Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2017-97-AC1, 2017

Geoscientific
Model Development
Discussions

Interactive comment on “Evaluation of Integrated Assessment Model hindcast experiments: A case study of the GCAM 3.0 land use module” *by* Abigail C. Snyder et al.

Abigail C. Snyder et al.

abigail.snyder@pnnl.gov

Received and published: 26 August 2017

Anonymous 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.

Author’s 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.

Reviewer comment: 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.

Reviewer comment: - 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), 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 restructuring. 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: Other minor points are below. - 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 restructuring 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.

Author's 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: - Line 11 P14: 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.

Anonymous 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

Author's 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.

Reviewer comment: 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?

Reviewer comment: 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.

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

âĂć "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."

âĂć 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: this has been added to the revised manuscript. 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 7 of the paper and is publicly available at https://github.com/JGCRI/LandHindcastPaper.

Short comment: 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.

Please also note the supplement to this comment:
https://www.geosci-model-dev-discuss.net/gmd-2017-97/gmd-2017-97-AC1-supplement.pdf

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-97, 2017.