Response to referee 1

Authors' Comment: We express our thanks to the reviewer for the thorough and constructive comments. We have re-structured our manuscript to focus the presentation of our story. We respond to each point below.
Authors' changes: While the article has been significantly re-organized and expanded from the initial submission, almost all of the original content remains.

Reviewer Comment: This paper proposes a methodology for evaluating the hindcast results generated by integrated assessment models (IAMs) or land use models. As a case study, GCAM land use results are evaluated. The authors found that global aggregates are not sufficient for evaluating IAMs. Additionally, the deviation measures examined in this work successfully identity parametric and structural changes that may improve land allocation decisions in GCAM. The suggested future work is involving some 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.
Authors' Response: The reviewer's summary of our paper has helped clarify a re-structuring of our article to better communicate our aim.
Authors' changes: The focus of the paper is now more explicitly on presenting a set method for evaluation of IAM hindcast experiments rather than any particular model improvement. The application of the method to re-analyze past GCAM hindcast results (reference below) is intended to demonstrate that the evaluation method in this paper expands the insight for model improvement relative to the originally used evaluation metrics. Specifically, we communicate a narrowed focus: This article is focused on presenting an evaluation scheme for any variable (with observational data available) resulting from an IAM hindcast experiment. The question of how to more holistically evaluate models as complex as IAMs is an area for future research. The results of this work indicate to us that no single, quick evaluation measure is possible and sector by sector evaluation may be necessary. In particular, we find that global aggregates are truly not a sufficient measure, and we believe this is a valuable finding for the IAM community. The introduction has been expanded to more clearly explain this work's motivation and aim. We pre-define the goals of an evaluation. We then outline a tractable family of metrics that can meet these goals in section 2 beginning on page 3. The past GCAM hindcast experiment is described and reanalyzed entirely in its own section (now section 3 beginning on page 6) with this new evaluation scheme to demonstrate that 1) the evaluation goals are met and 2) the resulting application highlights GCAM's strengths and weakness in a more detailed manner than the original skill scores used in past works.

Reviewer Comment: The overall text is well written, and the logic is understandable. However, there are several concerns before publishing. Here, I listed some points that must be modified or improved. - The way how they use the metrics and to draw the conclusions which argue about the potential model improvement seems not comprehensive and quite naïve. The essence of this hindcast experiment exercise land use is surely determined by the crop demand and trade together with the yield information (either direct observation or extrapolation). At least without the assessment of demand reproducibility, it would be difficult to make conclusions.
Authors' Response: We agree with the reviewer that the demand side is fundamental to understanding all aspects of GCAM's performance. Applying our evaluation method to this sector of GCAM will be a rich area of future work. We believe that our restructuring and renewed focus on model evaluation, however, puts such an examination outside the scope of this paper. We believe that our re-structuring and clearer focus on presenting a method for hindcast evaluation, rather than GCAM specific improvements, implicitly addresses the lack of comprehensiveness as well. The GCAM specific improvements are now more clearly an example of the types of insights that may be drawn from the evaluation method that is the focus of this paper, as well as an illustration that greater insight is possible with this evaluation method than globally aggregated skill scores.
Authors' changes: We have explicitly addressed the non-comprehensive nature by first, combining all GCAM-specific background and results into a single section (section 3, beginning on page 6 of the revised manuscript), and second, expanding the GCAM-specific results section (section 3.2 page 9) to note our motivation in presenting the particular results selected. We have also added plots of the full results to the repository cited in the data availability section (full result tables of results have always

been available in the repository).

Reviewer Comment: The data chosen to display looks arbitrarily decided and not comprehensive. Starting from global total is fine. Then, the analysis went to specific crop (wheat) and country (USA). Looking from one of the objectives of this study which identifies model improvements, the analysis should be comprehensive. Based on the discussion in the last conclusion part which takes broader issues like FAO data things, the comprehensiveness seems important here also.
Authors' Response: Agreed.
Authors' changes: See above.

Reviewer Comment: Although the paper says that neither of Fujimori et al.'s techniques are compatible with their goals and methodology, the objective in Fujimori et al. seems quite similar to this paper's method. It is because the method in Fujimori et al. clearly states that "the regression method is focusing on the bias in the discrepancies between the simulation results and statistics by regions and years to identify which regions and years for each variable have large discrepancy." The authors should discuss what is the advantage and disadvantage of the proposed method in this paper.
Authors' Response: agreed.
Authors' changes: We have expanded this explanation in the context of our re-structuring. This is done in the paragraph beginning on page 2, line 31.

Reviewer Comment: GCAM model description should be more enriched. It is because in the latter part, they discuss about producer price, logit exponent and trade and so on. At least those things should be clearly described.
Authors' Response: We agree that the GCAM model description should be enriched and have done so in section 3.1, beginning on page 7.
Authors' changes: We have moved our section describing GCAM to after the detail of our evaluation method; we have combined it with the section describing the data of the first GCAM hindcast experiment that we re-analyze, and we expand some aspects of our explanation. We have more clearly cited the papers in which those definitions are provided, but feel that the repetition of the full content of those papers is unnecessary given the now-narrowed focus of the paper. We also clarify our explanation of producer prices.

Reviewer Comment: The carbon price already exists in the real world around 2010, and is it taken into account? This might have been discussed in their paper Calvin et al 2017 but as far as reading GCAM papers, the land use part is really sensitive to carbon price and sometimes looks unrealistic. It would be better to validate that part.
Authors' Response: Similar to the reviewer's observation regarding the importance of trade in evaluating all aspects of an IAM hindcast experiment, we agree that carbon prices would be a key avenue for future investigation but fall outside the scope of this paper. The availability of historical data for Carbon Price on land against which to validate model performance may also pose a problem.

Reviewer Comment: Line 7 P1; about the description "this is key in the integrated assessment community, where there often are not multiple models conducting hindcast experiment", I think the fact that not multiple models conduct hindcast should not be the reason why they need absolute term evaluation. Even if hindcast is carried out many similar models, it should be evaluated independently (for example, macro econonometric models like DSGE do validate individually).
Authors' changes: We have changed the abstract and introduction to reflect the re-structuring and narrowed focus of the paper. In particular, two of our goals for an evaluation method are to develop measures that can be used absolutely for evaluation of a single experiment for a single model AND relatively to compare the results of multiple experiments for a single model or the same experiment repeated across multiple models to aid the community in inter-comparison studies. The correspondingly re-written sentence begins on page 1 line 7: "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."

Reviewer Comment: Line 22, P1; It would be better to specify "other model validation exercises"
Authors' changes: This sentence has been re-written to clarify our intent. The corresponding re-written sentence is on page 2 line 1, "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)."

Reviewer Comment: Line 3 P2; Are the references all GCAM 3.0?
Authors' Response: Unless otherwise noted, yes.
Authors' changes: We have added language to the GCAM description in Section 3.1, page 7 line 17.

Reviewer Comment: Line1 P14; I cannot understand this sentence. are USA producer prices used globally?
Authors' Response: we have rewritten this and moved it to our expanded GCAM background section 3.1.
Authors' changes: Beginning on page 7 line 31, the text now reads: "GCAM uses a global market price (where global supply equals global demand) to set producer prices used by economic agents in profit calculations underlying land allocation decisions. Currently, every land use region shares the same producer price, initially the US base year price for calibration. This is partly due to data availability, but 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."

Reviewer Comment: The sentence "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" is strange. From the model point of view, the yield in all four scenarios are given parameters. So the different between (AY, AYB) and (FY, FYB) are not the matter of information quantity difference from real world.
Authors' Response: we have expanded and clarified this explanation in the section describing the first GCAM hindcast experiment, section 3.1 page 8.
Authors' changes: Now beginning on page 8 line 4, "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."

Reviewer Comment: In conclusion, authors suddenly address about trade and no discussion in results part. It seems strange and would be better to discuss in the results part more and derive some summary in the conclusion part.
Authors' Response: This comment was a key motivation in our restructuring of the paper to reflect the narrowed focus on model evaluation (rather than improvement) and highlight the demonstrative role played by reanalysis of the GCAM land allocation hindcast experiment with respect to our evaluation method.

Response to referee 2

Authors' Comment: We express our thanks to the reviewer for the insightful comments.
Authors' changes: We have re-structured our manuscript to focus the presentation of our story. We respond to each point below. While the article has been re-organized and expanded from the initial submission, almost all of the original content

remains.

Reviewer Comment: This paper describes an experiment in which the GCAM model is calibrated to the historical baseyear of 1990 and ran forward to the year 2010 to simulate historic changes in land use. The experiment is done under four different assumptions, including or excluding the historic trends in yields and including or excluding the US renewable fuel standards. They authors conclude that history is best explained when trends in yield and the US renewable fuel standard are included in the assumptions of the model.

Authors' Response: The reviewer's summary highlighted our need to better communicate that the focus of our paper is presenting a set method for evaluation of IAM hindcast experiments rather than any particular model improvement; the application of the method to re-analyze the data from a past GCAM hindcast experiment is intended to demonstrate that the method outlined in this paper expands the insight for model improvement relative to the originally used evaluation metrics (reference below). More narrowly: This article is focused on presenting an evaluation scheme for any variable (with observational data available) resulting from an IAM hindcast experiment. The question of how to more holistically evaluate models as complex as IAMs is an area for future research. The results of this work indicate to us that no single, quick evaluation measure is possible and sector by sector evaluation may be necessary. In particular, we find that global aggregates are truly not a sufficient measure, and we believe this is a valuable finding for the IAM community

Authors' changes: The focus of the paper is now more explicitly on presenting a set method for evaluation of IAM hindcast experiments rather than any particular model improvement. The application of the method to re-analyze past GCAM hindcast results (reference below) is intended to demonstrate that the evaluation method in this paper expands the insight for model improvement relative to the originally used evaluation metrics. Specifically, we communicate a narrowed focus: This article is focused on presenting an evaluation scheme for any variable (with observational data available) resulting from an IAM hindcast experiment. The question of how to more holistically evaluate models as complex as IAMs is an area for future research. The results of this work indicate to us that no single, quick evaluation measure is possible and sector by sector evaluation may be necessary. In particular, we find that global aggregates are truly not a sufficient measure, and we believe this is a valuable finding for the IAM community. The introduction has been expanded to more clearly explain this work's motivation and aim. We pre-define the goals of an evaluation. We then outline a tractable family of metrics that can meet these goals in section 2 beginning on page 3. The past GCAM hindcast experiment is described and reanalyzed entirely in its own section (now section 3 beginning on page 6) with this new evaluation scheme to demonstrate that 1) the evaluation goals are met and 2) the resulting application highlights GCAM's strengths and weakness in a more detailed manner than the original skill scores used in past works.

Reviewer Comment: The first sentence of the abstract (but also the main introduction) shows that the authors suffer from a syndrome that is all too common among IAM modelers: selective amnesia. There are several examples of hindcasting-type experiments in the (broader) IAM community, even though they not always use the keyword 'hindcasting'. If the authors had thoroughly read the introduction of Fujimori et al. 2016, they would have found about five additional examples that would be valuable to cite in this paper.

Authors' Response: Thank you for noting these omissions.

Authors' changes: We have expanded our reference list as well as adjusted the abstract and introduction. (page 1 line 1, page 2 line 1, page 2 line 31).

Reviewer Comment: The described experiment is fairly simple and straightforward, but immediately raises three questions that are not satisfactory dealt with in the paper: 1) Would the GCAM model reproduce historic trends better if some key parameters had other values? 2) Can we use this analysis to draw conclusions about the influence of the US renewable fuel standard on global land use? 3) What does this study imply for applications in which the GCAM model is ran forward into the future?

For the first issue, the authors could identify a few key-parameters (e.g. elasticities) and assume a range of values. By running the hindcasting experiment with these different values, they would learn something about the behavior of the GCAM model itself and whether certain parameter settings better explain the historic trends.

Authors' Response: We agree with the reviewer that using this evaluation method in the future to analyze the results of model runs spanning a parameter or parameters (such as elasticities). We feel that this is outside the scope of our re-structured paper, however.

Authors' changes: We expand our intro (the paragraph beginning on page 2 line 6) to specify that, upon doing such an experiment, a definition of "better explaining" was necessary. This work seeks to provide such a quantified definition, and the restructuring reflects this focus. The results of the first hindcast experiment motivate the goals our evaluation method must meet and the reanalysis of the first data serves as a demonstrative example of how the evaluation method may be applied and the types of results that may come from it. In the now self-contained GCAM-specific section 3 (paragraph beginning on page 8 line 4), we clarify that the original paper does indeed take this approach for one parameter of interest (structure of exogenous yield inputs).

Reviewer Comment: The second issue would make the paper a lot more relevant to a non-modeling audience. If the US renewable fuel standard considerably changed land use trends, this should have had consequences for land use emissions and indirect land use change. The difference between the FY and FYB scenarios should be the impact of the renewable fuel standard. Since several existing studies already examine the impact of the US renewable fuel standard on land use, the authors should compare the results of their experiment to these studies.

Authors' Response: We agree with the reviewer that this is a fascinating avenue for future investigation. Similar to the notion of parameter estimation for elasticities, future experiments could be designed to investigate the most accurate way to implement this standard but are outside the scope of our restructured paper.

Authors' changes: We have expanded the GCAM description to detail the implementation of the fuel standards used in the first hindcast experiment, section 3.1 paragraph beginning on page 8 line 15.

Reviewer Comment: On the third question, the authors briefly discuss how future applications of GCAM could be improved by updating yield information. However, a more direct comparison between the (common) assumptions for future runs vs these historic scenarios would be valuable. What is the typical setup for a future run? The AY scenario? What does that set of assumptions imply for interpreting future results of the model? Do errors compound over time, and should users be worried about the long-term results of the model? Such questions are not discussed at the moment and would be a relevant addition to the final sections of the paper.

Authors' Response: We agree that these are key questions to address, and we have expanded different aspects of our GCAM-specific section 3 to detail each.

Authors' changes: The reference set up of GCAM is detailed in the paragraph beginning on page 7 line 24. The relation of the reference set up to the scenarios re-analyzed in this work is covered in the paragraph beginning on page 8 line 4. Finally, considerations of how the GCAM-specific results for the scenarios re-examined relate to other setups and sectors of GCAM are discussed in the paragraph beginning on page 17 line 19.

Reference: 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.

Response to short comments

Short comment 1: In particular, please note that for your paper, the following requirements have not been met in the Discussions paper:

- "All papers must include a section, at the end of the paper, entitled 'Code availability'. Here, either instructions for obtaining the code, or the reasons why the code is not available should be clearly stated. It is preferred for the code to be uploaded as a supplement or to be made available at a data repository with an associated DOI (digital object identifier) for the exact model version described in the paper. Alternatively, for established models, there may be an existing means of accessing the code through a particular system. In this case, there must exist a means of permanently accessing the

precise model version described in the paper. In some cases, authors may prefer to put models on their own website, or to act as a point of contact for obtaining the code. Given the impermanence of websites and email addresses, this is not encouraged, and authors should consider improving the availability with a more permanent arrangement. After the paper is accepted the model archive should be updated to include a link to the GMD paper."

- Inclusion of Code and/or data availability sections is mandatory for all papers and should be located at the end of the article, after the conclusions, and before any appendices or acknowledgments. For more details refer to the code and data policy. Thus, please add a Code availability section stating how the GCAM model can be accessed. Additionaly, please consider uploading the data set to a data repository as described above.

Authors' Response: All GCAM releases from 3.0 onward are available at: https://github.com/JGCRI/gcamcore/releases. Results from GCAM 3.0 simulations were used in this work. The land data output from GCAM 3.0 runs, as well as code for analyzing it and producing figures, has availability provided in Section 5 of the paper and is publicly available at https://github.com/JGCRI/Land

Short comment 2: GMD is strongly encouraging (but does not enforce) authors to provide persistent access to their program code and data used in the manuscript. Typically this is guaranteed through the use of a DOI which can be created for releases made in GitHub using Zenodo. Alternatively, the relevant data can be supplied as a supplement to the manuscript at GMD. In this spirit I would like to suggest to upload a tar-ball of https://github.com/JGCRI/LandHindcastPaper as a supplement and to state the license for the use of the data in the manuscript and in the supplement
Authors' Response: We will add such a supplement when uploading the revised manuscript.

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~~ 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. ~~This is key in the integrated assessment community, where there often are not multiple models conducting hindcast experiments to allow for model intercomparison~~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 ~~serve a second purpose~~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. ~~As a case study~~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 ~~,~~ ~~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.~~ 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. ~~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 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., 201 . 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; Edmonds and Reiley, 1985; 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. ~~Brief background on the land use system in GCAM is provided in Sect. 3.1 with more detailed information available in Wise et al. (2014) . The operating assumptions used are outlined in Sect. 4 and represent different levels~~ The different scenarios represent different extremes of information for ~~land allocation decisions available to the economic agents in the land use module. The~~ 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 ~~limits the insight for model improvement that they can provide by masking important deficiencies.~~ 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.

~~A hindcast experiment was also~~ 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; Krieg 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., 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.~~ 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.

~~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~~ 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. ~~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~~

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

observational data , and forecasts evolve in time from The scheme is then used to re-examine the land use data from Paper 1 to demonstrate application of the evaluation method and 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. resulting expanded results relative to application of skill scores.

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.

2 Evaluation methods

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

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(regardless whether RMSE or MAE), the types of conclusions that could be drawn would not differ appreciably with the specific measure chosenwhether 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, σ_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. 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~~ regional normalized RMSE	~~standard deviation around time mean of observation~~ standard deviation ar mean of observation
v'_{ij}	~~regional normalized centered RMSE~~ regional normalized centered RMSE	~~standard deviation around time mean of observation~~ standard deviation ar mean of observation
$\hat{e}_{ij}$	~~revised regional normalized RMSE~~ revised regional normalized RMSE	~~standard deviation around trend line of observation~~ standard deviation aro lline of observation

2.3 Informative performance benchmarks

While the time series statistics outlined in ~~Sect.~~ 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 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 timeseries~~

3 A case study of GCAM 3.0 land allocation

The data described below and analyzed in Section 3.1 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.

4 ~~Data~~

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., 20 . 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 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.

~~The data analyzed in Sect. 3.1 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~~ 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.

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

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

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 uses a global market price (where global supply equals global demand) to set producer prices used by economic agents in profit calculations underlying land allocation decisions. Currently, every land use region shares the same producer price, initially the US base year price for calibration. This is partly due to data availability, but 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) ~~were performed~~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 ~~United Nations Food and Agricultural Organization (FAO)~~ FAO (FAO, 2014) during the period 1990 to 2010 for the nine GCAM crops with ~~corresponding~~ 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.~~ .

4 ~~Results~~

3.1 Results

~~The metrics outlined in Sect. 2.2 are~~ A selection of results demonstrating how the evaluation method summarized in Table 1 can be used to analyze ~~the GCAM land allocation output previously examined in the first hindcast experiment.~~ 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 ~~as well as~~ and expand the previous GCAM land hindcast results ~~(Calvin et al., 2017).~~ 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.

3.2 ~~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.

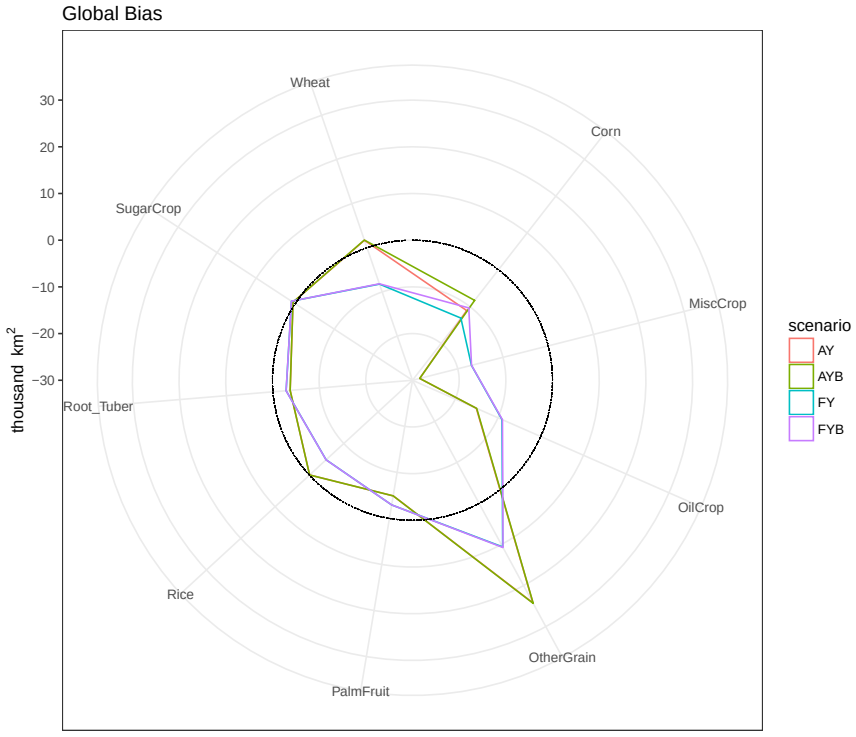

Figure 1. Global bias, b_{iG} (equation 5). The black circle corresponds to $b_{i,G} = 0$.

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.

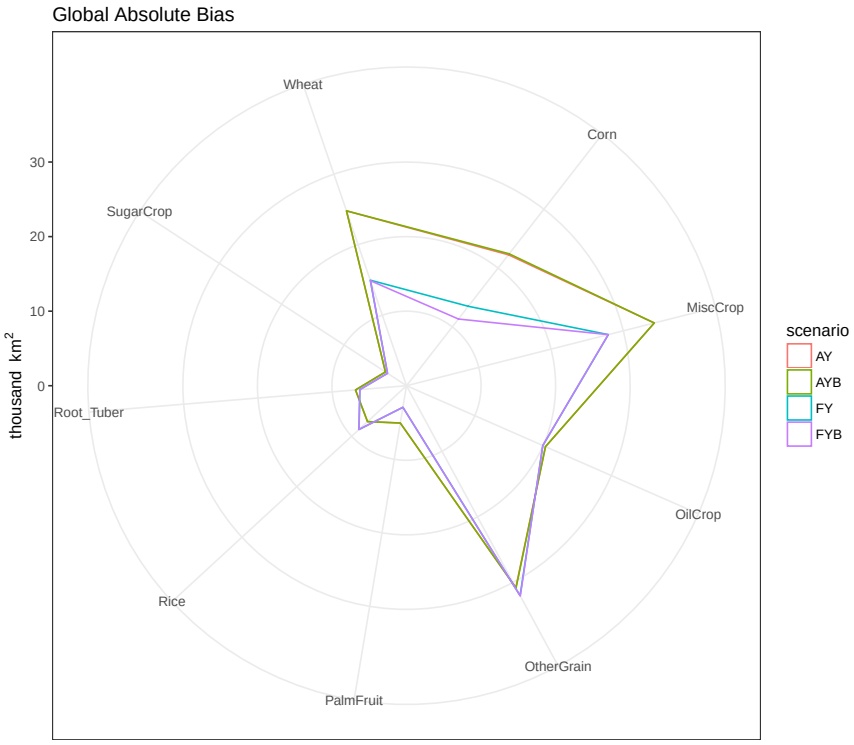

Figure 2. Global absolute bias, $|b_{iG}|$ (equation 6).

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.~~

~~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.~~

3.2 ~~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.

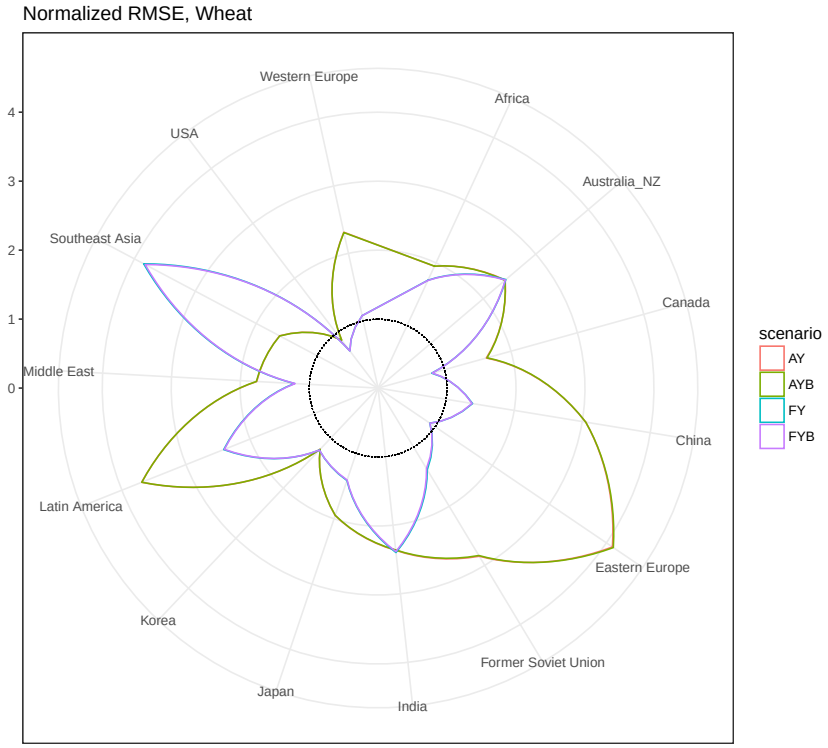

Figure 3. Normalized RMSE, e'_{ij} (equation 7), in each region for the land allocated to Wheat. The black circle is at the performance benchmark, $e'_{ij} = 1$, (equation 11). e'_{ij} compares RMSE error with the standard deviation of observation for each crop.

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.

~~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.

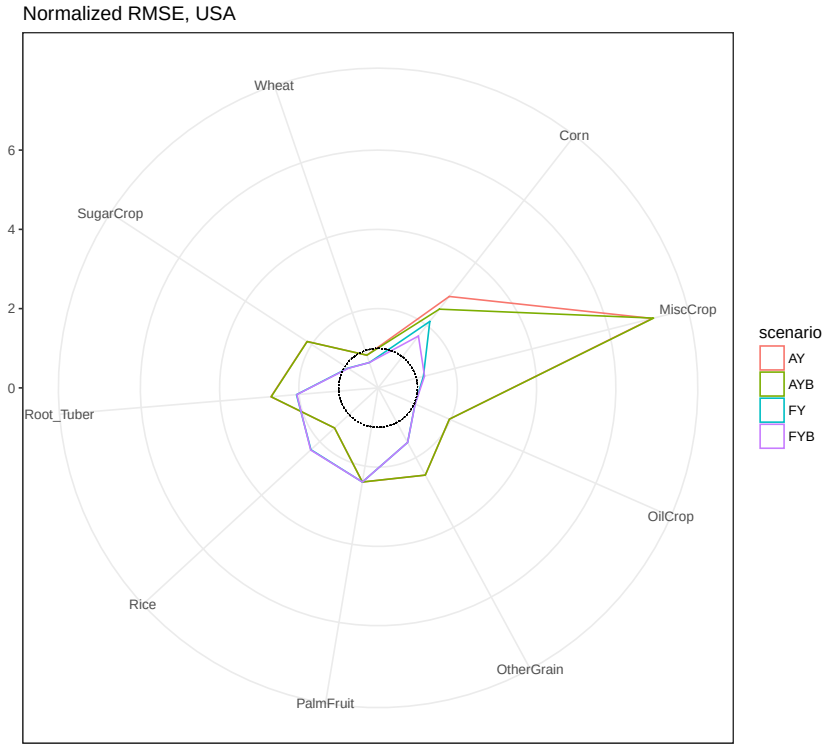

Figure 4. Normalized RMSE, e'_{ij} (equation 7), for each crop in the United States. A black circle is included for $e'_{ij} = 1$. e'_{ij} compares RMSE error with the standard deviation of observation for each crop.

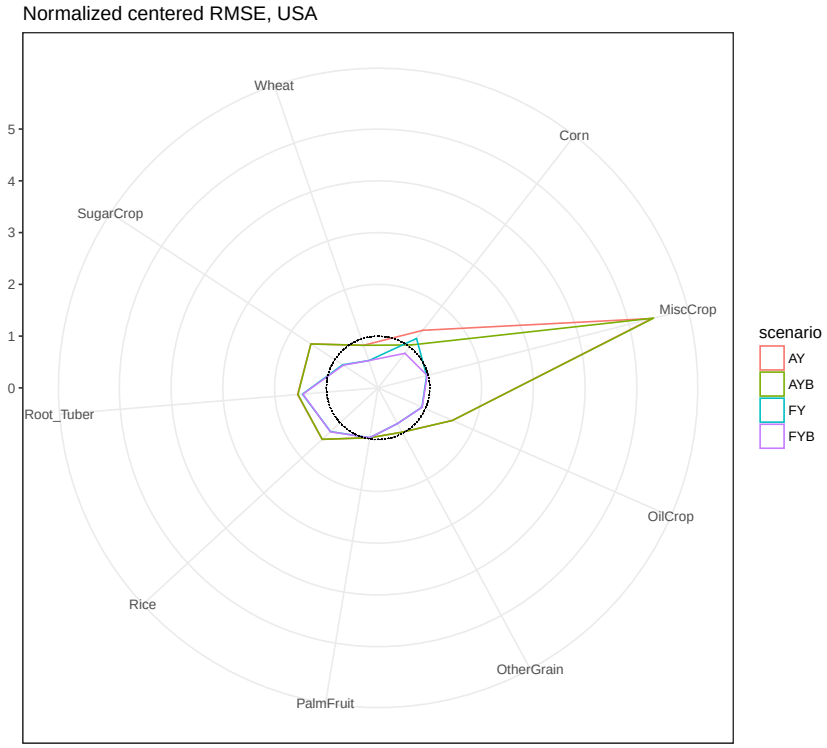

Figure 5. Normalized centered RMSE, v'_{ij} (equation 8), for each crop in the United States. A black circle is included for $v'_{ij} = 1$. v'_{ij} compares centered RMSE error with the standard deviation of observation for each crop.

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.

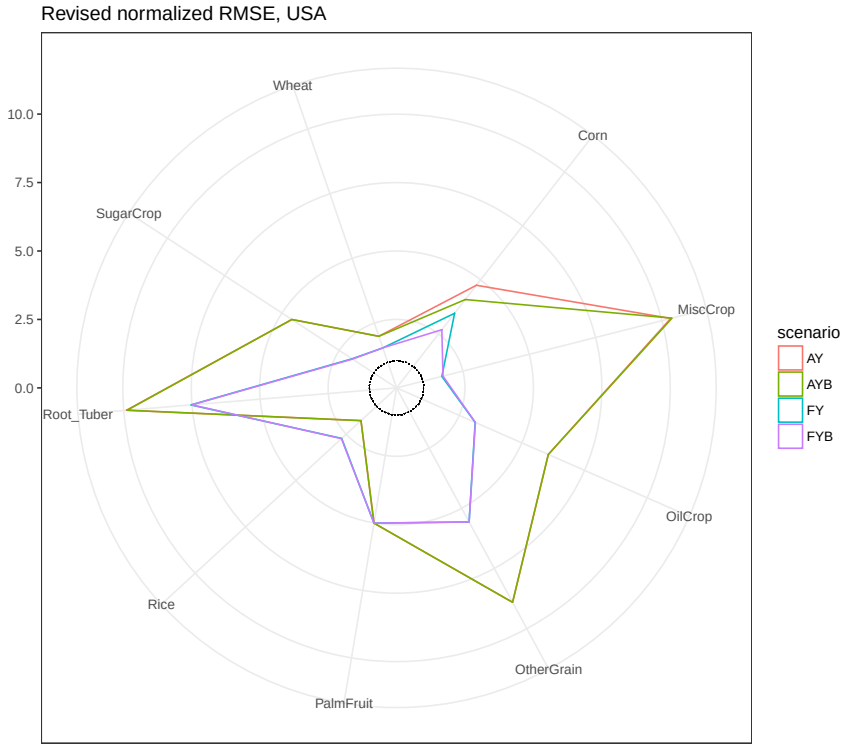

Figure 6. Revised normalized RMSE, $\hat{e}_{ij}$ (equation 10), for each crop in the United States. A black circle is included for $\hat{e}_{ij} = 1$. $\hat{e}_{ij}$ compares RMSE error with the standard deviation about the observed trend line for each crop.

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.~~ about year end yields to make planting decisions.

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.~~ using different distributions to govern competition.

~~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~~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 ~~scenarios incorporating forecasted yields (~~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. ~~At the other extreme, the scenarios using actual yield information (~~In contrast to the FY and FYB scenarios, the AY and AYB ~~)~~ scenarios 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~~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.

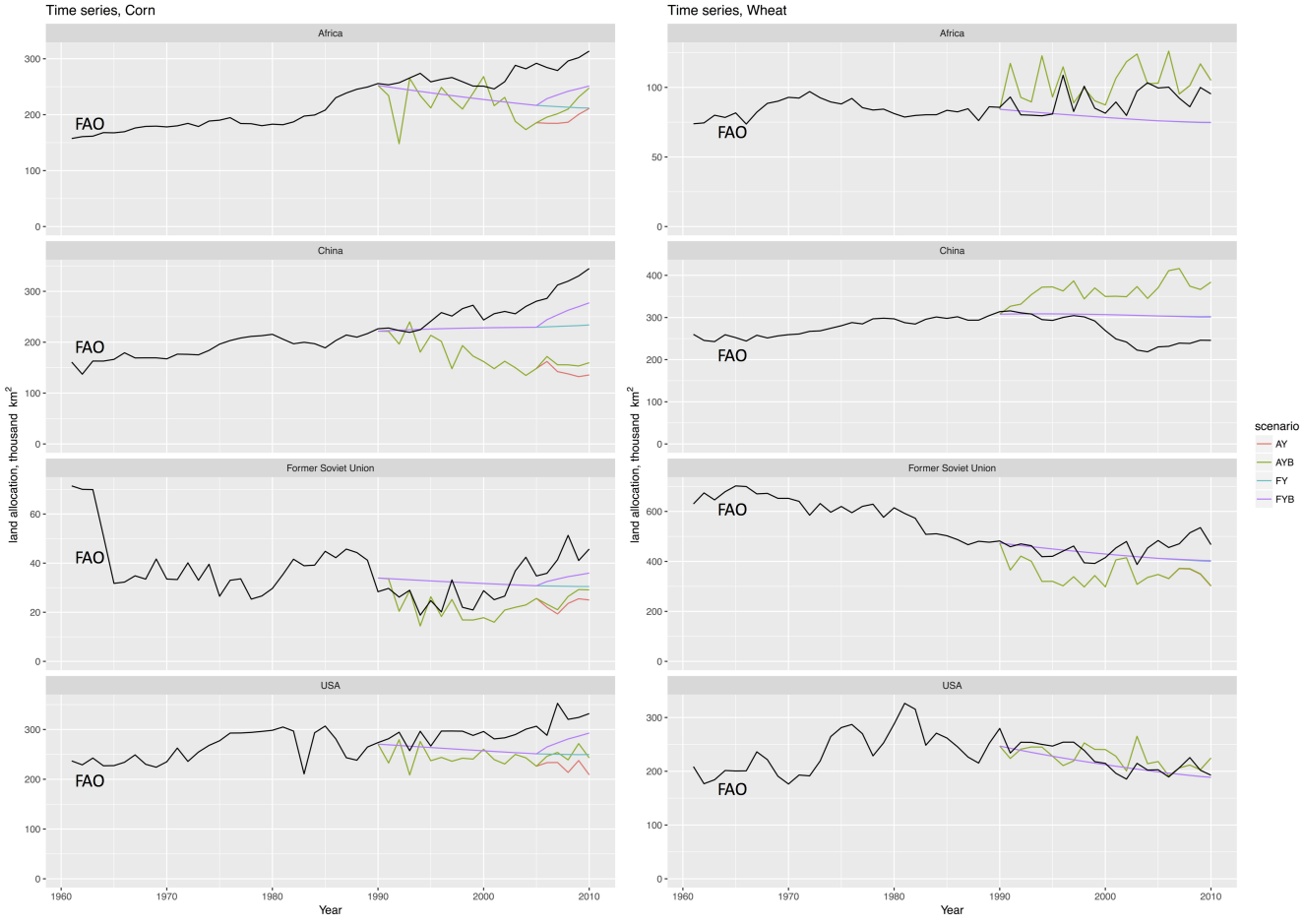

Figure 7. Time series for land allocated to Corn (left) and Wheat (right) in units of thousand km^2 across select regions. The black time series in each panel represents FAO observational data. The colored time series correspond to different GCAM scenarios.

~~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.~~

3.2 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 ~~(and therefore global demand)~~ well, as measured by global bias b_{iG}~~. However, since agricultural commodities are traded on the global market, there are~~ , Figure 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~~. 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.~~ , demonstrated in Figure 3.

~~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. 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.

~~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.~~ 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.

~~This work raises several questions. The first is~~

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 ~~a sense of~~ measurement uncertainty. When fitting trend lines to the FAO data ~~, it becomes~~ for use in the revised normalized RMSE metric, $\hat{e}_{i,j}$ (equation 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.

~~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~~

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. ~~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.~~

6 ~~Data availability~~

~~The data analyzed~~ Results from GCAM 3.0 simulations were used in this work~~is publicly available at~~. All GCAM releases from 3.0 onward are available at:
https://github.com/JGCRI/~~LandHindcastPaper~~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.

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.

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.