Author response on “A Bayesian Ensemble Data Assimilation to Constrain Model Parameters and Land Use Carbon Emissions”

Sebastian Lienert[1,2] and Fortunat Joos[1,2]

[1]Climate and Environmental Physics, Physics Institute, University of Bern, Bern, Switzerland
[2]Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland.

Correspondence to: Sebastian Lienert (lienert@climate.unibe.ch)

Contents

1 Reply to reviewer S. Zaehle

We want to thank the reviewer for the time and effort for the careful and very insightful review. In the following, we respond to the reviewer point by point, with our **responses in bold** and *quotations from the updated manuscript in cursive*. Please also consider the updated manuscript with track changes in the supplementary. Also note that we expanded the discussion section to include a paragraph on a potential bias in the fossil fuel emissions used for the deconvolution by including non-fuel uses.

Point-by-point response

Lienert and Joos apply a bayesian data assimilation framework to the LPX-Bern model in order to constrain a selection of model parameters using a range of local to global carbon and water cycle observations. In the manuscript, they describe the framework and illustrate the key model performance criteria. This framework allows them to provide a data-constrained simulation of the regional and global terrestrial carbon balance between 1860 and 2016, and in particular to estimate the land-use related carbon emission, including an uncertainty range. This is a very good study integrating multiple observations in a systematic and reproducable way to constrain a process-based global carbon cycle model. This system is not only used to produce a newly calibrated LPX-Bern version for future use, but also to provide useful insight into the magnitude (and particularly the uncertainty) of land-use emissions. Overall this is a valid contribution to Biogeosciences.

Thank you

Unfortunately, I have troubles following the method description. The description of the way, the parameter distributions are updated is remains fairly unclear. I recommend that the authors devote a special section in the Methods section to clarify a couple of points:

A separate section on prior selection is added to the manuscript.

a) how was the prior distribution of the parameters derived (literature ranges typically only allow to assume uniform distributions); b) how exactly is the ensemble updated after the metrics are calculated. Is it the probability distribution of each parameter, which is updates? This would lead to a new LHS set to be produced, and subsequent new model runs? Figure 1 would suggest that this is the case, but in this case, the new set would be dependent on the metric and metric weighting, which contradicts the statements made on P2 (also, it's computationally probably prohibitive). Or is each LHS sample weighted according to the model performance, and this weight then used to calculate the PDF of a modelled output? If that is the case, I don't understand the iterative nature of the LHS sample updates? Also, in this case, it would be good if the authors would elaborate on the way they've estimates the posteriori distributions.

We introduced a new subsection discussing the method used to obtain the prior parameter distribution and reordered the section for clarity. Please refer to section 2.3.2 (p.6 l.9-p.8 l.7 in the manuscript).

Figure 1 in this reply shows the evolution of the median parameter values and ranges of the ensembles with 200 and 300 members (T1-T6) and large ensembles with 1000 members (E1-E3), discussed in the new section in the manuscript. Only the parameters used in the final ensemble E3 are shown. In the small ensembles, different model configuration and parameters were tested. For instance, in ensemble T2 nitrogen limitation was not considered and thus the nitrogen

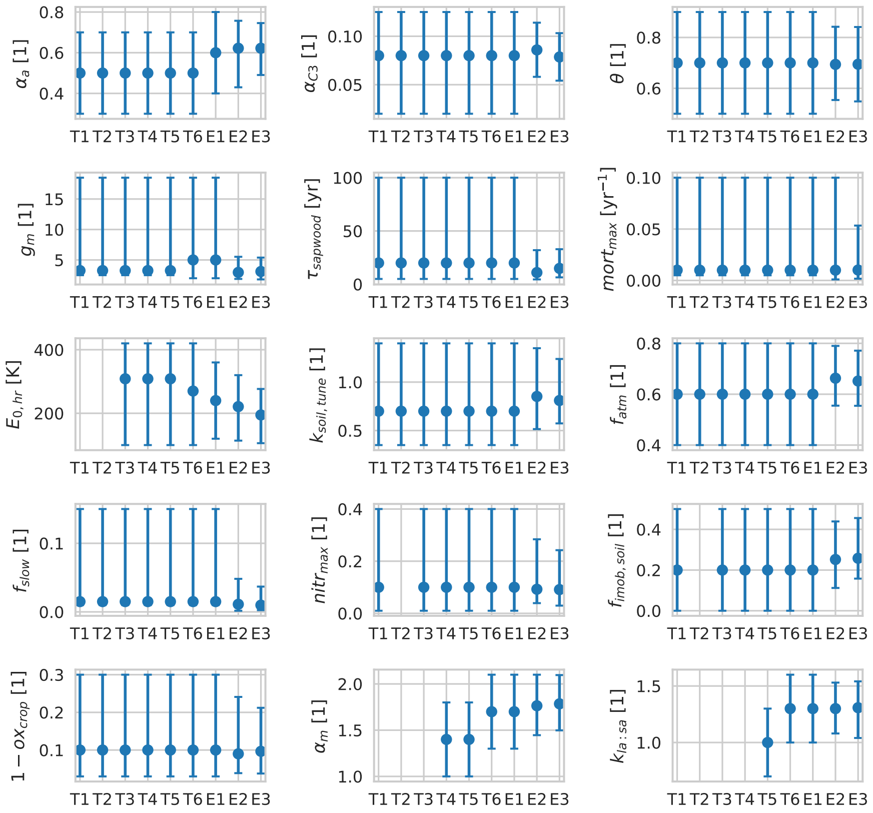

Figure 1. Median and 90% confidence intervals used for the prior distributions of the parameters of 9 ensembles. T1-T6 are ensembles with fewer members and E1 and E2 were precursors of the final ensemble E3.

cycle related parameters were not sampled. The factorial simulation and small ensembles informed the choice for the prior of ensemble E1, which is then iteratively improved to arrive at the prior of E3.

Given that the authors highlight the ability to change the cost-function and weighting as a key strength of their method, it would be also interesting if they would add a discussion point as to how robust they believe the posteriori parameter distributions are against their choice of metric & weighting.

We qualitatively assessed the robustness by reevaluating the ensemble for a subset of the observational targets. For this purpose we created multiple hierarchical weighting schemes, each missing one of the observational targets or a category of targets when compared to the default version, and looked at the induced changes in the parametrization of the best guess version. We added a section discussing this approach to the manuscript:

"We investigate the dependency of the constrained ensemble on the choice of the observational constraints by reevaluating the ensemble for a subset of observations. We created 19 weighting schemes, each missing one of the individual observational constraints (Figure 2 and table 2) and otherwise identical to the default scheme. Then the median skill weighted parameter

values of these ensembles are compared to the best-guess values of $M_{net,net}$ (section 3.3). The relative change in parameterization is less than 1% for 15 out of the 19 considered alternative weighting schemes. Leaving away the global vegetation and soil carbon constraints lead to moderate changes, notably to a change in the parameter for mortality ($mort_{max}$) of 4% and 2% respectively. Not including the soil carbon distribution in high latitudes lead to an increase of the parameter for the dependency of soil respiration on temperature ($E_{0,hr}$) of 2%. The largest changes in parameterization were observed when not considering the atmospheric deconvolution, most notably the sapwood-heartwood turnover time $\tau_{sapwood}$ decreased by 5%. When omitting entire categories in the benchmarking scheme, the changes in parametrization are larger than for omitting individual constraints, with parameter changes of up to 1% for the fluxes, 5% for the inventory and 6% for the transient category. This shows that the final parameterization is not overly sensitive to the inclusion or omission of a single observational product."

I have a number of further suggestions to improve the clarity of the manuscript: P1 L17: in the context of a data assimilation paper, the use of assimilated here is confusing. replace by stored?

Done

P2 L3: Add "Amongst others," at the beginning of the sentence

Done

P2 L4: unclear what uncertain prescribed LULCC processes are meant to be, perhaps give examples, or clarify that it's the representation of these processes that is uncertain

Done, the sentence now reads:

"In addition to uncertainties in the prescribed LULCC forcings and the representation of LULCC and other processes in DGVMs, the values of the applied parameters are subject to substantial uncertainties."

P2 L9 DA "should" be an integral part of model development, but unfortunately it is not always.

Replaced "is" by "should be".

P2 L10: Is the Houwelling reference appropriate here? This does mostly relate to inverse atmospheric modelling

Removed Houweling reference

P2 L12: Not sure that I understand sequentially correct here. Most DA methods would assimilate different data sources simulateneously. Also, I think cost-function is the more common term for metric in this context

Revised sentence to read: *"A drawback of these methods is that the sampling process is dependent on the choice of the cost function, the design of which is not trivial when assimilating multiple observations simultaneously."*

P2 L14 This sentence is a bit out of context in a paragraph on alternative DA methods, because benchmarking does in general not imply DA. It seems more logical to merge this sentence with the Paragraph starting in L25, and move the entire paragraph to L6 after Le Quere et al. 2016.

Moved, the paragraph now reads:

"Amongst others, Dynamic Global Vegetation Models (DGVMs) can be used to assess the contribution of LULCC to the terrestrial carbon budget (Le Quéré et al., 2016). The assessment of the performance of a given model version using observational

benchmarks has been actively discussed in the literature (Hoffman et al., 2017; Peng et al., 2014; Kelley et al., 2013; Luo et al., 2012; Blyth et al., 2011; Randerson et al., 2009) and different frameworks have been proposed. The selection of observational targets is vital to a successful assimilation of observational data. In order to constrain the contemporary carbon cycle, 14 data products are used, ranging from global inventories of carbon (Ciais et al., 2013) to spatially resolved satellite estimates of photosynthetically absorbed radiation (Gobron et al., 2006). The goal of the data set selection process was to have observations capturing the magnitudes of fluxes and inventories in the carbon cycle, as well as its transient response to anthropogenic perturbance. In addition to uncertainties in the prescribed LULCC forcings and the representation of LULCC and other processes in DGVMs, the values of the applied parameters are subject to substantial uncertainties. We use a Monte-Carlo-like data assimilation approach (Steinacher et al., 2013; Steinacher and Joos, 2016; Battaglia and Joos, 2017) to sample 15 key model parameters and construct a 1000-member model ensemble to investigate this parameter related uncertainty in the DGVM LPX-Bern. Furthermore, we establish a new reference version of the model."

L2 L18: As noted above, I have troubles following here: LHS simply provides a set of parameter combinations, in which each parameter is sampled given a specified distribution and notably, ensuring that there is no correlation amongst any of the parameters. LHS does not imply any model metric per se. The way the posterior distribution is derived from the prior distribution and the model metrics is unclear. How many iterations would be needed to arrive at a stable solution, what is the stopping criteria, and why is it possible to change the metric during the DA procedure? This would change the posterior distribution, and therefore impede convergence.

Please see the answer to the major points.

P2 L31: I wonder if the flow of the introduction would be more logical if one would first talk about the LULCC processes as in this paragraph, then about the benchmarking in the preceding paragraph, and only then about data assimilation?

We have restructured the introduction but slightly deviated from the reviewers suggestion to improve text flow. Please also see the attached manuscript with track changes.

P2 L31: While the (add) “net” land-atmosphere flux can “to some extent” be . . .

Done

P2 L32: add “residual” terrestrial carbon sink?

Done

P4 L21: I think it is worth highlighting that the strength of LHS over other sampling techniques is that the set of parameters in uncorrelated.

Done:

“..to generate an uncorrelated parameter ensemble of a given size.”

P4 L20: The text confuses MC parameter sampling techniques, which are indepedent of any purpose the sampling is made for, from MC Data assimilation techniques, which are not?

Changed “Monte Carlo sampling techniques” to “Monte Carlo data assimilation techniques”

P4 L26: the description of alpha_a should correspond to table 1, it is not FAPAR!

Done:

"The fraction of photosynthetically active radiation assimilitated at ecosystem level relative to leaf level, α_a .."

P4 L7: Literature range only allow to give uniform distribution. How where the nonuniform distribuition parameters obtained / estimated?

See answer to major points and attached manuscript.

P4 L7: I have trouble following from here on. Maybe this would become clearer, if first all the metrics and data sources were explained, and then the way the distributions are updated is clearer presented.

Revised this section (See answer to major points)

P8 L10: which winds were used for the transport? I assume that the winds were not interannually varying?

Yes, the transport matrix does not include interannual variability. Added:

This method does not include the interannual variability of the transport.

Additionally, we added an explanation to Figure 8 for clarity:

"As expected, the interannual variability in seasonal amplitude of CO_2 is not captured as the atmospheric transport model TM2 does not represent interannual variability in mass transport."

L9 L5: Inversion typically refers to the inverse modeling of atmospheric transport, whereas here - as far as I understand this, you simply take the land flux as the residual of the fossil fuel emission and ocean uptake.

We changed all occurrences of inversion to deconvolution.

P9 L10: Are these data sources not redundant with the global maps of total and soil C storage described earlier?

While the information of the global carbon content is also contained in the maps, we feel the inclusion of the additional, well established, global target is warranted by the importance of these targets. This is effectively increasing the weight of these targets.

P9 L27: I don't understand the reasoning for the duplication of ensembles with gross transitions. Please motivate.

We did not repeat the procedure to improve the prior distribution (See updated manuscript) for $\mathbf{M}_{gross,gross}$ and as such the prior and posterior distributions do not converge. Consequently we do not feel comfortable to use $\mathbf{M}_{gross,gross}$ as the basis for our estimates for $\mathbf{E}_{\mathrm{LUC}}$. As a compromise we introduced $\mathbf{M}_{gross,net}$, retaining the confidence in the performance of $\mathbf{M}_{net,net}$ and simply adding the important processes of shifting cultivation and wood harvest.

P9 L31: As noted above, I have difficulties following this description.

Section revised completely, see the answer to major points.

P10 L5-8: Is material for the introduction, not the results section

Removed Paragraph

P10 L 8-11 can be safely removed.

Removed Paragraph

Section 3: When giving numerical estimates, please add either range or standard deviation, whenever the number is based on the ensemble. I also think that the more logical arrangement of the Results sections would be to first talk about model performance, and then about the attribution of the net land flux to LULCC and residual.

Added the skill weighted 90% confidence interval for every reported number, except when reporting the median difference between two ensemble configurations. We agree that the suggested order of the result section is more logical, however we feel that the results on LULCC are of more interest to a broader range of readers, and thus prefer to lead with those results.

P13 L4: Why would an underestimation of the ELUC not affect your conclusions about ELUC?

The net land-atmosphere flux is underestimated because E$_{gross,net}$ features additional processes that lead to an increase in E_{LUC}, while the residual land sink remains constant. However if only consideringE_{LUC} we expect the magnitude of the residual land-sink and net land-atmosphere flux to be less important than for instance model performance in the vegetation carbon benchmarks (Li et al., 2017). For clarity we revised the sentence:

"A caveat of this choice is that the net land-atmosphere flux is underestimated in $M_{gross,net}$ because the residual land sink only responds to the lower E_{LUC} of $M_{net,net}$. However if only considering E_{LUC} we expect the magnitude of the residual land-sink and net land-atmosphere flux to be less important than model performance with respect to vegetation carbon (Li et al., 2017) and other benchmarks."

P17 L6: is the use of the word significant appropriate here?

Changed 'significant' to 'relevant'

P17 L8: why not?

Using the vegetation carbon distribution directly, would have been a valid choice. Exchanging the total carbon distribution for the vegetation carbon distribution in the hierarchical weighting scheme reveals that the median parameter values used for the best guess version change less than 0.5%.

P 18 L 3: Why is this different from the approach described in Section 2?

Sentence shortened and clarified to read:

"We compare the total land-atmosphere exchange flux to the results of the atmospheric CO$_2$ deconvolution in Fig. 11"

Conclusion Section: There is no need to repeat details of the methods or approach undertaken

Shortened conclusion section by removing sentences which go into too much detail.

Figure 1: Ensure all lines are visible

Adjusted legend

Table 1: Check units and definition for E0. This seems more like an activation energy to me (not a temperature sensitivity What are the units of the k_la:sa? Is this simply a scalar?

E$_0$ is defined according to Lloyd and Taylor 1994, which considers the effect of an activation energy which is varying with temperature. Using their representation it has the unit [K] and is strictly speaking not an activation energy. As such

we find the definition appropriate. k$_{la:sa}$ scales the PFT dependent leaf area to sapwood area and is as such unitless. The leaf area to sapwood area has the units [m^2/m^2]

2 Reply to reviewer J.-F. Exbrayat

We want to thank the reviewer for the time and effort for the careful and very insightful review. In the following, we respond to the reviewer point by point, with our **responses in bold** and *quotations from the updated manuscript in cursive*. Please also consider the updated manuscript with track changes and a high-resolution figure for this reply in the supplementary. Also note that we expanded the discussion section to include a paragraph on a potential bias in the fossil fuel emissions used for the deconvolution by including non-fuel uses.

Point-by-point response

The study presents an approach to constrain a DGVM with multiple observational streams of carbon stocks, gross and net fluxes. The authors rely on a latin hypercube stratified sampling to perturb model parameters and create several 1,000-member ensemble simulations of the terrestrial carbon cycle for the historical period. Results focus on the estimation of land-use and land-cover change emissions. This study is quite innovative in the context of the global terrestrial carbon cycle as model parameters are constrained globally.

Thank you

I have found several similarities between the method described here and the Generalised Likelihood Uncertainty Estimation C1 method used in hydrological sciences (Beven and Binley 1992).

We have added a reference to Beven and Binley in the introduction:

"Other approaches have also been investigated, such as using generalized likelihood function for model calibration and uncertainty estimation (Beven and Binley, 1992)"

First, my main criticism targets the description of the sampling method. It is very unclear how the prior probability distribution in Figure 1 and the new best-guess values in Table 1 have been obtained, and how the posterior distribution of the parameters is calculated. Is it based on the selection criterion used to exclude the less skilled model parameters (p7 l5-8)? If Figure 1 and Table 1 present results from the current manuscript they should be described in the corresponding section.

We restructured the method section and introduced a new subsection describing the explorative approach used to obtain the prior distribution. Additionally, we clarified the procedure to arrive at the posterior distribution in Section 2.3.2 (p.6 l.9-p.8 l.7 in the manuscript with track changes).

Figure 1 in this reply shows the evolution of the median parameter values and ranges of the ensembles with 200 and 300 members (T1-T6) and large ensembles with 1000 members (E1-E3), discussed in the new section in the manuscript. Only the parameters used in the final ensemble E3 are shown. In the small ensembles, different model configuration and parameters were tested. For instance, in ensemble T2 nitrogen limitation was not considered and thus the nitrogen cycle related parameters were not sampled. The factorial simulation and small ensembles informed the choice for the prior of ensemble E1, which is then iteratively improved to arrive at the prior of E3.

Second, I struggle to understand what experiments were actually undertaken. From section 2.6, it seems that three simulations are performed for each parameter set. These three simulations differ in the representation of LULCC: none (reference), gross or net transitions. Then, the results section reports the three different model configurations Mnet,net, Mgross,net, Mgross,gross while these are first described as three alternative skill weighted median.

To clarify we added the following text to section 2.5 (Former 2.6):

"For each of the parameter sets 4 transient simulations over the industrial period are performed: (i) a simulation with prescribed net transitions ($M_{net,net}$ and $M_{gross,net}$), (ii) a simulation with prescribed gross transitions ($M_{gross,net}$ and $M_{gross,gross}$), (ii) a run with landuse area fixed at preindustrial levels and (iv) a run with landuse including shifting cultivation held at preindustrial levels. The last two simulations are used purely diagnostic to determine E_{LUC}."

Third, I am unclear about the skill-weighted mean method. Simulations with either net or gross land-use configuration are likely to yield different results so it is hard for me to justify Mgross,net. I understand that the Mgross,net skill-weighted mean provides the best results compared to benchmarks (Table 3) but it could be an artefact, couldn't it?

As now explained in the revised MS (see our answer above), we did not perform the procedure for optimizing the prior distribution for M$_{gross,gross}$. The prior and posterior distributions of this configuration do not converge and as such we feel not confident in using it as the basis for our estimates for E_{LUC}. However, it is clear from literature that processes such as shifting cultivation and wood harvest are an important component of landuse change. As a compromise we use the optimized M$_{net,net}$ ensemble with the additional gross transition processes added, without the retuning of the model parameters.

Also, some parameter sets are likely to perform better in some regions and worse in other. Therefore, would a spatially-explicit weighting scheme (Schwalm et al., 2015; Exbrayat et al. 2018) be more suited to constrain the ensemble?

The use of spatially dependent parametrization offers numerous advantages, which include the potential to yield better performance with regard to observational data. However we believe that assessing the performance of an individual model using global parametrization, can still provide valuable insight in the terrestrial carbon cycle, as a potential caveat of regional parametrization are the additional degrees of freedom which could potentially lead to an over-fitting of the problem. We have added the following text to the discussion:

"An other avenue of increasing model performance is to introduce spatially explicit parametrization, as recently used in multi-model averaging studies (Exbrayat et al., 2018; Schwalm et al., 2015). A caveat of using this approach with a single model is a potential overfitting of the parameters."

Hereafter are some more specific comments p4 l6: CRU TS3.23 covers 1901-2014, so how are simulations performed for 1800-2014 (or is it 1800-2016 like in the abstract?) please clarify throughout the manuscript

Simulations are performed from 1800 to 2016 with recycled climate data from 1901-1930. Corrected wrong period 1901-2014 to 1901-2016 and changed the wrong reference from CRU TS3.23 to CRU TS3.25 (1901-2016). The recycling of the climate data is described at the end of section 2.2.

p4 l21: please define what model metrics

Added specification in sentence:

"..,the sampling is independent of the metrics used to assess model performance,.."

p5 l6: how have these distributions been chosen?

Please see update to sampling description.

p7 l1: please write MSEi rel

Done

p9 l28: 'LULUC'? please correct here and in several other places

Done, corrected LULUC to LULCC throughout the text

p10 l13-23: please include some information about the uncertainty displayed in the Figures here and throughout the text

We now report the skill weighted 90% confidence interval throughout the text, except for differences between different ensemble configurations.

p11 l1: please quantify 'slight'

The uptake from 1980 to 2016 amounts to 2.6 PgC. We now report the interval 1990-2016 and revised the sentence to read:

"The resulting total change in land carbon is negative, with a slight uptake of carbon at the end of the century, amounting to 9.3 (-0.9,22.2) PgC between 1990 and 2016"

p14 l11: see previous comment on the study period

The simulation spans 1800-2016, however spatial output was only saved after 1901 due to storage limitations.

p16 l3: an informative figure would a covariance matrix of the parameter sets' scores for each criterion

Figure 2 in this reply shows plots of the skill in individual observational targets for all parameter sets. A striking feature is the high correlation of the skill in total carbon map with skill in soil carbon map, which is not unexpected. There is no scatter plot with a lack of points in the upper right corner, i.e. no observational constraints are mutually exclusive. While we agree that this figure is informative, the sheer size and number of subplots make an inclusion in the manuscript or supplementary difficult. Please note that a version of the figure in vector graphic format is included in the supplementary of this reply.

p16 l12: according to Figure 8b and d, the model captures the seasonality but not the interannual variability. This is worth reporting (and explaining).

The interannual variability is not captured because the transport model used does not feature winds with interannual variability. Added sentence:

"As expected, the interannual variability in seasonal amplitude of CO_2 is not captured as the atmospheric transport model TM2 does not represent interannual variability in mass transport."

Fig 1: Mnet,net is not defined

Added:

"... ensemble with net land-use ($M_{net,net}$)"

Fig 3: please explain the sign convention as it seems at odd with figure 4 (ELUC in particular)

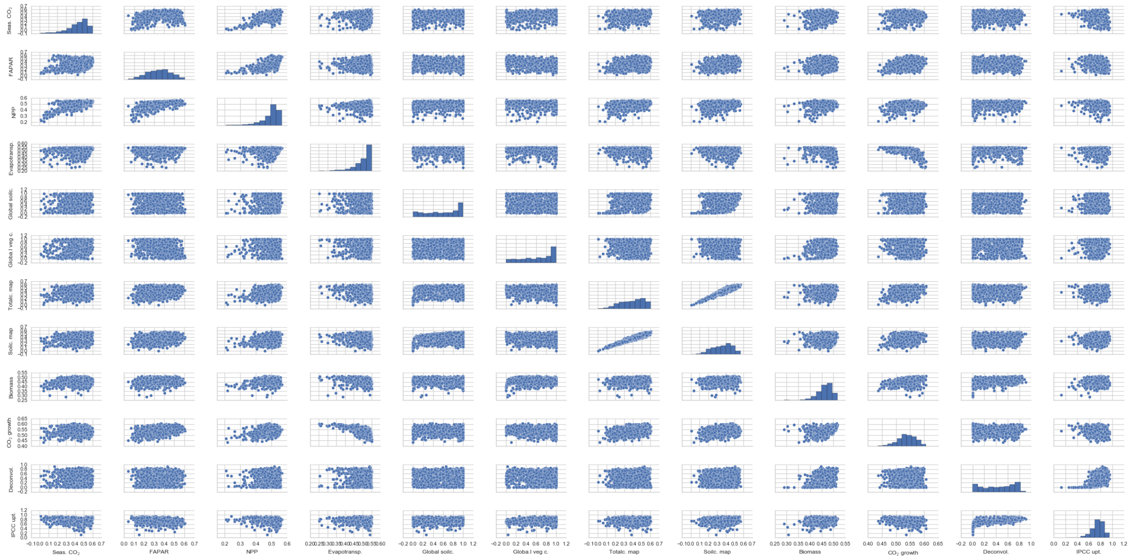

Figure 2. Skill in observational targets for all parameter sets. The diagonal shows a histogram of the skills for the targets, the off-diagonal shows the skill of two observational datasets in a scatter plot. A high resolution version of this plot is attached to the reply in the open discussion.

We updated Figure 4 to show a release of carbon to the atmosphere due to LULCC as positive, which is consistent with Figure 3 (And the rest of the text). Updated the figure caption and text to be consistent with this change.

Fig 7: this figure is very complicated. Why is it important to look at the whole ensemble, and the constrained one? Constraining the ensemble uncertainty is not a major point in the rest of the manuscript and uncertainties are not reported in most of the text.

We expanded the section explaining the constraining process (See answer above) and added additional confidence intervals for the numeric results. We revised the text in the first paragraph of 3.2 to better explain the figure:

In Fig. 7 a mapping of the MSE_{rel} to an individual skill score is displayed for the observational data-sets with a spatial structure, to demonstrate how well the median of the ensemble and the new version LPX v1.4 are able to simulate individual observations. The figure also demonstrates the success of the assimilation process: the skill scores for many individual targets are improved in the ensemble median and LPX v1.4 compared to LPX v1.2, the starting point of our work. As a consequence of our iterative prior selection (section 2.3.2) the median skill for an individual constraint is similar in the constrained ensemble compared to the unconstrained ensemble. In all but the fAPAR benchmark the skill is consistently higher than the minimum skill criterion. With the exception of the biomass measurements by (Keith et al., 2009) and the fAPAR benchmark, the maximum skill in the constrained ensemble is identical to the full ensemble. The reduced maximum skill in those benchmarks is due to an exclusion of singular runs excelling at this benchmark but performing badly in others.

Fig 8: please move the legend

Done

3 List of changes

Following is a short overview of the most important changes in the manuscript. Please also consider the more detailed point-by-point replies above and the manuscript with track changes below.

- Introduction: Improved text flow and added additional references.
- Method section - Model setup and spinup: Corrected used climate forcing
- Method section - Sampling and Constraining: Extensive changes: restructured, improved text and added new subsection “Selection of the prior distribution”.
- Method section - Observational constraints: Renamed the atomspheric deconvolution from inversion to deconvolution throughout the text.
- Method section - Definition of Land-Use emissions and the setup of the model ensembles: Clarified simulations performed.
- Result section: 90% confidence interval reported on numerical results in the text. Slight changes in the median value are due to changing to taking the ensemble median after performing the temporal summation/average on individual simulations instead of calculating the median for every year and then performing the temporal aggregate for consistency.
- Result section - Evaluation of ensemble performance with respect to observational targets: Improved description of Figure 7 in the text.
- Result section - Evaluation of ensemble performance with respect to observational targets: Added paragraph investigating the dependency of the constrained ensemble on the selection of observational constraints.
- Discussion - Benchmark performance and best guess version: Added new paragraph discussing the potential influence of neglecting non-fuel uses of fossil fuel on our atmospheric deconvolution.
- Discussion - Benchmark performance and best guess version: Added new paragraph mentioning spatially explicit parametrization.
- References: Removed superfluous URLs in addition to DOIs.
- Various small changes throughout the text, such as corrected spelling errors or moved legends in plots, please also consider the manuscript with track changes and the replies to the reviewers.

4 Manuscript with track changes

A Bayesian Ensemble Data Assimilation to Constrain Model Parameters and Land Use Carbon Emissions

Sebastian Lienert[1,2] and Fortunat Joos[1,2]

[1]Climate and Environmental Physics, Physics Institute, University of Bern, Bern, Switzerland
[2]Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland.

Correspondence to: Sebastian Lienert (lienert@climate.unibe.ch)

Abstract.

A dynamic global vegetation model (DGVM) is applied in a probabilistic framework and benchmarking system to constrain uncertain model parameters by observations and to quantify carbon emissions from land-use and land-cover change (LULCC). Processes featured in DGVMs include parameters which are prone to substantial uncertainty. To cope with these uncertainties Latin Hypercube Sampling (LHS) is used to create a 1000-member perturbed parameter ensemble, which is then evaluated with a diverse set of global and spatio-temporaly resolved observational constraints. We discuss the performance of the constrained ensemble and use it to formulate a new best-guess version of the model (LPX-Bern v1.4). The observationally constrained ensemble is used to investigate historical emissions due to LULCC (E_{LUC}) and their sensitivity to model parametrization. We find a global E_{LUC} estimate of 158 (108, 211) PgC (median and 90% confidence interval) between 1800 and 2016. We compare E_{LUC} to other estimates both globally and regionally. Spatial patterns are investigated and estimates of E_{LUC} of the ten countries with the largest contribution to the flux over the historical period are reported. We consider model versions with and without additional land-use processes (shifting cultivation and wood harvest) and find that the difference in global E_{LUC} is on the same order of magnitude as parameter induced uncertainty and in some cases could potentially even be offset with appropriate parameter choice.

1 Introduction

Due to constraining atmospheric CO_2 concentrations and the relatively well known CO_2 sink in the ocean it follows that about a fifth of anthropogenic CO_2 emissions is ~~assimilated by~~ stored in the terrestrial biosphere (Ciais et al., 2013). However, the partitioning of this land-atmosphere flux to effects from human-induced land-use and land-cover change (LULCC) and the transient change of the residual terrestrial sink remains highly debated (Schimel et al., 2015). It is estimated that approximately a third of the cumulative anthropogenic CO_2 emissions in the industrial period stem from the effects of LULCC (Arneth et al., 2017; Brovkin et al., 2013; Gerber et al., 2013; Houghton and Nassikas, 2017; McGuire et al., 2001; Mahowald et al., 2017; Pongratz and Caldeira, 2012; Sitch et al., 2015; Strassmann et al., 2008; Stocker et al., 2017, 2014; Peng et al., 2017). A better understanding of the mechanisms of the historical terrestrial carbon cycle is vital for more accurate future projections of the global carbon cycle and climate. In addition,a better understanding of the residual terrestrial sink can also help to improve our understanding of the terrestrial carbon cycle of the past, unperturbed by human influence.

Amongst others, Dynamic Global Vegetation Models (DGVMs) ~~can be used to assess~~ are used to quantify the contribution of LULCC to the terrestrial carbon budget (Le Quéré et al., 2016). The assessment of the performance of a given model using observational benchmarks is actively discussed in the literature (Hoffman et al., 2017; Peng et al., 2014; Kelley et al., 2013; Luo et al., 201 different frameworks have been proposed. In addition to ~~uncertain~~ uncertainties in the prescribed LULCC forcings and the representation of LULCC and other processes in DGVMs, the ~~parametrization of those processes is subject to further~~ values of the applied parameters are subject to substantial uncertainties. We use a Monte-Carlo-like data assimilation approach ~~(Steinacher et al., 2013; Steinacher and Joos, 2016; Battaglia and Joos, 2017)~~ (Steinacher et al., 2013; Steinacher and Joos, 2016; Battagli sample 15 key model parameters and construct a 1000-member model ensemble to investigate this parameter related uncertainty in the DGVM LPX-Bern ~~. Furthermore, we~~ and establish a new reference version of the model. 14 data products are used as observational constraints. These range from global inventories of carbon (Ciais et al., 2013) to spatially resolved satellite estimates of photosynthetically absorbed radiation (Gobron et al., 2006). The goal of the data set selection process is to have observations capturing the magnitudes of fluxes and inventories in the carbon cycle, as well as its transient response to the anthropogenic perturbation.

The assimilation of observations ~~is~~ should be an integral part of model development. Various approaches to incorporate constraining data exist, such as variational approaches minimizing a cost function using the adjoint of the model ~~(Houweling et al., 1999; Kato e~~ the use of ensemble Kalman filters ~~(Houweling et al., 1999; Lorenc, 2003; Gerber and Joos, 2013; Stöckli et al., 2011; Ma et al., 2017). However since these methods operate sequentially, the metric and relative weighting of observational products have to be chosen beforehand, which can prove difficult if multiple observational data sets are to be assimilated simultaneously. The assessment of the performance of a given model version using observational benchmarks has also been actively discussed in the literature (Hoffman et al., 2017; Peng et al., 2014; Kelley et al., 2013; Luo et al., 2012; Blyth et al., 2011; Randerson et al., 2009) and different frameworks have been proposed~~(Lorenc, 2003; Gerber and Joos, 2013; Stöckli et al., 2011; Ma et al., 2017). A drawback of these methods is that the sampling process is dependent on the choice of the cost function, the design of which is not trivial when assimilating multiple observations simultaneously. Other approaches have also been investigated, such as using generalized likelihood function for model calibration and uncertainty estimation (Beven and Binley, 1992). Here we employ the Latin Hypercube Sampling (LHS) (McKay et al., 1979) approach, as used successfully in previous studies ~~(Steinacher et al., 2013; Battaglia~~ It allows simultaneous stratified sampling of a range of parameters, given an appropriate prior parameter distribution, while offering the opportunity to change evaluation metrics a posteriori, thus enabling a sensible incorporation of multiple observational constraints. By improving the prior distribution iteratively it is possible to reasonably capture observations while considering trade-offs between the different targets. Additionally, this approach not only yields a best-guess of parameter values but also contains information about the associated uncertainties. A drawback of this technique is that it is not possible to increase the size of the ensemble after the initial sampling and if the range of the prior distribution is too large the algorithm has decreased computational efficiency.

~~The selection of observational targets is vital to a successful assimilation of observational data. In order to constrain the contemporary carbon cycle, 14 data products are used, ranging from global inventories of carbon (Ciais et al., 2013) to spatially resolved satellite estimates of photosynthetically absorbed radiation (Gobron et al., 2006). The goal of the data set selection process was to have observations capturing the magnitudes of fluxes and inventories in the carbon cycle, as well as its transient response to anthropogenic perturbance. The inclusion of further independent observations in this framework would further serve to reduce parameter uncertainty.~~

While the land-atmosphere carbon flux can to some extent be constrained by the other components of the global carbon cycle, the contribution of LULCC, and in turn the implied residual terrestrial carbon sink, are highly uncertain. Efforts to fill this knowledge gap have been made using bookkeeping approaches (Houghton et al., 2012; Hansis et al., 2015; Houghton and Nassikas, 2017) and bottom-up modeling approaches using DGVMs (McGuire et al., 2001; Stocker et al., 2014; Wilkenskjeld et al., 2014; Sitch et al., 2015). Bookkeeping models can offer valuable information on the magnitude of regional and global LULCC emissions (E_{LUC}), but they typically rely on time-invariant estimates of carbon densities and thus neglect the direct impact of climate change on vegetation. Observational data on carbon densities and response of the vegetation to LULCC effects can be directly incorporated in bookkeeping models. In contrast, DGVM model studies are able to produce highly resolved spatial results and consider changes to vegetation structure due to anthropogenic perturbance, but DGVMs have large uncertainties due to differences in process modeling and parametrization. Additionally, a number of LULCC processes are often neglected, such as the effect of gross land-cover transitions (shifting cultivation), management (wood harvest) or erosion. Studies investigating these processes generally have found that the inclusion of those processes leads to an increase in E_{LUC} (Arneth et al., 2017; Wilkenskjeld et al., 2014; Stocker et al., 2014). On the other hand, neglected processes such as human-induced erosion can have the opposite effect and reduce net E_{LUC} (Kosmas et al., 2007; Billings et al., 2010; Hoffmann et al., 2013; Wang et al., 2017). The effect of parameter uncertainty on these estimates is often only considered indirectly in the intercomparison of models. Here we investigate a parameter ensemble of a single DGVM, constrained by observation and provide direct estimates of parameter induced uncertainties in LULCC estimates. These uncertainties are put into context by investigating the effect of additional LULCC processes, such as shifting cultivation and wood harvest, as already investigated in previous studies (Stocker et al., 2014; Wilkenskjeld et al., 2014; Shevliakova et al., 2009). We rely here on the LUH2 v2h (Hurtt et al., 2018) land-cover data to force the DGVM LPX-Bern v1.4.

2 Methods

2.1 LPX-Bern

The Land Surface Processes and eXchanges (LPX-Bern) model (Spahni et al., 2013; Stocker et al., 2013; Keller et al., 2017) is a Dynamic Global Vegetation Model (DGVM) based on the Lund-Potsdam-Jena (LPJ) model (Sitch et al., 2003). It features coupled nitrogen, water and carbon cycles and distinguishes between different types of prescribed land-use classes: natural vegetation, peatland, cropland, pasture and urban land. The vegetation composition for a given land-use class is determined dynamically. Different plant functional types (PFTs), with given bioclimatic limits, compete for resources. Here ~~9~~ 8 tree PFTs and 2 herbaceous PFTs are used to describe natural vegetation, the same two generic herbaceous PFTs grow on pasture and cropland, and two moss PFTs, two flood tolerant tropical PFTs, and a flood-tolerant herbaceous PFT grow on peatlands.

Two different configurations are used to treat the transition between different classes of land-use. The simpler implementation adjusts the fractional land-use cover at the end of each year such that the prescribed area fractions are achieved, this computationally efficient configuration is referred to as net land-use. The more advanced gross land-use implementation also includes effects of shifting cultivation and wood management by prescribing all the transitions between different land-use classes and harvested wood (Stocker et al., 2014; Strassmann et al., 2008). Furthermore, it includes an additional land-use class, the so-called secondary forest, natural vegetation growing on abandoned pasture or cropland. A major drawback of this scheme is the significantly increased computational cost. Additionally, the implementation of gross land-use in LPX-Bern in the current version does not allow for the simultaneous simulation of peatlands. For both schemes a fraction oc_{frac} of the crops above-ground biomass is directly oxidized to the atmosphere, simulating crop harvest. 75% of heartwood and sapwood biomass production from forest conversion is assigned to decaying product pools, the remaining 25% are respired directly to the atmosphere as assumed harvest losses. Associated root and leaf mass are transferred to an below and above ground litter pool respectively. The biomass in the product pools is evenly split in a long-lived (mean lifetime 20 years) and a short-lived (mean lifetime 2 years) pool. In the gross LULCC setup biomass is harvested according to the prescribed forcing and the resulting heartwood is assigned to product pools using the same allocation rules as before.

2.2 Model setup and spinup

The model is run on a 1° x 1° global grid and forced with CRU TS3~~.23~~ .25 climate data (Harris et al., 2014) and global atmospheric CO_2 concentration from ice core reconstructions (Meure et al., 2006; Joos and Spahni, 2007) and direct atmospheric measurements after 1958 (Tans and Keeling). The Land-Use Harmonization LUH2 v2h (Hurtt et al., 2018) estimates for land-use patterns and transitions are prescribed to the model. Additionally nitrogen deposition (Lamarque et al., 2013) and fertilization (Zaehle et al., 2011) and the extent of northern hemisphere peatlands (Tarnocai et al., 2009) are prescribed. As described in section 2.3 we use an ensemble approach featuring 1000 simulations with different parameters. All ensemble members share a 1500 year spin-up run to pre-industrial conditions, using the median parameter values. To ensure the equilibration of each member an additional 300 year individual spin-up run, featuring an analytical equilibration of the soil carbon pools after 100 years, is performed. The model is then run transiently from 1800 to ~~2014~~ 2016 with recycled climate data (years 1901-1930) in the 19^{th} century.

2.3 Sampling and Constraining

The model parameter space is sampled using Latin Hypercube Sampling (LHS) (McKay et al., 1979) to create an ensemble of model configurations and assess model uncertainty. LHS is a stratified sampling method using chosen prior parameter distribution to generate ~~a~~ an uncorrelated parameter ensemble of a given size. In contrast to most Monte Carlo ~~sampling~~ data assimilation techniques, the sampling is independent of the ~~model metrics~~ metrics used to assess model performance, allowing to modify the metrics ~~without large~~ after the sampling without substantial computational effort. A drawback of this sampling

Table 1. Description of sampling parameters with values for LPX v1.2 and the new best guess version v1.4. If not otherwise indicated parameters are unitless.

Parameter	Description	LPX v1.2	LPX v1.4
α_a	Fraction of PAR assimilitated at ecosystem level relative to leaf level	0.5	0.6175
α_{C3}	Intrinsic quantum efficiency of CO_2 uptake in C3 plants	0.07	0.07660
θ	Co-limitation shape parameter	0.7	0.6937
g_m	Canopy conductance scaling parameter for water demand calculation	3.24	3.120
α_m	Priestley-Taylor coefficient in water demand calculation	1.394	1.786
$\tau_{sapwood}$	Sapwood-heartwood turnover time [yr]	20	15.33
$k_{la:sa}$	Allometric scaling parameter: leaf area to sapwood area	1.0	1.310
$mort_{max}$	Asymptotic maximum in mortality equation [yr^{-1}]	0.01	0.01016
$E_{0,hr}$	Temperature sensitivity of heterotrophic respiration [K]	308.56	190.16
f_{atm}	Fraction of litter entering atmosphere directly	0.6	0.6503
f_{slow}	Fraction of litter entering slow soil pool	0.015	0.009512
$k_{soil,tune}$	Tuning factor for soil decay	0.7	0.7965
$nitr_{max}$	Maximum nitrification rate	0.1	0.09096
$f_{imob,soil}$	Nitrogen imobilization in soil	0.0	0.2639
$1-ox_{crop}$	Fraction of direct oxidation of leaf turnover on cropland	0.1	0.09920

strategy is that it is not possible to increase the size of the ensemble after the initial sampling. ~~Kernel density estimations of the prior probability distribution (blue) and the posterior probability distribution of the constrained ensemble (M$_{net,net}$). The prior distribution was improved iteratively, resulting in near convergence of prior and posterior distribution. Vertical black bars indicate the parameter values used in LPX-Bern v1.2.~~

~~The~~ The generated ensemble is then constrained using an hierarchical weighting scheme of deviations to observational data sets to obtain a global skill score, rating each model member.

2.3.1 Description of the sampling parameters

Table 1 lists the selected sampling parameters ~~and their assumed prior distribution are given in Table 1~~as well as their old values in LPX v1.2 and new best guess values (LPX v1.4). The parameters were selected for their importance in various aspects of the model, 10 of the 15 parameters were also used by (Steinacher et al., 2013). The ~~parameter selection was further guided by an earlier study by Zaehle et al. (2005), investigating the relative importance of 36 parameters in the LPJ model. The~~ fraction of photosynthetically active radiation assimilated at ecosystem level relative to leaf level, α_a, the intrinsic quantum efficiency of CO_2 uptake in C3 plants, α_{C3} and θ the rubisco co-limitation shape parameter are of primary importance for the photosynthetic carbon assimilation. g_m and α_m are parameters in the empiric water demand calculation and have a direct impact on the hydrological cycle and consequently also the carbon assimilation. The sapwood-heartwood turnover time, $\tau_{sapwood}$, the maximum mortality parameter, $mort_{max}$, and the ratio between leaf area and sapwood area, $k_{la:sa}$, are vital for the allocation of the carbon to the different vegetation pools and thus also the overall vegetation carbon pool size. The fraction of the flux leaving the litter pools that is respired to the atmosphere directly and entering the slow soil pool, f_{atm} and f_{slow} influence soil and litter carbon inventories. These pools are further controlled by the temperature sensitivity of heterotrophic respiration $E_{0,hr}$, which is of special significance under changing climate, and a scaling factor for soil decomposition $k_{soil,tune}$, affecting the residence time of both the fast and the slow soil carbon pool. By using factorial simulations two important parameters for the nitrogen cycle were identified, the maximum nitrification rate, $nitr_{max}$, and the fraction governing immobilization of mineral nitrogen in the soil, $f_{imob,soil}$. Finally, the oxidation rate of crops ox_{crop}, representing the harvest of biomass on croplands, is directly linked to emissions from human land-use.

~~Each prior parameter distribution is sampled from either a normal~~

2.3.2 Selection of the prior distribution

The prior distribution used for LHS was derived in multiple steps following partly an explorative approach. An initial version of the ensemble with 1000 members was run using the 10 LPX parameters and distribution used by (Steinacher et al., 2013) and four additional parameters relating to the nitrogen cycle and oxidation rates in areas with anthropogenic land-use. The ensemble is sampled using normal and log-normal distributions with distribution parameters chosen such that the median matches the parameter value of LPX-Bern v1.2 and the 90% confidence interval matches plausible ranges or literature-based ranges where available. Normal distributions are used by default, log-normal ~~distribution. The prior distributions were chosen starting from literature-based plausible ranges . Then these initial priors were refined iteratively by applying the Bayesian approach, as described in this and the next section, in sequence using multiple ensembles and additional factorial simulations for selected parameters . The prior distributions (Table 1)of the featured ensemble are~~ distributions are used for parameters with asymmetric parameter ranges and parameters with values close to zero. This initial ensemble was evaluated against a subset of the observational constraints presented in section 2.4 and it was found that ensemble performance is poor, especially with respect to global atmosphere-land fluxes. Sensitivity of model outcomes to indvidual parameter values was explored by 76 factorial simulations where aditional parameters were varied. The information from these sensitivitiy simulations together with results on parameter sensitivity of an earlier sudy Zaehle et al. (2005) are used to identify key model parameters. In addition six ensembles of reduced size (four 200 members and two 300 members), featuring slightly different parameter combinations, were used to refine the median parameter values and their ranges. By evaluating these simulations the final set of parameter presented in section 2.3.1 was selected. The final iteration included the seqential computation of three observation-constrained ensembles with 1000 members each. The first of these three ensemble was calculated with priors based on the refined median parameter values. The median and 95% confidence ~~interval of the constrained posterior parameter distribution~~ of ~~an earlier ensemblerun. While during ensemble revisions some metrics have changed and forcing files were updated,~~ the ~~result of this iterative procedure leads to a~~ posterior distribution after observation assimilation as described in section 2.3.1 is then used as the prior of a new 1000 member parameter ensemble. This procedure is repeated one more time to arrive at the prior distributions used in the final ensemble and displayed in Figure 1. No formal convergence criterion is employed, since the computation and

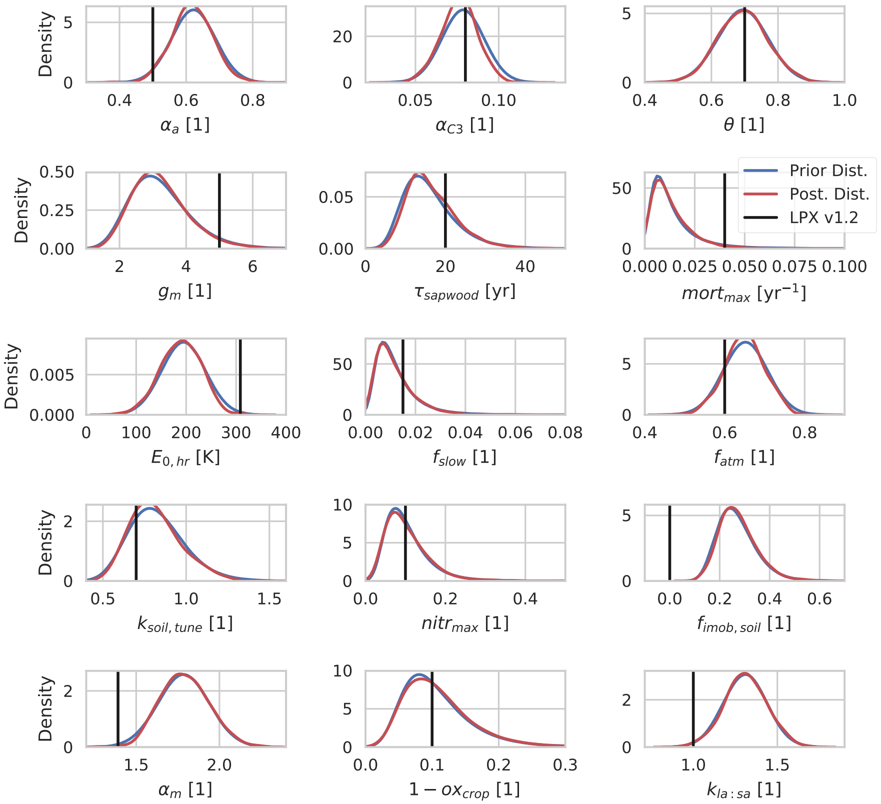

Figure 1. Kernel density estimations of the prior probability distribution (blue) and the posterior probability distribution (red) of the constrained ensemble with net land-use ($\mathrm{M}_{net,net}$). The prior distribution was improved iteratively, resulting in near convergence of prior and posterior distribution. Vertical black bars indicate the parameter values used in LPX-Bern v1.2.

evaluation of a single ensemble represents a considerable computational and analytical effort. The near convergence of the ~~prior and the posterior parameter distribution for~~ posterior (Section 2.3.1) and prior distribution of the final ensemble (~~Fig. 1) .~~

~~Description of sampling parameters with values for LPX v1.2 and the new best guess version v1.4. If not otherwise indicated parameters are unitless. Paramter Description LPX~~ Figure 1) indicates a near optimal solution for the parameter distribution in the context of the observational constraints and the associated skill score metric (section 2.3.1). In addition, this convergence of prior and posterior distribution also indicates that the final prior distribution is suited to adequatly sample the parameter space for our selection of observational constraints. The differences between the parameter value used in the older LPX-Bern v1.2 ~~LPX v1.4 α_a Fraction of PAR assimilitated at ecosystem level relative to leaf level 0.5 0.6175 α_{C3} Intrinsic quantum efficiency of CO_2 uptake in C3 plants 0.07 0.07660 θ Co-limitation shape parameter 0.7 0.6937 g_m Canopy conductance scaling parameter for water demand calculation 3.24 3.120 α_m Priestley-Taylor coefficient in water demand calculation 1.394 1.786 $\tau_{sapwood}$ Sapwood-heartwood turnover time yr20 15.33 $k_{la:sa}$ Allometric scaling parameter : leaf area to sapwood~~

~~area 1.0 1.310 $mort_{max}$ Asymptotic maximum in mortality equation yr^{-1}0.01 0.01016 $E_{0,hr}$ Temperature sensitivity of heterotrophic respiration K308.56 190.16 f_{atm} Fraction of litter entering atmosphere directly 0.6 0.6503 f_{slow} Fraction of litter entering slow soil pool 0.015 0.009512 $k_{soil,tune}$ Tuning factor for soil decay 0.7 0.7965 $nitr_{max}$ Maximum nitrification rate 0.1 0.09096 $f_{imob,soil}$ Nitrogen imobilization in soil 0.0 0.2639 $1-ox_{crop}$ Fraction of direct oxidation of leaf turnover on cropland 0.1 0.09920~~ and the best guess parameter values of the final ensemble (Figure 1; see also Table 1) provide a measure by how much individual parameters were revised during our iterative data assimilation. For completeness, we report that individual forcing data sets, such as the land use data, were updated and the set of observational constraints expanded during the course of the work.

2.4 ~~Skill scores~~

2.3.1 Skill scores and the posterior distribution

The performance of the final 1000 member model ensemble is evaluated using ~~a~~ the set of observational constraints ~~as~~ listed in Table 2. The model-data discrepancy for a given observational data set i and model run is estimated by the relative Mean Squared Error (MSE$^i_{rel}$)

$$\mathrm{MSE}^i_{rel} = \sum_j w_j \frac{(X^{mod,i}_j - X^{obs,i}_j)^2}{\sigma^2}. \tag{1}$$

w_j are the normalized weights of the data points j, which in the case of gridded data sets correspond to the grid cell area. $X^{mod,i}_j$ and $X^{sim,i}_j$ correspond to the ~~modeled~~ modelled and observed data points for constraint i respectively. In accordance with (Schmittner et al., 2009) and (Steinacher et al., 2013) the combined error σ^2 is approximated by the model-data variance of the model member with the smallest MSE$^i_{rel}$ of the ensemble. As a consequence, the smallest possible MSE$^i_{rel}$ using this approximation is one. If the observational error is known and larger than the variance, it is instead used as an estimate for the combined error, allowing a ~~theoretical~~ minimum MSE$^i_{rel}$ of zero.

The MSE~~$_{rel}$~~$^{i}_{rel}$ of all individual observational constraints is aggregated to a total error MSE$^{tot}_{rel}$ with a hierarchical weighting scheme shown in Fig. 2 and translated to a skill score S_m for each ensemble member m. We require that MSE$_{rel}$ is smaller than five for each of the individual observational data sets; otherwise S_m is set to 0.

$$S_m = \begin{cases} 0 & \exists i : \mathrm{MSE}^i_{rel} > 5 \\ exp(-\frac{1}{2}\mathrm{MSE}^{tot}_{rel}) & \text{else} \end{cases} \tag{2}$$

The size of the ensemble is further reduced by excluding runs with low skill scores, such that the remaining 667 runs have 99 % of the cumulative skill score $\sum_m S_m$ of all runs, which we term the constrained ensemble. The maximum achievable skill score is not 1 for spatially resolved data since it would correspond to a MSE$^{tot}_{rel}$ of 0, which is not achievable due to the approximation for the combined error, used in ~~some of the observational~~ the spatially resolved constraints. We did not renormalize skill score to a scale between 0 and 1.

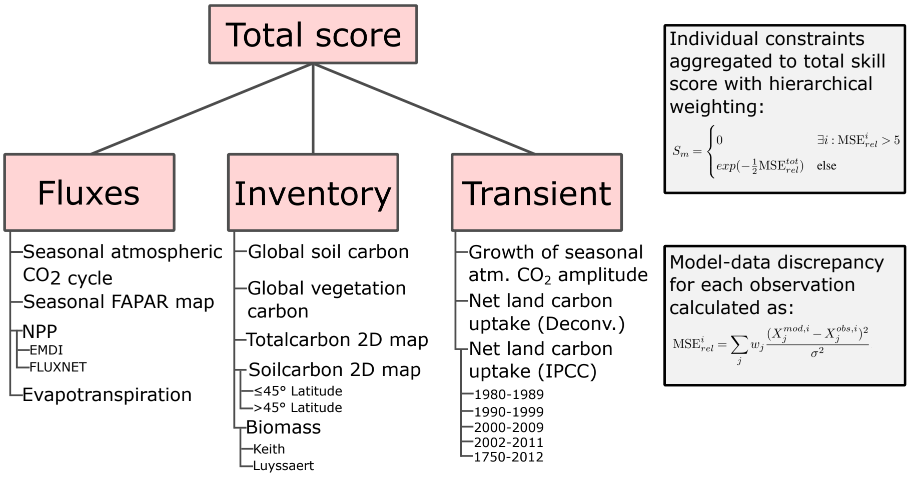
Total score
Fluxes
Inventory
Transient
Seasonal atmospheric CO_2 cycle
Seasonal FAPAR map
NPP
EMDI
FLUXNET
Evapotranspiration
Global soil carbon
Global vegetation carbon
Totalcarbon 2D map
Soilcarbon 2D map
≤45° Latitude
>45° Latitude
Biomass
Keith
Luyssaert
Growth of seasonal atm. CO_2 amplitude
Net land carbon uptake (Deconv.)
Net land carbon uptake (IPCC)
1980-1989
1990-1999
2000-2009
2002-2011
1750-2012
Individual constraints aggregated to total skill score with hierarchical weighting:
$S_m = \begin{cases} 0 & \exists i : \mathrm{MSE}^i_{rel} > 5 \\ exp(-\frac{1}{2}\mathrm{MSE}^{tot}_{rel}) & \text{else} \end{cases}$
Model-data discrepancy for each observation calculated as:
$\mathrm{MSE}^i_{rel} = \sum_j w_j \frac{(X_j^{mod,i} - X_j^{obs,i})^2}{\sigma^2}$

Figure 2. Hierarchical weighting scheme to aggregate the relative mean squared error of individual observational constraints to a total error which is then mapped to a total skill score.

The ~~skill score weighted normalized histogram of a~~ so-called posterior distribution of a parameter or quantity of interest is obtained by using the skill score weighted normalized histogram, which can be interpreted as a probability density function, of the constrained ensemble. The skill weighted median and confidence interval of a given quantity is then determined by transforming the histogram to a discrete cumulative density function using a cumulative sum and approximating the desired quantiles by a first-order interpolation. Throughout this paper we report the skill weighted median of numerical results along with the 5% and 95% quantiles, corresponding to the 90% confidence interval, in parentheses.

2.4 Observational constraints

The calculation of the MSE_{rel} requires the model and observational data to conform to the same structure. In the following, the required pre-processing will be outlined briefly. The seasonality of the fraction of absorbed photosynthetically active radiation (FAPAR) as simulated in the model is compared to a satellite-derived product (Gobron et al., 2006) which was regridded to the model grid and the MSE is calculated from the averaged monthly fields in the measurement period.

The modeled total and soil carbon distribution between 1982 and 2005 are compared to a data set based on observations (Carvalhais et al., 2014), regridded to the model resolution. The soil carbon map is divided into low and high latitudes regions in order to avoid potential biases from peat areas with very high soil carbon content.

For site level observed NPP (Multi-Biome NPP (Olson et al., 2013) and FLUXNET v3.1 (Luyssaert et al., 2009, 2007)), the site measurements are compared to the averaged modeled NPP of natural vegetation between 1931 and 1997 of the corresponding model grid cell. If multiple measurements are contained in the same grid cell they are averaged. Similarly, the site level measurements of biomass carbon (Keith et al., 2009; Luyssaert et al., 2009, 2007) are compared to the modeled natural

Table 2. Observations used to constrain the model ensemble

Category	Variable	Description	Reference
Fluxes	Seasonal atm. CO_2	Seasonal cycle at nine sites	(GLOBALVIEW-CO2, 2013)
Fluxes	NPP	Estimates of the 81 Multi-Biome Class A field measurements	(Olson et al., 2013)
Fluxes	NPP	Estimates of NPP at ∼140 FLUXNET sites	(Luyssaert et al., 2009, 2007)
Fluxes	fAPAR	SeaWiFS satelite product, regridded to model resolution	(Gobron et al., 2006)
Fluxes	Evapotranspiration	Merged evapotranspiration synthesis product from the LandFlux-EVAL	(Mueller et al., 2013)
Inventory	Total Carbon	Global distribution of total ecosystem carbon	(Carvalhais et al., 2014)
Inventory	Soil Carbon	Global distribution of total soil carbon	(Carvalhais et al., 2014)
Inventory	Vegetation Carbon	Biomass estimates at ∼140 FLUXNET sites	(Luyssaert et al., 2009, 2007)
Inventory	Vegetation Carbon	Biomass estimates at 136 sites	(Keith et al., 2009)
Inventory	Global Soil Carbon	Global inventory 1950 ± 450 ~~GtC~~ PgC	(Ciais et al., 2013)
Inventory	Global Vegetation Carbon	Global inventory 550 ± 100 ~~GtC~~ PgC	(Ciais et al., 2013)
Transient	Growth of CO_2 amplitude	Growth of seasonal atmospheric CO_2 amplitude at four sites	(GLOBALVIEW-CO2, 2013)
Transient	Land uptake (~~Inversion~~Deconvolution)	Global land uptake from atmospheric ~~inversion~~ deconvolution	~~(Boden et al., 2017; Battaglia and Joos, 2017)~~ study)
Transient	Land uptake (IPCC)	Global land uptake in five periods	(Ciais et al., 2013)

vegetation carbon, averaged between the periods 1950-2000 and 1931-1997 respectively. The biomass carbon of Luyssaert et al. (2009) is obtained by using a carbon to organic matter conversion factor of 0.475.

The TM2 (Kaminski et al., 1999), a global atmospheric tracer model was used to translate the gridded land-atmosphere flux to local anomalies in atmospheric CO_2. This method does not include the interannual variability of the transport. 9 sites from the GLOBALVIEW-CO2 database (GLOBALVIEW-CO2, 2013) were selected and the annual offset corrected seasonality of CO_2 in the period of 1980-2013 was compared. The influence of sea-air carbon exchange on the seasonal cycle and trend in atmospheric CO_2 are taken into account. This is done by prescribing net sea-to-air fluxes as simulated by the Bern3D model (standard setup) ~~(Battaglia and Joos, 2017; Roth et al., 2014; Ritz et al., 2011)~~(Battaglia and Joos, 2018; Roth et al., 2014; Ritz et al., 2011). The growth of the seasonal amplitude at a subset of four sites with high seasonality was used as a further constraint.

The modeled mean annual evapotranspiration between 1989-2005 was compared to the LandFLUX-EVAL evapotranspiration data product (Mueller et al., 2013).

The global terrestrial carbon flux is constrained by ~~an inversion~~a deconvolution, for which the global atmospheric CO_2 concentration, the median of an ensemble of simulated ocean-atmosphere fluxes ~~(Battaglia and Joos, 2017)~~(Battaglia and Joos, 2018), consistent with other estimates (Khatiwala et al., 2013; DeVries, 2014), and an inventory of anthropogenic CO_2 emissions (Boden et al., 2017) were used. The combined error in Equation 1 is estimated by propagating the 90% confidence interval of ocean-atmosphere fluxes and assuming a 5% uncertainty for the anthropogenic emissions (Ballantyne et al., 2015).

The estimates of global soil and vegetation carbon as given by IPCC (Ciais et al., 2013) are used as a global constraint. The observation-based estimates are compared to the average soil and vegetation carbon over the whole industrial period. Additionally, the estimates for the global land-atmosphere flux in the periods 1970-1979, 1980-1989, 1990-1999, 2000-2009 and 2002-2011, are compared to the simulated land-atmosphere fluxes over the same period. Since the model simulation starts only in the year 1800, the estimated land-atmosphere flux over the industrial period from 1750-2011 is compared with the model by approximating the flux of the period 1750-1800 with 1801-1850. For all global constraints, the uncertainties reported by IPCC are used as an estimate for the combined error in Equation 1.

2.5 Definition of Land-Use emissions and the setup of ~~three~~ the model ensembles

To quantify emissions from LULCC a second simulation featuring a time-invariant pre-industrial land-cover distribution and nitrogen fertilization is run for every ensemble member. In accordance with the TRENDY model intercomparison ((Sitch et al., 2015)), we define the emissions from LULCC as the difference of the change in carbon in the reference and fixed LULCC simulation. The change of carbon in the land system is calculated from the cumulative net biome production (NBP) including emissions from product pools. Since the additional simulations with fixed LULCC feature transient CO_2 and climate forcing, the direct impact of climate change and increasing CO_2 on E_{LUC} are considered, however unlike in coupled models (Strassmann et al., 2008) physical and biogeochemical feedbacks of LULCC on the climate are neglected. We refer to the literature (Strassmann et al., 2008; Pongratz et al., 2014; Stocker and Joos, 2015) for further discussion of differences in the definition of land-use fluxes.

For each of the parameter sets 4 transient simulations over the industrial period are performed: (i) a simulation with prescribed net transitions ($M_{net,net}$ and $M_{gross,net}$), (ii) a simulation with prescribed gross transitions ($M_{gross,net}$ and $M_{gross,gross}$), (ii) a run with landuse area fixed at preindustrial levels and (iv) a run with landuse including shifting cultivation held at preindustrial levels. The last two simulations are used purely diagnostic to determine E_{LUC}. E_{LUC} is investigated using three different ensemble configurations. $M_{net,net}$ labels the standard model version featuring only net ~~LULUC~~ LULCC transitions. $M_{gross,net}$ and $M_{gross,gross}$ feature modules for shifting cultivation and wood harvest (gross land-use) but lack northern peatlands due to technical limitations. $M_{gross,net}$ reuses the skill scores calculated for $M_{net,net}$ and $M_{gross,gross}$ features skill scores calculated on the basis of the gross land-use configuration.

For the $M_{net,net}$ ensemble and the $M_{gross,net}$ ensemble, the priors of the model parameters were improved iteratively during the development of our benchmark system. Consequently, the solutions for the model parameters and associated model outcomes converge. For example, the prior and the posterior probability distribution of the sampled parameters are nearly identical (Fig. 1). This provides strong support that an optimal solution for the sampled parameters has been found for the applied model structure and observational constraints. In contrast, the parameters of the $M_{gross,gross}$ ensemble were not improved iteratively, given the computational cost, and prior and posterior solutions do not converge.

3 Results

3.1 Land-Use Emissions

~~The magnitude of emissions of carbon due to changes in LULCC is hard to quantify and subject to large uncertainties (Li et al., 2017; Arneth et al., 2017; Houghton and Nassikas, 2017; Pongratz et al., 2011; Roman-Cuesta et al., 2016; Stocker et al., 2014;~~ The use of ~~an~~ the ensemble framework allows us to quantify both the magnitude and the uncertainty of land-use emissions in a model due to parameter spread. Following the procedure outlined in the method section, E_{LUC} is computed for every ensemble member. In this section, we first present E_{LUC}, total land-atmosphere fluxes and the residual land carbon sink on a global scale for the three ensemble configurations and then further analyze spatial patterns and regionally aggregated estimates.

3.1.1 Global Fluxes

Global aggregates of skill weighted median NBP, E_{LUC}, residual terrestrial sink and their respective cumulative fluxes, including a 90% confidence interval as an estimate for model parameter uncertainty, are shown in Fig. 3. For the standard model configuration $\mathrm{M}_{net,net}$, featuring net land-use, the total change in land carbon (i.e. cumulative NBP) is a release of ~~24.2~~ 24.4 (4.5, 44.0) PgC from 1860 to 1960 and an uptake of ~~25.5~~ 25.4 (8.4, 47.0) PgC from 1961 to 2016. The standard deviation of NBP increases from ~~0.8~~ 0.82 (0.65, 0.97) PgC/yr between 1860 and 1960 to ~~1.2~~ 1.18 (0.97, 1.38) PgC/yr in the latter period. The change in total carbon is discussed in more detail in section 3.3. E_{LUC} is positive throughout the whole industrial period, i.e. a source of carbon to the atmosphere. A temporary maximum of emissions is reached in the 1950s followed by relatively constant emissions until the 2000s where the emissions increase with enhanced variability. The cumulative emissions from 1860 to 2016 amount to ~~98.0~~ 96.9 (59, 138) PgC. The residual terrestrial sink, computed as the difference between NBP and E_{LUC}, shows a similar pattern of variability as NBP. While the residual terrestrial sink flux is negative in some years, the cumulative residual terrestrial sink generally increases steadily and amounts to ~~98.9~~ 98.1 (71.0, 132.3) PgC between 1860 and 2016.

In addition to the standard model configuration a second ensemble of a model configuration $\mathrm{M}_{gross,net}$ featuring modules for shifting cultivation and wood harvest (gross land-use) is employed. By using the skill scores $\mathrm{M}_{net,net}$, the parametrization remains identical allowing to compare the pure mechanistic difference between the two versions. The difference in median E_{LUC} between the net and gross land-use configuration is most pronounced in the second half of the 20th century and amounts to ~~43.7~~ 44.5 PgC between 1860 and 2016. The gross land-use ensemble simulates on average ~~0.41 GtC~~ 0.40 PgC yr^{-1} more emissions due to ~~LULUC~~ LULCC between 1950 and 2016. This result is compatible with the earlier study by (Stocker et al., 2014), which investigated land-use-change using an earlier version of LPX-Bern with a single parameter configuration. The residual terrestrial sink shows as expected a near identical behavior in the two versions. The resulting total change in land carbon is negative, with a slight uptake of carbon ~~after 1980.~~ at the end of the century, amounting to 9.3 (-0.9,22.2) PgC between 1990 and 2016.

A third model configuration $\mathrm{M}_{gross,gross}$ is obtained by recalculating the skill scores from the gross land-use results. As described in section 2.5, the priors of the $\mathrm{M}_{gross,gross}$ were not improved iteratively to yield convergence between prior and pos-

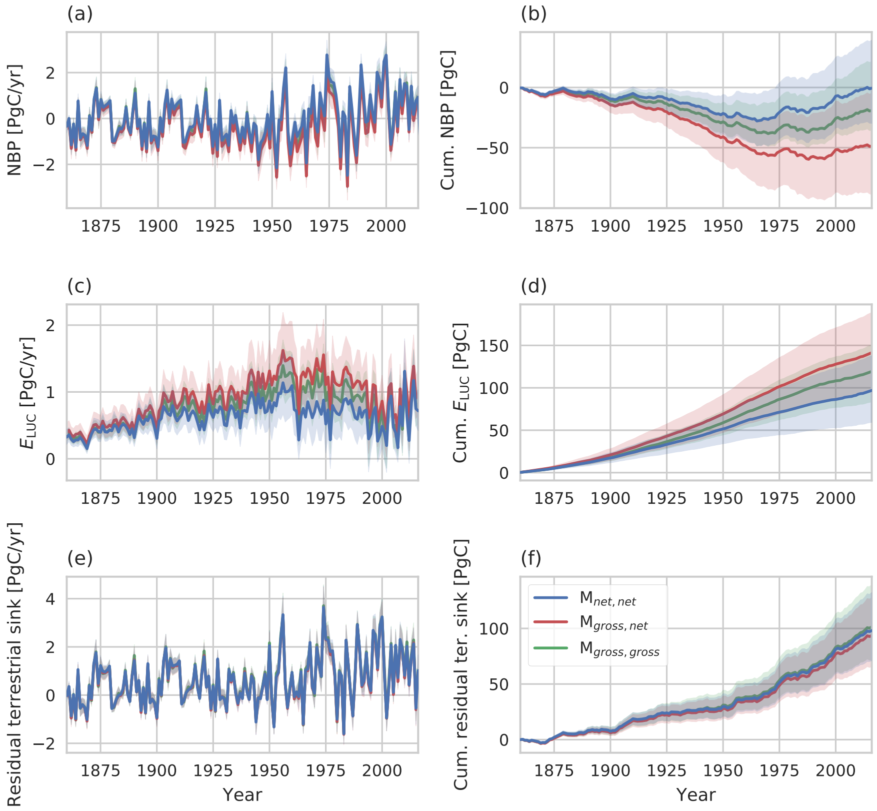

Figure 3. Skill weighted median Net Biome Production (NBP) (a), emissions due to ~~LULUC~~ LULCC E_{LUC} (c) and the residual terrestrial sink flux (e) and their respective cumulative fluxes (b,d,f) for the constrained ensemble with net land-use $\mathrm{M}_{net,net}$ (blue), additional gross land-use processes $\mathrm{M}_{gross,net}$ (red) and gross land-use with skill scores recalculated $\mathrm{M}_{gross,gross}$ (green). The shading corresponds to the 90% confidence interval.

terior solutions. This leaves only 200 runs in $\mathrm{M}_{gross,gross}$ in contrast to the 667 runs in $\mathrm{M}_{net,net}$ and consequently $\mathrm{M}_{gross,net}$. In addition, several important benchmarks such as vegetation carbon density are not simulated as well in $\mathrm{M}_{gross,gross}$ compared to $\mathrm{M}_{net,net}$ and $\mathrm{M}_{gross,net}$. Since NBP is constrained by observations, median cumulative NBP from 1860 to 2016 is only ~~17.5~~ 18.6 PgC smaller in the $\mathrm{M}_{gross,gross}$ than in the $\mathrm{M}_{net,net}$ ensemble. Surprisingly E_{LUC} is only 21.4 ~~GtC~~ PgC higher over the period from 1860 to 2016 for $\mathrm{M}_{gross,gross}$ than for the standard version $\mathrm{M}_{net,net}$. Why are E_{LUC} emissions so similar between these two ensemble versions with net and gross transitions? The residual sink is relatively insensitive to ~~parameterization~~ parametrization in LPX and the version with gross skill scores only has a moderately larger residual sink uptake of 7.8 PgC in the considered period, largely caused by a downward adjustment of the parameter $E_{0,hr}$ governing the temperature dependency in heterotrophic respiration to a median value of 151 K (190 K in $\mathrm{M}_{net,net}$; Table 1). In $\mathrm{M}_{gross,gross}$, global vegetation carbon inventory is only 417 (341, 506) PgC (average over the industrial period) compared to 468 ~~GtC~~ (358, 590) PgC in the $\mathrm{M}_{net,net}$

Table 3. Comparison of the skill weighted median emissions due to Land-use-change of the two constrained LPX parameter ensembles (90% confidence intervals in brakets) to the bookkeeping method and DGVM model ensemble of (Le Quéré et al., 2016). The uncertainty in the DGVM multi-model ensemble is given by the standard deviation across model members, for the bookkeeping method a best value judgement on the uncertainty of ±0.5 PgC yr^{-1} is provided.

	Mean E_{LUC} [PgC yr^{-1}]				
	1960-1969	1970-1979	1980-1989	1990-1999	2000-2009
LPX-Bern M$_{net,net}$	0.70 (0.33, 1.04)	0.69 (0.30, 1.06)	0.75 (0.40, 1.07)	0.55 (0.22, 0.83)	0.52 (0.21, 0.78)
LPX-Bern M$_{gross,net}$	1.22 (0.78, 1.64)	1.25 (0.77, 1.71)	1.19 (0.77, 1.57)	0.93 (0.54, 1.28)	0.74 (0.41, 1.05)
LPX-Bern M$_{gross,gross}$	1.02 (0.65, 1.32)	1.04 (0.65, 1.37)	0.99 (0.66, 1.27)	0.74 (0.37, 1.05)	0.59 (0.26, 0.87)
GCP2016 Bookkeeping	1.5 ± 0.5	1.3 ± 0.5	1.4 ± 0.5	1.6 ± 0.5	1.0 ± 0.5
GCP2016 DGVMs	1.2 ± 0.3	1.2 ± 0.3	1.2 ± 0.2	1.1 ± 0.2	1.3 ± 0.3

ensemble. The observational constraints for the net land carbon sink (Fig. 2, Table 2) are apparently better approximated for a smaller vegetation carbon stock in M$_{gross,gross}$. Vegetation carbon inventory is underestimated by about 20% compared to the observational benchmarks. The smaller vegetation carbon stock in M$_{gross,gross}$ leads to smaller E_{LUC} anything else kept equal. In addition, the amount of carbon harvested (ox_{crop}) is reduced from 90% in the standard M$_{net,net}$ ensemble to 83% in the M$_{gross,gross}$ ensemble. As a result of these two adjustments, E_{LUC} is smaller in the M$_{gross,gross}$ than in the M$_{gross,net}$ ensemble. If the relative importance of the land-atmosphere observational constraints is increased, the difference in E_{LUC} of M$_{gross,gross}$ and M$_{net,net}$ is decreased even further.

E_{LUC} as simulated by LPX-Bern is compared in Table 3 to a bookkeeping method and a DGVM model ensemble average from the Global Carbon Project (GCP, Le Quéré et al. (2016)). E_{LUC} in the net land-use configuration M$_{net,net}$ is considerably smaller than the estimates of the GCP with an average annual emissions of 0.64 (0.29, 0.95) PgC yr^{-1} between 1960 and 2009, compared to the 1.4 PgC yr^{-1} of the bookkeeping approach and the 1.2 PgC yr^{-1} of the multi-model DGVM approach. The emissions of the gross land-use configuration with gross skill scores are higher but still fairly low with 0.88 (0.51, 1.17) PgC yr^{-1}. The version featuring gross land-use with net skill scores yields higher land-use emissions with 1.07 (0.66, 1.46) PgC yr^{-1}, which is within the uncertainties of both estimates. The largest discrepancy between LPX and GCP is found in the 1990s and 2000s. The uncertainty in the parameter ensembles is comparable to the uncertainty in the multi-model ensemble of the GCP. The tendency to low emissions is a consequence of the ensemble favoring low emissions to match the observational total land-atmosphere flux, combined with a relatively weak residual terrestrial sink in LPX-Bern.

In the following the ensemble version with gross land-use and skill scores from the net land-use ensemble M$_{gross,net}$ is used to investigate the spatial structure of E_{LUC}. This is motivated by the much better representation of the vegetation carbon benchmark in the M$_{gross,net}$ ensemble than in the M$_{gross,gross}$ and a higher confidence in the overall benchmark performance of the M$_{net,net}$ ensemble. A caveat of this choice is that the net land-atmosphere flux is underestimated~~. However, this should not influence our conclusions on~~ in M$_{gross,net}$ because the residual land sink only responds to the lower E_{LUC} of M$_{net,net}$.

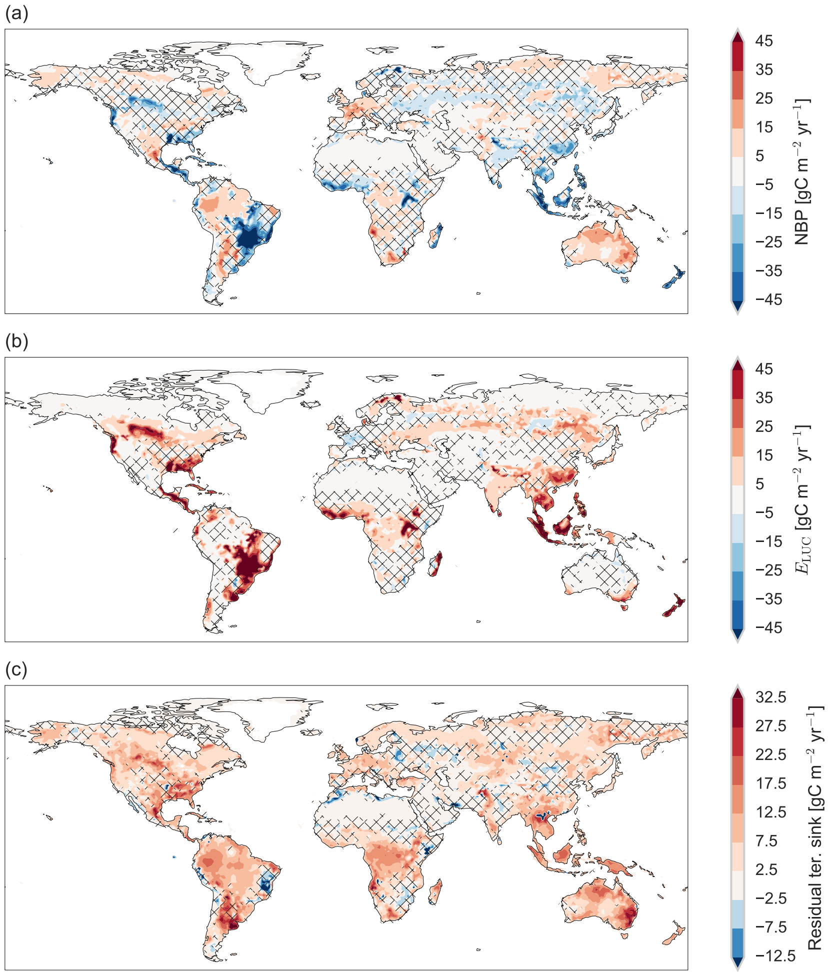
(a)
45
35
25
15
5
−5
−15
−25
−35
−45
NBP [gC m^{-2} yr^{-1}]
(b)
45
35
25
15
5
−5
−15
−25
−35
−45
E_{LUC} [gC m^{-2} yr^{-1}]
(c)
32.5
27.5
22.5
17.5
12.5
7.5
2.5
−2.5
−7.5
−12.5
Residual ter. sink [gC m^{-2} yr^{-1}]

Figure 4. Skill weighted median annual net biome production (NBP) (a), ~~atmosphere-land flux~~ emissions due to land-use change E_{LUC} (b) and the residual terrestrial sink flux (c) from 1901-2016 for the ensemble $\mathrm{M}_{gross,net}$. Areas, where the lower and upper limit of the 90% confidence interval have different signs, are hatched.

However if only considering E_{LUC} we expect the magnitude of the residual land-sink and net land-atmosphere flux to be less important than model performance with respect to vegetation carbon (Li et al., 2017) and other benchmarks.

3.1.2 Spatial Patterns and Regional Aggregates

The land-atmosphere fluxes show large regional differences (Fig. 4). The most pronounced feature of net atmosphere-land fluxes is the release of carbon due to deforestation in the Amazon rainforest and the regions close to the equator and a tendency to a net uptake of carbon at higher latitudes, such as central Europe. The calculated land-use ~~change flux is negative~~ emissions

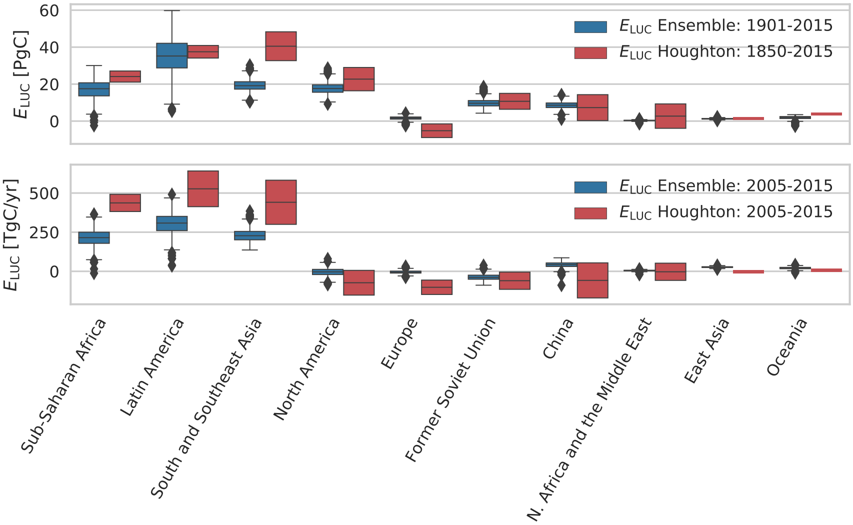

Figure 5. E_{LUC} as simulated by the ensemble $\mathrm{M}_{gross,net}$ compared to the Houghton and Nassikas (2017) estimates. The E_{LUC} of the ensemble was calculated for individual countries and then aggregated to 10 regions defined in Houghton and Nassikas (2017). The top panel shows the estimates for the emissions over the industrial period; for LPX gridded output is only available after 1901. The bottom panel shows the annual mean emissions from 2005 to 2015. The Houghton and Nassikas (2017) estimates include the reported uncertainties based on the standard deviation of five quasi-independent studies; for East Asia and Oceania no uncertainty is reported.

E_{LUC} are positive everywhere except central Europe and the west coast of Northern America, resulting in the expected overall emission of carbon due to land-use change. The residual carbon uptake, that is the total atmosphere-land flux minus the contribution of land-use change, shows a consistent uptake of carbon between 1901-2016, with the exception of some areas with high ensemble uncertainty. There are large regions where the 90% confidence interval in the ensemble does not agree on the sign, however, most of these areas feature low NBP.

The E_{LUC} of $\mathrm{M}_{gross,net}$ are aggregated to regions and compared to estimates of Houghton and Nassikas (2017) (Fig. 5). Since spatial output in LPX is only available after 1901 in LPX, the period 1850 to 2015 in Houghton and Nassikas (2017) is approximated by the interval 1901 to 2015. The global skill weighted median E_{LUC} from 1850 to 1900 amounts to 24.5 (16.9, 33.6) PgC. Overall the global median emissions between 1850 and 2015 in LPX amount to 144.5 (97.5, 192.7) PgC very close to the estimate Houghton and Nassikas (2017) of 145.5 ± 16.0 PgC. The largest discrepancy in the individual regions is found in South and Southeast Asia, where LPX yields lower emission estimates, which might be a consequence of the lack of tropical peatlands in the ensemble. In the recent decade from 2005 to 2015, the agreement is less pronounced. While the global annual flux simulated by LPX of 866 (552, 1181) TgC yr^{-1} is within the uncertainty of the independent estimate of 1113 ± 345 TgC yr^{-1}, the distribution of this flux to the regions shows some divergence. In LPX the tropical regions yield lower emissions, which is somewhat offset by a weaker sink effect in the temperate regions of North America, Europe, China and the former Soviet Union.

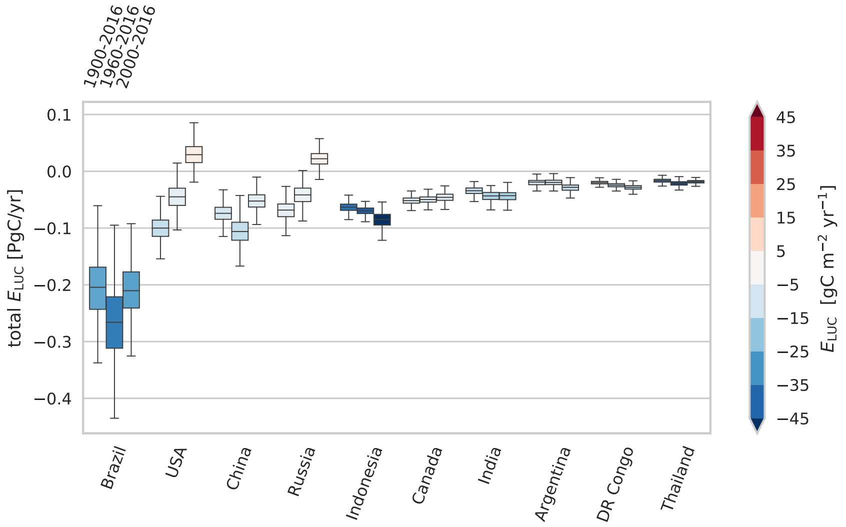

Figure 6. Overview of 10 countries with the highest overall contribution to emissions due to land-use change from 1901-2016 in the model ensemble $M_{gross,net}$. The three columns of the bar plot of each country show the total land-use change flux from 1900-2016, 1960-2016 and 2000-2016 respectively. The color of the bar plot corresponds to the land-use change flux per unit area from the respective country and period.

By using the NaturalEarthData administrative borders the E_{LUC} of $M_{gross,net}$ are aggregated to individual countries. The E_{LUC} of the ten countries with the largest contribution to total E_{LUC} from 1901-2016 are shown in Fig. 6. Brazil emitted the most carbon due to land-use change, because of the size of the country combined with the high emissions per unit area. The United States of America, China and Russia have moderate per unit area emissions but are a large contributor due to their sheer size. These 3 countries show a decrease of emissions in the 21st century, with the USA and Russia having negative emissions in the 2000s. Indonesia shows the largest per area emissions of the considered countries and emissions increase in the 2000s. The emissions in Indonesia are likely underestimated due to a lack of tropical peatlands in the ensemble.

3.2 Evaluation of ensemble performance with respect to observational targets

In this section, the performance of the net land-use ensemble members ($M_{net,net}$; $M_{gross,net}$ performance is nearly identical) in the different observational metrics are discussed. In Fig. 7 a mapping of the MSE_{rel} to an individual skill score is displayed for the observational data-sets with a spatial structure, to demonstrate how well the median of the ensemble and the new version LPX v1.4 are able to simulate individual observations. The figure also demonstrates the success of the assimilation process: the skill scores for many individual targets are improved in the ensemble median and LPX v1.4 compared to LPX v1.2, the starting point of our work. As a consequence of our iterative prior selection (section 2.3.2) the median skill for an individual constraint is similar in the constrained ensemble compared to the unconstrained ensemble. In all but the fAPAR benchmark the

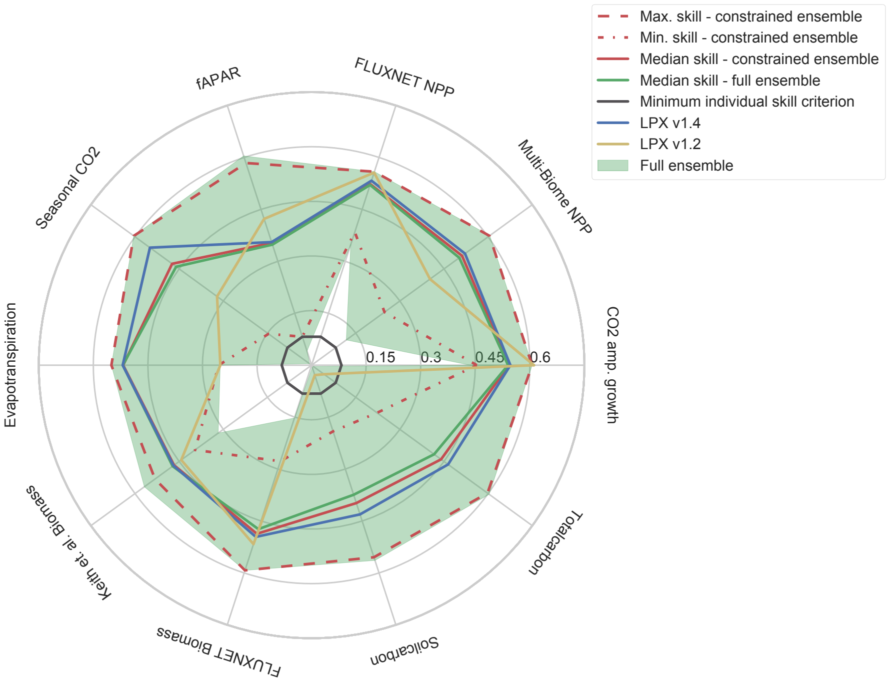

Figure 7. A mapping of the MSE_{rel} of observational targets with spatial structure to an individual skill $s = e^{-\frac{1}{2}MSE_{rel}}$ for the $M_{net,net}$ ensemble. The range of the unconstrained ensemble is shaded in green. The range and median of the constrained ensemble is outlined in red. The skill of the new model reference version 1.4 (blue) is compared to the skill of the older model version 1.2 (red). The minimum MSE_{rel} criterion is shown in black. The theoretically maximum achievable skill is 0.61 for the targets shown (section 2.3.1).

skill is consistently higher than the minimum skill criterion. With the exception of the biomass measurements by (Keith et al., 2009) and the fAPAR benchmark, the maximum skill in the constrained ensemble is identical to the full ensemble. The reduced maximum skill in those benchmarks is due to an exclusion of singular runs excelling at this benchmark but performing badly in others. ~~The median skill is as expected better or equal for the constrained ensemble than for the full ensemble. The minimum skill is significantly enhanced in the constrained ensemble. In all but the fAPAR benchmark the skill is consistently higher than the minimum skill criterion.~~ LPX v1.4, indicative of the $M_{net,net}$ ensemble performance, is compared to the observational targets in more detail in the supplementary Figures S1-S14.

As an illustration of the observational constraints, we consider the seasonal cycle of atmospheric CO_2 and the growth in the amplitude of the seasonal cycle of atmospheric CO_2. In Fig. 8 the median simulated values, as well as the 90% confidence interval, of the $M_{net,net}$ ensemble are compared to the atmospheric measurements (GLOBALVIEW-CO2, 2013) for a subset of 2 measurement sites, Alert (Nunavut, Canada) and Terceira Island (Azores, Portugal). The model ensemble is able to reproduce the seasonality pattern, as well as the increase in seasonal amplitude. As expected, the interannual variability in seasonal

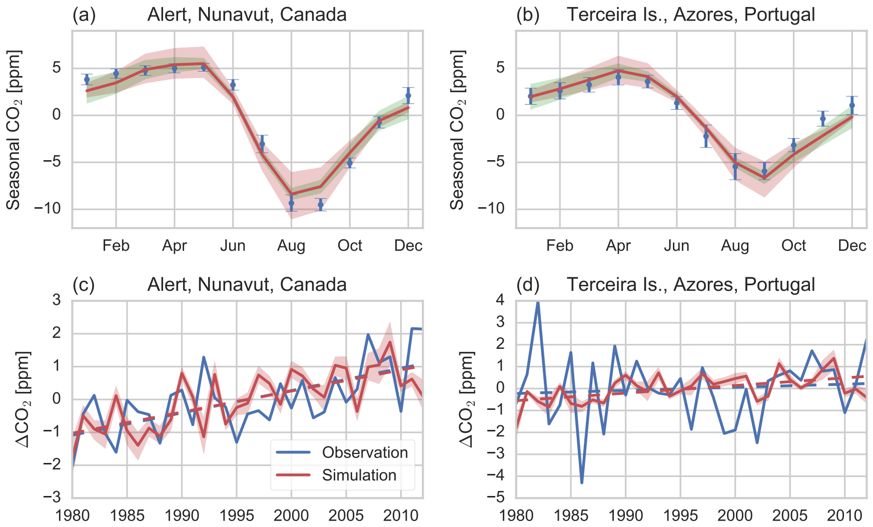

Figure 8. Panels (a) and (b): Seasonal cycle averaged from 1980 to 2013 at 2 measurement sites (GLOBALVIEW-CO2, 2013) (blue) compared to the median values of the $\mathrm{M}_{net,net}$ ensemble, with 90% confidence interval shaded in red. The standard deviations of the seasonal average are indicated with error bars for the measurements and green shading for the simulations. In panels (c) and (d) the growth in the amplitude of atmospheric CO_2 for the same 2 measurement sites (GLOBALVIEW-CO2, 2013) (blue) are compared to the median of the model ensemble, with the 90% confidence interval shaded in red. A linear fit indicated by dashed lines is included. The CO_2 concentration at a given site and time is computed with the TM2 transport model using simulated net land-atmosphere fluxes for each ensemble member and ocean-atmosphere fluxes from the Bern3D ocean model ~~(Battaglia and Joos, 2017)~~(Battaglia and Joos, 2018). The seasonal cycle of CO_2 is dominated by fluxes from the land, in particular, the northern hemisphere.

amplitude of CO_2 is not captured as the atmospheric transport model TM2 does not represent interannual variability in mass transport.

For the scalar targets, the median values and range of the full and constrained ensemble are compared in Fig. 9. The constrained ensemble shows a consistently improved performance for the uptake targets. In general, the targets are matched well for the 20th century but net land carbon uptake is underestimated in the model ensemble compared to the observational estimates in the beginning of the 21st century. Soil carbon and vegetation carbon inventory are matched well in the model, with a considerable decrease of model spread in the constrained ensemble. The median vegetation carbon of the constrained ensemble is lower than the full ensemble. This is due to a trade-off in the skill of land carbon uptake, increased vegetation carbon leads to a higher release of carbon due to deforestation.

Vegetation carbon inventory and spatial distribution are highly ~~significant~~ relevant for E_{LUC} estimates (Li et al., 2017). The sum of the vegetation carbon estimate and soil carbon estimate by Carvalhais et al. (2014) is used as a constraint for the total carbon, however, the individual vegetation carbon data is not used as a constraint. Nevertheless, the global vegetation carbon inventories of the two products are compatible with 422 (328,523) PgC for the vegetation carbon as simulated by LPX and 449

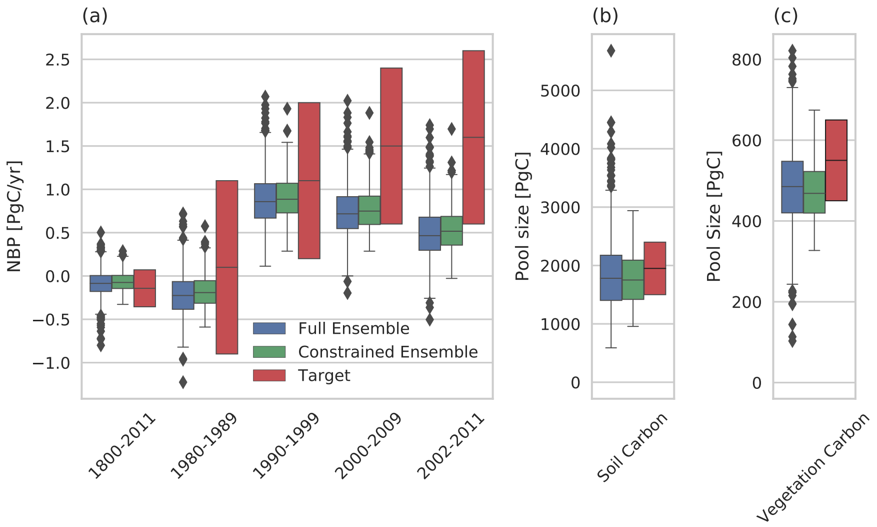

Figure 9. The value and uncertainty of the scalar targets (red) compared to an unweighted histogram of the full (blue) and constrained (green) ensemble $M_{net,net}$. Panel (a) shows the net biome production in 5 periods, panels (b) and (c) show the global soil and vegetation carbon inventories respectively.

PgC for the Carvalhais et al. (2014) estimate. The spatial patterns (Fig. 10) between simulated vegetation and the Carvalhais et al. (2014) estimates are fairly consistent with a correlation between the two products of $r^2 = 0.83$. LPX simulates somewhat more carbon in vegetation in the high latitude. The extent of areas with high vegetation density in tropical Africa is larger in LPX, but peak vegetation density in this area is lower than in the observational product. The vegetation carbon density in the model is somewhat lower in south-east Asia.

We compare the total land-atmosphere exchange flux to ~~an inversion using anthropogenic fossil fuel emissions (Boden et al., 2017),~~ the results of the atmospheric CO_2 ~~concentration (Rubino et al., 2013) and an ensemble simulation of the Bern3D ocean model~~ deconvolution in Fig. 11. The model ensemble shows lower emissions in the early 20th-century and slightly underestimates NBP in the latter half of the 20th-century compared to the ~~inversion~~deconvolution. The overall exchange of carbon over the industrial period is within the uncertainty of the estimate.

We investigate the dependency of the constrained ensemble on the choice of the observational constraints by reevaluating the ensemble for a subset of observations. We created 19 weighting schemes, each missing one of the individual observational constraints (Figure 2 and table 2) and otherwise identical to the default scheme. Then the median skill weighted parameter values of these ensembles are compared to the best-guess values of $M_{net,net}$ (section 3.3). The relative change in parametrization is less than 1% for 15 out of the 19 considered alternative weighting schemes. Leaving away the global vegetation and soil carbon constraints lead to moderate changes, notably to a change in the parameter for mortality ($mort_{max}$) of 4% and 2% respectively. Not including the soil carbon distribution in high latitudes lead to an increase of the parameter for the dependency

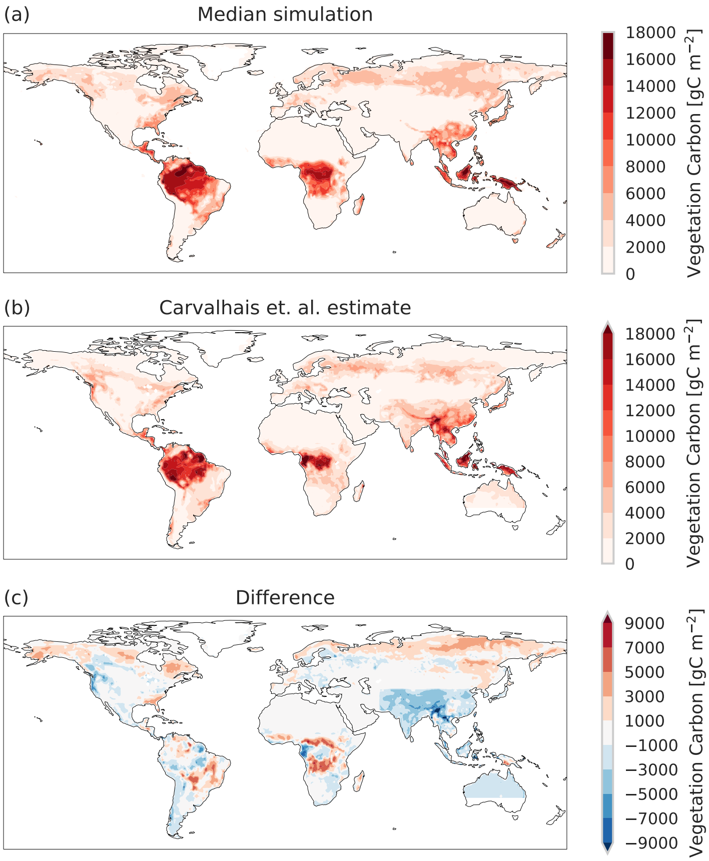

Figure 10. The skill weighted median M$_{net,net}$ vegetation carbon distribution averaged between 1982 and 2005 (a), compared to the Carvalhais et al. (2014) vegetation carbon estimate (b). The correlation of the estimates is $r^2 = 0.83$. The absolute difference is shown in panel (c).

of soil respiration on temperature ($E_{0,hr}$) of 2%. The largest changes in parametrization were observed when not considering the atmospheric deconvolution, most notably the sapwood-heartwood turnover time $\tau_{sapwood}$ decreased by 5%. When omitting entire categories in the benchmarking scheme, the changes in parametrization are larger than for omitting individual constraints, with parameter changes of up to 1% for the fluxes, 5% for the inventory and 6% for the transient category. This shows that the final parametrization is not overly sensitive to the inclusion or omission of a single observational product.

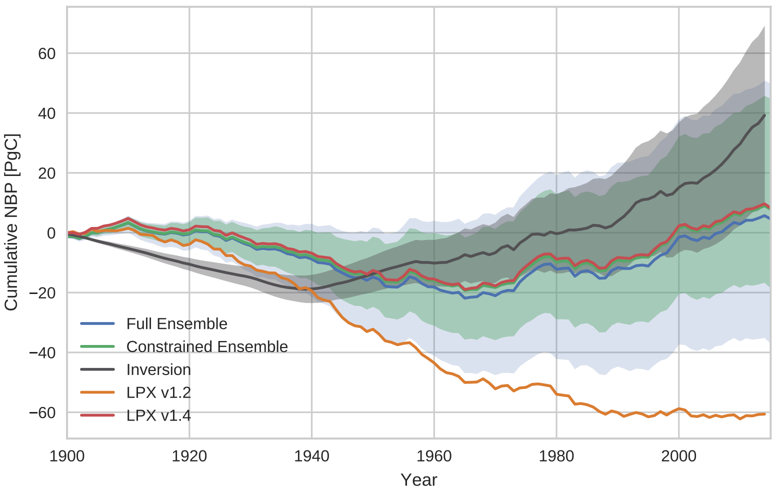

Figure 11. Cumultative net biome production (NBP) of the unconstrained (blue) and constrained (green) ensemble with 90% confidence interval shaded, LPX Version 1.2 (orange) and the new reference model version 1.4 (red). The result of a so-called "single deconvolution" is shown by the black line and grey range. In this ~~inversion~~deconvolution, the change in the land inventory is inferred from the records of atmospheric CO_2 and anthropogenic emissions from fossil fuel (and cement production) and ocean carbon uptake as estimated from an ensemble of simulations with the Bern3D model.

The unweighted kernel density estimates of the prior (full ensemble) and posterior (constrained ensemble) parameter distributions are shown in Fig. 1. The iterative procedure discussed in ~~2.3~~ 2.3.2 results in only slight changes of the posterior distribution with respect to the prior distribution. The median of the distributions is however substantially different from the initial parameter value used in LPX v1.2, the version used as a starting point for this study.

3.3 Parameters of the new reference model version

We use the constrained ensemble to establish a new reference model version, featuring a set of optimized parameters. The reference version will be used for model simulations where the use of an ensemble is not appropriate or required.

The skill weighted median parameter values of the constrained ensemble are used as a reference model and its parameter values are shown in Table 1. In Fig. 11 cumulative NBP is displayed for an older model version, the mean values of constrained and unconstrained model ensemble as well as a run with the new best guess parameters. The best guess version is similar to the mean behavior of the constrained ensemble, showing a net uptake of carbon in the latter half of the 20^{th} century, consistent with observations (Ciais et al., 2013). We note that the intermediate version v1.3 used in Keller et al. 2017 features similar parameter settings as determined here. This version simulated 20th century changes in carbon isotope discrimination and intrinsic water use efficiency in good agreement with tree-ring data. The severe underestimation of the land-carbon sink in older versions of LPX-Bern was a consequence of the introduction of new features and improvements in the code of LPX-

Bern, without subsequent retuning of the parametrization. The parameter changes are most pronounced in the temperature dependence of heterotrophic respiration $E_{0,hr}$ and α_m, a parameter associated with plant water demand. Both of these changes are not unexpected, as they increase the land carbon sink. In the case of heterotrophic respiration less carbon is lost due to increasing surface temperature and the increased water demand amplifies the CO_2 fertilization effect.

Overall the updated parametrization shows a well-balanced performance in the spatial benchmarks shown in Fig. 7. The older LPX version excels at singular metrics, namely the amplitude growth of CO_2 and the FLUXNET measurements, but breaks down at others, such as the spatial distribution of carbon and evapotranspiration. Furthermore, it also performs considerably worse in the scalar and ~~inversion~~ deconvolution targets.

The choice of using the skill weighted median parameters of the constrained ensemble instead of simply using the best performing parameter set for the reference version is motivated by its robustness and representativeness of the ensemble. While the best performing model member certainly possesses a higher skill score, its parameter values can depend strongly on the choice and weighting of the observational targets, whereas the median parameter values depend less on individual metrics.

4 Discussion

4.1 Land-atmosphere fluxes and E_{LUC}

The simultaneous assimilation of multiple observational constraints yields soil and vegetation stocks and distributions which are consistent with observations. The total land-atmosphere carbon flux is reproduced relatively well in the model configuration using net land-use $\mathrm{M}_{net,net}$. Comparing the land-atmosphere carbon flux to the independent flux estimates by Schimel et al. (2015) in the period 1990-2007, the tropical and southern fluxes are in good agreement to the atmospheric ~~inversion~~ deconvolution results with airborne constraint with a flux of 0.24 (-0.02,0.57) in LPX-Bern. The flux in the northern extratropical areas of 0.50 (0.37,0.63) is on the lower end but easily fulfills the mass balance.

The observed uncertainties of E_{LUC} due to parameter uncertainty in the DGVM LPX is on the same order of magnitude as structural uncertainties, such as including or not including modules for shifting cultivation and wood harvest. The effect of the inclusion of additional land-use processes can even be compensated by a change of parametrization, while still conforming to the observational benchmarks, indicating that it might be possible to capture the magnitude of E_{LUC}, while neglecting second order processes. The compensation of E_{LUC} occurs because the residual sink is less sensitive to parametrization changes than the E_{LUC} in LPX-Bern. This behavior has also lead to an E_{LUC} that is on the lower end of independent estimates (Le Quéré et al., 2016). A lack of large difference in E_{LUC} from model setups featuring gross and net land-use might seem in contrast with the result of other studies investigating these processes (Arneth et al., 2017; Wilkenskjeld et al., 2014; Stocker et al., 2014; Shevliakova et al., 2009), however, if we keep parametrization constant ($\mathrm{M}_{gross,net}$) we find the expected lower E_{LUC} for net land-use.

We investigated the magnitude and spatial distribution of E_{LUC} in the model configuration using the skill scores and parametrization from the standard net land-use configuration with additional processes of shifting cultivation and wood harvest

($M_{gross,net}$). This choice is motivated by the good performance of the net land-use ensemble in the observational benchmarks (Section 3.2 and Figures S1-S14).

A good correspondence between simulated fluxes and the estimates of Houghton and Nassikas (2017) in 10 regions during the industrial period is found. When comparing recent decades, LPX-Bern generally seems to simulate lower E_{LUC} than both the bookkeeping approach based estimate and the aggregated estimates in the GCP. The biggest disparity is comparatively low fluxes in the South and Southeast Asian regions in LPX-Bern, which are at least partially explained by the lack of tropical peatlands in this model configuration. The burning and draining of tropical peatlands is an important contribution to E_{LUC} in tropical regions (Maria Roman-Cuesta et al., 2016; Koh et al., 2011; Hooijer et al., 2010). The annual emissions estimate from draining peatlands used in Houghton and Nassikas (2017), increase from almost no emissions in 1980 to roughly 0.2 PgC yr^{-1} in 2015. The lack of tropical peatlands is also consistent with the underestimated soil carbon density in these regions when compared to Carvalhais et al. (2014). Other studies suggest higher historical E_{LUC}, such as the bookkeeping approach by Hansis et al. (2015), including shifting cultivation, with an estimate of 261 PgC between 1850-2005. Some of the difference between DGVM model results and bookkeeping approaches can be attributed to different definitions of LULCC emission (Pongratz et al., 2014; Stocker and Joos, 2015).

A recent study by Li et al. (2017) constrained E_{LUC} by using biomass observations. They derived a relationship between E_{LUC} and biomass in nine regions using the nine DGVMs in the TRENDY-v2 model intercomparison (Sitch et al., 2015) and applied empirical estimates for biomass carbon to arrive at a constrained E_{LUC} of 155 ± 50 PgC between 1901 and 2012. The result of 116 (77, 156) PgC as in this study is compatible, albeit somewhat lower. By neglecting all other constraints and exclusively using the global vegetation carbon by IPCC (Ciais et al., 2013) and the biomass map by (Carvalhais et al., 2014) (Also used as one of the constraints in Li et al. (2017)) as constraints, we arrive at a higher E_{LUC} of 130 (87, 179) PgC. This illustrates the importance of the biomass inventory for the magnitude of E_{LUC}.

E_{LUC} is not only influenced by uncertain model processes and parametrizations but also the underlying ~~LULUC~~ LULCC forcings (Goll et al., 2015). Peng et al. (2017) have shown that the choice of transition rules, governing how new land-use areas are allocated from previous areas, has a considerable effect on E_{LUC}. The effects of these uncertainties are not accounted for in this study since we only use one land-cover forcing product and one set of transition rules is used.

Overall the ensemble approach produces E_{LUC} estimates consistent with other independent estimates, albeit somewhat on the lower end of the range of estimates. This is a consequence of the constraining process favoring parametrization with low E_{LUC} over a high residual sink, which is discussed further in the next section.

4.2 Benchmark performance and best guess version

A hierarchical weighting scheme to compare a diverse set of constraints was employed, following earlier work (Steinacher et al., 2013). A set of 14 data sets (Fig. 2, Table 2) was selected to constrain the model's performance with regard to steady state carbon and water fluxes and carbon inventories as well as with regard to transient changes. Globally aggregated as well as spatially resolved information is used to constrain simulated spatial patterns and to robustly model global mean properties. The temporal focus is on the decadal-to-century time scales most relevant for projections of anthropogenic climate-carbon cycle changes and on the seasonal cycle of photosynthesis and the decadal amplification of the seasonal cycle in land-atmosphere fluxes (McGuire et al., 2001; Graven et al., 2013) which provide information on underlying processes. The iterative procedure for choosing the prior parameter distribution yielded an ensemble which performs well with respect to the selected metrics.

In addition to the weighting of model results with the global skill score, we employed a minimum skill criterion, discarding runs with very bad performance in a singular metric. This approach is somewhat comparable to pre-calibration methods, where implausible parameter spaces are also ruled out (Williamson et al., 2017; Holden et al., 2010; Edwards et al., 2011), and aims to sensibly reduce the size of the parameter space.

While the uptake of carbon by the terrestrial biosphere in the model ensemble is significantly larger than earlier versions of LPX, it is still in the lower range of estimates. A direct way of increasing the magnitude of change in land carbon is to change pool sizes, which is here restricted by other observational constraints. The inclusion of more processes, such as natural and human-induced erosion (Wang et al., 2017) could also increase the strength of the terrestrial sink, however other processes such as shifting cultivation lead to a decrease of the land carbon sink. A further possibility is the revising of established processes in the model. The climatic dependence of the auto- and heterotrophic respiration is an important component, mitigating the CO_2 fertilization effect. The implementation of a more refined module might decrease this negative feedback, thus increasing carbon storage and sink sensitivity.

The sink strength could potentially also be enhanced by including so far not included parameters and including additional constraints that discriminate between the different components of the land sink.

Fossil carbon emissions and thus the net biome production and the carbon sink inferred from the deconvolution may be biased high for the most recent decades. The fossil emissions are estimated from fossil-fuel production data, which include the fraction used for non-combustion purposes such as the production of plastics and asphalt. Boden et al. (2017) assume non-fuel uses equal to zero (Andres et al., 2012) since the products will eventually be oxidized as well. Geyer et al. (2017) estimate that 8.3 Pg of plastics were produced between 1950 and 2015 of which 2.6 Pg were in use in 2015, 0.8 Gt incinerated and 4.9 Gt discarded. This implies that between 2.6 and 7.5 Pg plastic may still be left unoxidized. This is relatively small compared to the residual terrestrial sink, estimated to be around 69 (51, 93) PgC for the period from 1950 to 2015 ($M_{net,net}$ in Figure 3). However, about half of the plastic was produced since 2000 and estimated production is about 0.4 Pg/yr in 2015. In addition, about 0.1 Pg of bitumen asphalt is produced annually. Considering that most of the molecular weight of plastics is from carbon, fossil CO_2 emissions and in turn the terrestrial sink are biased high by up to 0.5 PgC/yr in 2015. This potential bias may be compared to the residual terrestrial sink flux of 1.2 (0.8,1.7) PgC/yr during 2005 to 2015. Interestingly the deconvolution of the atmospheric and fossil CO_2 records suggest a recent acceleration in the trend of the net biome production (Figure 11); this acceleration may also be biased high. In conclusion, considering plastic and asphalt products brings the most recent trends in the net biome production from the deconvolution versus the LPX model in better agreement, while estimates of net biome production and the terrestrial sinks are hardly affected before 2000 CE.

The release of both spatially and temporally resolved carbon flux observations by using remote sensing, such as the Carbon Monitoring System Flux Pilot (CMS) project, featuring not only net fluxes but also gross production and respiration, is a very promising candidate for constraining the parameter space further. The spatial structure might restrict the apparent degree of freedom in partitioning the terrestrial sink in E_{LUC} and residual land carbon sink. δ^{13}C isotope measurements in vegetation also have the potential to be a useful additional constraint in land biosphere models (Keller et al., 2017).

An other avenue of increasing model performance is to introduce spatially explicit parametrization, as used in multi-model averaging studies (Exbrayat et al., 2018; Schwalm et al., 2015). A caveat of using this approach with a single model is a potential overfitting of the parameters.

The simultaneous assimilation of multiple observational constraints allowed to formulate a well rounded best guess version of the model. While this parameter version doesn't necessarily excel at every single benchmark, it shows a consistent performance amongst all different targets. This behavior leads us to believe that the best-guess version is well suited for simulations spanning long time spans, both for paleo and future research questions, where the use of a full parameter ensemble is not feasible. Furthermore, it can also be used in model intercomparison studies, where single realizations of different models are compared.

5 Conclusions

We successfully applied a multi-purpose model benchmark to a perturbed parameter ensemble ~~, obtained with a Monte-Carlo like sampling technique,~~ of a dynamic global vegetation model (DGVM). Specifically, we developed a "best-guess" model version and constrained the residual carbon sink flux and carbon emissions from anthropogenic land-use (E_{LUC}) over the industrial period. The general characteristics of the framework are as follow. (i) The framework permits a standardized model benchmarking (Hoffman et al., 2017; Kelley et al., 2013; Luo et al., 2012; Blyth et al., 2011) by comparing different models or model versions graphically and using statistical metrics (Stow et al., 2009) to a broad and diverse range of observations. (ii) The efficient Latin Hypercube sampling method (McKay et al., 1979) is used to explore the model parameter space and to set up and run perturbed parameter ensembles for a large set of model parameters. The advantage of the Latin Hypercube sampling is the representative sampling of different parameter combinations, whereas a shortcoming is that the sampling size has to be determined in advance. (iii) A hierarchical model weighting scheme is used to assimilate diverse observations. These may differ with respect to spatial and temporal resolution and quality and include observations from the local scale, such as data from individual biomass measurements or the seasonal CO_2 cycle at individual atmospheric sampling sites, up to global scale gridded data products such as satellite measurements of absorbed radiation by plants. A major advantage of this scheme compared to sequential assimilation techniques such as Ensemble Kalman Filters is that the influence of necessarily subjective choices (Rougier, 2007) on the results can be investigated a posteriori; in other words without performing costly additional simulations. The subjective choices may be of scientific nature such as whether an observational data set is considered or not or of more technical nature such as whether gridded data values are weighted by grid cell area or not. (iv) The applied modular framework is easily extendable to incorporate different or more observational constraints and to different mechanistic models including other DGVMs, ocean models (Battaglia et al., 2016) or Earth System Models (Steinacher et al., 2013; Steinacher and Joos, 2016)). (v) The Bayesian, skill-score weighted ensemble is able to constrain the median and uncertainty ranges of unknown or uncertain quantities such as carbon emissions from anthropogenic land-use, marine nitrous oxide production

~~(Battaglia and Joos, 2017)~~(Battaglia and Joos, 2018), or climate sensitivity metrics (Steinacher and Joos, 2016) (vi) Finally, the skill-score weighted ensemble is suitable for probabilistic projections including both likely and less likely model configurations and assumptions.

A new reference version of the LPX-Bern (v1.4) DGVM was established. We were able to show that the constrained ensemble, as well as a resulting best guess version, perform consistently well under a range of benchmarks (Table 2) while satisfying a minimum skill criterion in every single benchmark. The ~~best guess version was formulated using the weighted median parameter values of the constrained ensemble, instead of using the parametrization of the overall best run. This choice is motivated by the robustness of the resulting parameter values with respect to changes in the hierarchical weighting scheme and the in- or exclusion of individual observational targets and its representativeness of the perturbed parameter ensemble. The~~ new model version LPX-Bern v1.4 successfully simulates observation-based estimates of the cumulative net land uptake and release over the industrial period.

Many previous studies have investigated inherent uncertainties in ELUC estimates (Houghton et al., 2012; Goll et al., 2015; Peng et al., 2017). Our study aims to contribute to this ongoing discussion by providing DGVM E_{LUC} uncertainty estimates purely due to parameter uncertainty in an observationally constrained model ensemble using the LUH2 v2h (Hurtt et al., 2018) product. Overall the benchmarking scheme favors runs with low emissions due to a relatively low residual sink sensitivity in the model and constraining total land-atmosphere fluxes. We consider model ensembles with and without additional land-use processes (shifting cultivation and wood harvest) and find that the difference in global E_{LUC} is on the same order of magnitude as parameter induced uncertainty. The inclusion of shifting cultivation and wood harvesting increases emissions similar in magnitude to earlier studies (Stocker et al., 2014; Shevliakova et al., 2009) when applying the same model parameters, while in some cases these additional emissions could potentially even be offset with appropriate parameter choice. We attributed the fluxes to different countries and closer investigated the ten countries with the most emissions in the industrial period due to land-use and land-use change. Our land-use carbon emission estimates are similar to those of Houghton and Nassikas (2017) on the country level and overall consistent with other independent estimates on regional to global levels (Li et al., 2017; Le Quéré et al., 2016).

The observation-constrained DGVM ensemble and best guess version established in this work are ready for use in model intercomparison studies (Tian et al., 2018; Sitch et al., 2015) and longer time span paleo simulations. It may also be applied to quantify future terrestrial carbon fluxes and E_{LUC} for different shared socio-economic pathways. Additional new observational data streams may be implemented in our modular framework to further refine results.

Data availability. Model output is available upon request to the corresponding author (lienert@climate.unibe.ch).

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

Acknowledgements. We thank G. Battaglia for supplying the Bern3D model output and M. Scholze for providing the TM2 transport matrices. We would like to thank the data community for their efforts in providing high quality data sets. This work was supported by the Swiss National Science Foundation (#200020_172476).

References

Andres, R. J., Boden, T. A., Bréon, F. M., Ciais, P., Davis, S., Erickson, D., Gregg, J. S., Jacobson, A., Marland, G., Miller, J., Oda, T., Olivier, J. G., Raupach, M. R., Rayner, P., and Treanton, K.: A synthesis of carbon dioxide emissions from fossil-fuel combustion, Biogeosciences, 9, 1845–1871, https://doi.org/10.5194/bg-9-1845-2012, 2012.

Arneth, A., Sitch, S., Pongratz, J., Stocker, B. D., Ciais, P., Poulter, B., Bayer, A., Bondeau, A., Calle, L., Chini, L., Gasser, T., Fader, M., Friedlingstein, P., Kato, E., Li, W., Lindeskog, M., Nabel, J. E. M. S., Pugh, T. A. M., Robertson, E., Viovy, N., Yue, C., and Zaehle, S.: Historical carbon dioxide emissions due to land use changes possibly larger than assumed, Nature Geoscience, https://doi.org/10.1038/ngeo2882, 2017.

Ballantyne, A. P., Andres, R., Houghton, R., Stocker, B. D., Wanninkhof, R., Anderegg, W., Cooper, L. A., DeGrandpre, M., Tans, P. P., Miller, J. B., Alden, C., and White, J. W.: Audit of the global carbon budget: Estimate errors and their impact on uptake uncertainty, Biogeosciences, 12, 2565–2584, https://doi.org/10.5194/bg-12-2565-2015, 2015.

Battaglia, G. and Joos, F.: Hazards of decreasing marine oxygen : the near-term and millennial-scale benefits of meeting the Paris climate targets, pp. 1–21, 2017.

Battaglia, G. and Joos, F.: Marine N2O Emissions From Nitrification and Denitrification Constrained by Modern Observations and Projected in Multimillennial Global Warming Simulations, Global Biogeochemical Cycles, 32, 92–121, https://doi.org/10.1002/2017GB005671, 2018.

Battaglia, G., Steinacher, M., and Joos, F.: A probabilistic assessment of calcium carbonate export and dissolution in the modern ocean, Biogeosciences, 13, 2823–2848, https://doi.org/10.5194/bg-13-2823-2016, 2016.

Beven, K. and Binley, A.: The future of distributed models: Model calibration and uncertainty prediction, Hydrological Processes, 6, 279–298, https://doi.org/10.1002/hyp.3360060305, 1992.

Billings, S. A., Buddemeier, R. W., De, D., Van Oost, K., and Bohling, G.: A simple method for estimating the influence of eroding soil profiles on atmospheric CO2, Global Biogeochemical Cycles, 24, 1–14, https://doi.org/10.1029/2009GB003560, 2010.

Blyth, E., Clark, D. B., Ellis, R., Huntingford, C., Los, S., Pryor, M., Best, M., and Sitch, S.: A comprehensive set of benchmark tests for a land surface model of simultaneous fluxes of water and carbon at both the global and seasonal scale, Geoscientific Model Development, 4, 255–269, https://doi.org/10.5194/gmd-4-255-2011, 2011.

Boden, T., Marland, G., and Andres, R.: Global, Regional, and National Fossil-Fuel CO2 Emissions, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tenn., U.S.A., https://doi.org/10.3334/CDIAC/00001_V2017, 2017.

Brovkin, V., Boysen, L., Arora, V. K., Boisier, J. P., Cadule, P., Chini, L., Claussen, M., Friedlingstein, P., Gayler, V., Van den hurk, B. J., Hurtt, G. C., Jones, C. D., Kato, E., De noblet ducoudre, N., Pacifico, F., Pongratz, J., and Weiss, M.: Effect of anthropogenic land-use and land-cover changes on climate and land carbon storage in CMIP5 projections for the twenty-first century, Journal of Climate, 26, 6859–6881, https://doi.org/10.1175/JCLI-D-12-00623.1, 2013.

Carvalhais, N., Forkel, M., Khomik, M., Bellarby, J., Jung, M., Migliavacca, M., Mgru, M., Saatchi, S., Santoro, M., Thurner, M., Weber, U., Ahrens, B., Beer, C., Cescatti, A., Randerson, J. T., and Reichstein, M.: Global covariation of carbon turnover times with climate in terrestrial ecosystems, Nature, 514, 213–217, https://doi.org/10.1038/nature13731, 2014.

Ciais, P., Sabine, C., Bala, G., Bopp, L., Brovkin, V., Canadell, J., Chhabra, A., DeFries, R., Galloway, J., Heimann, M., Jones, C., Quéré, C. L., Myneni, R. B., Piao, S., and Thornton, P.: Carbon and Other Biogeochemical Cycles, Climate Change 2013 - The Physical Science Basis, pp. 465–570, https://doi.org/10.1017/CBO9781107415324.015, 2013.

DeVries, T.: Global Biogeochemical Cycles The oceanic anthropogenic CO 2 sink : Storage , air-sea fluxes , and transports over the industrial era, pp. 1–17, https://doi.org/10.1002/2013GB004739.Received, 2014.

Edwards, N. R., Cameron, D., and Rougier, J.: Precalibrating an intermediate complexity climate model, Climate Dynamics, 37, 1469–1482, https://doi.org/10.1007/s00382-010-0921-0, 2011.

Exbrayat, J.-F., Bloom, A. A., Falloon, P., Ito, A., Smallman, T. L., and Williams, M.: Reliability ensemble averaging of 21st century projections of terrestrial net primary productivity reduces global and regional uncertainties, Earth System Dynamics, 9, 153–165, https://doi.org/10.5194/esd-9-153-2018, https://www.earth-syst-dynam.net/9/153/2018/, 2018.

Gerber, M. and Joos, F.: An Ensemble Kalman Filter multi-tracer assimilation: Determining uncertain ocean model parameters for improved climate-carbon cycle projections, Ocean Modelling, 64, 29–45, https://doi.org/10.1016/j.ocemod.2012.12.012, 2013.

Gerber, S., Hedin, L. O., Keel, S. G., Pacala, S. W., and Shevliakova, E.: Land use change and nitrogen feedbacks constrain the trajectory of the land carbon sink, Geophysical Research Letters, 40, 5218–5222, https://doi.org/10.1002/grl.50957, 2013.

Geyer, R., Jambeck, J. R., and Law, K. L.: Production, use, and fate of all plastics ever made, Science Advances, 3, e1700 782, https://doi.org/10.1126/sciadv.1700782, http://advances.sciencemag.org/lookup/doi/10.1126/sciadv.1700782, 2017.

GLOBALVIEW-CO2: Cooperative Global Atmospheric Data Integration Project. 2013, updated annually. Multi-laboratory compilation of synchronized and gap-filled atmospheric carbon dioxide records for the period 1979-2012 (obspack_co2_1_GLOBALVIEW-CO2_2013_v1.0.4_2013-12-23)., 2013.

Gobron, N., Pinty, B., Aussedat, O., Chen, J. M., Cohen, W. B., Fensholt, R., Gond, V., Huemmrich, K. F., Lavergne, T., Mélin, F., Privette, J. L., Sandholt, I., Taberner, M., Turner, D. P., Verstraete, M. M., and Widlowski, J. L.: Evaluation of fraction of absorbed photosynthetically active radiation products for different canopy radiation transfer regimes: Methodology and results using Joint Research Center products derived from SeaWiFS against ground-based estimations, Journal of Geophysical Research Atmospheres, 111, https://doi.org/10.1029/2005JD006511, 2006.

Goll, D. S., Brovkin, V., Liski, J., Raddatz, T., Thum, T., and Todd-Brown, K. E.: Strong dependence of CO 2 emissions from anthropogenic land cover change on soil carbon parametrization and initial land cover, Global Biogeochemical Cycles, 29, 1–58, https://doi.org/10.1002/2014GB004988.Abstract, 2015.

Graven, A. H. D., Keeling, R. F., Piper, S. C., Patra, P. K., and Stephens, B. B.: Title : Enhanced seasonal exchange of CO 2 by northern ecosystems since 1960, 146, 1085–1090, 2013.

Hansis, E., Davis, S. J., and Pongratz, J.: Relevance of methodological choices for accounting of land use change carbon fluxes Special Section :, pp. 1–17, https://doi.org/10.1002/2014GB004997.Received, 2015.

Harris, I., Jones, P. D., Osborn, T. J., and Lister, D. H.: Updated high-resolution grids of monthly climatic observations - the CRU TS3.10 Dataset, International Journal of Climatology, 34, 623–642, https://doi.org/10.1002/joc.3711, 2014.

Hoffman, F. M., Koven, C. D., Keppel-Aleks, G., Lawrence, D. M., Riley, W. J., Randerson, J. T., Ahlström, A., Abramowitz, G., Baldocchi, D. D., Best, M. J., Bond-Lamberty, Kauwe, B. M. G. D., Denning, A. S., Desai, A. R., Eyring, V., Fisher, J. B., Fisher, R. A., Gleckler, P. J., Huang, M., Hugelius, G., Jain, A. K., Kiang, N. Y., Kim, H., Koster, R. D., Kumar, S. V., Li, H., Luo, Y., Mao, J., McDowell, N. G., Mishra, U., Moorcroft, P. R., Pau, G. S. H., Ricciuto, D. M., Schaefer, K., Schwalm, C. R., Serbin, S. P., Shevliakova, E., Slater, A. G., Tang, J., Williams, M., Xia, J., Xu, C., Joseph, R., and D. Koch: International Land Model Benchmarking (ILAMB) 2016 Workshop Report, pp. 1–, https://doi.org/10.2172/1330803., 2017.

Hoffmann, T., Mudd, S. M., Van Oost, K., Verstraeten, G., Erkens, G., Lang, A., Middelkoop, H., Boyle, J., Kaplan, J. O., Willenbring, J., and Others: Humans and the missing C-sink: erosion and burial of soil carbon through time, Earth Surface Dynamics, 1, 45, 2013.

Holden, P. B., Edwards, N. R., Oliver, K. I. C., Lenton, T. M., and Wilkinson, R. D.: A probabilistic calibration of climate sensitivity and terrestrial carbon change in GENIE-1, Climate Dynamics, 35, 785–806, https://doi.org/10.1007/s00382-009-0630-8, 2010.

Hooijer, A., Page, S., Canadell, J. G., Silvius, M., Kwadijk, J., Wösten, H., and Jauhiainen, J.: Current and future CO 2 emissions from drained peatlands in Southeast Asia, Biogeosciences, 7, 1505–1514, https://doi.org/10.5194/bg-7-1505-2010, 2010.

Houghton, R. A. and Nassikas, A. A.: Global and Regional Fluxes of Carbon from Land Use and Land-Cover Change 1850-2015, Global Biogeochemical Cycles, pp. 456–472, https://doi.org/10.1002/2016GB005546, 2017.

Houghton, R. A., House, J. I., Pongratz, J., Van Der Werf, G. R., Defries, R. S., Hansen, M. C., Le Quéré, C., and Ramankutty, N.: Carbon emissions from land use and land-cover change, Biogeosciences, 9, 5125–5142, https://doi.org/10.5194/bg-9-5125-2012, 2012.

Houweling, S., Kaminski, T., Dentener, F., Lelieveld, J., and Heimann, M.: Inverse modeling of methane sources and sinks using the adjoint of a global transport model Sander global methane emissions, Journal of Geophysical Research, 104, 26 137–26 160, https://doi.org/10.1029/1999JD900428, 1999.

Hurtt, G., Chini, L., Sahajpa, R., and Frolking, S.: Harmonization of global land-use change and management for the period 850-2100 (In prep), Geoscientific Model Development, 2018.

Joos, F. and Spahni, R.: Rates of change in natural and anthropogenic radiative forcing over the past 20 , 000 years, Proceedings of the National Academy of Sciences of the United States of America, 105, 1425–1430, https://doi.org/10.1073/pnas.0707386105, 2007.

Kaminski, T., Heimann, M., and Giering, R.: A coarse grid three-dimensional global inverse model of the atmospheric transport: 2. Inversion of the transport of CO 2 in the 1980s, Journal of Geophysical Research, 104, 18 555, https://doi.org/10.1029/1999JD900146, 1999.

Kaminski, T., Knorr, W., Schürmann, G., Scholze, M., Rayner, P. J., Zaehle, S., Blessing, S., Dorigo, W., Gayler, V., Giering, R., Gobron, N., Grant, J. P., Heimann, M., Hooker-Stroud, A., Houweling, S., Kato, T., Kattge, J., Kelley, D., Kemp, S., Koffi, E. N., Köstler, C., Mathieu, P. P., Pinty, B., Reick, C. H., Rödenbeck, C., Schnur, R., Scipal, K., Sebald, C., Stacke, T., Van Scheltinga, A. T., Vossbeck, M., Widmann, H., and Ziehn, T.: The BETHY/JSBACH Carbon Cycle Data Assimilation System: Experiences and challenges, Journal of Geophysical Research: Biogeosciences, 118, 1414–1426, https://doi.org/10.1002/jgrg.20118, 2013.

Kato, T., Knorr, W., Scholze, M., Veenendaal, E., Kaminski, T., Kattge, J., and Gobron, N.: Simultaneous assimilation of satellite and eddy covariance data for improving terrestrial water and carbon simulations at a semi-arid woodland site in Botswana, Biogeosciences, 10, 789–802, https://doi.org/10.5194/bg-10-789-2013, 2013.

Keith, H., Mackey, B. G., and Lindenmayer, D. B.: Re-evaluation of forest biomass carbon stocks and lessons from the world's most carbon-dense forests., Proceedings of the National Academy of Sciences of the United States of America, 106, 11 635–11 640, https://doi.org/10.1073/pnas.0901970106, 2009.

Keller, K. M., Lienert, S., Bozbiyik, A., Stocker, T. F., Churakova (Sidorova), O. V., Frank, D. C., Klesse, S., Koven, C. D., Leuenberger, M., Riley, W. J., Saurer, M., Siegwolf, R., Weigt, R. B., and Joos, F.: 20th century changes in carbon isotopes and water-use efficiency: tree-ring-based evaluation of the CLM4.5 and LPX-Bern models, Biogeosciences, 14, 2641–2673, https://doi.org/10.5194/bg-14-2641-2017, 2017.

Kelley, D. I., Prentice, I. C., Harrison, S. P., Wang, H., Simard, M., Fisher, J. B., and Willis, K. O.: A comprehensive benchmarking system for evaluating global vegetation models, Biogeosciences, 10, 3313–3340, https://doi.org/10.5194/bg-10-3313-2013, 2013.

Khatiwala, S., Tanhua, T., Mikaloff Fletcher, S., Gerber, M., Doney, S. C., Graven, H. D., Gruber, N., McKinley, G. A., Murata, A., Ríos, A. F., and Sabine, C. L.: Global ocean storage of anthropogenic carbon, Biogeosciences, 10, 2169–2191, https://doi.org/10.5194/bg-10-2169-2013, 2013.

Koh, L. P., Miettinen, J., Liew, S. C., and Ghazoul, J.: Remotely sensed evidence of tropical peatland conversion to oil palm, Proceedings of the National Academy of Sciences, 108, 5127–5132, https://doi.org/10.1073/pnas.1018776108, 2011.

Kosmas, C., Giraldez, J. V., da Silva, J. R. M., Merckx, R., Van Oost, K., Quine, T. A., Govers, G., De Gryze, S., Six, J., Harden, J. W., Ritchie, J. C., McCarty, G. W., and Heckrath, G.: The Impact of Agricultural Soil Erosion on the Global Carbon Cycle, Science, 318, 626–629, https://doi.org/10.1126/science.1145724, 2007.

Lamarque, J.-F., Dentener, F., McConnell, J., Ro, C.-U., Shaw, M., Vet, R., Bergmann, D., Cameron-Smith, P., Doherty, R., Faluvegi, G., Ghan, S. J., Josse, B., Lee, Y. H., MacKenzie, I. A., Plummer, D., Shindell, D. T., Stevenson, D. S., Strode, S., and Zeng, G.: Multi-model mean nitrogen and sulfur deposition from the Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP): evaluation historical and projected changes, Atmospheric Chemistry and Physics Discussions, 13, 6247–6294, https://doi.org/10.5194/acpd-13-6247-2013, 2013.

Le Quéré, C., Andrew, R. M., Canadell, J. G., Sitch, S., Ivar Korsbakken, J., Peters, G. P., Manning, A. C., Boden, T. A., Tans, P. P., Houghton, R. A., Keeling, R. F., Alin, S., Andrews, O. D., Anthoni, P., Barbero, L., Bopp, L., Chevallier, F., Chini, L. P., Ciais, P., Currie, K., Delire, C., Doney, S. C., Friedlingstein, P., Gkritzalis, T., Harris, I., Hauck, J., Haverd, V., Hoppema, M., Klein Goldewijk, K., Jain, A. K., Kato, E., Körtzinger, A., Landschützer, P., Lefèvre, N., Lenton, A., Lienert, S., Lombardozzi, D., Melton, J. R., Metzl, N., Millero, F., Monteiro, P. M. S., Munro, D. R., Nabel, J. E. M. S., Nakaoka, S. I., O'Brien, K., Olsen, A., Omar, A. M., Ono, T., Pierrot, D., Poulter, B., Rödenbeck, C., Salisbury, J., Schuster, U., Schwinger, J., Séferian, R., Skjelvan, I., Stocker, B. D., Sutton, A. J., Takahashi, T., Tian, H., Tilbrook, B., Van Der Laan-Luijkx, I. T., Van Der Werf, G. R., Viovy, N., Walker, A. P., Wiltshire, A. J., and Zaehle, S.: Global Carbon Budget 2016, Earth System Science Data, 8, 605–649, https://doi.org/10.5194/essd-8-605-2016, 2016.

Li, W., Ciais, P., Peng, S., Yue, C., Wang, Y., Thurner, M., Saatchi, S. S., Arneth, A., Avitabile, V., Carvalhais, N., Harper, A. B., Kato, E., Koven, C., Liu, Y. Y., Nabel, J. E. M. S., Pan, Y., Pongratz, J., Poulter, B., Pugh, T. A. M., Santoro, M., Sitch, S., Stocker, B. D., Viovy, N., Wiltshire, A., Yousefpour, R., and Zaehle, S.: Land-use and land-cover change carbon emissions between 1901 and 2012 constrained by biomass observations, Biogeosciences Discuss, 5194, 2017–186, https://doi.org/10.5194/bg-2017-186, 2017.

Lorenc, A. C.: The potential of the ensemble Kalman filter for NWP—a comparison with 4D-Var, Quarterly Journal of the Royal Meteorological Society, 129, 3183–3203, https://doi.org/10.1256/qj.02.132, 2003.

Luo, Y. Q., Randerson, J. T., Abramowitz, G., Bacour, C., Blyth, E., Carvalhais, N., Ciais, P., Dalmonech, D., Fisher, J. B., Fisher, R., Friedlingstein, P., Hibbard, K., Hoffman, F., Huntzinger, D., Jones, C. D., Koven, C., Lawrence, D., Li, D. J., Mahecha, M., Niu, S. L., Norby, R., Piao, S. L., Qi, X., Peylin, P., Prentice, I. C., Riley, W., Reichstein, M., Schwalm, C., Wang, Y. P., Xia, J. Y., Zaehle, S., and Zhou, X. H.: A framework for benchmarking land models, Biogeosciences, 9, 3857–3874, https://doi.org/10.5194/bg-9-3857-2012, 2012.

Luyssaert, S., Inglima, I., Jung, M., Richardson, A. D., Reichstein, M., Ppapale, D., Piao, S. L., Schulze, E. .-D. D., Wingate, L., Matteucci, G., Aragao, L., Aubinet, M., Beer, C., Bernhofer, C., Black, K. G., Bonal, D., Bonnefond, J. .-M. M. .-M., Chambers, J., Ciais, P., Cook, B., Davis, K. J., Dolman, A. J., Gielen, B., Goulden, M., Grace, J., Granier, A., Grelle, A., Griffis, T., Grünwald, T., Guidolotti, G., Hanson, P. J., Harding, R., Hollinger, D. Y., Hutyra, L. R., Kolari, P., Kruijt, B., Kutsch, W., Lagergren, F., Laurila, T., Law, B. E., Le Maire, G., Lindroth, A., Loustau, D., Malhi, Y., Mateus, J., Migliavacca, M., Misson, L., Montagnani, L., Moncrieff, J., Moors, E., Munger, J. W., Nikinmaa, E., Ollinger, S. V., Pita, G., Rebmann, C., Roupsard, O., Saigusa, N., Sanz, M. J., G., S., C., S., Smith, M. .-L. L., Tang, J., Valentini, R., Vesala, T., and Janssens, I. A.: CO2 balance of boreal, temperate, and tropical forests derived from a global database, Global Change Biol., 13, 2509–2537, https://doi.org/10.1111/j.1365-2486.2007.01439.x, 2007.

Luyssaert, S., Inglima, I., and Jung, M.: Global Forest Ecosystem Structure and Function Data For Carbon Balance Research, https://doi.org/10.3334/ORNLDAAC/949, 2009.

Ma, R., Zhang, L., Tian, X., Zhang, J., Yuan, W., Zheng, Y., Zhao, X., and Kato, T.: Assimilation of remotely-sensed leaf area index into a dynamic vegetation model for gross primary productivity estimation, Remote Sensing, 9, https://doi.org/10.3390/rs9030188, 2017.

Mahowald, N. M., Randerson, J. T., Lindsay, K., Munoz, E., Doney, S. C., Lawrence, P., Schlunegger, S., Ward, D. S., Lawrence, D., and Hoffman, F. M.: Interactions between land use change and carbon cycle feedbacks, Global Biogeochemical Cycles, 31, 96–113, https://doi.org/10.1002/2016GB005374, 2017.

Maria Roman-Cuesta, R., Rufino, M., Herold, M., Butterbach-Bahl, K., Rosenstock, T. S., Herrero, M., Ogle, S., Li, C., Poulter, B., Verchot, L., Martius, C., Stuiver, J., and De Bruin, S.: Hotspots of gross emissions from the land use sector: Patterns, uncertainties, and leading emission sources for the period 2000-2005 in the tropics, Biogeosciences, 13, 4253–4269, https://doi.org/10.5194/bg-13-4253-2016, 2016.

McGuire, a. D., Sitch, S., Clein, J. S., Dargaville, R., Esser, G., Foley, J., Heimann, M., Joos, F., Kaplan, J., Kicklighter, D. W., Meier, R. a., Melillo, J. M., Moore, B., Prentice, I. C., Ramankutty, N., Reichenau, T., Schloss, A., Tian, H., Williams, L. J., and Wittenberg, U.: Carbon balance of the terrestrial biosphere in the Twentieth Century: Analyses of CO 2 , climate and land use effects with four process-based ecosystem models, Global Biogeochemical Cycles, 15, 183–206, https://doi.org/10.1029/2000GB001298, 2001.

McKay, M. D., Beckman, R. J., and Conover, W. J.: Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code, Technometrics, 21, 239–245, https://doi.org/10.1080/00401706.1979.10489755, 1979.

Meure, C. M., Etheridge, D., Trudinger, C., Steele, P., Langenfelds, R., van Ommen, T., Smith, A., and Elkins, J.: Law Dome CO2, CH4 and N2O ice core records extended to 2000 years BP, Geophysical Research Letters, 33, https://doi.org/10.1029/2006GL026152, 2006.

Mueller, B., Hirschi, M., Jimenez, C., Ciais, P., Dirmeyer, P. A., Dolman, A. J., Fisher, J. B., Jung, M., Ludwig, F., Maignan, F., Miralles, D. G., McCabe, M. F., Reichstein, M., Sheffield, J., Wang, K., Wood, E. F., Zhang, Y., and Seneviratne, S. I.: Benchmark products for land evapotranspiration: LandFlux-EVAL multi-data set synthesis, Hydrology and Earth System Sciences, 17, 3707–3720, https://doi.org/10.5194/hess-17-3707-2013, 2013.

Olson, R. J., Scurlock, J. M. O., Prince, S. D., Zheng, D. L., and Johnson, K. R.: NPP Multi-Biome: NPP and Driver Data for Ecosystem Model-data Intercomparison, R2, https://doi.org/10.3334/ornldaac/615, 2013.

Peng, S., Ciais, P., Chevallier, F., Peylin, P., Cadule, P., Sitch, S., Piao, S., Ahlström, A., Huntingford, C., Levy, P., Li, X., Liu, Y., Lomas, M., Poulter, B., Viovy, N., Wang, T., and Wang, X.: Benchmarking the seasonal cycle of CO2 fluxes simulated by terrestrial ecosystem models, pp. 1–19, https://doi.org/10.1002/2014GB004931, 2014.

Peng, S., Ciais, P., Maignan, F., Li, W., Chang, J., Wang, T., and Yue, C.: Sensitivity of land use change emission estimates to historical land use and land cover mapping, Global Biogeochemical Cycles, 31, 626–643, https://doi.org/10.1002/2015GB005360, 2017.

Pongratz, J. and Caldeira, K.: Attribution of atmospheric CO2 and temperature increases to regions: Importance of preindustrial land use change, Environmental Research Letters, 7, https://doi.org/10.1088/1748-9326/7/3/034001, 2012.

Pongratz, J., Caldeira, K., Reick, C. H., and Claussen, M.: Coupled climate-carbon simulations indicate minor global effects of wars and epidemics on atmospheric CO2 between AD 800 and 1850, The Holocene, 21, 843–851, https://doi.org/10.1177/0959683610386981, http://hol.sagepub.com/cgi/doi/10.1177/0959683610386981, 2011.

Pongratz, J., Reick, C. H., Houghton, R. A., and House, J. I.: Terminology as a key uncertainty in net land use and land cover change carbon flux estimates, Earth System Dynamics, 5, 177–195, https://doi.org/10.5194/esd-5-177-2014, 2014.

Randerson, J. T., Hoffman, F. M., Thornton, P. E., Mahowald, N. M., Lindsay, K., Lee, Y. H., Nevison, C. D., Doney, S. C., Bonan, G., Stöckli, R., Covey, C., Running, S. W., and Fung, I. Y.: Systematic assessment of terrestrial biogeochemistry in coupled climate-carbon models, Global Change Biology, 15, 2462–2484, https://doi.org/10.1111/j.1365-2486.2009.01912.x, 2009.

Ritz, S. P., Stocker, T. F., Joos, F., Ritz, S. P., Stocker, T. F., and Joos, F.: A Coupled Dynamical Ocean–Energy Balance Atmosphere Model for Paleoclimate Studies, Journal of Climate, 24, 349–375, https://doi.org/10.1175/2010JCLI3351.1, 2011.

Roman-Cuesta, R. M., Herold, M., Rufino, M. C., Rosenstock, T. S., Houghton, R. A., Rossi, S., Butterbach-Bahl, K., Ogle, S., Poulter, B., Verchot, L., Martius, C., and De Bruin, S.: Multi-gas and multi-source comparisons of six land use emission datasets and AFOLU estimates in the Fifth Assessment Report, for the tropics for 2000-2005, Biogeosciences, 13, 5799–5819, https://doi.org/10.5194/bg-13-5799-2016, 2016.

Roth, R., Ritz, S. P., and Joos, F.: Burial-nutrient feedbacks amplify the sensitivity of atmospheric carbon dioxide to changes in organic matter remineralisation, Earth System Dynamics, 5, 321–343, https://doi.org/10.5194/esd-5-321-2014, 2014.

Rougier, J.: Probabilistic inference for future climate using an ensemble of climate model evaluations, Climatic Change, 81, 247–264, https://doi.org/10.1007/s10584-006-9156-9, 2007.

Rubino, M., Etheridge, D. M., Trudinger, C. M., Allison, C. E., Battle, M. O., Langenfelds, R. L., Steele, L. P., Curran, M., Bender, M., White, J. W. C., Jenk, T. M., Blunier, T., and Francey, R. J.: A revised 1000 year atmospheric δ 13 C-CO 2 record from Law Dome and South Pole, Antarctica, Journal of Geophysical Research: Atmospheres, 118, 8482–8499, https://doi.org/10.1002/jgrd.50668, 2013.

Schimel, D., Stephens, B. B., and Fisher, J. B.: Effect of increasing CO2 on the terrestrial carbon cycle, Proceedings of the National Academy of Sciences, 112, 436–441, https://doi.org/10.1073/pnas.1407302112, 2015.

Schmittner, A., Urban, N. M., Keller, K., and Matthews, D.: Using tracer observations to reduce the uncertainty of ocean diapycnal mixing and climate-carbon cycle projections, Global Biogeochemical Cycles, 23, https://doi.org/10.1029/2008GB003421, 2009.

Schwalm, C. R., Huntinzger, D. N., Fisher, J. B., Michalak, A. M., Bowman, K., Cook, R., El-masri, B., Hayes, D., Huang, M., Jacobson, A., Jain, A., King, A. W., Lei, H., and Liu, J.: Toward “ optimal ” integration of terrestrial biosphere model ensembles, Geophysical Research Letters, 42, 4418–4428, https://doi.org/10.1002/2015GL064002.Received, 2015.

Shevliakova, E., Pacala, S. W., Malyshev, S., Hurtt, G. C., Milly, P. C. D., Caspersen, J. P., Sentman, L. T., Fisk, J. P., Wirth, C., and Crevoisier, C.: Carbon cycling under 300 years of land use change: Importance of the secondary vegetation sink, Global Biogeochem. Cycles, 23, 1–16, https://doi.org/10.1029/2007GB003176, 2009.

Sitch, S., Smith, B., Prentice, I. C., Arneth, A., Bondeau, A., Cramer, W., Kaplan, J. O., Levis, S., Lucht, W., Sykes, M. T., Thonicke, K., and Venevsky, S.: Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model, Global Change Biology, 9, 161–185, https://doi.org/10.1046/j.1365-2486.2003.00569.x, 2003.

Sitch, S., Friedlingstein, P., Gruber, N., Jones, S. D., Murray-Tortarolo, G., Ahlström, A., Doney, S. C., Graven, H., Heinze, C., Huntingford, C., Levis, S., Levy, P. E., Lomas, M., Poulter, B., Viovy, N., Zaehle, S., Zeng, N., Arneth, A., Bonan, G., Bopp, L., Canadell, J. G., Chevallier, F., Ciais, P., Ellis, R., Gloor, M., Peylin, P., Piao, S. L., Le Quéré, C., Smith, B., Zhu, Z., and Myneni, R.: Recent trends and drivers of regional sources and sinks of carbon dioxide, Biogeosciences, 12, 653–679, https://doi.org/10.5194/bg-12-653-2015, 2015.

Spahni, R., Joos, F., Stocker, B. D., Steinacher, M., and Yu, Z. C.: Transient simulations of the carbon and nitrogen dynamics in northern peatlands: from the Last Glacial Maximum to the 21st century, Climate of the Past, 9, 1287–1308, https://doi.org/10.5194/cp-9-1287-2013, 2013.

Steinacher, M. and Joos, F.: Transient Earth system responses to cumulative carbon dioxide emissions: Linearities, uncertainties, and probabilities in an observation-constrained model ensemble, Biogeosciences, 13, 1071–1103, https://doi.org/10.5194/bg-13-1071-2016, 2016.

Steinacher, M., Joos, F., and Stocker, T. F.: Allowable carbon emissions lowered by multiple climate targets, Nature, 499, 197–201, https://doi.org/10.1038/nature12269, 2013.

Stocker, B., Feissli, F., and Strassmann, K.: Past and future carbon fluxes from land use change, shifting cultivation and wood harvest, Tellus B, 1, 1–15, https://doi.org/10.3402/tellusb.v66.23188, 2014.

Stocker, B. D. and Joos, F.: Quantifying differences in land use emission estimates implied by definition discrepancies, Earth System Dynamics, 6, 731–744, https://doi.org/10.5194/esd-6-731-2015, 2015.

Stocker, B. D., Roth, R., Joos, F., Spahni, R., Steinacher, M., Zaehle, S., Bouwman, L., Xu-Ri, and Prentice, I. C.: Multiple greenhouse-gas feedbacks from the land biosphere under future climate change scenarios, Nature Climate Change, 3, 666–672, https://doi.org/10.1038/nclimate1864, 2013.

Stocker, B. D., Yu, Z., Massa, C., and Joos, F.: Holocene peatland and ice-core data constraints on the timing and magnitude of CO2 emissions from past land use, Proceedings of the National Academy of Sciences, 114, 1492–1497, https://doi.org/10.1073/pnas.1613889114, 2017.

Stöckli, R., Rutishauser, T., Baker, I., Liniger, M. A., and Denning, A. S.: A global reanalysis of vegetation phenology, Journal of Geophysical Research: Biogeosciences, 116, 1–19, https://doi.org/10.1029/2010JG001545, 2011.

Stow, C. A., Jolliff, J., McGillicuddy, D. J., Doney, S. C., Allen, J. I., Friedrichs, M. A., Rose, K. A., and Wallhead, P.: Skill assessment for coupled biological/physical models of marine systems, Journal of Marine Systems, 76, 4–15, https://doi.org/10.1016/j.jmarsys.2008.03.011, 2009.

Strassmann, K. M., Joos, F., and Fischer, G.: Simulating effects of land use changes on carbon fluxes: Past contributions to atmospheric CO2 increases and future commitments due to losses of terrestrial sink capacity, Tellus, Series B: Chemical and Physical Meteorology, 60, 583–603, https://doi.org/10.1111/j.1600-0889.2008.00340.x, 2008.

Tans, P. and Keeling, R.: NOAA/ESRL and Scripps Institution of Oceanography, www.esrl.noaa.gov/gmd/ccgg/trends/scrippsco2.ucsd.edu/.

Tarnocai, C., Canadell, J. G., Schuur, E. A. G., Kuhry, P., Mazhitova, G., and Zimov, S.: Soil organic carbon pools in the northern circumpolar permafrost region, Global Biogeochem. Cycles, 23, https://doi.org/10.1029/2008GB003327, 2009.

Tian, H., Yang, J., Lu, C., Xu, R., Canadell, J. G., Jackson, R., Arneth, A., Chang, J., Chen, G., Ciais, P., Gerber, S., Ito, A., Huang, Y., Joos, F., Lienert, S., Messina, P., Olin, S., Pan, S., Peng, C., Saikawa, E., Thompson, R. L., Vuichard, N., Winiwarter, W., Zaehle, S., Zhang, B., Zhang, K., and Zhu, Q.: The global N2O Model Intercomparison Project (NMIP): Objectives, Simulation Protocol and Expected Products, Bulletin of the American Meteorological Society, pp. BAMS–D–17–0212.1, https://doi.org/10.1175/BAMS-D-17-0212.1, 2018.

Wang, Z., Hoffmann, T., Six, J., Kaplan, J. O., Govers, G., Doetterl, S., and Van Oost, K.: Human-induced erosion has offset one-third of carbon emissions from land cover change, Nature Climate Change, 7, 345–350, https://doi.org/10.1038/nclimate3263, 2017.

Wilkenskjeld, S., Kloster, S., Pongratz, J., Raddatz, T., and Reick, C. H.: Comparing the influence of net and gross anthropogenic land-use and land-cover changes on the carbon cycle in the MPI-ESM, Biogeosciences, 11, 4817–4828, https://doi.org/10.5194/bg-11-4817-2014, 2014.

Williamson, D. B., Blaker, A. T., and Sinha, B.: Tuning without over-tuning: Parametric uncertainty quantification for the NEMO ocean model, Geoscientific Model Development, 10, 1789–1816, https://doi.org/10.5194/gmd-10-1789-2017, 2017.

Zaehle, S., Sitch, S., Smith, B., and Hatterman, F.: Effects of parameter uncertainties on the modeling of terrestrial biosphere dynamics, Global Biogeochemical Cycles, 19, 1–16, https://doi.org/10.1029/2004GB002395, 2005.

Zaehle, S., Ciais, P., Friend, A. D., and Prieur, V.: Carbon benefits of anthropogenic reactive nitrogen offset by nitrous oxide emissions, Nature Geoscience, 4, 601–605, https://doi.org/10.1038/ngeo1207, 2011.