Supporting Information for

An analysis of methodological and spatial differences in global cropping systems models and maps

Weston Anderson[1*], Liangzhi You[1,2*], Stanley Wood[3], Ulrike Wood-Sichra[1], Wenbin Wu[2]

[1] International Food Policy Research Institute
2033 K Street, NW
Washington, DC 20006, USA

[2] Key Laboratory of Agri-informatics,
Ministry of Agriculture / Institute of Agricultural Resources and Regional Planning
Chinese Academy of Agricultural Sciences,
Beijing 100081, China

[3] Bill&Melinda Gates Foundation
500 Fifth Avenue North
Seattle, WA 98109, USA

*For Correspondence: Weston.b.Anderson@gmail.com, L.You@cgiar.org, +1-202-862-8168(phone), +1-202-467-4439(Fax)

Introduction

This supporting information provides details on the mathematical formulation of the downscaling procedures for each model discussed in the main text, details of the Gaussian filter analysis, the intermediate output for each analysis (the pixel-wise comparisons and the original cropping systems maps for each product), and the full analysis for maize and rice. The data is intended to provide readers with the complete details of each model specification and each analysis for those interested in a specific model methodology or output.

Appendix S1: Model Methodology

S1.1 M3 distribution of crop harvested areas and yields:

In each grid cell that had agricultural inventory data the map of crop area was calculated as follows:

$$fcrop_i = fcropland_i \left(\frac{crop_k}{cropland_k}\right) \quad \text{(S1)}$$

Where $fcrop_i$ is the harvested area of a specific crop in pixel i, $fcropland_i$ is the fraction of pixel i designated as cropland, $crop_k$ is the harvested area of a specific crop in statistical reporting unit k and $cropland_k$ is the amount of cropland in statistical reporting unit k. Yield was distributed uniformly across each grid cell as equivalent to the yield reported in the statistical reporting unit as a whole.

S1.2 MIRCA distribution of harvested area:

MIRCA primarily reconciles the differences between Siebert et al., 2007 dataset of areas equipped for irrigation (AEI), Cropland extent of Ramankutty et al., 2008, and the harvested area (HA) maps of the M3 dataset (Monfreda et al., 2008) to provide a monthly cropping map for irrigated and rainfed crops. The priorities used to reconcile inconsistencies between the datasets is outlined in Table A1. MIRCA first produces a Condensed Crop Calendar of harvested area for each sub-crop *c* and SRU k ($HAccc_{c,k}$). A sub-crop is used to represent multi-cropping systems or different sub-groups of a crop that grow at different points in the year. For a complete description of how the Condensed Cropping Calendar is produced, see Portman et al., 2008.

Table S1: Priorities in distributing the Condensed Crop Calendars to monthly growing area grids

Priority	Dataset	Goal
1	Area equipped for irrigation (Siebert et al., 2007)	In each month and grid cell the sum of irrigated crop-specific areas is lower than or equal to the area equipped for irrigation
2	Cropland extent (Ramankutty et al., 2008)	In each grid cell and month the sum of crop-specific irrigated and rainfed areas is lower than or equal to the cropland extent
3	Harvested crop area (Monfreda et al., 2008)	In each grid cell and for each crop class the annual sum of the irrigated and rainfed harvested crop area is equal to the total harvested area of the specific crop

Irrigated Crops:

The MIRCA process is a production system specific cell-wise approach to disaggregating harvested area by month. Area equipped for irrigation and cropland extent both include fallow land in their definition, so these classes need not be used completely so long as the annual harvested area is disaggregated and designated as either rainfed or irrigated. Area equipped for irrigation is prioritized over cropland extent and harvested area in the following process:

1. Calculate the irrigated harvested area (IHA) for subcrop *c* in cell *i* of month *m*

$$IHA_{c,i,m} = \frac{(HA_{c,i,m} \times fAEI_{i,m})}{\sum subcrop\ types_{i,m}} \quad \text{(S2)}$$

Where $fAEI_{i,m}$ is the fraction of pixel *i* equipped for irrigation in month *m*

2. Assign irrigated harvested areas to the monthly minimum of area equipped for irrigation and harvested area for subcrop *c* in cell *i* of month *m*

$$if\ AEI > 0\ and\ CroplandExtent > 0\ and\ HA_{c,i,m,} > 0,\ \ IHA_{c,i,m} = \min\left(AEI_{c,i,m}, HA_{c,i,m}\right) \quad \text{(S3)}$$

3. If there still exists HAccc_{ck} distribute irrigated growing areas to those cells that have cropland extent greater than zero and that are equipped for irrigation even if no $HA_{c,i,m}$ exists

$$if\ AEI > 0\ and\ CroplandExtent > 0\ ,\ \ IHA_{c,i,m} = remaining\ \text{HAccc}_{ck} \quad \text{(S4)}$$

4. If there still exists $\text{HAccc}_{c,k}$, distribute it to areas within cells that are equipped for irrigation, even if the cropland extent (and therefore $HA_{c,i,m,}$) is zero.

$$if\ AEI > 0\ and\ CroplandExtent = 0,\ \ IHA_{c,i,m} = remaining\ \text{HAccc}_{ck} \quad \text{(S5)}$$

Rainfed Crops:

Following the distribution of irrigated crop areas, rainfed crops were distributed. Rainfed annual crops were treated differently than rainfed permanent crops. Annual crops were allowed to grow on areas equipped for irrigation so long as they were available, while permanent crops were not.

5. Calculate the rainfed harvested area for crop *c* in pixel *i* of month *m* ($RHA_{c,i,m}$) by distributing rainfed crops to areas in which available cropland extent exceeds available area equipped for irrigation:

$$if\ CroplandExtent > AEI,\ \ RHA_{c,i,m} = \min(HA_{c,i,m}, CroplandExtent_i) \quad \text{(S6)}$$

6. If there still exists $HAccc_{c,k}$, expand suitable areas in cells with more cropland extent than area equipped for irrigation to 95% of the cell, leaving room to account for infrastructure.

$$if\ CroplandExtent > AEI,\ \ RHA_{c,i,m,MIRCA} = \min(HA_{c,i,m}, (0.95 * Area_i)) \quad \text{(S7)}$$

7. If there still exists $HAccc_{c,k}$, expand suitable areas in cells with either cropland extent or area equipped for irrigation to 95% of the cell, leaving room to account for infrastructure.

$$if\ CroplandExtent > 0\ \ or\ \ AEI > 0,\ \ RHA_{c,i,m,MIRCA} = \min(HA_{c,i,m}, (0.95 * Area_i)) \quad \text{(S8)}$$

The total harvested area of all rainfed and irrigated crops is therefore the sum of the IHA calculated in steps A2-A5 and the RHA calculated in steps A6-A8.

S1.3 SPAM harvested area and yield distribution

The SPAM model distributes available crop statistics using a cross-entropy approach that incorporates ancillary data on crop price, market access, biophysical suitability and expert elicitation. Shannon (1948) first introduced the concept of information entropy to measure the uncertainty of expected information in a system. Jaynes (1957) adopted the concept of information entropy and further proposed the principle of maximum entropy in statistical inference: the least informative probability distribution can be found by maximizing the entropy. In other words, without information to the contrary, all possible states of a system are equally likely. With respect to the concept's application to SPAM, Golan, Judge and Miller (1996) presents the formation of maximum entropy (or minimum cross-entropy) principle for use in parameter estimation problems.

Harvested Area calculation

SPAM distributes statistical information from allocation unit (e.g a country or a province) by using the cropping intensity of crop *j* in production system *l* to convert the reported harvested areas ($HA_{j,l}$) to physical areas ($CropArea_{jl}$) as:

$$CropArea_{jl} = \frac{HA_{jl}}{CroppingIntensity_{jl}} \quad \forall j, l \quad \text{(S9)}$$

SPAM next defines the area allocated to pixel *i* for crop *j* in production system *l* (A_{ijl}) using the share of the total physical area for crop *j* in production system *l* ($Share_{jl}$) and the physical area ($CropArea_{jl}$) as:

$$A_{ijl} = CropArea_{jl} \times Share_{jl} \times s_{ijl} \quad \forall i, j, l \tag{S10}$$

The minimum cross-entropy approach employed by the SPAM model calculates the area shares for crop *j* of pixel *i* in production system *l* as:

$$\min_{\{s_{ijl}\}} \left[\; CE\left(s_{ijl}, \pi_{ijl}\right) = \sum_i \sum_j \sum_l s_{ijl} \ln\left(s_{ijl}\right) - \sum_i \sum_j \sum_l s_{ijl} \ln\left(\pi_{ijl}\right) \right] \tag{S11}$$

Subject to the following constraints:

$$\sum_{i \in k} s_{ijl} = 1 \quad \forall j, i, k \tag{S12}$$

$$\sum_j \sum_l CropArea_{jl} \times Share_{jl} \times s_{ijl} \le CroplandExtent_i \quad \forall i \tag{S13}$$

$$CropArea_{jl} \times Share_{jl} \times s_{ijl} \le CropSuitableArea_{ijl} \quad \forall i, j, l \tag{S14}$$

$$\sum_{i \in k} \sum_l CropArea_{jl} \times Share_{jl} \times s_{ijl} = SubCropArea_{jk} \quad \forall k, j \in J \tag{S15}$$

$$\sum_{l \in L} CropArea_{jl} \times Share_{jl} \times s_{ijl} \le AEI_i \quad \forall i \tag{S16}$$

$$1 \ge s_{ijl} \ge 0 \quad \forall i, j, l \tag{S17}$$

Where *l* may be irrigated, rainfed high-input, rainfed subsistence or rainfed low input. $CroplandExtent_i$ is the total extent of cropland for pixel *i* and $CropSuitableArea_{ijl}$ is the area suitable for crop *j* at input level *l* in pixel *i*. $SubCropArea_{jk}$ is the crop area statistics for crop *j* in subnational SRU *k*. AEI_i is the area equipped for irrigation in pixel *i*. *J* is a set of commodities for which sub-national production statistics exist and *L* is a set of commodities within pixel *i* that are irrigated. π_{ijl} represents the prior estimate of area shares for crop *j* at input level *l* in pixel *i*.

The prior is developed using expert elicitation where available and elsewhere is calculated based on potential unit revenue, Rev_{ijl}.

$$Rev_{ijl} = Share_{jl} \times Price_j \times Access_{ij} \times SuitableYield_{ijl} \tag{S18}$$

Where $Price_j$ is the price of crop j $Access_{ij}$ is a measure of the physical accessibility of the market for crop j from pixel i. $SuitableYield_{ijl}$ is the agro-climatically suitable yield for crop j at input level l in pixel i. Then the prior allocation of crop area is estimated using irrigated area and cropland as follows.

$$PriorArea_{ijl} = AEI_i \times \frac{Rev_{ijl}}{\sum_j Rev_{ijl}} \quad \forall j, i, \ \forall l = irrigated \tag{S19}$$

$$PriorArea_{ijl} = \left(CroplandExtent_i - AEI_i - PriorArea_{ij,subsistence}\right) \times \frac{Rev_{ijl}}{\sum_l \sum_j Rev_{ijl}} \quad \forall j, i, \ \forall l = rainfed \tag{S20}$$

In the case of subsistence farming, the revenue measure is replaced by a measure of population density. The subsistence part of the sub-national crop area is then pre-allocated using rural population density as a weight.

$$PriorArea_{ij,subsistence} = SubCropArea_{jk} \times Percent_{jl} \times \frac{Pop_i}{\sum_{i \in k} Pop_i} \quad \forall j, i, l\text{=subsistence} \tag{S21}$$

After this pre-allocation, the prior is calculated by normalizing the allocated areas over the whole allocation unit:

$$\pi_{ijl} = \frac{PriorArea_{ijl}}{\sum_i PriorArea_{ijl}} \quad \forall i, j, l \tag{S22}$$

Yield Calculation

The calculation of yield is based on the statistical yield information for crop j within production system l for each SRU k. First, the average potential yield, $\bar{Y}_{jl}$, is calculated as:

$$\bar{Y}_{jlk} = \frac{\sum_i Suitability_{ijl} \times A_{ijl}}{\sum_i A_{ijl}} \quad \forall k \tag{S23}$$

$$Y_{ijl} = \frac{Suitability_{ijl} \times CropYield_{jlk}}{\bar{Y}_{jlk}} \tag{S24}$$

Where $CropYield_{jlk}$ is the statistical yield reported for crop *j* in production system *l* within SRU *k*. Then the production of crop *j* in production system *l,* and pixel *i,* $Prod_{ijl}$, could be calculated as the following:

$$Prod_{ijl} = \left(A_{ijl} \times CropingIntensity_{jl}\right) \times Y_{ijl} \tag{S25}$$

S1.4 GAEZ distribution of harvested area and yield:

The GAEZ model distributes available statistical data using an iterative rebalancing procedure that converges to the same answer as the cross entropy approach used by SPAM (Fischer et al., 2006). The GAEZ formulation includes distance to market, population density, ruminant livestock density, farming system zone and producer price by crop as a means of further disaggregating available national or sub-national statistics.

The iterative rebalancing algorithm used by the model is documented in detail in Fischer et al. (2006), but generally works by using multipliers (separated into rainfed *R* and irrigated *I*) for area (λ_j^R and λ_j^I) by crop *j*, for cropping intensity (ρ^R and ρ^I), and yield (μ_j^R and μ_j^I) by crop *j*. The algorithm updates the multipliers iteratively such that all constraints are met, and in the process produces grid-cell specific allocations of harvested area and production for rain-fed and irrigated land. ρ^R and ρ^I provide a measure of the discrepancy between the potential for multi-cropping and actual cropping intensity, while μ_j^R and μ_j^I represent the gap between actual and potential crop yields.

The GAEZ model uses a two-step nested process within each iteration of the rebalancing algorithm, by which land is broadly allocated into two sets of crops: Set I_1, for which the spatial distribution layer (ε_{ij}) exists for pixels i and crops *j,* and Set I_2, for which the spatial distribution layer does not exist. ε_{ij} is defined as a subset of the M3 dataset for selected crops in countries for which more than 50% of the data was derived from sub-national statistics. Shares of land are distributed to each set of irrigated ($SetShare_1^I$ and $SetShare_2^I$) and rainfed ($SetShare_1^R$ and $SetShare_2^R$) crops as follows:

$$SetShare_1^I = \frac{\sum_{j \in I_1^I} CroppingIntensity_i^I \times Yield_{ij}^I \times Price_j \times \lambda_j^I \times RelativeYield_{ij}^I}{\sum_{j \in I_1^I \cup I_2^I} CroppingIntensity_i^I \times Yield_{ij}^I \times Price_j \times \lambda_j^I \times RelativeYield_{ij}^I} \quad \text{(S26)}$$

$$SetShare_2^I = 1 - SetShare_1^I \quad \text{(S27)}$$

$$I_1^I = \{(j \in I_1) \wedge (\varepsilon_{ij} > 0) \wedge (RelativeYield_{ij}^I \geq \gamma_j^I)\} \quad \text{(S28)}$$

$$I_2^I = \{(j \in I_1) \wedge (RelativeYield_{ij}^I \geq \gamma_j^I)\} \quad \text{(S29)}$$

Where $RelativeYield_{ij}^I$ is defined in equation (S60), and γ_j^I is the crop allocation relative yield threshold for irrigated crop *j*. Similarly for rainfed crops:

$$SetShare_1^R = \frac{\sum_{j \in I_1^R} CroppingIntensity_i^R \times Yield_{ij}^R \times Price_j \times \lambda_j^R \times RelativeYield_{ij}^R}{\sum_{j \in I_1^R \cup I_2^R} CroppingIntensity_i^R \times Yield_{ij}^R \times Price_j \times \lambda_j^R \times RelativeYield_{ij}^R} \quad \text{(S30)}$$

$$SetShare_2^R = 1 - SetShare_1^R \quad \text{(S31)}$$

$$I_1^R = \{(j \in I_1) \wedge (\varepsilon_{ij} > 0) \wedge (RelativeYield_{ij}^R \geq \gamma_j^R)\} \quad \text{(S32)}$$

$$I_2^R = \{(j \in I_1) \wedge (RelativeYield_{ij}^R \geq \gamma_j^R)\} \quad \text{(S33)}$$

Where $RelativeYield_{ij}^R$ is defined in equation (A61), γ_j^R is the crop allocation relative yield threshold for rainfed crop *j*.

In the second step of the process, rainfed crop-specific area shares ($Share_{ij}^R$) and irrigated crop-specific physical area shares ($Share_{ij}^I$) are calculated for each grid cell *i* and crop *j* in the set of grid cells within a specific SRU *k* as follows:

$$Share_{ij}^I = SetShare_1^I \times \frac{(\varepsilon_{ij}^I \times \lambda_j^I)}{\sum_{k \in I_1^I} (\varepsilon_{ik}^I \times \lambda_k^I)}, \; j \in I_1^I \quad \text{(S34)}$$

$$Share_{ij}^I = 0, \; j \in I_1 \; \wedge \; j \notin I_1^I \quad \text{(S35)}$$

$$Share_{ij}^I = SetShare_2^I \times \frac{CroppingIntensity_{ij}^I \times Yield_{ij}^I \times Price_j \times \lambda_j^I \times RelativeYield_{ij}^I}{\sum_{k \in I_2^I} CroppingIntensity_{ik}^I \times Yield_{ik}^I \times Price_k \times \lambda_k^I \times RelativeYield_{ik}^I}, j \in I_2^I \quad \text{(S36)}$$

$$Share_{ij}^{I} = 0\,,\ j \in I_2\ \wedge\ j\ \notin I_2^{I} \tag{S37}$$

And similarly for rainfed crop areas:

$$Share_{ij}^{R} = SetArea_1^{R} \times \frac{(\varepsilon_{ij}^{R} \times \lambda_j^{R})}{\sum_{k \in I_1^{R}} (\varepsilon_{ik}^{R} \times \lambda_k^{R})}\ ,\ j \in I_1^{R} \tag{S38}$$

$$Share_{ij}^{R} = 0\,,\ j \in I_1\ \wedge\ j\ \notin I_1^{R} \tag{S39}$$

$$Share_{ij}^{R} = SetShare_2^{R} \times \frac{CroppingIntensity_{ij}^{R} \times Yield_{ij}^{R} \times Price_j \times \lambda_j^{R} \times RelativeYield_{ij}^{R}}{\sum_{k \in I_2^{R}} CroppingIntensity_{ik}^{R} \times Yield_{ik}^{R} \times Price_k \times \lambda_k^{R} \times RelativeYield_{ik}^{R}}\,, j \in I_2^{R} \tag{S40}$$

$$Share_{ij}^{R} = 0\,,\ j \in I_2\ \wedge\ j\ \notin I_2^{R} \tag{S41}$$

The crop-specific area shares are then used to calculate irrigated harvested areas (HA_{ij}^{I}) and rainfed harvested areas (HA_{ij}^{R}) for each pixel *i* and crop *j* in the set of grid cells within a specific SRU *k* as follows:

$$HA_{ij}^{I} = fCroplandExtent_i^{I} \times GridCellArea_i \times \left(\frac{\rho^{I} \sum_{k \epsilon crops} Share_{ij}^{I} \times CroppingIntensity_{ij}^{I}}{\sum_{k \epsilon crops} Share_{ij}^{I}}\right) \tag{S42}$$

$$HA_{ij}^{R} = fCroplandExtent_i^{R} \times GridCellArea_i \times \left(\frac{\rho^{R} \sum_{k \epsilon crops} Share_{ij}^{R} \times CroppingIntensity_{ij}^{R}}{\sum_{k \epsilon crops} Share_{ij}^{R}}\right) \tag{S43}$$

The constraints in the model are as follows:

Grid cell proportion of rainfed and irrigated land:

$$fCroplandExtent_i^{R} = fCroplandExtent_i^{T} - fCroplandExtent_i^{I},\ \forall i \tag{S44}$$

Where $fCroplandExtent_i^{T}$ is the proportion of total cropland in grid cell *i* (IIASA dataset, developed as part of GAEZ) and $fCroplandExtent_i^{I}$ is the proportion of each cropland grid cell that is equipped for irrigation (Seibert et al., 2007) dataset.

Cropland extent by grid cell:

$$CroplandExtent_i = fCroplandExtent_i^{T} \times GridCellArea_i,\ \forall i \tag{S45}$$

Where $fCroplandExtent_i^T$ is the total share of cropland (irrigated and rainfed) in each pixel

Grid cell cropping intensity for annual crops:

$$CroppingIntensity_i^I = \rho^I \times CroppingFactor_i^I \tag{S46}$$

$$CroppingIntensity_i^R = \rho^R \times CroppingFactor_i^R \tag{S47}$$

Where $CroppingFactor_i^I$ and $CroppingFactor_i^R$ correspond to the cultivation intensity class factor of irrigated and rainfed annual crops, respectively.

Total irrigated harvested area by crops:

$$HA_j^I = \propto_j^I \times HA_j, \;\; \forall j \tag{S48}$$

$$HA_j^R = \left(1 - \propto_j^I\right) \times HA_j, \;\; \forall j \tag{S49}$$

Where $\propto_j^I$ is the proportion of harvested area that is irrigated for crop *j*.

Harvested area and cropland extent (irrigated and rainfed) by grid cell:

$$HA_i^I = CroppingIntensity_i^I \times fCroplandExtent_i^I \times GridCellArea_i, \;\; \forall i \tag{S50}$$

$$HA_i^R = CroppingIntensity_i^R \times fCroplandExtent_i^R \times GridCellArea_i, \;\; \forall i \tag{S51}$$

Harvested area by pixel and crop:

$$HA_{ij}^I = CroppingIntensity_i^I \times Share_{ij}^I \times fCroplandExtent_i^I \times GridCellArea_i, \;\; \forall i, j \in annual\ crops \tag{S52}$$

$$HA_{ij}^I = CroppingIntensity^P \times Share_{ij}^I \times fCroplandExtent_i^I \times GridCellArea_i, \forall i, j \in perennial\ crops \tag{S53}$$

$$HA_{ij}^R = CroppingIntensity_i^R \times Share_{ij}^R \times fCroplandExtent_i^R \times GridCellArea_i, \;\; \forall i, j \in annual\ crops \tag{S54}$$

$$HA_{ij}^R = CroppingIntensity^P \times Share_{ij}^R \times fCroplandExtent_i^R \times GridCellArea_i, \forall i, j \in perennial\ crops \tag{S55}$$

Where $CroppingIntensity^P$ is the cropping intensity of perennial crops.

$$HA_j^I = \sum_i HA_{ij}^I, \;\; \forall j \tag{S56}$$

$$HA_j^R = \sum_i HA_{ij}^R, \;\; \forall j \tag{S57}$$

Grid cell irrigated yield:

$$Yield_{ij}^{I} = \mu_j^{I} \times PotentialYield_{ij}^{I,high}, \ \forall i, j \tag{S58}$$

$$Yield_{ij}^{R} = \mu_j^{R} \times \left(\left(1 - \psi_{ij}^{R}\right) \times PotentialYield_{ij}^{R,low} + \left(\psi_{ij}^{R} \times PotentialYield_{ij}^{R,high} \right) \right), \ \forall i, j \tag{S59}$$

Where $PotentialYield_{ij}^{I,high}$, $PotentialYield_{ij}^{R,high}$ and $PotentialYield_{ij}^{R,low}$ are the potential yield of crop j on grid cell i in a high input irrigated system, high input rainfed system and low input rainfed system, respectively. Potential yield information is derived from the GAEZ v3.0 database (Fischer et al., 2013). ψ_{ij}^{R} is a spatial location factor used to reflect differences in management intensity and input use derived from remote sensing, household survey data, information on farm size or market orientation of a household.

Grid cell relative yield factor:

$$RelativeYield_{ij}^{I} = {PotentialYield_{ij}^{I,high}} \Big/ {\max_{k \in gridcell} \left(PotentialYield_{kj}^{I,high} \right)} \tag{S60}$$

$$RelativeYield_{ij}^{R} = {PotentialYield_{ij}^{R,high}} \Big/ {\max_{k \in gridcell} \left(PotentialYield_{kj}^{R,high} \right)} \tag{S61}$$

Further information on the formulation used in the model may be found in the GAEZ v.3.0 documentation main text with details in Appendix A8 (Fischer et al., 2013). Details on the iterative rebalancing algorithm may be found in Fischer et al. (2006).

Appendix S2: Comparison of the Downscaling Methodologies

Cropland extent delineation

As a first step towards delineating crop specific harvested area and yield, each cropping system model defined a spatially explicit layer of cropland extent, representing the proportion of cropland in each 5-minute pixel globally. Because each subsequent step in the modeling process relies on the definition of cropland extent, the degree to which each pair of cropland extent products agree represents an upper bound of inter-model agreement on the spatial distribution of crop physical areas.

M3, MIRCA and SPAM all rely on the same base dataset for cropland extent: Ramankutty et al., (2008), which is an extension of Leff et al., (2004). Leff et al., (2004) synthesize satellite-derived land cover data and agricultural census data worldwide to assess the distribution of major crops across a global 5 arc minute grid in terms of the proportion of the total harvested area of each of the crops in each administrative unit. Following and improving on this work, Ramankutty et al. (2008) developed a new global land cover data set for croplands and pasture circa 2000 (at the same 5 arc minute resolution of the original dataset) by combining Boston University's MODIS-derived land cover data (Friedl et al., 2002) and SPOT VEGETATION based GLC2000 (Bartholome and Belward, 2005). Ramankutty et al. (2008) apply a multiple linear regression model to relate the combined satellite derived datasets to the agricultural statistics using a least squares error framework. The optimization is applied separately to six different regions of the world.

The cropland extent developed by Ramankutty et al. (2008) is used directly by M3 and with modifications by MIRCA and SPAM. By combining Ramankutty et al. (2008) and the global map of irrigation areas (GMIA), MIRCA produced a global dataset of monthly growing areas of 26 irrigated crops on the same 5 arc minutes grid. SPAM similarly reconciles the GMIA map of irrigated areas and Ramunkutty cropland extent by setting the cropland extent to be at least equal to the irrigated area in a preprocessing step. For more information on each of these methodologies, see Appendix S1.

GAEZ uses GLC2000 data and GMIA, but also considers a global land cover categorization (IFPRI, 2002), which is based on a reinterpretation of the Global Land Cover Characteristics Database v.2.0 (EROS Data Center, 2000), a layer of forest land from the Forest Resources Assessment of FAO (FAO, 2001), the IUCN-WCMC protected areas inventory (WPDA, 2009) and an estimate of land required for housing and infrastructure for the year 2000 derived from FAO-SDRN, based on LANDSCAN 2003 that were calibrated to UN 2000 population figures (Fischer et al., 2008; Bhaduri et al., 2002; Dobson et al., 2000). GAEZ runs a cross-sectional regression on the land cover distributions to derive weights, which are then applied in an iterative adjustment procedure to match estimated reference values such that the geographic and statistical data are consistent.

Suitability Constraints

GAEZ and SPAM further constrain potential crop distribution using biophysical and socioeconomic suitability prior to allocating the harvested area of each crop. M3 and MIRCA do not consider suitability criteria. SPAM directly uses the suitable area product from GAEZ, meaning that despite using different cropland extent products to constrain the distribution of crops, the two models use identical constraints on biophysically suitable land. The GAEZ suitability product integrates an extensive set of edaphic and climatic factors into its biophysical suitability analysis to produce a potential yield estimate and a suitability index by production system and crop. Further information on the suitability index analysis developed as part of the GAEZ model may be found in Fischer et al. (2013).

In addition to biophysical suitability criteria, both SPAM and GAEZ consider the socioeconomic factors that often constrain or encourage crop production. As a means of differentiating between low, medium and high input or management conditions, GAEZ divides the land into land use types. Land use types are derived using information on road infrastructure, livestock density, population density and distance to market. Low input, for example, relies on available human/livestock labor while high input is market oriented, using improved varieties, fertilizer, pesticides and machinery. Similar to GAEZ, SPAM explicitly models different production systems, which include high-input irrigated, high-input rainfed, low –input rainfed and subsistence (always low-input rainfed). SPAM includes data on crop prices and market access to construct a realistic market scenario in which there are not only biophysical barriers to producing crops, but social economic forces as well (see Appendix S1 for an explicit mathematical formulation).

Distribution of harvested area and yield

Perhaps the largest methodological differences between M3, MIRCA, SPAM and GAEZ are in the approaches used to downscale statistical data reported at administrative unit level into gridcell specific values. M3 uses the most straightforward method, allocating each crop evenly across potential cropland in each statistical reporting unit as the proportion of harvested area occupied by the crop to total harvested area in that reporting unit. Crop yield in each grid cell is assigned as being the same as the yield reported for the statistical unit as a whole. This approach implicitly assumes both environmental conditions and management/production systems are uniform across the cropland extents of each statistical reporting unit, or that there is insufficient information to characterize the spatial variations of crop production within a statistical unit. As a result, the distinct tolerances of individual crops to those spatial patterns are not incorporated in the downscaling procedure. This approach does not, furthermore, acknowledge the very significant differences between the yield levels of irrigated and rainfed production systems, nor of commercial and smallholder producers within these sometimes large and highly diverse statistical reporting units.

MIRCA primarily focuses on reconciling the differences between information derived from sub-national crop production statistics, M3 crop distributions and the Siebert et al. (2005) irrigated areas database. MIRCA deals only with harvested area and essentially uses the relative share of rainfed and irrigated cropland within each grid cell to break out M3 total crop areas into gridcell-specific rainfed and irrigated areas. The MIRCA model derives cropping intensity, which is used to convert harvested to physical area. MIRCA also includes use of numerous checks and adjustments to reconcile differences between the cropland area of Ramankutty et al. (2008) and the irrigated area estimates of Siebert et al. (2005) within each gridcell, given that total cropland area should at all times be greater than or equal to the irrigated cropland area. Similar to M3, MIRCA does not consider any form of suitability in its downscaling procedure.

The downscaling approaches of GAEZ and SPAM are predicated on the importance of attempting to take explicit account of available evidence of the spatial variation of production conditions within the cropland extent and of the significantly different yield resulting from each of those systems. Both GAEZ and SPAM use an approach that produces a result mathematically equivalent to that of a cross-entropy formulation, but GAEZ uses an iterative rebalancing procedure to adjust weighting factors until all constraints in the model are met, while SPAM uses a cross-entropy formulation. SPAM uses explicit cropping intensity from statistics and expert opinion to calculate cropping intensity while GAEZ derives a cropping intensity factor through the rebalancing procedure (see Appendix S1). Although the two models incorporate similar information (see Table 2, main text), the manner in which the information is used to constrain the model differs (see Appendix S1 for details on the mathematical formulation of each model). Additionally, GAEZ differs from SPAM in that it uses a "location factor" to incorporate spatially explicit information including geo-referenced household survey data. The prior in SPAM is used to capture spatially explicit information as well but the model does not include household survey data, instead leveraging the field presence of the CGIAR network to incorporate an extensive dataset of expert elicitations.

Input Data and Model Interdependencies

The major determinants of the potential reliability of downscaling efforts are (a) the quality of the cropland extent dataset indicating the physical extent and area intensity of cropland (e.g., share of cropland area in each 5 arc minute grid cell), and (b) the resolution and reliability of the sub-national crop statistics. Each model builds on a common set of available data as well as previous work in cropping systems modeling. Table 2 (main text) illustrates both the broad linkages and increasing sets of input data and assumptions that each of the M3, MIRCA, GAEZ and SPAM datasets relies upon.

National and Sub-National Statistics

All four datasets draw on FAOSTAT national data to provide control totals for cropland area, the harvested area, and yields of specific crops, while also spending considerable efforts to collect sub-national crop statistics to allow as detailed as possible disaggregation of national totals within sub-national administrative boundaries. Since MIRCA relies on M3 to provide its input data on the spatial allocation of the total area and average yield, it relies initially on the same sources of subnational crop statistics. The GAEZ model uses data from FAOSTAT as a constraint at the national level and – similar to MIRCA - uses the M3 sub-national statistics for select crops in countries that have sub-national statistics covering more than 50% of the country. SPAM relies on a separate collection of sub-national statistical data sources, focusing on increased coverage in developing countries.

M3 reports a total of 22,106 statistical reporting units globally, of which 56 were national, 2,299 were first level sub-national disaggregation (e.g., US state level), and 19,751 were second level (e.g. US county level) reporting units. SPAM reports 24,507 statistical units of which 251 were national, 2,758 were first level, and 21,498 were second level. SPAM focused its data collection efforts particularly in developing countries. For example in Africa the M3 and SPAM data sets were developed using around 300 and 4,150 second-level statistical reporting units respectively.

Extent of Irrigation

Those models that distinguish between rainfed and irrigated cultivations (MIRCA, SPAM and GAEZ) all use the GMIA v4.0, released in 2007 (Siebert et al., 2005), to identify the location and area intensity of irrigated production. However, the MIRCA and SPAM teams compiled information in the national and sub-national shares of different production systems and cropping intensities independently. MIRCA and SPAM both draw on FAO's AQUASTAT and national databases for gaining greater insights into national and crop-specific irrigation extents and practices, but MIRCA relies on a richer collection of national data, including a more complete collection of national/sub-national crop calendars and cropping intensities (in part because the goal of MIRCA is to produce monthly and not annual crop distribution maps). In contrast to MIRCA and SPAM, GAEZ relies on its own data for information about cultivation intensity of irrigated crops.

Ancillary Data

SPAM and GAEZ incorporate datasets beyond those used by M3 and MIRCA as a means of differentiating between production levels within cropping systems. The SPAM approach requires additional sets of data because it attempts further disaggregation of its rainfed production statistics amongst commercial and subsistence categories, and bases its approach to distribution of individual crops within the cropland extent on agronomic, economic and demographic principles and assumptions. These include crop area and production shares amongst irrigated production and large-scale/commercial and smallholder rainfed production, the spatial differences in the biophysical suitability of individual crops for irrigated and rainfed (commercial and subsistence) production, and estimates of the spatial patterns of population density as well as crop prices. GAEZ similarly divides the land into land use types to reflect variable management and input conditions. Data used to differentiate among land use types reflect the specific requirements of each and include road infrastructure, livestock density, population density and distance to market. Table 2 reflects the overlapping and separate ancillary datasets used by SPAM and GAEZ. In addition to available ancillary datasets, SPAM leverages the international network and field presence of CGIAR to undergo a systemic validation process. The feedback from this validation is used to inform future model simulations. This process is unique to the SPAM model.

Appendix S3: Gaussian Filter Sensitivity Analysis

The 2-dimensional Gaussian filter for pixel i may be expressed as:

$$g(x, y)_i = \left(\frac{1}{2\pi\sigma^2}\right)\left(e^{-\frac{x^2+y^2}{2\sigma^2}}\right) \quad \text{(S1)}$$

Where x is the distance from the horizontal axis, y is the distance from the vertical axis and σ is the standard deviation of the Gaussian distribution, used to control the kernel density as illustrated below

Figure S1: Differences in Cropland extent with Gaussian filters having kernel densities of 0 (pixel-level comparison), 1 (4 pixel radius), 2 (8 pixel radius), 3 (12 pixel radius) and 4 (16 pixel radius).

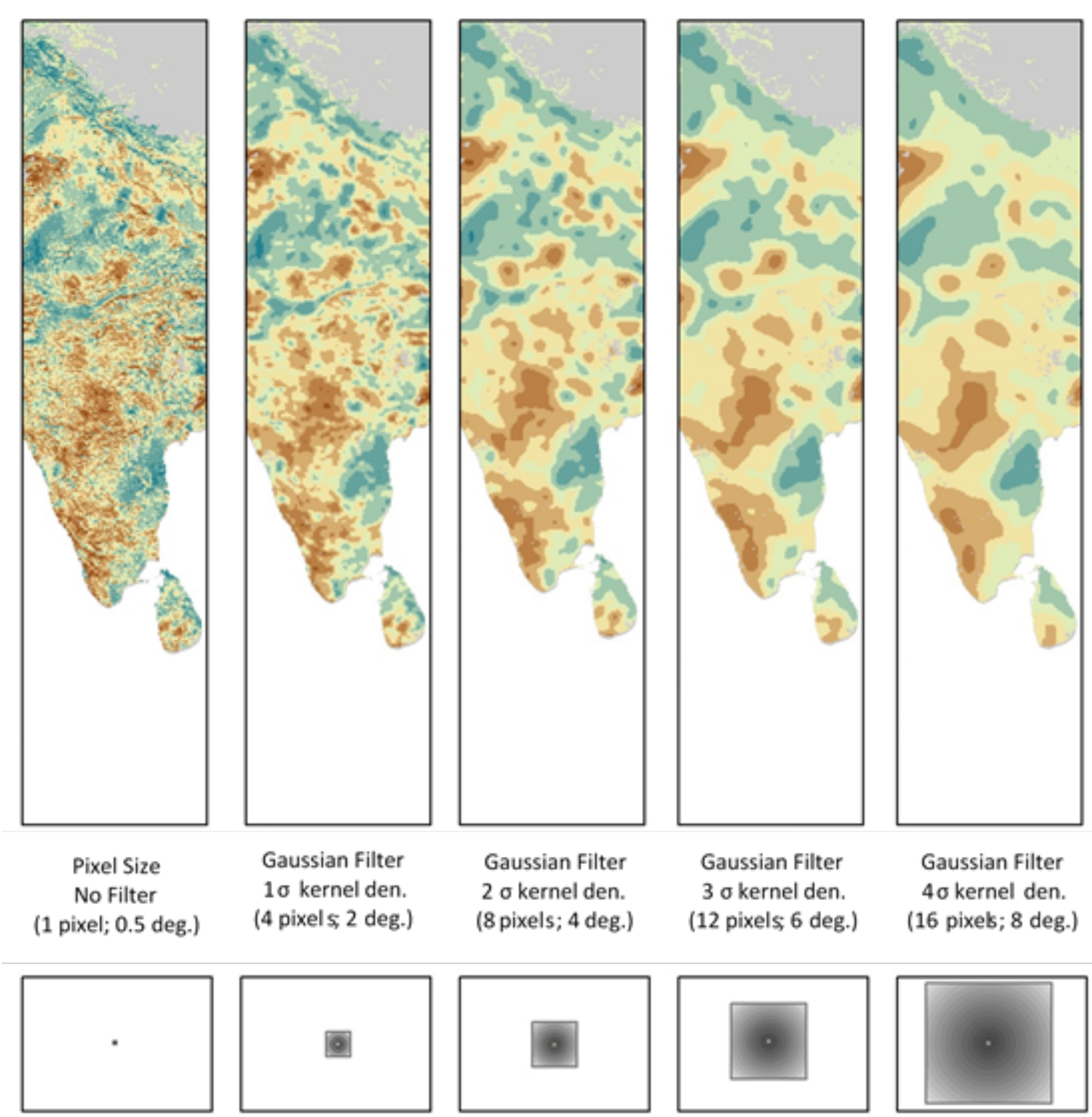
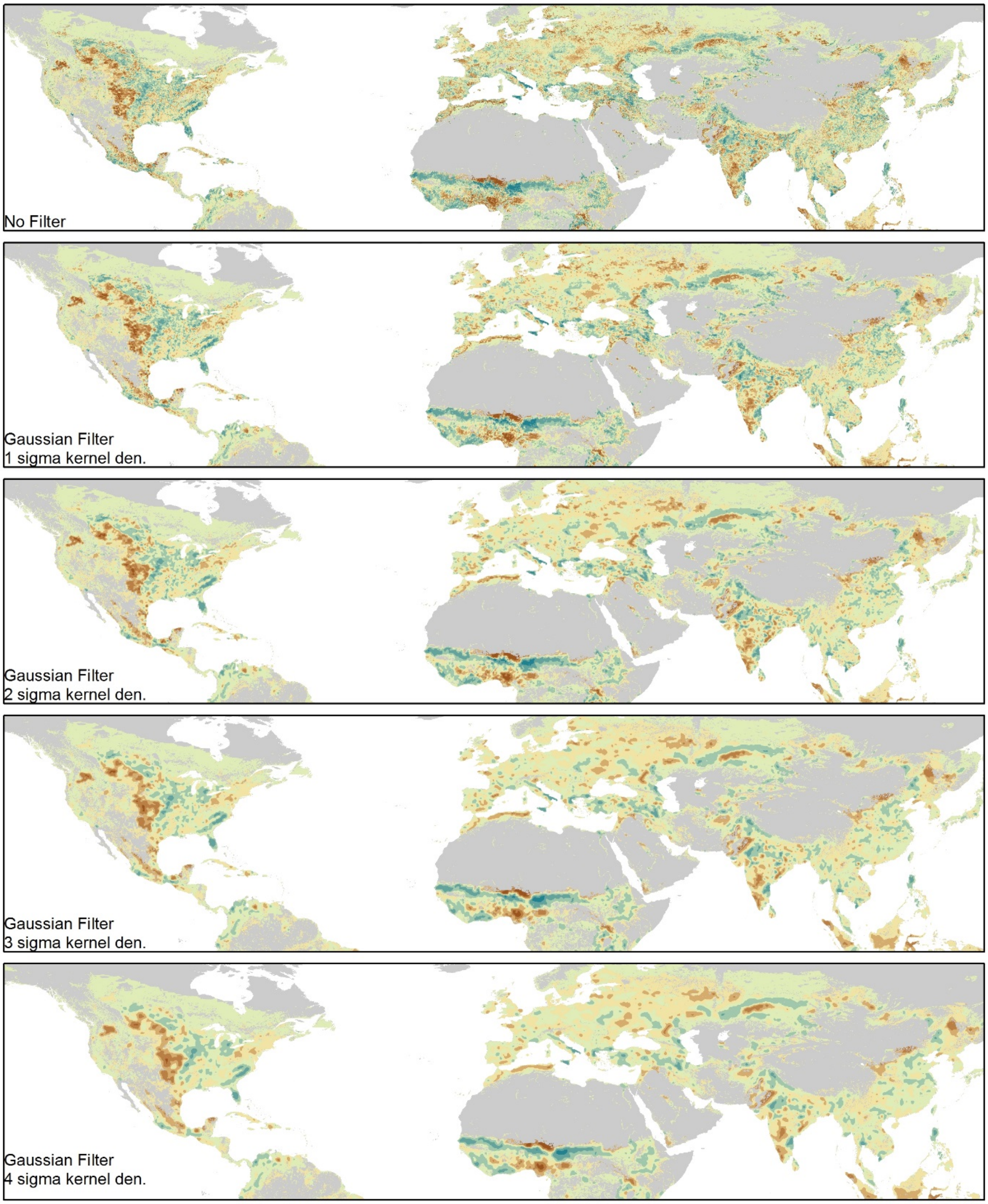


Figure S2: Differences in Cropland extent with Gaussian filters having kernel densities of 0 (pixel-level comparison), 1 (4 pixel radius), 2 (8 pixel radius), 3 (12 pixel radius) and 4 (16 pixel radius).



Appendix S4: Cropland extent supplementary material

Figure S1: Cropland extent of A) GAEZ and B) Ramankutty et al., (2008)

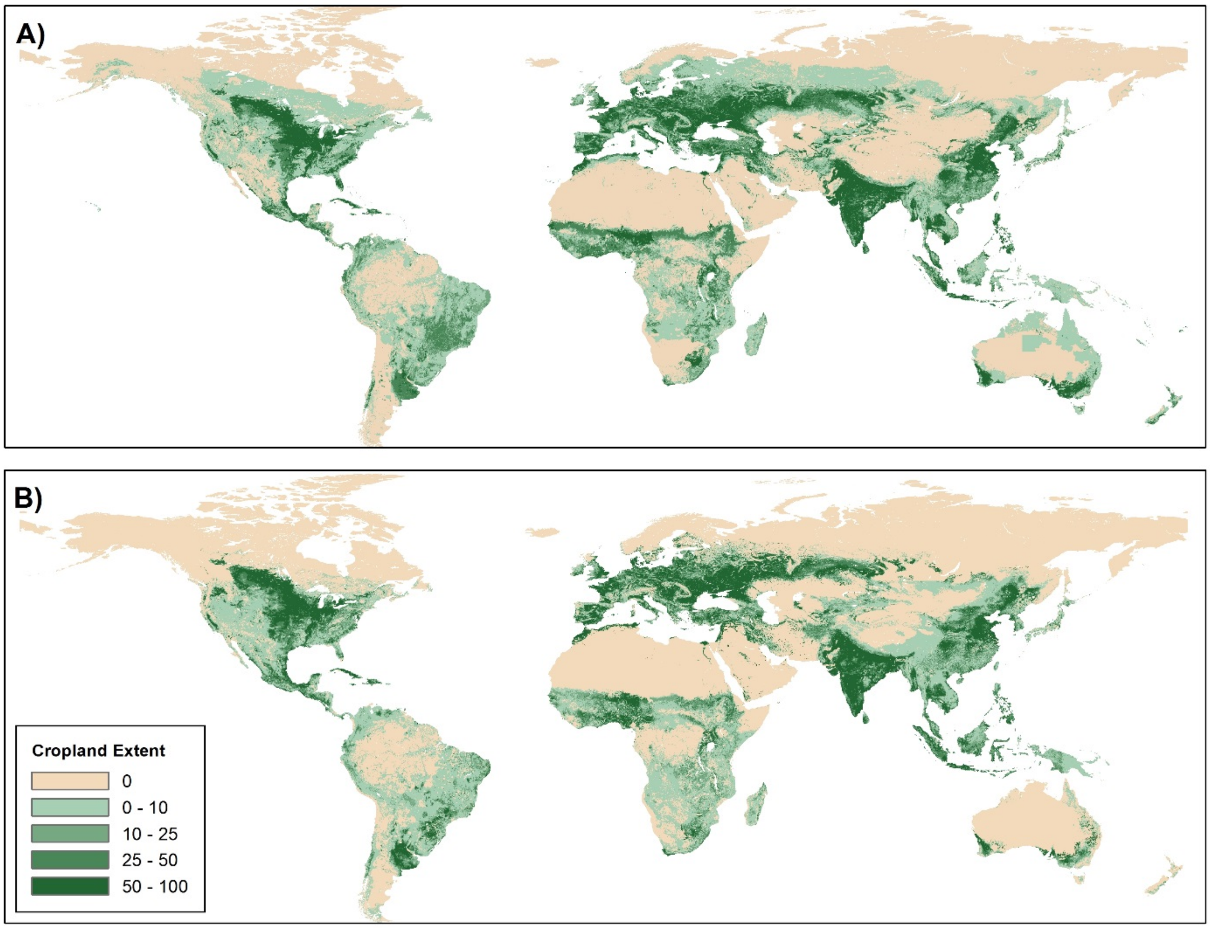


Figure S2: Pixel-wise cropland extent differences

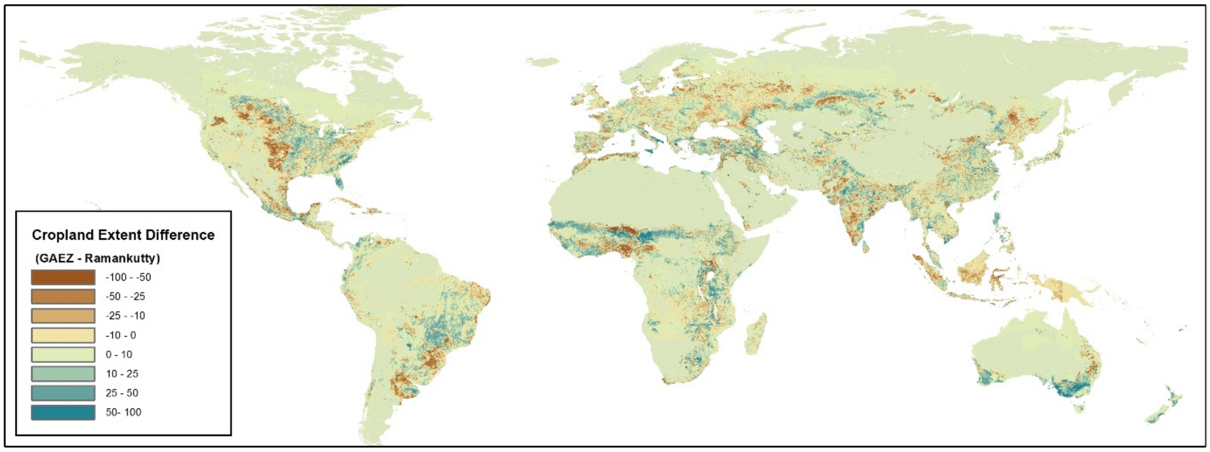
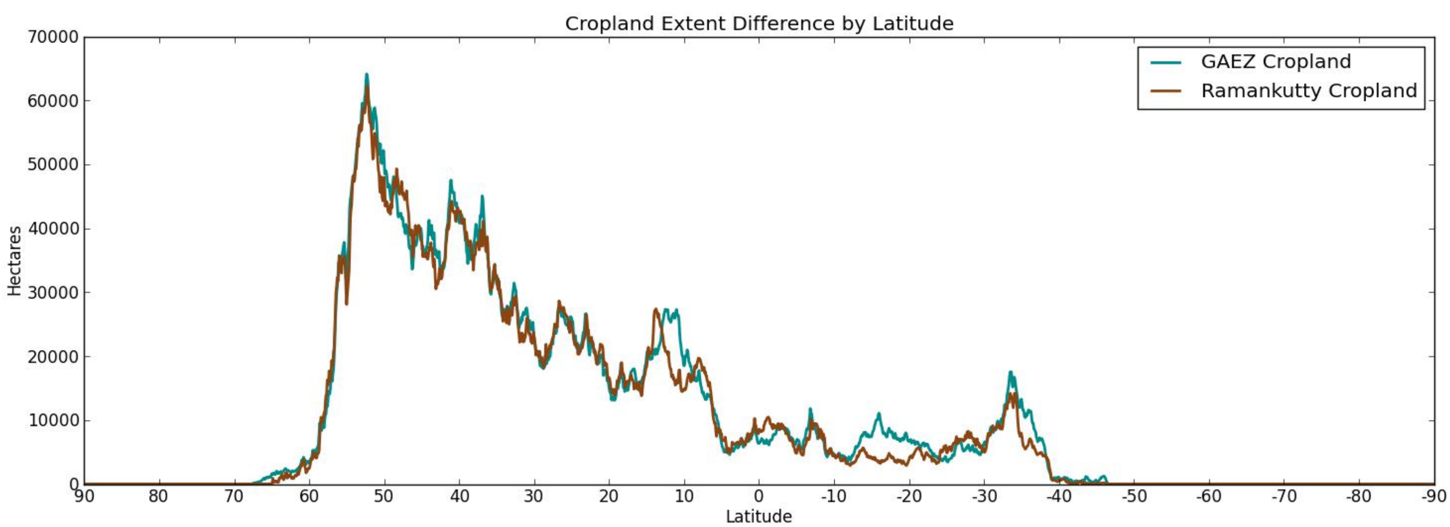
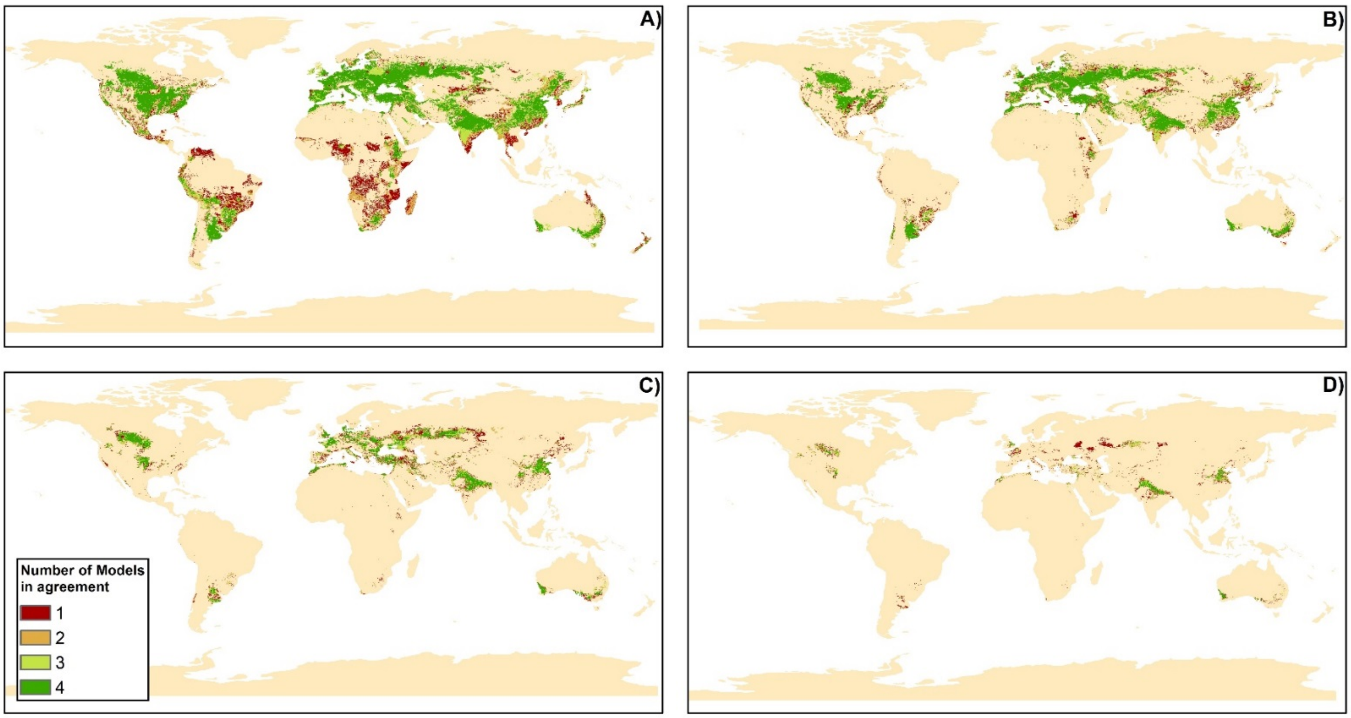


Figure S3: GAEZ and Ramankutty cropland Extent by Latitude



Appendix S5: Supplementary Figures for Wheat

Figure S1: Model agreement on magnitude of harvested area of wheat by threshold. A) Harvested Area > 0% of cell, B) Harvested Area > 1% of cell, C) Harvested Area > 10% of cell, D) Harvested Area > 25% of cell



Appendix S6: Supplementary Analysis for Rice and Maize

<u>S6.1 Rice harvested areas and yields</u>

Although there are minor departures from model consensus at higher thresholds, the majority of the discrepancies in the spatial distribution of rice are at the lowest threshold (see supplementary Figure S1 panel A). As discussed with wheat, these dissimilarities arise due to differences in model methodology. But in contrast to wheat, the models in disagreement are primarily the SPAM and MIRCA models (see supplementary Fig. S5 in Appendix S6.2), not only those that spread harvested area across plausible cropland, implying that the inconsistencies are in part attributable to the collection of supplementary subnational statistics as described in Section 4.1. At higher thresholds for harvested area the disagreement between products is minimal and will be explored further in the following analyses.

The harvested areas of rice appear to be less influenced by discrepancies in the cropland extent products, and instead differ as a function of downscaling method or input data. Evaluating the harvested areas of rice by latitude reveals that MIRCA predicts significantly more harvested area for rice north of 30N than do the other products (see Fig. S2). This divergence may stem from the fact that the MIRCA method of crop distribution does not consider biophysical limitations, or it may reflect differences in the input statistical crop yields at a sub-national level given that the above average estimation by MIRCA appears to be concentrated in eastern China (see Fig. S3).

With the exception of MIRCAs large harvested area in east China, the relation between GAEZ and each of the products that use Ramankutty cropland extent (M3, MIRCA and SPAM) is nearly identical (see Fig. S3). This may indicate that all three use similar sub-national rice data in China. M3, MIRCA and SPAM differ from GAEZ, for example, in their distribution of rice within India: GAEZ distributes more rice area to the southwest while other products distribute more rice to the northeast.

Maps of rice yields match even less well than did the wheat yields, differing significantly over all latitudes north of the equator (see Fig. S2). M3 predicts yields higher than SPAM and GAEZ over most latitudes but particularly in Asia (see Figs. S2 and S4). This can be seen in the skew of the distributions of the histograms in each pair-wise yield comparison. The differences between GAEZ and SPAM are most pronounced in China. SPAM predicts consistently lower yields than both M3 and GAEZ over nearly all of China.

Figure S5 illustrates the model-dependent differences, and resulting uncertainty, in calculating the yield gap (panel A) using both an absolute measure (panel C) and with regard to existing yields (panel B). Areas in which the yield gap uncertainty ratio approaches 1 signify areas in which uncertainty dominates the estimate of the yield gap. However, it is equally important to contextualize these uncertainties relative to existing yields and using an absolute measure of model difference. Areas displaying large values in all three panels indicate areas in which the model-estimated yield gaps disagree (panel C), where this disagreement is a significant proportion of the estimated yield gap (panel A) and in which the differences are important in the context of existing food production systems (panel B).

The model dependent uncertainty exceeds the estimated yield gap in significant parts of every rice-growing continent, meaning that uncertainty in the estimation of yields dominates the yield gap calculation in these regions. As with wheat, the differences relative to existing yields are decreased in some major producing areas, but remain significant in others. Of particular interest are the Ganges basin in India, Peru, parts of China and Indonesia, which display high values across all three indices.

S6.2 Maize harvested area and yield

Model comparisons for the harvested area of maize should be interpreted with some care as not all models measure identical quantities. M3, GAEZ and SPAM all measure maize as it is grown for grain only, while MIRCA measures maize as the sum of maize grown for silage and for grain. While there was insufficient model data to separate (or aggregate) the models quantities to align identically, a comparison between models is still useful for broadly identifying inter-model differences. For the yield analysis all models (M3, SPAM and GAEZ) did measure the same quantity.

The consensus analysis reveals that as compared with rice and wheat, the agreement between models on the extent of maize harvested area is generally good with a few notable exceptions. At the lowest threshold of harvested area, MIRCA is alone in designating growing area in much of Canada, Japan, the UK and Ireland while SPAM is the only model that indicates additional area in central Asia (see supplementary Fig. S6 panel A and supplementary Fig. S11 in Appendix S7.3). The SPAM differences are likely due to collection of additional national or sub-national statistics, while the MIRCA differences may arise due to a combination of the inclusion of silage and additional data collection. At the threshold of > 1% (panel B), model methodology dominates the differences: MIRCA indicates higher proportions of maize harvested area due to a more even distribution across statistical reporting units. In southeast China all models agree on the general extent of maize harvested area but disagree trivially on the spatial placement as evidenced by comparison with the results of the Gaussian analysis in this region. All models show a relative consensus on spatial extent of maize harvested area that covers >10% and > 25% of the cell (see panels C and D of S6).

As with the harvested area of wheat, the broad patterns of disagreement in the harvested areas of maize reflect differences in the cropland extent products in many regions. This relation is particularly apparent in South America, West Africa and North America (see Fig. S8). By latitude, MIRCA displays a larger harvested area than any other product at 50-60N (see Fig. S7). This may have to do with the lack of biophysical constraints in the distribution method or may be due to the inclusion of maize used for silage, as described earlier.

The maize yields show consistently different patterns both spatially and by lattitude. GAEZ, for example, displays consistently higher yields in the tropics (see Fig. S7), while M3 predicts significantly lower yields than either SPAM or GAEZ in Eastern United States (see Fig. S9). As evidenced by the skew in the histogram insets of the pair-wise comparisons, SPAM distributes maize yields to be significantly larger in a smaller number of cells (insets, Fig. S9).

Figure S10 illustrates the uncertainty ratio (panel A), the model-dependent differences in calculating the yield gap (panel C), and those differences relative to existing yields (panel C). Differences in the estimated yield gap exceed 1 tonne / ha over a majority of maize producing areas of the globe. As with both rice and wheat, uncertainty dominates the calculation of the yield gap in significant portions of every continent (see Fig. S10). Areas displaying large values in all three indices include parts of East and Southern Africa, Brazil, Mexico and Pakistan. These differences highlight the importance of continued efforts towards improving model estimates of harvested area and yield.

Figure S1: Model agreement on magnitude of harvested area of rice by threshold. A) Harvested Area > 0% of cell, B) Harvested Area > 1% of cell, C) Harvested Area > 10% of cell, D) Harvested Area > 25% of cell

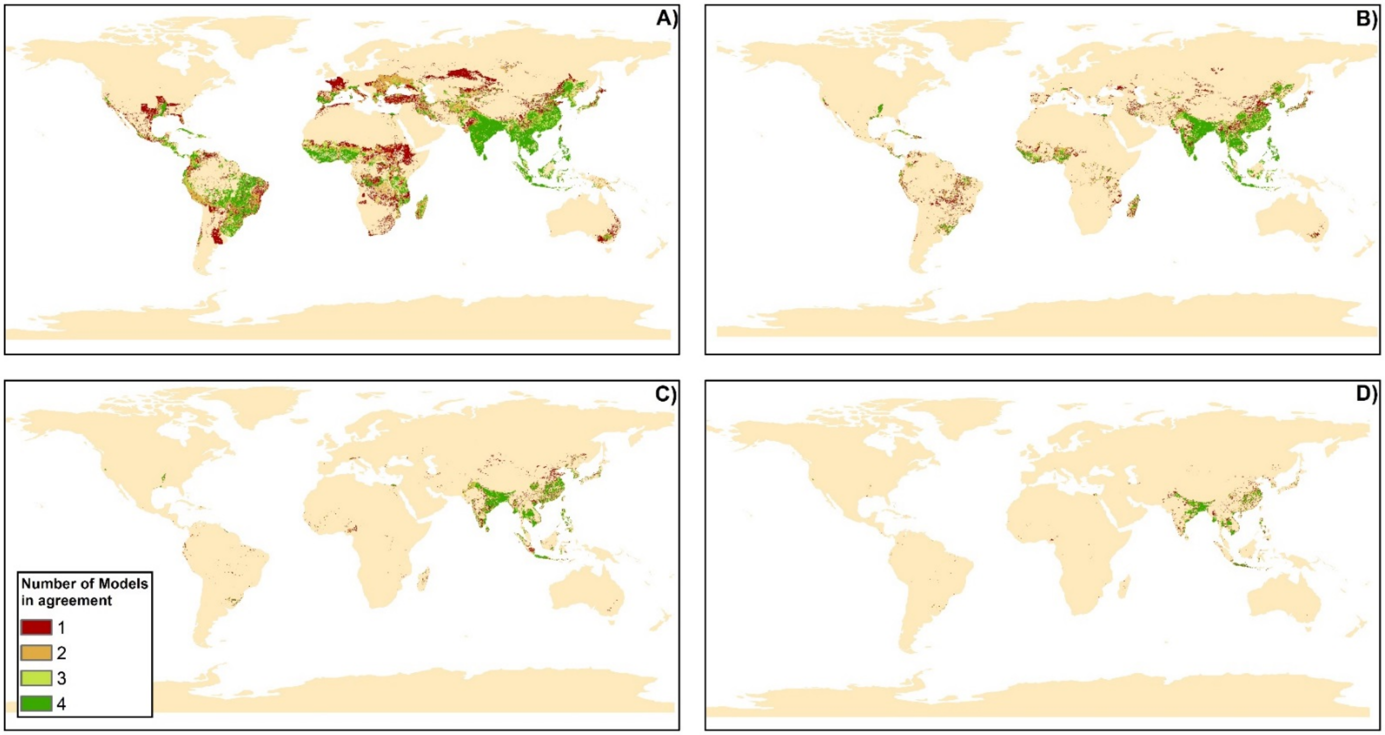


Figure S2: Rice harvested area and yield by latitude

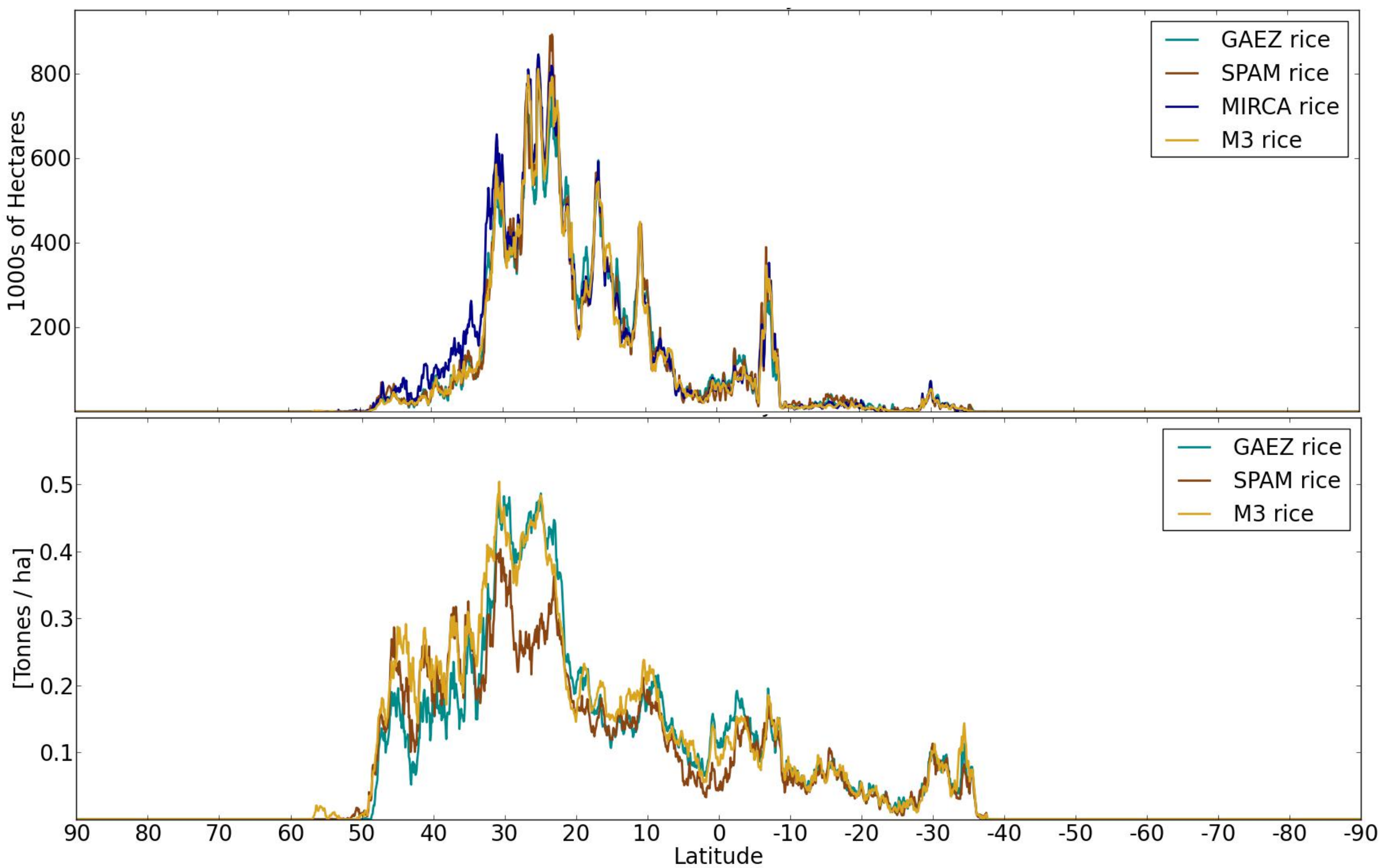


Figure S3: Comparison of rice harvested area by model following a Gaussian filter of three sigma kernel density. Histograms in each panel display the normalized percent of pixels as a function of harvested area, y-axis limits[0, 50%], x-axis limits[-5000, 5000] ha.

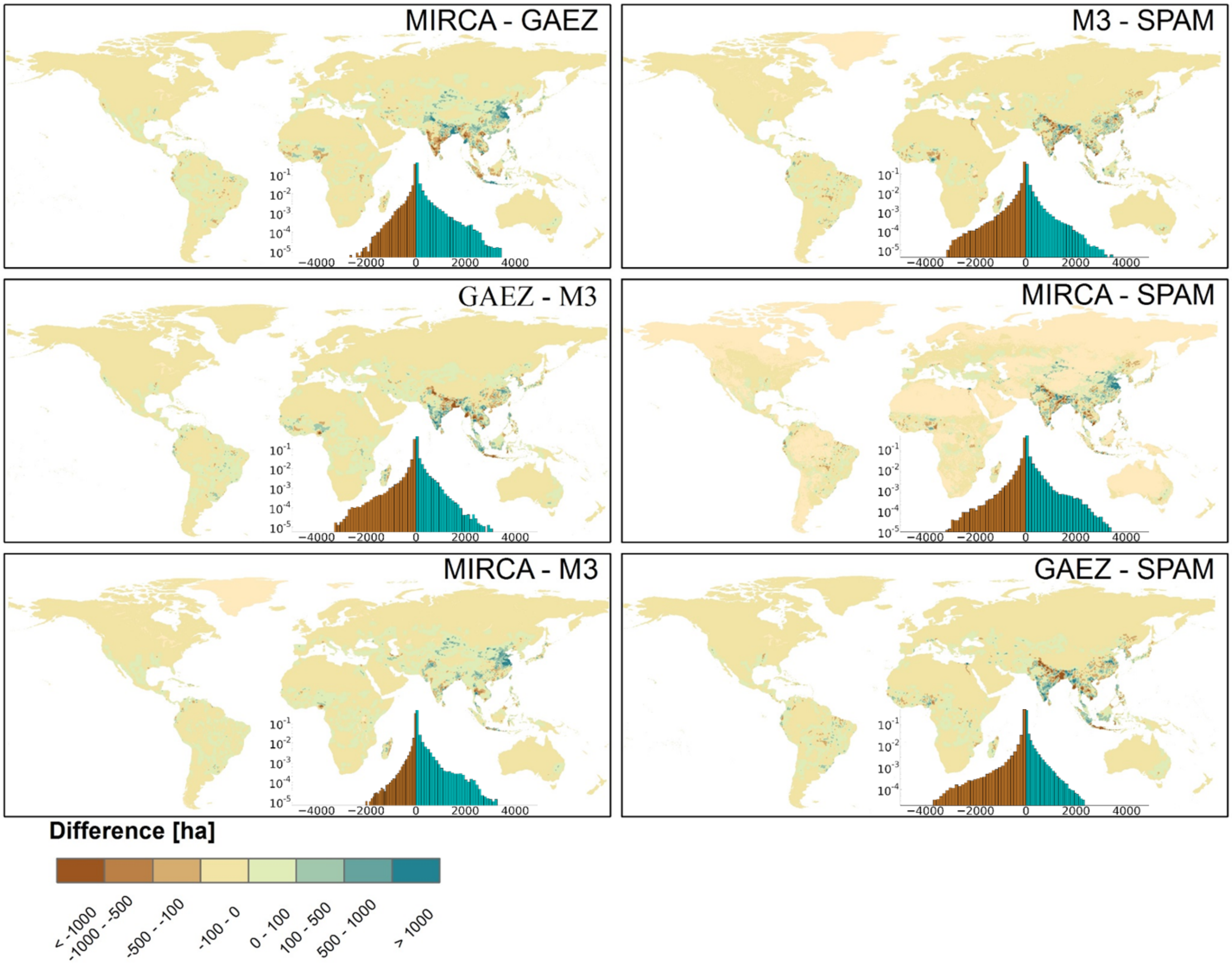


Figure S4: Comparison of rice yield by model following a Gaussian filter of three sigma kernel density. Histograms in each panel display the normalized percent of pixels as a function of yield, y-axis limits[0, 35%], x-axis limits[-20, 20] tonnes/ha.

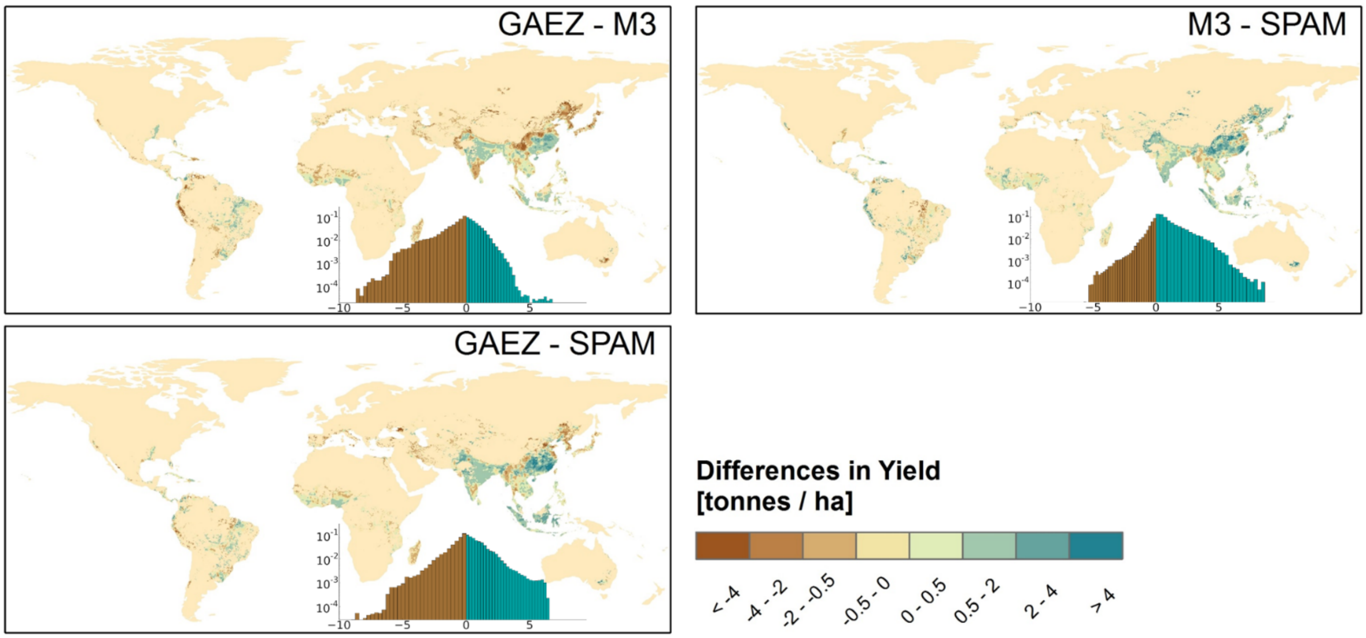


Figure S5: Implications of model differences for estimated rice yield gaps. A) Yield Gap Uncertainty Ratio: average model difference divided by average estimated yield gap B) average difference in estimated yield gap divided by existing yield C) average difference in estimated yield gap

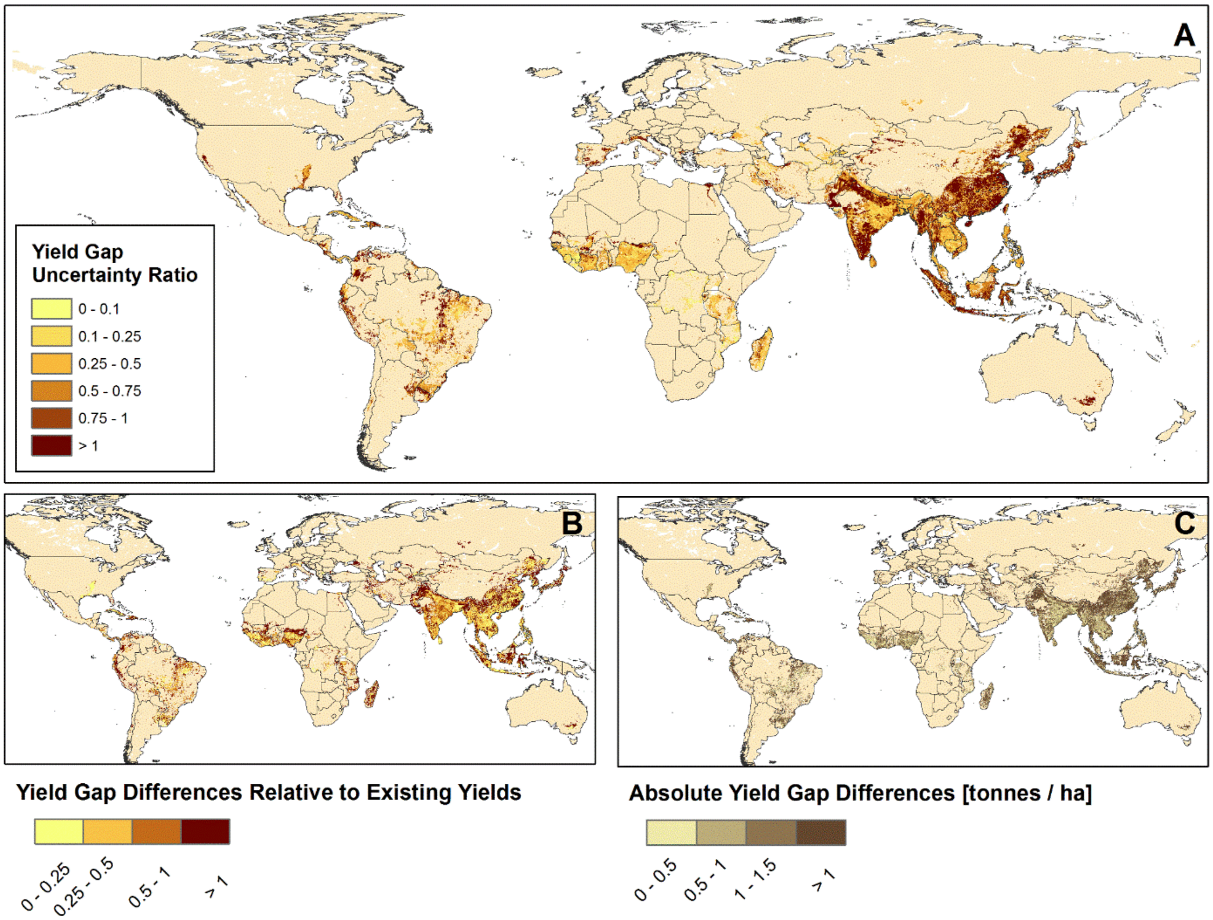
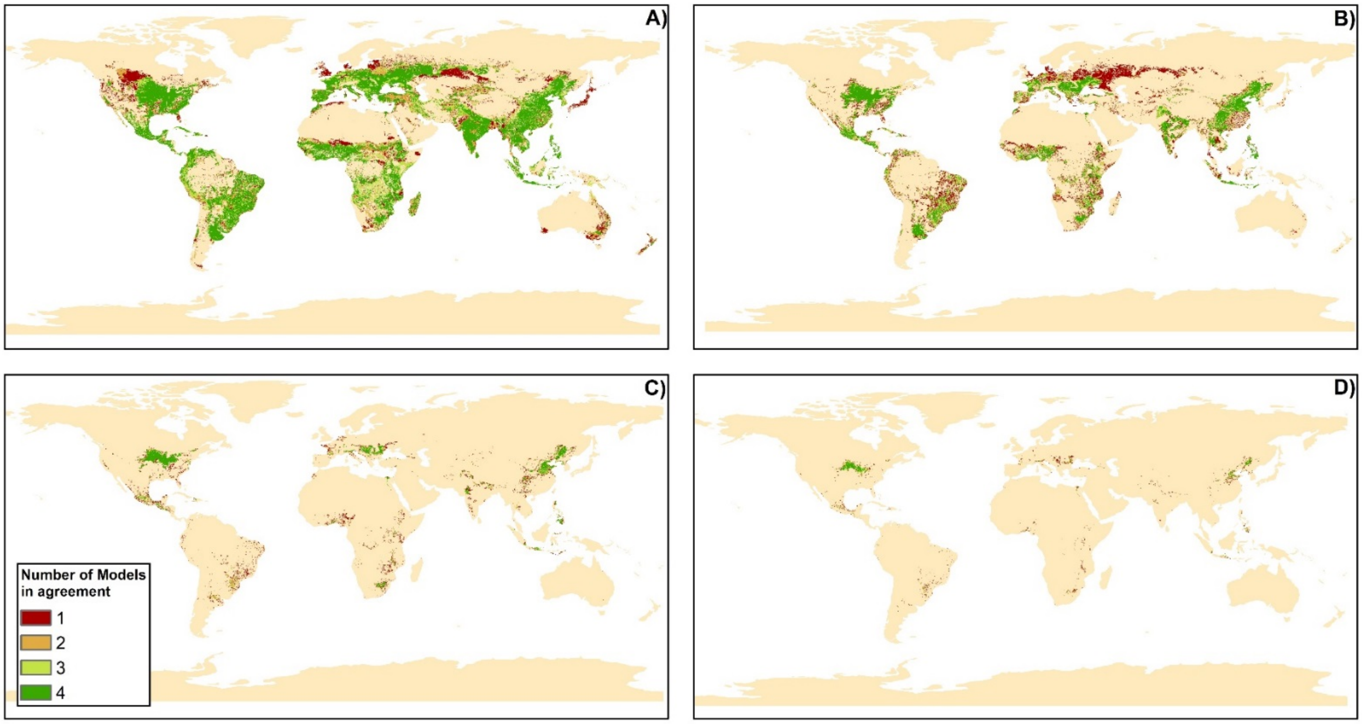


Figure S6: Model agreement on magnitude of harvested area of maize by threshold. A) Harvested Area > 0% of cell, B) Harvested Area > 1% of cell, C) Harvested Area > 10% of cell, D) Harvested Area > 25% of cell



628 *Figure S7: Maize harvested area and yield by latitude*

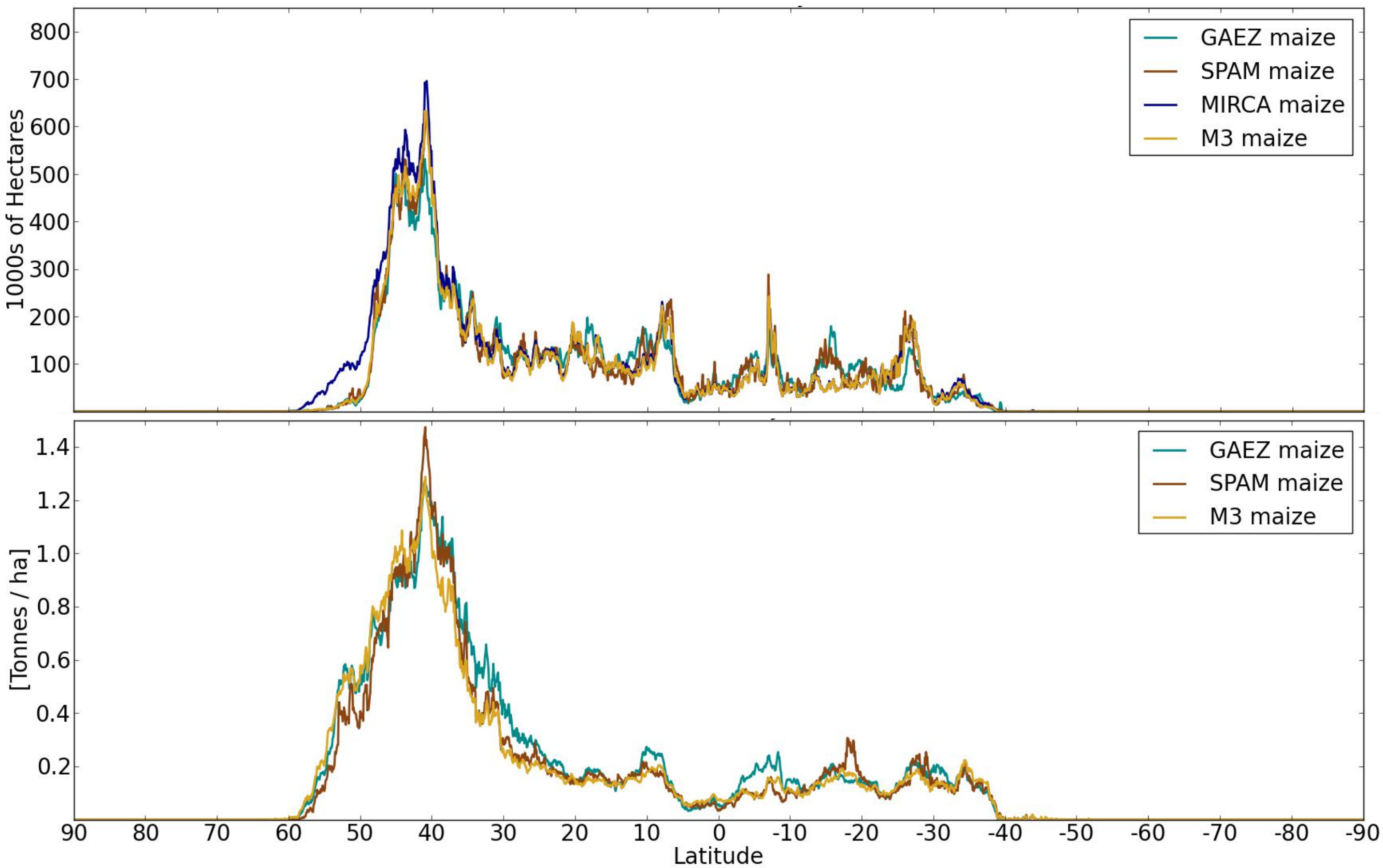


Figure S8: Comparison of maize harvested area by model following a Gaussian filter of three sigma kernel density. Histograms in each panel display the normalized percent of pixels as a function of harvested area, y-axis limits[0, 35%], x-axis limits[-5000, 5000] ha.

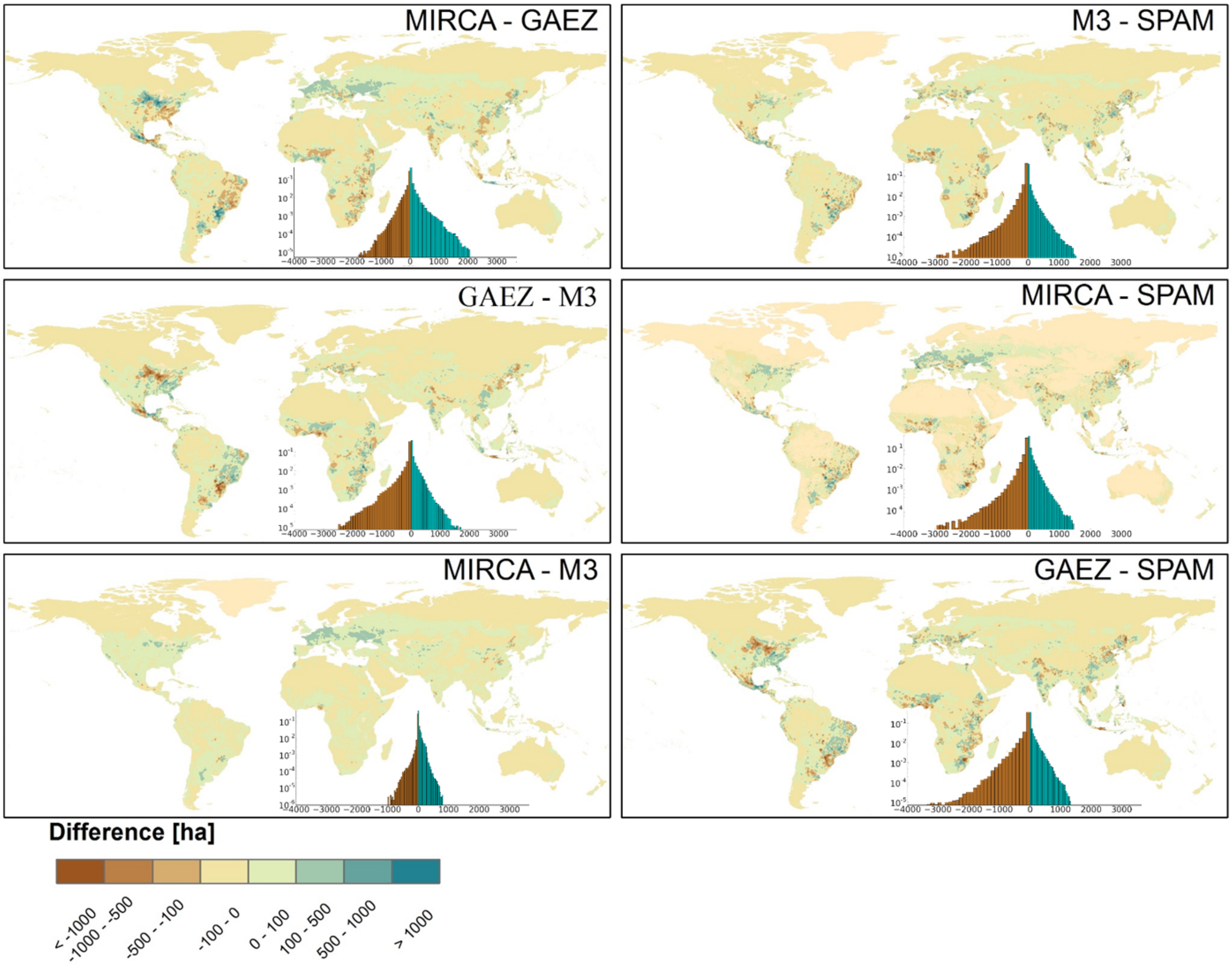


Figure S9: Comparison of maize yield by model following a Gaussian filter of three sigma kernel density. Histograms in each panel display the normalized percent of pixels as a function of yield, y-axis limits[0, 35%], x-axis limits[-20, 20] tonnes/ha.

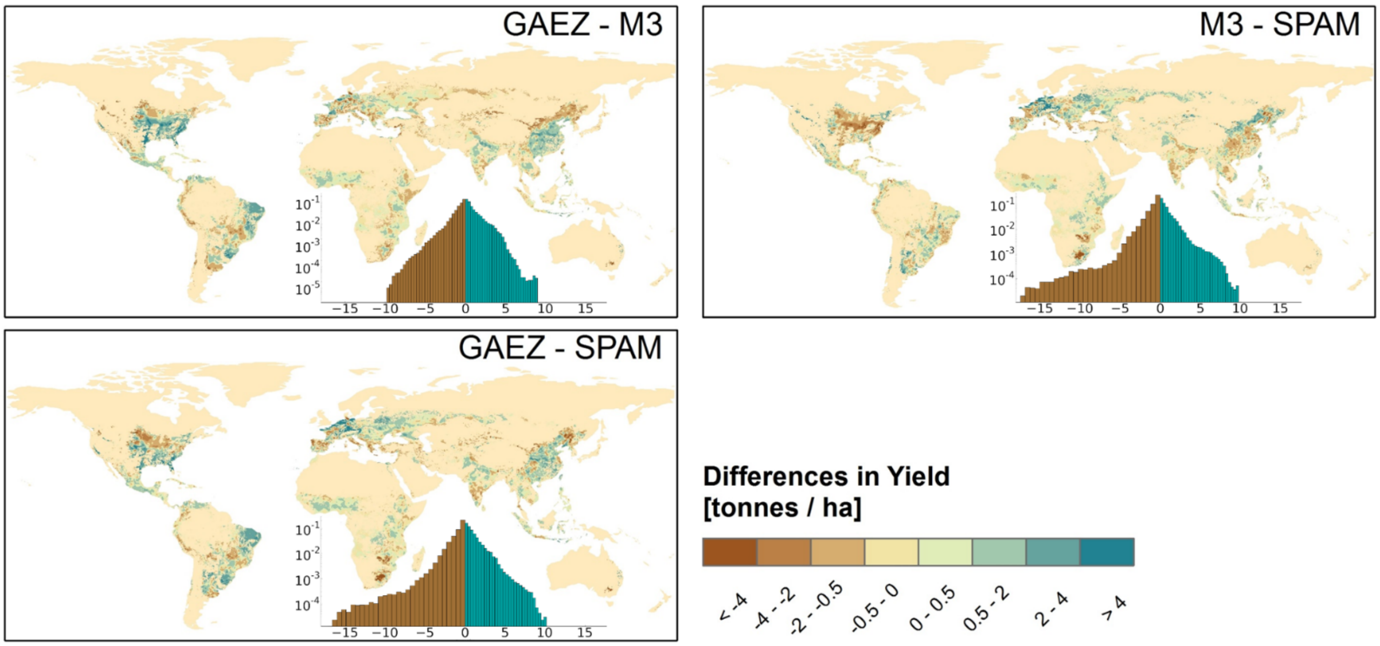
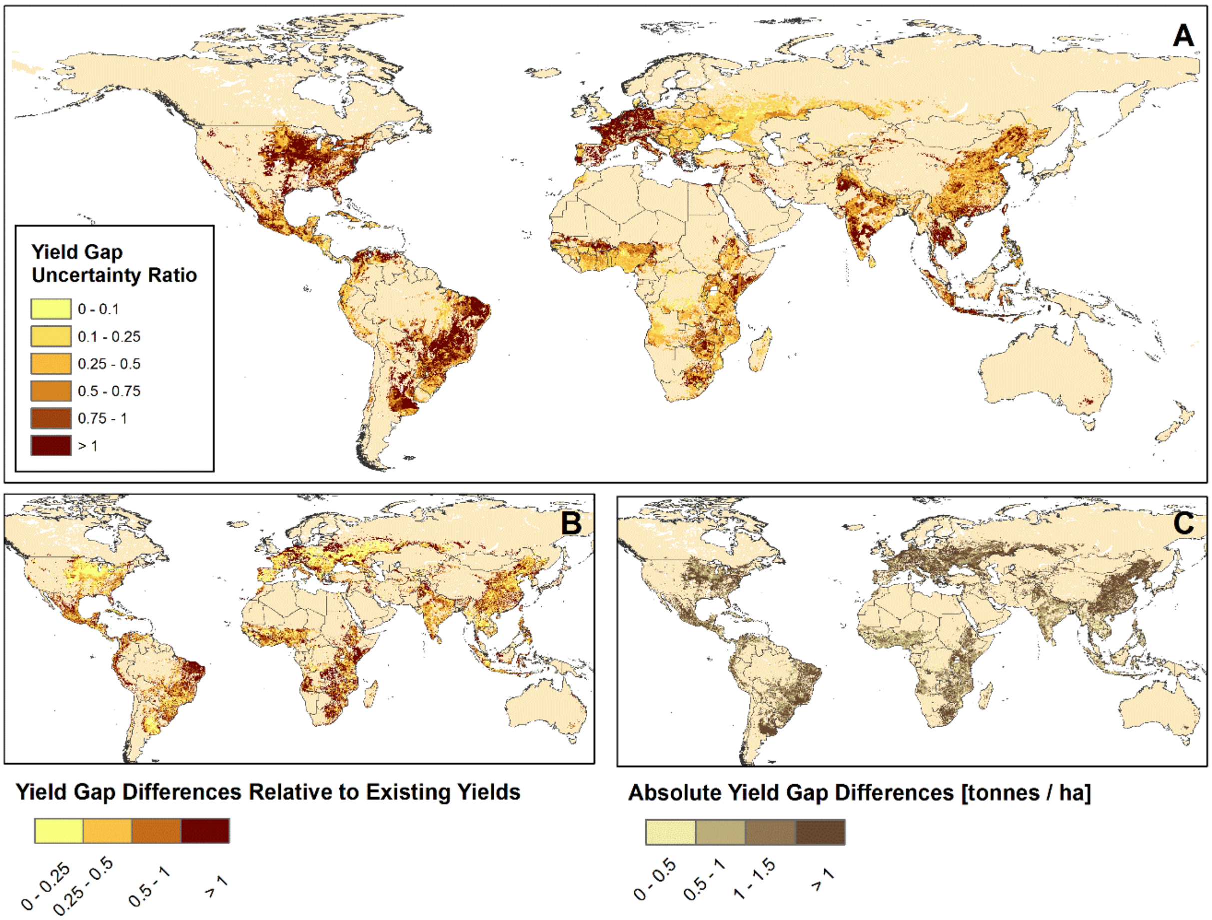


Figure S10: Implications of model differences for estimated maize yield gaps. A) Yield Gap Uncertainty Ratio: average model difference divided by average estimated yield gap B) average difference in estimated yield gap divided by existing yield C) average difference in estimated yield gap



Appendix S7: Supplementary pixel-wise figures for wheat, rice and maize

S7.1 Pixel-wise figures for wheat:

Figure S1: Wheat harvested area for A) M3, B) GAEZ, C) MIRCA and D) SPAM

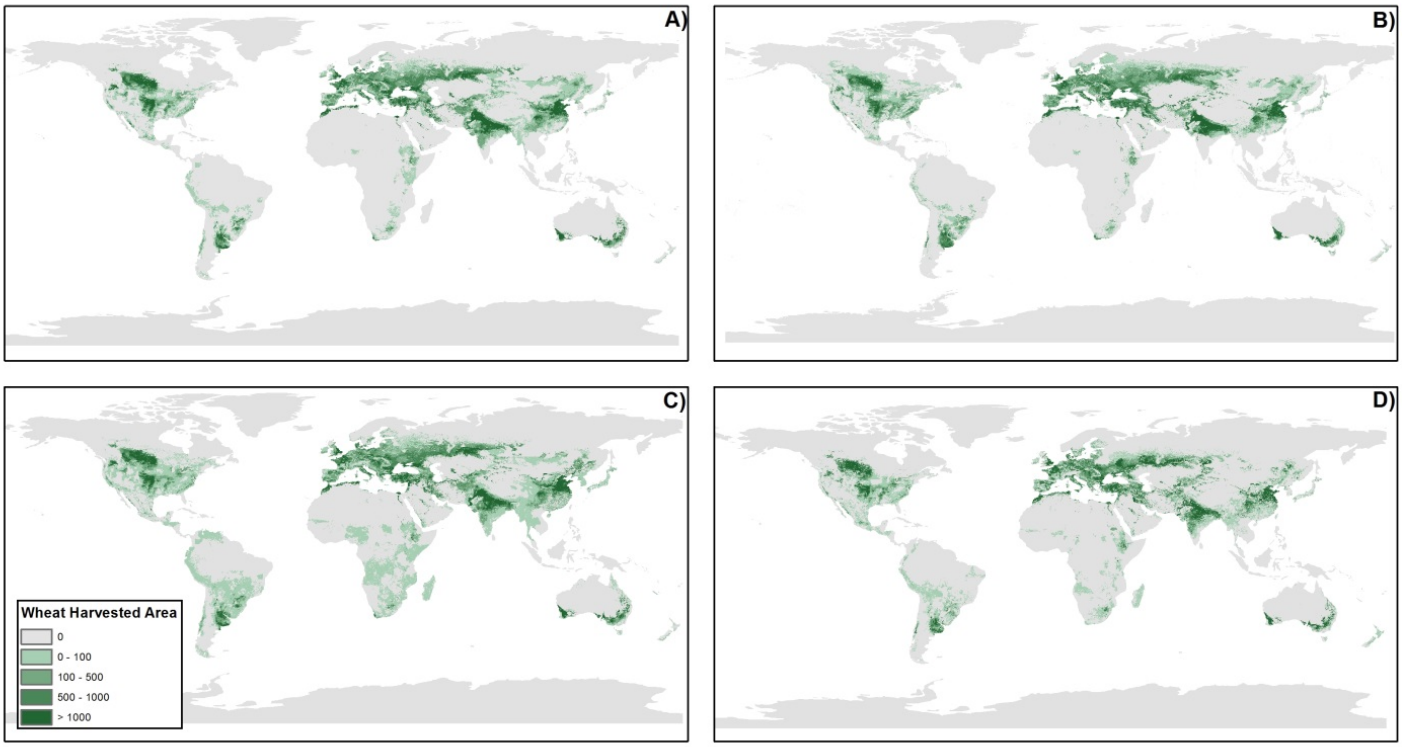


Figure S2: pixel-wise comparison of the wheat harvested area by model

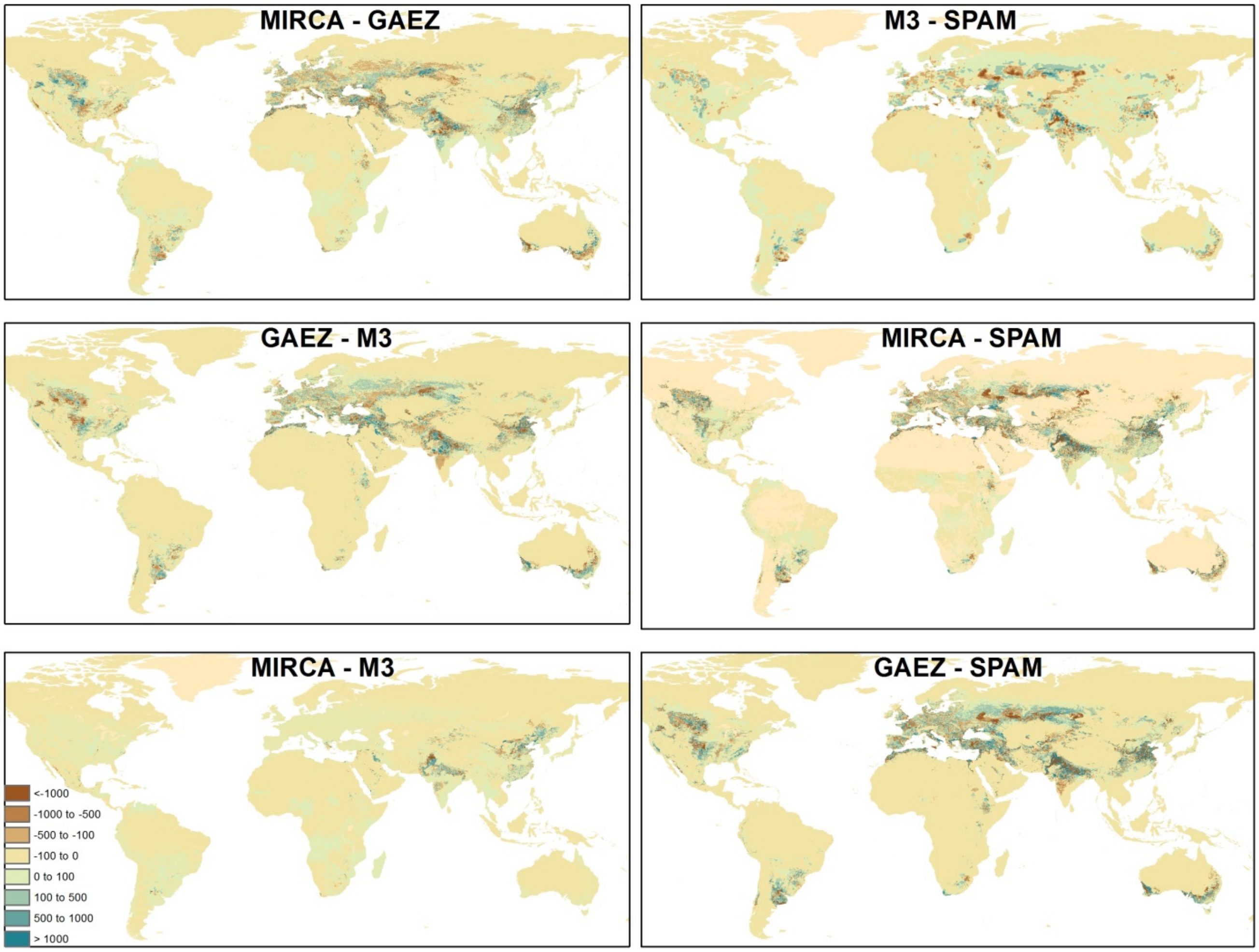


Figure S3: Wheat yield for A) M3, B) GAEZ, and C) SPAM

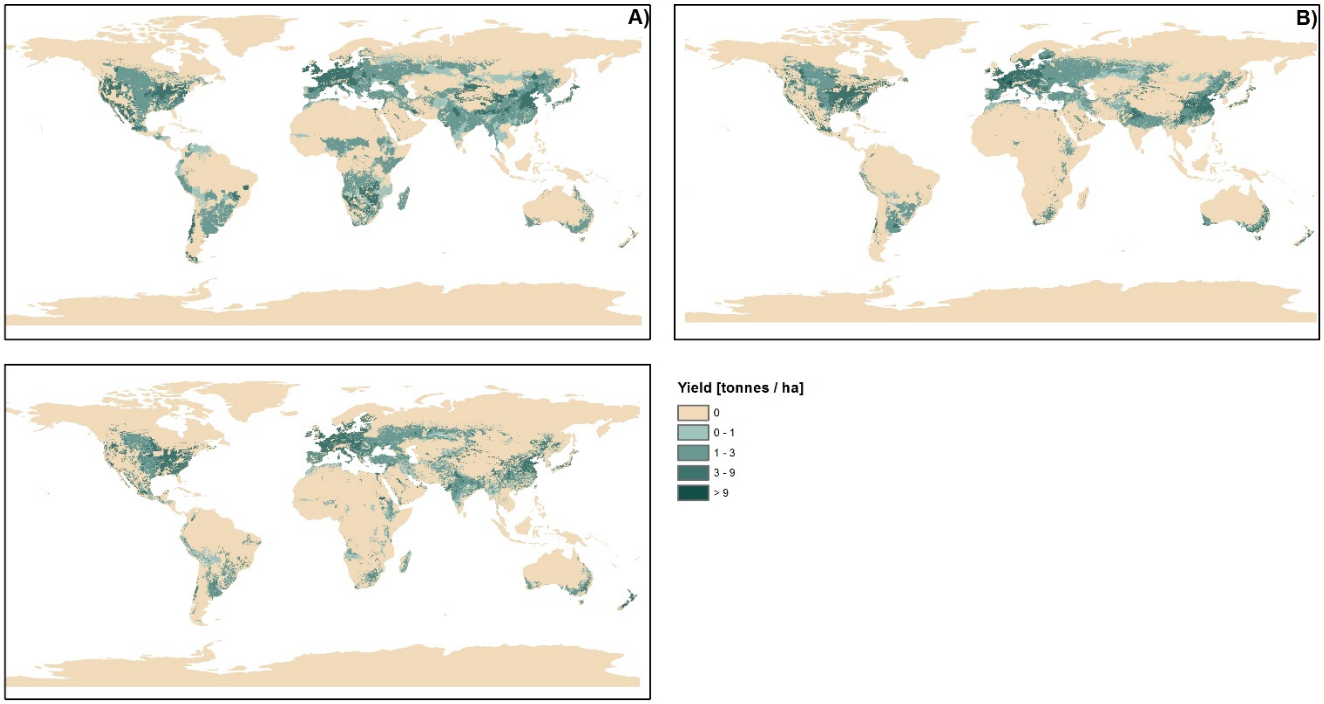
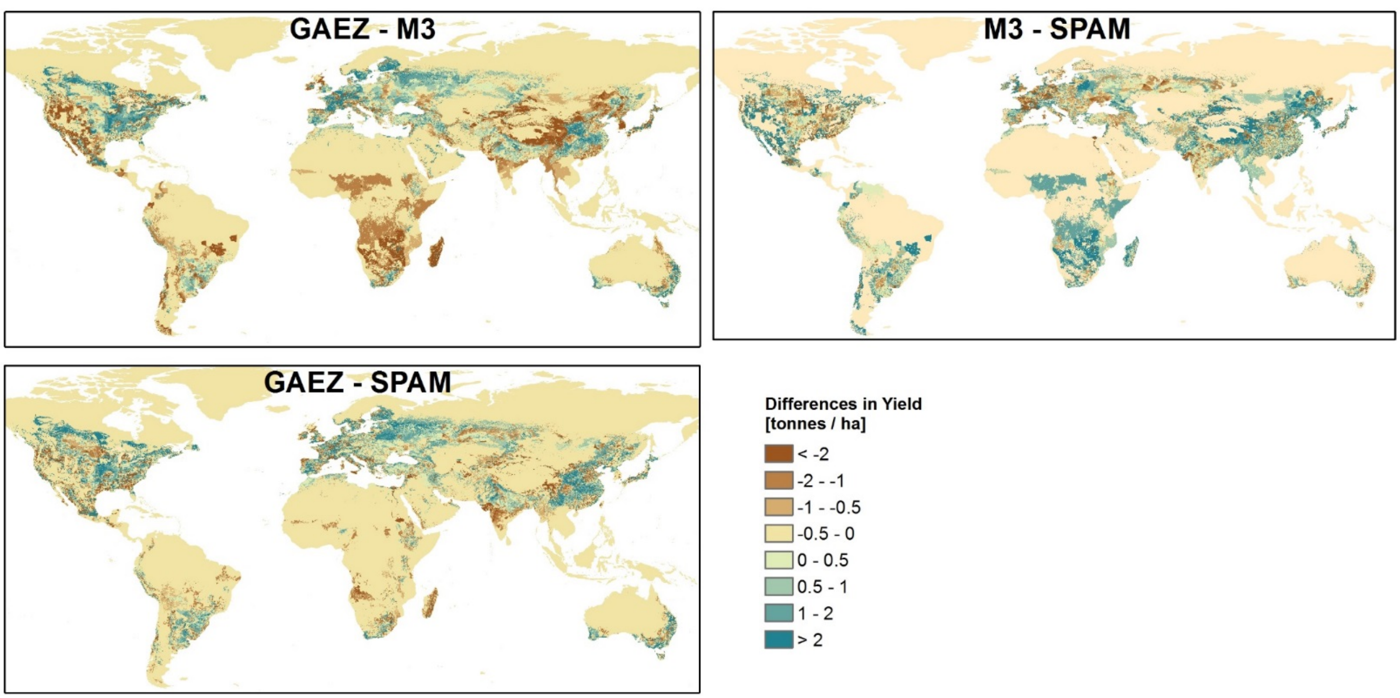


Figure S4: pixel-wise comparison of the wheat yield by model



S6.2 Pixel-wise figures for rice

Figure S5: Rice harvested area for A) M3, B) GAEZ, C) MIRCA and D) SPAM

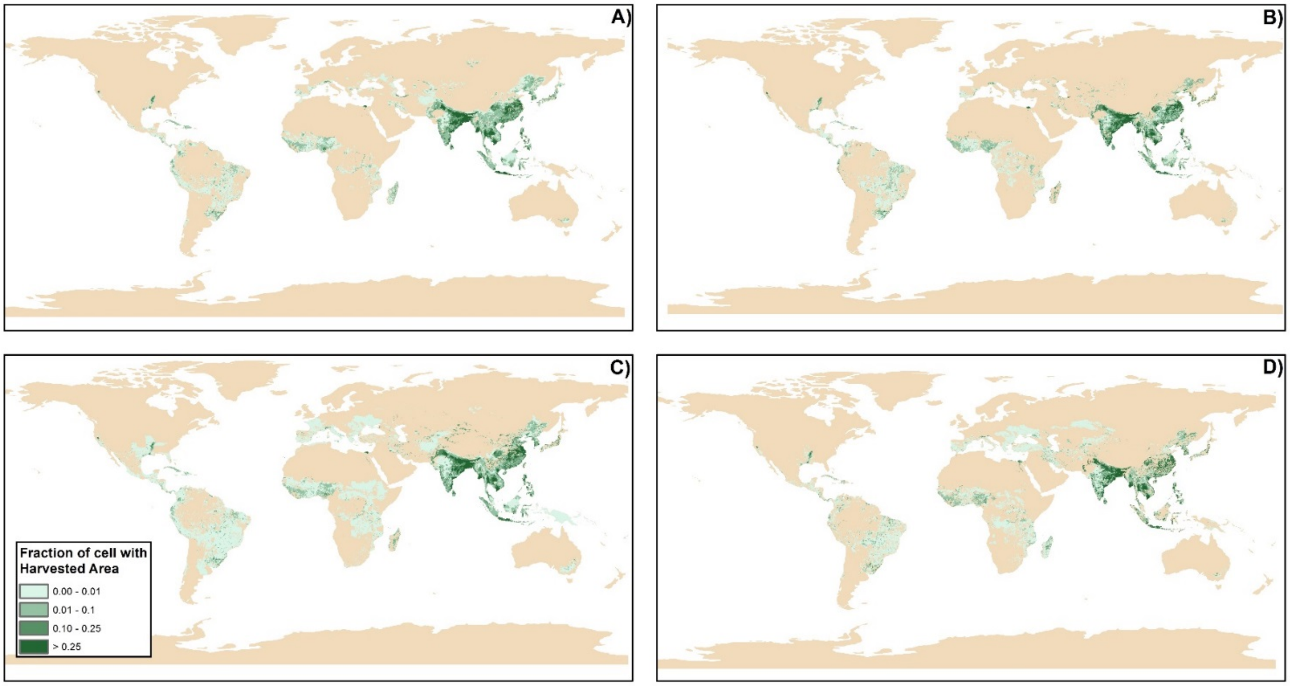


Figure S6: pixel-wise comparison of the rice harvested area by model

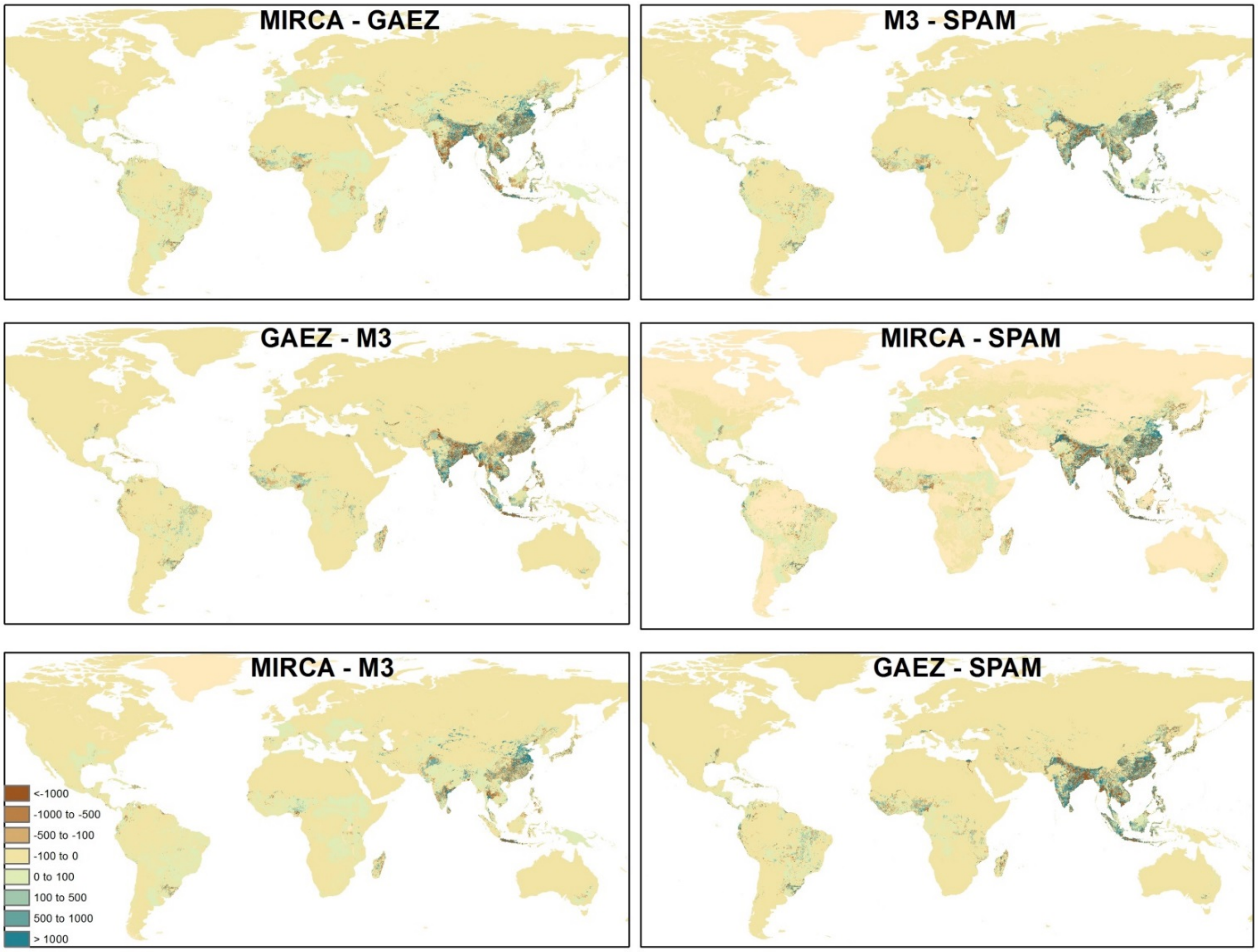


Figure S7: Rice yield for A) M3, B) GAEZ, and C) SPAM

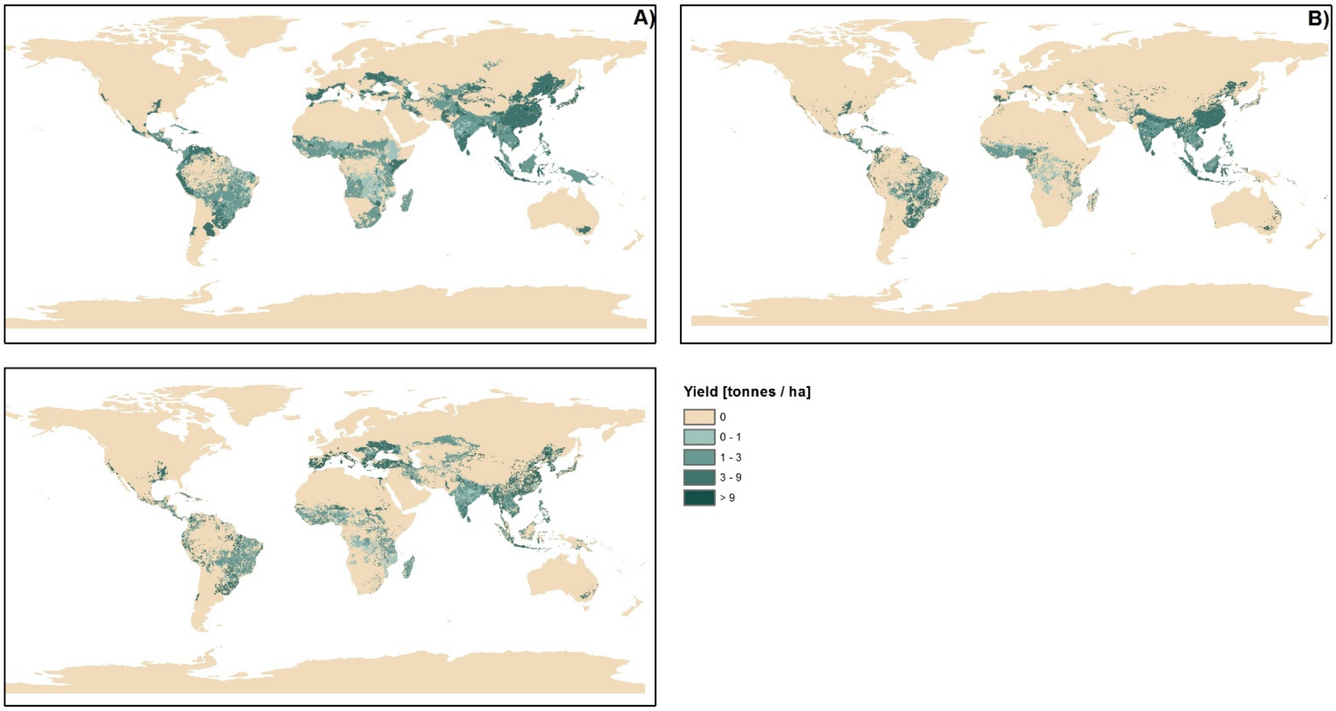
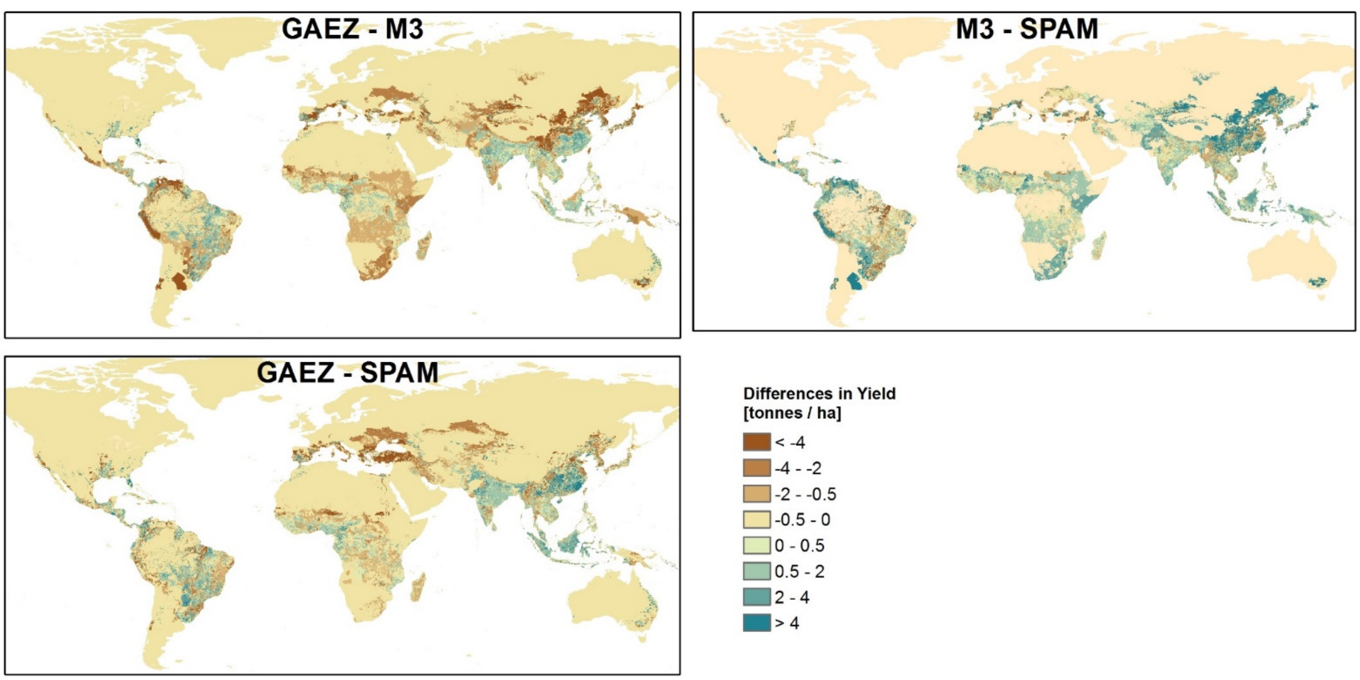


Figure S8: pixel-wise comparison of the rice yield by model



S7.3 Pixel-wise figures for maize

Figure S9: Maize harvested area for A) M3, B) GAEZ, C) MIRCA and D) SPAM

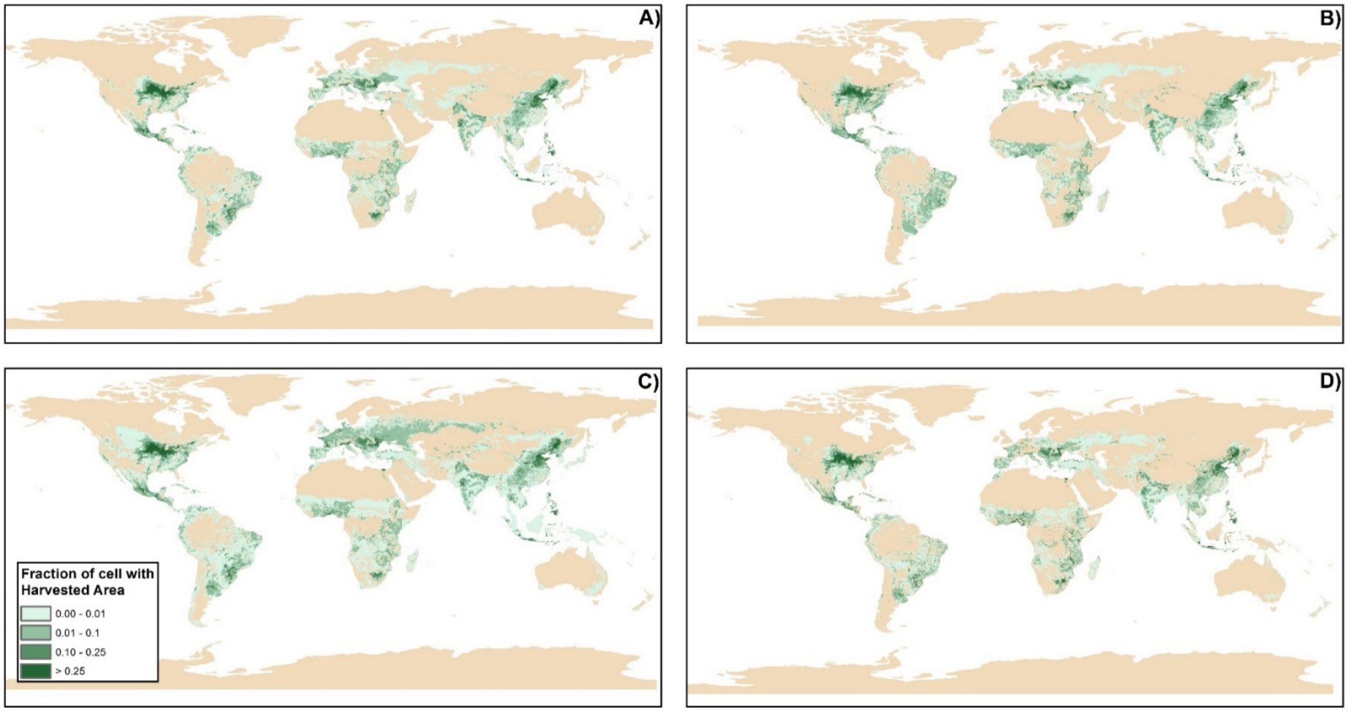


Figure S10: pixel-wise comparison of the maize harvested area by model

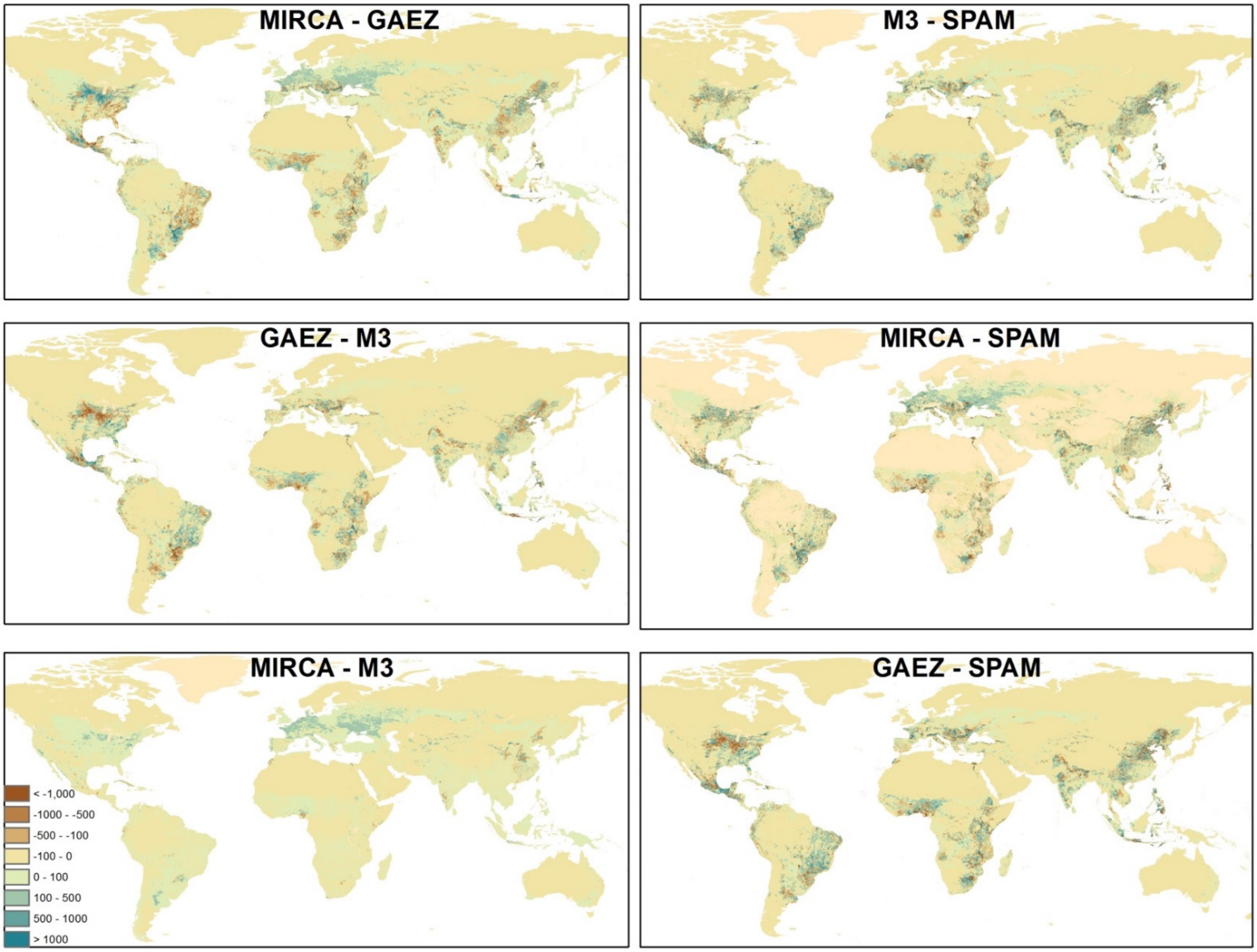


Figure S11: Maize yield for A) M3, B) GAEZ, and C) SPAM

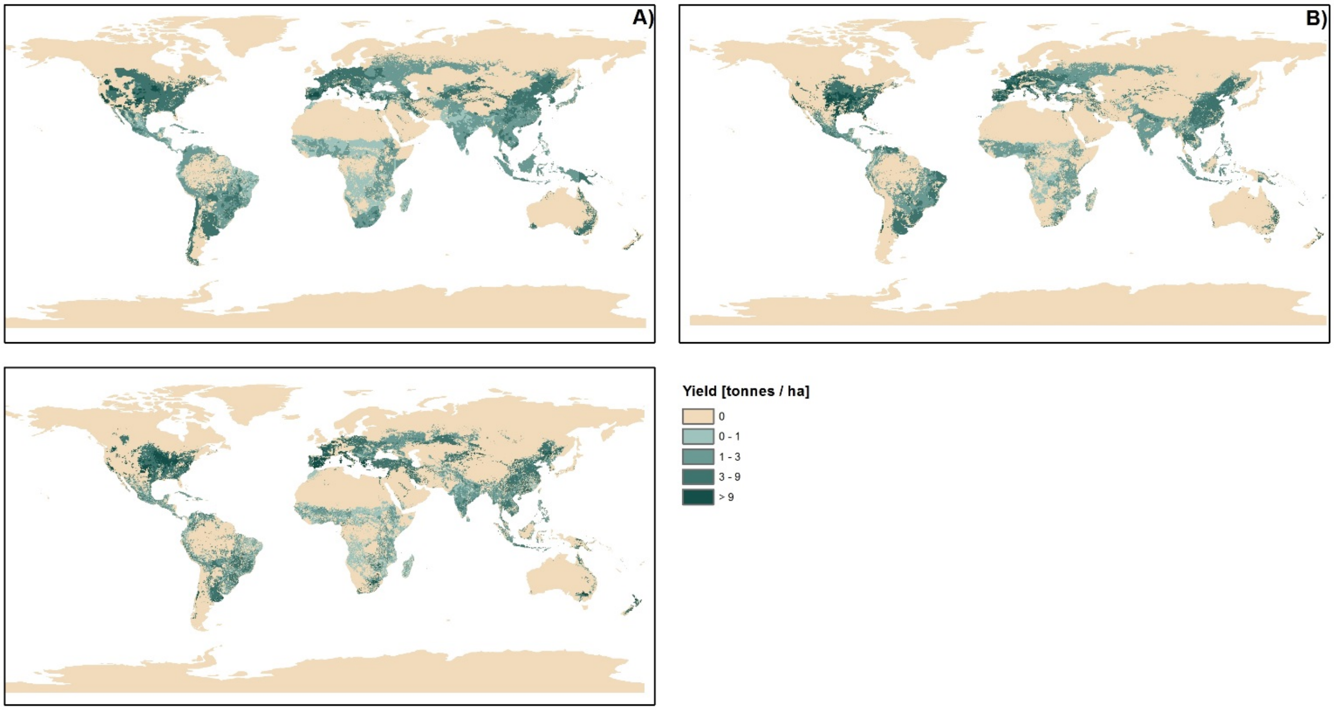
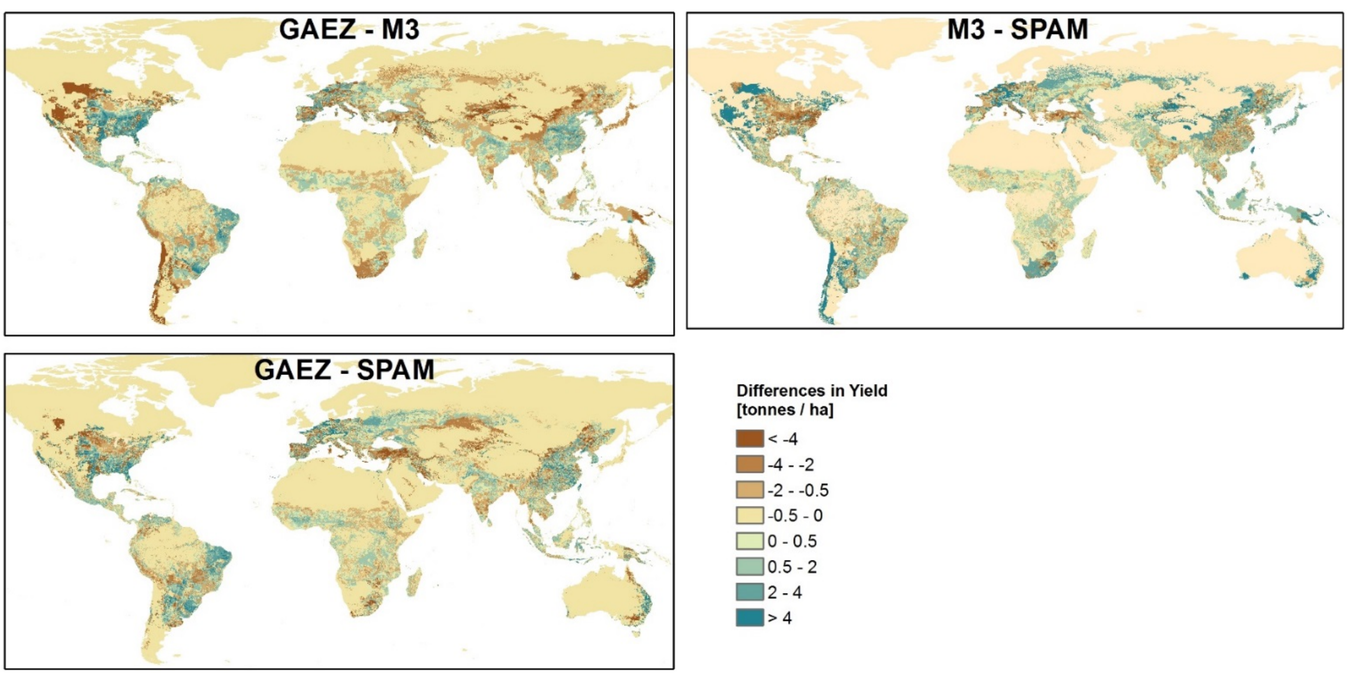


Figure S12: pixel-wise comparison of the maize yield by model



WORKS CITED

[1] Bartholome, E., & Belward, A. S. (2005), GLC2000: A new approach to global land cover mapping from Earth Observation data, *International Journal of Remote Sensing*, **26**, 1959– 1977.

[2] Bhaduri, B.B., Bright, E., Coleman, P. & Dobson, J. (2002) LandScan. Geoinformatics, 5, 34–37.

[3] Dobson, J.E., Brlght, E.A., Coleman, P.R. & Worley, B.A. (2000) LandScan: A Global Population Database for Estimating Populations at Risk. 66, 849–857.

[4] Fischer, G., Ermolieva, T., Ermoliev, Y. & Velthuizen, H.T. Van (2006) Spatial Recovering of Agricultural Values from Aggregate Information : Sequential Downscaling Methods. International Journal of Knowledge and Systems Sciences (ISKSS), 3.

[5] Fischer, G., F.O. Nachtergaele, S. Prieler, E.Teixeira, G.Tóth, H. van Velthuizen, L. Verelst, D. Wiberg (2013). Global Agro-Ecological Zones (GAEZ v3.0), http://gaez.fao.org/Main.html#. Accessed October 2013.

[6] Golan, A., Judge, G. & Miller, D., 1996. Maximum Entropy Econometrics: Robust Estimation with Limited Data, New York: John Wiley & Sons

[7] Jaynes, E.T. (1957) Information Theory and Statistical Mechanics. 320–330.

[8] Paris, Q. (1998) An analysis of ill-posed production problems using maximum entropy. American Journal of Agricultural Economics, 80, 124–138.

[9] Portmann, F., Siebert, S., Bauer, C. & Döll, P. (2008) Global dataset of monthly growing areas of 26 irrigated crops. Frankfurt Hydrology Paper 06, Institute of Physical Geography.University of Frankfurt, Frankfurt am Main, Germany

[10] Portmann, F.T., Siebert, S. & Döll, P. (2010) MIRCA2000-Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling. Global Biogeochemical Cycles, 24.

[11] Shannon, C.E. (1948) A Mathematical Theory of Communication. Bell System Technology Journal, 27, 379–423.