This article presents an ambitious modelling exercise that combines the LPJ-GUESS dynamic global vegetation model with the PLUM land-use model (now under the new name LandSyMM). While the model has been presented before in Alexander et al 2018, the coupled models have not been presented before from the LPJ-GUESS perspective which makes this an interesting article. Additionally, the large number of ecosystem service indicators make the presented analysis interesting on its own.

Before the article can be published in ESD however the structure of the article needs substantial improvement. While I acknowledge the challenge of describing such a complex model in a comprehensive as well as concise manner, I do think improvements can be made. I have three general points and a number of more detailed points that need to be addressed.

We thank the referee for their helpful comments.

The structure of the methods section is confusing. I would recommend to start with an overview of the models used and how they interact (a diagram might be helpful), then a detailed description of the different ecosystem service indicators analysed.
Also, I would recommend to reduce the size of the methods section by moving some of the information on detailed input for LPJ-GUESS and PLUM to the SI. On the other hand, much more information is required on the scenario setup. Details about the main assumptions should be presented. A table with an overview would be helpful. It is important that the reader does not depend on a different article to understand how the presented scenarios are defined.

The extra experimental scenarios that have been performed to improve understanding of the results are confusing. Maybe these can be added to a descriptive scenario table, either in the main text or the SI, and they could be given explicit names to make it easier to refer to them in the results section, e.g. SSP4-2.6-noCO2/noLUC. I think this is better than for example the long text 'in the constant climate+CO2 experiment' (line 295-296).

This section has been significantly reworked:

- Sect. 2.1 now focuses on LPJ-GUESS, with the text on ecosystem services having been moved to the Introduction and the new Sect. 2.5.
- Sect. 2.2 focuses (as before) on PLUM, now including some text about where it uses data from and gives data to LPJ-GUESS.
- Sect. 2.3 describes how the coupling works. This is necessarily very technical, but a flowchart figure is now provided for clarification.
- Sect 2.4, describing input data and scenarios, has been compressed significantly relative to the old Sects. 2.3.1–2. Technical information regarding data sources is now less prominent, with about half of the section serving instead to provide context about the SSPs and RCPs. For interested readers, the Supplementary Methods provide more technical detail.
- Sect. 2.5 focuses now solely on the ecosystem service indicators used in the study. Background information on the ecosystem services in question has been moved to the Introduction.

A LandSyMM overview diagram (Figure 1) has been added to Sect. 2.3. Experiment nomenclature has been standardized and is explained in the new Table 1 (Sect. 2.3).

The results and discussion section is rather lengthy and is (as the names correctly states) a mix between results and discussion. I think this reduces the clarity of your story and is not in line with the standard outline of scientific articles. I think the article would greatly benefit from splitting up this section into a clear description of the results (topic by topic) and subsequently a discussion of the results in the context of the literature (again, topic by topic).

We agree that it would be more in line with common practice to move such comparisons to a separate Discussion section, but while this works well for most papers, we believe it would not in our case. First, some of our results with regard to land use area are so striking (and different from results from similar work) that it makes sense to address them immediately. This also provides the reader with valuable context for interpreting the rest of our results. Finally, because this paper touches on so many different ecosystem services, it is necessarily rather long. Postponing the literature comparison to a separate Discussion section would necessitate spending space there reminding the reader of our own results; in the interest of keeping this paper as concise as possible, we have chosen to avoid that.

Detailed comments:

Line 9-10: this is the first time biodiversity is mentioned in the abstract, while it is presented as one of the major outcomes.

We have added "biodiversity" to the list at Line 5.

Line 15: please rephrase 'larger than today's by anywhere from' to something more concise and more academic, e.g. 'an increase ranging from 1.5 billion to 6 billion'

This clause now reads "with a population increase by 2100 ranging from 1.5 billion to nearly 6 billion people".

Line 57-102: this section describes the ecosystem service indicators presented in the article. This should be a separate section. In addition, a lot of the text is introduction to why these indicators are important. I think this reduces the clarity of the methods section which should be a more technical description of the indicators presented in the article. Maybe you can move part of the text (in a more concise way) to the introduction.
Line 93-102: I assume the biodiversity indicator is not a standard output of LPJ-GUESS, correct? It is now part of the LPJ-GUESS section which is misleading. It should be more clear that this is calculated based on the downscaled PLUM results.

Description of the analyzed ecosystem service indicators has been moved to the new Section 2.5. More introductory or background text about the ecosystem services has been moved to the Introduction.

Line 109-110: The Popp et al 2017 article describes a large number of SSP scenarios from 5 IAM models. Please specify which scenario from which model has been used and preferably refer to a paper that presents the results specifically for this model.

We now specify that the bioenergy demand comes from the MESSAGE-GLOBIOM model.

Line 115: this is a very detailed start of this section. Consider restructuring the sentence. Maybe move this entire section to the SI as very detailed information for main text.
Line 137-138: a summary of the climate and land use data used should be given in the text, not in the SI. Consider moving this text to the SI and summarizing input data in the main text.

The relevant text (now Sect. 2.3) has been overhauled, and an overview figure (now Fig. 1) has been added.

Line 161: make sure your language is more academic. Don't use terms like 'briefly' or 'attempts'.

In this matter of style, we disagree with the reviewer. We have left in "attempts" because it is an accurate description of what the algorithm does. "Briefly" primes the reader to expect a non-comprehensive explanation of the algorithm, improving reading flow. That said, most of this text has been expanded and moved to the new Supplementary Methods Sect. SM3, and "briefly" has been removed.

Line 178: what is minimum 'non-agricultural area'. This sounds like a PLUM-specific technical term, please rephrase.

This has been clarified: "These included input and transport costs, tariffs, and minimum non-agricultural area (which places an upper limit on the total fraction of a gridcell that PLUM can allocate to cropland and pasture)."

Line 203: 'cropland area expands about 10% between 2050,. . .'. Between 2050 and what?

This has been corrected to read "between 2050 and 2100".

Line 214: 'what PLUM calls "other management"'. Please use more academic language and avoid usage of model-specific technical terms.

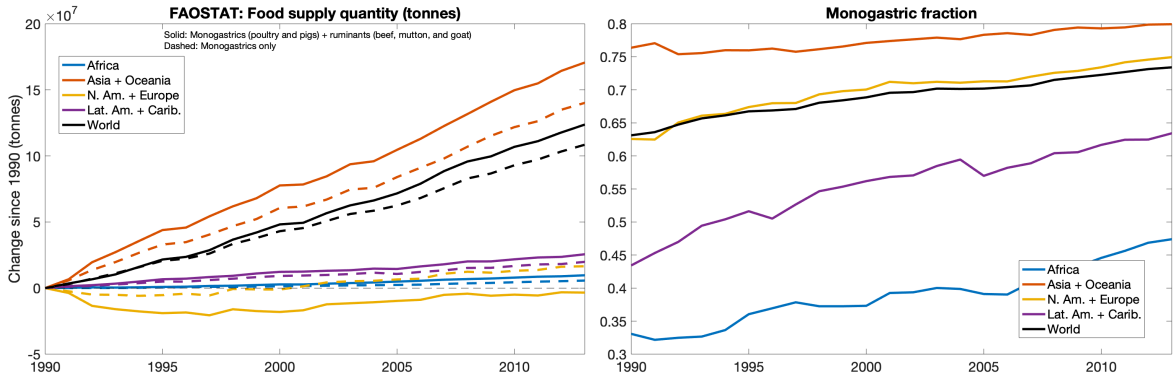
We have changed "what PLUM calls" to "PLUM's". However, we have left the reference to "'other management' intensity," since avoiding model-specific terms harms reproducibility. For readers not already familiar with the term, our explanatory parenthetical text—"(representing, e.g., pesticide application)"—should provide sufficient clarification.

Figure 1: why do you only show ruminants? Non-ruminants also have a very strong effect on the agricultural system due to high feed requirements.

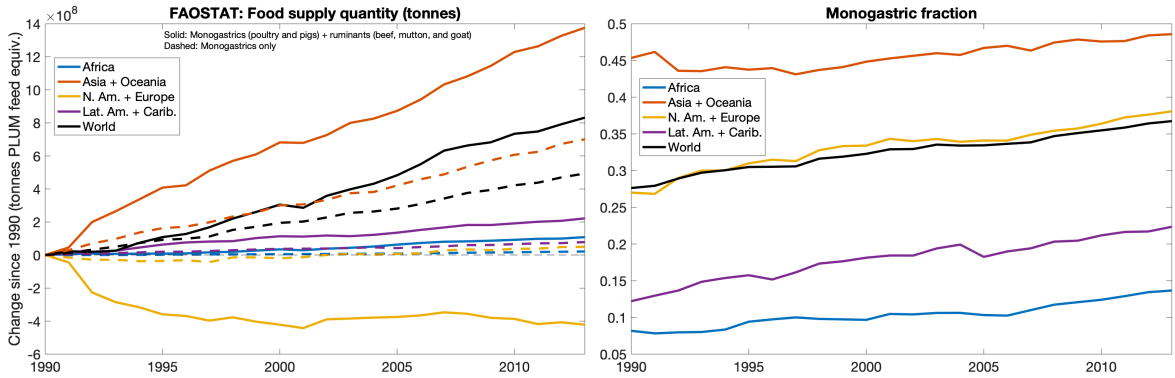
We originally excluded monogastrics from this figure in the interest of limiting its size, since the role of monogastrics demand is not discussed in the text, and interested readers could find its trajectory for each SSP in what is now Fig. SR2 (global demand trajectories for each commodity; formerly Fig. SR1). However, we have now added monogastrics demand to this bar graph (now Fig. 2).

Also, I am surprised about the very strong increase in ruminant demands. FAOSTAT actually shows that in recent decades demand for ruminants-based products has increased relatively little while monogastric-based products have increased much stronger.

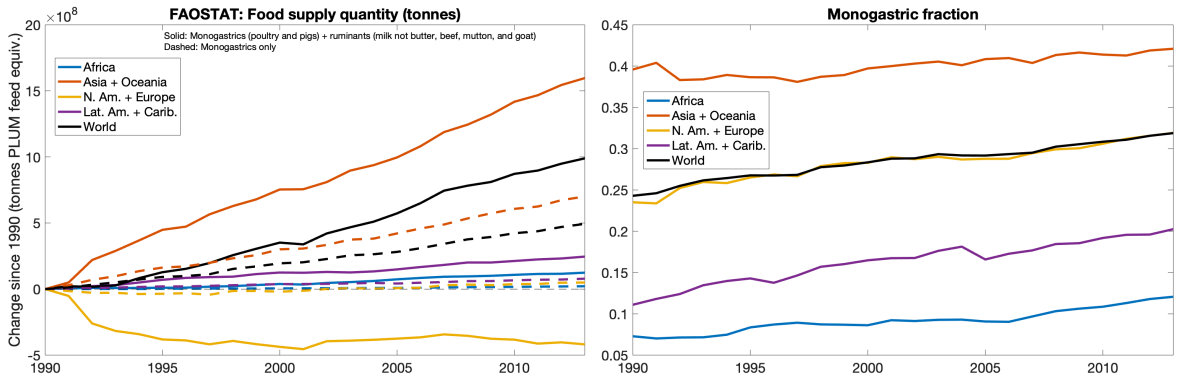
Monogastrics do make up a majority of meat supply by raw tonnage:



However, they make up a minority of meat supply once converted to the units of PLUM demand, which are tons feed equivalent: Ruminants require much more feed to produce a given weight of meat than do monogastrics. To convert, we multiply poultry meat, pig meat, mutton/goat meat, and beef respectively by 3.3, 6.4, 15.0, and 25.0 tons feed per ton product:

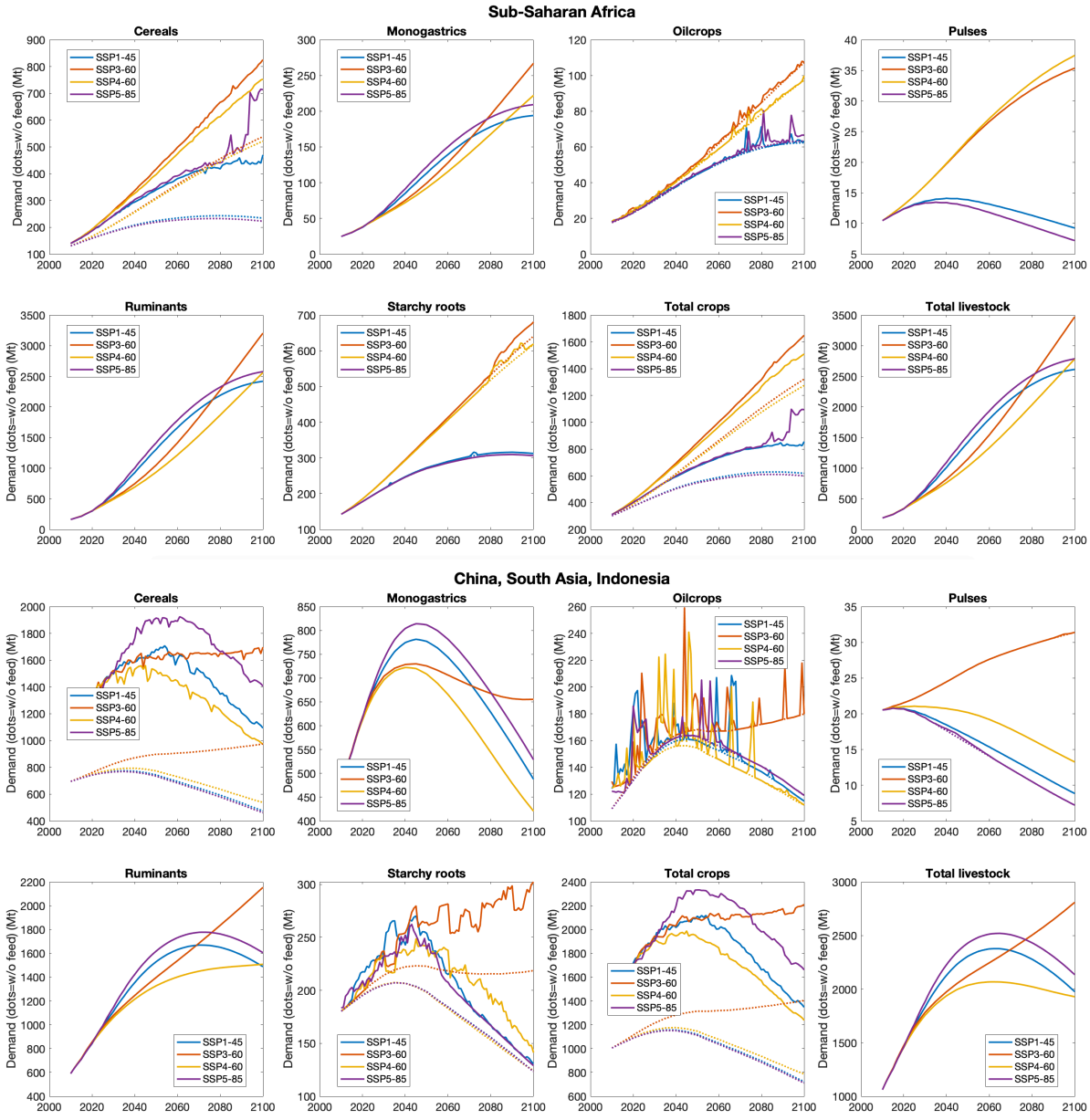


Additionally, ruminant feed requirements in PLUM include the production of milk, the consumption of which has been increasing more rapidly than ruminant meat. This further lowers the fraction of monogastric products in the feed-equivalent figures (0.7 tons feed per ton milk):

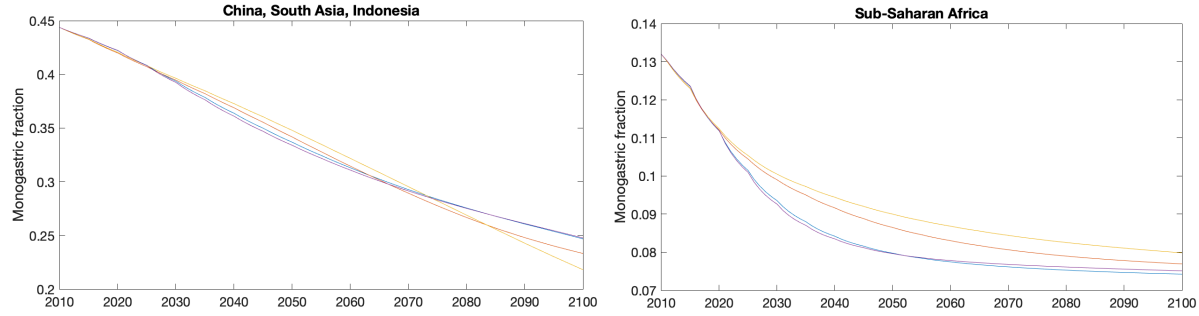


Furthermore, shifts in where demand is growing will change the makeup of the global livestock product landscape. The above figures show that recent trends in livestock meat demand have been driven by increasing consumption in Asia, where meat demand is mostly for monogastrics. In contrast, PLUM projects strong increases in livestock demand in sub-Saharan Africa as population and wealth increase there; that region has historically had a much lower fraction of livestock demand comprised of monogastrics. This is shown in the figures below, which present PLUM-projected demand for Sub-Saharan Africa as compared

with combined China, South Asia, and Indonesia. (Region groupings are different between these plots and previous. Also note that colors here stand for scenarios, whereas colors previously stood for regions.)



PLUM also projects that the fraction of livestock products (meat and milk) provided by monogastrics will decrease as these regions get wealthier:



Line 243: I was taught to avoid the term 'forecast' (sounds like something a fortune teller would say) but rather use the term 'project', but maybe this is a personal preference.

This has been changed to "projects".

Line 251-252: this sentence is extremely vague and unhelpful. Please make more explicit.

The last two sentences of this paragraph have been changed to the following: "While we do not expect LandSyMM's results to necessarily match those of other models, such a large, qualitative difference requires explanation. Several factors related to experimental setup and overall model structure likely contribute."

Line 276: it is impossible to have 500% or 700% higher land use, using current agriculture (~50 Mkm2) this would mean 250 or 350 Mkm2 which is more than the terrestrial area of the world.

These values refer to *relative* difference from the Alexander et al. (2018) differences. That is, the range of cropland area among scenarios in the present work is six times what the range was in Alexander et al. (2018). The text has been changed to read (new text in **bold**): "The **spread** in land-use area **projections** between the most extreme scenarios is much higher in this work than…"

Line 282: what do you mean with infrastructure efficiency? Also, if I understand correctly you make similar SSP-specific assumptions in PLUM (SM6). These should also lead to a higher spread in land-use projections, right?

PLUM does make some SSP-specific assumptions, but not the ones described in this sentence. The sentence has been changed to read (new/amended text in **bold**):

> As described above, IMAGE makes a number of assumptions (based on the SSP storylines) **that PLUM does not** regarding future deviations from historical 'business-as-usual' trends and relationships, including dietary shifts, **reductions in food losses during transport**, and forest conservation.

Line 276-278: I don't see why the RCPs lead to a much larger spread in scenario results?

After the cited sentence we have added the following: "The wide variation among the SSPs in population and economic growth trajectories, along with SSP-specific PLUM parameters (Sect. 2.3.2), contribute to this increased spread."

Also, is PLUM informed about the yield effects of climate change as these would impact for example trade and food security?

Informing PLUM about the yield effects of climate change (and changing CO_2 concentration) is indeed the reason we feed it with LPJ-GUESS-simulated potential yields. Text has been added throughout the manuscript to emphasize this point.

Line 325: is this realistic? Could not reduced feedback effects from lower evapotranspiration from forests in fact reduce runoff?

Line 336-337: how can increased agriculture reduce the risks on droughts? Please explain.

The end of the paragraph originally ending with Line 326 has been changed to read (new text in **bold**):

> Deforestation in central Africa, for example, is the primary driver of increasing mean annual runoff there **because of reduced evapotranspiration relative to existing vegetation. Note, however, that LandSyMM can only represent the effect of land cover change on evapotranspiration and runoff directly—to include the impact of these flux differences on rainfall would require a coupling with a climate model.**

We have also added the following text in the Methods: "The CMIP5 runs did include land-use change, but not the trajectories output by PLUM. As such, and as with all models that are not climate-coupled but rather use offline forcings, we do not consider the effects of our simulated land-use change on climate."

Line 348-349: I do not understand what the 'fraction of included land area' means and why it shows that not including routing is not a big issue (it sounds disconcerting to me). Please explain better.

In that sentence, "the fraction of included land area in any class" has been changed to "the results for any class".

Line 359-360: why is the estimate more consistent if there is less over application of N? In reality many countries exceed the N uptake rate of plants (most notably China). It sounds less realistic to me that N application only increases by 2%. Does this not imply a major break with historical trends?

The integrated assessment models whose output was used in Krause et al. (2017) actually simulated these very high levels of nitrogen input as resulting in very high yields—i.e., they did not simulate the real-life phenomenon of overapplication. We do not attempt to account for overapplication, either, but in Alexander et al. (2018) we showed that LandSyMM nevertheless does reproduce historical N application levels globally.

To clarify our point here, we have changed the text to read (edited text in **bold**):

> … used fertilizer information from IMAGE and MAgPIE**. Strong increases in fertilizer in those models resulted in strongly increased yields, but nitrogen limitation is alleviated at much lower levels in LPJ-GUESS. IMAGE and MAgPIE fertilization rates thus** often exceeded what plants in LPJ-GUESS could actually take up, resulting in high amounts of N loss. Coupling LPJ-GUESS with PLUM provides for a more internally consistent estimate of future N losses**, while still reproducing historical fertilizer application well (Alexander et al., 2018)**.

Line 369-370: you cannot state 'and other models' here if you refer to three articles that are if I am not mistaken all based on LPJ-GUESS. It is quite logical then that the estimates are similar. Please add an independent reference.

We have replaced this sentence with the following (edited/new text in **bold**):

> Global combined BVOC emissions over 2001–2010 totaled ~546 TgC yr^{-1} (~503 and ~43 TgC yr^{-1} for isoprene and monoterpenes, respectively), which compares well with estimates from LPJ-GUESS **using different land use scenarios (Arneth et al., 2008; Hantson et al., 2017; Szogs et al., 2017) and the MEGAN model (Sindelarova et al., 2014).**

Line 430: I don't think 'storylines' is the right word here as you don't calculate storylines but scenarios that are described by a certain storyline.

We have replaced "storylines" with "scenarios" there. To avoid repetition, in the previous sentence we replaced "scenarios" with "possible futures".

Line 435-439: would it be possible to draw stronger conclusions based on the scenario assumptions on how certain future developments should rather be avoided etc?

To maintain objectivity, we have decided to avoid such prescriptivist language.

SI:

Figure SR2: why are the starting points in 2010 so different for irrigation and fertilizer? This should be historic data I assume? Also, there is hardly any trends in the irrigation results, why is this the case?

This figure (now Fig. SR3) presents PLUM outputs pre-harmonization; we have added text to this effect in the caption. These raw PLUM outputs are not necessarily expected to align exactly with historical data, as evidenced by the need for the harmonization routine. Nor are they necessarily expected to align with each other at the beginning of the period, because of scenario-specific parameters in PLUM derived from the SSPs—particularly regarding the cost of irrigation and fertilizer. Text to this effect has been added to Sect. 2.2 and the caption of what is now Fig. SR3.

If PLUM were to begin with some "historical" parameter values and gradually phase in scenario-specific values, this would improve agreement among scenarios at the beginning of the future run. However, as there is no obviously "correct" way to design this phase-in, we have made the parsimonious decision to apply all scenario-specific parameters at the beginning.

Regarding irrigation, the following text has been added to the end of the second paragraph of Sect. 3.1:

> PLUM prescribes lower irrigation rates by the end of the century for most scenarios (Figs. 2, SR2). This is enabled by higher global mean rainfall in all RCP scenarios, as evidenced by the bars for runoff in Figure 4, as well as by improved water-use efficiency for crops other than C_4 cereals due to increased CO_2 concentrations. Crop demand increase in SSP3-60 outweighs these effects, however, resulting in higher irrigation in that scenario.

Figure SR3: I don't understand the top figure on livestock demand. The order of magnitude makes it likely the results are on crop production, but the title suggests this is total production (?) of livestock products. But the bottom figure shows feed for livestock so if the top figure is also about feed demand for livestock it would not be useful. Please explain and improve.

The former Fig. SR3 (now Fig. SR4) has been simplified to show only the fraction of ruminant food that is provided by feed crops, which is the only information required to support the corresponding assertion in the main text (i.e., that feed becomes much more important in raising ruminant livestock beginning around 2090).

Figure SR8: similar figures are shown for south asia and sub-saharan Africa. Why not consistently show results for all regions? Maybe in slight smaller panels and without the figures with maps in between? Also, why is this figure shown as a delta instead of absolute amounts and how can it be that the demand is so extremely jumpy for oil crops? This seems very unrealistic.

There are an enormous number of possible regions and region groupings that could have such figures made; we present only what is necessary to support specific assertions made in the main text. Similarly, we have presented percentage change rather than absolute amount because the former is more directly explanatory of assertions in the main text. Showing absolute amount would, in some cases, make relevant trends difficult to discern.

The "jumpiness" of individual crops is due to shifts in which crops are used as animal feed. These shifts are due primarily to changes in relative prices of the different crop commodities. Note that the dotted lines, which exclude demand for animal feed, are much more stable. It is indeed unrealistic to expect, e.g., oilcrop production to triple from one year to the next, as would be required to satisfy the demand increase seen in the US and Canada in the early 2040s (Fig. SR9, formerly Fig. SR8). For the purposes of our ecosystem services analysis, however, gross decadal trends in total agricultural area and management inputs are much more important than exactly what is being grown on cropland, and those gross trends are much smoother. Text explaining this has been added to Sect. 2.2.

Figure SR10: please write complete description instead of referring to another figure.

Complete descriptions have been added to the former Figs. SR 5–7, 10, and 13 (now SR 6–8, 11, and 14).