PNAS RESEARCH ARTICLE | EARTH, ATMOSPHERIC, AND PLANETARY SCIENCES OPEN ACCESS

Data-driven predictions of the time remaining until critical global warming thresholds are reached

Noah S. Diffenbaugh[a,1] and Elizabeth A. Barnes[b]

Edited by Michael Mann, The Pennsylvania State University, University Park, PA; received April 25, 2022; accepted December 14, 2022

Leveraging artificial neural networks (ANNs) trained on climate model output, we use the spatial pattern of historical temperature observations to predict the time until critical global warming thresholds are reached. Although no observations are used during the training, validation, or testing, the ANNs accurately predict the timing of historical global warming from maps of historical annual temperature. The central estimate for the 1.5 °C global warming threshold is between 2033 and 2035, including a ±1σ range of 2028 to 2039 in the Intermediate (SSP2-4.5) climate forcing scenario, consistent with previous assessments. However, our data-driven approach also suggests a substantial probability of exceeding the 2 °C threshold even in the Low (SSP1-2.6) climate forcing scenario. While there are limitations to our approach, our results suggest a higher likelihood of reaching 2 °C in the Low scenario than indicated in some previous assessments—though the possibility that 2 °C could be avoided is not ruled out. Explainable AI methods reveal that the ANNs focus on particular geographic regions to predict the time until the global threshold is reached. Our framework provides a unique, data-driven approach for quantifying the signal of climate change in historical observations and for constraining the uncertainty in climate model projections. Given the substantial existing evidence of accelerating risks to natural and human systems at 1.5 °C and 2 °C, our results provide further evidence for high-impact climate change over the next three decades.

global warming | UN Paris agreement | machine learning | AI for climate | CMIP6

Significance

The United Nations Paris Agreement aims to hold global warming well below 2 °C and pursue 1.5 °C. Given the clear evidence for accelerating climate impacts, the time remaining until these global thresholds are reached is a topic of considerable interest. We use machine learning methods to make truly out-of-sample predictions of that timing, based on the spatial pattern of historical temperature observations. Our results confirm that global warming is already on the verge of crossing the 1.5 °C threshold, even if the climate forcing pathway is substantially reduced in the near-term. Our predictions also suggest that even with substantial greenhouse gas mitigation, there is still a possibility of failing to hold global warming below the 2 °C threshold.

The United Nations Paris Agreement articulates the goal of "holding the increase in the global average temperature to well below 2 °C above preindustrial levels and pursuing efforts to limit the temperature increase to 1.5 °C above preindustrial levels" (1). While these global temperatures may not represent absolute physical thresholds, they are relevant for a broad range of climate risks (e.g., refs. 2–6), including impacts on human health (e.g., refs. 2, 6, and 7), economic growth (e.g., ref. 8), crop yields (e.g., refs. 2 and 9), coastal and small island communities (e.g., refs. 2, 5, and 10), terrestrial and marine ecosystems (e.g., refs. 2, 6, and 11), and the frequency, intensity, and cost of extreme climate events (e.g., refs. 12–16).

Given their policy relevance, and the strong scientific evidence for accelerating impacts, the time remaining until these global warming thresholds are reached has generated considerable interest in the scientific literature (e.g., refs.2 and 17), the policy community (e.g., ref. 18), and the public discourse (e.g., ref. 19). The uncertainty of this timing is particularly important, both for understanding the response of the climate system to external forcing and for a suite of climate risk management decisions—including mitigation decisions that depend on the pace of climate stabilization in response to decarbonization, and adaptation decisions that depend on the rate and magnitude of change in the regional and local climate.

There has already been substantial analysis of the likely timing of the 1.5 °C and 2 °C thresholds (e.g., refs. 2, 17, and 20). For example, the Intergovernmental Panel on Climate Change (IPCC) Special Report on 1.5 °C (SR1.5) concluded that "Global warming is likely to reach 1.5 °C between 2030 and 2052 if it continues to increase at the current rate" (2). More recently, the IPCC Sixth Assessment Report (AR6) concluded that "In all scenarios assessed here except SSP5-8.5, the central estimate of crossing the 1.5 °C threshold lies in the early 2030s, about 10 years earlier than the midpoint of the likely range (2030 to 2052) assessed in the SR1.5, which assumed continuation of the then-current warming rate." (17). The "early 2030s" timescale for crossing 1.5 °C is consistent with current estimates based on linear extrapolation of the recent global temperature trend (21). Further, although the IPCC SR1.5 concluded that emissions to date have not been sufficient to cause warming of 1.5 °C on their own (2), the relationship between cumulative emissions and global temperature change implies that one additional decade at the current

Author affiliations: [a]Doerr School of Sustainability, Stanford University, Stanford, CA 94305; and [b]Department of Atmospheric Science, Colorado State University, Fort Collins, CO 80523

Author contributions: N.S.D. and E.A.B. designed research; N.S.D. and E.A.B. performed research; E.A.B. contributed new reagents/analytic tools; N.S.D. and E.A.B. analyzed data; and N.S.D. and E.A.B. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: diffenbaugh@stanford.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2207183120/-/DCSupplemental.

Published January 30, 2023.

rate of annual emissions would be sufficient to create a one-third likelihood of exceeding the 1.5 °C threshold (20).

There is greater uncertainty in the level of cumulative emissions that will cause 2 °C of global warming—and hence greater sensitivity to the future forcing scenario and greater uncertainty in the time to reach the 2 °C threshold (20). However, the IPCC projects that high and extremely high forcing scenarios are likely and very likely, respectively, to cause 2 °C of global warming during the mid-21st century, while low and extremely low forcing scenarios are unlikely and extremely unlikely, respectively, to cause 2 °C of warming by the end of the 21st century (20).

In this study, we develop a data-driven approach to predict the time until the 1.5 °C and 2 °C thresholds ("time-to-threshold"). Our predictions use the spatial pattern of historical temperature observations as input to artificial neural networks (ANNs) that have been trained only on climate model simulations. Setting up the problem as a machine learning prediction task provides the opportunity to make truly out-of-sample predictions on the observations and to directly incorporate uncertainties into the prediction—providing a unique, data-driven approach to constraining climate model uncertainty. In addition, as discussed by Barnes et al. (22), this ANN framework is able to separate signal from noise without having to assume stationarity, detrend the observations, or rely on long control simulations to calculate the internal variability, thereby providing a unique and independent pathway for quantifying the signal of climate change in historical observations. Finally, to perform well, the ANN identifies nonlinear, regional temperature patterns that serve as the most reliable indicators of the time-to-threshold across climate models. We use explainable AI methods (XAI) to visualize these indicator patterns and assess the prediction process of the ANN (23–27).

Results

Our analysis combines global climate models, machine learning, and historical climate observations to predict the expected time until specific global warming thresholds will be reached under different climate forcing scenarios. Our framework consists of four elements: i) identifying the time at which the ensemble-mean forced response reaches the global warming threshold for each climate model in each forcing scenario; ii) training, validating, and testing an ANN for each global warming threshold in each forcing scenario using only the simulated maps of annual temperature anomalies and the simulated time until the global warming threshold is reached; iii) using the observed maps of historical annual temperature anomalies to predict the number of years until each global warming threshold will be reached in each forcing scenario; and iv) using explainable artificial intelligence (XAI) methods to understand the areas of the globe that are most important for the ANN's prediction. (See *Materials and Methods* for further details.)

Like the broader Coupled Model Intercomparison Project (CMIP6) ensemble (17), the subset of climate models that archive at least 10 realizations (*SI Appendix*, Table S1) exhibits a wide range of warming rates in the "High" (SSP3-7.0), "Intermediate" (SSP2-4.5), and "Low" (SSP1-2.6) climate forcing scenarios (Fig. 1*A*). However, although the rate of global-scale warming varies across the climate models, the magnitude and spatial pattern of temperature change is generally consistent for the different climate models at the time that a given global threshold is reached (e.g., Fig. 1*B* and ref. 20). As a result, while the threshold year varies substantially (e.g., 2011 to 2038 and 2021 to 2055 for 1.5 °C and 2 °C, respectively, in the High scenario; Fig. 1*C* and *SI Appendix*, Fig. S1), the ANNs are able to accurately predict the time-to-threshold consistently across the climate models (Fig. 1*C* and *SI Appendix*, Figs. S1–S3).

The ANN training, validation, and testing yield a machine learning prediction model with which we can use observed maps of annual temperature anomalies to predict the time-to-threshold for different global temperatures under different climate forcing scenarios. We begin by using the observed temperature maps to predict the time to the 1.1 °C threshold, which is the approximate current level of global warming (20). In the High scenario, we find that the predicted time-to-threshold for the 2021 observed temperature map is 1 y (with a ± 1σ range of −5 to +5 y), meaning that the predicted year for reaching a forced global temperature response of 1.1 °C is the year 2022 (2017 to 2027) (Fig. 2). Our central estimates for the time to the 1.1 °C threshold are also 0 to 1 y in the Low and Intermediate scenarios (*SI Appendix*, Figs. S4 and S5). In addition, we find that the slope of predicted time to 1.1 °C over the past 15 y has been very close to −1 y/year (Fig. 2 and *SI Appendix*, Figs. S4 and S5), documenting a steady march toward 1.1 °C consistent with the patterns learned by the ANN under the historical forcing simulations. The most prominent periods of variability in the 1.1 °C time-to-threshold time series are the period following the eruption of Mt. Pinatubo in 1991 and the 2010 to 2016 period that encompassed the so-called "global warming hiatus" (28) followed by the warmest year on record (Fig. 2).

Given the high fidelity of the out-of-sample, observations-based prediction of the time to the current level of global warming, we use our framework to predict the time to the 1.5 °C and 2 °C thresholds. For 1.5 °C, the observed pattern of annual temperature anomalies in 2021 leads to a predicted time-to-threshold of 2035 (2030 to 2040) in the High scenario, 2033 (2028 to 2039) in the Intermediate scenario, and 2033 (2026 to 2041) in the Low scenario (Fig. 3). For 2 °C, the observed pattern of annual temperature anomalies in 2021 leads to a predicted time-to-threshold of 2050 (2043 to 2058) in the High scenario, 2049 (2043 to 2055) in the Intermediate scenario, and 2054 (2044 to 2065) in the Low scenario. The slope over the past 15 y has been close to −1 y/year for both the 1.5 °C and 2 °C thresholds in the High scenario (Fig. 3), with steeper slopes in the Intermediate and Low scenarios (*SI Appendix*, Figs. S4 and S5). The Low scenario exhibits the greatest uncertainty in the time-to-threshold (Fig. 3), consistent with its lower signal-to-noise ratio. (These comparisons of temperature thresholds and forcing scenarios are generally robust across a range of random seeds, model realizations, and observational products; *SI Appendix*, Figs. S6–S8.)

We use an XAI attribution method [gradient multiplied by input; (29, 30)] to identify the most important regions of the globe for the ANN's prediction. This analysis reveals a number of areas that contributed to a shorter predicted time-to-threshold in both the ANN training and out-of-sample prediction (Fig. 4). These include continental regions such as areas of the Tibetan Plateau, western North America, and the Mediterranean region and northern Africa. They also include oceanic regions such as areas off the coast of southwestern Africa, southeastern Australia, and the Maritime Continent, along with areas of the Indian Ocean and the Southern Ocean. In addition, there are a number of areas that contributed to a longer predicted time-to-threshold in both the ANN training and out-of-sample prediction, such as areas of the Indian Peninsula, the Gulf of Guinea, and northern Australia.

Discussion

Our framework offers a number of unique and independent perspectives relative to previous estimates of the timing of global warming thresholds. First and foremost, our data-driven

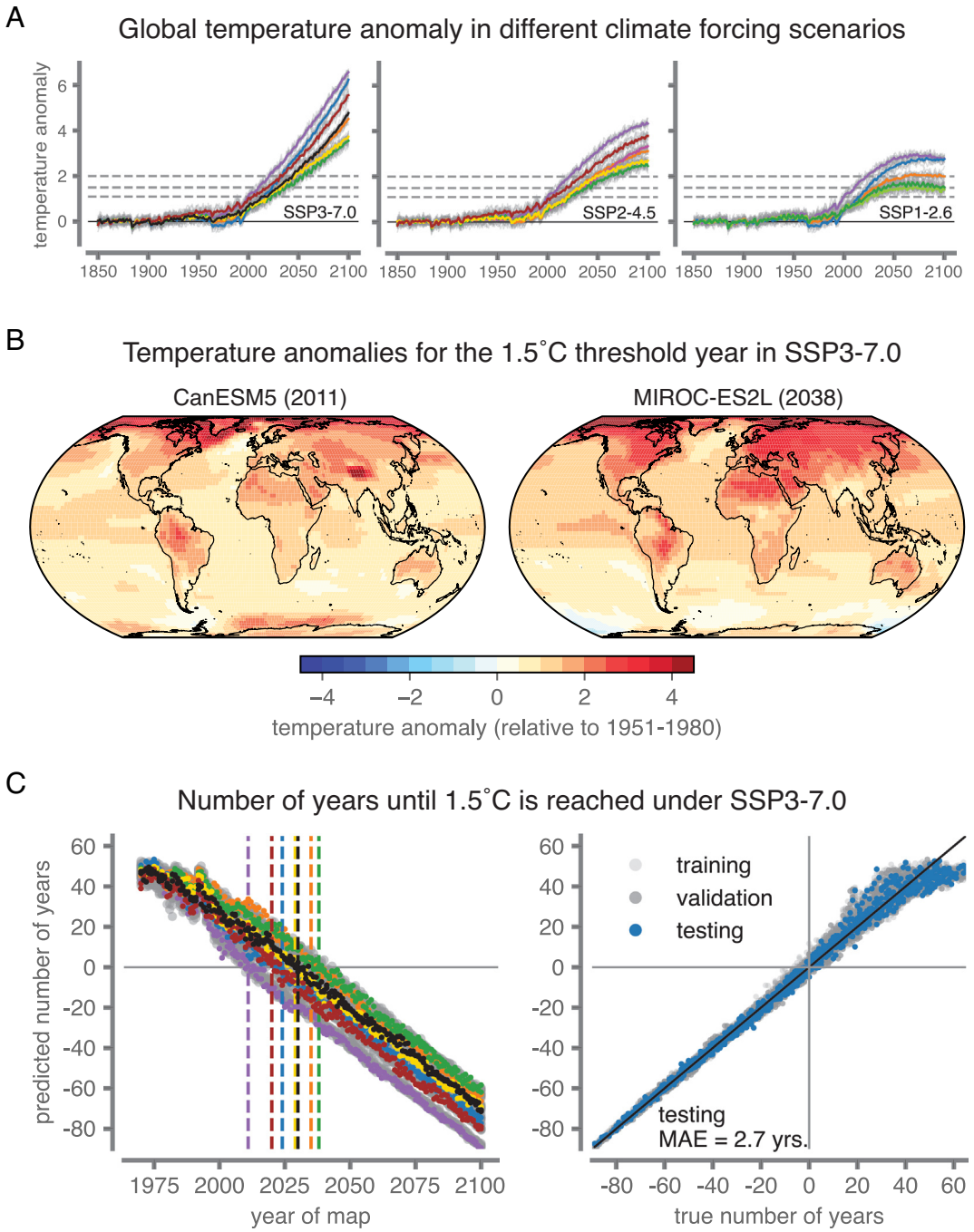

Fig. 1. Time to global warming thresholds in global climate model ensembles. (*A*) Global temperature change relative to the preindustrial baseline (1850 to 1899) for 10-member global climate model ensembles in the High (SSP3-7.0),Intermediate (SSP2-4.5) and Low (SSP1-2.6) climate forcing scenarios. Gray lines show individual realizations; colors show the mean of the respective 10 realizations for each global climate model. See *SI Appendix*, Table S1 for list of climate models used in each climate forcing scenario. (*B*) Maps of temperature anomalies for the "threshold year" (i.e., the year in which the ensemble-mean global warming reaches 1.5 °C) for the global climate models with the earliest and latest threshold years in SSP3-7.0. Anomalies are shown relative to the 1951 to 1980 baseline to match the baseline period of the temperature observations (see *Materials and Methods*). (*C*) Comparison of training, validation, and testing of the artificial neural network (ANN) trained on maps of annual temperature and a global warming threshold of 1.5 °C in SSP3-7.0. *Left* panel shows the predicted number of years until the 1.5 °C threshold for each annual temperature map in each global climate model. *Right* panel shows the comparison of training, validation, and testing for the predicted versus true number of years until the 1.5 °C threshold across the full global climate model ensemble (*SI Appendix*, Table S1). See *SI Appendix*, Figs. S1–S3 for additional temperature thresholds and scenarios.

approach predicts the time-to-threshold (including uncertainty) using the observed pattern of annual temperature anomalies as input to an ANN trained only on simulated temperatures. Our approach is thus distinct from extrapolating the recent observed global warming trend or quantifying the time at which different thresholds are reached across ensembles of climate model scenarios.

Despite these distinctions, our central estimates of the time until the 1.5 °C threshold (2033 to 2035, depending on the scenario; Fig. 3) are consistent with the IPCC's assessment [central estimate "in the early 2030s"; (20)]. Our estimates for 1.5 °C are also consistent with both extrapolation from the recent global temperature trend [central estimate of 2034; (21)] and more sophisticated filtering methods [which yield an estimate of 0.24 °C

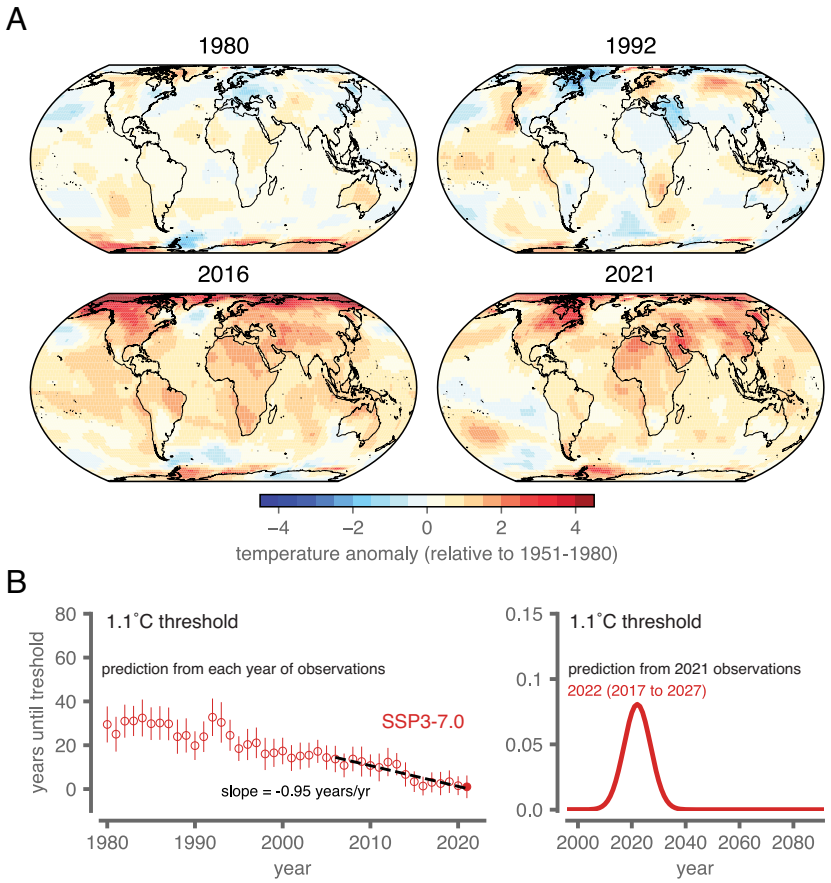

Fig. 2. Time to the current level of global warming predicted from observed maps of annual temperature anomalies. (*A*) Maps of observed annual temperature anomalies for selected individual years, including the first year of our observations-based prediction (1980), the year following the Pinatubo volcanic eruption (1992), the year with the highest global-mean temperature (2016), and the most recent year for which annual data are available (2021). (*B*) The time to 1.1 °C of global warming predicted from the observed map of annual temperature anomalies, using the artificial neural network (ANN) trained on a global warming threshold of 1.1 °C in the High climate forcing scenario (SSP3-7.0). *Left* panel shows the median prediction (and ±1σ range) for the observed map of annual temperature anomalies in each year from 1970 to 2021. The slope quantifies the rate of change of predicted time to 1.1 °C (with a perfect prediction exhibiting a slope of −1 y per year). The *Right* panel shows the distribution of predicted years in which 1.1 °C will be reached based on the observed map of annual temperature anomalies in 2021. Note that no historical temperature observations are used in training, validating, or testing the ANN.

of warming in the past decade; (31)]. Because our prediction is focused on the forced response, on which the actual annual global temperature variability will be overlain, our central estimate for 1.5 °C is also broadly consistent with the IPCC assessment that, by 2030, global temperature "in any individual year could exceed 1.5 °C relative to 1850 to 1900 with a likelihood between 40% and 60%" (17). The fact that our predictions for the 1.5 °C threshold are quite similar to these other approaches, and that the ANN is highly accurate in predicting the time to the current level of global warming (despite not being given any observations as input during training, validation, or testing) should increase confidence in our results—and the expectation that global warming will reach 1.5 °C in the next 1 to 2 decades.

Agreement with the IPCC assessment is more equivocal for the 2 °C threshold. On the one hand, our time-to-threshold estimates for 2 °C of 2050 (± 1σ: 2043 to 2058) in the High scenario and 2054 ((± 1σ: 2044 to 2065) in the Low scenario (Fig. 3) are not inconsistent with the IPCC AR6's evaluation of the CMIP6 ensemble for 2041 to 2060 in the High scenario (mean: 2.3 °C; 5 to 95%: 1.6 to 3.2 °C) and Low scenario (mean: 1.9 °C; 5 to 95%: 1.2 to 2.7 °C), respectively (Table 4.2 in ref. 17). In addition, the IPCC reports the multimodel GCM average global warming for 2081 to 2100 to be 2 °C in the Low scenario (17), and a number of GCMs show peak warming prior to the end of the century in the Low scenario—including well above 2 °C (*SI Appendix*, Table S2 and ref. 17). However, the IPCC AR6 also concludes that 2 °C of global warming is "unlikely" to be crossed during the 21st century in the Low forcing scenario (17). This apparent difference results at least in part from the fact that the IPCC AR6 synthesis assessment is "explicitly constructed by combining scenario-based projections with observational constraints based on past simulated warming, as well as an updated assessment of equilibrium climate sensitivity (ECS) and transient climate response (TCR)" (17).

It is important to emphasize that not all of the climate models reach 2 °C in the Low forcing scenario (Fig. 1 and ref. 17). Hence, one concern about our prediction of a very high likelihood of crossing the 2 °C threshold in the Low scenario could be that for this particular prediction, the ANN is only trained on that subset of the climate model ensemble that reaches 2 °C (*SI Appendix*, Table S1). We therefore conduct multiple analyses to test the possibility that this limitation biases our results. First, we make out-of-sample predictions using the maps of annual temperature from climate models that do not reach 2 °C in the Low scenario (*SI Appendix*, Fig. S9). These out-of-sample tests suggest that even when trained only on climate models that do reach 2 °C in the Low scenario, the ANN is able to identify out-of-sample lower-warming climate trajectories. However, based on the simulated maps of annual temperature in the early 2020s, the ANN still incorrectly predicts that the 2 °C threshold will be reached by the

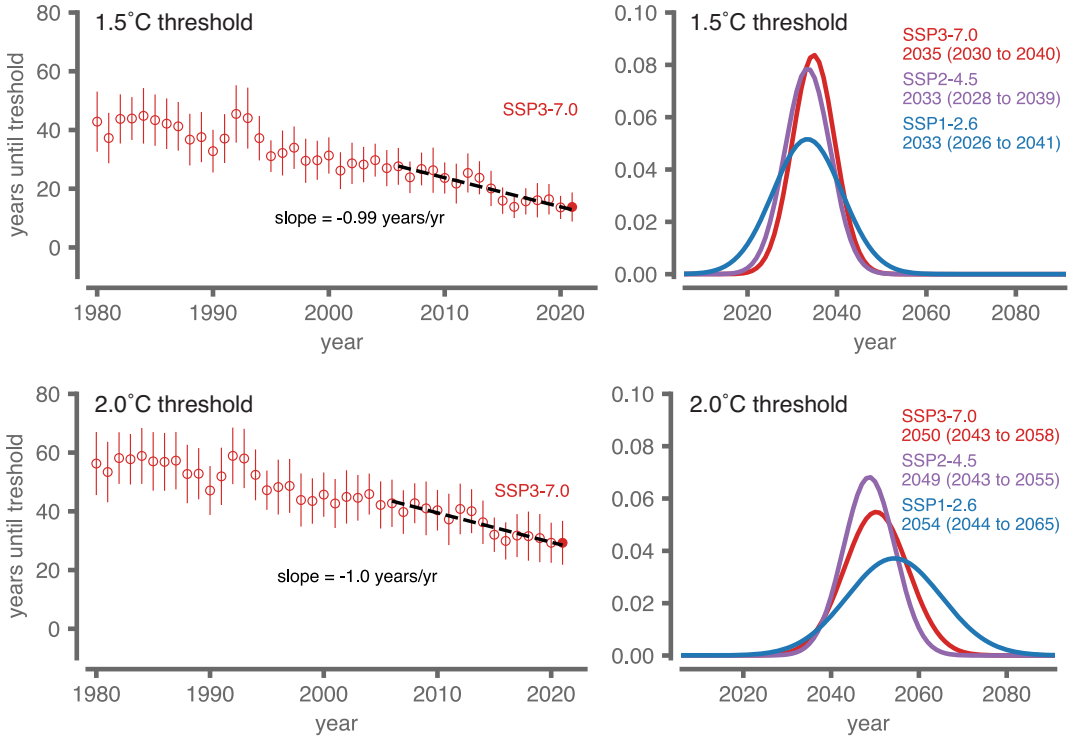

Fig. 3. Time to future global warming thresholds predicted from observed maps of annual temperature anomalies. As in Fig. 2*C*, but for the 1.5 °C and 2 °C thresholds in the High (SSP3-7.0), Intermediate (SSP2-4.5), and Low (SSP1-2.6) climate forcing scenarios. The global climate models used for each scenario/threshold combination are shown in *SI Appendix*, Table S1.

end of the 21st century in all three of these climate models. Therefore, while our main analysis suggests a higher likelihood that 2 °C will be reached under the Low scenario compared with the IPCC AR6 synthesis assessment, it does not rule out the possibility of avoiding 2 °C if the Low scenario is achieved.

The concern about only training on GCMs that reach the global temperature threshold is sharpened by the "hot model problem" (32), in which the CMIP6 ensemble contains a number of GCMs that are highly sensitive to external forcings. Therefore, as a second test, we quantify the sensitivity of our ANN prediction to the removal of the "hottest" GCMs (as measured by the highest TCR reported in ref. 33). For this test, we use all climate models that have archived at least 5 realizations in the Low scenario (*SI Appendix*, Table S2), yielding an ensemble that consists of more GCMs but fewer realizations of each GCM. The range of peak, forced global-mean warming in these climate models spans 1.6 °C to 3.0 °C in the 21st century of the Low scenario (*SI Appendix*, Fig. S10 and Table S2), allowing us to maximize the range of warming that is sampled while still capturing the effects of internal variability. To test the sensitivity of our results to the inclusion of "hot" GCMs, we repeat our analysis of the 1.5 °C threshold using ANNs trained on subsets of this ensemble. First, we find that the ANN trained on this ensemble yields a similar prediction for the time to the 1.5 °C threshold as our main result for the Low scenario, though with a slightly earlier central estimate (2031) and greater likelihood of earlier time-to-threshold (Fig. 2 and *SI Appendix*, Table S2). Second, we find that the exclusion of the 3 highest-TCR GCMs [which all have TCR of at least 2.3 (33)] delays the central estimate by 4 y, while the exclusion of the 5 highest-TCR GCMs [which all have TCR equal to or greater than the CMIP6 multimodel mean of 2.0 °C (33)] delays the central estimate by an additional 2 y (*SI Appendix*, Table S2). These results suggest that (i) the ANN predictions based on recent observed surface temperature maps are not highly sensitive to the inclusion of the hottest GCMs, and (ii) our estimate of a time-to-threshold of 2054 for the 2 °C threshold in the Low scenario (Fig. 2) is likely slightly earlier with the inclusion of high-TCR GCMs.

However, this second test is limited by the fact that it is—by necessity—focused on the time to the 1.5 °C threshold. Thus, as a third test, we train a separate ANN to ingest maps of annual-mean surface temperature anomalies and predict the amount of additional warming until the peak forced global-mean temperature is reached, again using all climate models that have archived at least 5 realizations in the Low scenario (*SI Appendix*, Fig. S10). Based on the map of observed annual temperature anomalies in 2021, the mean prediction is >1 °C of additional global warming under the Low scenario (in addition to what has already occurred up until 2021). While it yields uncertainty in the predicted amount of peak warming, this independent analysis supports the conclusion of a substantial likelihood that 2 °C could be reached under the Low scenario.

Another distinct feature of our framework is that the ANN focuses on specific regions to predict the time-to-threshold (Fig. 4 and *SI Appendix*, Figs. S11–S14). Some key regions appear to be the Indian Ocean, Tibetan Plateau, and western North America (Fig. 4). These regions can be interpreted as robust indicators of the time-to-threshold, as identified by the ANN, and are a complex combination of where the signal-to-noise ratio is large, the climate models generally agree on the patterns of change, and the relationships across different regions (i.e., grid points) can be leveraged. The Indian Ocean is a well-studied indicator of anthropogenic change since the signal-to-noise ratio there is relatively large (34). The ANN appears to be leveraging this region in its time-to-threshold prediction for both the climate models (Fig. 4*A*) and the observations (Fig. 4*B*). The contrasting signs of the XAI explanation over southwestern Africa and the adjacent ocean suggest that the ANN may be using land-sea contrast in this region

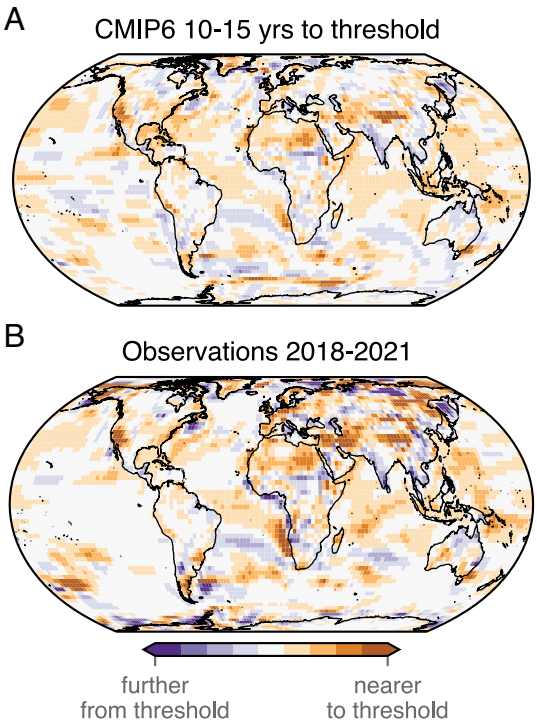

Fig. 4. Attribution heatmap of the most relevant regions for the artificial neural network (ANN) prediction of the time-to-threshold. (*A*) Global climate model ensemble. (*B*) Historical temperature observations for 2018 to 2021. Warm colors indicate shorter time to global warming threshold; cool colors indicate longer time to global warming threshold. Maps show results for the 1.5 °C threshold in the High (SSP3-7.0) climate forcing scenario. See *Materials and Methods* for additional details of the heatmap calculation and *SI Appendix*, Figs. S12 and S13 for additional scenarios.

when making its prediction. The lack of color over the eastern equatorial Pacific in Fig. 4 is equally notable, as it demonstrates that the ANN has learned to avoid using this region for its prediction. This can be explained by the fact that this is a region with large internal variability driven in large part by the El Nino Southern Oscillation. Thus, the network learns that it is not a reliable region for assessing the time-to-threshold.

When considering the areas identified in the XAI results, it is important to emphasize that these are regions that are most relevant for the ANN's prediction of the time until the forced response is reached. As a result, the XAI pattern does not mirror the spatial pattern of the mean forced response (17) nor the spatial pattern of historical warming (20). The spatial pattern of warming is a critical influence on climate sensitivity (35, 36), and various calculations of the climate sensitivity are generally reduced when using the pattern of observed historical warming (36–40). Some of this "pattern effect" is due to the spatial pattern of the response to the combination of historical forcings (37), including non-CO_2 forcings (38, 41). In addition, a substantial fraction of the pattern effect is likely "unforced" (i.e., due to internal climate variability) (37, 40). The combination of forced and unforced pattern effects results in reduced calculation of climate sensitivity from historical temperature observations (37).

Our ANN-based predictions using maps of historical temperature observations suggest that the 1.5 °C and 2 °C thresholds will arrive later than projected by the highest sensitivity climate models but sooner than projected by the lowest sensitivity climate models (Figs. 1–3 and *SI Appendix*, Fig. S10 and Table S2). In the case of our analysis, the ANN is "learning" which regions matter most for accurately predicting the time until a given forced response is reached, based on the patterns across multiple climate models. This ANN approach is distinct from learning the climate sensitivity via the pattern of warming. Hence, while our predictions use the observed maps of annual temperature anomalies as out-of-sample inputs, the observed pattern of long-term historical temperature change is not the indicator of the time-to-threshold. The relatively short time-to-thresholds predicted from recent annual temperature observations (e.g., Fig. 3) are a reflection of the regions the ANN has identified as being critical for the time until the forced response is reached (e.g., Fig. 4), even if the full spatial pattern of observed warming implies lower climate sensitivity (e.g., ref. 37).

With that said, a data-driven approach that learns from climate models necessarily requires that the models have some truth in their representation of the real world. Clearly, the ANN is only capable of learning from the data it is trained on and therefore may learn errors and biases present in the climate model simulations. We partially alleviate this issue by training over multiple climate models such that the ANN must learn reliable indicators of the time-to-threshold that apply across the climate models. Thus, it is not necessary that all of the patterns of variability and forced response are well-represented across the entire ensemble of climate models in order for the ANN to learn the reliable indicators, as long as biases are not common across all the climate models.

The fact that the framework provides a reasonable estimate for reaching the 1.1°C threshold (Fig. 2 and *SI Appendix*, Figs. S4 and S5) across multiple observational products (*SI Appendix*, Fig. S7) and multiple forcing scenarios (and hence multiple climate model combinations; *SI Appendix*, Table S1) provides support that the ANN is learning relationships that are applicable and relevant to the real world. This is nicely illustrated in Fig. 3: On the one hand, the ANN is not sensitive to the observed strong El Nino event of 1997/1998 since it has apparently learned that specific patterns of internal climate variability are not reliable indicators of the time-to-threshold. In contrast, following the Mt. Pinatubo eruption in 1991, the ANN predicts that the time-to-threshold is approximately 10 y later than in previous years; this reflects the cooling of the global surface temperatures due to volcanic aerosols (42), which the ANN incorrectly interprets as an indication of being further from the threshold temperature (see discussion in ref. 22). Note, however, that as the volcanic forcing weakens from 1991 to ~1995, the network readjusts its predictions to be more in-step with those prior to the eruption.

It is important to note that the specific resulting ANN for a given forcing scenario is a product of which ensemble members are used for training, the random initialization of the model weights, and the choices of the model architecture and hyperparameters. While justifications for our choices are provided in the *Materials and Methods*, here we note that model performance, time-to-threshold, and XAI results are very consistent across different observational products and a range of model choices and random seeds (*SI Appendix*, Figs. S6–S8, S14, S17, S18, and S20). The fact that the ANNs predict larger uncertainty ranges (and slightly larger sensitivity to random seed) under the Low scenario compared to the High and Intermediate scenarios is likely due to the smaller number of climate models included in the Low scenario training set (*SI Appendix*, Table S1), as well as the comparatively smaller forced response throughout the Low scenario simulation.

Finally, in comparing our results with previous assessments, it should be noted that the definition of the global warming threshold may vary across studies. For example, while the IPCC AR6 used a 20-y mean to define the global warming thresholds, we use the forced response across multiple climate model realizations.

Using the forced response enables us to most clearly distinguish the signal of global warming from the noise of internal climate variability. However, although we have included 10 realizations of each climate model in our main analysis (*SI Appendix*, Table S1), our results are still limited by the number of models and realizations that are available. This is especially true for lower forcing scenarios. For example, only 4 of the climate models that archived at least 10 realizations in SSP3-7.0 also archived at least 10 realizations in SSP1-2.6 (*SI Appendix*, Table S1), and the availability of at least 10 realizations in the very low SSP1-1.9 scenario was prohibitively limited. In addition, expanding the number of climate models in SSP1-2.6 to test the sensitivity to climate model TCR requires expanding the analysis to models with only 5 realizations (*SI Appendix*, Table S2). Further refinement of our results will require multiple large single-model ensembles in multiple scenarios, particularly for low and very low forcing trajectories.

Conclusions

The UN Paris Agreement aims to limit the impacts of climate change on natural and human systems by holding global warming below specific temperature thresholds (1). While there have been a number of assessments of the time until those thresholds are likely to be reached, ours uses machine learning methods to make truly out-of-sample predictions based on maps of historical temperature observations. Our results thus provide unique and largely independent predictions (with uncertainty) in different climate forcing scenarios. More generally, this framework also provides a unique, data-driven approach for quantifying the signal of climate change in historical observations and for using observations to constrain the uncertainty of future climate projections.

The fact that our central estimate for the time until 1.5 °C lies between 2033 and 2035 in the High, Intermediate, and Low forcing scenarios confirms that global warming is already on the verge of crossing the 1.5 °C threshold, even if the climate forcing pathway is substantially reduced in the near term. Our predictions also show a high probability of reaching the 2 °C threshold by mid-century in the High, Intermediate, and Low scenarios, suggesting that even with substantial greenhouse gas mitigation, there is still a possibility of failing to achieve the UN Paris goal of holding global warming well below the 2 °C threshold. Given the substantial literature quantifying accelerating risks for natural and human systems at 1.5 °C and 2 °C, our results suggest a high likelihood of high-impact climate change over the next three decades.

Materials and Methods

Datasets. We predict the time until various global warming thresholds are reached ("time-to-threshold") using ANNs trained on temperature anomalies simulated by multiple realizations of multiple global climate models. To do so, we draw on Phase 6 of CMIP6, which archives a large suite of coordinated global climate model simulations from modeling groups around the world (43). We compare the predicted time-to-threshold in three future climate forcing scenarios: SSP3-7.0, SSP2-4.5, and SSP1-2.6. These scenarios are identified as High, Intermediate, and Low forcing scenarios, respectively, by the IPCC (20) and include the two core scenarios used by the IPCC in projecting future climate change (17) and synthesizing climate-related risks (4).

To ensure a substantial sampling of internal climate variability for each climate model, we include CMIP6 models that have archived at least 10 realizations in a given scenario (*SI Appendix*, Table S1). In addition, to ensure equal model weighting within each scenario, we only include 10 realizations of each climate model (even though some have archived more than 10 realizations). Further, to maximize the number of 10-member ensembles in our analysis, we also include 10 realizations of SSP3-7.0 from the NCAR CESM2 large ensemble project (44, 45). (For CESM2, we use the smoothed biomass burning experiments, which we label as "CESM2-LE2-smbb".)

We use observed maps of annual temperature anomalies to make out-of-sample predictions of the time-to-threshold (see below). Our primary observational dataset is the Berkeley Earth Surface Temperature ("Berkeley") dataset (46). Berkeley provides monthly temperature anomalies from the 1951 to 1980 climatology on a 1° × 1° geographical grid. For both the climate models and observational data, we calculate annual-mean anomalies relative to the 1951 to 1980 climatology at each grid point. We then regrid the annual-mean anomalies to the common 2.5° × 2.5° geographical grid used by the IPCC (47).

To test the robustness to observational uncertainty, we repeat the time-to-threshold predictions using the NASA GISTEMP global gridded temperature dataset (48), along with the ERA5 (49) and NCEP/NCAR R1 (50) reanalysis products. These datasets exhibit similar global-scale temperature anomalies relative to the 1951 to 1980 baseline (*SI Appendix*, Fig. S15). We find that the results are robust for Berkeley, GISTEMP, and ERA5, with NCEP/NCAR R1 a clear outlier (*SI Appendix*, Figs. S7 and S8).

Global Warming Thresholds. There are a variety of possible definitions of a global temperature "threshold", including the time at which the multidecadal mean exceeds a temperature value (e.g., the IPCC uses a 20-y mean; ref. 17) or the first/last year that the annual temperature rises above/dips below a temperature value (e.g., the IPCC assesses the probability of individual years exceeding 1.5 °C; ref. 17). Therefore, the time to reach a particular threshold can vary substantially depending on which definition is used. In order to clearly distinguish the temperature response to external forcing from the noise of internal climate variability, we define the temperature threshold as the year in which the forced response (represented by the climate model ensemble mean) reaches the global warming threshold. We calculate the respective forced response for each climate model in a given climate forcing scenario. Note that the forced response in observations is much harder to quantify, given that we only have one realization of the real world. However, our machine learning approach circumvents this issue by learning reliable indicators across a range of climate models and thus does not require the observed forced response to be directly computed (23, 24).

To test the sensitivity of our predicted time-to-threshold to the definition of the temperature threshold, we repeat our analysis replacing the forced response with a 15-y smoothing of the ensemble-mean time series (using the scipy.signal.savgol_filter with a window length of 15 y and polynomial of order 3). We find that the results obtained using the ensemble-mean smoothing (*SI Appendix*, Fig. S16) are nearly identical to those obtained using the forced response (Fig. 3).

We predict the time-to-threshold for global warming of 1.1 °C, 1.5 °C, and 2 °C above the preindustrial baseline (1850 to 1899). The 1.5 °C and 2 °C thresholds are derived from the goals stated in the UN Paris Agreement (1). In addition, we use the 1.1 °C threshold to evaluate whether our framework is able to accurately predict the timing of the global warming that has already occurred (based on the IPCC's recent assessment; ref. 20).

ANNs. We train a feed-forward artificial neural network (ANN) to ingest maps of annual-mean temperature anomalies and predict the time-to-threshold (including quantification of uncertainty). The input to the network is a flattened map of 72 latitude by 144 longitude grid points, which results in an input vector of 10,368 units in length. The neural network is composed of two hidden layers with 25 units each and an output layer of two units. These two units represent the mean and SD of a conditional Gaussian distribution, where the network is trained to optimize its weights and biases to minimize the negative log-likelihood (51–53). *SI Appendix*, Fig. S19 shows the probability integral transform of the validation and testing predictions (e.g., refs. 54–56), demonstrating that the uncertainties (SDs) learned by the network are meaningful. We train via backpropagation using an Adam optimizer with a learning rate of 0.00001 and a batch size of 64 and apply early stopping on the validation loss using a patience of 50 epochs. We combat overfitting by applying L_2 regularization (regularization parameter of 10) between the input layer and the first layer. A ReLU activation function is applied throughout the network, except in connection to the output layer, which is linear. The model is coded and trained using Tensorflow 2.7.0, and Tensorflow-Probability 0.15.0.

We conduct a number of analyses to test the sensitivity of our predictions to various methodological choices. The network architecture and hyperparameters were

chosen as those that performed well on the validation data across multiple metrics (*SI Appendix*, Figs. S17 and S18). The specific networks shown in the paper are those with random seed 2247, which sets both the specific ensemble members used for training/validation/testing as well as the random ANN initialization. Although ANN performance varied slightly for different architectures and hyperparameter choices, the observations-based predictions are very consistent, providing further confidence in our results (*SI Appendix*, Fig. S20). In addition, the observations-based predictions are robust across varying random initializations of the network and which ensemble members were used for training (*SI Appendix*, Fig. S6).

Prediction of Time to Global Warming Thresholds Using Observed Maps of Annual Temperature Anomalies as Input to the ANNs. Once an ANN is trained, validated, and tested for a global warming threshold and climate forcing scenario, we can use any map of annual temperature anomalies to make out-of-sample predictions of the time-to-threshold. Therefore, we first train, validate, and test a separate ANN for each warming threshold in each scenario (*SI Appendix*, Figs. S1–S3). We then use the observed maps of annual temperature anomalies for each year from 1980 to 2021 as input to each of those ANNs. This yields a time series of predictions (with uncertainty) of the time-to-threshold for the 1.1 °C, 1.5 °C, and 2 °C thresholds in the High (SSP3-7.0), Intermediate (SSP2-4.5), and Low (SSP1-2.6) climate forcing scenarios (Figs. 2 and 3 and *SI Appendix*, Figs. S4 and S5). We emphasize that the ANN has not been given any observed climate data as input prior to this observations-based prediction, meaning that these are truly out-of-sample predictions. Further, neither the training nor prediction includes the simulated or observed global-mean temperature time series.

We note that different observational products suggest a range of global warming since the preindustrial era, including a "likely" range as high as 1.3 °C (20). Given this uncertainty, we repeat our analysis for a global warming threshold of 1.3 °C and find that the observational products again agree on the time-to-threshold (with NCEP/NCAR R1 again a clear outlier) (*SI Appendix*, Fig. S7 and S8). There are multiple reasons for this close agreement between observational products. One is that our method quantifies the time until the forced response is reached and not the long-term mean of an individual realization of the climate system. Another is that our analysis is based on annual temperature anomalies from the 1951 to 1980 baseline; while there are differences in the magnitude of global warming since the preindustrial between the datasets, there is a close agreement in the annual global temperature anomalies relative to the 1951 to 1980 baseline (which is well measured, especially relative to the preindustrial). NCEP/NCAR R1 shows the greatest difference in regional temperatures from the other datasets (*SI Appendix*, Fig. S15), and hence is the greatest outlier in the predicted time-to-threshold for both 1.1 °C and 1.3 °C (*SI Appendix*, Figs. S7 and S8).

Using XAI to Understand the Spatial Patterns that Are Most Relevant for the Prediction of the Time to Global Warming Thresholds. We apply multiple XAI methods to explore and visualize the spatial patterns in the input temperature maps that are most relevant for the network's prediction of the time-to-threshold. These XAI methods are designed to help understand the decision-making process of the network in order to gauge the trustworthiness in the network's prediction. For a more in-depth discussion of these methods and how they compare for geoscience applications, see refs. 26 and 57. We further use XAI here to explore particular regional patterns in the input maps that led to the network's accurate prediction. *SI Appendix*, Fig. S11 shows the results of Gradient, Gradient*Input, and Integrated Gradients. The Gradient method (58) is a sensitivity method and approximates the local gradient of the output relative to the input and can be interpreted as how a unit increase in the temperatures at each grid point in the input map would impact the prediction. Gradient*Input (29, 30) is an attribution method and approximates the marginal contribution of the input grid point to the prediction. As its name implies, this is computed as the product of the Gradient multiplied by the input map. Integrated Gradients (59) is another attribution method that identifies the contribution of each input grid point to the output starting from a user-defined reference vector. Here, we use a reference vector of zeros. Integrated Gradients is very similar to Gradient*Input except that it attempts to account for nonlinearities in the learned function.

For each input sample of interest, we compute a heatmap for each XAI method which acts as the "explanation" for the network's prediction. We then average the heatmaps together to produce the panels in Fig. 4 and *SI Appendix*, Figs. S11–S13. Our colormap convention is such that orange colors denote regions in the input maps that act to push the network to smaller time-to-threshold predictions (i.e., the input maps are closer to the threshold), while purple colors denote regions in the input maps that act to push the network to larger time-to-threshold predictions (i.e., the input maps are further from the threshold). We find that the regions that are emphasized by the network are very similar between the different observational products (*SI Appendix*, Fig. S14).

Data, Materials, and Software Availability. Code is available on GitHub at https://github.com/eabarnes1010/target_temp_detection (60) and is archived on Zenodo at the following DOI: https://doi.org/10.5281/zenodo.7510551 (61).

ACKNOWLEDGMENTS. We are grateful for insightful comments and feedback during the review and editorial process. We thank Dr. Patrick Keys for the helpful discussions. We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP6. We thank the climate modeling 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. We also thank the National Center for Atmospheric Research (NCAR) for access to the CESM2 large ensemble results (via the ESGF). Computational resources were provided by Stanford's Center for Computational Earth and Environmental Sciences and the Stanford Center for Research Computing, as well as Colorado State's Walter Scott, Jr. College of Engineering Asha Cluster. This research was supported, in part, by Stanford University and by the Regional and Global Model Analysis program area of the US Department of Energy's Office of Biological and Environmental Research as part of the Program for Climate Model Diagnosis and Intercomparison project.

1. UNFCCC Adoption of the Paris Agreement. I: Proposal by the President. Draft Decision CP.21 https://unfccc.int/resource/docs/2015/cop21/eng/l09r01.pdf (United Nations Office, Geneva, 2015)
2. M. Allen, M. Babiker, Y. Chen, H. C. de Coninck, "IPCC SR15: Summary for policymakers" in *IPCC Special Report Global Warming of 1.5 °C* (Intergovernmental Panel on Climate Change, 2018).
3. IPCC, "Summary for policymakers" in *Climate Change 2014: Impacts, Adaptation, and Vulnerability in Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, C. B. Field*et al.*, Eds. (Cambridge University Press, 2014), pp. 1–32.
4. IPCC, "Summary for policy makers" in *Climate Change 2022: Impacts, Adaptation and Vulnerability in Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, H. O. Pörtner*et al.*, Eds. (IPCC, 2022).
5. C.-F. Schleussner *et al.*, Differential climate impacts for policy-relevant limits to global warming: The case of 1.5 °C and 2 °C. *Earth Syst. Dyn.* **7**, 327–351 (2016).
6. O. Hoegh-Guldberg *et al.*, The human imperative of stabilizing global climate change at 1.5°C. *Science* **365**, eaaw6974 (2019).
7. K. L. Ebi *et al.*, Health risks of warming of 1.5 °C, 2 °C, and higher, above pre-industrial temperatures. *Environ. Res. Lett.* **13**, 63007 (2018).
8. M. Burke, W. M. Davis, N. S. Diffenbaugh, Large potential reduction in economic damages under UN mitigation targets. *Nature* **557**, 549–553 (2018).
9. D. B. Lobell *et al.*, Prioritizing climate change adaptation needs for food security in 2030. *Science* **319**, 607–610 (2008).
10. R. J. Nicholls *et al.*, Stabilization of global temperature at 1.5°C and 2.0°C: Implications for coastal areas. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **376**, 20160448 (2018).
11. R. Warren, J. Price, E. Graham, N. Forstenhaeusler, J. VanDerWal, The projected effect on insects, vertebrates, and plants of limiting global warming to 1.5 °C rather than 2 °C. *Science* **360**, 791–795 (2018).
12. S. I. Seneviratne, M. G. Donat, A. J. Pitman, R. Knutti, R. L. Wilby, Allowable CO2 emissions based on regional and impact-related climate targets. *Nature* **529**, 477–483 (2016).
13. N. S. Diffenbaugh, D. Singh, J. S. Mankin, Unprecedented climate events: Historical changes, aspirational targets, and national commitments. *Sci. Adv.* **4**, eaao3354 (2018).
14. E. M. Fischer, S. Sippel, R. Knutti, Increasing probability of record-shattering climate extremes. *Nat. Clim. Chang.* **11**, 689–695 (2021).
15. A. Dosio, L. Mentaschi, E. M. Fischer, K. Wyser, Extreme heat waves under 1.5 °C and 2 °C global warming. *Environ. Res. Lett.* **13**, 54006 (2018).
16. F. V. Davenport, M. Burke, N. S. Diffenbaugh, Contribution of historical precipitation change to US flood damages. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2017524118 (2021).
17. J.-Y. Lee *et al.*,"2021: Future Global Climate: Scenario-Based Projections and Near-Term Information" in *Climate Change 2021: The Physical Science Basis in Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, V. Masson-Delmotte *et al.*, Eds. (Cambridge University Press, Cambridge, UK and New York, NY,2021), pp. 553–672. 10.1017/9781009157896.006.
18. UNEP, Emissions gap report 2021: The heat is on–A world of climate promises not yet delivered" (United Nation Environment Program, 2021).
19. Z. Hausfather, "Analysis: When might the world exceed 1.5C and 2C of global warming?" *Carbon Brief* (2020). https://www.carbonbrief.org/analysis-when-might-the-world-exceed-1-5c-and-2c-of-global-warming/.

20. IPCC, "Summary for policymakers" in *Climate Change 2021: The Physical Science Basis in Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, V. Masson- Delmotte *et al.*, Eds. (Cambridge University Press, Cambridge, UK and New York, NY, 2021).
21. A. Amici, Global Temperature Trend Monitor: User Guide (Copernicus Climate Change Service, ECMWF, 2021).
22. E. A. Barnes, J. W. Hurrell, I. Ebert-Uphoff, C. Anderson, D. Anderson, Viewing forced climate patterns through an AI lens. *Geophys. Res. Lett.* **46**, 13389–13398 (2019).
23. J. K. Rader, E. A. Barnes, I. Ebert-Uphoff, C. Anderson, Detection of forced change within combined climate fields using explainable neural networks. *Earth Sp. Sci. Open Arch.* **36** (2021).
24. E. A. Barnes *et al.*, Indicator patterns of forced change learned by an artificial neural network. *J. Adv. Model. Earth Syst.* **12**, e2020MS002195 (2020).
25. Z. M. Labe, E. A. Barnes, Detecting climate signals using explainable AI with single-forcing large ensembles. *J. Adv. Model. Earth Syst.* **13**, e2021MS002464 (2021).
26. A. Mamalakis, I. Ebert-Uphoff, E. A. Barnes, Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. *Environmental Data Science*, **1**, E8 (2022). 10.1017/eds.2022.7
27. I. Ebert-Uphoff, K. Hilburn, Evaluation, tuning, and interpretation of neural networks for working with images in meteorological applications. *Bull. Am. Meteorol. Soc.* **101**, E2149–E2170 (2020).
28. I. Medhaug, M. B. Stolpe, E. M. Fischer, R. Knutti, Reconciling controversies about the 'global warming hiatus'. *Nature* **545**, 41–47 (2017).
29. A. Shrikumar, P. Greenside, A. Kundaje, "Learning important features through propagating activation differences" in *International Conference on Machine Learning* (PMLR, 2017), pp. 3145–3153.
30. A. Shrikumar, P. Greenside, A. Shcherbina, A. Kundaje, Not just a black box: Learning important features through propagating activation differences. arXiv [Preprint] (2016). https://arxiv.org/abs/1605.01713 Accessed 6 November 2022.
31. B. H. Samset *et al.*, Earlier emergence of a temperature response to mitigation by filtering annual variability. *Nat. Commun.* **13**, 1578 (2022).
32. Z. Hausfather, K. Marvel, G. A. Schmidt, J. W. Nielsen-Gammon, M. Zelinka, Climate simulations: Recognize the 'hot model'problem. *Nature* **605**, 26–29 (2022).
33. G. A. Meehl *et al.*, Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 Earth system models. *Sci. Adv.* **6**, eaba1981 (2020).
34. J. W. Hurrell, M. P. Hoerling, A. S. Phillips, T. Xu, Twentieth century north atlantic climate change. Part I: Assessing determinism. *Clim. Dyn.* **23**, 371–389 (2004).
35. K. C. Armour, C. M. Bitz, G. H. Roe, Time-varying climate sensitivity from regional feedbacks. *J. Clim.* **26**, 4518–4534 (2013).
36. T. Andrews *et al.*, Accounting for changing temperature patterns increases historical estimates of climate sensitivity. *Geophys. Res. Lett.* **45**, 8490–8499 (2018).
37. A. E. Dessler, Potential problems measuring climate sensitivity from the historical record. *J. Clim.* **33**, 2237–2248 (2020).
38. Y. Dong *et al.*, Biased estimates of equilibrium climate sensitivity and transient climate response derived from historical CMIP6 simulations. *Geophys. Res. Lett.* **48**, e2021GL095778 (2021).
39. C. Proistosescu, P. J. Huybers, Slow climate mode reconciles historical and model-based estimates of climate sensitivity. *Sci. Adv.* **3**, e1602821 (2017).
40. K. Marvel, R. Pincus, G. A. Schmidt, R. L. Miller, Internal variability and disequilibrium confound estimates of climate sensitivity from observations. *Geophys. Res. Lett.* **45**, 1595–1601 (2018).
41. K. Marvel, G. A. Schmidt, R. L. Miller, L. S. Nazarenko, Implications for climate sensitivity from the response to individual forcings. *Nat. Clim. Chang.* **6**, 386–389 (2016).
42. A. Robock, Volcanic eruptions and climate. *Rev. Geophys.* **38**, 191–219 (2000).
43. V. Eyring *et al.*, Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.* **9**, 1937–1958 (2016).
44. G. Danabasoglu *et al.*, The community earth system model version 2 (CESM2). *J. Adv. Model. Earth Syst.* **12**, e2019MS001916 (2020).
45. K. B. Rodgers *et al.*, Ubiquity of human-induced changes in climate variability. *Earth Syst. Dyn.* **12**, 1393–1411 (2021).
46. R. A. Rohde, Z. Hausfather, The Berkeley earth land/ocean temperature record. *Earth Syst. Sci. Data Discuss.* (2020), 10.5194/essd-2019-259.
47. IPCC, 2013: Summary for Policymakers in Climate Change 2013: The Physical Science Basis, Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, T. F. Stocker *et al.*, Eds. (Cambridge University Press, Cambridge, New York, USA, 2013).
48. GISTEMP, GISS Surface Temperature Analysis (GISTEMP) (Version 4, NASA Goddard Institute for Space Studies, New York, NY, 2021). Accessed 25 April 2022.
49. H. Hersbach *et al.*, The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).
50. E. Kalnay *et al.*, The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77**, 437–471 (1996).
51. O. Dürr, B. Sick, E. Murina, *Probabilistic Deep Learning: With Python, Keras and Tensorflow Probability* (Manning Publications, 2020).
52. E. A. Barnes, R. J. Barnes, N. Gordillo, Adding uncertainty to neural network regression tasks in the geosciences. arXiv [Preprint] (2021). https://arxiv.org/abs/2109.07250 Accessed 6 November 2022.
53. A. P. Guillaumin, L. Zanna, Stochastic-deep learning parameterization of ocean momentum forcing. *J. Adv. Model. Earth Syst.* **13**, e2021MS002534 (2021).
54. D. R. Bourdin, T. N. Nipen, R. B. Stull, Reliable probabilistic forecasts from an ensemble reservoir inflow forecasting system. *Water Resour. Res.* **50**, 3108–3130 (2014).
55. T. Gneiting, M. Katzfuss, Probabilistic forecasting. *Annu. Rev. Stat. Its Appl.* **1**, 125–151 (2014).
56. A. P. Dawid, Present position and potential developments: Some personal views statistical theory the prequential approach. *J. R. Stat. Soc. Ser. A.* **147**, 278–290 (1984).
57. A. Mamalakis, E. A. Barnes, I. Ebert-Uphoff, Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. *Artificial Intelligence for the Earth Systems* **1**, e220012 (2022).
58. K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv [Preprint] (2013). arXiv1312.6034.
59. M. Sundararajan, A. Taly, Q. Yan, "Axiomatic attribution for deep networks" in *International Conference on Machine Learning* (PMLR, 2017), pp. 3319–3328.
60. E. A. Barnes, Data from "Data-driven predictions of the time remaining until critical global warming thresholds are reached" *Github.* https://github.com/eabarnes1010/target_temp_detection. Deposited 6 January 2023.
61. E. A. Barnes, Data from "Data-driven predictions of the time remaining until critical global warming thresholds are reached" *Zenodo.* https://doi.org/10.5281/zenodo.7510551. Deposited 6 January 2023.