Geosci. Model Dev. Discuss., doi:10.5194/gmd-2017-97, 2017
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 1 June 2017


Geoscientific Model Development Discussions

# 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 rarely undertaken in the Integrated Assessment Model (IAM) community. 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. This is key in the integrated assessment community, where there often are not multiple models conducting hindcast experiments to allow for model intercomparison. The performance benchmarks serve a second purpose, providing information about the reasons a model may perform poorly on a given measure and therefore identifying opportunities for improvement. As a case study, 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. We find quantitative evidence that global aggregates alone are not sufficient for evaluating IAMs, such as GCAM, that require global supply to equal global demand at each time period. Additionally, the deviation measures examined in this work successfully identity parametric and structural changes that may improve land allocation decisions in GCAM. Future work will involve implementing the suggested improvements to the GCAM land allocation system identified by the measures in this work, using the measures to quantify performance improvement due to these changes, and, ideally, applying these measures to other sectors of GCAM and other land allocation models.

## 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. While the IAM community regularly undertakes other model validation exercises, hindcast experiments are relatively new to the community. 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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2017-97, 2017
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 1 June 2017


Geoscientific Model Development Discussions

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.

The Global Change Assessment Model version 3.0 (GCAM) (Calvin et al., 2011; Kim et al., 2006; Clarke et al., 2007; Edmonds and Reiley, 1985; Kyle et al., 2011) was recently used to conduct a hindcast experiment (Calvin et al., 2017). Calvin et al. 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. Brief background on the land use system in GCAM is provided in Sect. 2 with more detailed information available in Wise et al. (2014) . The operating assumptions used are outlined in Sect. 4 and represent different levels of information for land allocation decisions available to the economic agents in the land use module. The highly aggregated nature that makes the skill scores examined convenient also limits the insight for model improvement that they can provide by masking important deficiencies.

A hindcast experiment was also performed for the energy component of the AIM/CGE model (Fujimori et al., 2016). 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. Unfortunately, neither of these techniques are compatible with our goals and methodology. 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.

This paper seeks to explore metrics that are as simple to implement as skill scores, but that provide more usable information for model improvement. The work outlined below 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 provide little insight about the reasons a model may fail. A case study is performed reexamining the land use results of the first GCAM hindcast experiment (Calvin et al., 2017). The methods developed here are generalizable, and could be applied to other sectors within GCAM, other IAMs, or potentially land use models from other communities.

## 2 GCAM background

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.

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 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.

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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2017-97, 2017
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 1 June 2017


Geoscientific Model Development Discussions

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. Additional details are available in Wise et al. (2014).

## 3 Methods

This work explores the extent of conclusions that may be drawn from the root mean square error (RMSE) measure of model performance. 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 (regardless whether RMSE or MAE), the types of conclusions that could be drawn would not differ appreciably with the specific measure chosen. 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). 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.

### 3.1 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

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

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

$$e_{ij}^2 = b_{ij}^2 + v_{ij}^2, \quad (2)$$

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

$$b_{ij} = \overline{s^{ij}} - \overline{o^{ij}}, \quad (3)$$

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2017-97, 2017
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 1 June 2017


Geoscientific Model Development Discussions

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

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

### 3.2 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. 3.1. 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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2017-97, 2017
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 1 June 2017


Geoscientific Model Development Discussions

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. 4, 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).

**Table 1.** Statistics for model evaluation

| abbreviation: | description: | normalized by: |
|---|---|---|
| $b_{iG}$ | global bias | |
| $\lvert b_{iG}\rvert$ | global absolute bias | |
| $e'_{ij}$ | 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 line of observation |

### 3.3 Informative performance benchmarks

While the time series statistics outlined in Sect. 3.1 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.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2017-97, 2017
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 1 June 2017


Geoscientific Model Development Discussions

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 equation 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 equation 11.

A similar benchmark for $\hat{e}_{ij}$ would be $\hat{e}_{ij} < 1 \iff e_{ij} < \hat{\sigma}_o^{ij}$; the total error must be less than the observed standard deviation about the trend line. $\hat{e}_{ij} > 1$ indicates the trend line of the simulated time series likely does not match the trend line of the observed time series.

## 4 Data

The data analyzed in Sect. 5 is from the first GCAM land use system hindcast experiment (Calvin et al., 2017). 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).[2] The following four test cases (scenarios) were performed:

- 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 United Nations Food and Agricultural Organization (FAO) (FAO, 2014) during the period 1990 to 2010 for the nine GCAM crops with corresponding

[2]GCAM 3.0 divides land into 14 geopolitical regions; GCAM 4.3 uses a finer division of 32 geopolitical regions.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2017-97, 2017
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 1 June 2017




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, due to economic agents' over responsiveness resulting from unrealistically high levels of information for decision making.

# 5 Results

The metrics outlined in Sect. 3.2 are used to analyze the GCAM land allocation output previously examined in the first hindcast experiment. With this approach, we are able to verify as well as expand the previous GCAM land hindcast results(Calvin et al., 2017).

## 5.1 Global performance

Figure 1 shows the global bias (equation 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. GCAM requires that global supply equal global demand for each commodity in order to solve at each timestep. 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.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2017-97, 2017
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 1 June 2017


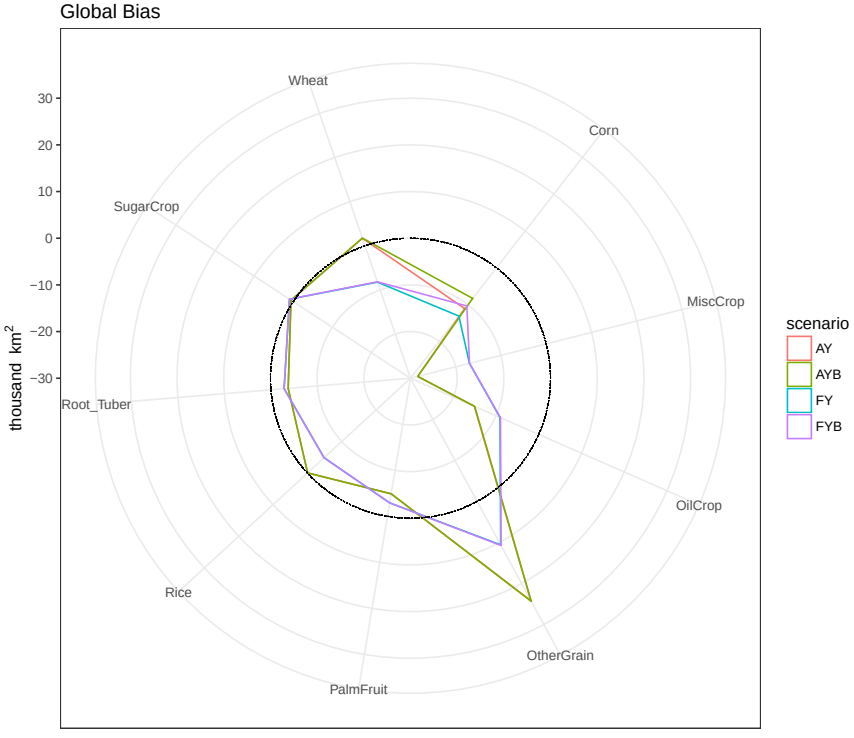


**Figure 1.** Global bias, $b_{iG}$ (equation 5). The black circle corresponds to $b_{i,G} = 0$.

Figure 2 shows the global absolute bias (equation 6). For each crop, the magnitude of the global absolute bias in Figure 2 is larger than the magnitude of the global bias in Figure 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 Figure 2. In other words, the FYB scenario is most successful at modeling global supply when cancellation across regions is prohibited.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2017-97, 2017
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 1 June 2017


Geoscientific Model Development Discussions
Open Access
EGU


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

The compensating errors across regions can be further studied by examining the normalized RMSE, $e'_{ij}$ (equation 7), for a single crop. Figure 3 displays the individual regional errors for Wheat. A black circle is included to denote the performance benchmark $e'_{ij} = 1$ (equation 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 in the next section.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2017-97, 2017
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 1 June 2017


Geoscientific Model Development Discussions
Open Access
EGU

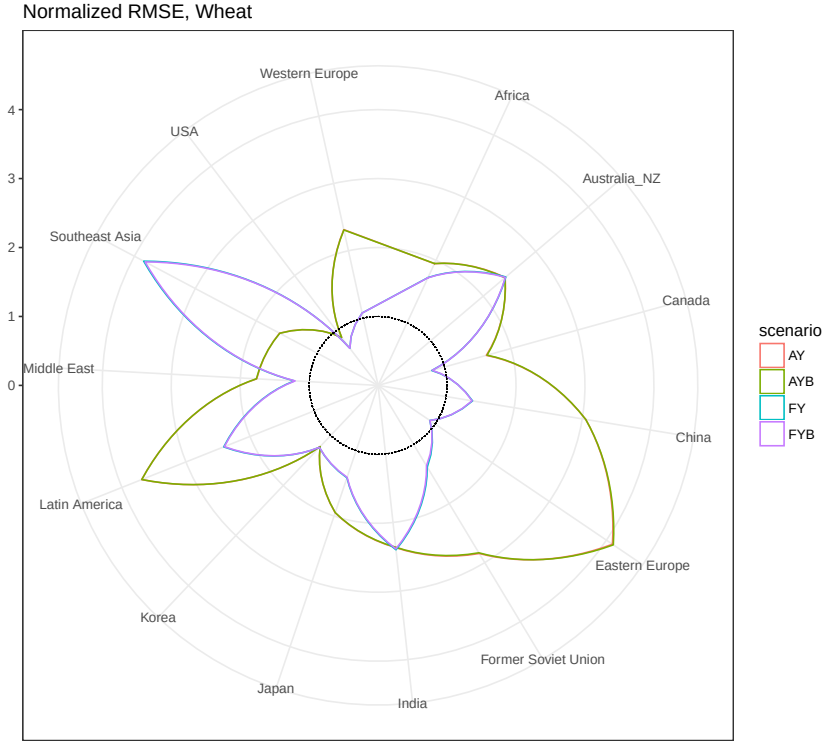


**Figure 3.** Normalized RMSE, $e'_{ij}$ (equation 7), in each region for the land allocated to Wheat. The black circle is at the performance benchmark, $e'_{ij} = 1$, (equation 11). $e'_{ij}$ compares RMSE error with the standard deviation of observation for each crop.

### 5.2 The role of bias

Because root mean square error decomposes into bias and centered root mean square error (equation 2), a sense of whether bias is too large can be gained from comparing $e'_{ij}$ (equation 7) and $v'_{ij}$ (equation 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.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2017-97, 2017
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 1 June 2017


Geoscientific Model Development Discussions
Open Access
EGU

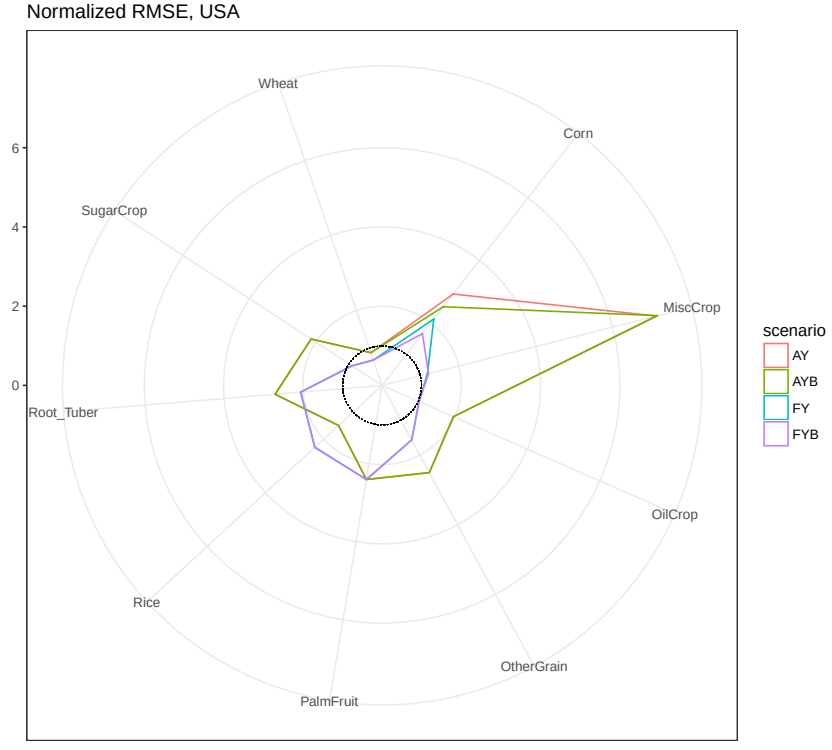


**Figure 4.** Normalized RMSE, $e'_{ij}$ (equation 7), for each crop in the United States. A black circle is included for $e'_{ij} = 1$. $e'_{ij}$ compares RMSE error with the standard deviation of observation for each crop.

Figure 5 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 Figure 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.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2017-97, 2017
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 1 June 2017


Geoscientific Model Development Discussions
Open Access
EGU

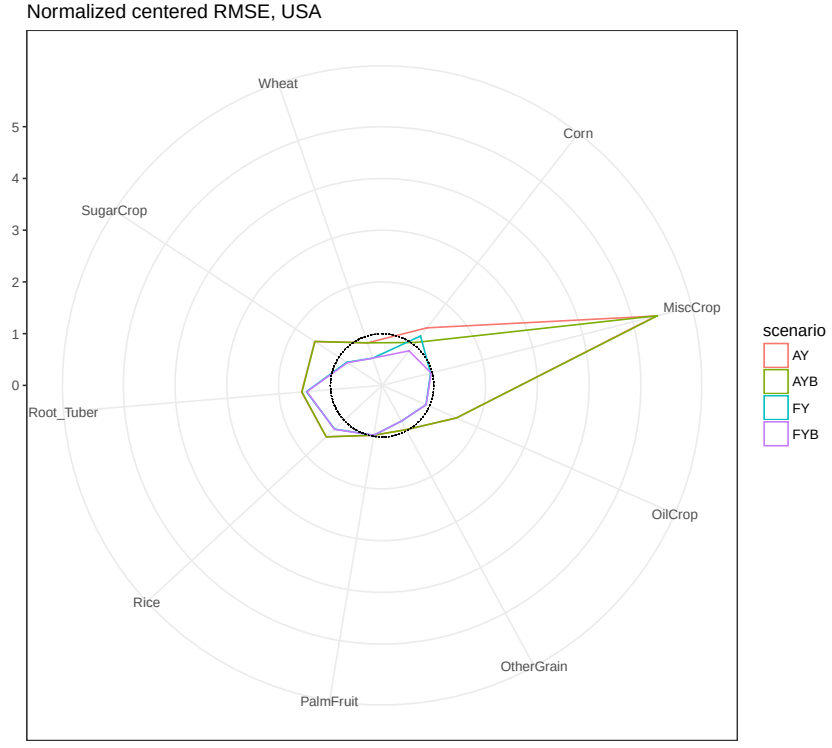


**Figure 5.** Normalized centered RMSE, $v'_{ij}$ (equation 8), for each crop in the United States. A black circle is included for $v'_{ij} = 1$. $v'_{ij}$ compares centered RMSE error with the standard deviation of observation for each crop.

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, $\hat{e}_{ij}$ (equation 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.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2017-97, 2017
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 1 June 2017


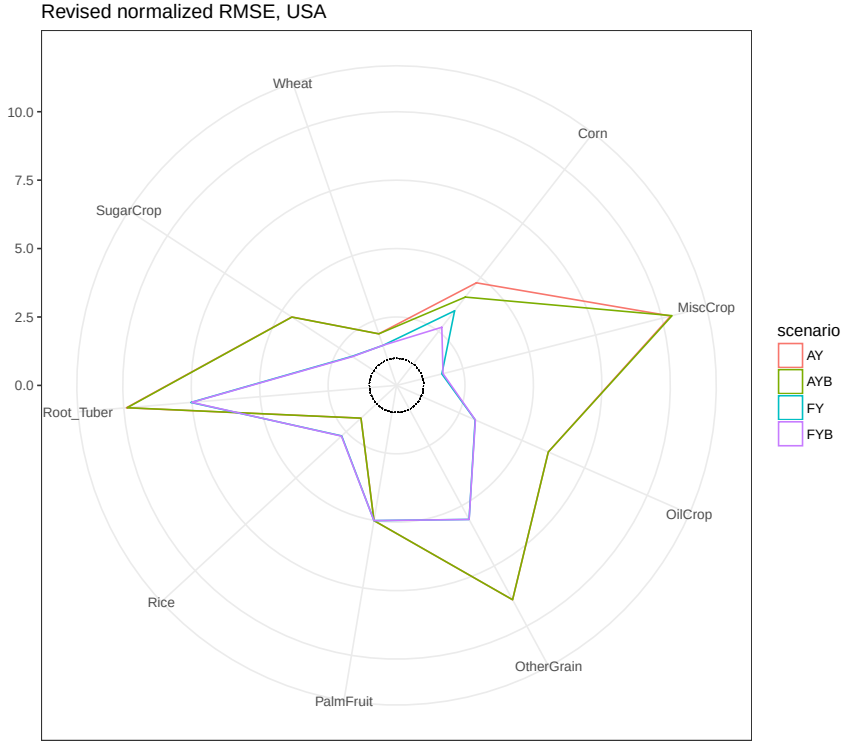


**Figure 6.** Revised normalized RMSE, $\hat{e}_{ij}$ (equation 10), for each crop in the United States. A black circle is included for $\hat{e}_{ij} = 1$. $\hat{e}_{ij}$ compares RMSE error with the standard deviation about the observed trend line for each crop.

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 Figure 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 (Figure 7, right) tends to be the most accurate, consistent with the results depicted in Figure 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 to make decisions. The economic agents in AY and AYB have access to year end yield information when making their land allocation decisions and still fail to match observation.

One possibility for the incorrect direction of simulated trends for 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 an absolute cost logit instead of the relative cost logit implemented here.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2017-97, 2017
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 1 June 2017




Additionally, every economic agent in GCAM uses USA producer prices in calculating land allocation for each crop. This is partly due to data availability, but could lead to incorrectly incorporating or missing impacts of policies like subsidies or crop insurance programs. That the AY (red) scenario displays different performance than the AYB (green) scenario reinforces this point: 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).

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 scenarios incorporating forecasted yields (FY and FYB) 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. At the other extreme, the scenarios using actual yield information (AY and AYB) lead to GCAM's land allocation being overly responsive, due to economic agents having more information than their real world counterparts. Therefore, an adaptive forecast, updating the forecast from the historical period with yields from each simulation year as the simulation progresses and weighting more recent observations more heavily, offers the best avenue for future testing.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2017-97, 2017
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 1 June 2017


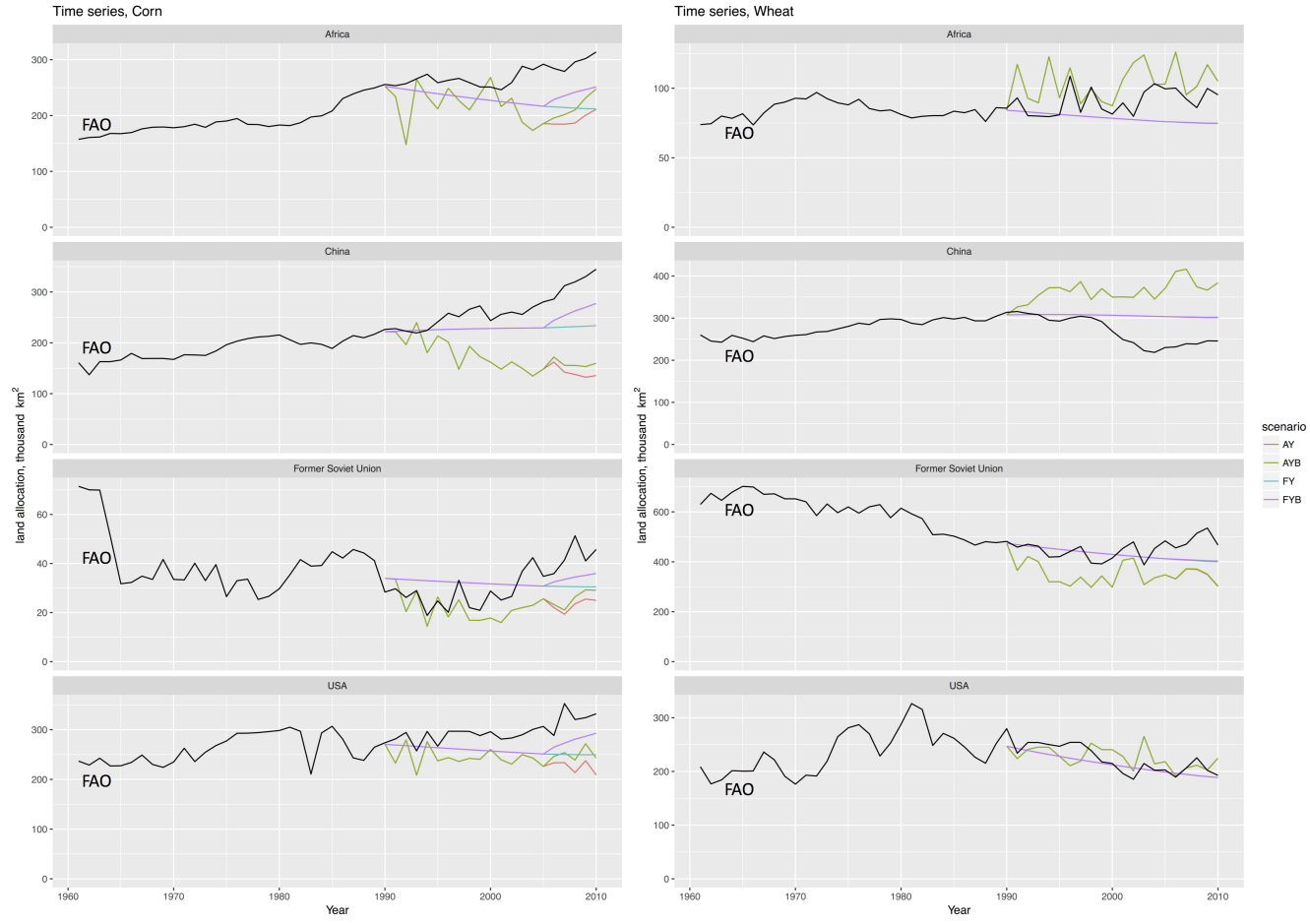


**Figure 7.** Time series for land allocated to Corn (left) and Wheat (right) in units of thousand km$^2$ across select regions. The black time series in each panel represents FAO observational data. The colored time series correspond to different GCAM scenarios.

## 6 Conclusions

This analysis confirms that, 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). This suggests a larger challenge in evaluating Integrated Assessment Models. 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 (and therefore global demand) well, as measured by global bias $b_{iG}$. However, since agricultural commodities are traded on the global market, there are no regional constraints on supply to supplement. As a result, there are numerous regional supply solutions that may satisfy the global constraint. This provides ample opportunity for error cancellation across regions. Any integrated assessment model requiring globally balanced supply and demand without additional regional constraints will



likely encounter this same issue. Because there is both additive cancellation (Figure 1) and regional compensation (Figure 3) of errors, replicating global aggregates is a necessary, but not sufficient, constraint on model performance. Additional model validation metrics are required.

While many of the performance benchmarks used in the climate modeling literature compare performance across models, the performance benchmarks identified for the measures implemented in this work allow the performance of GCAM to be evaluated in absolute terms, with context given by the intrinsic statistics of observational time series. This modification is necessary as no other IAMS have completed similar land use hindcast experiments to date. Therefore, there is no opportunity to examine the performance of GCAM relative to the performance of another IAM.

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. 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.

The types of results found here for GCAM land allocation are generally the extent of what can be achieved with deviation based measures of model performance. Together, the series of metrics highlights the strengths of the GCAM land use module and suggests specific structural changes to improve the modeling of land use.

This work raises several questions. The first is 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 a sense of measurement uncertainty. When fitting trend lines to the FAO data, it becomes 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 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.

A second question applicable to any IAM is how to evaluate the model as a whole. The GCAM land use module was used as a case study here and in past work. However, the land use module was not run in isolation. It interacts with all of the other systems modeled in GCAM and the current work provides no sense of the changes seen in other systems. The scheme implemented here could certainly be applied to each of the other systems (assuming observational data for the period is available), but the number of variables to examine may be large enough to be intractable. A remaining challenge is to develop a method to evaluate such a large system without the use of global aggregates.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2017-97, 2017
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 1 June 2017




## 7 Data availability

The data analyzed in this work is publicly available at https://github.com/JGCRI/LandHindcastPaper.

*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.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2017-97, 2017
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 1 June 2017




## References

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.

Edmonds, J. and Reiley, J.: Global Energy-Assessing the Future, 1985.

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.

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.

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.