Imputation of missing land carbon sequestration data in the AR6 Scenario Database

Ruben Prütz[1, 2, 3, 4], Sabine Fuss[1, 2, 4], Joeri Rogelj[3, 5, 6]

[1]Geography Department, Humboldt-Universität zu Berlin, Berlin, Germany
[2]Mercator Research Institute on Global Commons and Climate Change (MCC), Berlin, Germany
[3]Grantham Institute for Climate Change and the Environment, Imperial College London, London, United Kingdom
[4]Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, Potsdam, Germany
[5]Centre for Environmental Policy, Imperial College London, London, United Kingdom
[6]Energy, Climate and Environment Program, International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria

Correspondence to: Ruben Prütz (Pruetz@mcc-berlin.net)

Abstract. The AR6 Scenario Database is a vital repository of climate change mitigation pathways used in the latest IPCC assessment cycle. In its current version, many scenarios in the database lack information about the level of anthropogenic carbon dioxide removal via land sinks, as net negative CO_2 emissions and gross removals on land are not always separated and not consistently reported across models. This makes scenario analyses focusing on carbon dioxide removal challenging. We test and compare the performance of different regression models to impute missing data on land carbon sequestration for the global level and for several sub-global macro regions from available data on net CO_2 emissions in agriculture, forestry, and other land use. We find that a k-nearest neighbors regression performs best among the tested regression models and use it to impute and provide two publicly available imputation datasets [https://doi.org/10.5281/zenodo.13373539] (Prütz et al., 2024) on carbon dioxide removal via land sinks for incomplete global scenarios (n=404) and incomplete regional R10 scenario variants (n=2358) of the AR6 Scenario Database. We discuss the limitations of our approach, the use of our datasets for secondary assessments of AR6 scenario ensembles, and how this approach compares to other recent AR6 data reanalyses.

1 Introduction

Climate change mitigation pathways, created with integrated assessment models (IAMs), have come to take up a critical role in the assessment work of Working Group III of the Intergovernmental Panel on Climate Change (IPCC) (Guivarch et al., 2022b; Riahi et al., 2022). The AR6 Scenario Database hosted by the International Institute for Applied Systems Analysis (IIASA) contains climate change mitigation pathways compiled for and considered in the Working Group III Contribution to the IPCC Sixth Assessment Report (Byers et al., 2022; Kikstra et al., 2022).

In these pathways, carbon dioxide removal (CDR) from the atmosphere is primarily represented by bioenergy with carbon capture and storage (BECCS) and by carbon sequestration in land sinks – primarily via afforestation and reforestation (Riahi et al., 2022). Among the global scenarios in the AR6 Scenario Database that passed the vetting process (n=1202) (see Guivarch et al. (2022b) for details about the AR6 scenario vetting process), 419 pathways miss the variable for carbon sequestration on land (‘Carbon Sequestration|Land Use’), which complicates secondary analyses that investigate CDR implications across scenarios and models. A range of different secondary scenario ensemble evaluations based on data from the AR6 Scenario Database have been published in recent years, e.g., assessing the arising gap in CDR deployment (Lamb et al., 2024), determining the level and composition of residual emissions (Lamb, 2024), analysing the removal per land unit (Zhao et al., 2024), evaluating the attainability of mitigation scenarios (Warszawski et al., 2021), classifying emission pathways reflecting the climate objectives of the Paris Agreement (Schleussner et al., 2022), or exploring scenario characteristics driving CDR deployment (Prütz et al., 2023). All these analyses rely on proxy data or interim solutions to address the limited data availability of land carbon sequestration in the AR6 Scenario Database.

Two such interim solutions to account for this data gap are documented in the literature, namely, the use of net negative CO_2 emissions in agriculture, forestry, and other land use (AFOLU) as a lower bound proxy variable for CDR via land sinks (Prütz et al., 2023; Schleussner et al., 2022; Warszawski et al., 2021), and a criteria-based scenario filtering and exclusion approach to ensure a consistent selection of scenarios with similar reporting of CDR via land sinks (Prütz et al., 2023). Both approaches have limitations in depicting CDR via land sinks adequately and comprehensively (Ganti et al., 2024). A more recent approach is based on a reanalysis of land CO_2 fluxes using the reduced-complexity compact earth system model OSCAR v3.2 (Gidden et al., 2023). While the AR6 reanalysis dataset by Gidden et al. manages to resolve several of the data issues linked to CDR via land sinks – specifically, aligning the removal baseline and improving the consistency across scenarios – it still combines gross and net CO_2 fluxes on land in their land sink CDR variable, resulting in both positive and negative CDR values, which conflicts with the concept and clean definition of anthropogenic CDR from the atmosphere (Matthews et al., 2021). In the AR6 Scenario Database, CDR is conventionally reported in positive numbers. Also, while being very comprehensive, the reanalysed dataset by Gidden et al. is limited to a subset (n=914) of all global and vetted scenarios (n=1202) of the AR6 Scenario Database, while also providing reanalysed scenario data for five sub-global macro regions (R5 level). Figure 1 compares the available land sink CDR data of the AR6 Scenario Database to the reanalysed variable by Gidden et al. and the net-negative AFOLU CO_2 proxy, showing the discrepancy of the net-negative AFOLU CO_2 proxy and the negative values for land sink CDR of the reanalysis.

Here, we test and compare the performance of several different regression models to impute missing data on land carbon sequestration (Land CDR) based on available data on net CO_2 emissions in AFOLU for both global scenarios and the R10 regions in the AR6 Scenario Database. We use the best performing regression model to impute missing data for 404 global scenarios and 2358 sub-global scenario variants across the R10 regions and provide two imputation datasets, which are made publicly available. Lastly, we discuss our approach's use cases and limitations and detail how our approach compares to the two above-mentioned interim solutions and the recent reanalysis of the AR6 Land CDR data. In the following, we refer to CDR via land sinks or carbon sequestration on land as Land CDR. Table 1 gives an overview and description of key variables in this analysis.

Table 1. Overview of the analysis variables

Variable	Description

<table>
<tr><td>'Carbon Sequestration|Land Use'</td><td>This variable from the AR6 Scenario Database is defined as the “total carbon dioxide sequestered through land sinks (e.g., afforestation, soil carbon enhancement, biochar)”. This is the target variable that we impute for incomplete scenarios. In this analysis, we refer to this variable as AR6 Land CDR.</td></tr>
<tr><td>'AR6 Reanalysis|OSCARv3.2|Carbon Removal|Land|Direct'</td><td>This variable from the reanalysis by Gidden et al. is intended to depict CDR through land sinks, similar to the AR6 Land CDR. However, the baseline CO_2 flux substantially differs compared to the AR6 Land CDR, as the data was aligned to national greenhouse gas inventories. This variable contains both positive and negative values, which suggests that it is showing net instead of gross removal. In this analysis, we refer to this variable as Gidden et al. Land CDR (direct).</td></tr>
<tr><td>'Emissions|CO2|AFOLU'</td><td>This variable from the AR6 Scenario Database is defined as the net “CO_2 emissions from agriculture, forestry and other land use (IPCC category 3)”. This is the predictor variable that we use to predict the target variable. In this analysis, we refer to this variable as net AFOLU CO_2 emissions.</td></tr>
<tr><td>| 'Emissions|CO2|AFOLU' < 0 |</td><td>This variable shows the net CO_2 removal from agriculture, forestry and other land use, based on the negative values in the variable net AFOLU CO_2 emissions. This variable has been used in several studies as lower bound proxy for AR6 Land CDR. We refer to this variable as net negative AFOLU CO_2.</td></tr>
<tr><td>'Imputed|Carbon Sequestration|Land Use'</td><td>This is one of two variables in the imputation datasets provided in this analysis. This variable contains the predicted values from our data imputation without further adjustment.</td></tr>
<tr><td>'Imputed & Proxy|Carbon Sequestration|Land Use'</td><td>This is one of two variables in the imputation datasets provided in this analysis. This variable contains the predicted values from our data imputation. For scenarios in which the predicted Land CDR is lower than the net negative AFOLU CO_2, we replaced all predicted removals with the values from the net negative AFOLU CO_2 and indicated this adjustment.</td></tr>
</table>

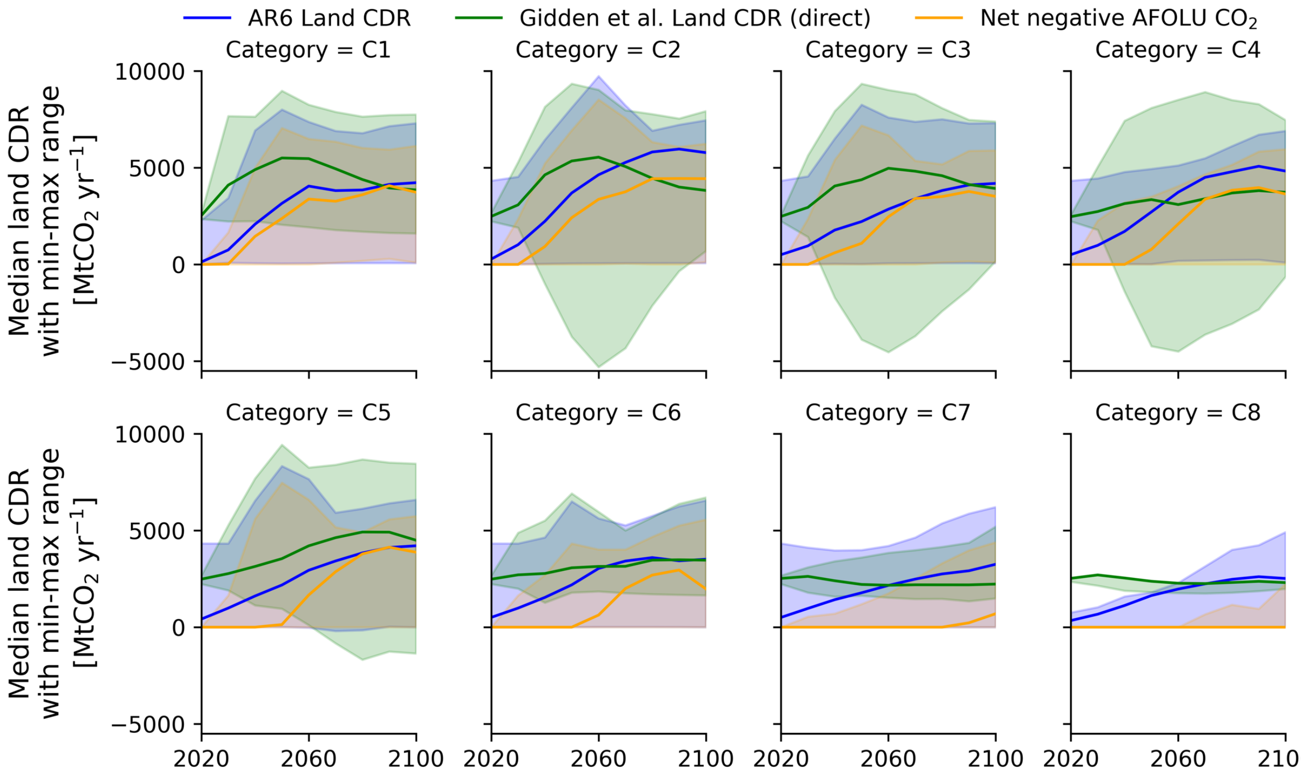

Figure 1. Comparison of available AR6 Land CDR data (‘Carbon Sequestration|Land Use’) with the Land CDR reanalysis by Gidden et al. (‘AR6 Reanalysis|OSCARv3.2|Carbon Removal|Land|Direct’) and the AR6 net negative AFOLU CO_2 emissions (based on negative values in ‘Emissions|CO2|AFOLU’) as a lower bound proxy for AR6 Land CDR across AR6 scenario categories. Only scenarios available for all three variables were considered in the figure (scenarios n=725). The Land CDR scenarios in the reanalysis by Gidden et al. are aligned with national greenhouse gas inventories, shown by the difference in baseline in 2020 compared to the other two variables. The solid lines show the median across scenarios while the shaded area shows the min-max range. Note: We follow the convention of the AR6 Scenario Database, to report CDR in positive numbers, whereas the Land CDR variable in the reanalysis by Gidden et al. shows both positive and negative CDR numbers. An overview of AR6 scenario categories (C1-8) is provided in Table 2.

2 Methods

Overview. In our analysis, we used different regression models to predict missing data on AR6 Land CDR (target variable: ‘Carbon Sequestration|Land Use’) for 404 incomplete global scenarios and for 2358 incomplete sub-global scenario variants across R10 regions based on available scenario data on AFOLU CO_2 emissions (predictor variable: ‘Emissions|CO2|AFOLU’). AFOLU CO_2 emissions were chosen as predictor variable due to good data availability in the AR6 Scenario Database and because this variable is conceptually most closely related to AR6 Land CDR among the variables in the AR6 Scenario Database – the variable for AFOLU CO_2 emissions represents the net CO_2 fluxes corresponding to the gross variable for Land CDR, as defined in Table 1. The AR6 R10 region classification comprises 10 macro regions plus one additional region for “rest of the world” (Figure 4b), resulting in a total of 11 macro regions, which we considered in our analysis. While the AR6 R10 classification allows for a comparison of regions across models and scenarios, not all regions are available for all models and scenarios, e.g., only a small subset of models has the category "rest of the world" (R10ROWO). For our analysis, we used the exact R10 regional classification as assigned in the AR6 Scenario Database without excluding or adjusting regions for individual scenarios or models.

As an initial step, we selected all vetted scenarios from the AR6 Scenario Database for which both the predictor and the target variable are available at the global level (n=783) and across the R10 regions (n=6162). Among the vetted global scenarios (n=1202) in the AR6 Scenario Database, 15 scenarios from the model REMIND 1.6 do not report AFOLU CO_2 emissions, which is why we could not include these scenarios in our imputation. Among the vetted regional scenario variants (n=8531) across the R10 regions in the AR6 Scenario Database, 11 regional variants of scenario EN_INDCi2100 from the model GEM-E3 V2021 do not report AFOLU CO_2 emissions, which is why we could not include these scenario variants in our imputation. Figure 2 provides a simplified conceptual overview of the scenario selection, exclusion, and imputation workflow.

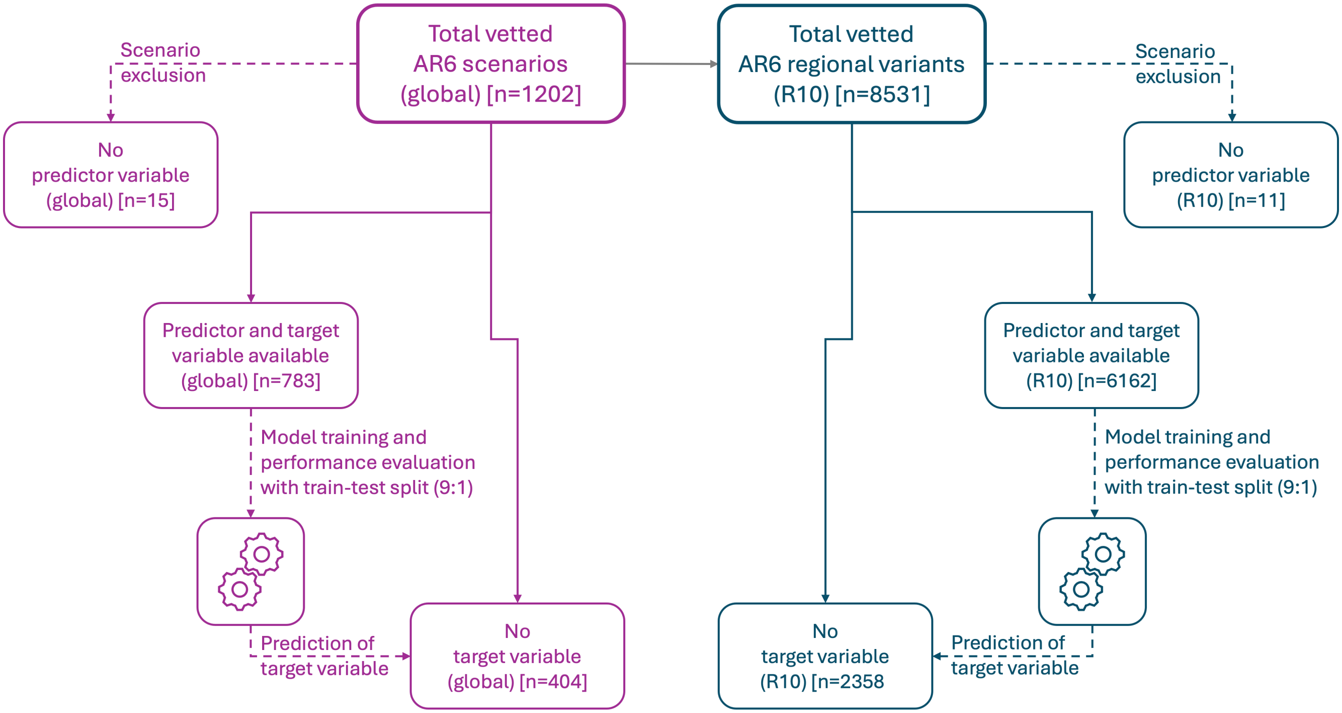
Total vetted AR6 scenarios (global) [n=1202]
Total vetted AR6 regional variants (R10) [n=8531]
Scenario exclusion
No predictor variable (global) [n=15]
Scenario exclusion
No predictor variable (R10) [n=11]
Predictor and target variable available (global) [n=783]
Predictor and target variable available (R10) [n=6162]
Model training and performance evaluation with train-test split (9:1)
Model training and performance evaluation with train-test split (9:1)
Prediction of target variable
No target variable (global) [n=404]
No target variable (R10) [n=2358
Prediction of target variable

Figure 2. Conceptual overview of the scenario selection, exclusion, and imputation workflow for the global scenarios and for regional scenario variants at AR6 R10 level. The numbers in brackets indicate the respective number of scenarios. Dashed lines indicate a process, while solid lines depict the origin of a scenario subset. Details about the model selection, training, and performance evaluation are provided in the Methods section.

We split both the global and the regional scenario datasets into training and testing sets (9:1) for our regression analysis to have a large dataset for training the models while still having a sufficiently large testing dataset to evaluate the prediction performance and to validate the models. The training set was used to fit the predictor variable to the target variable to train the regression models, and the testing set was then used to evaluate the prediction performance of the trained regression models. The regression models were separately trained on the global scenario data and the regional scenario variants as the scale of AR6 Land CDR deployment differs substantially between the global and the regional level. Regional scenario variants for model training were treated as one large training set rather than splitting the data by R10 region before training. For both the global scenarios and regional scenario variants, we did not distinguish between AR6 scenario categories during the model training process to keep the number of training data points as large as possible to optimize the models' performances. The AR6 scenario category (C1-8) classification is based on the scenarios' global warming levels from low warming of 1.5 °C with no or limited temporary temperature overshoot (C1) to high warming of more than 4 °C within this century (C8) (Guivarch et al., 2022a). An overview of the AR6 scenario categories is provided in Table 2.

Table 2. Overview of scenario categories as in Guivarch et al. (2022a)

Category	Description
C1	Scenarios limiting warming to 1.5 °C in 2100 (>50% probability) with no or limited overshoot (≤67% exceedance probability of 1.5 °C)
C2	Scenarios returning to warming of 1.5 °C in 2100 (>50% probability) after a high overshoot (>67% exceedance probability of 1.5 °C)
C3	Scenarios limiting warming to 2 °C throughout this century (>67% probability)
C4	Scenarios limiting warming to 2 °C throughout this century (>50% probability)
C5	Scenarios limiting warming to 2.5 °C throughout this century (>50% probability)
C6	Scenarios limiting warming to 3 °C throughout this century (>50% probability)
C7	Scenarios limiting warming to 4 °C throughout this century (>50% probability)
C8	Scenarios exceeding warming of 4 °C within this century (≥50% probability)

Regression models. We considered and compared four commonly used regression models in our analysis: gradient boosting, decision tree, random forest, and a k-nearest neighbors regression model. In the initial stage, a more extensive set of commonly used regression models, including linear regression and multilayer perceptron regression, were tested, among which the four above-mentioned models were selected for further hyperparameter tuning due to their superior performance compared to other regression models in the initial set, based on the performance evaluation metrics described below.

For all models, we use the machine learning scikit-learn library for Python (Pedregosa et al., 2011). In the following, the four considered regression models are briefly described, while more detail is provided in the referenced seminal works and the scikit-learn documentation of the respective models including the mathematical representations of the underlying algorithms.

A decision tree model is a supervised learning method to predict a target variable based on decision rules derived from a predictor variable. The model produces piecewise approximations of the target variable through a series of binary data splits based on values of the predictor variable. For continuous predictor variables, as in our case net AFOLU CO_2 emissions, the decision tree model iteratively selects thresholds for the predictor variable to split the data into a group above and below the respective threshold and then averages the target variable values per group to come up with a prediction. At each node of the tree, multiple potential splits are evaluated, and the imposed threshold that minimizes the prediction error is selected to increase the accuracy of the prediction of the target variable. This process continues recursively, splitting the data at each node until the tree reaches the leaf node, where no further splits are possible or no further reduction in the prediction error is achieved. At the leaf node, the tree makes a final prediction of the target variable – in our case, the expected AR6 Land CDR (Breiman et al., 1984). A gradient boosting regression model is an ensemble method which sequentially combines multiple simple models, called weak learners (typically decision trees as described above), which correct the previous models' predictions to reduce the error and improve the final model (Friedman, 2001). A random forest model is also an ensemble method which combines multiple decision trees as described above, but unlike gradient boosting, the trees in a random forest model run in parallel instead of sequentially. Each decision tree works independently, and their individual predictions are averaged to produce the final prediction (Breiman, 2001). The k-nearest neighbors model is not based on decision trees. Instead, it uses the proximity (similarity) of a scenario to a number (k) of neighbors (similar scenarios) to make predictions. For a given scenario, the model identifies the k nearest data points of the predictor variable in the feature space and then averages the target variable values of these neighbors to come up with a prediction (Goldberger et al., 2004).

Performance optimization and evaluation. From the machine learning library scikit-learn, we used grid search for our regression model hyper-parameter optimization, and bootstrapping to estimate the variability in prediction performance for different subsamples of our training and testing data (Pedregosa et al., 2011). Grid search is an algorithm commonly used for regression problems, which allows to efficiently run regression models in different setups using all possible hyper-parameter combinations to eventually select the best performing model setup. The selection of hyper-parameter options for the model optimization was driven by the observed model performance and computation time. We used bootstrapping to explore how the prediction performance of our optimized regression models varied based on different resamples (n=1000) of the training and testing data, allowing us to better evaluate the robustness of the perceived performance of the tested models.

The model performance was determined based on four widely applied evaluation metrics, namely R-squared, mean absolute error, median absolute error, and maximum absolute error. These four evaluation metrics are briefly described in the following. R-squared was used to explore how well the tested regression model captured the relationship between the predicted variable and the actual variable in the validation dataset, indicating the goodness of their fit. R-squared can range from zero to one with higher values indicating better fits. The other three evaluation metrics instead indicate absolute error, meaning the absolute difference between the predicted and the actual variable in the validation dataset throughout 2020-2100 – lower error values indicate more accurate predictions of the regression models. As the absolute error differs across variable pairs of the global scenarios (n=79) and the regional scenario variants (n=617) in the testing dataset, we reported the mean, median and maximum error across the considered scenarios. Mean and median error are useful to estimate the prediction models' overall performance whereas the maximum error is used to indicate the extreme in lower-end performance, based on the most inaccurately predicted scenario. Ultimately, the best performing model (k-nearest neighbors) was used to impute the missing AR6 Land CDR data for incomplete global scenarios (n=404) and incomplete sub-global scenario variants (n=2358) across the R10 regions in the AR6 Scenario Database. The performance of the four considered regression models and the selection of the k-nearest neighbors model is discussed further under Results.

Data postprocessing. For all imputed scenarios, the predicted target variable was compared to their predictor variable to identify cases where imputed CDR on land is smaller than the respective net negative AFOLU CO_2 emissions, as this conceptual error was partly also perceived in the AR6 Scenario Database. The two imputation datasets for the global scenarios and R10 regional variants contain two data sheets. The first data sheet contains unadjusted imputation outputs. In contrast, the second sheet accounts for the conceptual error described above by replacing conceptually inconsistent predictions with their respective net negative AFOLU CO_2 emissions for all years in affected scenarios to provide a lower bound proxy for AR6 Land CDR – implications are explained in the discussion section. The code to implement the analysis and the global and regional imputation datasets are publicly available at [https://doi.org/10.5281/zenodo.13373539].

3 Results

We show the performance of the four tested regression models along the four above-described evaluation metrics based on the testing set used for the regression model validation for the global scenarios (Figure 3a) and their regional scenario variants (Figure 4a). Overall, the k-nearest neighbors regression model performs best, as it resembles the actual variable most accurately, while keeping mean, median and absolute difference between the predicted variable and the actual variable comparatively low throughout 2020-2100 for both the global scenarios and the regional scenario variants. It also shows relatively low variance in performance across the bootstrapping results (Figure 3a and 4a). While the gradient boosting regression performs comparatively well for the prediction of the global scenarios and slightly better concerning the maximum absolute error (Figure 3a), the k-nearest neighbors regression outperforms the gradient boosting regression regarding mean and median absolute error for the prediction of the incomplete regional scenario variants (Figure 4a). While the overall performance of these two regression models is similar, the k-nearest neighbors model was chosen to produce the two imputation datasets of this study, as the gradient boosting model partly predicted slightly negative values in the target variable, which is conceptually inconsistent with a clean definition of Land CDR, which should have a uniform removal sign. The other two regression models perform less well than the k-nearest neighbors and gradient boosting regressions. Overall, all models show a slight performance drop for R-squared around 2020-2060, with more stable or increased performance thereafter – we have found no convincing explanation for this slight temporal variation in performance.

In absolute terms, the mean, median, and maximum errors are larger for the evaluated global scenarios than for their regional scenario variants – this is expected due to the substantially higher levels of Land CDR deployment on the global level compared to the R10 regions. On the global level, mean error of the k-nearest neighbors model is consistently below 200 $MtCO_2\ yr^{-1}$ and for the median error consistently below 40 $MtCO_2\ yr^{-1}$ – on the R10 region level we see mean error consistently below 15 $MtCO_2\ yr^{-1}$ and median error close to zero. For both the global and regional level, the mean and median absolute difference between the predicted variable and the actual variable is judged to be reasonably low, based on the k-nearest neighbors regression, and the absolute difference between the actual and predicted variable is substantially smaller than between the actual variable and the net negative AFOLU CO_2 emissions as a lower bound proxy for comparison (see also Figure 3b and 4b). As a point of reference, the median AR6 Land CDR deployment across available scenarios of the scenario categories C1-8 in the AR6 Scenario Database is 1253 $MtCO_2\,yr^{-1}$ for 2020-2060 and 3570 $MtCO_2\,yr^{-1}$ for 2060-2100 – across the R10 regions median deployment is 39 $MtCO_2\,yr^{-1}$ and 179 $MtCO_2\,yr^{-1}$ respectively. This means that the median error accounts to around 1% of the median AR6 Land CDR deployment throughout the timeseries for the original global scenarios and even lower for the original regional scenario variants in the regression validation dataset. However, while the regression model seems to perform well overall based on the regression evaluation dataset, the observed maximum error suggests substantially worse performance in extreme cases, when looking at the scenario with the highest absolute error (Figure 3a and 4a).

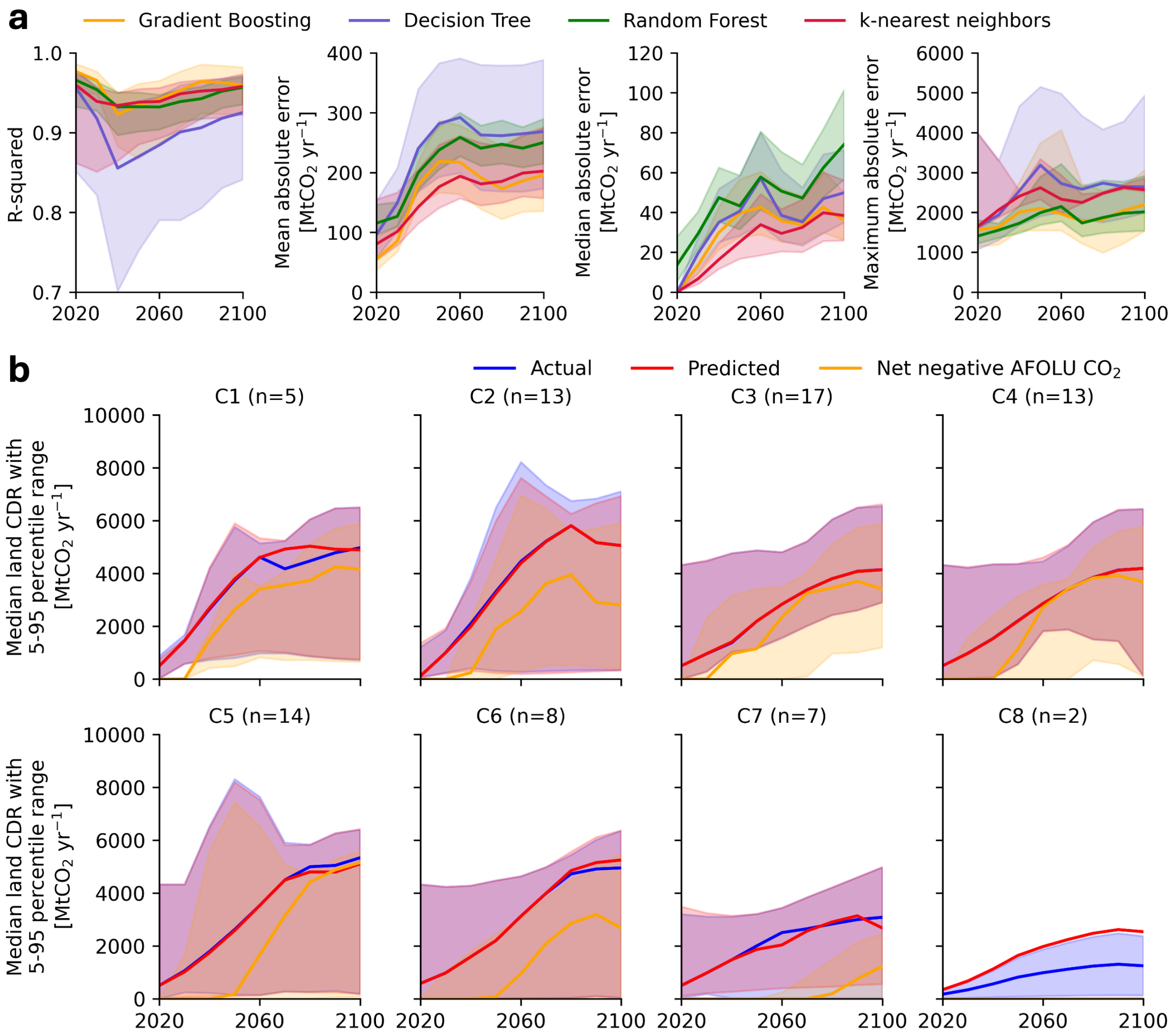

Figure 3. Prediction performance for the global scenario data. Panel (a) shows the performance of tested regression models to predict missing AR6 land removal data based on the used regression validation dataset (scenarios n=79). Performance across the four evaluation metrics is shown as the median (solid line) and 5-95 percentile range (shaded area) of the bootstrapping results (n=1000) for each of the four tested regression models. The performance results refer to the comparison between the predicted variable compared to the actual variable in the regression validation dataset. Panel (b) shows the actual ('Carbon Sequestration|Land Use') versus predicted Land CDR and the AR6 net negative AFOLU CO_2 emissions (based on negative values in 'Emissions|CO2|AFOLU') as a lower bound proxy for AR6 Land CDR across AR6 scenario categories in the regression validation dataset (scenarios n=79). The predicted data in the figure is based on the k-nearest neighbors regression. The solid lines show the median across scenarios while the shaded area shows the 5-95 percentile range. Note: We follow the convention of the AR6 Scenario Database, to report CDR in positive numbers. An overview of AR6 scenario categories (C1-8) is provided in Table 2.

Figure 3b and 4b show Land CDR across the global scenarios and their regional scenario variants in the regression validation dataset, considering the actual variable for AR6 Land CDR, the predicted Land CDR using the k-nearest neighbors regression, and the net negative AFOLU CO_2 emissions as a lower bound proxy for comparison. Considering the scenarios in the global and regional regression validation datasets, the predicted variable appears to be a better proxy variable for missing AR6 Land CDR than the net negative AFOLU CO_2 emissions proxy, as the predicted variable better resembles the shape of the actual variable and shows less absolute error throughout 2020-2100. While the predicted variable resembles the actual variable well across all eight AR6 scenario categories, Figure 3b suggests some variance in performance across these categories – for C8 scenarios the drop in resemblance of the actual variable is most visible. This is at least partly due to the small number of underlying scenarios of this category in the regression validation dataset at the global level (n=2). The prediction performance across the different R10 regions is comparatively consistent, as shown in Figure 4b. In some cases, the actual variable and the predicted Land CDR are smaller than the net negative AFOLU CO_2 emissions proxy, e.g., visible in some instances for the R10 region R10EUROPE in Figure 4b. This highlights a conceptual error in the underlying data, which is further discussed in the subsequent section.

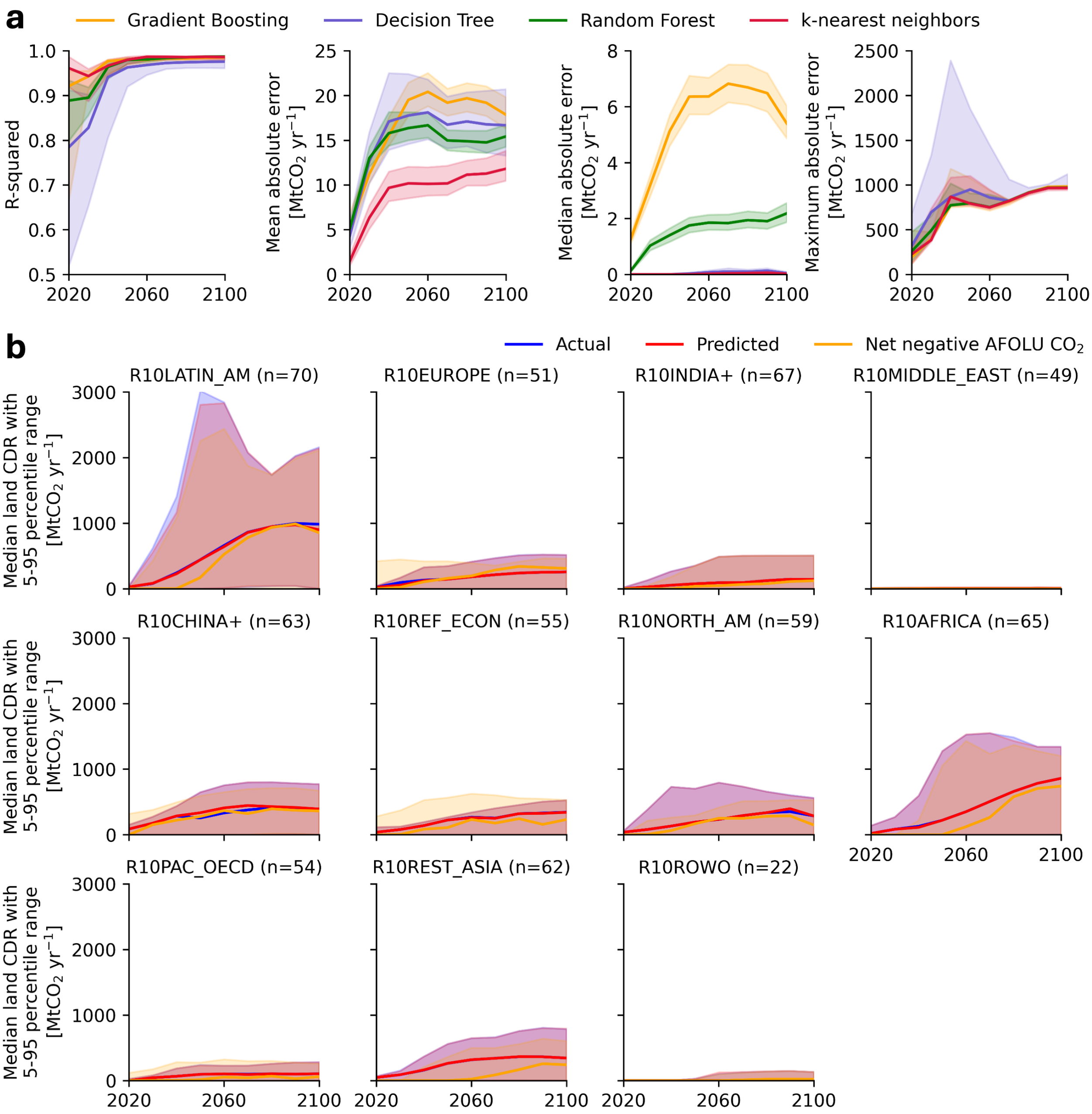

Figure 4. Prediction performance for the R10 regions scenario data. Panel (a) shows the performance of tested regression models to predict missing AR6 land removal data based on the used regression validation dataset (regional scenario variants n=617). Performance across the four evaluation metrics is shown as the median (solid line) and 5-95 percentile range (shaded area) of the bootstrapping results (n=1000) for each of the four tested regression models. The performance results refer to the comparison between the predicted variable compared to the actual variable in the regression validation dataset. Panel (b) shows the actual ('Carbon Sequestration|Land Use') versus predicted Land CDR and the AR6 net negative AFOLU CO$_2$ emissions (based on negative values in 'Emissions|CO2|AFOLU') as a lower bound proxy for AR6 Land CDR across AR6 R10 regions in the regression validation dataset (regional scenario variants n=617). The predicted data in the figure is based on the k-nearest neighbors regression. The solid lines show the median across scenarios while the shaded area shows the 5-95 percentile range. Note: We follow the convention of the AR6 Scenario Database, to report CDR in positive numbers.

4 Discussion and conclusion

In this study, we tested and compared four regression models to impute missing AR6 scenario data on Land CDR based on available data on net AFOLU CO_2 emissions. The tested k-nearest neighbors regression model performed best and was used to impute the missing AR6 Land CDR data for incomplete global scenarios (n=404) and incomplete sub-global scenario variants (n=2358) across the R10 regions. The global and regional imputation datasets are publicly available at: [https://doi.org/10.5281/zenodo.13373539].

While we effectively resemble and impute AR6 Land CDR data for incomplete scenarios, our imputed datasets do not resolve underlying inconsistencies in the reporting of AR6 Land CDR in the AR6 Scenario Database. The original data in the AR6 Scenario Database for the variable 'Carbon Sequestration|Land Use' is based on different reporting methodologies across IAMs, and land CO_2 fluxes are not always consistently and explicitly split into net negative CO_2 emissions and gross removals (Ganti et al., 2024; Prütz et al., 2023). Different baselines for today's land removal are also perceived across scenarios, as shown in Figure 1. For several scenarios in the AR6 Scenario Database, net negative AFOLU CO_2 emissions are larger than the reported AR6 Land CDR, which indicates conceptual errors as Land CDR is a gross variable, which can only be larger or equal to net negative AFOLU CO_2 emissions (Byers et al., 2022; Prütz et al., 2023). The issues of inconsistent removal baselines and net negative CO_2 emissions being larger than gross removal are partly also perceived in our imputed datasets, as we use data from the AR6 Scenario Database to train our model.

To address the latter problem, we provide an unadjusted imputation dataset as well as an adjusted imputation dataset for which we replaced conceptually inconsistent predictions (net negative CO_2 emissions being larger than gross removal) with their respective net negative AFOLU CO_2 emissions for all years in the affected scenarios to provide a lower bound proxy for AR6 Land CDR in the global and regional imputation dataset. We adjusted 106 global and 1594 regional scenario variants and indicated in the adjusted imputation dataset for which scenarios the adjustment was made.

We emphasize that our global and regional imputed datasets are imperfect and that the persisting issue of net negative CO_2 emissions and gross removals on land not always being separated and consistently reported across models, must be considered when using our data imputation. The here applied approach to infill missing data is purely based on statistical relationships and is not intended to replace further improvements in comprehensively reporting Land CDR in the next generation of mitigation scenarios produced with process-based models. Nevertheless, Figure 3b and 4b show that our imputed Land CDR variable is a markedly better proxy than the use of net-negative CO_2 emissions, which was partly used in previous studies (Prütz et al., 2023; Schleussner et al., 2022; Warszawski et al., 2021) – both in terms of resembling the removal curve and reducing absolute error. Our imputation is also a better alternative to omitting a large part of the scenario space that does not report AR6 Land CDR.

Concerning use cases, we believe our global and regional imputed datasets on AR6 Land CDR are most useful for analyses that aim to use the largest possible set of both original and imputed global scenarios (n=783+404) or regional R10 scenario variants (n=6162+2358) and a uniform carbon removal sign. Such scenario ensemble assessments are relevant to better understand a range of different aspects concerning Land CDR in climate change mitigation pathways. Several specific use cases have been highlighted above, including an assessment of the arising gap in CDR deployment (Lamb et al., 2024), an analysis of residual emissions including the land sector (Lamb, 2024), estimations of land-per-removal (Zhao et al., 2024) or evaluations of the attainability of mitigation scenarios, which rely on Land CDR (Warszawski et al., 2021).

So far, such analyses rely on insufficient proxy data or interim solutions to address the limited data availability of land carbon sequestration in the AR6 Scenario Database and could benefit from the here provided more comprehensive dataset on Land CDR across scenarios. Based on the evaluation of mean, median, and maximum absolute error of the here used regression model, it is advisable to use our dataset for analyses that rely on a large ensemble of scenarios, e.g., all scenarios of a certain scenario category or even several categories. This is because the prediction results from the regression model are more reliable for scenario ensembles than for individual scenarios, which may show larger error, as shown by the maximum error in Figure 3a and 4a – arguably, it is generally advisable to aim to use scenario ensembles instead of individual scenarios to better capture uncertainties and diverse underlying assumptions, which may lead to more robust and credible analysis outcomes (Guivarch et al., 2022b). The reanalysis discussed above by Gidden et al. is perceived to be more suitable in terms of consistency and accuracy of today's removals and for direct comparisons of scenario data and national greenhouse gas inventories (NGHGI). While our imputation dataset contains Land CDR data for the scenario starting year 2020, today’s and also historical emissions and removals are better captured and more comprehensively discussed by the Global Carbon Project (Friedlingstein et al., 2023) – the merit of our imputation dataset lies in the future timesteps of scenarios. As the here used imputation approach is purely based on statistical relationships between the predictor and target variable, it can also be applied to data availability problems in other domains to infill missing data, given that sufficient data is available to train and evaluate the model and that the models’ performance is judged to be adequate. Ultimately, we hope this study can be a valuable and complementary addition to the existing approaches addressing the Land CDR data gap in the AR6 Scenario Database.

Acknowledgments

We thank the Integrated Assessment Modeling Consortium (IAMC) and the International Institute for Applied Systems Analysis (IIASA) for their valuable work on collecting and hosting quantitative integrated assessment scenarios from and for the research community. We also thank the scikit-learn team for providing the free software machine learning library, that made this analysis easily implementable. The authors acknowledge funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101003687 (PROVIDE), and No 951542 (GENIE). R.P. acknowledges funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101081521 (UPTAKE). J.R. acknowledges funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101003536 (ESM2025). S.F. acknowledges funding from the German Ministry for Education and Research under grant agreement No 01LS2101F (CDRSynTra).

Code and data availability

The analysis code and the global and regional imputed datasets are publicly available at: [https://doi.org/10.5281/zenodo.13373539] (Prütz et al., 2024).

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

Author contributions. R.P. led the study and conceptualization, with supervision by S.F. and J.R.. R.P. implemented the analysis and wrote the original draft. All authors reviewed and edited the paper.

References

Breiman, L.: Random Forests, Mach Learn, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J.: Classification And Regression Trees, Routledge, https://doi.org/10.1201/9781315139470, 1984.

Byers, E., Krey, V., Kriegler, E., Riahi, K., Schaeffer, R., Kikstra, J., Lamboll, R., Nicholls, Z., Sandstad, M., Smith, C., van der Wijst, K., Lecocq, F., Portugal-Pereira, J., Saheb, Y., Stromann, A., Winkler, H., Auer, C., Brutschin, E., Lepault, C., Müller-Casseres, E., Gidden, M., Huppmann, D., Kolp, P., Marangoni, G., Werning, M., Calvin, K., Guivarch, C., Hasegawa, T., Peters, G., Steinberger, J., Tavoni, M., van Vuuren, D., Al -Khourdajie, A., Forster, P., Lewis, J., Meinshausen, M., Rogelj, J., Samset, B., and Skeie, R.: AR6 Scenarios Database, https://doi.org/10.5281/zenodo.5886912, April 2022.

Friedlingstein, P., O'Sullivan, M., Jones, M. W., Andrew, R. M., Bakker, D. C. E., Hauck, J., Landschützer, P., Le Quéré, C., Luijkx, I. T., Peters, G. P., Peters, W., Pongratz, J., Schwingshackl, C., Sitch, S., Canadell, J. G., Ciais, P., Jackson, R. B., Alin, S. R., Anthoni, P., Barbero, L., Bates, N. R., Becker, M., Bellouin, N., Decharme, B., Bopp, L., Brasika, I. B. M., Cadule, P., Chamberlain, M. A., Chandra, N., Chau, T.-T.-T., Chevallier, F., Chini, L. P., Cronin, M., Dou, X., Enyo, K., Evans, W., Falk, S., Feely, R. A., Feng, L., Ford, D. J., Gasser, T., Ghattas, J., Gkritzalis, T., Grassi, G., Gregor, L., Gruber, N., Gürses, Ö., Harris, I., Hefner, M., Heinke, J., Houghton, R. A., Hurtt, G. C., Iida, Y., Ilyina, T., Jacobson, A. R., Jain, A., Jarn\'{\i}ková, T., Jersild, A., Jiang, F., Jin, Z., Joos, F., Kato, E., Keeling, R. F., Kennedy, D., Klein Goldewijk, K., Knauer, J., Korsbakken, J. I., Körtzinger, A., Lan, X., Lefèvre, N., Li, H., Liu, J., Liu, Z., Ma, L., Marland, G., Mayot, N., McGuire, P. C., McKinley, G. A., Meyer, G., Morgan, E. J., Munro, D. R., Nakaoka, S.-I., Niwa, Y., O'Brien, K. M., Olsen, A., Omar, A. M., Ono, T., Paulsen, M., Pierrot, D., Pocock, K., Poulter, B., Powis, C. M., Rehder, G., Resplandy, L., Robertson, E., Rödenbeck, C., Rosan, T. M., Schwinger, J., Séférian, R., et al.: Global Carbon Budget 2023, Earth Syst Sci Data, 15, 5301–5369, https://doi.org/10.5194/essd-15-5301-2023, 2023.

Friedman, J. H.: Greedy function approximation: A gradient boosting machine., The Annals of Statistics, 29, 1189–1232, https://doi.org/10.1214/aos/1013203451, 2001.

Ganti, G., Gasser, T., Bui, M., Geden, O., Lamb, W. F., Minx, J. C., Schleussner, C.-F., and Gidden, M. J.: Evaluating the near- and long-term role of carbon dioxide removal in meeting global climate objectives, Commun Earth Environ, 5, 377, https://doi.org/10.1038/s43247-024-01527-z, 2024.

Gidden, M. J., Gasser, T., Grassi, G., Forsell, N., Janssens, I., Lamb, W. F., Minx, J., Nicholls, Z., Steinhauser, J., and Riahi, K.: Aligning climate scenarios to emissions inventories shifts global benchmarks, Nature, 624, 102–108, https://doi.org/10.1038/s41586-023-06724-y, 2023.

Goldberger, J., Hinton, G. E., Roweis, S., and Salakhutdinov, R. R.: Neighbourhood Components Analysis, in: Advances in Neural Information Processing Systems, 2004.

Guivarch, C., Kriegler, E., Portugal-Pereira, J., Bosetti, V., Edmonds, J., Fischedick, M., Havlík, P., Jaramillo, P., Krey, V., Lecocq, F., Lucena, A., Meinshausen, M., Mirasgedis, S., O'Neill, B., Peters, G. P., Rogelj, J., Rose, S., Saheb, Y., Strbac, G., Strøm, A. H., and Zhou, N.: IPCC, 2022: Annex III: Scenarios and modelling methods, in: IPCC: Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Shukla, P. R., Skea, J., Slade, R., Khourdajie, A. Al, van Diemen, R., McCollum, D., Pathak, M., Some, S., Vyas, P., Fradera, R., Belkacemi, M., Hasija, A., Lisboa, G., Luz, S., and Malley, J., 2022a.

Guivarch, C., Le Gallic, T., Bauer, N., Fragkos, P., Huppmann, D., Jaxa-Rozen, M., Keppo, I., Kriegler, E., Krisztin, T., Marangoni, G., Pye, S., Riahi, K., Schaeffer, R., Tavoni, M., Trutnevyte, E., van Vuuren, D., and Wagner, F.: Using large ensembles of climate change mitigation scenarios for robust insights, Nat Clim Chang, 12, 428–435, https://doi.org/10.1038/s41558-022-01349-x, 2022b.

Kikstra, J. S., Nicholls, Z. R. J., Smith, C. J., Lewis, J., Lamboll, R. D., Byers, E., Sandstad, M., Meinshausen, M., Gidden, M. J., Rogelj, J., Kriegler, E., Peters, G. P., Fuglestvedt, J. S., Skeie, R. B., Samset, B. H., Wienpahl, L., van Vuuren, D. P., van der Wijst, K.-I., Al Khourdajie, A., Forster, P. M., Reisinger, A., Schaeffer, R., and Riahi, K.: The IPCC Sixth Assessment Report WGIII climate assessment of mitigation pathways: from emissions to global temperatures, Geosci Model Dev, 15, 9075–9109, https://doi.org/10.5194/gmd-15-9075-2022, 2022.

Lamb, W. F.: The size and composition of residual emissions in integrated assessment scenarios at net-zero CO2, Environmental Research Letters, 19, 044029, https://doi.org/10.1088/1748-9326/ad31db, 2024.

Lamb, W. F., Gasser, T., Roman-Cuesta, R. M., Grassi, G., Gidden, M. J., Powis, C. M., Geden, O., Nemet, G., Pratama, Y., Riahi, K., Smith, S. M., Steinhauser, J., Vaughan, N. E., Smith, H. B., and Minx, J. C.: The carbon dioxide removal gap, Nat Clim Chang, 14, 644–651, https://doi.org/10.1038/s41558-024-01984-6, 2024.

Matthews, J. B. R., Möller, V., Van Diemen, R., Fuglestvedt, J. S., Masson-Delmotte, V., Méndez, C., Semenov, S., and Reisinger, A.: Annex VII: Glossary., in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, 2215–2256, https://doi.org/10.1017/9781009157896.022, 2021.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 2011.

Prütz, R., Strefler, J., Rogelj, J., and Fuss, S.: Understanding the carbon dioxide removal range in 1.5 °C compatible and high overshoot pathways, Environ Res Commun, 5, 41005, https://doi.org/10.1088/2515-7620/accdba, 2023.

Prütz, R., Fuss, S., and Rogelj, J.: Imputation of missing IPCC AR6 data on land carbon sequestration, https://doi.org/10.5281/zenodo.13373539, February 2024.

Riahi, K., Schaeffer, R., Arango, J., Calvin, K., Guivarch, C., Hasegawa, T., Jiang, K., Kriegler, E., Matthews, R., Peters, G. P., Rao, A., Robertson, S., Sebbit, A. M., Steinberger, J., Tavoni, M., and van Vuuren, D. P.: Mitigation pathways compatible with long-term goals, in: IPCC: Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Shukla, P. R., Skea, J., Slade, R., Khourdajie, A. Al, Diemen, R. van, McCollum, D., Pathak, M., Some, S., Vyas, P., Fradera, R., Belkacemi, M., Hasija, A., Lisboa, G., Luz, S., and Malley, J., Cambridge University Press, Cambridge / New York, 2022.

Schleussner, C.-F., Ganti, G., Rogelj, J., and Gidden, M. J.: An emission pathway classification reflecting the Paris Agreement climate objectives, Commun Earth Environ, 3, 135, https://doi.org/10.1038/s43247-022-00467-w, 2022.

Strefler, J., Bauer, N., Humpenöder, F., Klein, D., Popp, A., and Kriegler, E.: Carbon dioxide removal technologies are not born equal, Environmental Research Letters, 16, 74021, https://doi.org/10.1088/1748-9326/ac0a11, 2021.

Warszawski, L., Kriegler, E., Lenton, T. M., Gaffney, O., Jacob, D., Klingenfeld, D., Koide, R., Costa, M. M., Messner, D., Nakicenovic, N., Schellnhuber, H. J., Schlosser, P., Takeuchi, K., Van Der Leeuw, S., Whiteman, G., and Rockström, J.: All options, not silver bullets, needed to limit global warming to 1.5 °C: a scenario appraisal, Environmental Research Letters, 16, 64037, https://doi.org/10.1088/1748-9326/abfeec, 2021.

Zhao, X., Mignone, B. K., Wise, M. A., and McJeon, H. C.: Trade-offs in land-based carbon removal measures under 1.5 °C and 2 °C futures, Nat Commun, 15, 2297, https://doi.org/10.1038/s41467-024-46575-3, 2024.