ARTICLES
https://doi.org/10.1038/s41558-021-01168-6

nature climate change

Check for updates

Machine-learning-based evidence and attribution mapping of 100,000 climate impact studies

Max Callaghan[1,2]✉, Carl-Friedrich Schleussner[3,4,5], Shruti Nath[3,6], Quentin Lejeune[3], Thomas R. Knutson[7], Markus Reichstein[8,9], Gerrit Hansen[10], Emily Theokritoff[3,4,5], Marina Andrijevic[3,4,5], Robert J. Brecha[3,11], Michael Hegarty[3], Chelsea Jones[3], Kaylin Lee[3], Agathe Lucas, Nicole van Maanen[3,4,5], Inga Menke[3], Peter Pfleiderer[3,4,5], Burcu Yesil[3] and Jan C. Minx[1,2]

Increasing evidence suggests that climate change impacts are already observed around the world. Global environmental assessments face challenges to appraise the growing literature. Here we use the language model BERT to identify and classify studies on observed climate impacts, producing a comprehensive machine-learning-assisted evidence map. We estimate that 102,160 (64,958–164,274) publications document a broad range of observed impacts. By combining our spatially resolved database with grid-cell-level human-attributable changes in temperature and precipitation, we infer that attributable anthropogenic impacts may be occurring across 80% of the world's land area, where 85% of the population reside. Our results reveal a substantial 'attribution gap' as robust levels of evidence for potentially attributable impacts are twice as prevalent in high-income than in low-income countries. While gaps remain on confidently attributabing climate impacts at the regional and sectoral level, this database illustrates the potential current impact of anthropogenic climate change across the globe.

There is overwhelming evidence that the impacts of climate change are already being observed in human and natural systems[1]. These effects are emerging in a range of different systems and at different scales, covering a broad range of research fields from glaciology to agricultural science and from marine biology to migration and conflict research[2]. The evidence base for observed climate impacts is expanding[3], and the wider climate literature is growing exponentially[4,5]. Systematic reviews and systematic maps offer structured ways to collectively identify and describe this evidence while maintaining transparency, attempting to ensure comprehensiveness and reduce bias[6]. However, their scope is often confined to very specific questions covering no more than dozens to hundreds of studies.

In the climate science community, evidence-based assessments of observed climate change impacts are performed by the Intergovernmental Panel on Climate Change (IPCC)[2]. Since the first assessment report (AR) of the IPCC in 1990, we estimate that the number of studies relevant to observed climate impacts published per year has increased by more than two orders of magnitude (Fig. 1a). Since the third AR, published in 2001, the number has increased tenfold. This exponential growth in peer-reviewed scientific publications on climate change[4,5] is already pushing manual expert assessments to their limits. To address this issue, recent work has investigated ways to handle big literature in sustainability science by scaling systematic review and map methods to large bodies of published research using technological innovations and machine-learning methods[7–11]. Much of this work builds on a related literature that has applied natural language processing (NLP) techniques to problems of evidence synthesis in the health sciences[12–14].

Fully utilizing the available knowledge on emerging climate change impacts is key to informing global policy processes[15] as well as regional and local risk assessments and on-the-ground action on climate adaptation[16,17]. While the global policy process may be served well with literature assessments presenting results aggregated on the level of continents or world regions[2,18], informing climate adaptation typically requires more highly localized and contextualized information on climate impacts[19,20].

Another core challenge of literature reviews and assessments of observed climate impacts relates to the question of whether climate impacts can be attributed to anthropogenic forcing[21]. While anthropogenic climate change signals have been identified in observed trends in a number of variables[21], including temperature[22], precipitation[23], sea level rise[24] and water resources[25], and selected extreme weather[26] events, the confidence in these assessments is still subject to substantial regional variations and remains relatively tentative at smaller spatial scales even if very high confidence levels can be reached for larger-scale (for example, global scale) attribution findings. Confidence also strongly depends on the variable being considered and specifically decreases further down the impact chain, that is, for indicators of changes in human and natural systems that are driven by changes in other climate impact variables[21].

[1]Mercator Research Institute on Global Commons and Climate Change, Berlin, Germany. [2]Priestley International Centre for Climate, University of Leeds, Leeds, UK. [3]Climate Analytics, Berlin, Germany. [4]Integrative Research Institute on Transformations of Human-Environment Systems, Humboldt University, Berlin, Germany. [5]IRI THESys and Geography Faculty, Humboldt University, Berlin, Germany. [6]Institute of Atmospheric and Climate Sciences, ETH Zürich, Zürich, Switzerland. [7]NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA. [8]Department of Biogeochemical Integration, Max Planck Institute for Biogeochemistry, Jena, Germany. [9]Michael Stifel Center Jena for Data-Driven and Simulation Science, Jena, Germany. [10]Robert Bosch Stiftung GmbH, Berlin, Germany. [11]Hanley Sustainability Institute, Renewable and Clean Energy Program and Physics Department, University of Dayton, Dayton, OH, USA. ✉e-mail: Callaghan@mcc-berlin.net

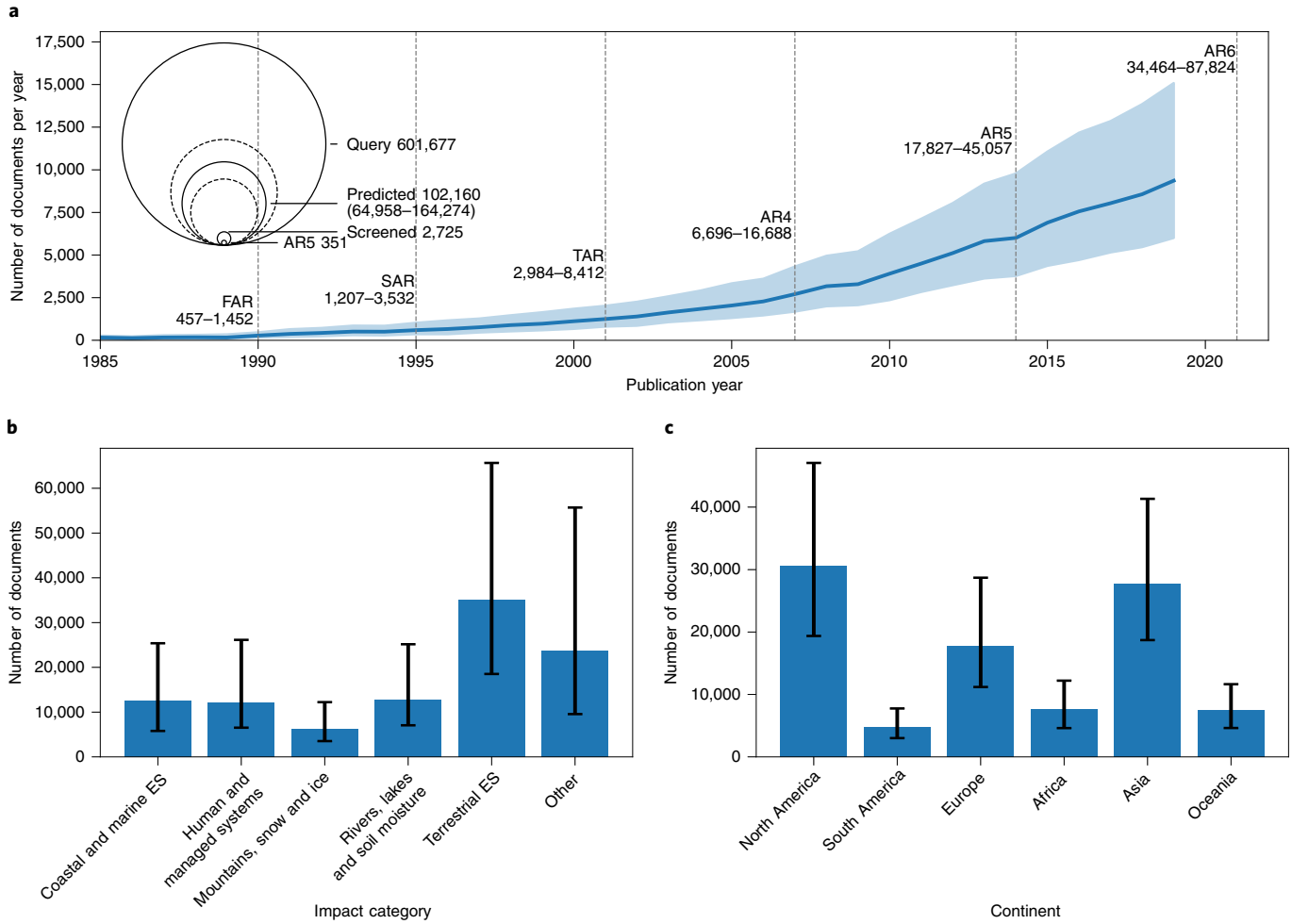


Fig. 1 | Results of the machine-assisted literature review. All results shown are based on our search queries and subsequent classification by the machine-learning pipeline. Uncertainty ranges denote the number of studies whereby the mean ±1 s.d. for the range of predictions for relevance and category membership obtained via bootstrapping is greater than 0.5 **a**, Growth in the scientific literature relevant to observed climate impacts over the past 30 years (cumulative totals for IPCC assessment periods are highlighted for reference). Inset: numbers of documents considered in the total query and in the IPCC AR5 WGII Tables 18.5–18.9. **b**,**c**, The estimated number of studies for each impact category (**b**) and continent (**c**) in our database (note that uncertainty bars consider uncertainty over relevance as well as impact category). ES, ecosystem; FAR, First Assessment Report; SAR, Second Assessment Report; TAR, Third Assessment Report.

In addition, methodological approaches and robustness criteria for climate change attribution differ widely among studies and disciplines, requiring expert judgement on a case-by-case basis to compile a comprehensive evidence base.

This points towards the added value of joining the body of evidence documenting regional or local-scale studies about climate impacts linked to common climate drivers such as temperature and precipitation change to a spatially resolved detection/attribution database of those variables.

Using Bidirectional Encoder Representations from Transformers (BERT), a state-of-the-art deep-learning language representation model[27], we develop a machine-learning pipeline to identify, locate and classify studies on observed climate impacts at a scale beyond that which is possible manually (Extended Data Fig. 1). We combine this spatially resolved dataset with an approach to attributing observed trends in surface temperature and precipitation at the grid-cell level (5°×5° and 2.5°×2.5° cells, respectively) to human influence on the climate. In doing so, we establish a new paradigm for assessing the impacts of climate change across human and natural systems.

Mapping over 100,000 impact studies

We searched two large bibliographic databases (Web of Science and Scopus) using an inclusive and transparent search method to systematically identify the literature on climate impacts. We assessed comprehensiveness by ensuring that our search string returned all references from tables 18.5–18.9 in the Fifth Assessment Report (AR5) Working Group II (WGII), which deal with the detection and attribution of climate impacts. Recent breakthroughs in NLP have extended the capabilities of text classification. BERT is a deep-learning language model trained using semi-supervised learning on massive corpora to represent text where word representations depend on context. Such models are able, to some extent, to capture the context-dependent meanings of texts. The pretrained model can be fine tuned on downstream tasks and has achieved state-of-the-art results across a range of NLP tasks. Using training data assembled by collaboratively screening and coding 2,373 abstracts, we use supervised machine learning, fine tuning the smaller and faster BERT variant DistilBERT[28], to classify (also on the basis of the abstract text) documents relevant to understanding the observed impacts of climate change in general and to predict

the human or natural systems for which they document impacts (the impact categories), as well as the climate variable(s) driving the documented impacts. Uncertainty estimates for the predictions are derived from bootstrapping. We employ a nested cross-validation approach to hyperparameter tuning, model selection and classifier evaluation and find that our binary inclusion classifier achieves an average F1 score (the harmonic mean of precision and recall) of 0.71 and receiver operating curve area under the curve (ROC AUC) score of 0.92. The prediction of impact type is achieved with an average macro F1 score of 0.84 while the prediction of climate driver is achieved with an average F1 score of 0.79 (see Methods and Extended Data Figs. 1–5 for a detailed explanation of the labelling, machine-learning approach and classifier performance).

Our query returned 601,677 unique documents (Fig. 1a), many more than would have been possible to screen by hand. We estimate that 102,160 (64,958–164,274) of these documents are relevant to understanding the observed impacts of climate change in general, judging from the spread of inclusion/exclusion predictions obtained from our model via bootstrapping (Fig. 1a). This base of relevant publications has grown substantially through the IPCC assessment cycles; 46,426 (34,464–87,824) articles have been published in the sixth assessment cycle so far. This represents more than twice the number of studies published during the AR5 period.

We used a geoparser pretrained using neural networks[29] to extract structured geographic information from the titles and abstracts of the studies in our database. Although the number of relevant studies in North America, Asia and Europe is much higher than in South America, Africa and Oceania, there is a large body of relevant studies available on all continents (Fig. 1c). Adjusted for population (Supplementary Fig. 1), the number of papers focusing on Oceania far exceeds the size of the literature devoted to other continents, with Africa and Asia receiving the least attention per million inhabitants. The relevant publications are also unevenly distributed across impact categories, with by far the largest number of studies, 34,974 (18,516–65,631), documenting impacts on terrestrial and freshwater ecosystems (Fig. 1b). However, the category with the comparably smallest coverage—mountains, snow and ice—still has 6,306 (3,526–12,225) studies.

In contrast to the map of observed impacts produced by the IPCC, we do not include only papers that formally attribute impacts to observed trends in climate. Instead, we take a more comprehensive approach reflecting that our objective is to map all possibly relevant studies on climate-related changes, rather than a list of studies where the relationship between an observed climate trend and specific impacts has been demonstrated with high confidence, or even linked to human influence on the climate. This includes studies attributing impacts to observed trends in climate variables, even where the authors do not attribute these trends to human influence, such as, for example, a study documenting the influence of the date of snowmelt on the phenology and population growth of mammals[30]. In addition, we include studies that provide evidence on the sensitivity of human or natural systems to climate metrics, such as how heart disease mortality responds to variations in temperature[31]. Finally, we include documents describing the impacts of extreme events and studies that detect significant trends in climate variables or climate extremes[32], regardless of whether these trends are in line with the expected effects of anthropogenic climate change. We exclude all studies that describe only potential or modelled impacts of future climate change.

Combining geolocated literature with climate information

To add context on the role of anthropogenic climate change in driving impacts, or more precisely the role of historical changes in anthropogenic climate forcing agents such as greenhouse gases and aerosols, we combine our literature database of studies selected using machine learning with spatially explicit analysis of detectable and attributable trends in two key climate variables. Combining evidence from climate model simulations and observational datasets allows identification of trends probably attributable in part to anthropogenic climate change for near-surface temperature and precipitation at the level of 5° (temperature) or 2.5° (precipitation) grid cells[22,23]. In this article, we apply this methodology to analyse trends from 1951 to updated observational data until 2018 for temperature (Fig. 2a) and until 2016 for precipitation (Fig. 2b). Grid cells in categories ± 2 or ± 3 show where trends cannot be explained by internal variability and are either consistent with or greater than the expected change in climate model simulations that include anthropogenic forcing agents. We infer that these cells display detectable and at least partly attributable trends (see Methods and Extended Data Figs. 6–8 for more details).

We next resolve the structured geographic information extracted from our studies, which ranges from continental scale to individual watersheds or communities, to sets of grid cells (Extended Data Fig. 9 and Methods). We can then derive the weighted number of studies per grid cell according to the number of grid cells to which each study relates. By combining studies related to temperature or precipitation with the gridded information on attributable trends in temperature and precipitation, this provides a necessary (though not necessarily sufficient) condition for a systematic two-step attribution to anthropogenic activities of the impacts predicted by the classifier[33]. Where studies documenting impacts associated with changes in temperature or precipitation co-occur with attributable trends in those variables, we claim that there is at least preliminary evidence for attributable impacts in these areas. This approach is similar in nature to the 'joint attribution' applied in IPCC AR4[34,35].

In general, we note that this type of automated assessment procedure is no substitute for careful assessment by experts but can identify large numbers of studies for a region that may point towards attributable human influence on impacts. Confidence in multi-step attribution claims depends on confidence in the attribution of the individual components (steps) along with the confidence or limitation in linking the different steps in the proposed causal chain[35]. One limitation of the partially automated two-step attribution approach is that we cannot verify that every temperature or precipitation trend cited in impact studies matches, in sign, magnitude or period, those attributed to human influence by the regional detection and attribution studies for temperature[22] and precipitation[23]. This is a greater problem for studies driven by precipitation, where both wetting and drying trends occur with greater temporal variation, although these make up the minority of partially attributed studies and grid cells. We also note that not all studies in our database document impacts in response to trends in climate variables. Where impacts are attributed to extreme events or variation in temperature or precipitation, the fact that recent trends in temperature or precipitation can be attributed to human influence provides important context but does not allow robust attribution of those impacts. These factors limit confidence in our cases of potential attribution of impacts to anthropogenic forcing. Our approach could be extended with more fine-grained analysis of studies or with attribution of additional signals in climate variables to make more robust attribution statements.

For 80% of global land area (excluding Antarctica), trends in temperature and/or precipitation can be attributed at least in part to human influence on the climate (purple cells, Fig. 2c). Using gridded population density data[36], we calculate that this covers 85% of the world's population. The majority of land grid cells show attributable warming trends, with exceptions where trends cannot be robustly distinguished from internal variability (white cells, category 0) or where there is insufficient data to establish trends (grey cells). For precipitation, attributable wetting and drying trends are found with greater geographical variation. There are also more grid cells where a trend in precipitation cannot be established, or where

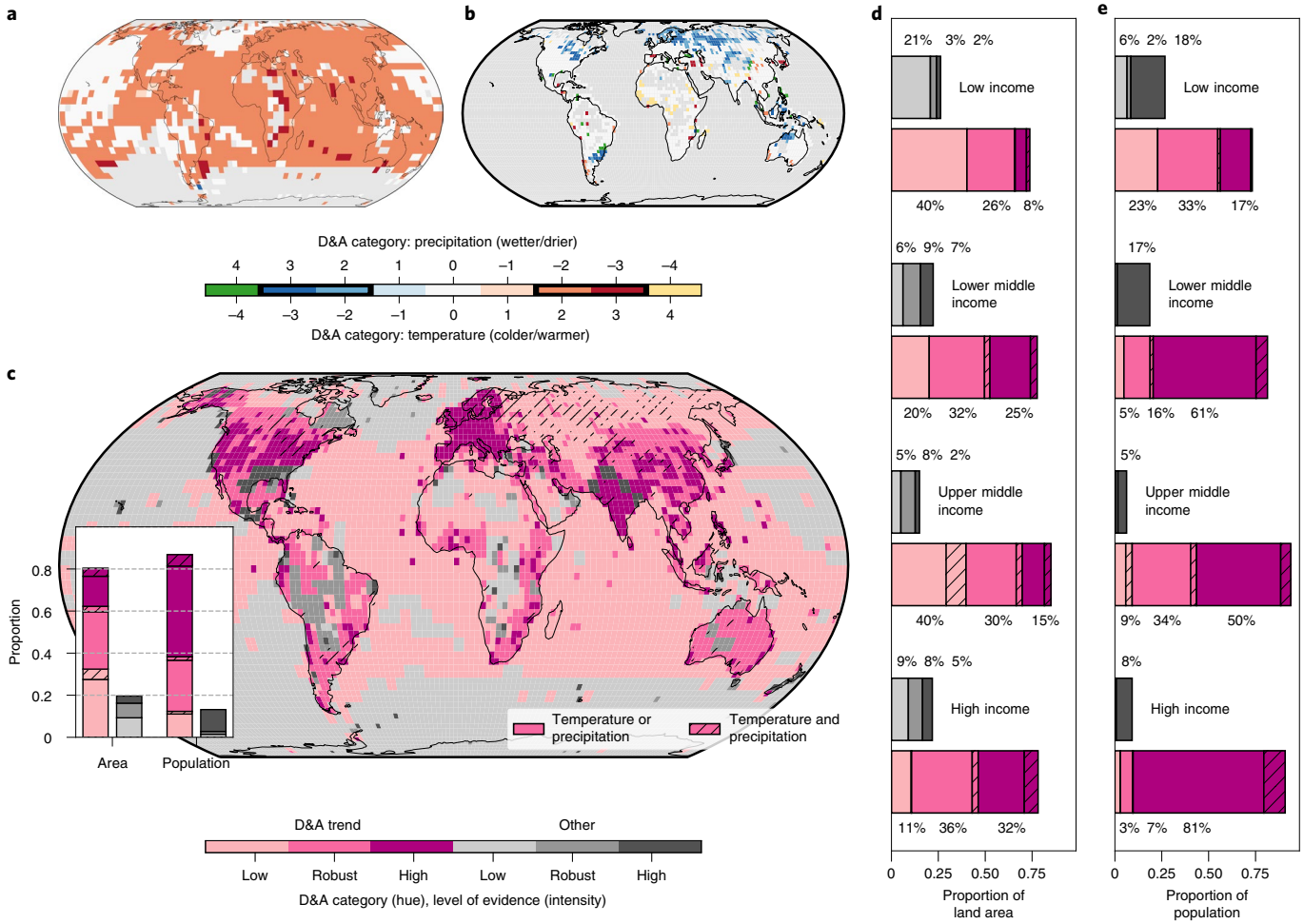


Fig. 2 | Potential attribution of impact studies to regional anthropogenic temperature and precipitation trends. a,**b**, Model-based assessment of the attribution of regional temperature for the time span 1951–2018 (**a**) and precipitation trends for the time span 1951–2016 (**b**) to human influence. Cooling/warming or drying/wetting trends in the regions marked as categories ±2 and ±3 are assessed as attributable in part to human influence (Methods). **c**, Global map of area-weighted studies coloured by the existence of detectable and attributable (D&A) trends (purple for attributable trends in at least one variable, cross-hatched for attributable trends in both variables, grey for no attributable trends) and indicating the localized evidence density (Low: <5 weighted studies; Robust: 5–20 weighted studies; High: >20 weighted studies). **d**,**e**, The proportion of land area (**d**) and population (**e**) with each grid-cell type, grouped by country income category.

the observed trend is opposite in sign to that simulated by climate model historical simulations (green and yellow cells, ±4).

Although most of the world's population resides in areas where trends in temperature and or precipitation can be at least partially attributed to human influence, there is substantial geographical variation in the degree to which the impacts of temperature and precipitation on human and natural systems have been studied. We characterize areas with fewer than 5 weighted studies per grid cell as displaying low levels of evidence, areas with 5–20 weighted studies as robust levels of evidence and areas with more than 20 weighted studies as high levels of evidence.

For 48% of global land area (hosting 74% of global population), we find robust or high levels of evidence of impacts on human and natural systems colocated with attributable temperature or precipitation trends (Fig. 2c). Areas with this combination of evidence are indicated by the darker purple cells. These constitute almost all grid cells in western Europe, North America, and South and East Asia, and there are parts of all continents that have similar pockets of substantial preliminary evidence.

However, for 33% of global land area (hosting 11% of global population), although there is evidence that long-term trends in precipitation and temperature are attributable at least in part to human influence, there is relatively little evidence in the existing literature about how these trends impact human and natural systems (Fig. 2c lightest purple shading). This imbalance suggests, in line with research measuring climate impacts using remote sensing[37], that the lack of evidence in individual studies is because these locations are less intensively studied, rather than because there is an absence of impacts in these areas. Parts of western Africa and southeastern, western and northern Asia contain several light purple grid cells where there is evidence to suggest that the climate (temperature and/or precipitation) has changed because of human influence, but there is little evidence on how this may be impacting human and natural systems. These demonstrable evidence gaps suggest a lack of impacts research commensurate with current knowledge of how the local climate (temperature and/or precipitation) is changing.

Some of the spatial features can be explained by the geographical characteristics. Among the regions with limited evidence are vast, sparsely populated and difficult-to-reach areas with a comparable uniform biosphere and climate such as Siberia or the Saharan desert. But beyond these features, our results clearly reveal a substantial 'attribution gap'. We find that 23% of the population of low-income

countries live in areas with low impact evidence despite at least partially attributable trends in temperature and/or precipitation (Fig. 2d). In high-income countries, this figure is only 3%. A density of 5 or more studies per grid cell with attributable impacts is 1.76 times as prevalent by population for high-income countries (88%) as for low-income countries (50%), while a density of 20 or more studies with attributable impacts is more than 4 times as prevalent (81% compared with 17%).

In the remaining grey grid cells (Fig. 2c), trends in precipitation and temperature have not been attributed to human influence on the climate according to the methodology in refs. [18,19], as applied to CMIP6 models. This does not rule out the possibility that some trends in precipitation or temperature have occurred in these regions that have been driven, at least in part, by human influence on the climate. However, due to various factors, such as lack of adequate observational data, high levels of natural variability compared with the climate change signal or limitations in modelling or estimated climate forcings, some observed changes that include anthropogenic contributions may not yet be attributable at the grid-cell level. This categorization of individual grid points may well change as new observational data are collected, as models improve, as the global climate continues to warm or as detection/attribution methodologies improve. Darker grey grid cells (10% of analysed land area) indicate where there are no detectable trends in temperature or precipitation that can be attributed to human influence at a grid-cell level but where there nevertheless appears to be substantial evidence that local trends in some climate variables lead to impacts on human and natural systems. For example, many studies refer to the impacts of temperature in the state of Western Australia, but of the 40 grid cells in the state, an attributable temperature trend can be demonstrated for 22 cells. For 16 of the remaining cells, a lack of data means that a detectable trend cannot be established, and for the remaining 2 cells, no attributable trend can be established.

The lightest grey cells (17% of land area) describe areas where we do not detect anthropogenic influence on regional temperature or precipitation and find few publications about the impacts of temperature or precipitation on human and natural systems. Apart from high latitudes and over the ocean, these cells are primarily in Africa. For example, in the light grey patch over the central part of sub-Saharan Africa, limitations of observed data, models or low signal-to-noise imply that we are unable to attribute temperature or precipitation trends to human influence on the climate using the methodologies employed here (Extended Data Fig. 4); further, we have identified few studies analysing the impacts of climate change on human and natural systems in those regions. These evidence gaps constitute substantial blind spots in understanding of climate impacts and, in some cases, understanding of attributable anthropogenic influence on regional precipitation and/or temperature.

In total, 57,366 studies discuss impacts related to a driver that our analysis suggests can be attributed in part to human influence on the climate in at least one grid cell to which the study refers. We find hundreds of partially or mostly attributable studies (where there are attributable trends in the relevant climate variable for at least 1% or more than 50% of grid cells, respectively) in each impact category across all continents (Fig. 3, indicated by the darker green and purple bars). This figure ranges from 268 (143–514) studies of impacts on mountains, snow and ice in Africa to 7,835 (4,308–13,552) studies of impacts on terrestrial ecosystems in North America. Wide confidence intervals here reflect the compound uncertainty deriving from classification of relevance, impact and driver.

Our analysis also allows quantification of how the share of research on each impact category varies from continent to continent. For example, research on human and managed systems makes up 12% of all research globally, but only 10% of research in Europe, compared with 19% in Africa. This focus on human and managed systems in Africa is remarkable given that the absolute numbers of studies in Africa (1,466) is similar to that in Europe (1,799) despite the vast difference in total numbers of studies between the two continents. This greater share of research in Africa documenting impacts in human and managed systems may reflect the high vulnerability of particularly sub-Saharan Africa to climate impacts[38].

Discussion and conclusion

We develop a two-step attribution process that combines a transparent and reproducible[39,40] machine-learning approach to identifying studies on observed climate impacts with model-based assessments of detectable anthropogenic contributions to historical temperature and precipitation trends. Using machine learning to scale up evidence synthesis allows us to map 100,000 studies of climate impacts, providing a comprehensive picture of the evidence base. Bringing together these two lines of evidence on climate change and climate impacts provides a new bridge between the climate science community and the impacts, adaptation, and vulnerabilities communities, and highlights the synergistic nature of their approaches.

Our spatially resolved approach allows for a systematic provision of regional to local, sector-specific climate impact information to local or regional experts and adaptation practitioners. This offers perspectives for a new climate service supporting the uptake of scientific information in local contexts and providing relevant information for adaptation action. Second, the quantification of an 'attribution gap' highlights the need for more research on climate impacts in low-income countries. Furthermore, the automated nature of the assessment allows for continuous updating of the database, creating a 'living' evidence map that can also be improved and extended by incorporating additional sources of relevant publications (for example, non-English-speaking evidence or improved/expanded regional detection/attribution studies) and targeted assisted learning in regional or topical areas of interest.

The compiled database is vast but neither complete nor perfect. Our systematic query-based literature search is extensive but will also exclude some relevant studies. The selection and categorization of studies was achieved using machine learning, meaning that results are subject to additional uncertainties, which compound for each level of classification. Further, documents were coded only at the abstract level, and only the abstracts were used as inputs to our classifiers. Given the relative simplicity of the types of information we extract (focusing on the impact area studied and the documented driver), we expect them to be covered in the abstract, which provides the condensed summary of the study's findings. Applying classifiers to noisy full texts that contain contextual information and related research as well as the results and topic of a study would greatly increase the risk of false positives. We thus find our approach well justified for such high-level syntheses.

The database we assemble will also incorrectly exclude some relevant documents and contain some documents that have been incorrectly included or incorrectly coded, but the approach enables us to report both classifier performance and associated uncertainties. In addition, some included studies may be of low quality as no process for critical appraisal (a key component of formal systematic reviews) was followed either by human reviewers or in the machine-learning pipeline. In the case of systems subject to other anthropogenic interference such as the global biosphere, managed systems such as agriculture or human systems themselves, identifying a robust climate change driver requires careful assessment of other socioeconomic factors[41,42], adding additional levels of complexity[43].

The two-step attribution process is also applied only for the subset of papers that provide evidence on impacts driven by temperature and precipitation. Exploring the role of human influence for studies analysing the effects of factors other than trends in mean temperature or precipitation as the main driver would require additional attribution strategies, but these could, in principle, be

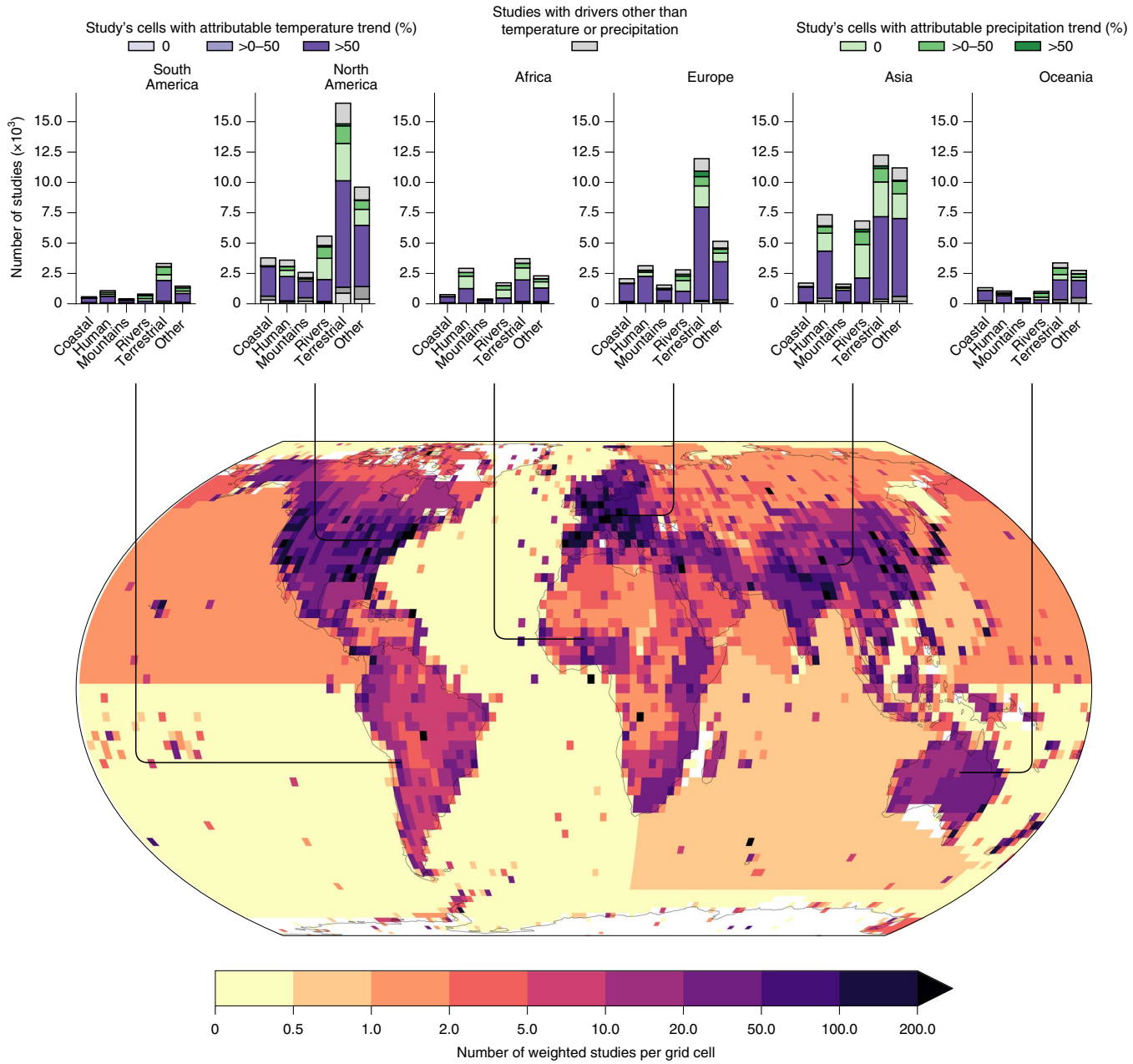


Fig. 3 | A global density map of climate impact evidence. Map colouring denotes the number of weighted studies per grid cell for all evidence on climate impacts ($N = 77{,}785$). Bar charts show the number of studies per continent and impact category. Bars are coloured by the climate variable predicted to drive impacts. Colour intensity indicates the percentage of cells a study refers to where a trend in the climate variable can be attributed (partially attributable: >0% of grid cells, mostly attributable: >50% of grid cells).

combined with individual studies in similar ways. There is a growing literature on attributable human influence on a number of climate metrics at the regional scale as well as extreme events[44–46] and, therefore, much scope for expansion of this approach. Finally, we note that plausible causal chains of cascading impacts are not covered by our attribution approach (such as temperature driving an increase in drought, leading to reduced agricultural yields) except where studies address each part of the causal chain.

These caveats highlight that the type of machine-learning-assisted evidence map we present here is no substitute for careful assessment by experts, either in the context of a gold-standard systematic review[47] or in IPCC assessments. However, in an age of 'big literature'[7,9], it is an invaluable complement. The use of machine learning means we consider more evidence than would otherwise be feasible, showing where evidence appears to be more prevalent and where important gaps can be observed. While traditional assessments can offer relatively precise but incomplete pictures of the evidence, our machine-learning-assisted approach generates an expansive preliminary but quantifiably uncertain map. Further, it enables us to provide an automated, living systematic map of climate impacts that can be readily updated. Ultimately, we hope that our global, living, automated and multi-scale database will help to jump start a host of reviews of climate impacts on particular topics or particular geographic regions.

Machine-learning pipelines as developed here will be useful to prepare the IPCC for the age of big literature by scaling systematic evidence mapping approaches. However, our results also show how synthesis and transparency can be lifted to new levels by combining

hitherto disparate lines of evidence and reporting classifier performance as well as associated uncertainties. If science advances by standing on the shoulders of giants, in times of ever-expanding scientific literature, giants' shoulders become harder to reach. Our computer-assisted evidence mapping approach can offer a leg up.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41558-021-01168-6.

Received: 31 May 2021; Accepted: 31 August 2021;
Published online: 11 October 2021

References

1. Cramer, W. et al. in *Climate Change 2014: Impacts, Adaptation, and Vulnerability* (eds Field, C. B. et al.) 979–1037 (Cambridge Univ. Press, 2014).
2. IPCC *Climate Change 2014: Impacts, Adaptation, and Vulnerability* (eds Field, C. B. et al.) (Cambridge Univ. Press, 2014).
3. Hansen, G. The evolution of the evidence base for observed impacts of climate change. *Curr. Opin. Environ. Sustain.* **14**, 187–197 (2015).
4. Haunschild, R., Bornmann, L. & Marx, W. Climate change research in view of bibliometrics. *PLoS ONE* **11**, e0160393 (2016).
5. Grieneisen, M. L. & Zhang, M. The current status of climate change research. *Nat. Clim. Change* **1**, 72–73 (2011).
6. Haddaway, N. R. & Pullin, A. S. The policy role of systematic reviews: past, present and future. *Springer Sci. Rev.* **2**, 179–183 (2014).
7. Callaghan, M. W., Minx, J. C. & Forster, P. M. A topography of climate change research. *Nat. Clim. Change* **10**, 118–123 (2020).
8. Porciello, J., Ivanina, M., Islam, M., Einarson, S. & Hirsh, H. Accelerating evidence-informed decision-making for the Sustainable Development Goals using machine learning. *Nat. Mach. Intell.* **2**, 559–565 (2020).
9. Nunez-Mir, G. C., Iannone, B. V. III, Curtis, K. & Fei, S. Evaluating the evolution of forest restoration research in a changing world: a "big literature" review. *New For.* **46**, 669–682 (2015).
10. Westgate, M. J. et al. Software support for environmental evidence synthesis. *Nat. Ecol. Evol.* **2**, 588–590 (2018).
11. Lamb, W. F., Creutzig, F., Callaghan, M. W. & Minx, J. C. Learning about urban climate solutions from case studies. *Nat. Clim. Change* **9**, 279–287 (2019).
12. Cohen, A. M. An effective general purpose approach for automated biomedical document classification. *AMIA Annu. Symp. Proc.* **2006**, 161–165 (2006).
13. Marshall, I. J., Kuiper, J., Banner, E. & Wallace, B. C. Automating biomedical evidence synthesis: RobotReviewer. In *Proc. Association for Computational Linguistics Meeting* 7–12 (The Association for Computational Linguistics, 2017).
14. Baclic, O. et al. Challenges and opportunities for public health made possible by advances in natural language processing. *Can. Commun. Dis. Rep.* **46**, 161–168 (2020).
15. Schleussner, C.-F. & Fyson, C. L. Scenarios science needed in UNFCCC periodic review. *Nat. Clim. Change* **10**, 272 (2020).
16. Fankhauser, S. Adaptation to climate change. *Annu. Rev. Resour. Econ.* **9**, 209–230 (2017).
17. Bedsworth, L. W. & Hanak, E. Adaptation to climate change. *J. Am. Plann. Assoc.* **76**, 477–495 (2010).
18. IPCC *Special Report on Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation* (Cambridge Univ. Press, 2012).
19. Hallegatte, S. & Mach, K. J. Make climate-change assessments more relevant. *Nature* **534**, 613–615 (2016).
20. Conway, D. et al. The need for bottom-up assessments of climate risks and adaptation in climate-sensitive regions. *Nat. Clim. Change* **9**, 503–511 (2019).
21. Hansen, G. & Stone, D. Assessing the observed impact of anthropogenic climate change. *Nat. Clim. Change* **6**, 532–537 (2016).
22. Knutson, T. R., Zeng, F. & Wittenberg, A. T. Multimodel assessment of regional surface temperature trends: CMIP3 and CMIP5 twentieth-century simulations. *J. Clim.* **26**, 8709–8743 (2013).
23. Knutson, T. R. & Zeng, F. Model assessment of observed precipitation trends over land regions: detectable human influences and possible low bias in model trends. *J. Clim.* **31**, 4617–4637 (2018).
24. Nerem, R. S. et al. Climate-change-driven accelerated sea-level rise detected in the altimeter era. *Proc. Natl Acad. Sci. USA* **115**, 2022–2025 (2018).
25. Gudmundsson, L., Leonard, M., Do, H. X., Westra, S. & Seneviratne, S. I. Observed trends in global indicators of mean and extreme streamflow. *Geophys. Res. Lett.* **46**, 756–766 (2019).
26. Padrón, R. S. et al. Observed changes in dry-season water availability attributed to human-induced climate change. *Nat. Geosci.* **13**, 477–481 (2020).
27. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. Preprint at https://arxiv.org/abs/1810.04805 (2019).
28. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. Preprint at https://arxiv.org/abs/1910.01108 (2020).
29. Halterman, A. Mordecai: full text geoparsing and event geocoding. *J. Open Source Softw.* **2**, 91 (2017).
30. Lane, J. E., Kruuk, L. E. B., Charmantier, A., Murie, J. O. & Dobson, F. S. Delayed phenology and reduced fitness associated with climate change in a wild hibernator. *Nature* **489**, 554–557 (2012).
31. Zhang, Y. Q., Yu, C. H. & Bao, J. Z. Acute effect of daily mean temperature on ischemic heart disease mortality: a multivariable meta-analysis from 12 counties across Hubei Province, China. *Zhonghua Yu Fang Yi Xue Za Zhi* **50**, 990–995 (2016).
32. Barry, A. A. et al. West Africa climate extremes and climate change indices. *Int. J. Climatol.* **38**, e921–e938 (2018).
33. Hegerl, G. C. et al. Good practice guidance paper on detection and attribution related to anthropogenic climate change. In *Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Detection and Attribution of Anthropogenic Climate Change* (eds Stocker, T. F. et al.) (IPCC, 2010).
34. Rosenzweig, C. et al. in *Climate Change 2007: Impacts, Adaptation and Vulnerability* (eds Parry, M. L. et al.) 79–131 (Cambridge Univ. Press, 2007).
35. Rosenzweig, C. et al. Attributing physical and biological impacts to anthropogenic climate change. *Nature* **453**, 353–357 (2008).
36. *Gridded Population of the World, Version 4 (GPWv4): Population Density* Revision 11 (CIESIN, 2018).
37. Frank, D. et al. Effects of climate extremes on the terrestrial carbon cycle: concepts, processes and potential future impacts. *Glob. Change Biol.* **21**, 2861–2880 (2015).
38. Schleussner, C.-F. et al. 1.5 °C hotspots: climate hazards, vulnerabilities, and impacts. *Annu. Rev. Environ. Resour.* **43**, 135–163 (2018).
39. Peng, R. D. Reproducible research in computational science. *Science* **334**, 1226–1227 (2011).
40. Müller-Hansen, F., Callaghan, M. W. & Minx, J. C. Text as big data: develop codes of practice for rigorous computational text analysis in energy social science. *Energy Res. Soc. Sci.* **70**, 101691 (2020).
41. Shepherd, T. G. Storyline approach to the construction of regional climate change information. *Proc. R. Soc. A* **475**, 20190013 (2019).
42. Rosenzweig, C. & Neofotis, P. Detection and attribution of anthropogenic climate change impacts. *Wiley Interdiscip. Rev. Clim. Change* **4**, 121–150 (2013).
43. Mengel, M., Treu, S., Lange, S. & Frieler, K. ATTRICI 1.1—counterfactual climate for impact attribution. *Geosci. Model Dev.* https://doi.org/10.5194/gmd-14-5269-2021 (2021).
44. Gudmundsson, L. et al. Globally observed trends in mean and extreme river flow attributed to climate change. *Science* **371**, 1159–1162 (2021).
45. Diffenbaugh, N. S. Verification of extreme event attribution: using out-of-sample observations to assess changes in probabilities of unprecedented events. *Sci. Adv.* **6**, eaay2368 (2020).
46. Herring, S. C., Christidis, N., Hoell, A., Hoerling, M. P. & Stott, P. A. *Explaining Extreme Events of 2019 from a Climate Perspective* (American Meteorological Society, 2021).
47. *Cochrane Handbook for Systematic Reviews of Interventions* (John Wiley & Sons, 2019).

Methods

Outline. An overview of each of the steps taken in this study is given in Extended Data Fig. 1. These are outlined briefly here and explained in detail in the following sections. Over 600,000 documents were retrieved from bibliographic databases using a query. Then 2,373 of these documents were screened for relevance and coded for impact type and driver by human reviewers. The implicit inclusion and coding decisions for a further 351 documents were extracted from Tables 18.5–18.9 in the contribution of WGII to the AR5 of the IPCC[1]. Machine-learning classifiers were trained to predict relevance of documents using the titles and abstracts and were evaluated using nested cross-validation. The best performing classifier was then fit with all labelled documents using bootstrapping to make predictions with confidence intervals for the relevance of the remaining documents. Those documents predicted to be irrelevant were discarded, as were documents labelled by reviewers as irrelevant. Multilabel classifiers were then trained using the remaining labelled relevant documents and were assessed in a similar fashion using cross-validation. Predictions for impact type and driver were then made for the remaining unlabelled documents. Geographical entities were extracted from the included studies using a geoparser, and each entity was matched to the set of 2.5° grid cells overlapping it. Observed trends in precipitation and temperature were collected for 2.5° and 5° grid cells and compared with climate models to assess whether observed trends were detectable (that is, unusual compared with natural variability, and in the same direction as simulated by historical forcing climate model simulations) and at least partially attributable to human influence on the climate, as discussed in the following. Finally, documents predicted to be driven by temperature or precipitation were extracted from the database of studies and merged with the grid-cell attribution datasets so that each document could be characterized by the presence of human-attributable climate trends in the grid cells it referred to, and each grid cell could be characterized by the number of studies referring to it.

Search, screening and coding. *Search strategy.* Potentially relevant documents were assembled by developing a query to search bibliographic databases. To validate the query, we tested this against a set of records known to be relevant. Tables 18.5–18.9 in the AR5 WGII[1] contain the studies considered in their assessment of the observed impacts of climate change. After extracting these references, we built a query that would return all of the references in the tables that specifically referred to the role of climate change (rather than of counterfactual explanations for impacts). The query is reproduced in the Supplementary Information (in the format for Web of Science; the same query was used for Scopus) and consists of three lists of keywords linked with Boolean "AND"s. The first set of keywords refer to climate and climate variables, the second to impacts and the third to observations and attribution.

The query was performed on Scopus and the following citation indices from the Web of Science Core Collection:

- Science Citation Index Expanded (SCI-EXPANDED), 1900–present
- Social Sciences Citation Index (SSCI), 1900–present
- Arts & Humanities Citation Index (A&HCI), 1975–present
- Conference Proceedings Citation Index – Science (CPCI-S), 1990–present
- Conference Proceedings Citation Index – Social Science & Humanities (CPCI-SSH), 1990–present
- Emerging Sources Citation Index (ESCI), 2015–present

The queries were updated on 19 October 2020; Web of Science returned 411,194 documents, and Scopus returned 476,778 documents. The total number of records after deduplication through fuzzy title and publication-year matching using trigram similarity was 601,667. The queries were imported into a database and deduplicated using the NACSOS review platform[48].

Inclusion and exclusion criteria. We take a broad definition of climate impacts to include all studies relevant to understanding the observed impacts of climate change. This includes the following (with some documents belonging to more than one category):

- Studies that explicitly link impacts to climate change (8% of coded studies)
- Studies that link impacts to trends in climate drivers such as temperature or precipitation (42% of coded studies)
- Studies that link impacts to extreme climate events (6% of coded studies)
- Studies that link impacts to variation in climate drivers (39% of coded studies)
- Studies that document regional or local climate trends (11% of coded studies)

Documents that provide only evidence of likely future impacts of climate change were excluded.

With this broad definition of climate impacts evidence, we do not claim that each study alone is evidence of the impacts of climate change. Rather, taken together, and in the context of observations and climate models, this collection of included studies constitutes the evidence base necessary for understanding climate impacts.

Coding impacts and drivers. Where documents were selected for inclusion, reviewers coded the attribution category, the climate impacts and the drivers (where appropriate) for each paper. Impacts and their drivers were chosen from a selection of 75 specific categories, which were aggregated according to the hierarchy of categories included in the supplementary file category_aggregation.csv. Ninety-three percent of included studies coded impacts in one or more of the five broad impact categories used by IPCC AR5:

- Mountains, snow and ice (11.42% of included studies)
- Rivers, lakes and soil moisture (21.27% of included studies)
- Terrestrial ecosystems (33.13% of included studies)
- Coastal and marine ecosystems (13.21% of included studies)
- Human and managed systems (21.42% of included studies)

Remaining studies documented only trends in climate variables without reference to any of these systems.

Screening and coding. A total of 2,373 documents were screened by members of the author team using the NACSOS platform[48], of which 1,125 were included as relevant and coded for impacts and drivers. The median number of documents coded per user was 133, and the mean was 173.

In addition, documents extracted from the tables 18.5–18.9 in AR5 WGII were automatically labelled as relevant and tagged with the broad impact categories corresponding to the table in which they were found.

To mitigate a highly unbalanced sample (few relevant documents among many irrelevant documents), and to make best use of reviewing resources, some documents were selected for screening using an adapted active learning pipeline. With active learning, a classifier (see following section for details) is trained using existing screening decisions to predict the relevance of documents yet to be reviewed. Usually, reviewers screen subsequent documents in decreasing order of predicted relevance, and the classifier is periodically updated with the new data that have been generated. Given that our goal was not to screen all relevant documents but to generate useful labels efficiently, we created samples with relevance predictions greater than 0.2, 0.3 and 0.4 to exclude documents with a low likelihood of being relevant. Documents were first screened by a small group of reviewers who developed the categorization scheme for impacts and drivers. A subsequent set of documents was screened by all reviewers, and differences in coding were discussed and alterations recorded. Reviewers were then split into teams corresponding with the AR5 impact categories according to expertise and screened documents predicted to be rather relevant (>0.33) to the given category. Each team screened a sample of documents and discussed differences in screening and coding decisions. Teams reached average Cohen's Kappa scores between 0.66, indicating substantial agreement, and 1.0, indicating full agreement[49]. After this initial round of double coding, reviewers proceeded to screen documents individually. Additional documents were selected for screening using keyword searches (https://github.com/mcallaghan/regional-impacts-map/blob/master/literature_identification/category_keywords.ipynb)[50] to identify documents from infrequently appearing subcategories.

Because the documents selected using the methods described are unlikely to be representative of the full set of documents returned by the query, we also screened 732 documents drawn at random, which we used for validation.

Machine-learning classifiers for inclusion, impact type and drivers. We first trained a binary classifier to predict the inclusion/exclusion decision given by reviewers. We use a nested cross-validation (CV) procedure (Extended Data Fig. 2) to optimize parameter settings and evaluate the performance of a support vector machine (SVM) classifier[51] as well as a pretrained DistilBERT model fine tuned with our labelled dataset[28]. SVMs have a long history of applications in evidence synthesis[12], while the BERT[27] model recently achieved state-of-the-art results in a variety of NLP challenges and has begun to be used in evidence synthesis pipelines[8]. However, large language models such as BERT can have non-trivial climate impacts[52], motivating our decision to use the lighter and faster DistilBERT, which retains '97% of its language understanding'[28] with greatly reduced computational resource usage.

In our nested cross-validation procedure, we first separate those documents that were drawn at random from the population of documents identified by the query from the remaining unrepresentative documents. Only randomly selected documents are used in validation and test sets to ensure that the estimation of the performance of the classifier on the whole dataset is not biased. In the outer fold of the cross-validation loop, a separate test set is drawn from the randomly selected documents for each fold, k, and all other documents are assigned to the test set. The inner cross-validation loop draws k inner validation sets from the remaining random documents in the training set and allocates all other documents in the training set to an inner training set. The inner loop is used to optimize hyperparameters for each model using grid search: a model is initialized with each combination of hyperparameters and fit on each inner training set and evaluated on each inner validation set. The combination of hyperparameters with the best mean F1 score across inner folds is selected as the best model. This model is fit with the training data from the outer cross-validation and evaluated with the test data. The outer cross-validation thus returns k scores for each metric, which we report in the following. We note that our cross-validation approach, while transparent, robust and thorough, is computationally expensive and that alternative procedures such as random search may provide similar results at lower computational cost, or minor improvements at the same cost[53].

In principle, additional improvements to the model may also be generated through additional pre-training[54] using the unlabelled corpus of climate-relevant abstracts. Pre-training BERT-like models on climate science corpora remains an area for future investigation.

We evaluated our binary inclusion/exclusion classifiers with five inner and outer folds. DistilBERT clearly outperformed SVM across all metrics, achieving an average F1 score of 0.71 and an average ROC AUC score of 0.92 (Extended Data Fig. 3). A final DistilBERT model configuration was chosen using the same procedure on the outer folds. Each combination of parameter settings was tested on each outer fold, and the combination of parameter settings with the highest mean F1 score was selected.

This final model was used to predict the relevance of all remaining documents. To create a confidence interval for each prediction, five versions of the final model were trained on five folds of the data. Upper and lower estimates for each document are given by the mean plus or minus one standard deviation. All documents where the lower estimate was below 0.5 were excluded from the study.

We then trained multilabel classifiers to predict the impact category and the driver category of included documents. Classifiers parameters were optimized and classifiers evaluated with the same nested cross-validation method using only those labelled documents that were included. Because documents selected for screening using the active learning process are broadly representative of the documents to which the multilabel classifiers are applied, all documents selected in this manner are also used for validation. Due to the lower number of documents, and lower number of documents drawn from a random sample in this set, we used a smaller k value of 3 for cross-validation. We treat each class equally and optimize using the macro F1 score. For the prediction of impact categories, DistilBERT outperforms SVM, achieving a macro-averaged F1 score of 0.84 and a macro-averaged ROC AUC score of 0.95 (Extended Data Fig. 4.). For classification of climate drivers, we optimize for the macro-averaged F1 score for the categories temperature and precipitation. DistilBERT outperforms SVM, achieving an average F1 score of 0.79 and an average ROC AUC score of 0.86. Where no individual class has a prediction larger than 0.5, documents are classed as 'Other systems'.

Detection and attribution. To put our database of impact studies in context, we match studies with grid-cell-level detection and attribution of temperature and precipitation trends to human influence on the climate.

Updating attribution of temperature and precipitation trends. We followed a previously published methodology[22,23] used to attribute observed temperature and precipitation trends to human influence around the globe, at the level of typical climate model grid cells (5° grid boxes for temperature and 2.5° grid boxes for precipitation). The different resolutions are based on the available observed datasets, which we did not re-grid for our project. The method relies on a comparison of gridbox-scale trends in observational datasets for temperature (HadCRUT4 v4.6[55]) and precipitation (GPCC v2018 https://psl.noaa.gov/data/gridded/data.gpcc.html) with those produced in climate model runs from Coupled Model Intercomparison Project Phase 6 (CMIP6)[56]. The CMIP6 runs simulate climate changes over the historical period under the influence of either all forcings (both natural and anthropogenic, referred to as 'ALL') or natural forcings only (referred to as 'NAT').

We analysed the outputs of these simulations from ten CMIP6 models: MIROC6, IPSL-CM6A-LR, CanESM5, HadGEM3-GC31-LL, CNRM-CM6-1,GFDL-ESM4, CCESS-ESM1-5, BCC-CSM2-MR, NorESM2-LM and CESM2. The model selection was based on the availability of ALL and NAT as well as 'piControl' runs (simulating internal climate variations in the absence of external forcings, apart from a constant solar forcing). The analysis provides a test of the ability of the corresponding ALL simulations to reproduce the regional trends in annual mean temperature and precipitation against observational data[57]. For some models, the ALL simulations were not available after 2014, in which case we combined them with the first few years of the ssp585 simulations of future climate conditions to match the length of the observational data.

Linear trends over the 1951–2018 (for temperature) and 1951–2016 (for precipitation) periods were computed over each grid cell with adequate data for each observational dataset, following the criteria of refs. [7,8] (Extended Data Fig. 6a,b). For temperature, we computed a linear trend for each ensemble member of the HadCRUT4 dataset, from which observed trend distributions were derived. Precipitation trends were not computed over grid cells where less than 20% of data was available for the first or last 10% of the observed time series or where the entire time series had less than 70% of data available. For temperature, we divide the trend period into five roughly equal periods and require that each period has at least 20% temporal coverage for annual means. We consider an annual mean as available if at least 40% of the months are available for the year.

To be compared with the observational data, for each model the data from both the ALL and NAT runs were first re-gridded onto the observational grids (5°×5° for temperature and 2.5°×2.5° for precipitation), excluding times and grid locations where observed data were missing, before linear trends were computed over each grid cell in which adequate temporal coverage was available (Extended Data Fig. 6c,d). For each model, we then assessed the potential effect of internal variability by computing trends of the length being investigated in 50 random samples of the corresponding piControl runs from each model. The model control runs had beforehand been corrected for any long-term drift and the anomaly series adjusted by a factor to ensure consistency of low-frequency variability between model control runs and estimated internal variability from observations (further discussed in the following). We then combined the resulting trend distributions from the piControl runs with the trends computed in the ensemble mean of ALL and NAT runs. Following previous studies[22,23], the final trend distribution for temperature was based on an aggregate distribution of all constructed model trend distributions (and thus included the spread of different model ensemble means) whereas for precipitation, an average distribution of model trends across the ensemble was used (that is, the distribution had the average characteristics of the ten CMIP6 models).

Attribution categories were assigned to grid cells (Extended Data Fig. 6e,f) on the basis of where their observed trend (or trend distribution in the case of temperature) lay relative to the final trend distributions derived from the ALL and NAT runs. Over the grid cells where an observed trend was in the same direction (sign) as the mean of the ALL trend distribution and was outside the trend distribution 5th–95th percentile range for the NAT simulations, the observed trend was categorized as −3 (+3), −2 (+2) or −1 (+1) depending on whether it was significantly stronger, the same, or weaker than the simulated decrease (increase). Categories −3 (+3) and −2 (+2) are defined as decreases (increases) that are detectable and at least partially attributable to anthropogenic forcing, according to our methodology. Categories −1 (+1) are detectable but not attributable. If the observed trend was significantly different from the NAT distribution, but was in the opposite direction to the mean of the All-Forcing distribution, it was categorized as −4 (observed decrease, modelled increase) or +4 (observed increase, modelled decrease). All observed trends (or trend distributions, in the case of temperature) that intersected with the 5th–95th percentile range of the corresponding trend distributions derived from the NAT runs were categorized as non-detectable, or indistinguishable from natural variability (category 0). Note that for cases where observed trends or trend distributions had a different sign of the mean trend from that of the trend distribution derived from the ALL runs, but were within the range of the Nat run distribution, the corresponding grid cells were also categorized as non-detectable (category 0).

Once the grid cells were categorized, in the case of temperature the results were re-gridded to a 2.5°×2.5° grid to allow superposition with the categories obtained for precipitation.

Our analysis requires the internal variability for each grid location and variable to be estimated via model control runs. To compare observed estimated internal variability and trends with those generated by the model control runs, Extended Data Figs. 7 and 8 show fractional difference maps for estimated internal low-frequency variability (model versus observed) for each model individually and for the ensemble mean of the modelled variability (the latter being most relevant for our analysis, which is based on combined estimated variability across the models). The observed low-frequency internal variability is estimated by subtracting the multimodel ensemble All-Forcing change from the observations and computing the standard deviation of the annual residuals, after application of a seven-year running mean filter. For models, we use the simulated variability from the various control runs, again smoothed with the seven-year running mean smoother. The averaged internal low-frequency variability comparison plot for precipitation (Extended Data Fig. 7, top panel) shows reds in most regions, indicating that by this measure of internal low-frequency variability, the CMIP6 models tend to overestimate observed variability levels. So our detection results for precipitation will tend to be conservative while, conversely, the ability of All-Forcing to be consistent with observations will tend to be liberal because the modelled spread is relatively wide. However, blue regions are evident in Extended Data Fig. 7 in some tropical regions, including over Africa and South America, indicating an undersimulation of internal low-frequency variability there. We took the internal variability comparisons versus observed estimated internal variability in Extended Data Fig. 7 and adjusted the control-run variability and trends by the ratio of observed s.d./model s.d. before computing our assessment categories. Results without this variability adjustment (not shown) are broadly similar but show more category −4 (unexplained trends of incorrect sign) over Africa, where internal low-frequency variability appears to be underestimated in models according to this analysis; unadjusted results show slightly less detectable human influence in middle and high latitudes, where internal variability is apparently overestimated in models.

For surface temperature (Extended Data Fig. 8) the internal variability comparison results versus observed estimates are similar to those of Knutson et al.[22] for CMIP3 and CMIP5 with a mixture of results: models tend to simulate more internal variability than the observed estimate in northern mid- to high latitudes, typically less than observed over most other ocean regions at lower latitudes and mixed results over land regions. Whether we include the grid-point-scale adjustment of simulated internal variability in our detection/attribution analysis or not, the results are similar (unadjusted control-run-based assessment not shown). For the assessment of 1951–2018 observed trends (Extended Data Fig. 6), there are some additional regions with detectable anthropogenic warming compared with Knutson et al.[22], but that is as expected since the Knutson et al. analysis examined trends only through 2010. With the termination of the 'global

warming hiatus' around 2014, the additional recent years have been adding to an ongoing strengthening warming signal and leading to even greater assessed area with detectable anthropogenic warming. In Extended Data Fig. 6 and elsewhere in the study, we use the adjusted control-run results for our assessments for both temperature and precipitation.

Spatial resolution of studies. To match these data with the finest-scale resolution of our database, we resolved each study to the set of 2.5° grid cells contained by the smallest geographical entity extracted from each paper's title and abstract using the geoparser Mordecai[29]. For each study, we calculated the proportion of the grid cells that this entity corresponds to in which an attributable trend for each variable can be found. For example, Extended Data Fig. 9a,b shows that 20 out of Sudan's 27 grid cells show an attributable anthropogenic warming trend, so each study referring to Sudan and documenting impacts predicted to be driven by temperature receives a precipitation trend proportion value of 20/27. Such a study would therefore add towards the dark red bars in Fig. 3, which count studies where an attributable temperature trend can be demonstrated for more than 50% of the grid cells the study refers to.

We also calculate a weighted number of studies for each grid cell by adding 1 divided by the number of grid cells a study refers to each of those grid cells, and repeating this procedure for all identified relevant studies. Extended Data Fig. 9c,d shows 11 studies that refer to impacts predicted to be driven by temperature trends in Sudan, where Sudan is the smallest geographical entity mentioned. Each grid cell in Sudan therefore receives 11/27 weighted studies. Given that some geographical entities were too small to hold one 2.5° grid cell, their longitude–latitude values were interpolated to the nearest grid cell instead and the grouped studies apportioned to that one grid cell. Because four additional studies refer to Khartoum, we add 4/1 to the weighted studies value in the grid cell containing Khartoum.

Data availability

The results of this study are made available in a public repository[58].

Code availability

The code used to produce these results is made available in a public repository[50].

References

48. Callaghan, M., Müller-Hansen, F., Hilaire, J. & Lee, Y. T. NACSOS: NLP assisted classification, synthesis and online screening. *Zenodo* https://doi.org/10.5281/zenodo.4121526 (2020).
49. McHugh, M. L. Interrater reliability: the kappa statistic. *Biochem. Med.* **22**, 276–282 (2012).
50. Callaghan, M. Machine learning-based evidence and attribution mapping of 100,000 climate impact studies - code. *Zenodo* https://doi.org/10.5281/ZENODO.5327409 (2021).
51. Chang, C.-C. & Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**, 27 (2011).
52. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: can language models be too big? In *Proc. ACM Conference on Fairness, Accountability, and Transparency* 610–623 (Association for Computing Machinery, 2021); https://doi.org/10.1145/3442188.3445922
53. Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **13**, 281–305 (2012).
54. Gururangan, S. et al. Don't stop pretraining: adapt language models to domains and tasks. Preprint at https://arxiv.org/abs/2004.10964 (2020).
55. Morice, C. P., Kennedy, J. J., Rayner, N. A. & Jones, P. D. Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: the HadCRUT4 data set. *J. Geophys. Atmos.* https://doi.org/10.1029/2011JD017187 (2012).
56. Eyring, V. et al. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.* **9**, 1937–1958 (2016).
57. Beusch, L., Gudmundsson, L. & Seneviratne, S. I. Crossbreeding CMIP6 Earth system models with an emulator for regionally optimized land temperature projections. *Geophys. Res. Lett.* **47**, e2019GL086812 (2020).
58. Callaghan, M. et al. Machine learning-based evidence and attribution mapping of 100,000 climate impact studies - data. *Zenodo* https://doi.org/10.5281/ZENODO.5257271 (2021).

Acknowledgements

M.C. is supported by a PhD stipend from the Heinrich Böll Stiftung. J.C.M. acknowledges funding from the ERC-2020-SyG GENIE (grant ID 951542). S.N. and Q.L. acknowledge funding from the German Federal Ministry of Education and Research (BMBF) and the German Aerospace Center (DLR) via the LAMACLIMA project as part of AXIS, an ERANET initiated by JPI Climate (http://www.jpi-climate.eu/AXIS/Activities/LAMACLIMA, last access: 26 August 2021, grant no. 01LS1905A), with co-funding from the European Union (grant no. 776608). M.R. acknowledges support by the ERC-SyG USMILE (grant ID 85518). R.J.B. acknowledges support from the EU Horizon2020 Marie-Curie Fellowship Program H2020-MSCA-IF-2018 (proposal no. 838667 -INTERACTION). We thank F. Zeng for providing preliminary temperature and precipitation trend assessment results for our project. We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP6. We thank the climate modelling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access and the multiple funding agencies who support CMIP6 and ESGF.

Author contributions

M.C., J.C.M. and C.-F.S. designed the research. M.C. developed the coding platform and machine-learning pipeline to identify studies, with advice from M.R. M.C, C-F.S., G.H., Q.L. and E.T. developed the codebook and coordinated screening and coding. M.C., Q.L., S.N. and C-F.S. conceptualized the link to detection and attribution data. S.N. performed the univariate detection and attribution analysis of temperature and precipitation trends and assessment of internal variability, in consultation with T.R.K, who designed the methodology for these calculations. M.C. and S.N. designed and implemented the matching of studies with detection and attribution data. M.C., C-F.S., S.N., Q.L, G.H., E.T., M.A., R.J.B., M.H., C.J., K.L., A.L., N.v.M., I.M., P.P. and B.Y. contributed to screening and coding studies. M.C., C-F.S., J.C.M., Q.L. and S.N. wrote the manuscript with contributions from all authors.

Competing interests

The authors declare no competing interests.

Additional information

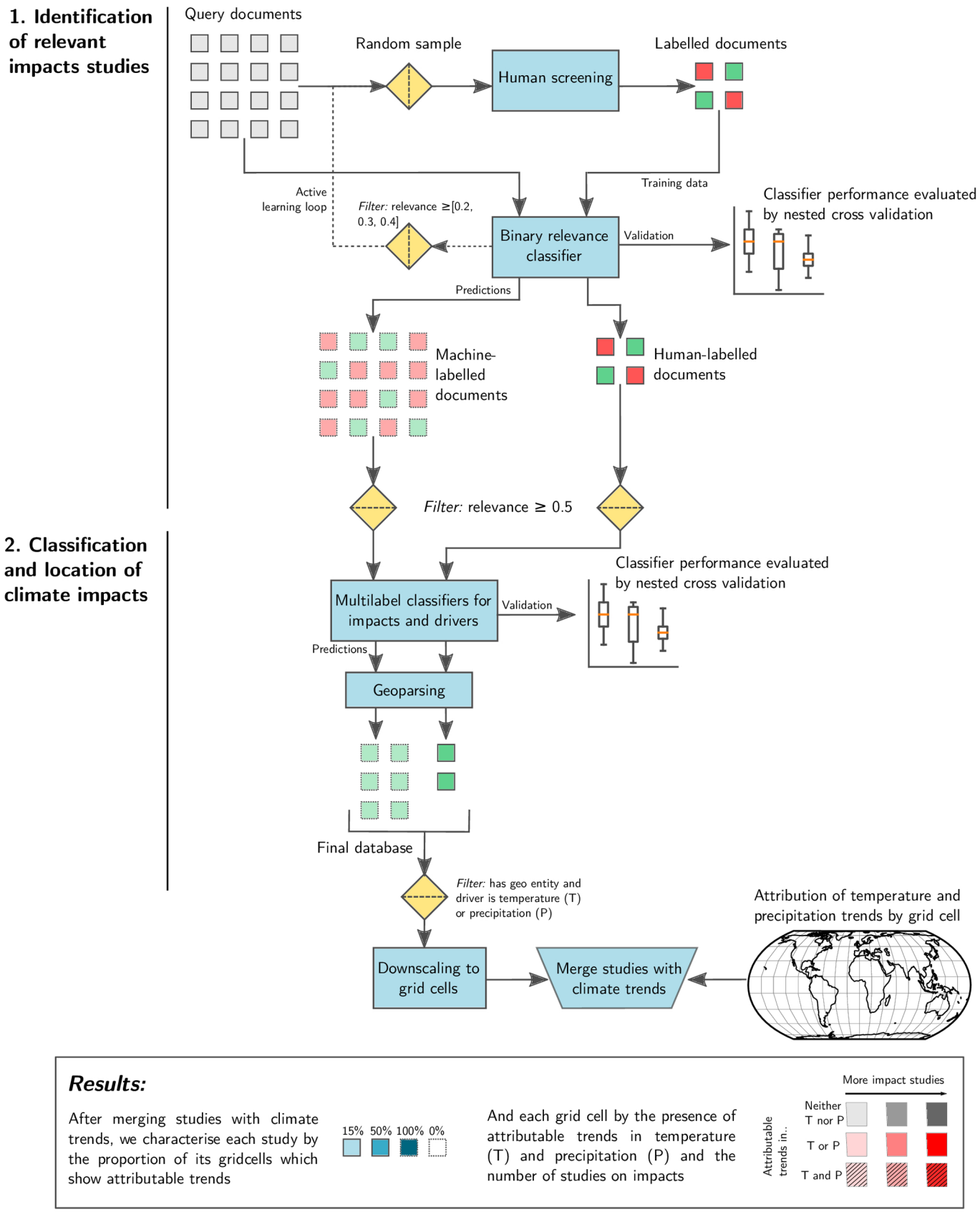
Extended data is available for this paper at https://doi.org/10.1038/s41558-021-01168-6.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41558-021-01168-6.

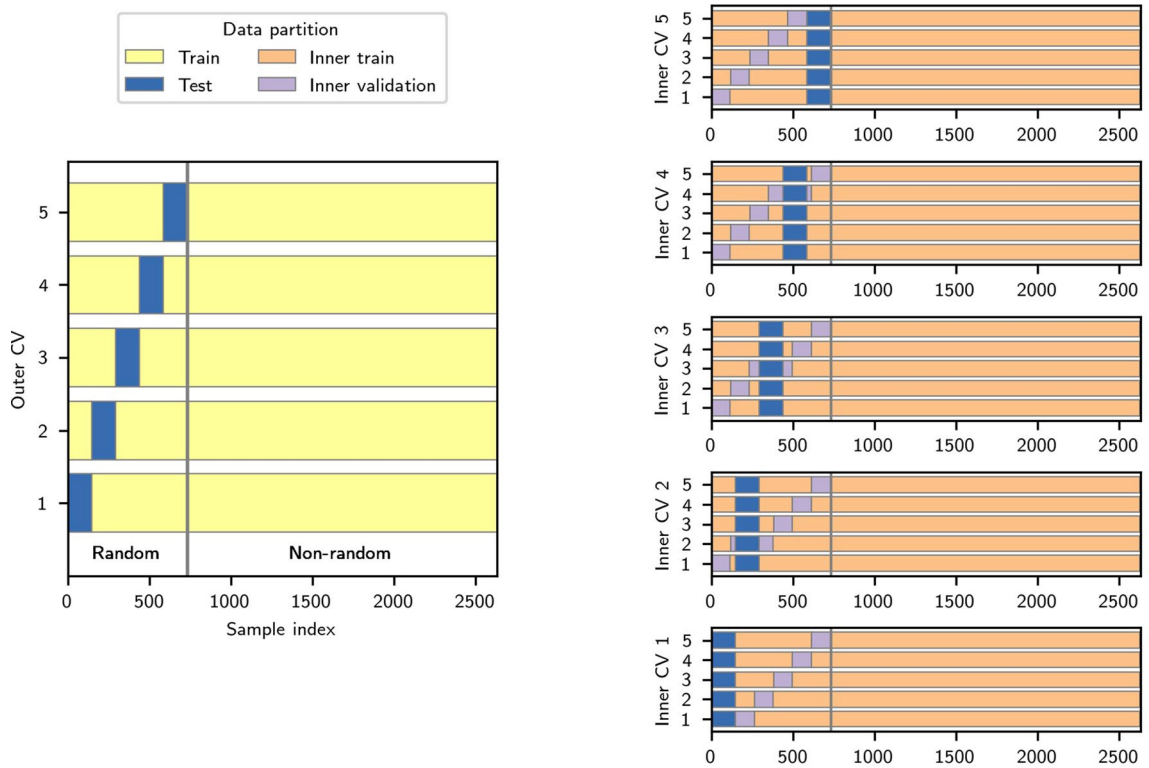
Correspondence and requests for materials should be addressed to Max Callaghan.

Peer review information *Nature Climate Change* thanks Abeed Sarker and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

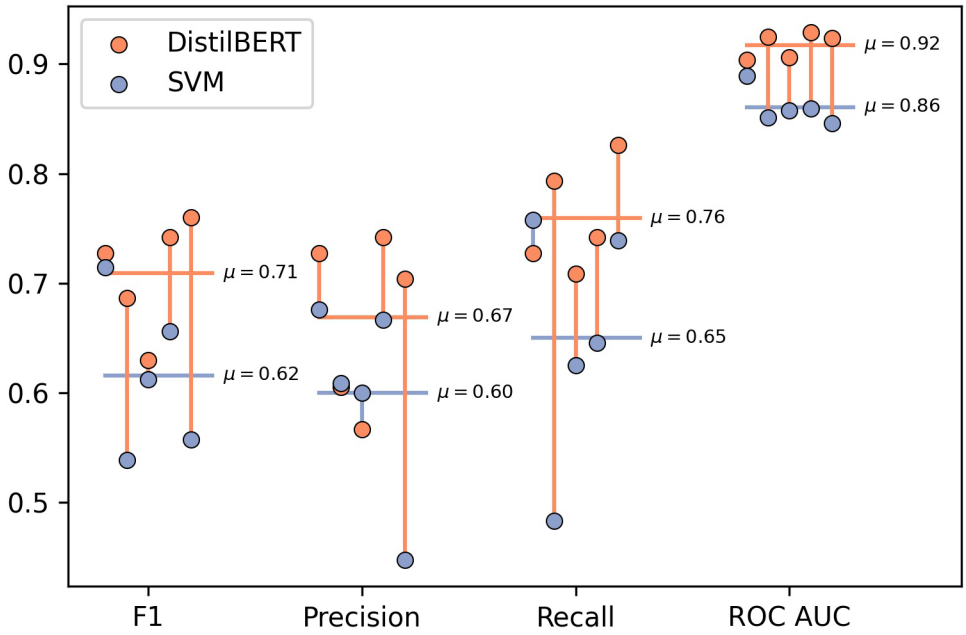
Reprints and permissions information is available at www.nature.com/reprints.



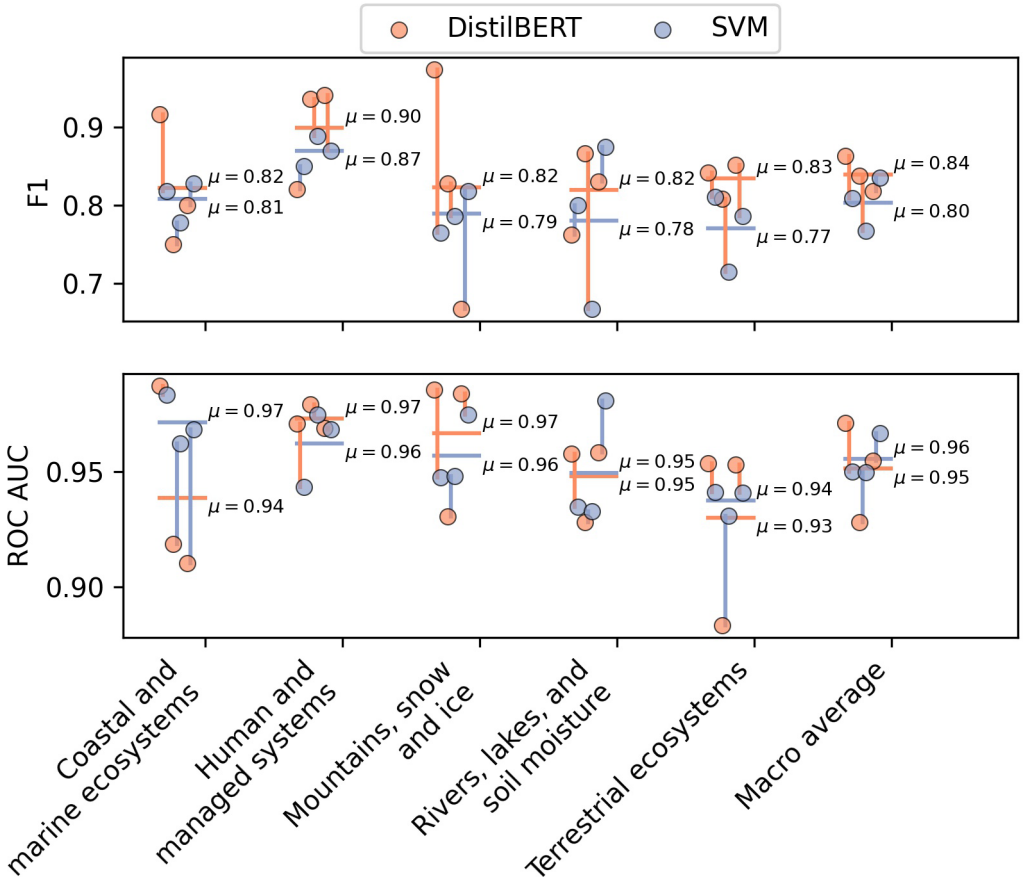
Extended Data Fig. 1 | A visual representation of the workflow of our machine learning assisted attribution map. Squares represent documents (not to scale), boxes represent the steps taken. Documents are screened by hand, and those labels are used to generate predictions and machine label documents. These machine-labelled documents are matched by location with information from observations and climate models on the detection and attribution of trends in temperature and precipitation.



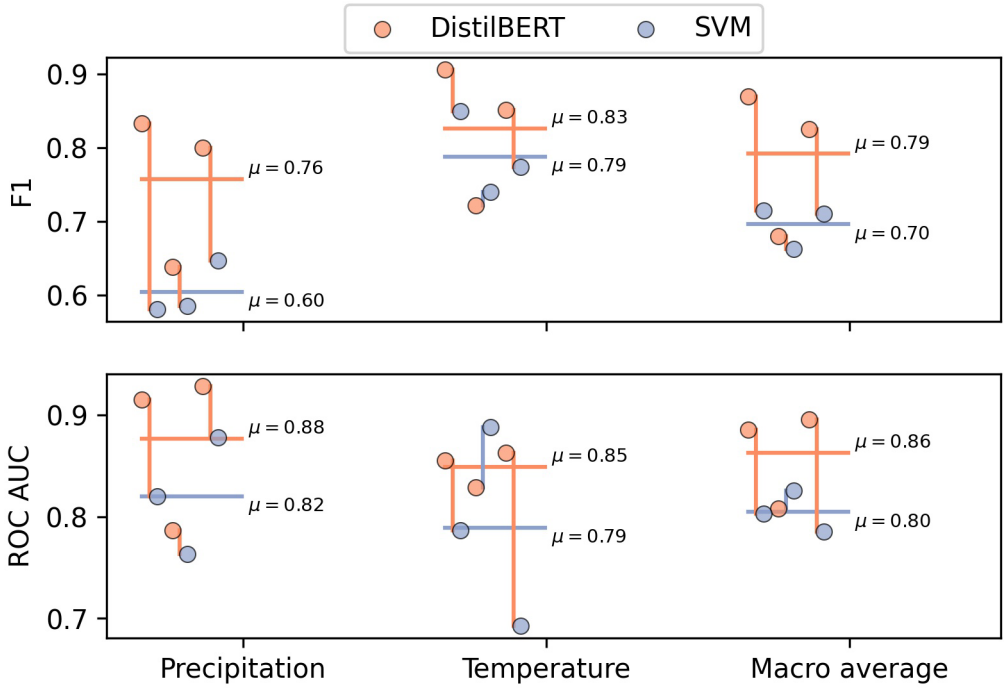
Extended Data Fig. 2 | Nested cross validation (CV) procedure for the binary relevance classifier. Models are fit using training documents and evaluated on validation/test documents. The inner CV loop is used to search for optimal hyperparameter settings, which are then evaluated on the outer test sets.



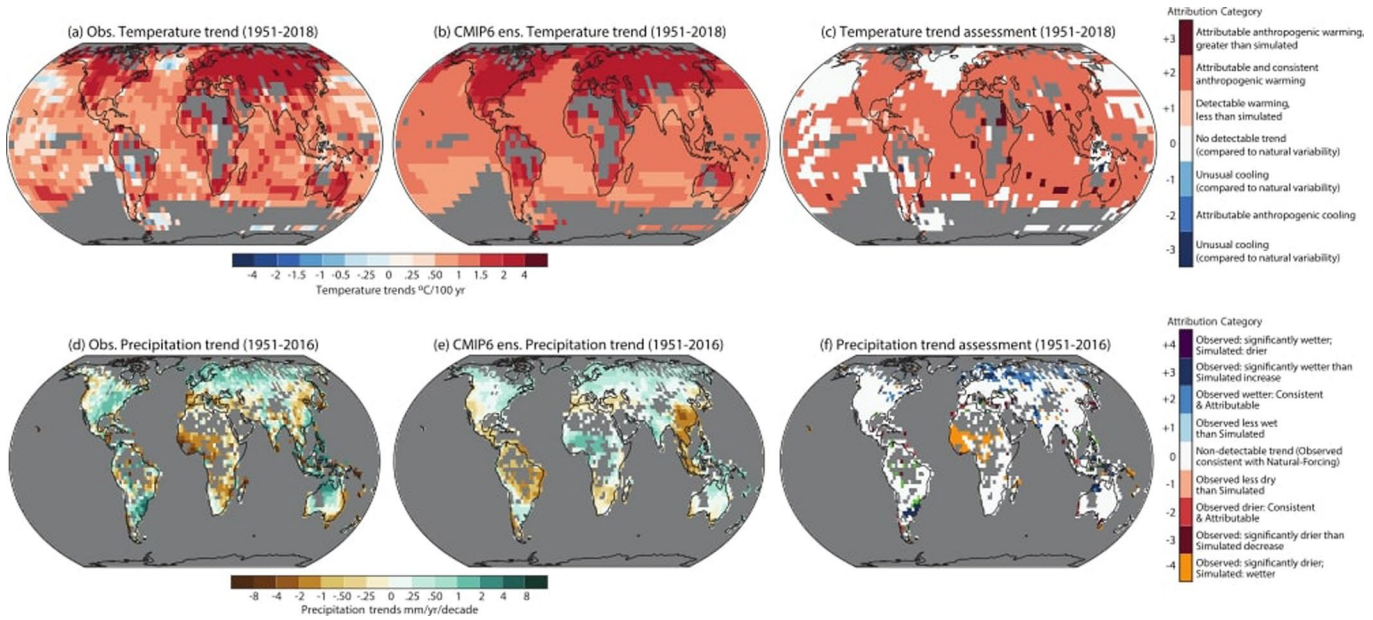
Extended Data Fig. 3 | Performance metrics for the binary inclusion/exclusion classifier. Each pair of dots represents the scores for a distinct cross-validation fold. Horizontal lines show the mean score across folds.



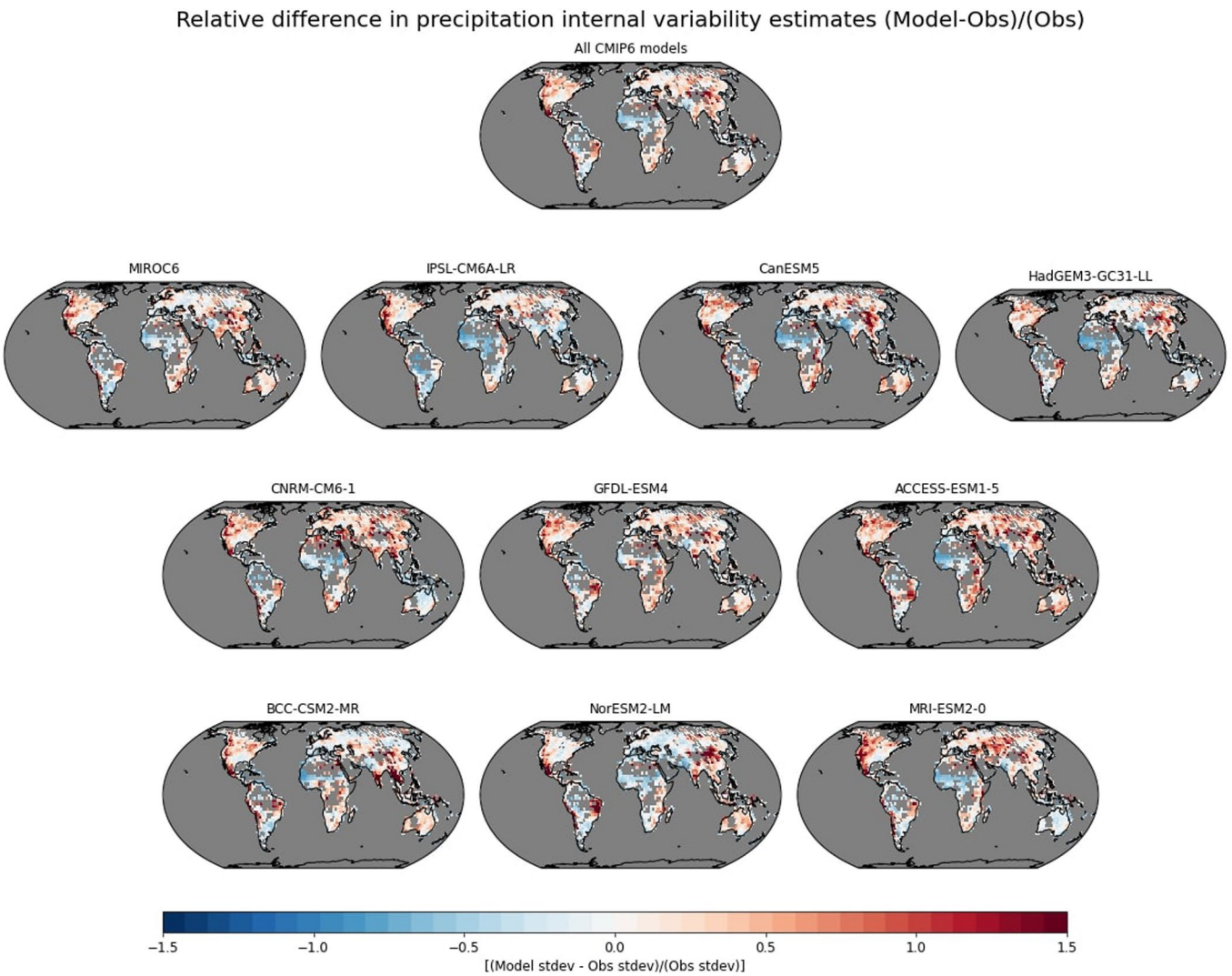
Extended Data Fig. 4 | Receiver operating curve area under the curve scores (ROC AUC) and F1 scores for the classification of impact categories. Each pair of dots represents the scores for a distinct cross-validation fold. Horizontal lines show the mean score across folds.



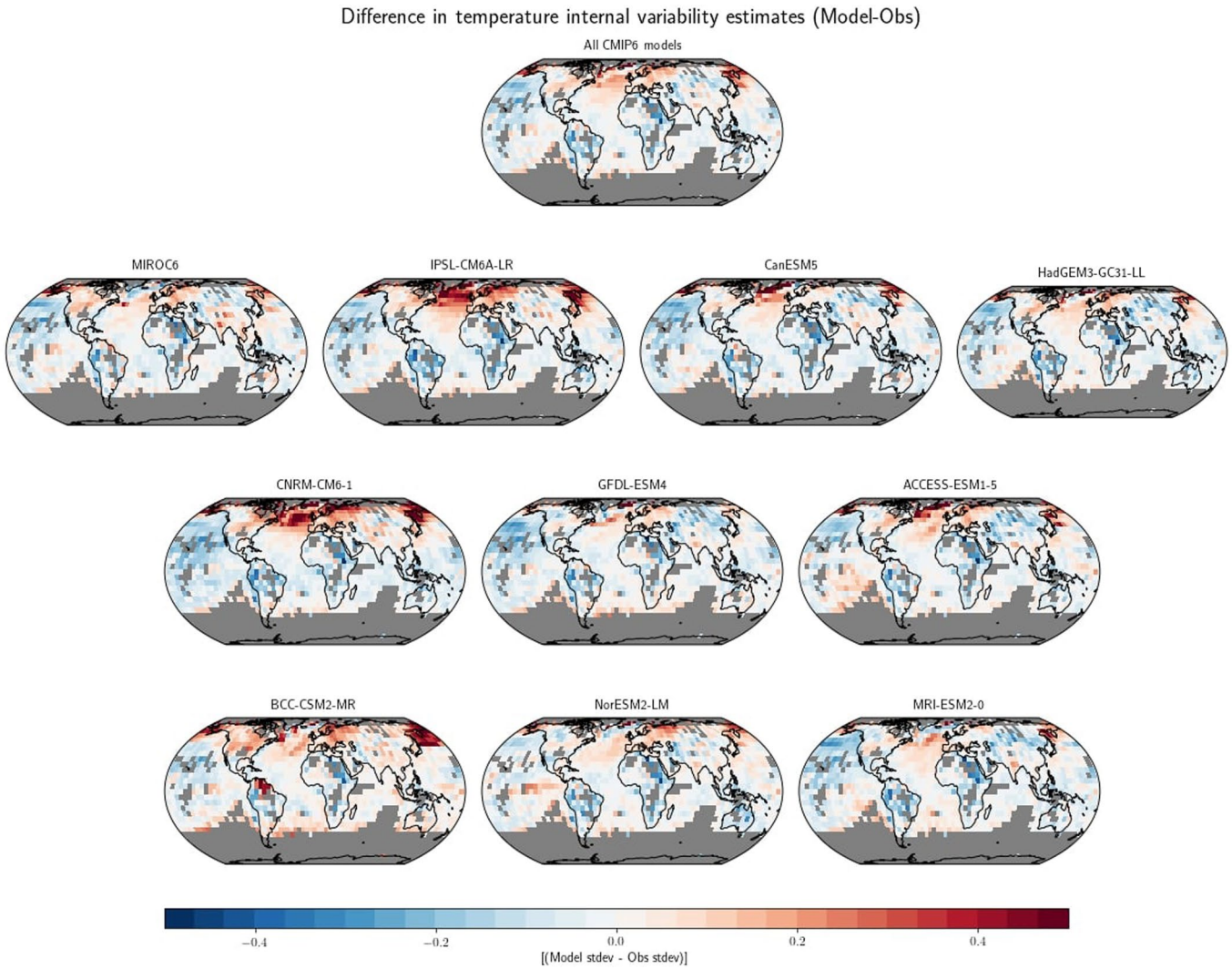
Extended Data Fig. 5 | Receiver operating curves area under the curve scores (ROC AUC)(ROC) and F1 scores for the classification of drivers. Each pair of dots represents the scores for a distinct cross-validation fold. Horizontal lines show the mean score across folds.



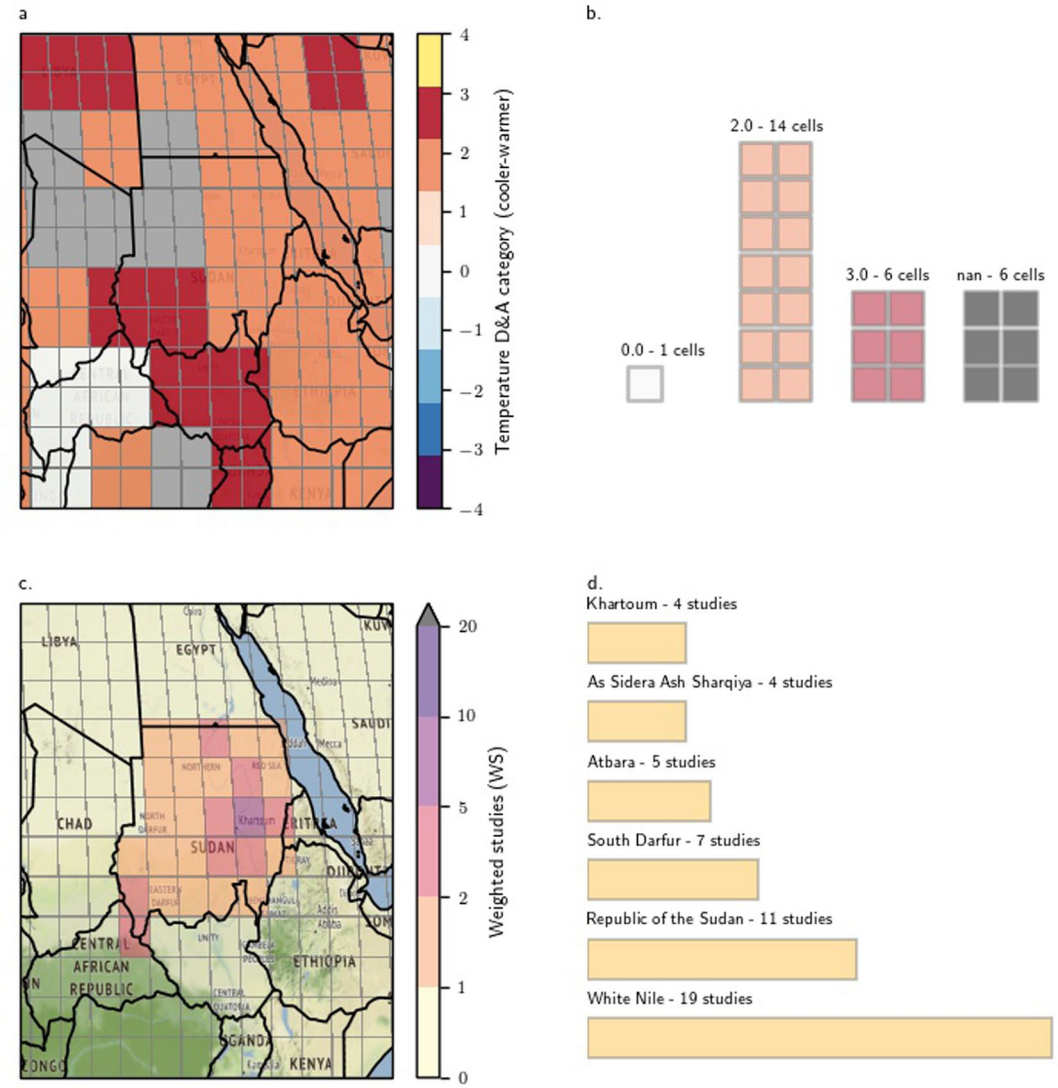
Extended Data Fig. 6 | Geographical distribution of surface trends. Temperature from 1951 to 2018 (left) and precipitation trends from 1951 to 2016 (right) in **(a),(b)** observations and **(c),(d)** CMIP6 10-model ensemble mean all-forcing runs. Bottom panels **(e),(f)** show observations categorised into attribution categories, following refs. [8,7], respectively. Observed cooling/warming or drying/wetting trends that–after accounting for internal climate variability–are inconsistent with the simulated response to natural forcings but consistent with the simulated response to both natural and anthropogenic forcings are indicated by categories -/+2. This is clearest case of changes that are at least partially attributable to anthropogenic forcing, according to the CMIP6 ensemble. Categories -/+1 have detectable observed changes, but are not assessed as attributable to anthropogenic forcing because the observed changes are significantly less than those simulated in the average all-forcing runs. Categories -/+3 have detectable changes and are assessed as at least partly attributable anthropogenic forcing, although the observed changes are inconsistent with the all-forcing runs. That is, they are in the same direction as, but are significantly stronger than, the mean of the all-forcing runs. Categories -/+4 represents cooling/warming or drying/wetting trends that are inconsistent with the simulated response to natural forcings but whose sign is opposite to that of the average simulated all-forcing response; category 0 represents trends that are not distinguishable from natural variability alone. Categories -/+4 and 0 are considered to be examples of non-detectable trends).



Extended Data Fig. 7 | Fractional difference between average CMIP6 modeled low-frequency standard deviation of annual mean precipitation vs observed precipitation. To estimate the internal low-frequency variability for both models and observations, the observed time series were detrended and low-pass filtered with a 7-year running mean filter prior to computing the standard deviations while for the models we used the full available control runs (7-yr running mean filtered) to estimate the internal low-frequency variability for each model. The top panel shows the multi-model ensemble standard deviation comparison while the ten individual panels below it show the comparison for each individual CMIP6 model used in the study. The fraction difference was computed as: [(Model st. dev. - Observed st. dev.) / (Observed st. dev.)].



Extended Data Fig. 8 | Difference between average CMIP6 modeled low-frequency standard deviation (°C) of annual mean surface air temperature vs observed surface temperature. To estimate the internal low-frequency variability for both models and observations, the observed time series were detrended and low-pass filtered with a 7-year running mean filter prior to computing the standard deviations while for the models we used the full available control runs (7-year running mean filtered) to estimate the internal low-frequency variability for each model. The top panel shows the multi-model ensemble standard deviation comparison while the ten individual panels below it show the comparison for each individual CMIP6 model used in the study.



Extended Data Fig. 9 | An illustration of the spatial resolution and weighting methodology. Detection and attribution categories for temperature in East Africa; **b**. the number of grid cells of each type in Sudan; **c**. weighted studies for each grid cell in Sudan; **d**. The number of studies referring to each extracted geographical location in Sudan.