# Evaluation of Integrated Assessment Model hindcast experiments: A case study of the GCAM 3.0 land use module

Abigail C. Snyder[1], Robert P. Link[1], and Katherine V. Calvin[1]

[1]Joint Global Change Research Institute, Pacific Northwest National Laboratory, College Park, MD 20740

*Correspondence to:* Abigail Snyder (abigail.snyder@pnnl.gov)

**Abstract.** Hindcasting experiments (conducting a model forecast for a time period in which observational data is available) are being undertaken increasingly often by the Integrated Assessment Model (IAM) community, across many scales of models. When they are undertaken, the results are often evaluated using global aggregates or otherwise highly aggregated skill scores that mask deficiencies. We select a set of deviation based measures that can be applied at different spatial scales (regional versus global) to make evaluating the large number of variable-region combinations in IAMs more tractable. We also identify performance benchmarks for these measures, based on the statistics of the observational dataset, that allow a model to be evaluated in absolute terms rather than relative to the performance of other models at similar tasks. An ideal evaluation method for hindcast experiments in IAMs would feature both absolute measures for evaluation of a single experiment for a single model and relative measures to compare the results of multiple experiments for a single model or the same experiment repeated across multiple models, such as in community intercomparison studies. The performance benchmarks highlight the use of this scheme for model evaluation in absolute terms, providing information about the reasons a model may perform poorly on a given measure and therefore identifying opportunities for improvement. To demonstrate the use of and types of results possible with the evaluation method, the measures are applied to the results of a past hindcast experiment focusing on land allocation in the Global Change Assessment Model (GCAM) version 3.0. The question of how to more holistically evaluate models as complex as IAMs is an area for future research. We find quantitative evidence that global aggregates alone are not sufficient for evaluating IAMs that require global supply to equal global demand at each time period, such as GCAM. The results of this work indicate it is unlikely that a single evaluation measure for all variables in an IAM exists, and therefore sector by sector evaluation may be necessary.

## 1 Introduction

Integrated assessment models (IAMs) couple human and physical Earth systems to explore the impacts of economic and environmental policies (Parson and Fisher-Vanden, 1997; Parson et al., 2007). IAMs are usually calibrated to a historical base year and simulate forward in time by incorporating changes in quantities such as population, GDP, technology, and policy to produce outputs that include land use, emissions, and commodity prices. Hindcast experiments use a model to produce a forecast simulation over a time period for which observational data is available. The ability to compare simulation data with observational data presents new opportunities for understanding a model's strengths and identifying avenues for improvement, and raises new research questions to explore. A variety of hindcast studies in IAMs of varying scale have used different metrics for evaluation studies, often driven by the research question of interest (Calvin et al., 2017; Fujimori et al., 2016; Baldos and Hertel, 2013; Beckman et al., 2011; van Ruijven et al., 2010b, a; Kriegler et al., 2015). However, no community standard for evaluation of IAMs currently exists, making it more difficult to compare results of hindcast experiments from different models. This work outlines goals for evaluating IAM hindcast experiments.

The Global Change Assessment Model version 3.0 (GCAM) (Calvin et al., 2011; Kim et al., 2006; Clarke et al., 2007; Kyle et al., 2011) was recently used to conduct a hindcast experiment (Calvin et al., 2017). Calvin et al., hereafter referred to as Paper 1, used skill scores (Reichler and Kim, 2008; Taylor, 2001; Schwalm et al., 2010) to compare performance of the land use module of GCAM under structurally different operating assumptions to an observational data set. The different scenarios represent different extremes of information for decision making given to the GCAM economic agents. One finding of this hindcast experiment with GCAM 3.0 was that the highly aggregated nature that makes the skill scores examined convenient also masks important deficiencies, limiting the insight they can provide for model development. A key question raised by this experiment, and which this work examines in greater detail, is how to actually define "improvement". The ease of use of skill scores has to be balanced with illuminating as many model deficiencies as possible. Only once a definition of improvement has been decided upon can parameter estimation studies be undertaken, as ranging over parameter values is only a useful task if one can quantitatively identify the parameter values that give the best agreement with historical data.

From this work, four goals for development of an IAM hindcast evaluation scheme were identified. A desirable evaluation method will proved information about the absolute performance of a single model run and may be used to measure relative performance of multiple model runs (from a single model or across many models of the same variables). Additionally, we seek a method that can describe multiple aspects of model performance at multiple scales, providing a flexible organizational structure for analyzing the large amount of data generated by IAMs while investigating particular hypotheses of interest. And finally, the method should include at least one metric that can be used as a cost function in future Monte Carlo-style parameter estimation experiments. Given these goals, it is unlikely that a single metric could be arrived at to satisfy all four. Rather, a condensed set of related metrics that together accomplish all four goals is sought for *evaluating* IAMs. The result of applying the set of metrics to model runs may be interpreted to identify future avenues for model *improvement* of a particular IAM. The implementation of such improvements is outside the scope of this paper.

Our evaluation goals are not independent of each other. A metric that provides absolute performance insight can be calculated for multiple model runs and compared to provide relative performance information. A metric evaluating a particular aspect of model performance may be used to estimate parameters to improve that aspect of model performance.

Several other works in the IAM hindcasting literature (Baldos and Hertel, 2013; Beckman et al., 2011; van Ruijven et al., 2010b, a; Kriegler et al., 2015; Fujimori et al., 2016) do not meet all four of our goals. For example, in the hindcast experiment performed for the energy component of the AIM/CGE model, Fujimori et al. present two statistics: a regression technique and an error statistic for global aggregates. The regression technique identifies regions and variables for which model performance may be improved. While the regression technique can produce desirable region-specific information about model performance and shortcomings for multiple variables, it unfortunately cannot be leveraged as a performance metric for future Monte Carlo-style parameter estimation exercises. It is also difficult to efficiently and comprehensively compare the regression results of multiple different scenarios to evaluate whether one scenario represents an overall better performance than another.

A common finding to both of these hindcast experiments is that global performance of a variable is often substantially better than the performance in individual regions.Therefore, while this work will explore global aggregates as previous analyses did, we find that global aggregates alone are not sufficient to evaluate IAMs that require global supply to equal global demand at each time period. GCAM is only one example of such an IAM.

The analysis scheme outlined below is designed with the four evaluation goals in mind and focuses on deviation based measures of model performance and the extent of conclusions that may be drawn from them. While many other model performance statistics exist, many operate on a pass/fail basis and therefore provide little insight about the reasons a model may fail. The scheme is then used to re-examine the land use data from Paper 1 to demonstrate application of the evaluation method and the resulting expanded results relative to application of skill scores.

## 2 Evaluation methods

A proposed scheme to meet the four evaluation goals inspired by past IAM hindcasting experiments is outlined below. This work explores the extent of conclusions that may be drawn from the root mean square error (RMSE) measure of model performance and finds that different uses of RMSE allow the possibility of addressing all four evaluation goals. While arguments against RMSE in favor of mean absolute error (MAE) exist (Legates and McCabe, 1999; Willmott and Matsuura, 2005), RMSE is chosen because it can be decomposed into errors from different sources (Murphy, 1988; Weglarczyk, 1998; Taylor, 2001). If only a single deviation measure were being examined, the types of conclusions that could be drawn would not differ appreciably whether RMSE or MAE is used. However the ability to decompose RMSE provides unique opportunities to understand different aspects of simulation performance.

Indices of agreement are popular in the literature and generally involve the comparison of a deviation measure between simulated and observed time series with some reference measure (Nash and Sutcliffe, 1970; Garrick et al., 1978; Willmott, 1981; Legates and McCabe, 1999; Willmott et al., 2012). Common reference measures include deviation measures between the observed data points and the mean of observations, or deviation measures between the observed data points and a baseline or naive model of the variable being simulated. Consistent with the idea of examining different reference measures, we normalize the root mean square error in different ways to capture different facets of model performance. Other members of the geoscientific modeling community are also moving to assess model performance with multiple normalized statistics, although we differ in specific techniques (Luo et al., 2012). These indices of agreement are particularly useful for evaluating model scenario performance in absolute terms due to the informative performance benchmarks outlined in Section 2.3. Other goodness-of-fit statistics such as correlation or a reduced chi-squared statistic were not chosen because they offer less information to guide improvements when a model displays poor performance.

### 2.1 Background: root mean square error decomposition

In the statistics outlined below, the value of variable $i$ in region $j$ at timestep $t$ is denoted by $s_t^{ij}$ for simulation and $o_t^{ij}$ for observation. Each time series contains $N$ discrete time points. The deviation measure of error chosen for model evaluation is the root mean square error, denoted for variable $i$ in region $j$ by

$$e_{ij} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(s_t^{ij} - o_t^{ij})^2}. \tag{1}$$

Root mean square error is the total deviation error in the model, decomposed as follows:

$$e_{ij}^2 = b_{ij}^2 + v_{ij}^2, \tag{2}$$

where $b_{ij}$ represents bias and $v_{ij}$ represents errors due to variability. Bias of variable $i$ in region $j$ is given by

$$b_{ij} = \overline{s^{ij}} - \overline{o^{ij}}, \tag{3}$$

where $\overline{s^{ij}}$ is the mean of the simulated time series and $\overline{o^{ij}}$ is the mean of the observed time series. The errors due to variability are those remaining after bias is accounted for by subtracting the means of the simulation and observation. The centered root mean square error quantifies this error and is denoted by

$$v_{ij} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left[(s_t^{ij} - \overline{s^{ij}}) - (o_t^{ij} - \overline{o^{ij}})\right]^2}. \tag{4}$$

### 2.2 Metrics for model evaluation

Past hindcast experiments in Integrated Assessment Models have implied that errors across regions cancel, leading to better performance at the global level than in most regions (Calvin et al., 2017; Fujimori et al., 2016). We define the time series for the global region, $G$, by concatenating the time series for each individual region. Therefore, for $J$ total regions whose time series each contain $N$ data points, the global time series contains $JN$ data points. To quantify the extent to which cancellation across regions occurs, bias is examined at the global level in two ways. First, the bias for the global region is examined, noting that it is mathematically equivalent to averaging the individual region biases:

$$b_{iG} = \overline{s^{iG}} - \overline{o^{iG}} = \frac{1}{J}\sum_{j=1}^{J} b_{ij}. \tag{5}$$

Second, we define global absolute bias as:

$$|b_{iG}| = \frac{1}{J}\sum_{j=1}^{J} |b_{ij}|. \tag{6}$$

By comparing the magnitudes of equations 5 and 6, the extent of cancellation occurring across regions may be quantified for each variable $i$.

At the regional level, normalization provides context for interpreting the errors in Sect. 2.1. The conventional normalization of root mean square uses the standard deviation of the observed time series, $\sigma_o^{ij}$. Normalized RMSE of variable $i$ in region $j$ is given by

$$e'_{ij} = \frac{e_{ij}}{\sigma_o^{ij}}. \tag{7}$$

$e'_{ij}$ gives a dimensionless measure: total error as a fraction of the standard deviation of observation of variable $i$ in region $j$. Similarly, the centered RMSE may be normalized by the standard deviation of observation, to give the errors due to variability as a fraction of the observed standard deviation. Normalized centered RMSE of variable $i$ in region $j$ is given by

$$v'_{ij} = \frac{v_{ij}}{\sigma_o^{ij}}. \tag{8}$$

The normalization used in equations 7 and 8 compares deviation measures to the observed variance about the temporal mean. However, that variance encompasses the trend line behavior. Therefore, we also normalize RMSE for variable $i$ in region $j$ by the observed variance about the trend line, following the convention of comparing deviation measures to a selected baseline to provide more targeted information about model performance (Garrick et al., 1978; Willmott, 1984; Legates and McCabe, 1999).

For each variable $i$ in each region $j$, let $\hat{y}(t)$ be the trend line fitted to the observational data, with $\hat{y}_t$ the values at the discrete time steps considered. Then we define the standard deviation of observation about the trend line as

$$\hat{\sigma}_o^{ij} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left[(o_t^{ij} - \hat{y}_t) - (\overline{o_t^{ij} - \hat{y}_t})\right]^2} \tag{9}$$

For the true trend line, $\hat{y}(t)$, the mean $\overline{o_t^{ij} - \hat{y}_t} = 0$. However, in numerically fitting the trend line, the mean is often not precisely 0. We can then define revised normalized RMSE by normalizing with the standard deviation about the trend line rather than about the time mean as follows:

$$\hat{e}_{ij} = \frac{e_{ij}}{\hat{\sigma}_o^{ij}} \tag{10}$$

One advantage of this refined measure is that $\hat{e}_{ij}$ penalizes poor simulation of the observed trend line more heavily than $e'_{ij}$. Another advantage is that, if the trend line is believed to be true to reality, the variance about the trend line will encapsulate natural variations (such as those due to weather) as well as observational uncertainty.

For the GCAM land use case study defined in Sect. 3.1, FAO observational data for each crop-region combination was individually detrended using the function loess.as from the R package fANCOVA (Wang, 2010) to fit the LOESS trend line, selecting the bias-corrected Akaike information criterion (AICC) method for generating the span parameter (Hurvich et al., 1998).

### 2.3 Informative performance benchmarks

While the time series statistics outlined in Section 2.1 have clear values corresponding to perfect model performance (i.e. a value of 0), specific criteria for acceptable and good model performance are more difficult to define objectively. In this section, we outline ways in which to contextualize the values achieved by each statistic outlined above to identify opportunities for model improvement.

For $e'_{ij}$ and $e_{ij}$, a helpful performance benchmark is defined as

$$e'_{ij} = \frac{e_{ij}}{\sigma_o^{ij}} < 1 \iff e_{ij} < \sigma_o^{ij} \tag{11}$$

Recall that the definition of standard deviation is $\sigma_o^{ij} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(o_t^{ij} - \overline{o^{ij}})^2}$. The right hand side of this equation is also what the root mean square error would be for a model taking $s_t^{ij} = \overline{o^{ij}}$ at each time step $t$. Satisfying Eq. (11) gives some sense of whether total error is small enough without achieving a perfect value of 0. It is popular to say that if $e'_{ij} > 1$, using the mean of the observed time series as a model leads to better performance than the current model. This interpretation is identical to that of the Nash-Sutcliffe Efficiency (Nash and Sutcliffe, 1970; Garrick et al., 1978; Legates and McCabe, 1999). However, for a nonstationary distribution of observations, the observed mean can only be calculated after the simulation period and therefore cannot be used as a model. When $e'_{ij} > 1$, either the bias or the variability component of RMSE (or both) is too large. Therefore, when $e'_{ij} > 1$, it is most useful to examine if $v'_{ij} < 1$. In this case, improving bias may allow the model to satisfy Eq. (11).

# 3 A case study of GCAM 3.0 land allocation

The data described below and analyzed in Section 3.2 is from the first GCAM land use system hindcast experiment, Paper 1. The land allocation data is re-analyzed using the method outlined in Table 1 in order to determine whether this method is more likely to achieve our four goals than the skill scores originally used. This demonstration is why we have chosen to re-evaluate existing experiments rather than repeat or develop new experiments in a more up to date version of GCAM. The full complement of resulting statistics and figures are available online with code and data, see Section 5.

## 3.1 GCAM background and data for re-analysis

GCAM is an Integrated Assessment Model capturing the interactions between human and earth systems.[1] GCAM includes energy, economic, and land use sectors that interact with each other and with a climate model. It is designed for long term forecasting and is typically operated in five year timesteps. Model behavior is calibrated to a historical base year using observational data, and forecasts evolve in time from the base year. Therefore, social, economic, and environmental policies in place during the base year are implicitly reflected in GCAM's performance. Policies that begin later, or change over time, must be more thoughtfully included, often explicitly.

Full details of the GCAM land use system, including equations, are provided in Wise et al. (2014) as well as in the online documentation[1]. Full details of different aspects of GCAM's structure are published in a variety of papers (Calvin et al., 2011; Kim et al., 2006; Clarke et al., 2007; Kyle et al., 2011). Briefly, the land use system of GCAM has a nested structure. In each sub-region within a geopolitical region, a nested structure is implemented with data specific to the sub-region. The land

[1]Documentation available at http://jgcri.github.io/gcam-doc/.

allocation choice at each branch in the nest is parameterized to reflect that sub-region's characteristics and may vary in response to economic, policy, and technological changes.

Economic agents in each sub-region operate to maximize the difference between revenue (including any taxes and subsidies) and the cost of production. The land use system assumes a distribution of costs, where the amount of land allocated for each use is actually the probability that land type is most profitable within its nest and avoiding winner-take-all behavior. That is, land is allocated to various possible uses via a logit distribution function at each branch of the nest. All references to GCAM within this work may be assumed to refer to GCAM version 3.0, unless otherwise specified.

Historical data prior to 1990 was used to calibrate GCAM 3.0, and then GCAM was run for a period from 1990 to 2010 without using additional historical data (i.e., GCAM is used to forecast agricultural land use from 1990 to 2010). There are nine GCAM crops (of 12) with historical data reported by the United Nations Food and Agricultural Organization (FAO) (FAO, 2014) during the period 1990 to 2010. The same analysis scheme outlined in Section 2 and demonstrated here could just as easily be used to examine any variable output by an IAM with historical data available for validation.

The reference set up of GCAM 3.0 (and all subsequent versions to date) for forecast into the 21st century uses smoothed FAO projections of yields as exogenous yield input information that is used by GCAM to simulate land allocation. The smoothing is performed as a five year rolling average including past and future years (i.e. the smoothed 2040 data point is generated as the average of data from 2038-2042).

Because GCAM requires global supply to equal global demand to solve for market prices at each time step, it is possible for GCAM economic agents are implicitly optimizing land allocation to meet global demand at minimum cost, even though GCAM is a dynamic recursive rather than an optimization model. When the economic agents are given unrealistic fore-knowledge of the impacts of weather events, for example, this implicit optimization may become particularly problematic. GCAM endogenously calculates a global market price (where global supply equals global demand) during the simulation period. This global market price is used to set producer prices used by economic agents in profit calculations underlying land allocation decisions, and every land use region shares the same global producer price. A global market price is needed for model calibration in the base year. Since such data does not currently exist, the USA producer price is used as the global price for calibration. This choice could lead to incorrectly incorporating or missing impacts of policies like subsidies or crop insurance programs. On the demand side, the price is sterilized in the GCAM calibration procedure.

Paper 1 featured experiments designed to investigate the possibility of unrealistic implicit optimization and examined two extremes of exogenous yield inputs via different parameterizations. The extremes also emphasize different aspects of the GCAM reference set up, and so the reference setup behavior is assumed to lie between the behaviors of the two extremes. The first extreme features increased variability in exogenous yield inputs compared to the GCAM reference. This is referred to as the Actual Yield case: GCAM makes planting decisions (allocates land) in 2005 based on knowing what the yield at the end of the year in 2005 will be, a case of economic agents having unrealistic levels of information for making planting decisions. There is no smoothing at all, and there is no explicit memory of past years' performance. The other extreme features a lack of variability and no updates to exogenous yield inputs during the simulation period 1990-2010, as opposed to the reference set up. This is referred to as the Forecast Yield case: a linear regression is fit to the historical yields over 1961-1990 and extrapolated linearly for the simulation period 1990-2010. There is no variation about this linear trend and economic agents have no fore-knowledge, contrasting the Actual Yield case.

To examine the impact of missing or incorrectly characterizing a policy, Paper 1 examined the US Renewable Fuel Standards implemented in 2005. The standards, among other things, increased demand for corn. GCAM runs without any implementation of the policy were compared with GCAM runs in which the increased demand for corn was explicitly included. Future scenarios interested in deeper analysis of the impacts of the US Renewable Fuel Standards may use a more detailed implementation or may make use of the metrics outlined in Section 2 to perform a Monte Carlo style parameter estimation for parameters related to the fuel standards.

These considerations result in the following four test cases (scenarios) examined in Paper 1:

- GCAM makes annual land allocations given data for population, income, and actual crop yields (denoted AY);
- GCAM makes annual land allocations given data for population, income, actual crop yields, and includes an estimate of the additional demand for corn resulting from the implementation of the U.S. Renewable Fuel Standards (denoted AYB);
- GCAM makes annual land allocations given data for population and income, but crop yields are forecasted based on an annual time trend for the years 1961 to 1990 (denoted FY);
- GCAM makes annual land allocations given data for population and income, crop yields are forecasted based on an annual time trend for the years 1961 to 1990, and includes an estimate of the additional demand for corn resulting from the implementation of the U.S. Renewable Fuel Standards (denoted FYB).

The simulated regional data in each of these four scenarios is compared to data reported by the FAO (FAO, 2014) during the period 1990 to 2010 for the nine GCAM crops with FAO data available. Calvin et al. found that the case FYB performed as well or better than the other scenarios across the skill scores considered: Reichler-Kim (Reichler and Kim, 2008), Normalized Mean Absolute Error (Schwalm et al., 2010; Luo et al., 2012), and Taylor Skill (Schwalm et al., 2010; Luo et al., 2012). Scenarios AY and AYB generally performed the worst.

### 3.2 Results

A selection of results demonstrating how the evaluation method summarized in Table 1 can be used to analyze multiple aspects of model performance at multiple scales and how the metrics may be used to make the analysis of the large amounts of data produced by IAMs more tractable are presented. The results presented were chosen both to illustrate the general types of insights that may be drawn from application of the evaluation scheme and to highlight the GCAM areas of strong performance and weak performance, with the full results for all variables at all scales by all metrics lying somewhere in between the results presented in this section. Each metric in Table 1 is used to re-examine the Paper 1 data, demonstrating the interactive and complementary nature of the metrics selected. With this approach, we are able to verify and expand the previous GCAM land hindcast results arrived at using skill scores in Paper 1. The analysis scheme does appear more capable of achieving all four evaluation goals than the skill scores. The full complement of resulting statistics and figures are available online with code and data, see Section 5 for details.

Figure 1 shows the global bias, Eq. (5), which is equivalent to the average of each individual region's bias. Because it is a signed quantity, a black circle is included at $b_{i,G} = 0$ for visual reference. Each scenario models global supply well for each crop with observational data available, as measured by global bias $b_{iG}$. The primary exceptions are that the scenarios AY (red) and AYB (green) model MiscCrop and OtherGrain poorly. This is not surprising, given that each of those crops is an aggregate of a large number of real world crops, varying across regions.

Figure 2 shows the global absolute bias, Eq. (6). For each crop, the magnitude of the global absolute bias in Fig. 2 is larger than the magnitude of the global bias in Fig. 1, indicating that errors are canceling across regions. Because there are no regional constraints on supply to supplement the requirement that global supply equal global demand, there are numerous regional supply solutions that may satisfy the global constraint. This provides ample opportunity for error cancellation across regions in any Integrated Assessment Model with a similar global constraint.

The FYB scenario (purple) displays the smallest absolute bias for all crops, with the exception of Rice and OtherGrain, in Fig. 2. In other words, the FYB scenario is most successful at modeling global supply when cancellation across regions is prohibited.

The compensating errors across regions can be further studied by examining the normalized RMSE, $e'_{ij}$, Eq. (7), for a single crop. Fig. 3 displays the individual regional errors for Wheat. A black circle is included to denote the performance benchmark $e'_{ij} = 1$, Eq. (11). With the exception of Southeast Asia, the forecast yield scenarios (FY, blue, and FYB, purple) outperform the scenarios using actual yield information (AY, red, and AYB, green). Scenarios FY and FYB show that compensating performance is occurring: the good performance in Canada, Eastern Europe, and USA is balanced by the poorer performance in Australia New Zealand, India, Latin America, and Southeast Asia. Similar trends hold when examining other crops.

To further understand the role of compensating errors in GCAM land allocation, the role of bias as a contributing factor is examined. Because root mean square error decomposes into bias and centered root mean square error, Eq. (2), a sense of whether bias is too large can be gained from comparing $e'_{ij}$, Eq. (7) and $v'_{ij}$, Eq. (8). If $e'_{ij} > 1$ and $v'_{ij} < 1$, bias may be considered a problematic source of errors. This is generally what occurs in GCAM.

Figure 4 displays the normalized RMSE, $e'_{ij}$, for each crop in the United States. A black circle is included for $e'_{ij} = 1$. In the FYB scenario (purple), $e'_{ij} > 1$ for every crop except Wheat.

Figure 5 displays the normalized centered RMSE, $v'_{ij}$, for each crop in the United States. A black circle is included for $v'_{ij} = 1$.

The FYB scenario (purple) displays $v'_{ij} < 1$ for all crops except Rice and Root Tuber. Compared with the larger values of $e'_{ij}$ in Fig. 4, this indicates that bias is a major contributing factor to performance issues. This general trend - that scenario FYB performs best and that bias is the major contributor to model performance issues for most crops - holds across regions.

It would be preferential for the bias to be improved intrinsically through structural or parametric model changes, rather than through bias correction techniques. Therefore, we examine which factors contribute to bias. The revised normalized RMSE, $\hat{e}_{ij}$, Eq. (10), compares GCAM performance to variations of the observed time series about the trend line. Figure 6 displays this metric for each crop in the USA. A black circle is included for $\hat{e}_{ij} = 1$. Each crop in each scenario misses the trend line behavior. With the exception of Rice, scenario FYB (purple) comes closest to capturing the trend line behavior. This result holds for most crops in most regions. Therefore, scenario FYB is one possible starting place in making structural improvements to GCAM.

To further examine the ways in which simulations may improve at capturing trend lines, time series for Corn (left) and Wheat (right) for multiple regions are depicted in Fig. 7. The black curves are FAO observational data for land allocation in each region, and the colored time series correspond to the different GCAM scenarios.

The time series for both Corn and Wheat illustrate a key issue: GCAM tends to incorrectly simulate whether land allocation should increase or decrease in time. The FYB scenario for Wheat (Fig. 7, right) tends to be the most accurate, consistent with the results depicted in Fig. 6. It is of note that the actual yield scenarios (AY, red, and AYB, green) are also susceptible to inaccurate discrimination between increasing and decreasing land allocation, showing that it is not improved by economic agents in GCAM having perfect information about year end yields to make planting decisions.

One possibility for the incorrect direction of simulated trends is that the parameters involved in the land allocation decision may be improved, by changing the calibration process and/or by using parameter estimation to adjust the logit exponents governing competition. Another option may be to explore the impacts of using different distributions to govern competition.

That the AY (red) scenario displays different performance than the AYB (green) scenario reinforces the importance of careful implementation of policies: explicitly including the effects of policies (such as in AYB) leads to different performance than assuming policies are implicitly included in the information provided to the model (as in AY, a case where real world yields that should implicitly reflect the increased demand due to the US Renewable Fuel Standards).

Finally, the time series for Corn in the Former Soviet Union and Wheat in China both suggest an opportunity for structural changes to improve the land allocation performance of GCAM. The yields for both of these crops display different slopes during the simulation period than the historical period. Therefore, the extension of the historical yield trends used in the FY and FYB scenarios has no hope of correctly capturing the different yield behavior during the simulation period. In turn, GCAM has no hope of capturing the different land allocation decisions in response to those yield changes. In contrast to the FY and FYB scenarios, the AY and AYB scenarios lead to GCAM's land allocation being very responsive to variability in yield inputs. One hypothesis is that this is because the economic agents in GCAM have unrealistic access to year end harvest amounts when making their planting decisions. This local yield input information may allow GCAM to meet global demand without matching historical data due to the lack of regional supply constraints.

### 3.3 GCAM-specific conclusions

Using the evaluation method outlined in Table 1, we expand the results presented in Paper 1. Like many IAMs, GCAM requires that global supply equal global demand for each commodity in each time period. The FYB scenario in GCAM models global supply well, as measured by global bias $b_{iG}$, Fig. 1. GCAM, at least, has no regional constraints on supply to supplement the global supply and demand constraint. As a result, there are numerous regional supply solutions that may satisfy the global constraint. This provides ample opportunity for error cancellation across regions, demonstrated in Fig. 3.

We find that the main opportunity to improve land allocation decisions in GCAM is to make structural and parametric changes to improve the trend line for each simulated time series and therefore improve bias. The scenario using yields forecasted from the historical period and modeling the U.S. Renewable Fuel Standards (scenario FYB) generally performs the best across all metrics and is the most reasonable starting point to begin model improvements. Specifically, updating the yield forecast as new information becomes available each year in the simulation period would allow the yield to capture changes occurring during the simulation period while avoiding the over-responsiveness of the scenarios using actual yields as inputs (scenarios AY and AYB). Changes to parameters, calibration methods, and data sources for producer prices may also improve the land use system's ability to discern whether land allocation trend lines should increase or decrease in time for a given crop-region combination. The metrics in Table 1 may be used for parameter estimation studies. In using GCAM to forecast into the future (where an AY scenario is not possible), providing the ability to adapt to shifts in yield occurring during a simulation period and the ability to better predict whether a land allocation trend line should increase or decrease in response to a yield shift would both be improvements.

Because the GCAM reference exogenous yield inputs lie between the two extremes examined in Paper 1 and here, one expects a hindcast experiment with the reference set up to have errors between those of the AY and FY cases. However, because the reference scenario has exogenous yield inputs based on FAO *forecasts* of yields, it is possible that the reference scenario may perform substantially worse than any of the cases examined in this work. This could occur if FAO forecasts of yields are dramatically inaccurate. Because planting decisions are not subject to the kind of vintaging seen with power plant construction, it is unlikely that errors will compound in an unexpected ways. A planting decision (in GCAM) only lasts for the year in which it occurs. A power plant construction lasts for 30+ years. This lack of vinatging makes it simpler to evaluate the land sector than other sectors of GCAM. Therefore, while the evaluation method outlined in this work can still be applied to sectors that feature vintaging, the results must be interpreted much more carefully. It's possible that additional metrics may have to be implemented for sectors with vintaging, and rigorous studies designed to specifically test the extent to which vintaging causes errors to compound may be undertaken in the future.

## 4 Conclusions

Examination of past hindcasting exercises in the IAM community has suggested that global aggregate metrics are often not well-suited to evaluating IAM hindcast performance. This work has outlined a suite of metrics designed to counteract this problem, and has demonstrated that the family of metrics presented is able to provide richer insight into model performance than global skill scores by re-evaluating the results of a past hindcast experiment in GCAM.

Further, applying the evaluation method outlined in Table 1 allows insight into evaluating IAMs beyond GCAM. While global results in GCAM are largely consistent with observations, cancellation of errors is present at the global level, a finding implied by previous hindcasting work in two different IAMs (Calvin et al., 2017; Fujimori et al., 2016). Any IAM requiring globally balanced supply and demand without additional regional constraints will likely encounter this same issue. This suggests a larger challenge in evaluating Integrated Assessment Models: replicating global aggregates is a necessary but in no way sufficient constraint on model performance. Indeed, many IAMs force global supply to equal global demand, and so global aggregates of many variables in IAMs simply reflect this forced behavior. Therefore, a family of validating metrics is found to be necessary in evaluation of IAM hindcast experiments. The option to evaluate results both relatively and absolutely should lead to more robust model improvements in the future by identifying the best performing scenarios for a single model, as well as aid the IAM community in conducting hindcast intercomparison studies.

A sector by sector application of a family of metrics may be necessary for evaluation of an IAM hindcast experiment as a whole. Future research into more tractable methods for simultaneous evaluation of all IAM sectors without masking deficiencies as global aggregates do is necessary to determine if this is the case. Such work is complicated by the lack of historical data against which to validate many IAM variables. Additionally, one may question whether the observational data being used for validation is reliable. Collecting global economic data is difficult and there is no opportunity for repeated measurements to obtain measurement uncertainty. When fitting trend lines to the FAO data for use in the revised normalized RMSE metric, $\hat{e}_{ij}$, Eq. (10), it became clear that in at least some regions the data may not be a reflection of reality. Namely, for some crops in Korea and Japan (among other regions), there is almost no variation about the trend line. There also was no available FAO data to validate three crops and other land types modeled by GCAM. Therefore, a better sense of observational uncertainty is necessary before parameter estimation based on observational data can take place.

## 5 Data and code availability

The data analyzed in this work is publicly available at https://github.com/JGCRI/LandHindcastPaper. This repository includes all input data, the R scripts for calculating all statistics and the results of those calculations, and the R scripts for generating all plots of statistics and the resulting plots.

Results from GCAM 3.0 simulations were used in this work. All GCAM releases from 3.0 onward are available at: https://github.com/JGCRI/gcamcore/releases.

*Author contributions.* A.C. Snyder analyzed the data. A.C. Snyder, R.P. Link, and K.V. Calvin prepared the manuscript.

*Competing interests.* The authors declare that they have no conflict of interest.

*Acknowledgements.* This research was based on work supported by the U.S. Department of Energy (DOE), Office of Science, Biological and Environmental Research as part of the Integrated Assessment Research program. The Pacific Northwest National Laboratory is operated for DOE by Battelle Memorial Institute under contract DE-AC05-76RLO1830.

## References


Baldos, U. L. C. and Hertel, T. W.: Looking back to move forward on model validation: insights from a global model of agricultural land use, Environmental Research Letters, 8, 034 024, 2013.

Beckman, J., Hertel, T., and Tyner, W.: Validating energy-oriented CGE models, Energy Economics, 33, 799–806, 2011.

Calvin, K., Clarke, L., Edmonds, J., Eom, J., Hejazi, M., Kim, S., Kyle, P., Link, R., Luckow, P., Patel, P., et al.: GCAM wiki documentation, Pacific Northwest National Laboratory, 2011.

Calvin, K., Wise, M., Kyle, P., Clarke, L., and Edmonds, J.: A Hindcast Experiment Using the GCAM 3.0 Agriculture and Land-use Module, Climate Change Economics, 8, 1750 005, 2017.

Clarke, L., Lurz, J., Wise, M., Edmonds, J., Kim, S., Smith, S., and Pitcher, H.: Model documentation for the minicam climate change science program stabilization scenarios: Ccsp product 2.1 a, Pacific Northwest National Laboratory, PNNL-16735, 2007.

FAO: FAOSTAT, Food and Agriculture Organization of the United Nations, 2014.

Fujimori, S., Dai, H., Masui, T., and Matsuoka, Y.: Global energy model hindcasting, Energy, 114, 293–301, 2016.

Garrick, M., Cunnane, C., and Nash, J.: A criterion of efficiency for rainfall-runoff models, Journal of Hydrology, 36, 375–381, 1978.

Hurvich, C. M., Simonoff, J. S., and Tsai, C.-L.: Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 60, 271–293, 1998.

Kim, S. H., Edmonds, J., Lurz, J., Smith, S., and Wise, M.: The Object-oriented Energy Climate Technology Systems (ObjECTS) framework and hybrid modeling of transportation in the MiniCAM long-term, global integrated assessment model, Energy J, 27, 63–91, 2006.

Kriegler, E., Petermann, N., Krey, V., Schwanitz, V. J., Luderer, G., Ashina, S., Bosetti, V., Eom, J., Kitous, A., Méjean, A., et al.: Diagnostic indicators for integrated assessment models of climate policy, Technological Forecasting and Social Change, 90, 45–61, 2015.

Kyle, G. P., Luckow, P., Calvin, K. V., Emanuel, W. R., Nathan, M., and Zhou, Y.: GCAM 3.0 agriculture and land use: data sources and methods, Tech. rep., Pacific Northwest National Laboratory (PNNL), Richland, WA (US), 2011.

Legates, D. R. and McCabe, G. J.: Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation, Water resources research, 35, 233–241, 1999.

Luo, Y., Randerson, J. T., Abramowitz, G., Bacour, C., Blyth, E., Carvalhais, N., Ciais, P., Dalmonech, D., Fisher, J. B., Fisher, R., et al.: A framework for benchmarking land models, Biogeosciences, 9, 2012.

Murphy, A. H.: Skill scores based on the mean square error and their relationships to the correlation coefficient, Monthly weather review, 116, 2417–2424, 1988.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I—A discussion of principles, Journal of hydrology, 10, 282–290, 1970.

Parson, E. A. and Fisher-Vanden, K.: Integrated assessment models of global climate change, Annual Review of Energy and the Environment, 22, 589–628, 1997.

Parson, E. A., Burkett, V., Fisher-Vanden, K., Keith, D., Mearns, L., Pitcher, H., Rosenzweig, C., and Webster, M.: Global-change scenarios: their development and use, 2007.

Reichler, T. and Kim, J.: How well do coupled models simulate today's climate?, Bulletin of the American Meteorological Society, 89, 303, 2008.

Schwalm, C. R., Williams, C. A., Schaefer, K., Anderson, R., Arain, M. A., Baker, I., Barr, A., Black, T. A., Chen, G., Chen, J. M., et al.: A model-data intercomparison of CO2 exchange across North America: Results from the North American Carbon Program site synthesis, Journal of Geophysical Research: Biogeosciences, 115, 2010.

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, Journal of Geophysical Research: Atmospheres, 106, 7183–7192, 2001.

van Ruijven, B., de Vries, B., van Vuuren, D. P., and van der Sluijs, J. P.: A global model for residential energy use: uncertainty in calibration to regional data, Energy, 35, 269–282, 2010a.

van Ruijven, B., van der Sluijs, J. P., van Vuuren, D. P., Janssen, P., Heuberger, P. S., and de Vries, B.: Uncertainty from model calibration: applying a new method to transport energy demand modelling, Environmental modeling & assessment, 15, 175–188, 2010b.

Wang, X.: fANCOVA: Nonparametric Analysis of Covariance, R package version 0.5-1., http://CRAN.R-project.org/package=fANCOVA, 2010.

Weglarczyk, S.: The interdependence and applicability of some statistical quality measures for hydrological models, Journal of Hydrology, 206, 98–103, 1998.

Willmott, C. J.: On the validation of models, Physical geography, 2, 184–194, 1981.

Willmott, C. J.: On the evaluation of model performance in physical geography, in: Spatial statistics and models, pp. 443–460, Springer, 1984.

Willmott, C. J. and Matsuura, K.: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, Climate research, 30, 79–82, 2005.

Willmott, C. J., Robeson, S. M., and Matsuura, K.: A refined index of model performance, International Journal of Climatology, 32, 2088–2094, 2012.

Wise, M., Calvin, K., Kyle, P., Luckow, P., and Edmonds, J.: Economic and physical modeling of land use in GCAM 3.0 and an application to agricultural productivity, land, and terrestrial carbon, Climate Change Economics, 5, 1450 003, 2014.

**Table 1.** Statistics for model evaluation

| abbreviation: | description: | normalized by: | notes: |
|---|---|---|---|
| $b_{iG}$ | global bias | | lacks absolute performance info |
| $\|b_{iG}\|$ | global absolute bias | | lacks absolute performance info |
| $e'_{ij}$ | regional normalized RMSE | standard deviation around time mean of observation | |
| $v'_{ij}$ | regional normalized centered RMSE | standard deviation around time mean of observation | |
| $\hat{e}_{ij}$ | revised regional normalized RMSE | standard deviation around trend lline of observation | |

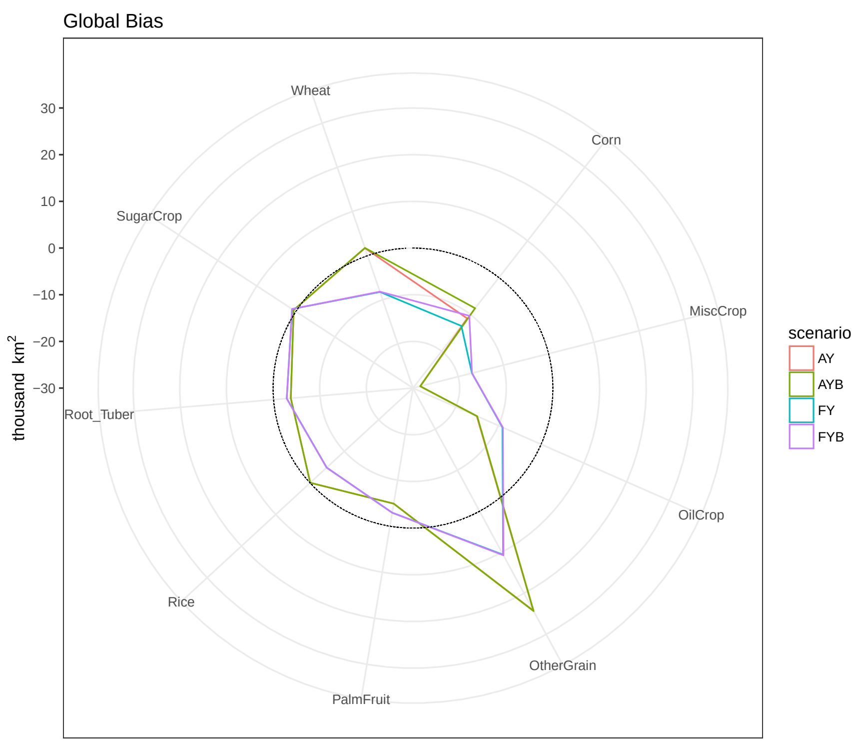


**Figure 1.** Global bias, $b_{iG}$, Eq. (5). The black circle corresponds to $b_{i,G} = 0$.


**Figure 2.** Global absolute bias, $|b_{iG}|$, Eq. (6).

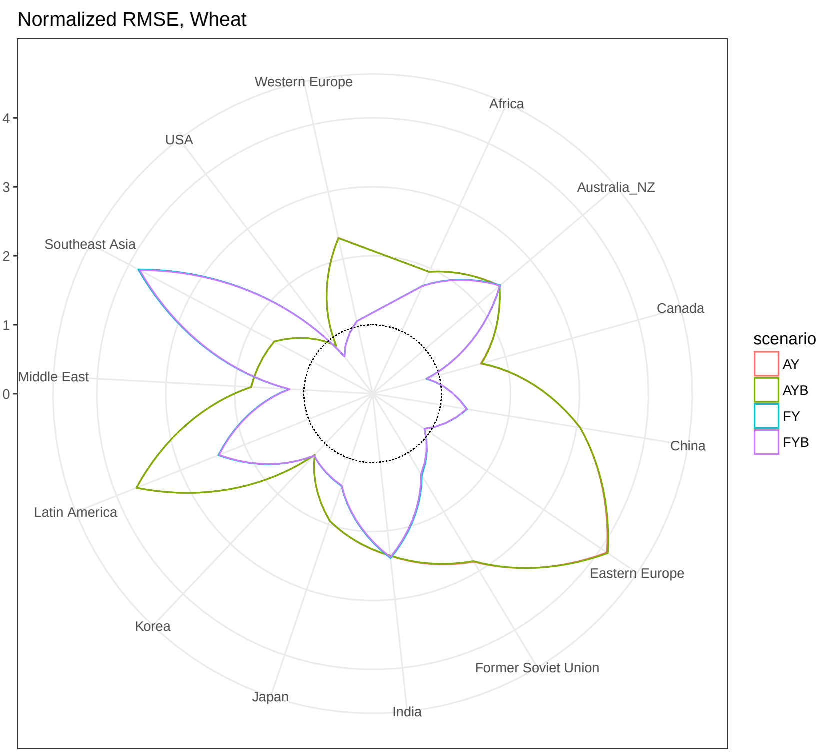


**Figure 3.** Normalized RMSE, $e'_{ij}$, Eq. (7), in each region for the land allocated to Wheat. The black circle is at the performance benchmark, $e'_{ij} = 1$, Eq. (11). $e'_{ij}$ compares RMSE error with the standard deviation of observation for each crop.

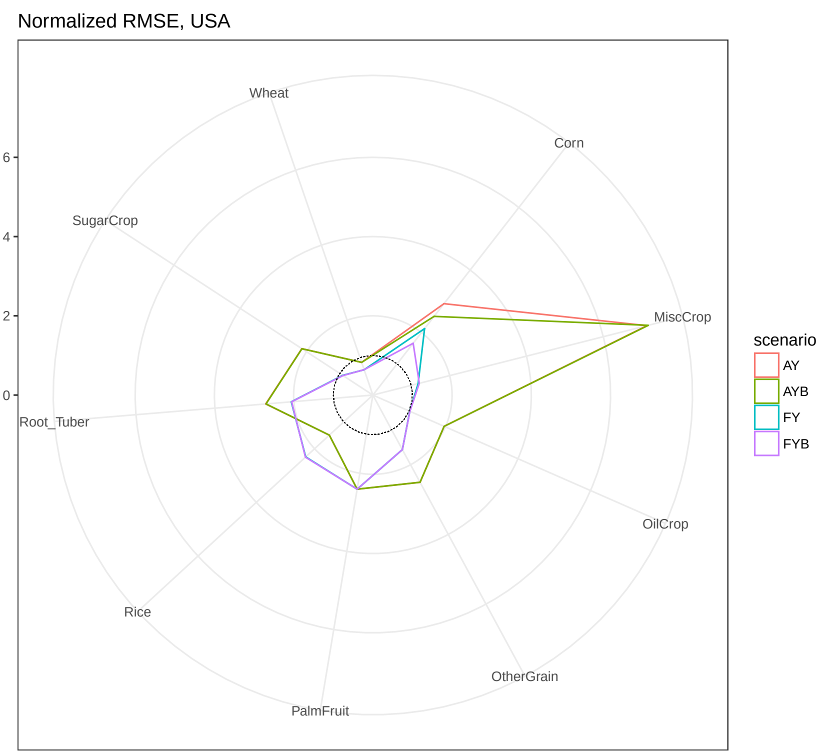


**Figure 4.** Normalized RMSE, $e'_{ij}$, Eq. (7), for each crop in the United States. A black circle is included for $e'_{ij} = 1$. $e'_{ij}$ compares RMSE error with the standard deviation of observation for each crop.

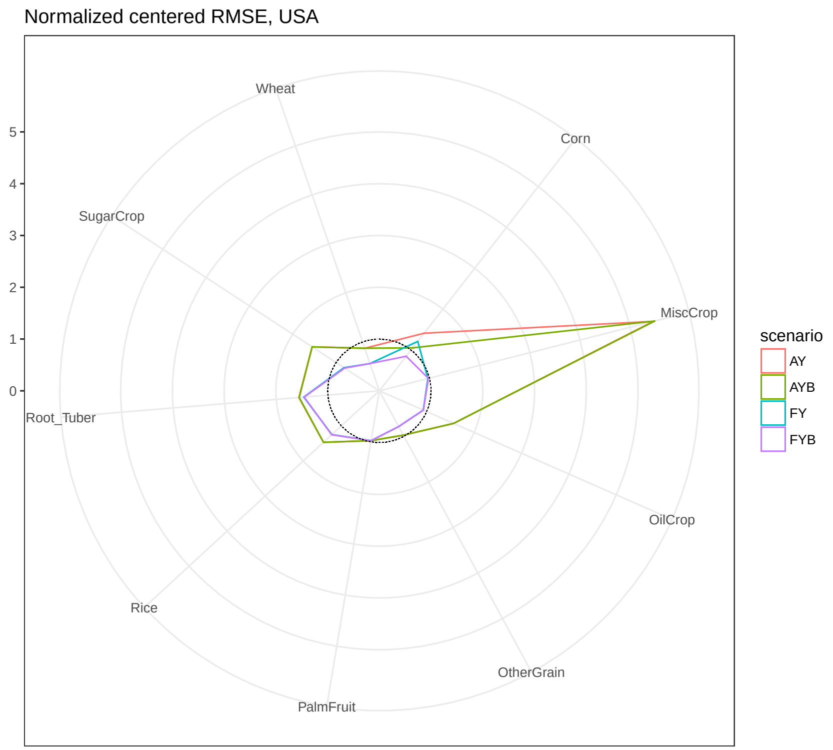


**Figure 5.** Normalized centered RMSE, $v'_{ij}$, Eq. (8), for each crop in the United States. A black circle is included for $v'_{ij} = 1$. $v'_{ij}$ compares centered RMSE error with the standard deviation of observation for each crop.


**Figure 6.** Revised normalized RMSE, $\hat{e}_{ij}$, Eq. (10), for each crop in the United States. A black circle is included for $\hat{e}_{ij} = 1$. $\hat{e}_{ij}$ compares RMSE error with the standard deviation about the observed trend line for each crop.

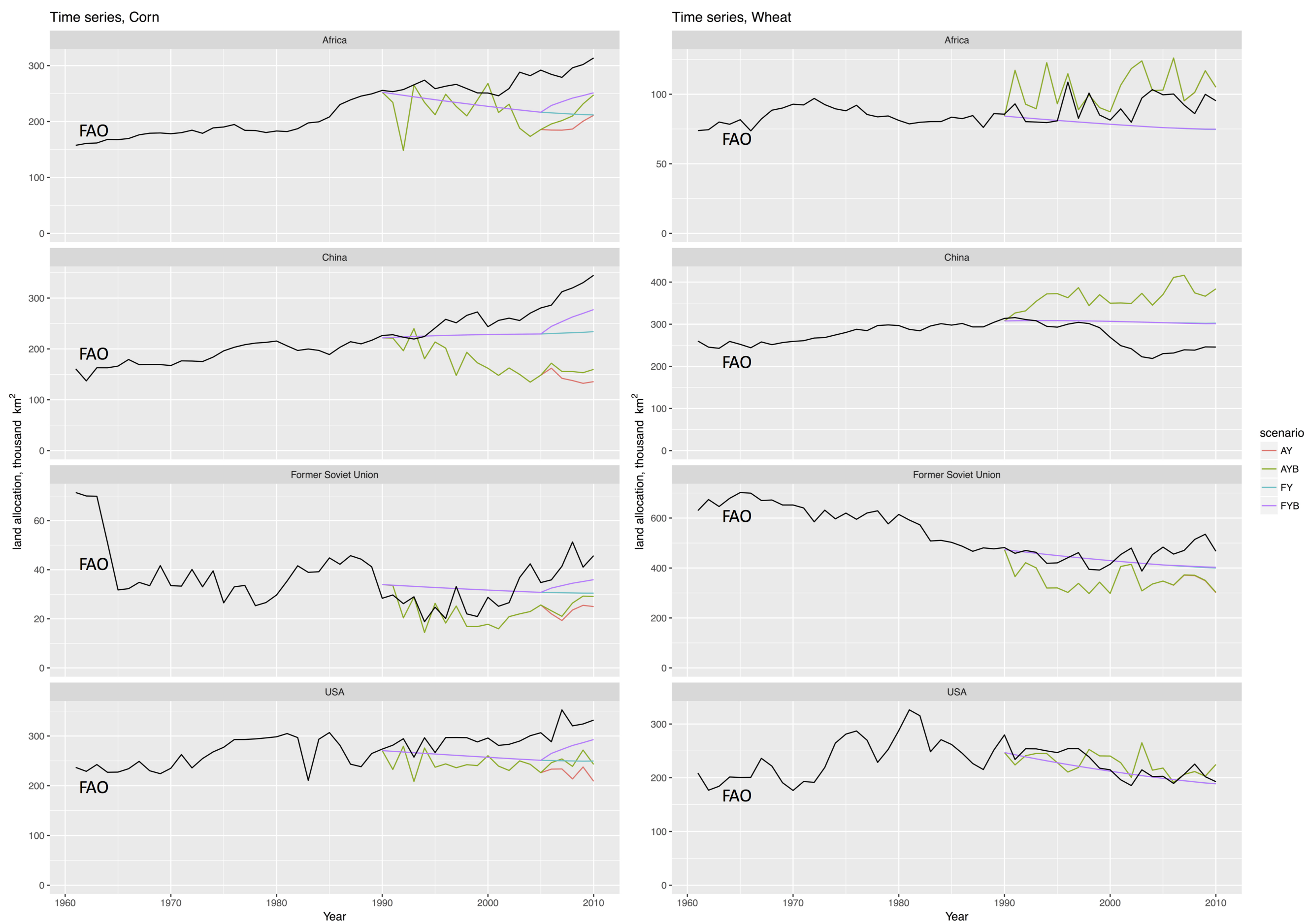


**Figure 7.** Time series for land allocated to Corn (left) and Wheat (right) in units of thousand km$^2$ across select regions. The black time series in each panel represents FAO observational data. The colored time series correspond to different GCAM scenarios.