Chapter 2
Technical Annex

Laura Diaz Anadon (University of Cambridge)
Jan C. Minx (Potsdam Institute for Climate Impact Research, PIK)
Marlene Bültemann (Potsdam Institute for Climate Impact Research, PIK)
Friedemann Gruner (Potsdam Institute for Climate Impact Research, PIK)
Finn Müller-Hansen (Potsdam Institute for Climate Impact Research, PIK)
Liz Pflugbeil (University of Cambridge)
D. Cale Reeves (University of Cambridge)
Tim Repke (Potsdam Institute for Climate Impact Research, PIK)

Chapter scientist: *Friedemann Gruner (Potsdam Institute for Climate Impact Research, PIK)*

Cite as: *Diaz Anadon, L., Minx, J. C., Bültemann, M., Gruner, F., Müller-Hansen, F., Pflugbeil, L., Reeves, D. C., Repke, T. Chapter 2: Research and development, in* ***The State of Carbon Dioxide Removal 3rd Edition 2026*** *(eds. Edwards, M. R. et al.). DOI: https://doi.org/10.17605/OSF.IO/9G5YW (2026)*

Technical Annex | Chapter 2

A2.1 Method for tracking early-stage R&D investments through third-party research grants

Research grants explicitly focused on CDR

This report assesses early-stage, third-party R&D investments in CDR using data on research projects granted by funding bodies as listed in the Dimensions database.[1,2] The database has global coverage and includes thousands of public and private funders, even though most of the funding is from public sources. The database is one of the largest grants databases, but the comprehensiveness of its coverage is difficult to assess. Extensive keyword searches were conducted for each of the CDR methods considered to download an initial set of about 20,100 grants potentially relevant to CDR research. A machine-learning classifier (ClimateBERT based on DistilRoBERTa[3]), fine-tuned on a large set of annotated scientific abstracts, was then used to differentiate between CDR grants and grants related to CDR. A manually annotated test sample was used to evaluate this classifier, and good performance (F1-score = 0.84) was found. A multi-class model was used to annotate the CDR methods that each grant covers, which worked with moderate performance (F1-score = 0.72). To identify grants on newly added CDR methods (biomass sinking, biomass burial, mineral products, harvested wood products, direct ocean capture and bio-oil storage), we additionally used a zero-shot classification approach prompting GPT 5.1 via the batch API. We evaluated this approach to find a similar performance as the ClimateBERT classification.

Other aspects analysed in the report – including the value of grants, the country of the funder and the receiving research organization, and the research fields – are provided directly in the Dimensions data. Data on the amount of funding were missing for 37% of the projects; these data were imputed using the average project funding. We compared these to estimates gained by imputing missing values based on median funding values as well as CDR method-specific mean values, which gave very similar estimates. The 10th and the 90th percentile of the project value distribution were used to estimate a range that should reflect the uncertainties in the calculations for estimating the total funding.

The update of grants data for the 3rd Edition of this report revealed that there are emerging data gaps in the Dimensions grants data which have not been there before. From 2022 onwards, data from a few major funders from China, South Africa and Russia are missing. We therefore excluded all grants (about 1,000 in total) from these funders for plots that show trends over time (Fig. 2.1) to not distort the trends shown.

Grants co-funding CDR research

To assess the importance of grants that support CDR research but do not mention CDR explicitly in their high-level summaries ("co-funding grants"), the analysis proceeded in three steps. First, a comprehensive keyword search was conducted for each CDR method to retrieve an initial set of 171,417 research publications from the Dimensions database. Using titles and abstracts, the same machine-learning relevance classifier as used for the CDR grants was then applied to identify both CDR relevance and methods covered. This resulted in a dataset of 75,319 CDR publications. Second, for all publications acknowledging funding (24%), the underlying grants were extracted and downloaded from Dimensions, yielding 14,573 unique research grants. In the third step, these grants were classified as either CDR grants or co-funding grants using the machine-learning approach, resulting in a final dataset of 10,960 co-funding grants.

About 17% of the grants in the final dataset lacked information on funding amounts. Missing values were imputed using the median grant amount, with the 10th and 90th percentiles used to define the uncertainty range. Finally, to estimate how much funding from co-funding grants could be attributed to CDR research publications, each grant's total funding volume was weighted by the share of CDR publications among all publications linked to that grant in the Dimensions database. Note that grants with a funding volume above US$1 billion were excluded because they correspond to large framework programmes (e.g. US national laboratory maintenance) that are not the type of targeted research grants the analysis focuses on. Depending on whether these outliers are included or not, the estimates for funding amounts that can be linked to co-funding grants change by an order of magnitude.

A2.2 Method for tracking scientific research on CDR

Number and methods of CDR publications

This report uses an AI-based approach to identify research publications on CDR in the English-language scientific literature. The methodology for this indicator follows Lück et al. (2025)[4] and the methodology used in the last edition of the State of CDR, albeit with some changes. First, combinations of search queries were designed for each CDR method based on a comprehensive list of keywords. Compared to the previous edition, we revised all queries to cover all fields of meta-data that can be searched in the scholarly databases (title, abstract and keywords). The search strings are validated against a set of studies included in the IPCC Sixth Assessment Report, ensuring that these studies were returned by the literature search and further refined with manually annotated data. For this edition, we also added new queries for six CDR methods: direct ocean capture, biomass sinking, biomass burial, mineral products, bio-oil storage, and wood products. These were evaluated against a validation set of 10-20 articles per CDR method. Using all search strings, about 412,000 records (after deduplication) were retrieved from four bibliographic databases: OpenAlex, Web of Science, Dimensions and Scopus. Note that the analysis in *The State of CDR* 1st Edition queried the Web of Science and Scopus, while the 2nd Edition relied only on the open-access database OpenAlex. The results in the two previous editions are therefore not directly comparable with the results of this edition.

For this edition, we expanded the dataset of manually screened and labelled abstracts developed in the last edition by manually assessing the suitability for inclusion (relevant/irrelevant) and the specific CDR method being studied for the new CDR methods considered in this edition. The labelled data were then used to train state-of-the-art machine-learning classifiers to predict a total of 119,000 relevant CDR research publications as well as the CDR methods covered within them. This automated approach enables a comprehensive search for scientific literature in bibliographic databases while still ensuring a high level of precision in the identification of relevant studies (see Table A.2.1). To identify the newly added six CDR methods, we applied a new approach, using zero-shot classification based on prompting a large language model (GPT 5.2 via the batch API). We evaluated the approach by manually labelling a validation dataset of 100 abstracts from each of the six queries and compared manually and LLM-generated labels to find a good enough performance. To reduce the rate of false positives, we only included records that were retrieved with the search query for the respective CDR method.

	Name (Annotation counts*)	Model	Precision	Recall	F1
Inclusion	Relevant (yes=3,968, no=4,161)	CLIMATEBERT	*89%*	*94%*	*91%*
CDR Method	BECCS (yes=331, no=4,998)	CLIMATEBERT	*90%*	*92%*	*91%*
	Biochar (yes=613, no=4,725)	TINYBERT	*98%*	*100%*	*99%*
	Biomass burial (yes=12, no=323)	LLM**	*80%*	*69%*	*72%*
	Biomass sinking (yes=12, no=323)	LLM	*80%*	*33%*	*47%*
	Bio-oil storage (yes=5, no=330)	LLM	--	--	--
	CDR (general) (yes=105, no=5,233)	CLIMATEBERT	*100%*	*50%*	*67%*
	Coastal wetland restoration (yes=343, no=4,954)	CLIMATEBERT	*98%*	*100%*	*99%*
	DACCS (yes=234, no=5,103)	SCIBERT	*83%*	*97%*	*89%*
	Direct ocean capture (yes=21, no=314)	LLM	*76%*	*76%*	*76%*
	Enhanced weathering (land-based) (yes=164, no=5,175)	SCIBERT	*74%*	*96%*	*84%*
	Forest-based CDR (yes=230, no=5,239)	SCIBERT	*72%*	*94%*	*82%*
	Harvested wood products (yes=54, no=281)	LLM	*82%*	*78%*	*80%*
	Mineral products (yes=47, no=288)	LLM	*70%*	*70%*	*70%*
	Ocean alkalinity enhancement (extended) (yes=467, no=7,745)	SCIBERT	*75%*	*93%*	*83%*
	Ocean fertilization or artificial upwelling (yes=92, no=5,247)	CLIMATEBERT	*100%*	*93%*	*96%*
	Peatland restoration (yes=163, no=5,172)	SCIBERT	*96%*	*88%*	*91%*
	Soil carbon sequestration (yes=465, no=4,849)	SCIBERT	*82%*	*82%*	*82%*

Table A2.1 *Overview of model performance for different labelling tasks.*

*Notes: *Count of manual annotations used for the training/cross-validation of the classifier. **Because the positive manual annotations are not very prevalent in our sample for the newly added CDR methods, the evaluation metrics come with a high uncertainty.*

Impact of CDR publications

To measure the impact of scientific publications on CDR in the scientific literature, we conduct an analysis of citations. Beyond simple statistics, we apply a methodology that compares CDR publications to publications on other low-carbon technologies from the same years and journals, as developed by Tripodi et al. (2024).[5] First, the method matches each CDR publication to a corresponding publication on low-carbon technology from the same year and journal. As in the original paper by Tripodi et al., we only consider publications that already had at least 5 years to accrue citations. Drawing on publication meta-data from the Dimensions database, this results in approximately 17,200 matched pairs of publications between 2000 and 2021 with citation counts as of January 2026. Second, the method applies a generalized linear model (GLM) based on the negative binomial distribution to estimate how the association of publications to different CDR methods affects citation counts. The model controls for common bibliometric factors known to influence citation counts such as the year of publication, the number of authors and the open access status of the publication.

A2.3 Methods for tracking patents on CDR

This report uses patent data from the new, openly available EPO Technology Intelligence Platform maintained by the European Patent Office. Granted patent data from all jurisdictions between the years 2005 and 2025 which was available by the 18th of January 2026 is considered. To classify the CDR methods, a retrieval augmented generator (RAG) powered by an offline version of the 8B parameter llama3 model is used, which classifies each technology several times to improve reliability. The data source for the RAG was the table of CDR methods for this edition. In order to use the technology classifier technique, all titles and abstracts in languages other than English were first translated to English using Google Translate. The classes used in the technology classifier technique are updated to align with the updated table of CDR methods in *The State of CDR* 3rd Edition, and the description for each method is also taken from this table. Given the methodological changes, the analysis results from this edition are not directly comparable to the results of the last edition.

To consider high-value patents, as in the last edition of this report and following Probst et al. (2021)[6], international patent families (DOCDB) are used that cover an invention in at least two jurisdictions. For the climate mitigation patent families, the Cooperative Patent Classification scheme is used, which provides the Y02 classification for "technologies or applications for mitigation or adaptation against climate change".[7] From the Y02 classifications, Y02A is excluded, since it refers to technologies for adaptation to climate change. The other Y02 classifications (Y02B, C, D, E, P, T, W) cover clean technologies in fields including energy, buildings, information and communication technologies, manufacturing, transportation, waste management and CO_2 capture and storage.

For the overall trend of CDR patent families, only those patent families that have at least one mention of a CDR technology by the technology classifier are included. To identify the weighted share of patent families per CDR method, each patent family was assigned to a CDR method weighted by the amount of times it was classified as this method and not as any other. For the comparison of regions, the patents were assigned to countries by the residence of the inventors. Not all patents of a family had a listed residence for each inventor, but each included patent family has at least one listed residence. Residences were then weighted such that the share of countries across a patent sums to one. The patenting office is not used since patents are territorial and often used by the applicants to gain access to a lucrative market.

The patent application and data gathering process leads to truncated data, for example, no CDR patent families were identified in 2025. Consequently, data from 2023 onwards should not be seen as representative and even data between 2020–2022 are likely to change subsequently. CCU and CCS are not included, unless the patent refers to either BECCS or DACCS.

A2.4 Methods for calculating growth rates

We aim to ensure comparability of growth rates across indicators by applying consistent calculation methods. For time periods longer than three years, growth rates are computed by fitting a linear function to the logarithmized data, equivalent to fitting an exponential function to the original data. The slope of the resulting linear function then corresponds to the exponent of the exponential function and represents the average exponential growth rate over the entire period considered. In cases where only two adjacent years are compared, we use the simple percentage difference as the growth rate.

References

1. Hook, D. W., Porter, S. J. & Herzog, C. Dimensions: Building Context for Search and Evaluation. *Front. Res. Metr. Anal.* **3**, (2018).
2. Herzog, C., Hook, D. & Konkiel, S. Dimensions: Bringing down barriers between scientometricians and data. *Quant. Sci. Stud.* **1**, 387–395 (2020).
3. Webersinke, N., Kraus, M., Bingler, J. A. & Leippold, M. ClimateBert: A Pretrained Language Model for Climate-Related Text. Preprint at https://doi.org/10.48550/arXiv.2110.12010 (2022).
4. Lück, S. et al. Scientific literature on carbon dioxide removal revealed as much larger through AI-enhanced systematic mapping. *Nat. Commun.* **16**, 6632 (2025).
5. Tripodi, G. et al. The public use of early-stage scientific advances in carbon dioxide removal: a science-technology-policy-media perspective. *Environ. Res. Lett.* **19**, 114009 (2024).
6. Probst, B., Touboul, S., Glachant, M. & Dechezleprêtre, A. Global trends in the invention and diffusion of climate change mitigation technologies. *Nat. Energy* **6**, 1077–1086 (2021).
7. European Patent Office. CPC – Cooperative Patent Classification. https://www.cooperativepatentclassification.org/sites/default/files/cpc/scheme/Y/scheme-Y.pdf (2026).