https://doi.org/10.5194/essd-2024-68
Preprint. Discussion started: 14 March 2024

Open Access Earth System Science Data Discussions

Imputation of missing IPCC AR6 data on land carbon sequestration

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

[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]Centre for Environmental Policy, Imperial College London, London, United Kingdom
[5]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, several scenarios in the database lack information about the level of gross carbon removal on land, as net and gross removals on land are not always separated and consistently reported across models. This makes scenario analyses focusing on carbon removals challenging. We test and compare the performance of different regression models to impute missing data on land carbon sequestration from available data on net CO_2 emissions in agriculture, forestry, and other land use. We find that a gradient boosting regression performs best among the tested regression models and provide a publicly available imputation dataset [https://doi.org/10.5281/zenodo.10696654] (Prütz et al., 2024) on carbon removal on land for 404 incomplete scenarios in the AR6 Scenario Database. We discuss the limitations of our approach, its use cases, and how this approach compares to other recent AR6 data re-analyses.

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) (Riahi et al., 2022; Guivarch 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 on land – primarily via afforestation and reforestation (Riahi et al., 2022). Among the scenarios in the AR6 Scenario Database that passed the vetting process (n=1202) (see Guivarch et al. (2022) 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. This gap requires the use of proxy data and interim solutions. Two such interim solutions to account for this data gap are documented in the literature, including the use of net negative CO_2 emissions in agriculture, forestry, and other land use (AFOLU) as a proxy variable for land-based CDR (Warszawski et al., 2021; Schleussner et al., 2022; Prütz et al., 2023) or

https://doi.org/10.5194/essd-2024-68
Preprint. Discussion started: 14 March 2024

scenario filtering and exclusion (Prütz et al., 2023). Both approaches have limitations in depicting gross carbon removals on land adequately and comprehensively (Ganti et al., 2024). A more recent approach is based on a re-analysis of land CO_2 fluxes using the earth system model OSCAR v3.2 (Gidden et al., 2023). While the AR6 re-analysis dataset by Gidden et al. manages to resolve several of the data issues linked to carbon removal on land, it still combines gross and net CO_2 emissions on land in their land-based CDR variable, resulting in both positive and negative CDR values, which conflicts with the concept and clean definition of gross CDR. Also, while being very comprehensive, the re-analyzed 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. Figure 1 compares the available land carbon sequestration data of the AR6 Scenario Database to the re-analyzed 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-based CDR of the re-analysis.

Here, we test and compare the performance of several different regression models to impute missing data on gross land carbon sequestration based on available data on net CO_2 emissions in AFOLU. We use the best performing regression model to impute missing data for 404 scenarios and to provide an imputation dataset, which is 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 re-analysis of the AR6 land carbon removal data.

https://doi.org/10.5194/essd-2024-68
Preprint. Discussion started: 14 March 2024

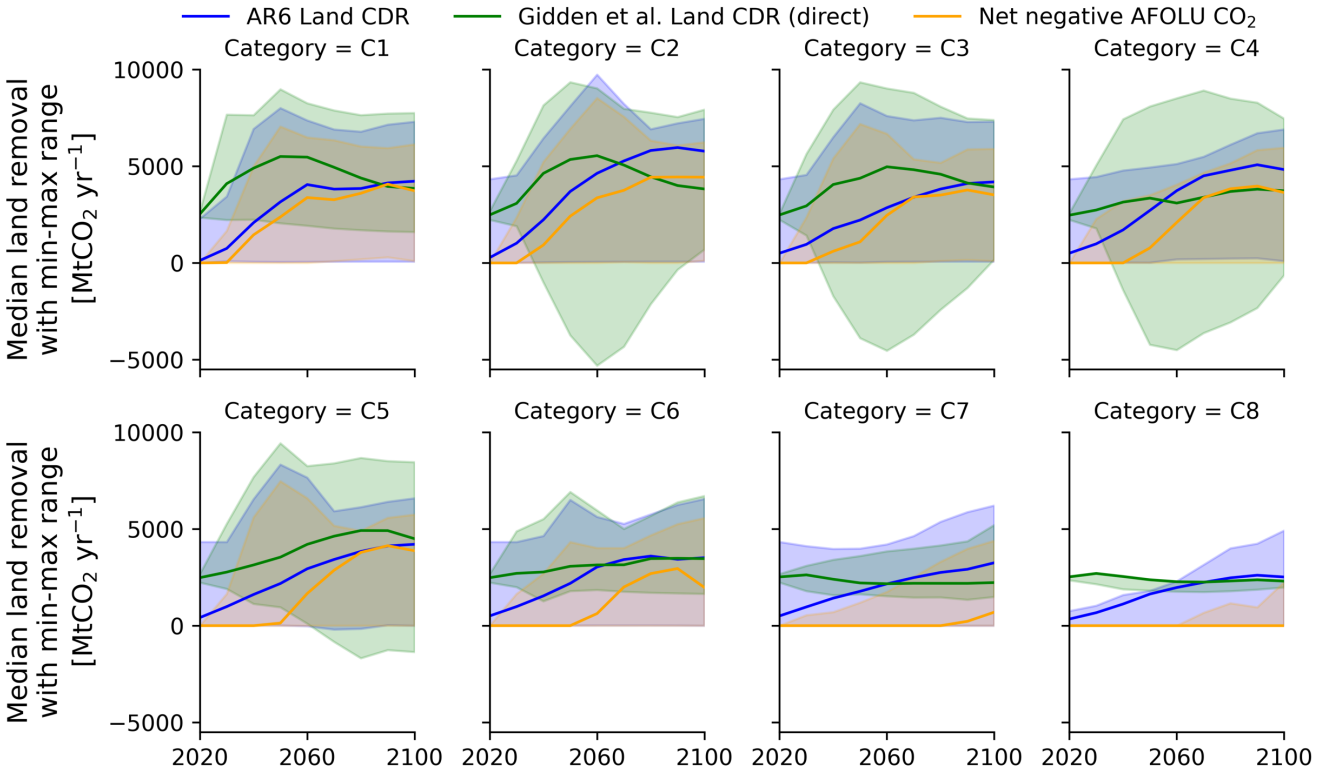

Figure 1: Comparison of available AR6 land-based CDR data ('Carbon Sequestration|Land Use') with the land carbon removal re-analysis by Gidden et al. ('AR6 Reanalysis|OSCARv3.2|Carbon Removal|Land|Direct') and the AR6 net negative AFOLU CO_2 emissions (based on negative values in 'Emissions|CO2|AFOLU') as a conservative proxy for land-based CDR across AR6 scenario categories. Only scenarios available for all three variables were considered in the figure (scenarios n=725).

2 Methods

In our analysis, we used different regression models to predict missing AR6 data on gross land carbon sequestration (dependent variable: 'Carbon Sequestration|Land Use') for 404 scenarios based on available scenario data on AFOLU CO_2 emissions (independent variable: 'Emissions|CO2|AFOLU'). As an initial step, we selected all vetted scenarios from the AR6 Scenario Database for which both the independent and the dependent variable are available (n=783). Among the vetted scenarios (n=1202) in the AR6 Scenario Database, 15 scenarios from REMIND 1.6 do not report AFOLU CO_2 emissions, which is why we could not include these scenarios in our imputation.

We then split this dataset into training and testing sets (9:1) for our regression analysis. The training set was used to fit the dependent variable to the independent variable to train the regression model, and the testing set was then used to evaluate the prediction performance of the trained regression model.

We considered and compared four regression models in our analysis: gradient boosting, decision tree, random forest, and a k-nearest neighbor regression model. In the initial stage, a more extensive set of different regression models, including linear

https://doi.org/10.5194/essd-2024-68
Preprint. Discussion started: 14 March 2024

regression and multilayer perceptron regression, were tested, among which the four models mentioned above were selected for further hyperparameter tuning due to their superior performance compared to other regression models in the initial set. We used grid search for our regression model hyper-parameter optimization, using the machine learning library scikit-learn (Pedregosa et al., 2011). The selection of hyper-parameter options for the model optimization was driven by the observed model performance and computation time.

The model performance was evaluated based on four metrics: R-squared, mean absolute error, median absolute error, and maximum absolute error. R-squared was used to explore how well the tested regression model captured the relationship between the dependent and independent variable, while mean, median and maximum absolute errors were used to evaluate the absolute difference between the dependent and the independent variable throughout 2020-2100. Ultimately, the best performing model (gradient boosting regression) was used to impute the missing gross land carbon sequestration data for 404 incomplete scenarios in the AR6 Scenario Database.

For two time steps of two imputed scenarios, negative values of up to -3 Mt CO_2 yr^{-1} were predicted, which is conceptually false and likely explained by the remaining model error. These values are only slightly below the conceptual minimum and, therefore, set to zero. For all imputed scenarios, the predicted dependent variable was compared to their independent 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 imputation dataset contains 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 false predictions with their respective net negative AFOLU CO_2 emissions as a conservative proxy for land-based CDR – implications are explained in the discussion section. The code to implement the analysis and the imputation dataset are publicly available at [https://doi.org/10.5281/zenodo.10696654].

3 Results

Figure 2 shows 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. Overall, the gradient boosting regression model performs best, as it describes the relationship between the dependent and independent variable most accurately, while keeping mean, median and absolute difference between the two variables comparatively low throughout 2020-2100. While the k-nearest neighbor regression performs comparatively well or slightly better concerning the mean and median absolute error, the gradient boosting regression outperforms the k-nearest neighbor regression regarding R-squared and the maximum absolute error. Also, the performance of the gradient boosting regression is most consistent when varying the ratio between the training and the testing set. The other two regression models perform less well than the gradient boosting and k-nearest neighbor regressions. Overall, all models show a slight performance drop around 2020-2060, with more stable or increased performance thereafter – we have found no convincing explanation for this slight temporal variation in performance.

https://doi.org/10.5194/essd-2024-68
Preprint. Discussion started: 14 March 2024

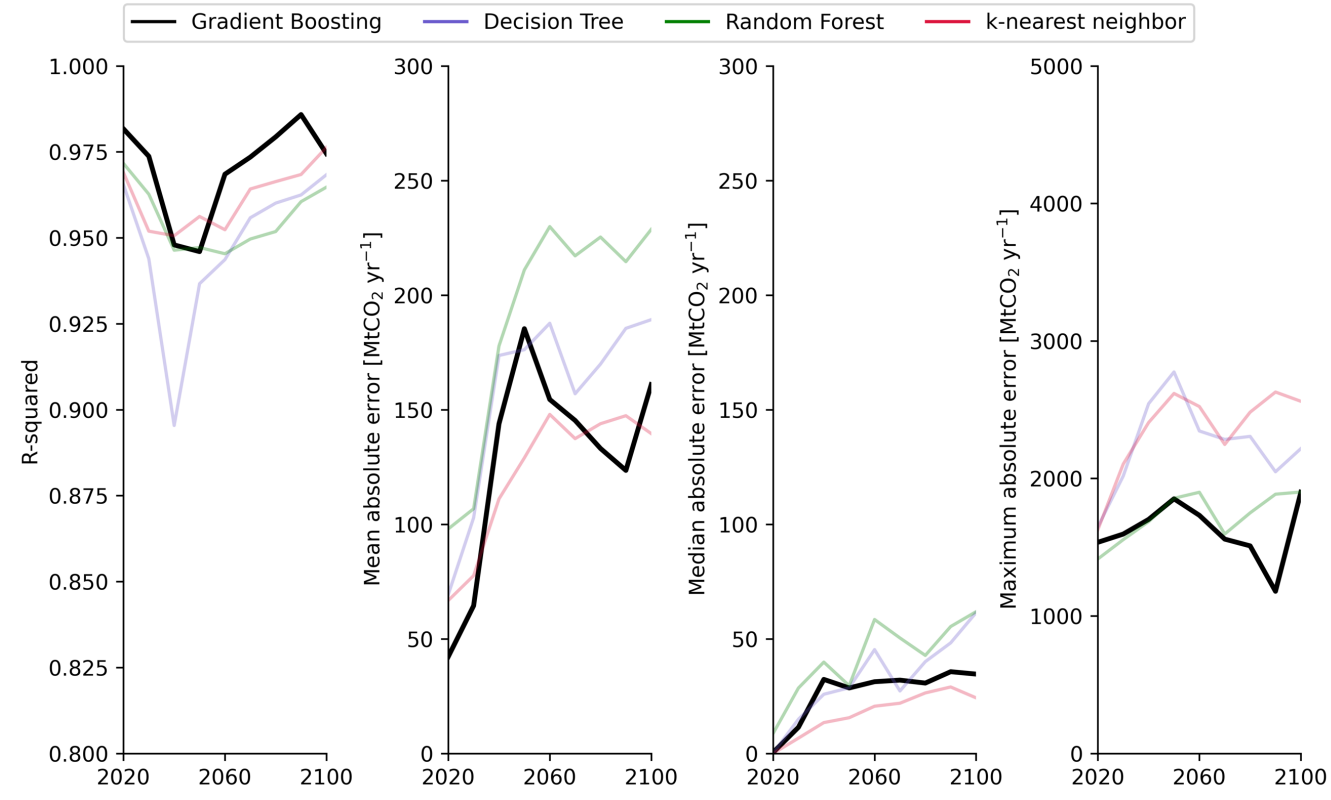

Figure 2: Performance of tested regression models to predict missing AR6 land removal data based on the used regression validation dataset (scenarios n=79).

Figure 3 shows the carbon removal on land across the scenarios in the regression validation dataset, considering the actual AR6 variable for carbon sequestration on land, the predicted carbon sequestration on land using the gradient boosting regression, and the net negative AFOLU CO_2 emissions as a conservative proxy for comparison. Considering the scenarios in this regression validation dataset, the predicted variable appears to be a better proxy variable for missing AR6 land carbon sequestration 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 3 suggests some variance in performance across these categories – for C8 scenarios, the drop in resemblance of the actual variable is most visible.

https://doi.org/10.5194/essd-2024-68
Preprint. Discussion started: 14 March 2024

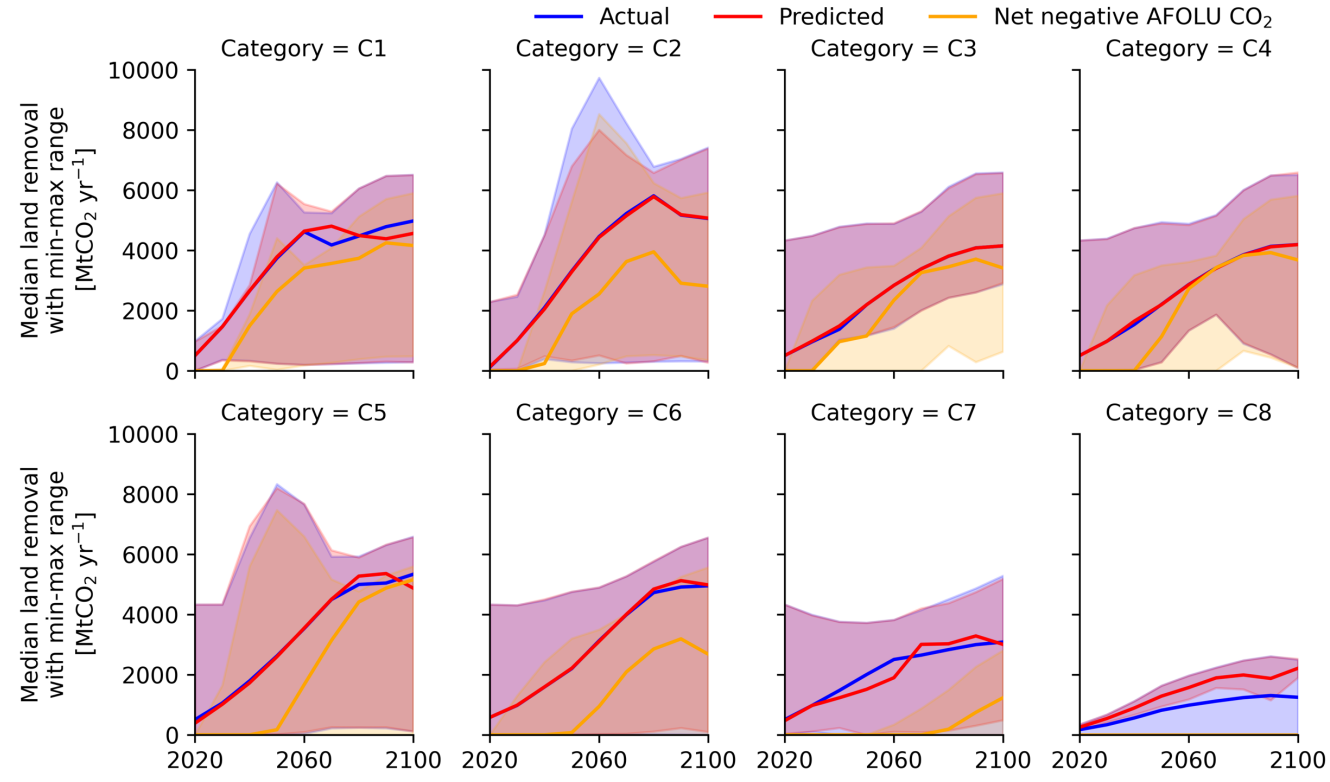

Figure 3: Actual ('Carbon Sequestration|Land Use') versus predicted land-based CDR and the AR6 net negative AFOLU CO_2 emissions (based on negative values in 'Emissions|CO2|AFOLU') as a conservative proxy for land-based CDR across AR6 scenario categories in the regression validation dataset (scenarios n=79). The predicted data in the figure is based on the gradient boosting regression.

4 Discussion and conclusion

In this study, we tested and compared four regression models to impute missing AR6 scenario data on land carbon sequestration based on available data on net AFOLU CO_2 emissions. The tested gradient boosting regression model performed best and was used to impute the missing land carbon sequestration data for 404 incomplete scenarios. The imputation dataset is publicly available at: [https://doi.org/10.5281/zenodo.10696654].

While we effectively resemble and impute land carbon sequestration data for 404 incomplete scenarios, our imputed dataset does not account for perceived land sequestration related data issues in the AR6 Scenario Database beyond data availability. The use of the variable 'Carbon Sequestration|Land Use' is further complicated as different reporting methodologies were used across IAMs, and land CO_2 fluxes are not always consistently and explicitly split into emissions and removals (Ganti et al., 2024). 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 carbon sequestration on land, which indicates conceptual errors as carbon sequestration on land is perceived to be a gross variable (Byers et al.,

https://doi.org/10.5194/essd-2024-68
Preprint. Discussion started: 14 March 2024

2022; Prütz et al., 2023). The issue of inconsistent removal baselines and net removal being larger than gross removal (which is the case for less than a quarter of all scenarios) is partly also perceived in our imputed dataset, 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 false predictions (net removal being larger than gross removal) with their respective net negative AFOLU CO_2 emissions as a conservative proxy for land-based CDR. We emphasize that our imputed dataset is imperfect and that the remaining data issues highlighted above must be considered when using our data imputation. Nevertheless, Figure 3 shows that our imputed land-based CDR variable is a markedly better proxy than the use of net-negative CO_2 emissions, which was partly used in previous studies (Schleussner et al., 2022; Warszawski et al., 2021; Prütz et al., 2023) – 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 carbon sequestration on land.

We believe our imputed dataset on land carbon removal is most useful for analyses that aim to use the largest possible set of both original and imputed scenarios (n=783+404) and a uniform carbon removal sign. However, the re-analysis mentioned above by Gidden et al. is perceived to be more useful in terms of consistency and accuracy of today's removals and for direct comparisons of scenario data and national greenhouse gas inventories (NGHGI). Ultimately, we hope this study can be a valuable and complementary addition to the existing approaches addressing the land carbon sequestration 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). RP acknowledges funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101081521 (UPTAKE). JR acknowledges funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101003536 (ESM2025).

Code and data availability

The analysis code and the imputed dataset are publicly available at: [https://doi.org/10.5281/zenodo.10696654] (Prütz et al., 2024).

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

https://doi.org/10.5194/essd-2024-68
Preprint. Discussion started: 14 March 2024

Author contributions. RP led the study and conceptualization, with supervision by SF and JR. RP implemented the analysis and wrote the original draft. All authors reviewed and edited the paper.

References

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.

Ganti, G., Gasser, T., Bui, M., Geden, O., Lamb, W., Minx, J., Schleussner, C.-F., and Gidden, M.: Carbon dioxide removal deployment consistent with global climate objectives, (Preprint), https://doi.org/https://doi.org/10.21203/rs.3.rs-3719978/v1, 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.

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, 2022.

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.

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, J. Mach. Learn. Res., 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.10696654, 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

https://doi.org/10.5194/essd-2024-68
Preprint. Discussion started: 14 March 2024

190 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.

195 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, Environ. Res. Lett., 16, 64037, https://doi.org/10.1088/1748-9326/abfeec, 2021.