortdims = c(0.05, 0.015), colvec = cols2,
            srt = c(270, 0), horiz = TRUE, textside = "bottom",
            legend.pos = c(4, 6), leg.adj = list(c(4.2, 0), c(-0.5, -0.5)),
            cex.val = 1.1, textcol = lcol, bordercol = lcol)
tc <- rect_coords(0.2, 0.8, nct, constEWorNS = 0.047, resEWorNS = 0.012)
xs <- c(tc[[1]][1, 1], sapply(tc, function(x) x[2, 1]))
ys <- sapply(tc, function(x) x[1, 2])
text(xs, ys, labels = brks2, srt = 270, adj = c(-0.1, 0.5), cex = 1.1)
dev.off()
```



9.2 Select modeled yields for agent-based model

From earlier simulation, to give idea about variability

```
dssat <- fread(full_path(p_data, "abm/dssat-yields.csv"))
dssat_red <- copy(dssat[])

dts <- c(2007288, 2007303, 2007318, 2007333, 2007348, 2007363, 2008013, 2008028)
dssat[PDAT %in% dts, ]

plot(1:31, 1:31, ylim = range(dssat$HWAH), pch = "", xlab = "YEAR",
     ylab = "HWAH")
points(dssat[PDAT %in% dts & FNAM != "ZASP0001", HWAH])
```

```r
boxplot(dssat[PDAT %in% dts & FNAM == "ZASP0001", HWAH])
boxplot(dssat[PDAT %in% dts & FNAM == "ZASP0002", HWAH])
boxplot(dssat[PDAT %in% dts & FNAM == "ZASP0003", HWAH])
dssat[PDAT %in% dts, mean(HWAH), by = FNAM]
dssat[PDAT %in% dts, mean(HWAH), by = .(CU, FNAM)]

plot(1:8, 1:8, ylim = range(dssat$HWAH), pch = "", xlab = "YEAR",
     ylab = "HWAH")
for(i in unique(dssat$FNAM)) {
  dssat[PDAT %in% dts & CU == 1 & FNAM == i, lines(HWAH)]
}

# provincial ag census statistics from 2007
prodyld <- cbind.data.frame("md" = as.character(abm_md$name_2),
                            "pha" = c(6219, 6203, 6391, 9404),
                            "prod" = c(14781, 17434, 24366, 25578))
prodyld$yld <- prodyld$prod / prodyld$pha

outylds <- copy(dssat[PDAT %in% c(2007303, 2007348, 2008013) &
                   FNAM != "ZASP0001", ])
outylds[PDAT %in% dts, mean(HWAH), by = .(FNAM)]
outylds[PDAT %in% dts, mean(HWAH), by = CU]
outylds[PDAT %in% dts, mean(HWAH), by = .(CU, FNAM)]
write.csv(outylds, file = full_path(p_data, "abm/abm-yields.csv"))
```

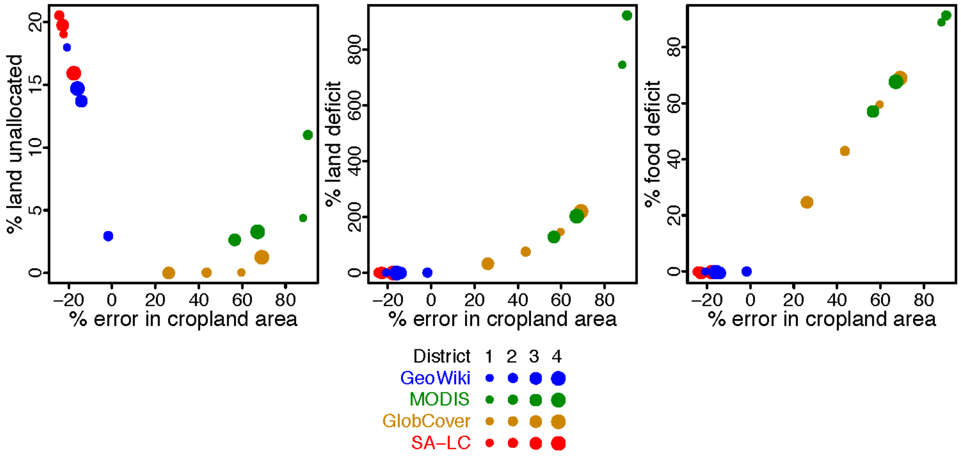
9.3 Downscaling and agent allocation

Performed separately and described in the methods and supplemental.

9.4 Analysis of agent allocation errors

```r
abias <- data.table(read_excel(full_path(p_data, "abm/stats.xlsx")))
# abias <- abias[Alg == 1]
setkey(abias, MD)
hhs <- data.table(read_excel(full_path(p_data, "abm/agent-bias2.xlsx"),
                             sheet = 2))
setkey(hhs, MD)
abias <- abias[hhs]
abias

# calculate district level bias as percent
pctbias <- abias[var == "ag_area",
                 lapply(list(sa30, globmu, modmu, geow), function(x) {
                   (gti - x) / gti * 100
                 })]

# HHs without land as percent total households
noland <- abias[var == "HHs_no_land",
                lapply(list(sa30, globmu, modmu, geow), function(x) {
                  x / HHs * 100
                })]

# unallocated farmland
unalloc <- round(abias[var == "ag_area_unalloc", snms, with = FALSE] /
  abias[var == "ag_area", snms, with = FALSE] * 100, 2)

# land deficit farmland
land_def <- round(abias[var == "land_def", snms, with = FALSE] /
  abias[var == "ag_area", snms, with = FALSE] * 100, 2)

# households couldn't find enough land
not_enough <- round(abias[var == "HHs_cant_find", snms, with = FALSE] /
                      (abias[var == "HHs_no_land",
                             lapply(.SD, function(x) HH_gen - x),
                             .SDcol = c(snms, "HH_gen")][, snms, with = FALSE]),
                    2)

# production deficits
hhprod <- abias[var == "prod", gti / HHs]
mdprodlc <- abias[var == "prod",
                  lapply(.(sa30, globmu, modmu, geow), function(x) x / HHs)]
prdef <- abias[var == "prod", lapply(.(sa30, globmu, modmu, geow), function(x) {
  (gti - x) / gti * 100
})]  # district total deficit
hhprodlc <- abias[var == "prod",
                  lapply(.(sa30, globmu, modmu, geow), function(x) x / HHs)]
hhprod_def <- (hhprod - hhprodlc) / hhprod * 100  # household deficit

cexs <- seq(1, 2, 1 / 3)
# pdf(full_path(p_fig, "agent-bias.pdf"), height = 5, width = 3)
pdf(full_path(p_fig, "agent-bias.pdf"), height = 2.5, width = 5)
cx <- c(0.9, 1, 0.6)
# png(full_path(p_fig, "agent-bias.png"), height = 350, width = 700)
# par(mfrow = c(2, 1), mar = c(0.5, 2.5, 0.5, 1), mgp = c(1, 0.05, 0),
#     tcl = -0.2, oma = c(3.5, 0, 0, 0))
par(mfrow = c(1, 3), mar = c(4, 2, 0.5, 0), mgp = c(0.9, 0.1, 0),
    tcl = -0.2, oma = c(3.5, 0, 0, 0.1))
scols <- c("red", "orange3", "green4", "blue")
cols <- c(rep("red", 4), rep("orange3", 4), rep("green4", 4), rep("blue", 4))
plot(unlist(pctbias), unlist(unalloc), col = cols, cex = cexs, pch = 20,
     ylab = "% land unallocated", xlab = "% error in cropland area",
     cex.lab = cx[2], cex.axis = cx[1])
plot(unlist(pctbias), unlist(land_def), col = cols, cex = cexs, pch = 20,
     ylab = "% land deficit", xlab = "% error in cropland area",
     cex.lab = cx[2], cex.axis = cx[1])
plot(unlist(pctbias), unlist(hhprod_def), col = cols, cex = cexs, pch = 20,
     ylab = "% food deficit", xlab = "% error in cropland area",
     cex.lab = cx[2], cex.axis = cx[1])
par(xpd = NA)

yst <- seq(0.05, 0.23, 0.18 / 4)
xy <- cbind(grconvertX(rep(0.5, 5), from = "ndc"),
            grconvertY(yst, from = "ndc"))
```

```
for(i in 2:length(yst)) {
  if(i == 5) {
    leg <- c("District", 1:4)
  } else {
    leg <- rep("", 4)
  }
  legend(x = xy[i, 1], y = xy[i, 2], legend = rep("",4), #legend = leg,
         pch = 20, cex = cx[3],
         horiz = TRUE, adj = c(4, -5.75),
         col = scols[i - 1], pt.cex = cexs, bty = "n")#cex = cexs,)
}
yst <- seq(0.06, 0.24, 0.18 / 4)
xy <- cbind(grconvertX(rep(0.505, 5), from = "ndc"),
            grconvertY(yst, from = "ndc"))
text(x = xy[, 1], y = xy[, 2], labels = c(mcap, "District"),
     #srt = 90, #pos = 2,
     #adj = c(-1, 1),
     pos = 2, col = c(scols, "black"), cex = cx[1])
xst <- seq(0.53, 0.60, 0.07 / 3)
xy <- cbind(grconvertX(xst, from = "ndc"),
            grconvertY(rep(yst[length(yst)], 4), from = "ndc"))
text(x = xy[, 1], y = xy[, 2], labels = 1:4, pos = 2, cex = cx[1])
dev.off()
```



10 Miscellaneous additional calculations

```
library(rgdal)
library(SAcropland)
library(rgeos)

p_dat <- full_path(full_path(proj_root("SAcropland"), "SAcropland"),
                   "external/ext_data")
```

Area of sub-Saharan Africa and South Africa's share of that area

```
eco <- readOGR(dsn = full_path(p_dat, "africa_ecofloristic_zones.sqlite"),
               layer = "africa_ecofloristic_zones")
eco@proj4string <- CRS(projection(raster(nrow = 1, ncol = 1)))
af <- readOGR(dsn = full_path(p_dat, "africa_countries_alb.sqlite"),
               layer = "africa_countries_alb")
af@data <- af@data[, c(1, 3:4)]
ecoalb <- spTransform(eco, af@proj4string)

# cut down countries
nms <- af$cntry_name[af$region == "Northern Africa"]
nms <- nms[nms != "Sudan"]  # keep Sudan
ssa <- af[!af$cntry_name %in% nms, ]
ssa <- ssa[-grep("Island|Sao|Verde|Mauritius|Helena|Comoros|Portugal|Seychell",
                 as.character(ssa@data$cntry_name)), ]
ssa <- gBuffer(ssa, width = 0)
# plot(ssa)

desert <- ecoalb[grep("desert", ecoalb$gez_term), ]
# plot(ecoalb[names(which.max(gArea(desert, byid = TRUE) / 10000 / 100)), ])
sahara <- as.numeric(names(which.max(gArea(desert, byid = TRUE))))

nosahara <- ecoalb[-sahara, ]
plot(nosahara, col = "red")

ovs <- which(!is.na(over(nosahara, ssa)))
ssaeco <- nosahara[ovs, ]
ssaeco <- ssaeco[-1, ]  # drop a couple Saharan mountains
ssaeco <- ssaeco[-1, ]
# plot(ssaeco[ssaeco@data$gez_term == "Tropical mountain system", ][1, ])
plot(ssa)
plot(ssaeco, col = "red", add = TRUE)

# calculate area of remainder
afarea <- gArea(ssaeco) / 10000
# plot(af[af$cntry_name == "South Africa", ])
saarea <- gArea(af[af$cntry_name == "South Africa", ]) / 10000
saarea / afarea
```

```
a <- sample(1:100, size = 20)
b <- sample(1:80, size = 20)
mean(a - b)
sum(a) / sum(b)
```